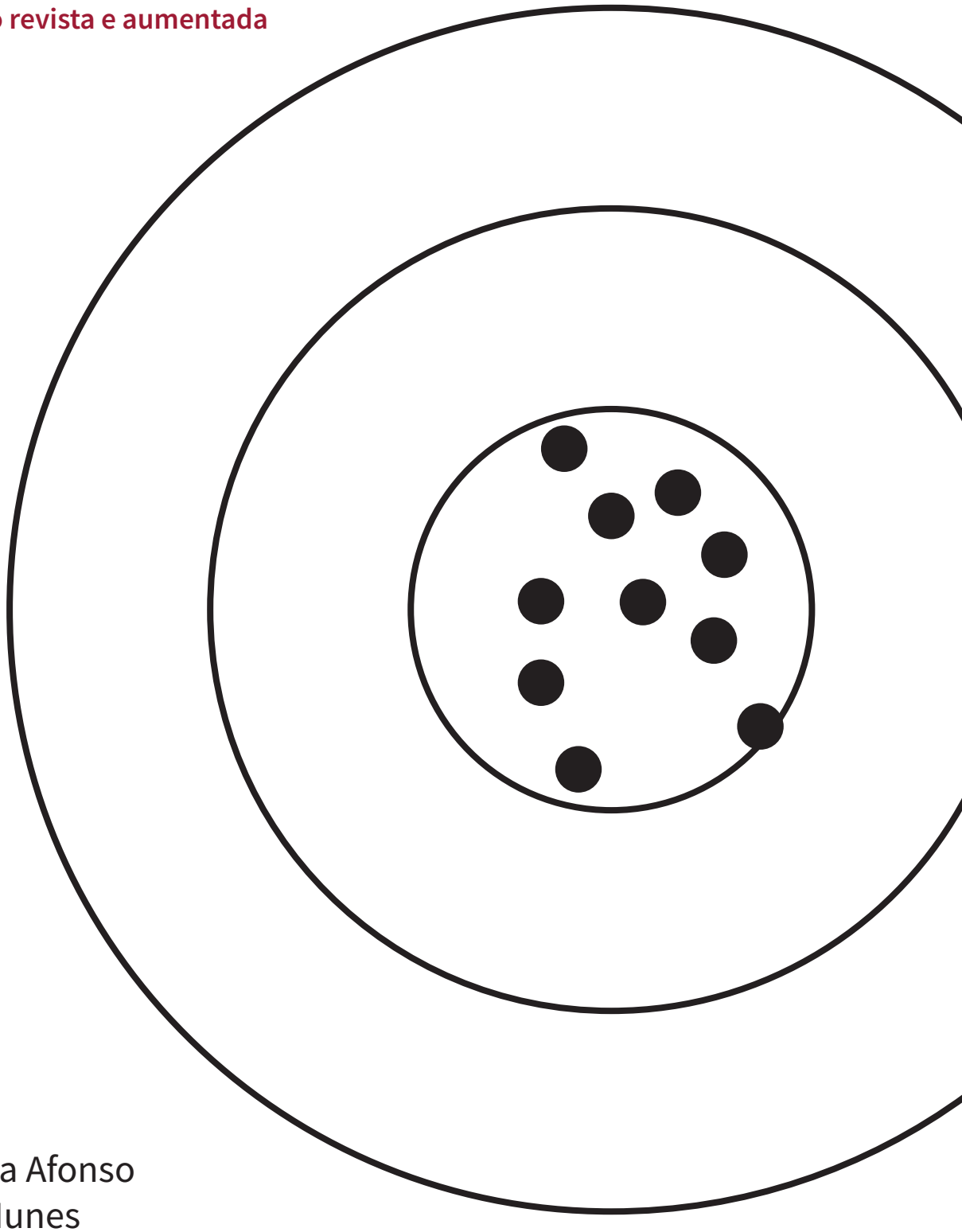
PROBABILIDADES E ESTATÍSTICA
Aplicações e Soluções em SPSS

Versão revista e aumentada



Anabela Afonso
Carla Nunes

UNIVERSIDADE DE ÉVORA

PROBABILIDADES E ESTATÍSTICA

Aplicações e Soluções em SPSS

Versão revista e aumentada

Anabela Afonso
Carla Nunes

UNIVERSIDADE
DE ÉVORA

Probabilidades e Estatística

Aplicações e Soluções em SPSS

Versão revista e aumentada

Anabela Afonso, Carla Nunes

Editora: Universidade de Évora

Capa: Divisão de Comunicação da Universidade de Évora

ISBN: 978-972-778-123-2

Índice

1 Prefácio

As Probabilidades e a Estatística têm vindo a ganhar um peso crescente no nosso quotidiano, repleto de informação que é necessário compreender.

Empresas, organismos e a própria sociedade civil constatam cada vez mais a necessidade da "quantificação" dos mais variadíssimos aspetos em se encontram inseridos. As necessidades são cada vez mais claras e exigentes assim como as respostas requeridas: Quanto? Onde? Quando? Como? Porquê? Com que impacto? Com que incerteza? Cenários possíveis?

A atual utilização da Estatística em todas as áreas, o desenvolvimento de *softwares* específicos e o facto de ser, cada vez mais, de todos e para todos, torna hoje indispensável que o ensino da Estatística se faça não só de uma forma robusta, mas também rápida, acessível e organizada.

Esta obra pretende capacitar os leitores com um conhecimento base de diferentes tópicos de Probabilidades e Estatística, que lhes permita ler/entender os conceitos relacionados com a utilização destas matérias na sua área, para posteriormente aplicarem corretamente as técnicas apropriadas e interpretarem os seus resultados.

Face à heterogeneidade dos destinatários (designadamente alunos dos 1º, 2º e 3º ciclos e pós-graduações) e ao consequente desnível em termos de bases matemáticas, que dificulta uma generalização do grau de aprofundamento dos conceitos teóricos e possíveis aplicações, este trabalho assume um perfil essencialmente teórico-prático, com uma abordagem simplista, mas rigorosa, para um público futuro utilizador da Estatística e não investigador da Estatística.

Cada capítulo é composto por uma primeira apresentação dos conceitos teóricos, seguida por um conjunto de exercícios resolvidos que auxiliam o leitor no seu estudo, resolvidos manualmente e, também sempre que possível, com o auxílio do SPSS. Para terminar cada capítulo são propostos exercícios, de diversos graus de dificuldades, cujas soluções são conjuntamente apresentadas.

No final do livro apresentamos um caso de estudo relacionado com a análise do Inquérito Nacional de Saúde, com base numa amostra de 1000 casos, gentilmente disponibilizada pelo Instituto Nacional de Saúde Dr. Ricardo Jorge e que se encontra disponibilizada na página http:// http://evunix.uevora.pt/~aafonso/. Este caso de estudo, resolvido em SPSS, permite fazer uma revisão dos conceitos mais utilizados, dando especial ênfase à aplicação e interpretação dos resultados.

Foi intenção deste trabalho referenciar todo o material bibliográfico utilizado, existem, porém, muitos exercícios que têm sido utilizados ao longo de vários anos nas universidades onde lecionamos e cuja origem já não nos é possível identificar.

Esta obra é uma versão revista e aumentada das duas edições anteriores. Nesta edição, tal como nas 2 anteriores, agradecemos toda a colaboração prestada por colegas, professores, alunos e amigos que foram imprescindíveis para a realização e melhoria deste trabalho. Um agradecimento especial ao meu filho Gonçalo pela precisa ajuda na edição da versão atual. A todos, o mais sincero OBRIGADA.

Este livro foi feito para os alunos. Pretende ser um contributo para o ensino das probabilidades e estatística e para a sua correta utilização. Esperamos que apreciem e que vos seja muito útil.

Anabela Afonso
Carla Nunes

1 Introdução

1.1 Noções base da Estatística

No nosso quotidiano é cada vez mais necessário tomar decisões rápidas e bem fundamentadas. As **probabilidades** e **estatística** podem ser pensadas como a ciência de aprendizagem a partir de dados, fornecendo métodos que auxiliam o processo de tomada de decisão através da análise dos dados disponíveis. Por outras palavras, é uma área do conhecimento que inclui os instrumentos necessários para recolher, organizar ou classificar, apresentar e interpretar conjuntos de dados.

A estatística divide-se em duas áreas:

- **Estatística descritiva:** conjunto de técnicas apropriadas para recolher, organizar, reduzir e apresentar dados estatísticos.
- **Inferência estatística:** conjunto de técnicas que, com base na informação amostral, permite caracterizar uma certa população, requerendo o conhecimento das probabilidades. As principais técnicas utilizadas são:
  - *Estimação*: visa determinar o valor dos parâmetros desconhecidos.
  - *Testes de hipóteses*: visa testar suposições acerca das características de uma certa população.

Estatística	Estatística Descritiva	
	Inferência Estatística	Estimação Testes de hipóteses

1.1.1 Distinção entre população e amostra

Muitas vezes não é desejável nem viável inquirir todos os elementos da população que se pretende estudar, especialmente quando o número de elementos da população é muito elevado. Daí que se inquira um subgrupo que seja representativo da população, ou seja, recolhe-se uma amostra. Na Figura 1.1 pretende-se ilustrar a representatividade da amostra e na Figura 1.2 a distinção entre população e amostra.

Figura 1.1: *Reflexão sobre a inferência estatística. (Fonte: http://alea-estp.ine.pt)*

Ao grupo de todos os elementos que se pretende estudar e que possuem uma característica (ou mais) em comum chama-se **população**.

As medidas relativas a uma população designam-se por **parâmetros** que, usualmente, são desconhecidos (mas fixos) e que, portanto, pretendem-se conhecer.

O subgrupo da população selecionado para análise é designado por **amostra**.

As medidas relativas à amostra designam-se por **estatísticas**. O valor destas estatísticas varia de amostra para amostra (logo é uma variável aleatória (v. a.)).

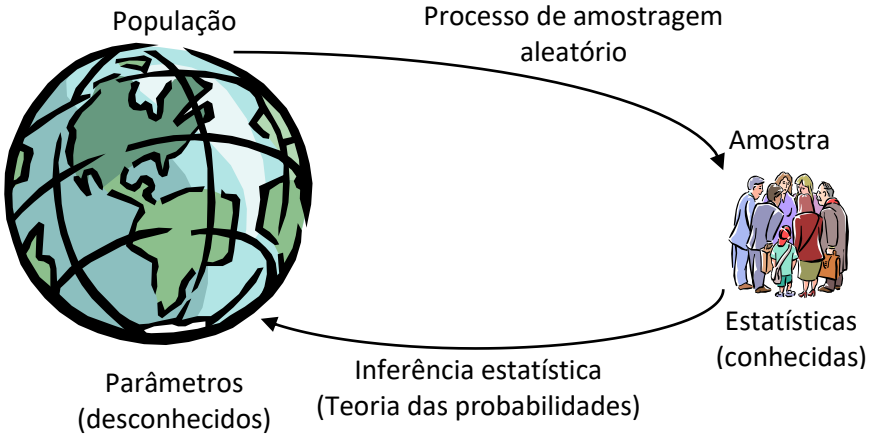


Figura 1.2: *População versus amostra.*

Na Tabela 1.1 apresentam-se as principais medidas estatísticas de interesse e a respetiva notação estatística.

Tabela 1.1: Notação.

Medida estatística	População (parâmetro)	Amostra (valor observado)	Amostra (estatística - v. a.)
Dimensão	N	n	--
Média	μ	$\bar{x}$	$\bar{X}$
Proporção	p	$\bar{p}$	$\bar{P}$
Variância	σ^2	s^2	S^2
Desvio padrão	σ	s	S
Coeficiente de correlação	ρ	r	R

1.1.2 Amostragem

Denomina-se por **amostragem** o processo utilizado para selecionar uma amostra a partir de uma população.

Esta seleção pode ser realizada recorrendo a dois tipos de métodos:

- **Probabilísticos**: cada um dos elementos da população tem hipóteses de ser incluído na amostra, sendo possível medir com rigor qual a probabilidade de tal suceder, através do cálculo de probabilidades. Exemplos:
  - Amostragem aleatória simples: com reposição e sem reposição;
  - Amostragem estratificada;
  - Amostragem por grupos e outras.

- **Não probabilísticos** também designados por **amostragem dirigida**: não permitem definir com rigor ou calcular as probabilidades de inclusão dos diferentes elementos da população na amostra. Estes processos são de um modo geral mais económicos e expeditos. Exemplos:
  - Amostragem por conveniência;
  - Amostragem subjetiva, entre outras.

Técnicas de amostragem	Probabilísticas	Aleatória Simples Sistemática Estratificada Por grupos Outras
	Não probabilísticas	Por conveniência Por julgamento (intencional) Por quotas "Bola de neve"

Através da amostragem podem-se retirar conclusões sobre a população a partir da amostra. É essencial que a forma como a amostra irá ser recolhida seja corretamente definida e organizada. A forma como são colocadas as questões podem originar que a informação recolhida não seja relevante para o estudo que se pretende realizar.
As etapas que compreendem a seleção da amostra, de forma a garantir que os objetivos são atingidos, são:

1. Definição dos objetivos do estudo.
2. Definição da população alvo: grupo de todos os indivíduos sobre os quais se pretendem tirar conclusões.
3. Decisão sobre os dados a observar.
4. Escolher a técnica de amostragem a utilizar para recolher a amostra e o método de recolha de dados (questionário, entrevista, …).
5. Calcular a dimensão da amostra.
6. Amostrar, ou seja, recolher a amostra.

1.1.3 Unidade estatística e dados estatísticos

Um conceito base da **estatística** é a definição dos indivíduos a estudar (unidades estatísticas) e que se pretende estudar sobre eles (características). Na Figura 1.3 é apresentado um esquema com o objetivo de clarificar estes conceitos.

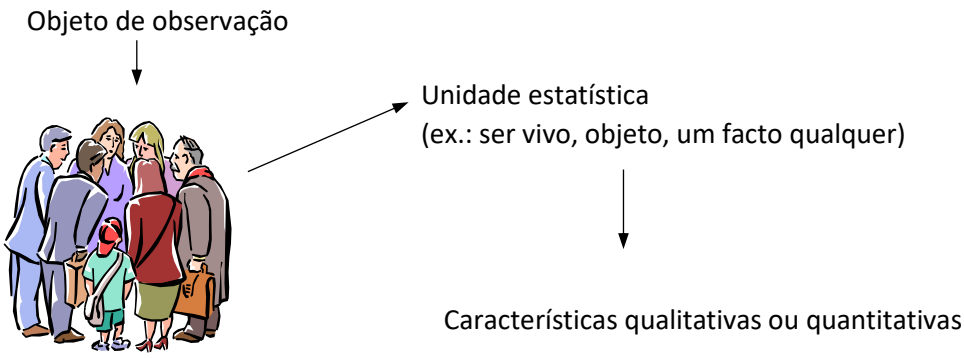


Figura 1.3: *Distinção entre unidade e característica.*

Designa-se por **unidade estatística**, ou **elemento**, qualquer indivíduo, objeto ou facto que é objeto da observação ou das conclusões.

Chama-se **dado estatístico** ao resultado da observação, que pode ser de tipo qualitativo ou quantitativo, das unidades estatísticas que compõem um determinado conjunto.

As características **qualitativas** revestem diferentes modalidades ou categorias.

As características **quantitativas** revestem diferentes intensidades ou valores.

Uma **variável** é uma característica relativa a todos os indivíduos (ou unidades estatísticas). O valor desta característica varia com as observações.

1.1.4 Classificação dos dados segundo a sua natureza

Os dados estatísticos podem ser do tipo:

- **Qualitativo** (ou **categórico**)**:** os dados podem ser separados em diferentes categorias que se distinguem por características não numéricas.
  - **Escala nominal:** quando os dados estão divididos por categorias que não possuem ordem.
    Exemplos: Classificação dos leitores de um determinado jornal pelo sexo – feminino ou masculino; Classificação dos frequentadores de pousadas pelo distrito de residência – Évora, Lisboa, Faro, Porto, ...; etc.
  - **Escala ordinal:** quando os **dados estão** divididos por categorias que obedecem a uma sequência com significado.
    Exemplo: Opinião sobre as aulas de estatística – muito boa, boa, razoável, má, muito má.
- **Quantitativo** (ou **numérico**)**:** consistem em números que representam contagens ou medições. Diz-se que os dados são de tipo **contínuo** quando podem tomar um número infinito não numerável de valores, e são de tipo **discreto** quando se podem enumerar os valores que podem tomar.
  Exemplos: temperatura do ar, número de pessoas que entram por hora num banco, número de erros por página num livro, etc.

Dados estatísticos	Qualitativos	Escala nominal
		Escala ordinal
	Quantitativos	Discretos
		Contínuos

1.1.5 Metodologia para a resolução de um problema estatístico

A resolução de um problema estatístico é composta por várias etapas. No esquema seguinte (Figura 1.4) pretende-se ilustrar os principais passos a percorrer.

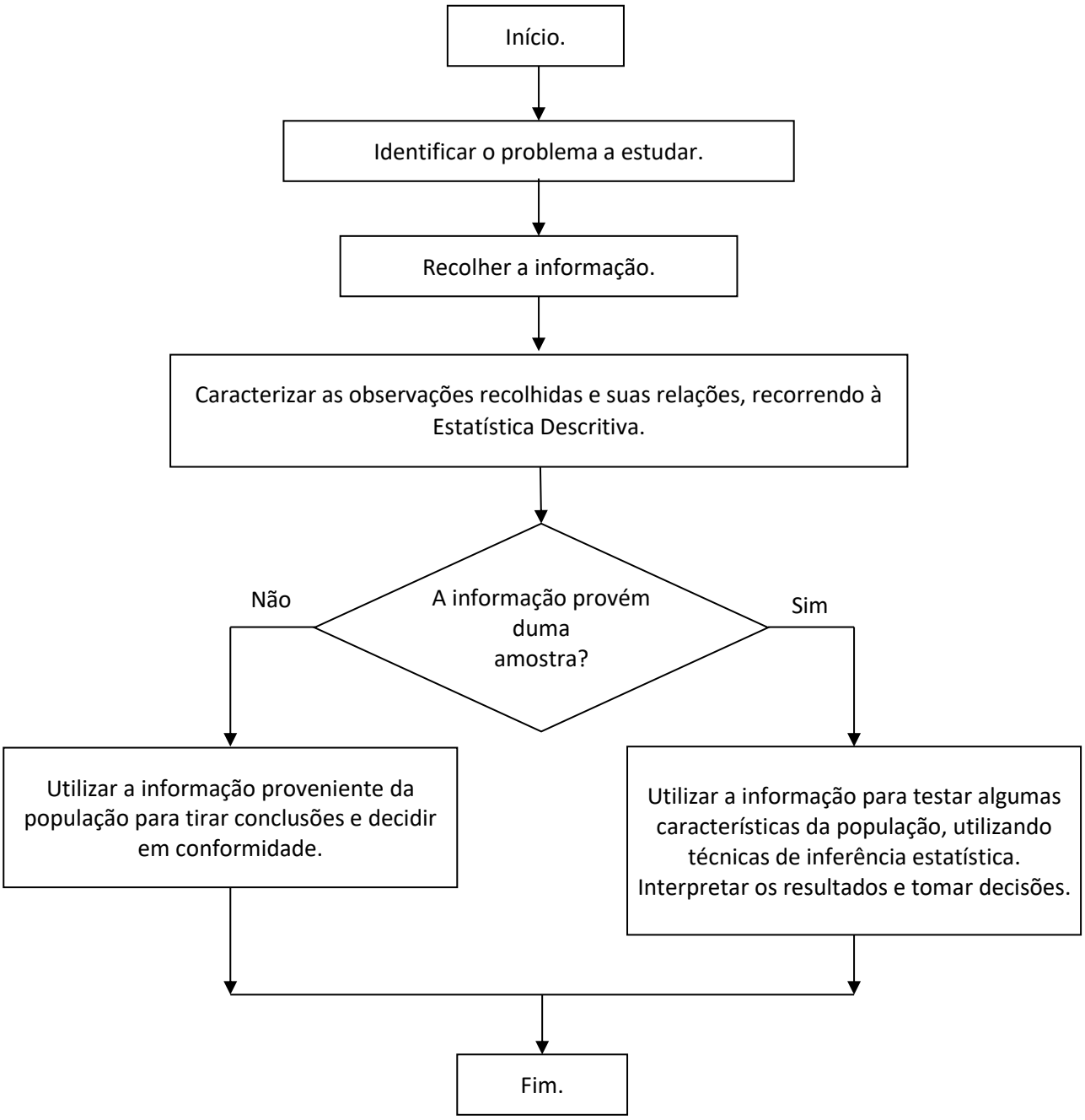


Figura 1.4: *Etapas para a resolução de um problema estatístico.*

1.2 Exercícios propostos

1. Para a realização um estudo sobre o hábito de fumar dos jovens portugueses do ensino superior.
    a) Identifique:
        i. A população.
        ii. Uma amostra.
        iii. As unidades estatísticas/indivíduos.
        iv. Os dados estatísticos.
    b) Na elaboração do questionário como é que formularia a questão sobre o consumo de tabaco, de forma a obter uma variável medida numa escala:
        i. Nominal?
        ii. Ordinal?
        iii. Quantitativa?
    c) Resumidamente, diga como exploraria/descreveria a informação que obteria em cada uma das situações apresentadas em b), por aplicação do questionário.

2. O excerto que se apresenta é um exemplo de um instrumento de recolha de dados:

Nome:_____
N.º de aluno: _____
1. Sexo: ☐ Masculino ☐ Feminino 2. Idade: ____ anos
3. Curso: _____ 4. Ano do curso: ____
5. Almoça na cantina da universidade?
☐ Sim (responder a 6) ☐ Não (responder a 7)
6. Qual a sua opinião sobre a qualidade da comida?
☐ Muito boa ☐ Boa ☐ Razoável ☐ Má ☐ Muito má
7. Onde costuma almoçar?
☐ Em casa ☐ No restaurante/café ☐ No bar da universidade ☐ Outro

Indique as variáveis presentes neste questionário e classifique-as.

3. Classifique cada uma das seguintes características em dado discreto ou contínuo:
    a) A estatura de um aluno.
    b) O número de erros num ditado.
    c) O tempo de espera (atraso) para uma consulta médica.
    d) O resultado efetivamente obtido na frequência.
    e) A nota final do aluno na disciplina.
    f) Número de exames médicos realizados num ano, por pessoa.

2 Estatística Descritiva

Neste capítulo são apresentadas as formas de sintetizar a informação recolhida através de tabelas, gráficos e medidas estatísticas, de modo a melhor caracterizar e interpretar essa informação.

2.1 Formas de representação tabular e gráfica

2.1.1 Tabela de frequências para dados univariados

Uma forma de resumir um conjunto de dados, composto por n observações, é através de uma **tabela de frequências**. Esta tabela disponibiliza um acesso rápido ao número, à percentagem ou proporção de elementos observados com uma determinada característica ou valor ou intervalo de valores (as chamadas classes de valores).

Uma **tabela de frequências** relaciona as categorias ou classes de valores com o número de ocorrências (frequências absolutas) e com a proporção (frequência relativa) de observações que pertencem a cada categoria ou classe.

As categorias ou classes de valores devem ser:

1. *Mutuamente exclusivas*, ou seja, cada valor observado só poderá pertencer a uma das categorias ou classes;
2. *Exaustivas*, ou seja, as categorias ou classes devem compreender todos os valores observados.

Notação: A notação utilizada nas tabelas de frequências é:

K	número de categorias/valores distintos/classes de valores que os dados assumem;
n_i	frequência absoluta da categoria/valor/classe de valores i, $i = 1, \dots, K$;
$n = \sum_{i=1}^{K} n_i$	dimensão do conjunto de dados, ou seja, número total de observações;
$f_i = \frac{n_i}{n}$	frequência relativa da categoria/valor/classe de valores i;
$N_i = \sum_{k=1}^{i} n_k$	frequência absoluta acumulada da categoria/valor/classe de valores i ;
$F_i = \frac{N_i}{n} = \sum_{k=1}^{i} f_k$	frequência relativa acumulada da categoria/valor/classe de valores i.

2.1.1.1 Dados qualitativos ou quantitativos discretos

A construção de uma tabela de frequências para *dados qualitativos* ou *quantitativos discretos* (Tabela 2.1) depende da definição das seguintes colunas:

1ª Coluna: Todas as K categorias ou valores distintos, x_i', que os dados assumem;

2ª Coluna: As frequências absolutas, n_i, ou seja, o número de vezes que cada categoria (valor) foi observada(o);

3ª Coluna: As frequências relativas, f_i, ou seja, a proporção de vezes que cada categoria (valor) foi observada(o);

Se os dados forem qualitativos ordinais ou quantitativos discretos são acrescentadas as colunas:

4ª Coluna: As frequências absolutas acumuladas, N_i, ou seja, o número de ocorrências das categorias (valores) inferiores ou iguais à categoria (valor) atual;

5ª Coluna: As frequências relativas acumuladas, F_i, ou seja, a proporção de ocorrências das categorias (valores) inferiores ou iguais à categoria (valor) actual.

Tabela 2.1: *Tabela de frequências para dados qualitativos ordinais ou quantitativos discretos.*

Categorias (x_i')	Freq. abs. (n_i)	Freq. rel. (f_i)	Freq. abs. acum. (N_i)	Freq. rel. acum. (F_i)
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
...	...	...	...	...
x_K	n_K	f_K	$N_K = n$	$F_K = 1$
Total	n	1		

Observação: Para dados qualitativos na escala nominal não se apresentam as frequências (absolutas e relativas) acumuladas (4ª e 5ª colunas).

Exemplo: Num estudo para analisar a ocorrência de acidentes de trabalho num determinado hospital, em 397 profissionais de saúde verificou-se que 16 não sofreram qualquer acidente, 32 tiveram 1 acidente, 89 reportaram 2 acidentes, 137 sofreram 3 acidentes, 98 sofreram 4 acidentes e 25 profissionais reportaram 5 acidentes.

A tabela de frequências associada à informação anterior é apresentada na Tabela 2.2.

Tabela 2.2: *Tabela de frequências relativa ao número de acidentes por profissional.*

N.º de acidentes por profissional (x_i')	N.º de profissionais (n_i)	Prop. de profissionais (f_i)	N.º acum. de profissionais (N_i)	Prop. acum. de profissionais (F_i)
0	16	0,0403	16	0,0403
1	32	0,0806	48	0,1209
2	89	0,2242	137	0,3451
3	137	0,3451	274	0,6902
4	98	0,2469	372	0,9370
5	25	0,0630	397	1,0000
Total	397	1,0000		

2.1.1.2 Dados quantitativos contínuos

Quando os dados são do tipo *quantitativo contínuo* então é necessário definir K classes de valores, que constituem as categorias dos dados em estudo. Para construir estes intervalos de classe existem vários métodos possíveis. Por exemplo, se interessa comparar os resultados de um estudo com os resultados de outro estudo, é fundamental que se utilizem as mesmas classes para ser possível efetuar as comparações. A forma como se definem as classes condiciona os resultados que apenas são válidos para a classificação efetuada. Seja qual for o método utilizado é aconselhável não obter um número muito elevado nem muito reduzido de classes (habitualmente $5 \leq K \leq 20$).

Exemplo de um método de construção de classes:

1º Determinar o número K de classes a construir, com base nas n observações, fazendo (regra de Sturges):

$$K = \left[\frac{\ln(n)}{\ln(2)}\right] + 1,$$

onde $[\text{número}]$ representa a parte inteira do número obtido (por ex: [5,1] = 5 e [5,9] = 5).

2º Determinar a amplitude a do conjunto de dados fazendo:

$$a = \text{máximo} - \text{mínimo}.$$

3º Determinar a amplitude ac de cada uma das classes fazendo:

$$ac = \frac{a}{K}.$$

4º Construir as classes c_i da seguinte forma:

$$c_1 = [\text{mínimo}; \text{mínimo} + ac[$$
$$c_2 = [\text{mínimo} + ac;\ \text{mínimo} + 2 \times ac[$$
$$\ldots$$
$$c_K = [\text{mínimo} + (K-1) \times ac;\ \text{mínimo} + K \times ac].$$

Exemplo: O Sr. Nobre decidiu dedicar-se à criação de leitões, que vende quando atingem os dois meses de idade e pesam mais de 9kg. Pretendendo fazer um estudo sobre os lucros obtidos com essa atividade, resolveu pesar 60 leitões com dois meses de idade, tendo obtido os seguintes resultados:

4,1	5,8	5,8	6,1	6,7	7,0	7,0	7,5	7,5	7,5
7,7	8,2	8,3	8,5	8,7	8,8	9,0	9,0	9,1	9,1
9,1	9,2	9,2	9,2	9,2	9,4	9,4	9,4	9,5	9,5
9,7	9,8	10,0	10,0	10,2	10,2	10,3	10,6	10,6	10,8
10,9	10,9	11,0	11,1	11,1	11,6	11,7	11,8	11,8	11,8
12,0	12,2	12,2	12,3	12,5	12,6	12,7	14,0	14,2	14,8

N.º de observações: $n = 60$

N.º de classes: $K = \left[\frac{ln(60)}{ln(2)}\right] + 1 = [5{,}9] + 1 = 5 + 1 = 6$

Amplitude total: $a = 14{,}8 - 4{,}1 = 10{,}7$

Amplitude das classes: $ac = \frac{10{,}7}{6} \approx 1{,}8$

Classes: $c_1 = [4{,}1; 5{,}9[$; $c_2 = [5{,}9; 7{,}7[$; $c_3 = [7{,}7; 9{,}5[$;
$c_4 = [9{,}5; 11{,}3[$; $c_5 = [11{,}3; 13{,}1[$; $c_6 = [13{,}1; 14{,}9]$

Para construir uma tabela de frequências para *dados quantitativos contínuos* (Tabela 2.3) colocar na:

1ª Coluna: As K classes de valores;

2ª Coluna: Os pontos médios, x_i', das classes $c_i = [LI_i; LS_i[$, onde LI_i e LS_i representam, repectivamente, os limites inferior e superior da classe i:

$$x_i' = \frac{LI_i + LS_i}{2},$$

ou seja, o ponto que fica no meio do intervalo da classe

3ª Coluna: As frequências absolutas, n_i, ou seja, o número de elementos cujo valor observado pertence à classe em estudo;

4ª Coluna: As frequências relativas, f_i, ou seja, a proporção de elementos cujo valor observado pertence à classe em estudo;

5ª Coluna: As frequências absolutas acumuladas, N_i, ou seja, o número de elementos cujo valor observado é inferior ao limite superior da classe em estudo, LS_i;

6ª Coluna: As frequências relativas acumuladas, F_i, ou seja, a proporção de elementos cujo valor observado é inferior ao limite superior da classe em estudo, LS_i.

Tabela 2.3: *Tabela de frequências para dados quantitativos contínuos.*

Classes (c_i)	Ponto médio (x_i')	Freq. abs. (n_i)	Freq. rel. (f_i)	Freq. abs. acum. (N_i)	Freq. rel. acum. (F_i)
$[LI_1; LS_1[$	x_1'	n_1	f_1	N_1	F_1
$[LI_2; LS_2[$	x_2'	n_2	f_2	N_2	F_2
...	...	...	...	...	...
$[LI_K; LS_K]$	x_K'	n_K	f_K	$N_K = n$	$F_K = 1$
Total		n	1		

Exemplo: Retomando o exemplo anterior relativo aos pesos dos leitões, na Tabela 2.4 apresenta-se a tabela de frequências associada aos pesos observados.

Tabela 2.4: *Tabela de frequências relativa aos pesos dos leitões.*

Pesos em kg (c_i)	Ponto médio (x_i')	N.º de leitões (n_i)	Prop. de leitões (f_i)	N.º ac. de leitões (N_i)	Prop. ac. de leitões (F_i)
[4,1; 5,9[	5	3	0,0500	3	0,0500
[5,9; 7,7[	6,8	7	0,1167	10	0,1667
[7,7; 9,5[	8,6	18	0,3000	28	0,4667
[9,5; 11,3[	10,4	17	0,2833	45	0,7500
[11,3; 13,1[	12,2	12	0,2000	57	0,9500
[13,1; 14,9]	14	3	0,0500	60	1,0000
Total		60	1,0000		

2.1.2 Representação gráfica de dados univariados

Os gráficos mais utilizados para representar os dados são:

- **Gráfico circular** – dados qualitativos;
- **Gráfico de barras** – dados qualitativos ou quantitativos discretos;
- **Gráfico de frequências acumuladas** – dados qualitativos na escala ordinal ou quantitativos discretos;
- **Histograma** – dados quantitativos contínuos;
- **Polígono de frequências** – dados quantitativos;
- **Polígono de frequências acumuladas** – dados quantitativos contínuos;
- **Caixa-de-bigodes** – dados não agrupados quantitativos.

2.1.2.1 Gráfico circular

Um **gráfico circular** é constituído por um círculo dividido em tantas fatias quantas as categorias da variável. O tamanho das fatias é determinado pelo número ou percentagem/proporção de observações nas categorias, i. e., pelas frequências absolutas, n_i, ou pelas relativas, f_i.

Na Figura 2.1 apresenta-se um exemplo genérico de um gráfico circular.

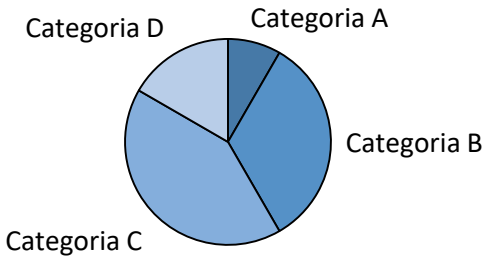


Figura 2.1: *Gráfico circular.*

2.1.2.2 Gráfico de barras

Um **gráfico de barras** é um diagrama de barras (usualmente verticais), sendo cada barra associada a cada uma das categorias da variável. A altura das barras é determinada pelas frequências absolutas, n_i, ou as relativas, f_i.

Na Figura 2.2 apresenta-se um exemplo genérico de um gráfico de barras.

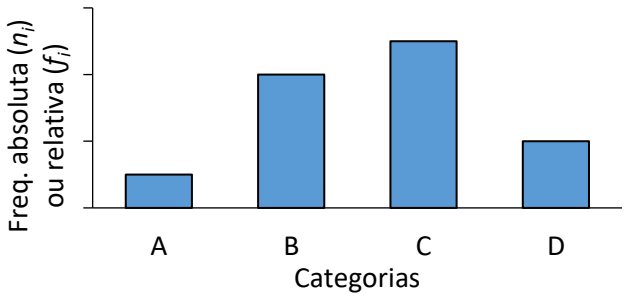


Figura 2.2: *Gráfico de barras.*

2.1.2.3 Gráfico de frequências acumuladas

Um **gráfico de frequências acumuladas**, ou diagrama integral, é um gráfico de linhas onde são representadas as frequências absolutas, N_i, ou relativas, F_i, acumuladas. Este gráfico apresenta a frequência acumulada até cada uma das categorias/valores, notando que até à primeira categoria/valor a frequência acumulada é nula. Para categorias/valores superiores à(ao) última(o) a frequência acumulada toma o valor n, se forem representadas as frequências N_i, ou 1, se forem representadas as frequências F_i.

Na Figura 2.3 apresenta-se um exemplo genérico de um gráfico de frequências acumuladas.

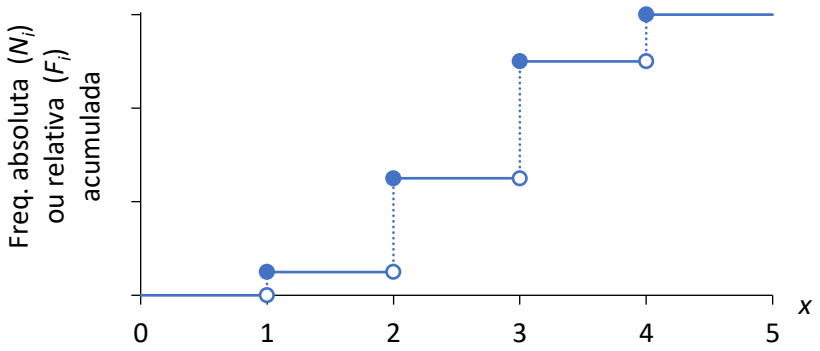


Figura 2.3: *Gráfico de frequências acumuladas.*

2.1.2.4 Histograma

Um **histograma** é um gráfico de barras verticais adjacentes, com uma barra associada a cada uma das classes da variável. A base de cada barra é proporcional à amplitude da respetiva classe e a área proporcional às frequências absolutas, n_i, ou relativas, f_i.

Na Figura 2.4 apresenta-se um exemplo genérico de um histograma.

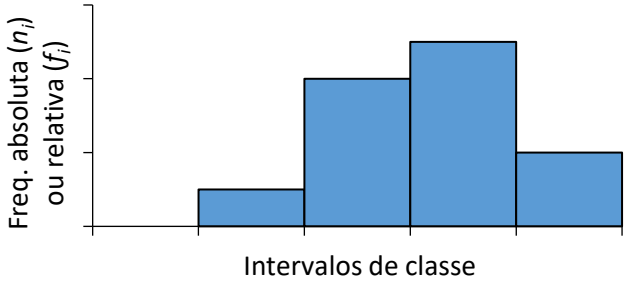


Figura 2.4: *Histograma.*

Quando as classes têm amplitudes diferentes é necessário transformar as frequências absolutas ou relativas, para que se verifique a proporcionalidade entre a altura das barras e a sua base e se garanta que a área é igual a n ou a 1 (com base nas frequências absolutas e relativas, respetivamente). Assim as frequências absolutas e relativas a representar são:

$$n_i^* = \frac{n_i}{a_i} \quad \text{e} \quad f_i^* = \frac{f_i}{a_i}$$

onde a_i é a amplitude da classe i.

2.1.2.5 Polígono de frequências

Um **polígono de frequências** é um gráfico de linhas onde são representadas as frequências absolutas, n_i, ou relativas, f_i, nos pontos médios das classes. Para fechar o polígono é necessário criar uma classe adicional em cada um dos extremos, com amplitude igual à classe adjacente e com frequência nula.

Na Figura 2.5 apresenta-se um exemplo genérico de um polígono de frequências.

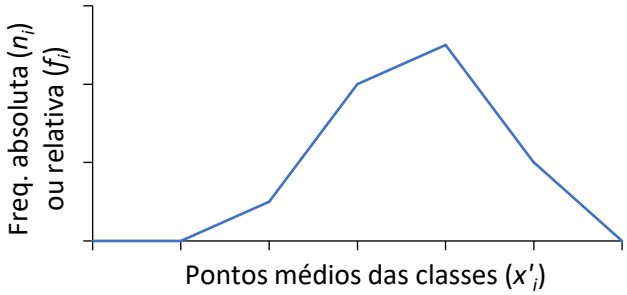


Figura 2.5: *Polígono de frequências.*

A área sob o polígono deverá ser igual à área do histograma, pelo que quando as classes têm amplitudes diferentes é necessário transformar as frequências absolutas ou relativas conforme já foi referido anteriormente (ver 2.1.2.4).

2.1.2.6 Polígono de frequências acumuladas

Um **polígono de frequências acumuladas**, ou polígono integral, é um gráfico de linhas onde são representadas frequências absolutas, N_i, ou relativas, F_i, acumuladas. A frequência acumulada para valores inferiores ao limite inferior da primeira classe é nula. A frequência acumulada para valores superiores ao limite superior da última classe é n, se forem representadas as frequências N_i, ou 1, se forem representadas as frequências F_i.

Na Figura 2.6 apresenta-se um exemplo genérico de um polígono de frequências acumuladas.

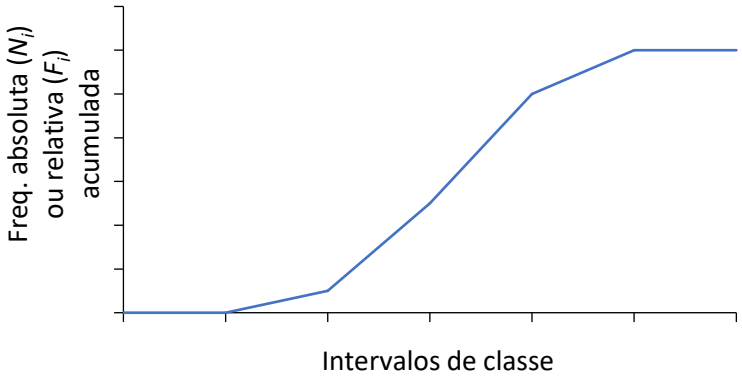


Figura 2.6: *Polígono de frequências acumuladas.*

2.1.2.7 Caixa de bigodes

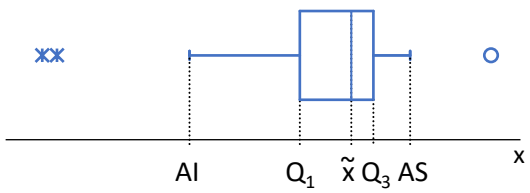
Da caixa de bigodes (ou diagrama de caixa ou *boxplot*) podem-se extrair as seguintes características de um conjunto de dados:

- Localização[†];
- Dispersão[†];
- As(simetria)[†];
- Valores atípicos (ou anómalos ou *outliers*).

[†] Ver Medidas descritivas, secção 2.2.

Uma **caixa de bigodes** é um gráfico que contém por um retângulo, dividido em duas partes, que situa os quartis. Os bigodes da caixa situam os pontos adjacentes inferior e superior, ou seja, o menor e maior valores observados que ainda não são considerados observações atípicas. Os asteriscos identificam os valores atípicos, ou seja, os valores observados muito pequenos e muito grandes (com ordens de grandeza que implicam que sejam classificados como valores anómalos).

Na Figura 2.7 apresenta-se um exemplo genérico de uma caixa de bigodes.



AI – Menor valor não atípico (ponto adjacente inferior)
Q_1 – 1º quartil†
$\tilde{x}$ – Mediana†
Q_3 – 3º quartil†
AS – Maior valor não atípico (ponto adjacente superior)
o – Valores atípicos moderados
$*$ – Valores atípicos severos

Figura 2.7: *Caixa de bigodes.*

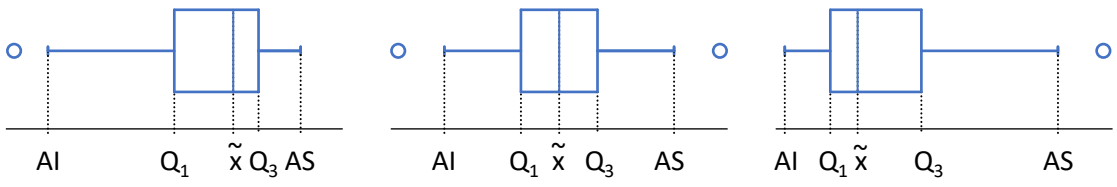
O ponto adjacente inferior (superior) corresponde ao menor (maior) valor observado que ainda está dentro da barreira de *outliers* $[BI; BS]$:

- Barreira inferior: $BI = Q_1 - 1{,}5 \times (Q_3 - Q_1)$;
- Barreira superior: $BS = Q_3 + 1{,}5 \times (Q_3 - Q_1)$.

Todos os valores da amostra que não pertencem ao intervalo $[BI; BS]$ designam-se por **atípicos**, e classificam-se em (Murteira *et al.*, 2007):

- Valores atípicos severos inferiores:
  valores inferiores a $Q_1 - 3 \times (Q_3 - Q_1)$;
- Valores atípicos moderados inferiores:
  valores superiores a $Q_3 - 3 \times (Q_3 - Q_1)$ e inferiores a $Q_1 - 1{,}5 \times (Q_3 - Q_1)$;
- Valores atípicos moderados superiores:
  valores superiores a $Q_3 + 1{,}5 \times (Q_3 - Q_1)$ e inferiores a $Q_3 + 3 \times (Q_3 - Q_1)$;
- Valores atípicos severos superiores:
  valores superiores a $Q_3 + 3 \times (Q_3 - Q_1)$.

A (as)simetria da distribuição é indicada pela caixa central e pelo comprimento dos "bigodes" (Figura 2.8).



a) Enviesada à direita. *b) Simétrica.* *c) Enviesada à esquerda*

Figura 2.8: *Exemplos de assimetria na caixa de bigodes.*

2.1.3 Tabela de contingência para dados bivariados

Uma forma de resumir um conjunto bivariado de dados, composto por n observações, é através da **tabela de contingência** (Tabela 2.5).

Uma **tabela de contingência** é uma representação de dados bivariados, quer de tipo qualitativo quer do tipo quantitativo, que podem ser classificados segundo dois critérios. Nesta tabela as linhas correspondem a um dos critérios e as colunas correspondem ao outro critério. No interior da tabela, as células correspondem ao número de observações, n_{ij}, que satisfazem simultaneamente o critério da linha i e da coluna j.

Tabela 2.5: *Tabela de contingências.*

	Critério Y						
Critério X	Y_1	Y_2	...	Y_j	...	Y_J	Total (X)
X_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1C}	$n_{1.}$
X_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2C}	$n_{2.}$
...	...	...	...	...	...	...	...
X_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iC}	$n_{i.}$
...	...	...	...	...	...	...	...
X_L	n_{L1}	n_{L2}	...	n_{Lj}	...	n_{LC}	$n_{L.}$
Total (Y)	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.C}$	n

Exemplo: Os Serviços Sociais de uma determinada Universidade têm que gerir as refeições fornecidas aos alunos e funcionários dessa universidade. Na Tabela 2.6 apresenta-se o número de alunos que almoçaram, jantaram ou almoçaram e jantaram em cada um dos refeitórios num determinado mês.

Tabela 2.6: *Tabela de contingências relativa ao número de alunos que almoçaram, jantaram ou almoçaram e jantaram nos refeitórios da universidade.*

	Refeitório				Total
Refeição	A	B	C	D	(por refeição)
Só almoço	300	100	150	50	600
Só jantar	50	80	0	0	130
Almoço e jantar	500	800	0	0	1300
Total (por refeitórios)	850	980	150	50	2030

2.1.4 Representação gráfica de dados bivariados

2.1.4.1 Gráfico de barras lado a lado

Um **gráfico de barras lado a lado** é um conjunto de diagramas de barras (usualmente verticais), em que cada conjunto corresponde ao gráfico de barras de uma variável qualitativa (por ex. Y) em cada uma das categorias da outra variável qualitativa (X). A altura das barras é determinada pelas frequências absolutas, n_i, ou relativas, f_i, da variável Y na categoria da variável X.

Na Figura 2.9 apresenta-se um exemplo genérico de um gráfico de barras lado a lado.

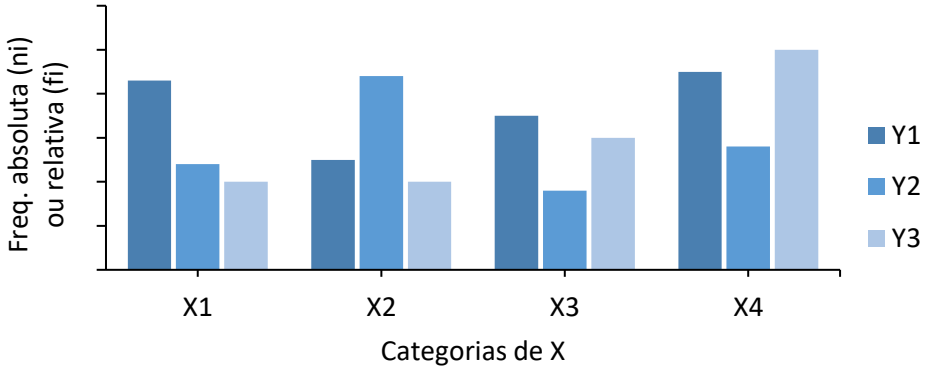


Figura 2.9: *Gráfico de barras lado a lado.*

2.1.4.2 Gráfico de barras empilhadas

Um **gráfico de barras empilhadas a 100%** é um conjunto de diagramas de barras (usualmente verticais), em que cada conjunto corresponde ao gráfico de barras empilhadas de uma variável qualitativa (por ex. Y) em cada uma das categorias da outra variável qualitativa (X). A altura das barras é determinada pelas frequências absolutas, n_i, ou relativas, f_i, da variável Y na categoria da variável X.

Na Figura 2.10 apresenta-se um exemplo genérico de um gráfico de barras empilhadas.

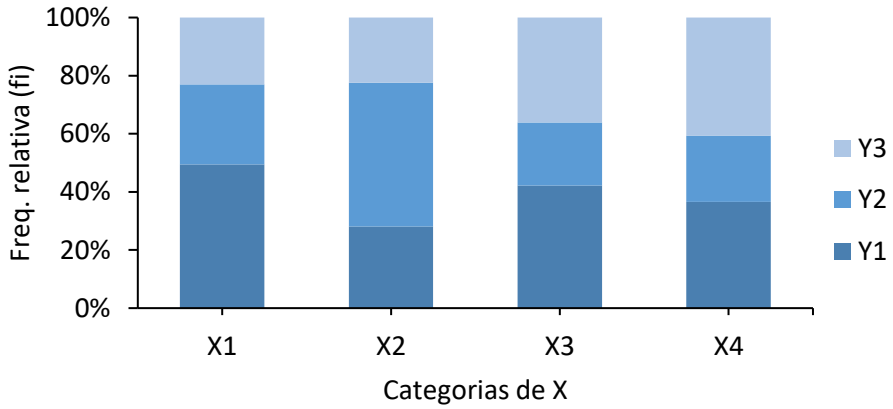


Figura 2.10: *Gráfico de barras empilhadas a 100%.*

2.1.4.3 Caixas de bigodes lado a lado

A representação **caixas de bigodes lado a lado** corresponde a um conjunto de caixas de bigodes dispostas lado a lado, representando cada uma das caixas a variável quantitativa em cada uma das categorias da variável qualitativa.

Na Figura 2.11 apresenta-se um exemplo genérico de caixas de bigodes lado a lado.

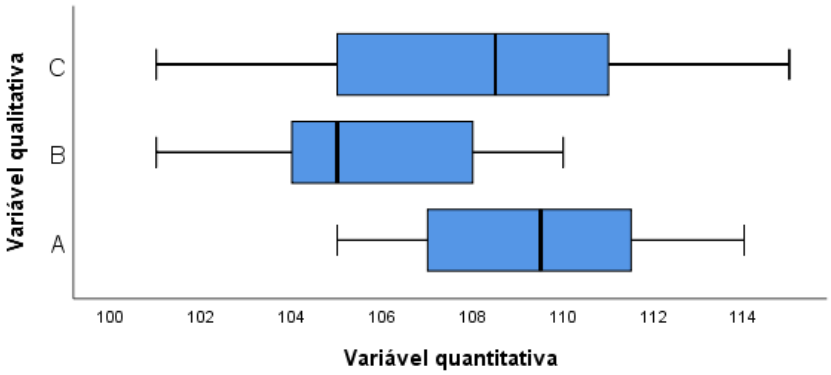


Figura 2.11: *Caixas de bigodes lado a lado.*

2.1.4.4 Gráfico de dispersão

Designa-se por **gráfico de dispersão** a representação gráfica num sistema de eixos cartesianos dum conjunto de observações (emparelhadas) de duas variáveis quantitativas X e Y:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

Na Figura 2.12 apresenta-se um exemplo genérico de um gráfico de dispersão.

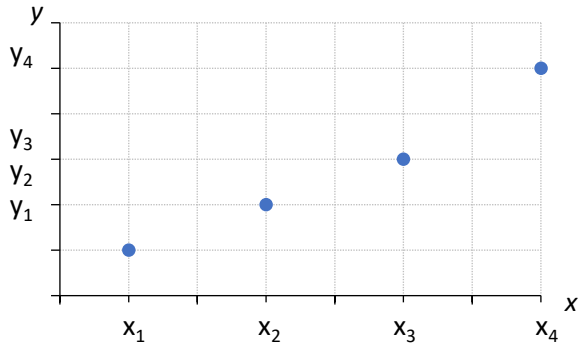


Figura 2.12: *Gráfico de dispersão.*

2.1.4.5 Gráfico quantil-quantil

Designa-se por **gráfico quantil-quantil**, ou Q-Q plot, a representação gráfica num sistema de eixos cartesianos dos quantis[†] de duas variáveis X e Y:

$$\left(Q_p^*, Q_p^*\right), \quad 0 \leq p \leq 1.$$

O gráfico quantil-quantil permite determinar se dois conjuntos de dados são provenientes da mesma população, através da comparação dos seus quantis, i.e., da proporção de observações que são menores ou iguais a um determinado valor. Se os conjuntos de dados forem provenientes da mesma população, então os pontos devem posicionar-se aproximadamente sobre uma reta com declive 1. Quanto mais afastados estiverem os pontos dessa reta maior a evidência de que os conjuntos de dados são provenientes de populações diferentes.

Na Figura 2.13 apresenta-se um exemplo genérico de um gráfico quantil-quantil.

[†] Ver definição de quantil, secção 2.2.1.2.1.

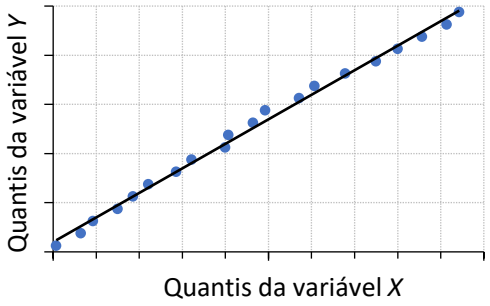


Figura 2.13: *Gráfico quantil-quantil.*

Estes gráficos são bastante úteis para verificar se o conjunto de dados pode ser proveniente de uma população com distribuição Normal (ou Gaussiana)[‡].

2.1.5 Exercícios resolvidos

2.1.5.1 Dados qualitativos na escala nominal

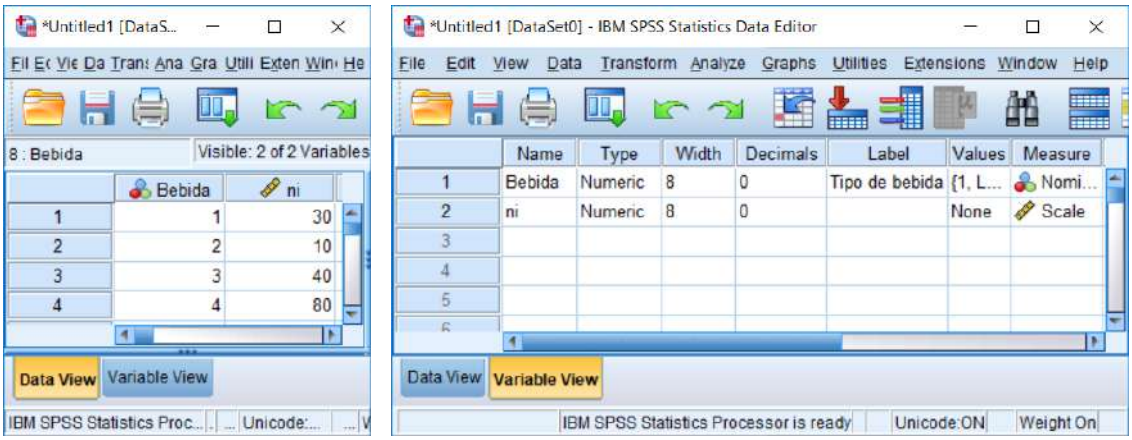
No jardim-de-infância *O Parque da Pequenada* questionaram-se as crianças com mais de 3 anos relativamente ao tipo preferido de bebida. Das 160 crianças inquiridas, 30 indicaram o leite como a bebida preferida, 10 referiram a água, 40 disseram os sumos naturais e 80 referiram os refrigerantes.

a) Defina e classifique a variável em estudo.
b) Construa a tabela de frequências.
c) Qual é a frequência absoluta de alunos que preferem água?
d) Qual é a frequência relativa de alunos que preferem refrigerantes?
e) Represente graficamente a informação anterior.

Resolução:

a) A variável em estudo é o tipo preferido de bebida pelas crianças com mais de 3 anos do jardim-de-infância *O Parque da Pequenada*. É uma variável qualitativa nominal.

b) (SPSS)



(SPSS) Data → Weight Cases…
(⊙ Weight cases by: Frequency Variable: ni)

[‡] Ver distribuição normal, secção 5.2.2.

(SPSS) Analyse → Descriptive Statistics → Frequencies…
(Variable: Bebida; ☑ Display frequency tables)

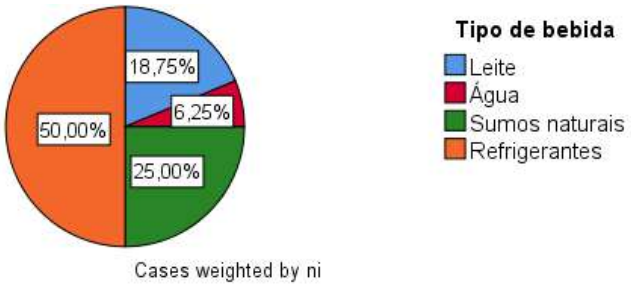
Tipo preferido de bebida

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Leite	30	18,8	18,8	18,8
	Água	10	6,3	6,3	25,0
	Sumos naturais	40	25,0	25,0	50,0
	Refrigerantes	80	50,0	50,0	100,0
	Total	160	100,0	100,0	

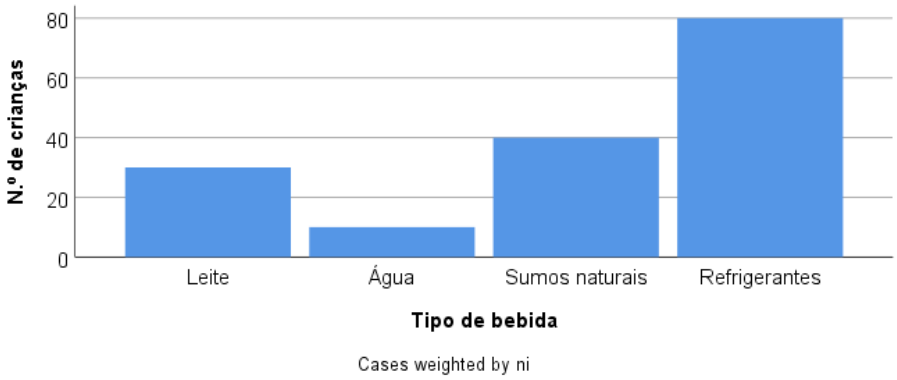
c) Frequência absoluta dos alunos que preferem água: 10 crianças.

d) Frequência relativa dos alunos que preferem refrigerantes: 50% ou 0,5.

e) (SPSS) Graphs → Legacy Dialogs → Pie… → Summaries for groups of cases
(Slices Represent: ⊙ N of cases ou ⊙ % of cases; Define Slices by: Bebida)



(SPSS) Graphs → Legacy Dialogs → Bar… → Simple; Summaries for groups of cases
(Bars Represent: ⊙ N of cases ou ⊙ % of cases; Category Axis: Bebida)



Como se pode observar pela análise da tabela de frequências e das figuras verificou-se que o mais frequente foi os alunos terem indicado como bebida preferida os refrigerantes e que uma minoria mencionou a água.

2.1.5.2 Dados qualitativos na escala ordinal

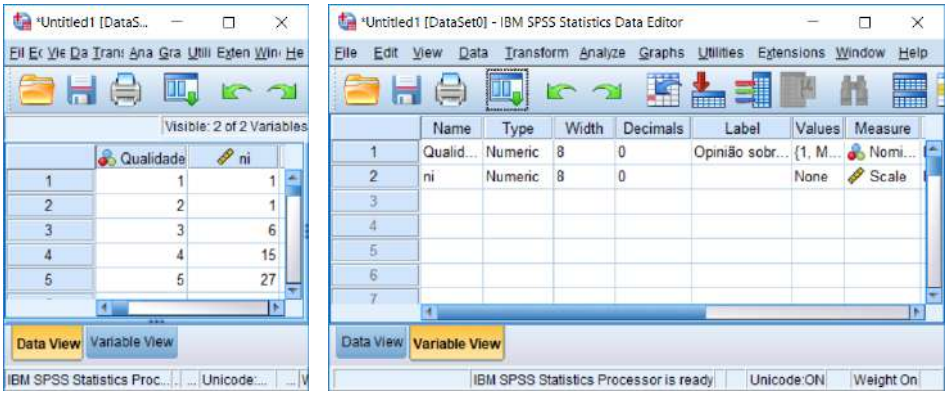
Num inquérito realizado utentes de um determinado Centro de Saúde sobre a qualidade do serviço, 27 afirmaram que o serviço foi muito bom, 15 afirmaram que foi bom, 6 disseram que foi médio, 1 considerou que foi mau e 1 muito mau.

a) Identifique a variável em estudo.
b) Construa a tabela de frequências.
c) Represente graficamente a informação anterior.

Resolução:

a) A variável em estudo é a opinião sobre a qualidade do serviço num determinado Centro de Saúde.

b) (SPSS)



(SPSS) Data → Weight Cases…
(⊙ Weight cases by: Frequency Variable: ni)

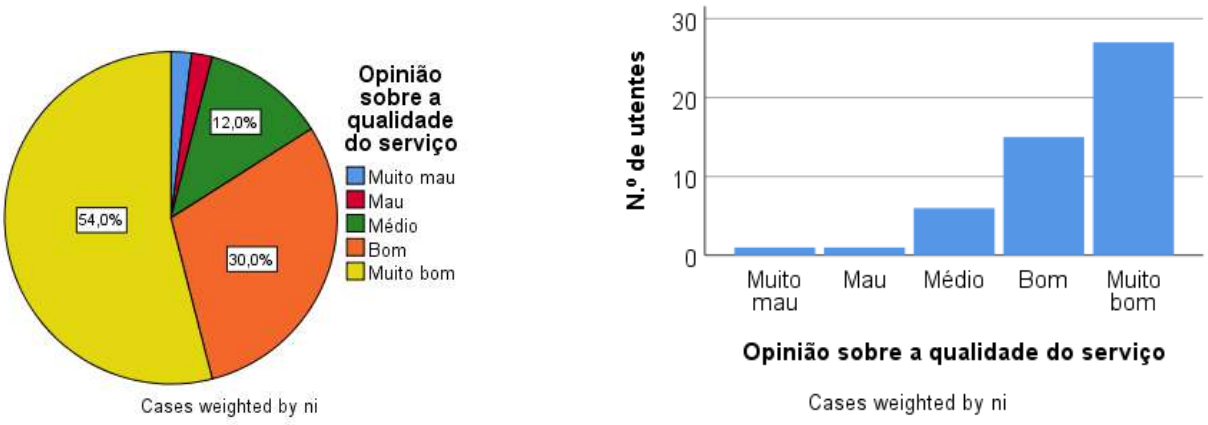
(SPSS) Analyse → Descriptive Statistics → Frequencies…
(Variable: Qualidade; ☑ Display frequency tables)

Qualidade do serviço prestado num determinado Centro de Saúde

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Muito mau	1	2,0	2,0	2,0
	Mau	1	2,0	2,0	4,0
	Médio	6	12,0	12,0	16,0
	Bom	15	30,0	30,0	46,0
	Muito bom	27	54,0	54,0	100,0
	Total	50	100,0	100,0	

c) (SPSS) Graphs → Legacy Dialogs → Pie… → Summaries for groups of cases
(Slices Represent: ⊙ N of cases ou ⊙ % of cases; Define Slices by: Qualidade)

(SPSS) Graphs → Legacy Dialogs → Bar… → Simple; Summaries for groups of cases
(Bars Represent: ⊙ N of cases ou ⊙ % of cases; Category Axis: Qualidade)



a) Gráfico circular. *b) Gráfico de barras.*

Figura 2.14: *Representação gráfica das frequências simples da opinião dos utentes sobre a qualidade do serviço.*

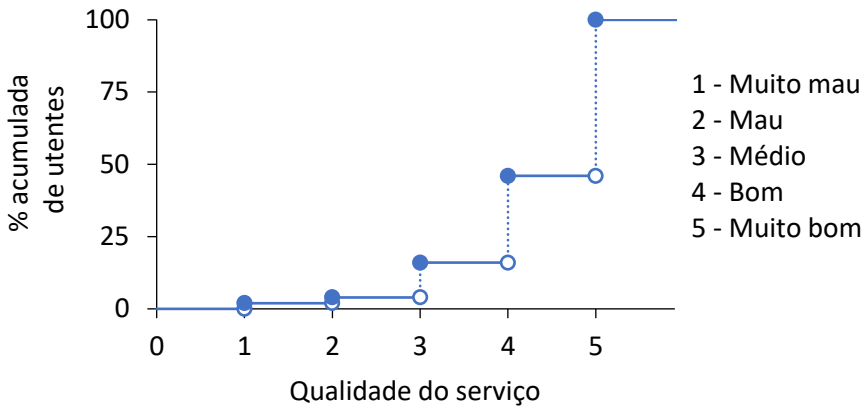


Figura 2.15: *Gráfico de frequências acumuladas da opinião sobre a qualidade do serviço.*

Da análise da tabela de frequências e da Figura 2.14, verificou-se que o mais comum é os utentes considerarem que a qualidade do serviço prestado nesse Centro de Saúde é muito boa, e muito poucos utentes consideraram que a qualidade do serviço é má ou muito má.

2.1.5.3 Dados quantitativos discretos

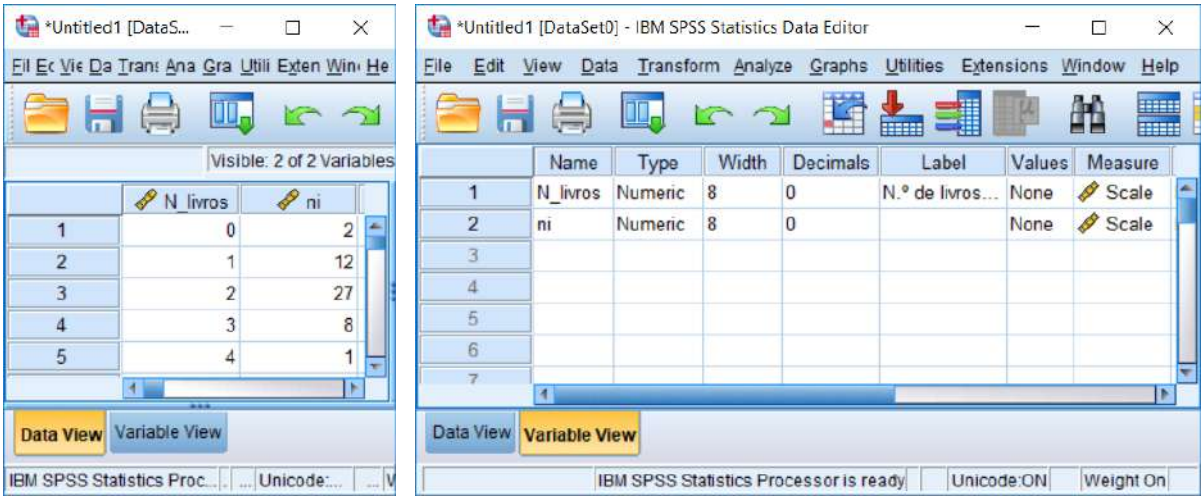
Numa aula teórica de Estatística perguntou-se aos 50 alunos presentes quantos livros leram durante as férias, tendo-se obtido os seguintes resultados: 1 aluno leu 4 livros, 8 leram 3 livros, 27 leram 2 livros, 12 leram apenas 1 livro e os restantes alunos não leram qualquer livro.

a) Identifique a variável em estudo.
b) Construa a tabela de frequências.
c) Represente graficamente a informação anterior.

Resolução:

a) A variável em estudo é o número de livros lidos durante as férias, pelos alunos.

b) (SPSS)



(SPSS) Data → Weight Cases…
(⊙ Weight cases by: Frequency Variable: ni)

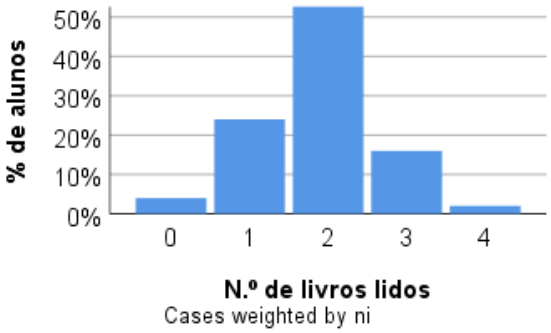
(SPSS) Analyse → Descriptive Statistics → Frequencies…
(Variable: N_Livros; ☑ Display frequency tables)

N.º de livros lidos durante as férias de verão

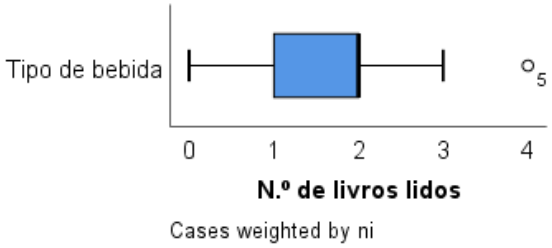
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	2	4,0	4,0	4,0
	1	12	24,0	24,0	28,0
	2	27	54,0	54,0	82,0
	3	8	16,0	16,0	98,0
	4	1	2,0	2,0	100,0
	Total	50	100,0	100,0	

c) (SPSS) Graphs → Legacy Dialogs → Bar… → Simple; Summaries for groups of cases (Bars Represent: ⊙ N of cases ou ⊙ % of cases; Category Axis: N_Livros)

(SPSS) Graphs → Legacy Dialogs → Boxplot… → Simple; Summaries of separate variables (Boxes Represent: N_livros)



a) Gráfico de barras.



b) Caixa de bigodes.

Figura 2.16: *Representação gráfica das frequências simples do número de livros lidos durante as férias de verão, pelos alunos que frequentam a disciplina Estatística.*

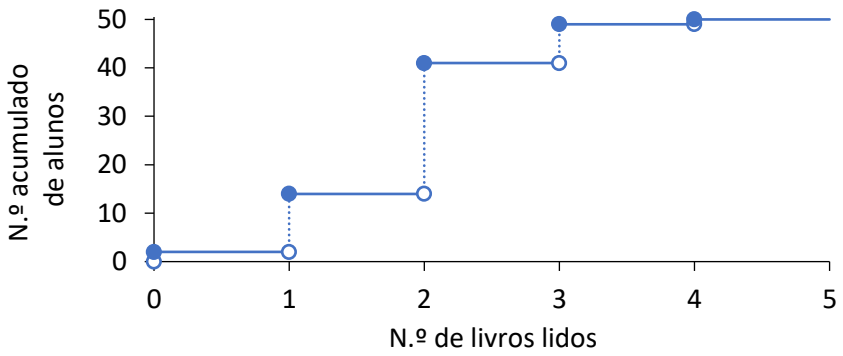


Figura 2.17: *Gráfico de frequências acumuladas do número de livros lidos durante as férias de verão, pelos alunos que frequentam a disciplina Estatística.*

Da análise da tabela de frequências, da Figura 2.16 e da Figura 2.17, verificou-se que o mais usual foi os alunos terem lido 2 livros durante as férias e que apenas 1 aluno leu 4 livro, sendo esse valor considerado um valor atípico.

2.1.5.4 Dados quantitativos contínuos

Enquadrado na política de urbanização duma determinada câmara municipal, relacionada com a definição obrigatória de espaços verdes, foram aprovados 100 projetos de construção de pequenos jardins públicos nesse concelho. As áreas, em m^2, dos referidos jardins distribuem-se da seguinte forma:

Áreas (em m^2)	[300; 600[	[600; 900[	[900; 1200[	[1200; 1500[	[1500; 1800]
N.º de jardins	50	30	9	5	6

a) Identifique a variável em estudo.
b) Construa a tabela de frequências.
c) Represente graficamente a informação anterior.

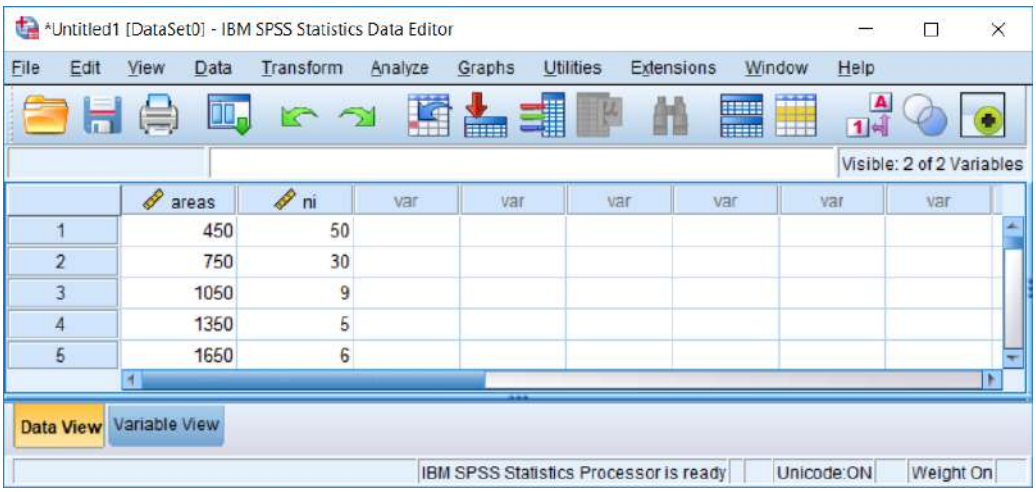
Resolução:

a) A variável em estudo é a área dos jardins públicos cujos projetos foram aprovados.

b)

Áreas (em m^2)	Ponto médio	N.º de jardins	% de jardins	N.º acum. de jardins	% acum. de jardins
[300; 600[	450	50	50	50	50
[600; 900[	750	30	30	80	80
[900; 1200[	1050	9	9	89	89
[1200; 1500[	1350	5	5	94	94
[1500; 1800]	1650	6	6	100	100
Total		100	100		

c) (SPSS)



(SPSS) Data → Weight cases
(⊙ Weight cases by Frequency Variable: ni)
Graphs → Legacy Dialogs → Histogram…
(Variable: areas)

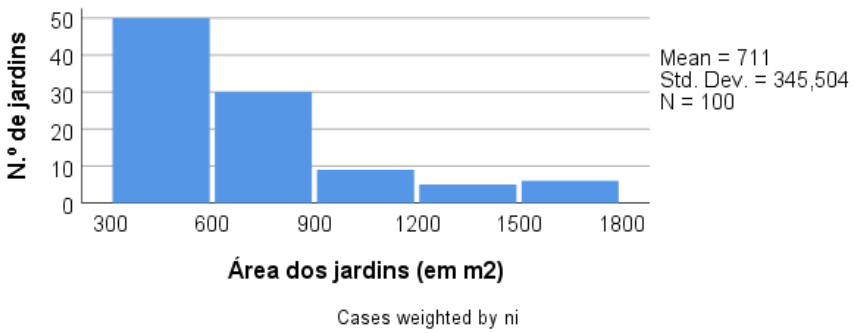
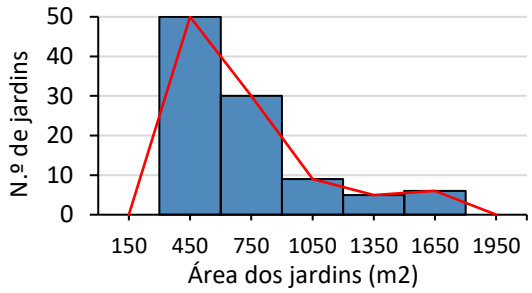
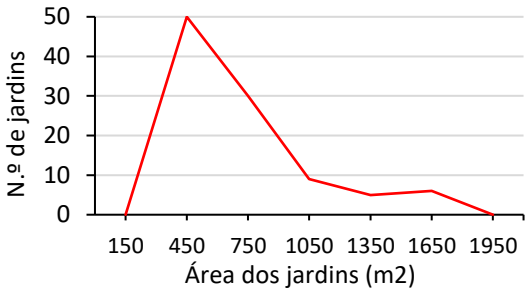


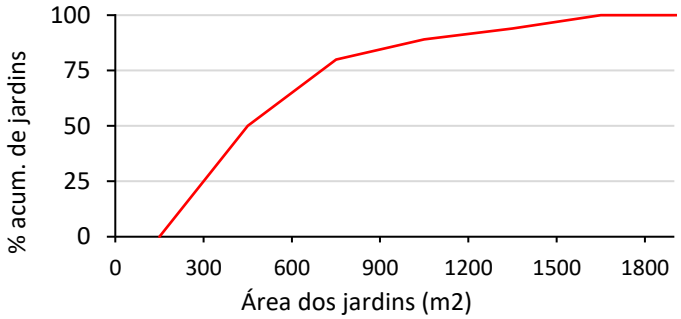
Figura 2.18: *Histograma das frequências simples das áreas dos jardins públicos.*



a) Histograma e polígono de frequências.



b) Polígono de frequências.



d) Polígono de frequências acumuladas

Figura 2.19: *Representação gráfica das frequências simples e acumuladas das áreas dos jardins públicos.*

Através da tabela de frequências, da Figura 2.18 e da Figura 2.19, verifica-se que grande parte dos jardins públicos aprovados têm área inferior a 600 m^2, e são poucos os jardins com mais de 900 m^2.

2.1.5.5 Dados bivariados

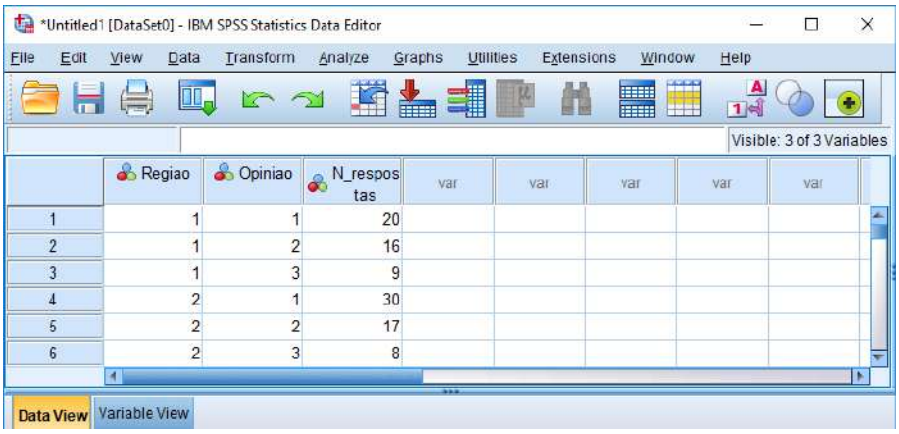
2.1.5.5.1 Duas variáveis qualitativas

Foi realizado um inquérito acerca da opinião dos consumidores sobre um novo detergente, residentes na capital e no interior. Dos 45 residentes na capital, 20 disseram que a opinião era boa, 16 regular e 9 estavam insatisfeitos. Dos 55 residentes no interior, 30 disseram que a opinião era boa, 17 regular e 8 estavam insatisfeitos

a) Represente a informação numa tabela de contingência.
b) Represente graficamente a informação.

Resolução:

a) (SPSS)



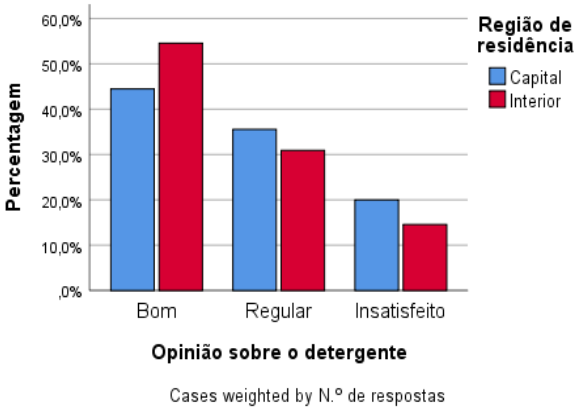
(SPSS) Data → Weight Cases…
(⊙ Weight cases by: N_respostas)

(SPSS) Analyse → Descriptive Statistics → Crosstabs…
(Row(s): Regiao; Columns(s): Opiniao)

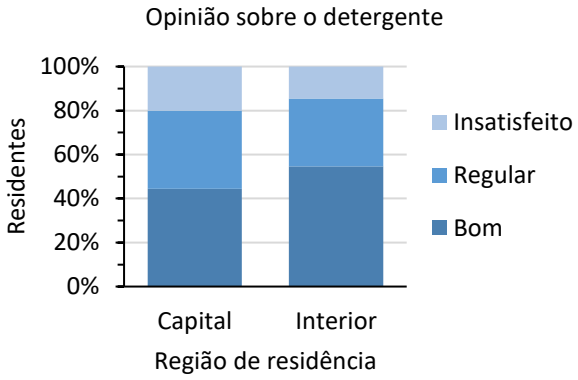
Região de residência * Opinião sobre o detergente Crosstabulation
Count

		Opinião sobre o detergente			
		Bom	Regular	Insatisfeito	Total
Região de residência	Capital	20	16	9	45
	Interior	30	17	8	55
Total		50	33	17	100

b) (SPSS) Graphs → Bar …
(Clustered; ⊙ Summaries for groups of cases; Define;
Bars Represent: ⊙ % of cases (ou N of cases); Category Axis: Opiniao; Define Clusters by: Regiao)



a) Gráfico de barras lado a lado



b) Gráfico de barras empilhadas

Figura 2.20: *Exemplo de gráficos de barras possíveis.*

2.1.5.5.2 Uma variável quantitativa e outra qualitativa

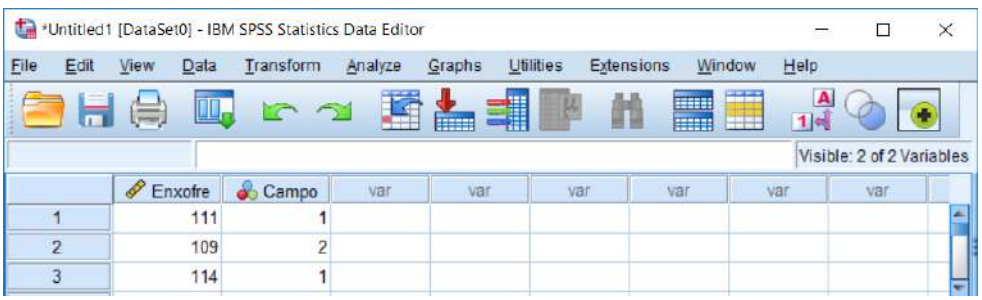
Um determinado método de análise permite determinar o conteúdo de enxofre no petróleo bruto. Os ensaios efetuados em 10 e 8 amostras aleatórias de 1 kg de petróleo bruto, provenientes de furos pertencentes, respetivamente, aos campos A e B, revelaram os seguintes resultados (em gramas):

Campo A:	111	114	105	112	107	109	112	110	110	106
Campo B:	109	103	101	105	106	108	110	104		

Represente graficamente a informação.

Resolução:

(SPSS)



(SPSS) Graphs → Legacy Dialogs → Boxplot…
(Simple; ⊙ Summaries for groups of cases; Define;
Variable: Enxofre; Category Axis: Campo)

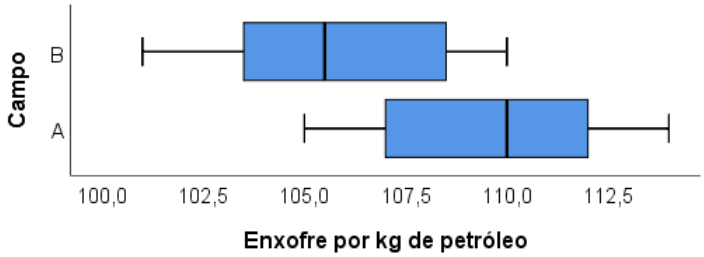


Figura 2.21: *Caixas de bigode lado a lado.*

2.1.5.5.3 Ambas as variáveis quantitativas

Os dados que se apresentam na tabela seguinte representam as notas obtidas, por 15 alunos, na disciplina Estatística e o número de horas dedicadas, por semana, ao estudo.

Aluno	Nota	N.º de horas	Aluno	Nota	N.º de horas	Aluno	Nota	N.º de horas
1	10	3	6	12	10	11	15	16
2	10	4	7	11	4	12	12	11
3	8	2	8	12	7	13	8	9
4	13	13	9	12	13	14	17	20
5	11	6	10	14	7	15	18	20

Construa uma tabela de contingência possível para este conjunto de dados e represente graficamente a informação.

Resolução:

a) (SPSS) Analyse → Descriptive Statistics → Crosstabs…
(Row(s): Horas_Estudo; Columns(s): Nota)

Tabela 2.7: *Exemplo de uma tabela de contingência possível com base num agrupamento de dados previamente elaborado.*

N.º de horas de estudo * Nota na disciplina Crosstabulation
Count

		Nota				
		[0; 10[	[10;14[	[14; 17[	[17; 20]	Total
N.º de horas de estudo	[0; 5[	1	3	0	0	4
	[5; 10[	1	2	1	0	4
	[10; 15[	0	4	0	0	4
	[15; 20]	0	0	1	2	3
Total		2	9	2	2	15

De notar que esta tabela também poderia ser feita com base nos dados originais, mas ficaria muito extensa e com pouca utilidade.

b) (SPSS) Graphs → Legacy Dialogs → Scatter/Dot… → Simple Scatter
(Y Axis: Nota; X Axis: Horas_Estudo)

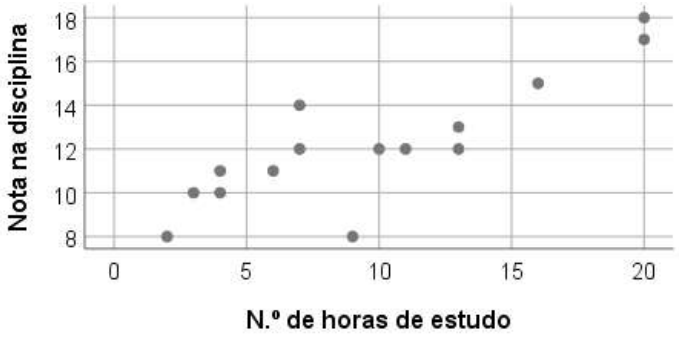


Figura 2.22: *Gráfico de dispersão.*

Através da Tabela 2.7 (é apenas um exemplo possível) verifica-se que grande parte dos alunos teve nota suficiente. Pela análise da Figura 2.22 parece existir uma relação entre a nota na disciplina e o número de horas de estudo, onde os alunos que mais horas dedicaram ao estudo tiveram melhores notas e as piores notas estão associadas a poucas horas de estudo.

2.2 Medidas descritivas

As medidas descritivas fundamentais que se vão estudar numa distribuição de frequências são:

- **Localização:** localizam os valores observados na distribuição.
  - *Tendência central*: média, mediana e moda;
  - *Tendência não central*: quantis (percentis, decis e quartis).
- **Dispersão:** medem o grau de dispersão dos dados em torno de um valor médio.
  - *Absoluta*: amplitude total, amplitude interquartil, desvio padrão e variância;
  - *Relativa*: coeficiente de variação e coeficiente de dispersão.
- **Assimetria:** medem o grau de afastamento da simetria da distribuição.
  - Grau de assimetria de Pearson, grau de assimetria de Bowley, coeficiente de assimetria de Fisher e coeficiente de assimetria amostral.
- **Achatamento:** medem a intensidade das frequências na vizinhança dos valores centrais.
  - Coeficiente de kurtosis, coeficiente percentil de kurtosis e coeficiente de kurtosis amostral.

Tabela 2.8: *Resumo das estatísticas mais úteis para cada escala de medida.*
(Fonte: Adaptado de Pestana e Cageiro, 2014)

Escala nominal	Escala ordinal	Escala quantitativa
Moda.	Moda, Quantis, Amplitude interquartil.	Moda[†], Quantis, Média, Amplitude Interquartil, Amplitude total, Desvio padrão, Variância, Coeficiente de variação, Assimetria, Achatamento.

[†] Note-se que numa variável contínua, em dados não agrupados, por definição, é fácil estar-se perante um fenómeno em que os valores medidos tendam a não se repetir (exemplo, taxa de juro considerando a existência de casas decimais). Nestes casos não faz sentido definir a moda, pois haverá múltiplas modas e esta medida não descreverá nenhuma característica importante dos dados.

Seguidamente apresentam-se as fórmulas de cálculo das principais medidas descritivas distinguindo, quando necessário, as fórmulas de cálculo para:

- Dados agrupados (em categorias ou intervalos, usualmente organizados numa tabela de frequências) e não agrupados;
- Dados discretos e contínuos, sendo as fórmulas para os dados discretos válidas para os dados na escala ordinal.

Por questões práticas, usualmente relacionadas com o uso de programas informáticos de estatística, são utilizados valores para codificar as variáveis qualitativas.

2.2.1 Medidas de localização

As **medidas de localização** informam sobre a localização de alguns valores importantes da distribuição. Estas medidas classificam-se em dois tipos:

- **Medidas de tendência central:** representam os fenómenos pelos seus valores médios, em torno dos quais tendem a concentrar-se os valores observados.
- **Medidas de tendência não central:** fornecem a localização dos valores da variável.

2.2.1.1 Tendência central

2.2.1.1.1 Média aritmética

A **média aritmética** ou, abreviadamente, média, $\overline{x}$, é a medida de localização mais correntemente utilizada.

Dados não agrupados quantitativos	Dados agrupados quantitativos
$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\overline{x} = \frac{1}{n}\sum_{i=1}^{K} n_i x_i' = \sum_{i=1}^{K} f_i x_i'$

Observação: Quando a distribuição é assimétrica[†] é muitas vezes utilizada a **média aparada**, que consiste em ignorar uma percentagem das maiores e menores observações e calcular a média aritmética com base nas restantes observações.

Quando o conjunto de dados é relativo à população, a média aritmética é representada por μ.

2.2.1.1.2 Moda

A moda identifica a categoria, valor ou classe de valores mais comuns no conjunto de dados. Quando a variável é contínua não faz sentido indicar o valor mais observado, mas a classe de valores com maior frequência. Caso seja necessário produzir um valor indicador para a moda, uma alternativa eleger um valor dentro da classe de valores com maior frequência, recorrendo por exemplo à fórmula de King (apresentada de seguida).

[†] Ver Medidas de assimetria, secção 2.2.4.

A **moda**, $\hat{x}$, é o valor que ocorre com maior frequência.

Dados não agrupados ou agrupados qualitativos ou discretos	Dados agrupados contínuos
Valor/categoria que surge com maior frequência.	1º Identificar a classe modal, C_{Mo}, i.e., a classe com maior frequência; 2º Caso seja necessário, o valor da moda é dado por: $\hat{x} = LI_{C_{Mo}} + a_{C_{Mo}} \frac{\Delta_1}{\Delta_1 + \Delta_2}$ $\Delta_1 = n_{C_{Mo}} - n_{C_{Mo-1}}$ e $\Delta_2 = n_{C_{Mo}} - n_{C_{Mo+1}}$, ou $\Delta_1 = f_{C_{Mo}} - f_{C_{Mo-1}}$ e $\Delta_2 = f_{C_{Mo}} - f_{C_{Mo+1}}$

Notação:

$LI_{C_{Mo}}$	limite inferior da classe modal,
$a_{C_{Mo}}$	amplitude da classe modal,
$n_{C_{Mo}}$	frequência absoluta da classe modal,
$n_{C_{Mo-1}}$	frequência absoluta da classe anterior à classe modal,
$n_{C_{Mo+1}}$	frequência absoluta da classe posterior à classe modal,
$f_{C_{Mo}}$	frequência relativa da classe modal,
$f_{C_{Mo-1}}$	frequência relativa da classe anterior à classe modal,
$f_{C_{Mo+1}}$	frequência relativa da classe posterior à classe modal.

Observação: Quando as classes têm amplitudes diferentes devem-se utilizar as frequências simples corrigidas, n_i^* ou f_i^*, conforme já foi referido (ver 2.1.2.4).

A moda pode ser determinada graficamente através do histograma, conforme ilustrado na figura seguinte.

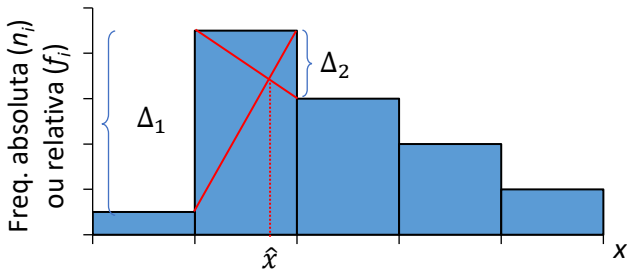


Figura 2.23: *Determinação gráfica da moda.*

2.2.1.1.3 Mediana

Quando os dados são contínuos a mediana corresponde ao valor, no eixo das ordenadas, que divide a área do histograma em duas partes iguais. Pode ser determinada graficamente através do polígono de frequências acumuladas (Figura 2.24).

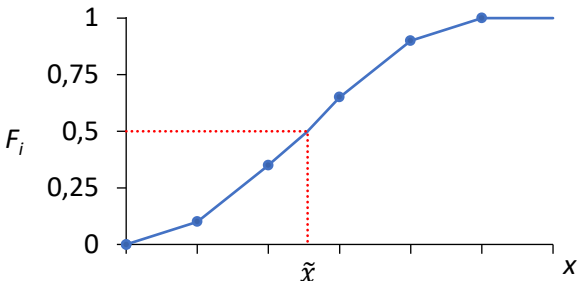


Figura 2.24: *Determinação gráfica da mediana.*

A **mediana**, $\tilde{x}$, é o valor na amostra ordenada que tem tantos valores inferiores ou iguais como superiores ou iguais a ele.

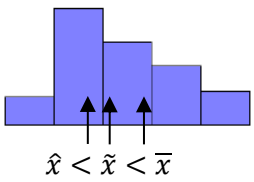
Dados ordinais, discretos ou não agrupados contínuos	Dados agrupados contínuos
1º Ordenar a amostra por ordem crescente; 2º O valor da mediana é dado por: $\tilde{x} = \begin{cases} \frac{x_{\frac{n}{2}:n} + x_{\frac{n}{2}+1:n}}{2}, & \text{se } n \text{ par} \\ x_{\frac{n+1}{2}:n}, & \text{se } np \text{ ímpar} \end{cases}$	1º Identificar a classe mediana, C_{Me}, cujo $N_i = n/2$ ou $F_i = 0{,}5$, i.e., a classe que contém a mediana. 2º O valor da mediana é dado por: $\tilde{x} = LI_{C_{Me}} + a_{C_{Me}} \frac{0{,}5n - N_{C_{Me-1}}}{n_{C_{Me}}}$ $= LI_{C_{Me}} + a_{C_{Me}} \frac{0{,}5 - F_{C_{Me-1}}}{f_{C_{Me}}}$

Notação:

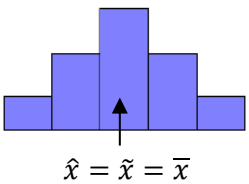
$x_{i:n}$	i-ésima observação na amostra ordenada de dimensão n,
$LI_{C_{Me}}$	limite inferior da classe mediana,
$a_{C_{Me}}$	amplitude da classe mediana,
$N_{C_{Me-1}}$	frequência absoluta acumulada da classe anterior à classe mediana,
$n_{C_{Me}}$	frequência absoluta da classe mediana,
$F_{C_{Me-1}}$	frequência relativa acumulada da classe anterior à classe mediana,
$f_{C_{Me}}$	frequência relativa da classe mediana.

2.2.1.1.4 Comparação entre a média, mediana e moda

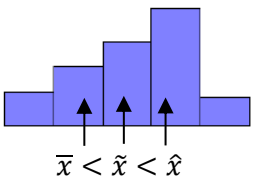
Através da análise do histograma é possível inferir sobre a assimetria da distribuição (Figura 2.25).



b) Assimetria positiva (Moda < Mediana < Média).



a) Simetria (Moda = Mediana = Média).



c) Assimetria negativa (Média < Mediana < Moda).

Figura 2.25: *Exemplos de assimetria no histograma.*

A média é calculada com base em todos os valores observados, tornando-a sensível a valores anómalos (Figura 2.26). A mediana, considerando que seu cálculo é feito apenas com base nos dois valores centrais, é mais robusta (menos sensível) a valores extremos.

Figura 2.26: *Comparação entre a mediana e a média. (Fonte: http://www.alea.pt).*

2.2.1.2 Tendência não central

2.2.1.2.1 Quantis

O **quantil Q_p^*** divide a amostra em duas partes tais que:

- Na 1ª parte $100 \times p\%$ dos elementos são menores ou iguais a Q_p^*;
- Na 2ª parte $100 \times (1-p)\%$ dos elementos são maiores ou iguais a Q_p^*;

$$Q_p^* = \begin{cases} \text{Quartis} = Q_i, & p = \frac{i}{4}, i = 1,2,3; \\ \text{Decis} = D_i, & p = \frac{i}{10}, i = 1,2,\dots,9; \\ \text{Percentis} = P_i, & p = \frac{i}{100}, i = 1,2,\dots,99. \end{cases}$$

Dados ordinais, discretos ou não agrupados contínuos	Dados agrupados contínuos
1º Ordenar a amostra por ordem crescente; 2º O valor do quantil é dado por: $Q_p^* = \begin{cases} \frac{x_{np:n} + x_{np+1:n}}{2}, & \text{se } np \text{ inteiro;} \\ x_{[np]+1:n}, & \text{se } np \text{ não inteiro;} \end{cases}$ onde $[np]$ = maior inteiro contido em np.	1º Identificar a classe quantil, C_Q, cujo $N_i = np$ ou $F_i = p$, i.e., a classe que contém o quantil p. 2º O valor do quantil é dado por: $Q_p^* = LI_{C_Q} + a_{C_Q} \frac{np - N_{C_{Q-1}}}{n_{C_Q}}$ $= LI_{C_Q} + a_{C_Q} \frac{p - F_{C_{Q-1}}}{f_{C_Q}}$

Notação:

$x_{i:n}$	i-ésima observação na amostra ordenada de dimensão n,
LI_{C_Q}	limite inferior da classe quantil,
a_{C_Q}	amplitude da classe quantil,
$N_{C_{Q-1}}$	frequência absoluta acumulada da classe anterior à classe quantil,
n_{C_Q}	frequência absoluta da classe quantil,

$F_{C_{Q-1}}$ frequência relativa acumulada da classe anterior à classe quantil,
f_{C_Q} frequência relativa da classe quantil.

Observações:

- $Q_1 = P_{25}$;
- $Q_2 = \tilde{x} = D_5 = P_{50}$;
- $Q_3 = P_{75}$;
- $D_i = P_{i\times 10}$;
- Murteira *et al.* (2007) sugerem que o cálculo dos quantis, no caso dos dados não agrupados em classes de intervalos, se proceda da seguinte forma:
  1º Calcular o valor de $r = 1 + (n-1)p$;
  2º Se r inteiro então $Q_p^* = x_{r:n}$;
  Se r não inteiro então $r = a + d$ onde a é a parte inteira e d a parte décimal e então
  $$Q_p^* = x_{a:n} + d(x_{a+1:n} - x_{a:n}) = (1-d)x_{a:n} + dx_{a+1:n}.$$

Os quantis podem ser determinados graficamente através do polígono de frequências acumuladas (Figura 2.27).

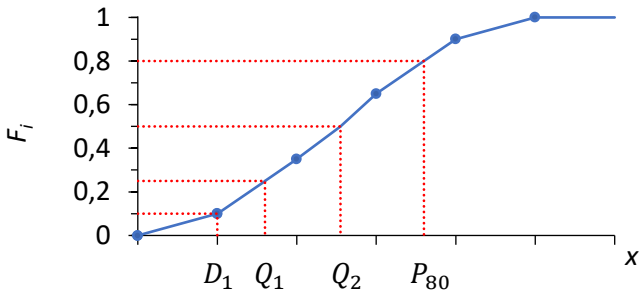


Figura 2.27: Determinação gráfica de diversos quantis.

2.2.2 Medidas de dispersão

Muitas vezes verifica-se que duas distribuições podem ter o mesmo valor central, mas apresentam distribuições completamente distintas, não sendo por isso possível uma caracterização correta dos valores observados apenas a partir desse valor central. Por exemplo, podem existir duas turmas com médias de 15 valores: na primeira turma todos os alunos tiveram nota 15, ao passo que na segunda turma metade dos alunos teve 10 valores e a outra metade teve 20 valores. Em termos médios estas turmas são idênticas, mas apresentam variações de valores completamente distintas, sendo então necessário recorrer às medidas de dispersão para uma caracterização mais completa.

As medidas de dispersão utilizam-se para estudar a concentração dos valores de um conjunto de dados em torno de uma medida de localização central do conjunto. Desta forma, quanto menor for a dispersão desses valores em relação à medida de localização central, mais representativa será essa medida desse conjunto de dados.

2.2.2.1 Medidas absolutas

As **medidas de dispersão absoluta** dependem das unidades em que a variável é expressa, pelo que não servem para comparar duas ou mais distribuições relativamente à dispersão.

2.2.2.1.1 Amplitude total

A **amplitude total**, $\boldsymbol{a}$, é a diferença entre a observação maior e a mais pequena:

$$a = x_{n:n} - x_{1:n} = \text{máximo} - \text{mínimo}.$$

2.2.2.1.2 Amplitude interquartil

A **amplitude interquartil**, $\boldsymbol{AIQ}$, é a diferença entre o 3º quartil e o 1º quartil. Corresponde a um intervalo que engloba 50% das observações centrais:

$$AIQ = Q_3 - Q_1.$$

Observação: Uma vez que esta medida não é afetada por valores extremos, deve ser utilizada para distribuições altamente assimétricas[†].

2.2.2.1.3 Desvio médio absoluto

O **desvio médio absoluto**, $\boldsymbol{dm}$, é a média dos desvios absolutos entre os valores observados e a média.

Dados não agrupados quantitativos	Dados agrupados quantitativos
$dm = \frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \overline{x}\rvert$	$dm = \frac{1}{n}\sum_{i=1}^{K} n_i\lvert x_i' - \overline{x}\rvert$

Observação: Esta medida só assume valores não negativos e quanto maior o seu valor maior a dispersão.

2.2.2.1.4 Variância

A **variância amostral**, $\boldsymbol{s^2}$, é a média dos quadrados dos desvios entre os valores observados e a média.

Dados não agrupados quantitativos	Dados agrupados quantitativos
$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{K} n_i(x_i' - \overline{x})^2$
$= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - n\overline{x}^2\right)$	$= \frac{1}{n-1}\sum_{i=1}^{K}\left(n_i x_i'^2 - n\overline{x}^2\right)$

Observações:

- Esta medida só assume valores não negativos e quanto maior o seu valor maior a dispersão.
- Para calcular a variância populacional, σ^2, basta substituir no denominador da variância amostral $n-1$ por n.
- A variância tem como desvantagem o facto de ser expressa em unidades ao quadrado, o que torna difícil a sua interpretação, razão pela qual se utiliza o desvio padrão.

[†] Ver Medidas de assimetria, secção 2.2.4.

2.2.2.1.5 Desvio padrão

O **desvio padrão amostral**, $\boldsymbol{s}$, é a medida de dispersão mais utilizada. O valor desta medida é obtido fazendo $\sqrt{\text{variância}}$.

Dados não agrupados quantitativos	Dados agrupados quantitativos
$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} n_i\left(x_i' - \overline{x}\right)^2}$
$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - n\overline{x}^2\right)}$	$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(n_i x_i'^2 - n\overline{x}^2\right)}$

Observações:

- Esta medida só assume valores não negativos e quanto maior o seu valor maior a dispersão.
- Para calcular o desvio padrão populacional, σ, basta substituir no denominador do desvio padrão amostral $n-1$ por n.
- Propriedades para dados com distribuição aproximadamente normal[†]:
  - Aproximadamente 68% dos dados estão no intervalo $[\overline{x} - s;\ \overline{x} + s]$;
  - Aproximadamente 95% dos dados estão no intervalo $[\overline{x} - 2s;\ \overline{x} + 2s]$;
  - Aproximadamente 100% dos dados estão no intervalo $[\overline{x} - 3s;\ \overline{x} + 3s]$.

2.2.2.2 Medidas relativas

As **medidas de dispersão relativa** não dependem das unidades em que a variável é expressa, pelo que são úteis para comparar duas ou mais distribuições relativamente à dispersão.

O **coeficiente de variação**, $\boldsymbol{CV}$, mede o grau de concentração em torno da média, em valor percentual:

$$CV = \frac{s}{\overline{x}} \times 100.$$

O **coeficiente de dispersão**, $\boldsymbol{CD}$, mede o grau de concentração em torno da média. É dado pelo quociente entre o desvio padrão e a média:

$$CD = \frac{s}{\overline{x}}.$$

Observações:

- Estes coeficientes só podem ser calculados quando a variável toma valores de um só sinal, isto é, todos os valores são todos positivos ou são todos negativos.
- Para valores inferiores a 50% do coeficiente de variação (ou 0,5 do coeficiente de dispersão) a média será tanto mais representativa quanto menor o valor deste coeficiente. Consequentemente, valores superiores a 50% do coeficiente de variação (ou 0,5 do coeficiente de dispersão) indicam uma pequena representatividade da média.

[†] Ver distribuição Normal, secção 5.2.2.

2.2.3 Momentos

2.2.3.1 Momento de ordem r em relação a um valor fixo V

O momento de ordem r em relação a um valor fixo V, $\boldsymbol{m'_{r,V}}$, é a média dos desvios, elevados à ordem r, entre os valores observados e um valor fixo V.

Dados não agrupados quantitativos	Dados agrupados quantitativos
$m'_{r,V} = \frac{1}{n}\sum_{i=1}^{n}(x_i - V)^r$	$m'_{r,V} = \frac{1}{n}\sum_{i=1}^{K} n_i(x'_i - V)^r$

Observações:

- Os momentos teóricos representam-se por μ em vez de m e correspondem ao caso em que se conhece toda a população.
- Designa-se por **momento central de ordem r**, ou **r -ésimo momento central** ou **momento de ordem r em relação à média**, m_r, quando $V = \overline{x}$ e verifica-se que:
  - O 1º momento central é sempre nulo, i. e., $m_1 = 0$;
  - O 2º momento central está relacionado com a variância amostral:

$$m_2 = \frac{n-1}{n}s^2;$$

  - Na população, o 2º momento central, μ_2, é igual à variância populacional, i. e.,

$$\mu_2 = \sigma^2;$$

  - Numa distribuição simétrica, todos os momentos centrais de ordem ímpar são nulos.
- Designa-se por **momento de ordem r em relação à origem**, ou **r -ésimo momento** ou **momento de ordem r**, $\boldsymbol{m'_r}$, quando $V = 0$. Para este caso particular verifica-se que:
  - O 1º momento em relação à origem é igual à média, i. e., $m'_1 = \overline{x}$.

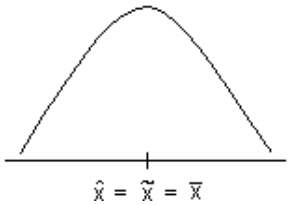
2.2.3.2 Relação entre os momentos

Entre os momentos em relação à média e os momentos em relação a um valor arbitrário V estabelecem-se as seguintes relações:

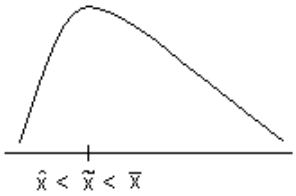
$$m_2 = m'_{2,V} - \left(m'_{1,V}\right)^2;$$
$$m_3 = m'_{3,V} - 3m'_{1,V}m'_{2,V} + 2\left(m'_{1,V}\right)^3;$$
$$m_4 = m'_{4,V} - 4m'_{1,V}m'_{3,V} + 6\left(m'_{1,V}\right)^2 m'_{2,V} - 3\left(m'_{1,V}\right)^4$$

2.2.4 Medidas de assimetria

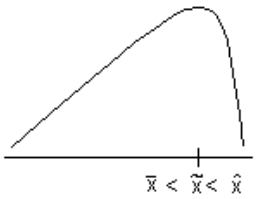
As medidas de assimetria permitem medir o grau de afastamento da simetria da distribuição (Figura 2.28). Quando no conjunto de dados predominam os valores menores (maiores) a diz-se que a distribuição é assimétrica positiva (negativa) e tem uma "cauda" à direita (esquerda). Caso contrário a distribuição é simétrica.



a) Distribuição simétrica ou não enviesada.



b) Distribuição assimétrica positiva ou enviesada à esquerda.



c) Distribuição assimétrica negativa ou enviesada à direita.

Figura 2.28: *Tipos de assimetria.*

O coeficiente de Fisher é o coeficiente de assimetria teórico, que representa o verdadeiro valor da assimetria da distribuição, e que só deve ser usado quando se conhece toda a população. Os coeficientes de Pearson e de Bowley são empíricos e têm como principal vantagem a sua facilidade de cálculo, hoje em dia ultrapassada pela utilização frequente de programas de estatística.

De referir que para certos conjuntos de dados, os coeficientes podem indiciar conclusões diferentes. Por exemplo, o coeficiente de Bowley não é sensível à existência de valores atípicos.

2.2.4.1 Coeficiente de assimetria de Fisher

O **grau de assimetria de Fisher**, $\boldsymbol{\gamma_1}$, é dado por:

$$\gamma_1 = \frac{\mu_3}{\sigma^2},$$

onde μ_3 representa o 3º momento teórico. O sinal de γ_1 é o sinal da assimetria.

2.2.4.2 Grau de assimetria de Pearson

O **grau de assimetria de Pearson**, $\boldsymbol{g_P}$, é dado por:

$$g_P = \frac{\overline{x} - \hat{x}}{s}, -3 < g_P < 3.$$

Para $g_P \approx 0$ a distribuição é simétrica; para $g_P \approx 3$ a distribuição é assimétrica positiva; para $g_P \approx -3$ a distribuição é assimétrica negativa.

Observação: O grau de assimetria de Pearson só pode ser utilizado quando a distribuição é unimodal, ou seja, só tem uma moda.

2.2.4.3 Grau de assimetria de Bowley

O **grau de assimetria de Bowley**, $\boldsymbol{g_B}$, é dado por:

$$g_B = \frac{(Q_3 - \tilde{x}) - (\tilde{x} - Q_1)}{Q_3 - Q_1}, -1 < g_B < 1.$$

Para $g_B \approx 0$ a distribuição é simétrica; para $g_B \approx 1$ a distribuição é assimétrica positiva; para $g_B \approx -1$ a distribuição é assimétrica negativa.

Observação: O grau de assimetria de Bowley deve ser utilizado quando se desconhece a média e o desvio padrão.

2.2.4.4 Coeficiente de assimetria amostral

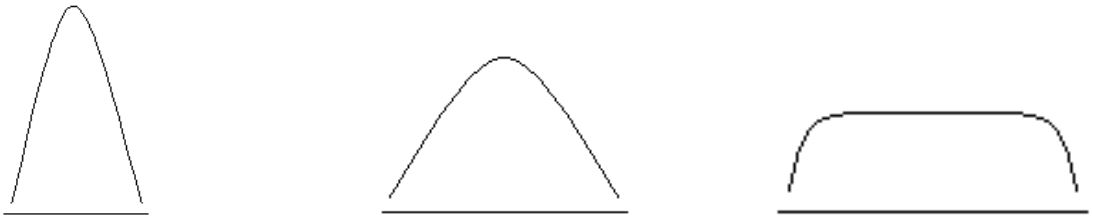
O **coeficiente de assimetria amostral** utilizado por vários softwares, como sejam SPSS, Excel e SAS, $\boldsymbol{g_a}$, é dado por:

$$g_a = \frac{n^2 m_3}{(n-1)(n-2)s^3}.$$

O sinal de g_a é o sinal da assimetria.

2.2.5 Medidas de achatamento

As medidas de achatamento dão-nos uma indicação sobre a intensidade das observações em torno dos valores centrais (Figura 2.29), ou seja, se é mais achatada (platicúrtica) ou mais pontiaguda (leptocúrtica) quando comparada a distribuição normal[†] (mesocúrtica).



a) Distribuição leptocúrtica. *b) Distribuição mesocúrtica.* *c) Distribuição platicúrtica.*

Figura 2.29: *Tipos de achatamento.*

O coeficiente de Kurtosis é o coeficiente de achatamento teórico que representa o verdadeiro valor do achatamento da distribuição, pelo que só deve ser usado quando se conhece toda a população. O coeficiente percentil de Kurtosis é empírico e tem como principal vantagem a sua facilidade de cálculo, hoje em dia ultrapassada pela utilização frequente de programas de estatística. De referir que estes coeficientes podem classificar o achatamento dum conjunto de dados de forma diferente. Por exemplo, o coeficiente percentil de kurtosis não é sensível à existência de valores muito grandes ou muito pequenos no conjunto de dados.

2.2.5.1 Coeficiente de kurtosis

O **coeficiente de kurtosis**, $\boldsymbol{\gamma_2}$, é dado por:

$$\gamma_2 = \frac{\mu_4}{\sigma^4},$$

onde μ_4 representa o 4º momento teórico. Para $\gamma_2 > 3$ a distribuição é leptocúrtica; para $\gamma_2 \approx 3$ a distribuição é mesocúrtica; para $\gamma_2 < 3$ a distribuição é platicúrtica.

Os achatamentos das curvas simétricas são tomados em relação ao da curva Normal para a qual $\gamma_2 = 3$.

[†] Ver distribuição Normal, secção 5.2.2.

2.2.5.2 Coeficiente percentil de kurtosis

O **coeficiente percentil de kurtosis**, $\boldsymbol{k_P}$, é dado por

$$k_P = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}.$$

Para $k_P < 0{,}263$ a distribuição é leptocúrtica; para $k_P \approx 0{,}263$ a distribuição é mesocúrtica; para $k_P > 0{,}263$ a distribuição platicúrtica.

2.2.5.3 Coeficientes de kurtosis amostral

O coeficiente de kurtosis habitualmente utilizado pelos softwares, tais como SPSS, Excel e SAS, $\boldsymbol{k_a}$, é dado por:

$$k_a = \frac{n^2(n+1)m_4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)}.$$

Para $k_a > 0$ a distribuição é leptocúrtica; para $k_a \approx 0$ a distribuição é mesocúrtica; para $k_a < 0$ a distribuição é platicúrtica.

2.2.6 Medidas de concentração

Considere-se a tabela de frequências (Tabela 2.9), onde y_i é o total da característica correspondente aos indivíduos ou elementos da i-ésima classe.

Tabela 2.9: *Tabela de frequências.*

Classes	Ponto médio (x_i')	Freq. absolutas (n_i)	$y_i = n_i x_i'$
$[LI_1; LS_1[$	x_1'	n_1	y_1
$[LI_2; LS_2[$	x_2'	n_2	y_2
…	…	…	…
$[LI_K; LS_K]$	x_K'	n_K	y_K
Total		n	

Definam-se as frequências acumuladas,

$$p_i = \frac{\sum_{j=1}^{i} n_j}{n} = \frac{N_i}{n} = F_i \quad \text{e} \quad q_i = \frac{\sum_{j=1}^{i} n_j x_j'}{\sum_{j=1}^{K} n_j x_j'} = \frac{\sum_{j=1}^{i} y_j}{\sum_{j=1}^{K} y_j}, \quad i = 1, 2, \ldots, K,$$

onde:

- p_i representa a proporção de indivíduos que possuem a característica com uma intensidade inferior ao limite superior da i-ésima classe, LS_i;
- q_i representa a proporção da totalidade da característica possuída pelos indivíduos que possuem a característica com uma intensidade inferior ao limite superior da i-ésima classe, LS_i.

Os valores p_i e q_i, $i = 1, 2, \ldots, K$, satisfazem as relações: $p_i \geq q_i$; $0 \leq p_i \leq 1$; $0 \leq q_i \leq 1$.

2.2.6.1 Curva de Lorenz

A **curva de Lorenz** obtém-se representando os pontos $(p_i, q_i), i = 1, 2, \dots, K$, num sistema de eixos cartesianos e unindo os mesmos por meio de segmentos de recta.

Se houver igual distribuição, os valores p_i e q_i são iguais e a curva de Lorenz degenera na diagonal que se designa por **reta de igual distribuição**. A área compreendida entre a reta de igual distribuição e a curva de Lorenz é designada por **área de concentração**. Quanto maior for esta área mais elevada será a concentração. Na Figura 2.30 apresenta-se o aspeto genérico da curva de Lorenz.

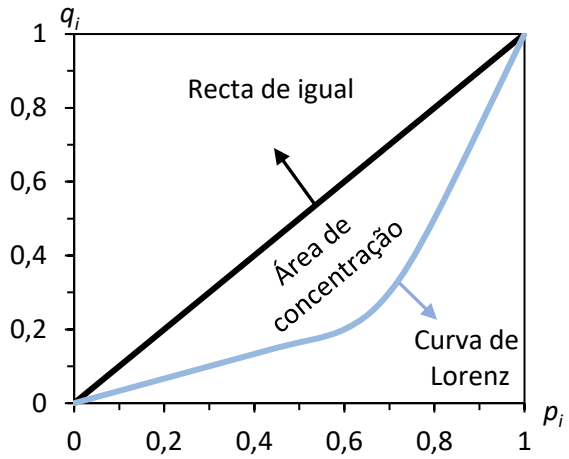


Figura 2.30: *Curva de Lorenz.*

2.2.6.2 Índice de concentração de Gini

O **índice de concentração de Gini**, IG, mede a concentração de uma determinada característica numa população. É dado por:

$$IG = \frac{\sum_{i=1}^{K-1}(p_i - q_i)}{\sum_{i=1}^{K-1} p_i} = 1 - \frac{\sum_{i=1}^{K-1} q_i}{\sum_{i=1}^{K-1} p_i}.$$

Características:

- $IG = 0$ quando há igual distribuição, i. e., $p_i = q_i$;
- $IG = 1$ quando a concentração for máxima, i. e., $q_i = 0$;
- Cresce com o aumento de concentração da característica em estudo.

2.2.7 Covariância e correlação

A covariância e a correlação são medidas estatísticas que permitem medir o grau de associação entre duas variáveis emparelhadas, ou seja, valores de duas variáveis que foram obtidos sobre o mesmo indivíduo ou unidade. Na aplicação destas medidas, tal como com todas as outras já apresentadas anteriormente, é necessário ter em conta o tipo de variáveis em estudo e também, em alguns casos, a própria distribuição dos dados.

2.2.7.1 Covariância

A **covariância amostral** entre x e y (amostra bivariada), $\boldsymbol{s_{xy}}$, mede o grau de associação linear entre duas variáveis quantitativas.

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right).$$

Observações:

- A covariância tem como desvantagem o facto de ser expressa nas unidades de medida das variáveis, o que dificulta a sua interpretação, razão pela qual se utiliza o coeficiente de correlação linear de Pearson.
- Para calcular a covariância populacional, σ_{xy}, basta substituir no denominador da covariância amostral $n-1$ por n.

2.2.7.2 Coeficiente de correlação de Pearson

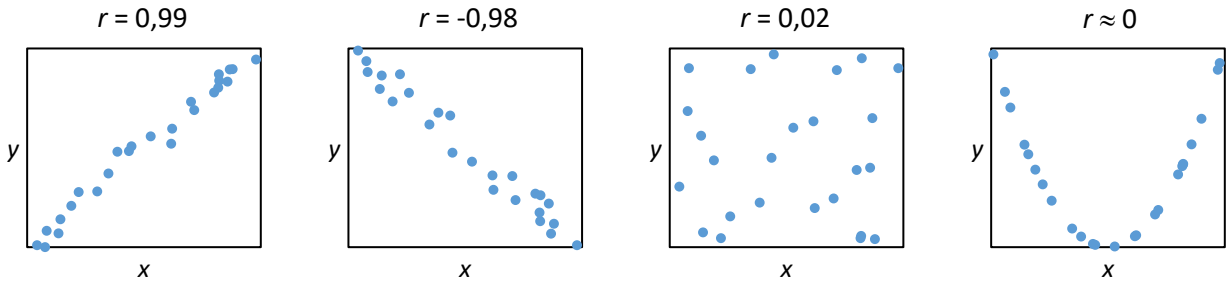
O **coeficiente de correlação de Pearson**, $\boldsymbol{r}$, mede o grau de associação linear entre x e y (amostra bivariada quantitativa) e é dado por:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, -1 < r < 1.$$

Para $r \approx -1$ existe associação linear negativa muito alta; para $r \approx 0$ não existe associação linear entre as variáveis; para $r \approx 1$ existe associação linear positiva muito alta.

No contexto inferencial, a abordar posteriormente, este coeficiente assume a normalidade dos dados.

Na Figura 2.31 apresentam-se vários exemplos de diferentes relações e os respetivos coeficientes de correlação de Pearson.



a) Correlação linear positiva elevada. | *b) Correlação linear negativa elevada.* | *c) Inexistência de correlação.* | *d) Inexistência de correlação linear.*

Figura 2.31: *Tipos de relação.*

Pestana e Gageiro (2014) sugerem, apenas por convenção, as seguintes interpretações com base nos valores absolutos de r: se $0 \leq |r| < 0{,}2$ não existe correlação ou é desprezável; se $0{,}2 \leq |r| < 0{,}7$ a correlação é moderada; se $0{,}7 \leq |r| < 0{,}9$ a correlação é forte; e no extremo se $|r| \geq 0{,}9$ a correlação é muito forte. Existem outras sugestões de classificação, mas esta parece ser a atualmente mais utilizada, reforçando-se sempre a importância de previamente se fazer uma análise gráfica desta relação.

Observação: Quando a correlação é significativa o primeiro passo a efetuar é investigar se não será apenas uma associação espúria, ou seja, não se consegue estabelecer qualquer relação causal entre as variáveis (Figura 2.32).

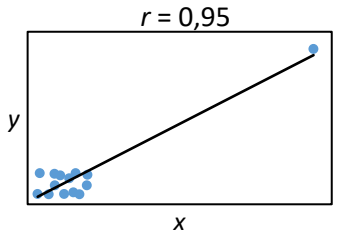


Figura 2.32: *Exemplo de correlação espúria.*

Exemplo de correlação espúria (Pestana e Silvio, 2002, pág. 146 e 147):

> *"Uma publicação anti-clerical que ficou célebre mostrava claramente que o número de crimes nas cidades inglesas tinha crescido com o aumento do número de pastores anglicanos, durante o século XIX. Ainda que os dados fossem correctos, e a correlação próxima de 1, inferir uma relação de causa a efeito é obviamente um disparate. A revolução industrial alterou a demografia consideravelmente, houve um aumento populacional importante no século XIX com grande concentração nas cidades – e o que é facto razoável é considerar que o número de crimes nas cidades aumentou com a concentração populacional, tal como o número de padres (e polícias, e advogados, etc.) aumentou o crescimento populacional. Os dois fenómenos estão fortemente correlacionados com o aumento populacional (esse sim, os explica), pelo que estão correlacionados entre si: mas não fica estabelecida qualquer relação causal entre eles."*

Quando a relação entre as variáveis não é linear ou pelo menos uma delas não é quantitativa deve-se utilizar outro coeficiente de correlação (Murteira *et al.*, 2007).

2.2.7.3 Coeficiente de correlação de Spearman

O **coeficiente de correlação de Spearman**, r_S, mede o grau de associação entre entre x e y (amostra bivariada) ordinais ou qualitativas (quando não for adequada a aplicação do coeficiente de correlação de Pearson). Este coeficiente não é sensível a assimetrias nem a valores atípicos e é dado por:

$$r_S = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2 \, , -1 \leq r_S \leq 1.$$

onde d_i é a diferença entre os valores de ordem de x e y.

Para $r_S \approx -1$ existe associação negativa muito alta entre as ordenações das variáveis; para $r_S \approx 0$ não existe associação entre as ordenações das variáveis; para $r_S \approx 1$ existe associação positiva muito alta entre as ordenações das variáveis.

No contexto inferencial, a abordar posteriormente, este coeficiente não exige a normalidade dos dados.

2.2.8 Exercícios resolvidos

2.2.8.1 Dados qualitativos na escala nominal

Determine e interprete a moda dos seguintes dados (Tabela 2.10) referentes ao exercício da secção 2.1.5.1.

Tabela 2.10: *Tabela de frequências relativa ao tipo preferido de bebida pelas crianças com mais de 3 anos do jardim-de-infância O Parque da Pequenada.*

Tipo de Bebida	x_i'	N.º de crianças (n_i)	Prop. de crianças (f_i)
Leite	1	30	0,188
Água	2	10	0,063
Sumos naturais	3	40	0,250
Refrigerantes	4	80	0,500
Total		160	1,0000

Resolução:

Moda: $\hat{x} = 4 =$ Refrigerantes (categoria com maior frequência).
O tipo de bebida que estas crianças preferem é o refrigerante.

(SPSS) Analyse → Descriptive Statistics → Frequencies…
(Variable: Bebida; Statistics: ☑ Mode)

Statistics
Tipo de bebida

N	Valid	160
	Missing	0
Mode		4

2.2.8.2 Dados qualitativos na escala ordinal

Determine e interprete a mediana, a moda, os quartis, o 4º decil, os percentis 15 e 46 e a amplitude interquartil, dos dados apresentados anteriormente na secção 2.1.5.2 cuja tabela de frequências (Tabela 2.11) é a que se segue.

Tabela 2.11: *Tabela de frequências relativa à qualidade do serviço prestado num determinado Centro de Saúde.*

Qualidade do Serviço	x_i'	N.º de utentes (n_i)	Prop. de utentes (f_i)	N.º ac. de utentes (N_i)	Prop. ac. de utentes (F_i)
Muito mau	1	1	0,02	1	0,02
Mau	2	1	0,02	2	0,04
Médio	3	6	0,12	8	0,16
Bom	4	15	0,30	23	0,46
Muito bom	5	27	0,54	50	1,00
Total		50	1,00		

Resolução:

Mediana:

$$n = 50 \text{ (é par)} \rightarrow \tilde{x} = \frac{x_{25:50} + x_{26:50}}{2} = \frac{5+5}{2} = 5.$$

Portanto, metade dos utentes considera que no mínimo a qualidade do serviço foi muito boa.

Moda:

$$\hat{x} = 5 = \text{Muito bom (categoria com maior frequência).}$$

O mais frequente foi os utentes considerarem que a qualidade do serviço prestado nesse Centro de Saúde é muito boa.

1º Quartil:

$$p = \frac{1}{4} = 0{,}25 \rightarrow np = 12{,}5 \text{ (não é inteiro)} \rightarrow Q_1 = x_{[12,5]+1:50} = x_{12+1:50} = x_{13:50} = 4.$$

25% dos utentes consideram que a qualidade do serviço prestado foi no máximo (inferior ou igual) boa e, portanto, os restantes 75% consideram que a qualidade do serviço foi no mínimo (superior ou igual) boa.

3º Quartil:

$$p = \frac{3}{4} = 0{,}75 \ \rightarrow \ np = 37{,}5 \text{ (não é inteiro)} \rightarrow Q_3 = x_{[37,5]+1:50} = x_{37+1:50} = x_{38:50} = 5$$

75% dos utentes consideram que a qualidade do serviço prestado foi no máximo (inferior ou igual) muito boa e, portanto, os restantes 25% consideram a qualidade do serviço prestado foi no mínimo (superior ou igual) muito boa.

4º Decil ou 40º Percentil:

$$p = \frac{4}{10} = 0{,}4 \ \rightarrow \ np = 20 \text{ (é inteiro)} \rightarrow D_4 = P_{40} = \frac{x_{20:50} + x_{21:50}}{2} = \frac{4+4}{2} = 4.$$

40% dos utentes consideram que a qualidade do serviço prestado foi no máximo (inferior ou igual) boa e, portanto, os restantes 60% consideram que o nível da qualidade do serviço prestado foi no mínimo (superior ou igual) boa.

15º Percentil:

$$p = \frac{15}{100} = 0{,}15 \ \rightarrow \ np = 7{,}5 \text{ (não é inteiro)} \rightarrow P_{15} = x_{[7,5]+1:50} = x_{7+1:50} = x_{8:50} = 3.$$

15% dos utentes consideram que a qualidade do serviço prestado foi no máximo (inferior ou igual) média e, portanto, os restantes 85% consideram que a qualidade do serviço foi no mínimo (superior ou igual) média.

46º Percentil:

$$p = \frac{46}{100} = 0{,}46 \ \rightarrow \ np = 23 \text{ (é inteiro)} \rightarrow P_{46} = \frac{x_{23:50} + x_{24:50}}{2} = \frac{4+5}{2} = 4{,}5.$$

46% dos utentes consideram que a qualidade do serviço foi no máximo (inferior ou igual) boa e, portanto, os restantes 54% consideram que a qualidade do serviço prestado foi muito boa.

Amplitude interquartil:

$$AIQ = Q_3 - Q_1 = 5 - 4 = 1.$$

Quando se ignoram 25% das piores opiniões e 25% das opiniões melhores, a diferença entre as restantes opiniões é de apenas uma categoria, ou seja, a opinião é muito semelhante.

(SPSS) Analyse → Descriptive Statistics → Frequencies…

(Variable: Qualidade; Statistics: ☑ Median; ☑ Mode; ☑ Quartiles; ☑ Percentile(s): 40, 15, 46)

Statistics

Opinião sobre a qualidade do serviço

N	Valid	50
	Missing	0
Median		5,00
Mode		5
Percentiles	15	3,00
	25	4,00
	40	4,00
	46	4,46
	50	5,00
	75	5,00

Observação: Nos quantis, a fórmula de cálculo utilizada no SPSS não é igual à descrita neste livro, pelo que podem verificar-se ligeiras diferenças entre os resultados obtidos.

2.2.8.3 Dados quantitativos discretos

Considere a tabela de frequências (Tabela 2.12) apresentada na secção 2.1.5.3:

Tabela 2.12: *Tabela de frequências relativa ao número de livros lidos durante as férias de verão.*

N.º de livros (x_i')	N.º de alunos (n_i)	Prop. de alunos (f_i)	N.º acum. de alunos (N_i)	Prop. acum. de alunos (F_i)
0	2	0,04	2	0,04
1	12	0,24	14	0,28
2	27	0,54	41	0,82
3	8	0,16	49	0,98
4	1	0,02	50	1,00
Total	50	1,00		

Determine e interprete as medidas de localização de tendência central, os quantis, o 1º e 9º decis, os percentis 82 e 95, a amplitude interquartil e a amplitude total, a variância e o desvio padrão, o coeficiente de variação e dispersão, o 2º, 3º e 4º momentos centrais, o 2º momento, o 2º momento em relação a 4 livros e estude a assimetria e achatamento dos dados.

Resolução:

Média (= 1º Momento):

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{5} n_i x_i' = \frac{(2\times 0)+(12\times 1)+(27\times 2)+(8\times 3)+(1\times 4)}{50} = \frac{94}{50} = 1{,}88.$$

Em média os alunos leram mais do que um livro (1,88) livros nas férias de verão.

Moda:

$\hat{x} = 2$ (categoria com maior frequência).

O mais comum foi os alunos terem lido 2 livros.

Mediana (= 2º Quartil = 5º decil = 50º Percentil):

$$n = 50 \text{ (é par)} \rightarrow \tilde{x} = \frac{x_{25:50} + x_{26:50}}{2} = \frac{2+2}{2} = 2.$$

Metade dos alunos leu no máximo 2 livros.

1º Quartil (= 25º Percentil):

$$p = \frac{1}{4} = 0{,}25 \rightarrow np = 12{,}5 \text{ (não é inteiro)} \rightarrow Q_1 = x_{[12,5]+1:50} = x_{12+1:50} = x_{13:50} = 1$$

25% dos alunos leram no máximo 1 livro.

3º Quartil (= 75º Percentil):

$$p = \frac{3}{4} = 0{,}75 \rightarrow np = 37{,}5 \text{ (não é inteiro)} \rightarrow Q_3 = x_{[37,5]+1:50} = x_{37+1:50} = x_{38:50} = 2.$$

75% dos alunos leram no máximo 2 livros, i.e., 25% dos alunos leram no mínimo 2 livros.

1º Decil (= 10º Percentil):

$$p = \frac{10}{100} = 0{,}1 \ \rightarrow \ np = 5 \text{ (é inteiro)} \rightarrow P_{10} = \frac{x_{5:50} + x_{6:50}}{2} = \frac{1+1}{2} = 1.$$

10% dos alunos leram no máximo 1 livro.

9º Decil (= 90º Percentil):

$$p = \frac{90}{100} = 0{,}9 \ \rightarrow \ np = 45 \text{ (é inteiro)} \rightarrow D_4 = \frac{x_{45:50} + x_{46:50}}{2} = \frac{3+3}{2} = 3.$$

90% dos alunos leram no máximo 3 livros, i.e., 10% dos alunos leram pelo menos 3 livros.

82º Percentil:

$$p = \frac{82}{100} = 0{,}82 \ \rightarrow \ np = 41 \text{ (é inteiro)} \rightarrow P_{82} = \frac{x_{41:50} + x_{42:50}}{2} = \frac{2+3}{2} = 2{,}5.$$

82% dos alunos leram no máximo 2 livros e os restantes 18% leram no mínimo (pelo menos) 3 livros.

95º Percentil:

$$p = \frac{95}{100} = 0{,}95 \ \rightarrow \ np = 47{,}5 \text{ (não é inteiro)} \rightarrow P_{95} = x_{[47{,}5]+1:50} = x_{48:50} = 3.$$

95% dos alunos leram no máximo 3 livros, i.e., 4% dos alunos leram pelo menos 3 livros.

Amplitude interquartil:

$$AIQ = Q_3 - Q_1 = 2 - 1 = 1.$$

Das 50 observações, a dispersão das 25 observações centrais é de 1 livro, ou seja, excluindo os 25% de alunos que menos livros leram e os 25% de alunos que mais livros leram, entre os restantes existe uma diferença de 1 livro entre o que mais livros leu e o que menos livros leu durante as férias.

Amplitude:

$$a = x_{50:50} - x_{1:50} = 4 - 0 = 4.$$

A diferença entre o número máximo e o número mínimo de livros lidos pelos alunos nas férias é de 4 livros.

Variância amostral:

$$s^2 = \frac{1}{49}\sum_{i=1}^{5} n_i (x_i' - \overline{x})^2 = \frac{2(0-1{,}88)^2 + 12(1-1{,}88)^2 + \cdots + 1(4-1{,}88)^2}{49} = 0{,}6384$$

ou

$$s^2 = \frac{1}{49}\left(\sum_{i=1}^{5} n_i x_i'^2 - 50\overline{x}^2\right) = \frac{2\times 0^2 + 12\times 1^2 + \cdots + 1\times 4^2 - 50\times 1{,}88^2}{49} = 0{,}6384.$$

Desvio padrão amostral:

$$s = \sqrt{s^2} = \sqrt{0{,}6384} = 0{,}799.$$

O desvio *típico* em relação ao número médio de livros lidos é 0,8 livros.

Coeficiente de variação:

$$CV = \frac{s}{\overline{x}} \times 100 = \frac{0{,}799}{1{,}88} \times 100 = 42{,}5\% < 50\%.$$

A média é representativa dos dados ($CV < 50\%$).

2º Momento central:

$$m_2 = \frac{n-1}{n} s^2 = \frac{1}{50}\sum_{i=1}^{5} n_i (x_i' - \overline{x})^2 = \frac{31{,}28}{50} = 0{,}6256.$$

3º Momento central:

$$m_3 = \frac{1}{50}\sum_{i=1}^{5} n_i(x_i' - \overline{x})^3 = \frac{2(0-1{,}88)^3 + 12(1-1{,}88)^3 + \cdots + 1(4-1{,}88)^3}{50} = -0{,}0131.$$

4º Momento central:

$$m_4 = \frac{1}{50}\sum_{i=1}^{5} n_i(x_i' - \overline{x})^4 = \frac{2(0-1{,}88)^4 + 12(1-1{,}88)^4 + \ldots + 1(4-1{,}88)^4}{50} = 1{,}2995.$$

2º Momento:

$$m_2' = \frac{1}{50}\sum_{i=1}^{5} n_i x_i'^2 = \frac{2\times 0^2 + 12\times 1^2 + 27\times 2^2 + 8\times 3^2 + 1\times 4^2}{50} = 4{,}16$$

ou

$$m_2 = m_2' - m_1'^2 \Leftrightarrow m_2' = m^2 + m_1'^2 = 0{,}6256 + 1{,}88^2 = 4{,}16.$$

2º Momento em relação a 4:

$$m_{2,4}' = \frac{1}{50}\sum_{i=1}^{5} n_i(x_i' - 4)^2 = \frac{2(0-4)^2 + 12(1-4)^2 + \cdots + 1(4-4)^2}{50} = 5{,}12.$$

Grau de assimetria de Pearson:

$$g_P = \frac{\overline{x} - \hat{x}}{s} = \frac{1{,}88 - 2}{0{,}799} = -0{,}1502.$$

A distribuição é quase simétrica, apresentando uma ligeira assimetria negativa.

Grau de assimetria de Bowley:

$$g_B = \frac{(Q_3 - \tilde{x}) - (\tilde{x} - Q_1)}{Q_3 - Q_1} = \frac{(2-2) - (2-1)}{2-1} = -1.$$

Considerando apenas a parte central dos dados, a distribuição é assimétrica negativa.

Coeficiente de assimetria amostral:

$$g_a = \frac{n^2 m_3}{(n-1)(n-2)s^3} = \frac{50^2 \times (-0{,}0131)}{49\times 48\times 0{,}799^3} = -0{,}027.$$

Pode-se considerar que a distribuição é simétrica, ou seja, a proporção de alunos que leram poucos livros é idêntica à dos que leram muitos livros.

Coeficiente percentil de kurtosis:

$$k_P = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})} = \frac{2-1}{2(3-1)} = 0{,}25.$$

Pode-se considerar que a distribuição é mesocúrtica.

Coeficiente de kurtosis amostral:

$$k_a = \frac{n^2(n+1)m_4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)} = \frac{50^2\times 51\times 1{,}2995}{49\times 48\times 47\times 0{,}799^3} - \frac{3\times 49^2}{48\times 47} = 0{,}485.$$

A distribuição é aproximadamente mesocúrtica (valor próximo de zero), i. e., verifica-se uma percentagem acentuada de alunos que leram um número de livros perto do número médio, e esta percentagem vai diminuindo gradualmente à medida que o número de livros lidos se afasta do número médio.

(SPSS) Analyse → Descriptive Statistics → Frequencies…
(Variable: N_Livros; Statistics: ☑ Mean; ☑ Median; ☑ Mode; ☑ Quartiles; ☑ Percentile(s): 10, 90, 82, 95; ☑ Std. Deviation; ☑ Variance; ☑ Range; ☑ Minimum; ☑ Maximum; ☑ Skewness; ☑ Kurtosis)

Statistics

N.º de livros lidos

N	Valid	50
	Missing	0
Mean		1,88
Median		2,00
Mode		2
Std. Deviation		,799
Variance		,638
Skewness		-,027
Std. Error of Skewness		,337
Kurtosis		,485
Std. Error of Kurtosis		,662
Range		4
Minimum		0
Maximum		4
Percentiles	10	1,00
	25	1,00
	50	2,00
	75	2,00
	82	2,82
	90	3,00
	95	3,00

2.2.8.4 Dados quantitativos contínuos

Considere os dados da secção 2.1.5.4, sobre as áreas dos jardins dos projetos aprovados, cuja tabela de frequências é a que se segue (Tabela 2.13).

Tabela 2.13: *Tabela de frequências relativa à área dos jardins.*

Áreas (m^2)	Ponto médio	N.º de jardins	% de jardins	N.º acum. de jardins	% acum. de jardins
[300; 600[	450	50	50	50	50
[600; 900[	750	30	30	80	80
[900; 1200[	1050	9	9	89	89
[1200; 1500[	1350	5	5	94	94
[1500; 1800]	1650	6	6	100	100
Total		100	100		

Determine e interprete as seguintes medidas estatísticas: média, mediana, moda, amplitude interquartil, 1º e 7º decis, 36º e 90º percentis, variância, desvio padrão, coeficientes de dispersão e variação, 2º, 3º e 4º momentos centrais, 2º momento e 2º momento relativamente a 450 m^2. Estude a assimetria e achatamento da distribuição. Construa a curva de Lorenz, calcule o índice de Gini e interprete.

Resolução:

Média (= 1º momento):

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{5} n_i x_i' = \frac{(50 \times 450) + (30 \times 750) + \cdots + (6 \times 1650)}{100} = 711.$$

Em média os jardins têm 711 m^2 de área.

Moda: classe modal:

$$C_{Mo} = [300; 600[\text{ (classe com maior frequência).}$$

O mais comum é os jardins terem entre 300 e 600 m^2 de área.

$$\hat{x} = LI_{C_{Mo}} + a_{C_{Mo}} \frac{\Delta_1}{\Delta_1 + \Delta_2} 300 + 300 \frac{50}{50 + 20} = 514{,}3,$$

onde $\Delta_1 = n_{C_{Mo}} - n_{C_{Mo-1}} = 50 - 0 = 50$ e $\Delta_2 = n_{C_{Mo}} - n_{C_{Mo+1}} = 50 - 30 = 20$.
O mais usual é os jardins terem 514,3 m^2 de área. Note-se que esta medida, moda em variáveis contínuas, usualmente não tem utilidade prática relevante.

Mediana (= 2º Quartil = 5º Decil = 50º Percentil):
$np = 100 \times 0{,}5 = 50$ → classe mediana: $C_{Me} = [300; 600[$;

$$\tilde{x} = Q_2 = P_{50} = Q^*_{0{,}50} = LI_{C_{Me}} + a_{C_{Me}} \frac{0{,}5n - N_{C_{Me-1}}}{n_{C_{Me}}} = 300 + 300 \frac{50 - 0}{50} = 600.$$

Metade dos jardins têm uma área inferior ou igual a 600 m^2.

1º Quartil (= 25º Percentil):
$np = 100 \times 0{,}25 = 25$ → classe 1º quartil: $C_Q = [300; 600[$;

$$Q_1 = P_{25} = Q^*_{0{,}25} = LI_{C_Q} + a_{C_Q} \frac{np - N_{C_{Q-1}}}{n_{C_Q}} = 300 + 300 \frac{25 - 0}{50} = 450.$$

25% dos jardins têm uma área inferior ou igual a 450 m^2.

3º Quartil (= 75º Percentil):
$np = 100 \times 0{,}75 = 75$ → classe 3º quartil: $C_Q = [600; 900[$;

$$Q_3 = P_{75} = Q^*_{0{,}75} = LI_{C_Q} + a_{C_Q} \frac{np - N_{C_{Q-1}}}{n_{C_Q}} = 600 + 300 \frac{75 - 50}{30} = 850.$$

75% dos jardins têm uma área inferior ou igual a 850 m^2 e os restantes 25% têm no mínimo 850 m^2.

Amplitude interquartil:

$$AIQ = Q_3 - Q_1 = 850 - 450 = 400.$$

Das 100 observações, a dispersão das 50 observações centrais é de 400 m^2. Excluindo 25% dos jardins com menor área e 25% dos jardins com maior área, entre os restantes jardins a diferença entre o maior e o menor jardim é de 400 m^2.

1º Decil (= 10º Percentil):
$np = 100 \times 0{,}1 = 10$ → classe 1º decil: $C_Q = [300; 600[$;

$$D_1 = P_{10} = Q^*_{0{,}10} = LI_{C_Q} + a_{C_Q} \frac{np - N_{C_{Q-1}}}{n_{C_Q}} = 300 + 300 \frac{10 - 0}{50} = 360.$$

10% dos jardins têm uma área inferior ou igual a 360 m^2.

7º Decil (= 70º Percentil):
$np = 100 \times 0{,}7 = 70$ → classe 7º decil: $C_Q = [600; 900[$;

$$D_7 = P_{70} = Q^*_{0{,}70} = LI_{C_Q} + a_{C_Q} \frac{np - N_{C_{Q-1}}}{n_{C_Q}} = 600 + 300 \frac{70 - 50}{30} = 800.$$

70% dos jardins têm uma área inferior ou igual a 800 m^2.

36º Percentil:
$np = 100 \times 0{,}36 = 36$ → classe 36º percentil: $C_Q = [300; 600[$

$$P_{36} = Q^*_{0{,}36} = LI_{C_Q} + a_{C_Q} \frac{np - N_{C_{Q-1}}}{n_{C_Q}} = 300 + 300 \frac{36 - 0}{50} = 516.$$

36% dos jardins têm uma área inferior ou igual a 516 m^2.

90º Percentil (= 9º Decil):

$np = 100 \times 0{,}9 = 90 \rightarrow$ classe 90º percentil: $C_Q = [1200; 1500[$;

$$P_{90} = D_9 = Q^*_{0,90} = LI_{C_Q} + a_{C_Q}\frac{np - N_{C_{Q-1}}}{n_{C_Q}} = 1200 + 300\frac{90 - 89}{5} = 1260.$$

90% dos jardins têm uma área inferior ou igual a 1260 m^2, ou 10% dos jardins têm pelo menos 1260 m^2.

Variância amostral:

$$s^2 = \frac{\sum_{i=1}^{5} n_i x_i'^2 - 100\overline{x}^2}{99} = \frac{50 \times 450^2 + \ldots + 6 \times 1650^2 - 100 \times 711^2}{99} = 119372{,}727.$$

Desvio padrão amostral:

$$s = \sqrt{s^2} = \sqrt{119372{,}727} = 345{,}504.$$

O desvio *típico* em relação à área média dos jardins é 345,504 m^2.

Coeficiente de variação:

$$CV = \frac{s}{\overline{x}} \times 100 = \frac{345{,}504}{711} \times 100 = 48{,}6\% < 50\%.$$

A média é moderadamente representativa dos dados.

2º Momento central:

$$m_2 = \frac{n-1}{n}s^2 = \frac{99}{100} \times 119372{,}727 = 118179.$$

3º Momento central:

$$m_3 = \sum_{i=1}^{5} \frac{n_i(x_i' - \bar{x})^3}{100} = \frac{50(450 - 711)^3 + \cdots + 6(1650 - 711)^3}{100} = 57356262.$$

4º Momento central:

$$m_4 = \sum_{i=1}^{5} \frac{n_i(x_i' - \bar{x})^4}{100} = \frac{50(450 - 711)^4 + \cdots + 6(1650 - 711)^4}{100} = 58491761877.$$

2º Momento:

$$m_2' = \sum_{i=1}^{5} \frac{n_i x_i'^2}{100} = \frac{50 \times 450^2 + 30 \times 750^2 + \cdots + 6 \times 1650^2}{100} = 623700.$$

ou alternativamente

$$m_2 = m_2' - m_1'^2 \Leftrightarrow m_2' = m^2 + m_1'^2 = 118179 + 711^2 = 623700.$$

2º Momento em relação a 450:

$$m_{2,450}' = \sum_{i=1}^{5} \frac{n_i(x_i' - 450)^2}{100} = \frac{50(450 - 450)^2 + \cdots + 6(1650 - 450)^2}{100} = 186300.$$

Grau de assimetria de Pearson:

$$g_P = \frac{\overline{x} - \hat{x}}{s} = \frac{711 - 514{,}3}{345{,}504} = 0{,}569.$$

A distribuição apresenta uma ligeira assimetria positiva, ou seja, verifica-se um maior número de jardins com áreas pequenas do que com áreas elevadas.

Grau de assimetria de Bowley:

$$g_B = \frac{(Q_3 - \tilde{x}) - (\tilde{x} - Q_1)}{Q_3 - Q_1} = \frac{(850 - 600) - (600 - 450)}{850 - 450} = 0{,}25.$$

A distribuição apresenta uma ligeira assimetria positiva.

Coeficiente percentil de kurtosis:

$$k_P = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})} = \frac{850 - 450}{2(1260 - 360)} = 0{,}222 < 0{,}263.$$

A distribuição apresenta uma ligeira leptocurtose, ou seja, as áreas dos jardins tendem a estar um pouco mais próximas da área média do que seria normal.

De referir que o objetivo principal deste coeficiente é comparar o comportamento desta distribuição em relação a uma distribuição Normal. Neste exercício, apesar da distribuição não ser simétrica (logo não Normal), optou-se por calcular apenas o coeficiente percentil de kurtosis, visto este ser menos sensível à existência de valores atípicos e para ilustrar a sua fórmula de cálculo.

(SPSS) Analyse → Descriptive Statistics → Frequencies…
(Variable: Areas; Statistics: ☑ Values are group midpoints; ☑ Mean; ☑ Median; ☑ Quartiles; ☑ Percentile(s); 10, 70, 36, 90; ☑ Variance; ☑ Std. Deviation; ☑ Skewness; ☑ Kurtosis)

Statistics

Áreas

N	Valid	100
	Missing	0
Mean		711,0000
Median		637,5000[a]
Std. Deviation		345,50359
Skewness		1,433
Std. Error of Skewness		,241
Kurtosis		1,312
Std. Error of Kurtosis		,478
Range		1200,00
Percentiles	10	.[b,c]
	25	450,0000
	36	532,5000
	50	637,5000
	70	826,9231
	75	903,8462
	90	1285,7143

a. Calculated from grouped data.
b. The lower bound of the first interval or the upper bound of the last interval is not known. Some percentiles are undefined.
c. Percentiles are calculated from grouped data.

Coeficiente de assimetria amostral:

$$g_a = skewness = 1{,}433.$$

A distribuição é assimetria positiva, ou seja, metade dos jardins têm área superior à área média o que significa que se observaram mais jardins com áreas maiores do que menores.

Coeficiente de achatamento amostral:

$$k_a = Kurtosis = 1{,}312.$$

A distribuição é leptocúrtica, ou seja, há uma elevada concentração de jardins com área próxima da área média observada.

Curva de Lorenz:

Áreas (m2)	x_i'	N.º de jardins (n_i)	N.º ac. de jardins $(F_i = p_i)$	Área total ocupada pelos jardins da classe i $(n_i x_i')$	Área total ocupada pelos jardins com área menor ou igual à da classe i $\left(\sum_{j=1}^{i} n_j x_j'\right)$	Proporção da área total ocupada pelos jardins com área menor ou igual à da classe i $\left(q_i = \frac{\sum_{j=1}^{i} n_j x_j'}{\sum_{j=1}^{5} n_j x_j'}\right)$
[300; 600[	450	50	0,50	22500	22500	0,3165
[600; 900[	750	30	0,80	22500	45000	0,6329
[900; 1200[	1050	9	0,89	9450	54450	0,7658
[1200; 1500[	1350	5	0,94	6750	61200	0,8608
[1500; 1800]	1650	6	1,00	9900	71100	1,0000
Total		100	4,13	71100		3,576

Da análise dos resultados obtidos na última coluna da tabela anterior, verifica-se que:

- 50% dos jardins ocupam 31,65% da área total aprovada.
- 80% dos jardins ocupam 63,29% da área total aprovada.
- 89% dos jardins ocupam 76,58% da área total aprovada.
- 94% dos jardins ocupam 86,08% da área total aprovada.

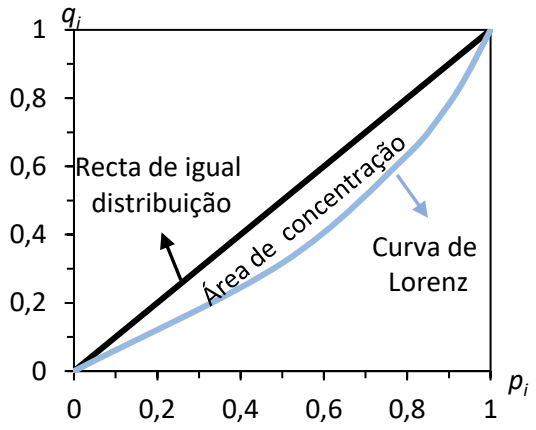


Figura 2.33: *Curva de Lorenz.*

Pela curva de Lorenz (Figura 2.33) pode-se concluir que a área aprovada para espaços verdes está distribuída quase de igual forma pelas diferentes áreas dos jardins. Os jardins de elevada dimensão, os que têm área superior a 900 m^2, ocupam 36,71% da área total e correspondem a 20% dos jardins aprovados, ou seja, ocupam um pouco mais da área aprovada do que os jardins de dimensão reduzida.

Índice de concentração de Gini:

$$IG = 1 - \frac{\sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = 1 - \frac{\sum_{i=1}^{4} q_i}{\sum_{i=1}^{4} p_i} = 1 - \frac{2{,}576}{3{,}13} = 0{,}177.$$

A área aprovada para espaços verdes está quase distribuída de igual forma pelas diferentes dimensões dos jardins.

2.2.8.5 Dados bivariados

1. Mediram-se e pesaram-se 10 alunos do curso de Psicologia. Na Tabela 2.14 apresentam-se as alturas, em metros (m.) e em polegadas (pol.), e os pesos, em quilogramas (kg.) e libras (lb.) (adaptado de Pestana e Velosa, 2002):

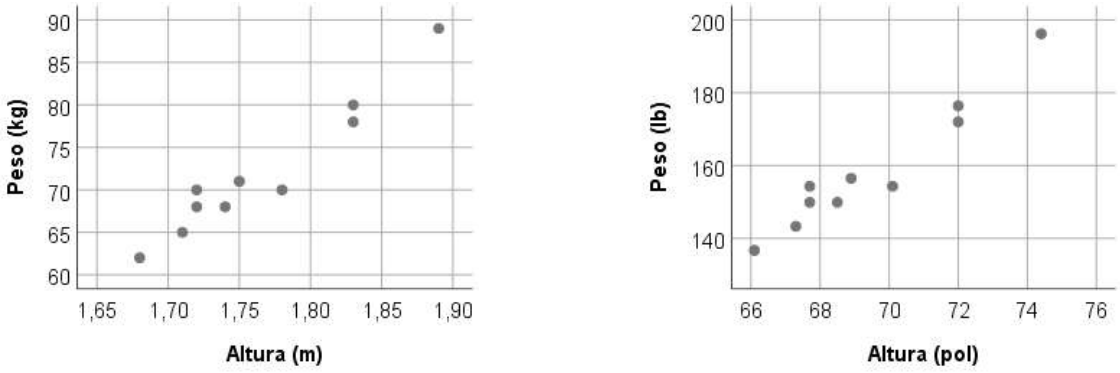
Tabela 2.14: *Pesos e alturas de 10 alunos do curso de Psicologia.*

Aluno	1	2	3	4	5	6	7	8	9	10
Altura (m.)	1,74	1,83	1,68	1,89	1,72	1,72	1,75	1,78	1,83	1,71
Peso (kg.)	68	80	62	89	68	70	71	70	78	65
Altura (pol.)	68,5	72	66,1	74,4	67,7	67,7	68,9	70,1	72	67,3
Peso (lb.)	149,9	176,4	136,7	196,2	149,9	154,3	156,5	154,3	172	143,3

a) Represente graficamente a altura *vs.* peso.
b) Calcule a altura e o peso médio, as respetivas variâncias.
c) Calcule o coeficiente de correlação de Pearson entre a altura e o peso.

Resolução:

a) (SPSS) Graphs → Legacy Dialogs → Scatter/Dot… → Simple Scatter
(Y Axis: Peso; X Axis: Altura)



a) Gráfico de dispersão - altura (m.) vs peso (kg.).

b) Gráfico de dispersão - altura (pol.) vs peso (lb.).

Figura 2.34: *Gráfico de dispersão - altura vs. peso.*

Como se pode verificar, a alteração da unidade de medida não altera o gráfico de dispersão (Figura 2.34).

b) Médias:

X – altura em metros: $\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1{,}74 + 1{,}83 + \ldots + 1{,}71}{10} = 1{,}765.$

Y – peso em quilogramas: $\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n} = \frac{68 + 80 + \ldots + 65}{10} = 72{,}1.$

X – altura em polegadas: $\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{68{,}5 + 72 + \ldots + 67{,}3}{10} = 69{,}47.$

Y – peso em libras: $\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n} = \frac{149{,}9 + 176{,}4 + \ldots + 143{,}3}{10} = 158{,}95.$

Portanto, a altura média dos alunos é 1,76 metros, ou seja, 69,47 polegadas. O peso médio dos alunos é 72,08 kg, ou seja, 158,95 libras.

Variâncias amostrais:

X – altura em metros:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}\right) = \frac{1{,}74^2 + 1{,}83^2 + \cdots + 1{,}71^2 - 10 \times 1{,}76^2}{9} = 0{,}004.$$

Y – peso em quilogramas:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}\right) = \frac{68^2 + 80^2 + \cdots + 65^2 - 10 \times 72{,}1^2}{9} = 64{,}322.$$

X – altura em polegadas:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}\right) = \frac{68{,}5^2 + \cdots + 67{,}3^2 - 10 \times 69{,}47^2}{9} = 6{,}789.$$

Y – peso em libras:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}\right) = \frac{149{,}9^2 + \cdots + 143{,}3^2 - 10 \times 158{,}95^2}{9} = 312{,}823.$$

(SPSS) Analyse → Descriptive Statistics → Descriptives…
(Variable: Altura, Peso; Options: ☑ Mean; ☑ Variance)

Descriptive Statistics

	N	Mean	Std. Deviation	Variance
Altura (m)	10	1,7650	,06621	,004
Peso (kg)	10	72,100	8,0201	64,322
Altura (pol)	10	69,470	2,6056	6,789
Peso (lb)	10	158,950	17,6868	312,823
Valid N (listwise)	10			

c) Covariância amostral:

1ª situação: (X – altura em metros e Y – peso em quilogramas)

$$s_{xy} = \sum_{i=1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{n-1} = \frac{(1{,}74 - 1{,}76)(68 - 72{,}1) + \cdots + (1{,}71 - 1{,}76)(65 - 72{,}1)}{9} = 0{,}514,$$

ou

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right) = \frac{1{,}74 \times 68 + 1{,}83 \times 80 + \cdots + 1{,}71 \times 65 - 10 \times 1{,}76 \times 72{,}1}{9} = 0{,}514.$$

2ª situação: (X – altura em polegadas e Y – peso em libras)

$$s_{xy} = \frac{(68{,}5 - 69{,}47)(149{,}9 - 158{,}95) + \cdots + (67{,}3 - 69{,}47)(143{,}3 - 158{,}95)}{9} = 44{,}583,$$

ou

$$s_{xy} = \frac{68{,}5 \times 149{,}9 + \ldots + 67{,}3 \times 143{,}3 - 10 \times 69{,}47 \times 158{,}95}{9} = 44{,}583.$$

Portanto, existe associação linear crescente entre a altura e o peso, ou seja, espera-se que os indivíduos que apresentam altura acima da média tenham um peso acima da média, e os indivíduos com altura abaixo da média tenham um peso abaixo da média.

Coeficiente de correlação de Pearson:

1ª situação: (X – altura em metros e Y – peso em quilogramas)

$$r = \frac{s_{xy}}{s_x s_y} = \frac{0{,}5136}{\sqrt{0{,}0044}\sqrt{64{,}3358}} = 0{,}968.$$

2ª situação: (X – altura em polegadas e Y – peso em libras)

$$r = \frac{s_{xy}}{s_x s_y} = \frac{44{,}5826}{\sqrt{6{,}789}\sqrt{312{,}8228}} = 0{,}968.$$

Portanto, existe relação linear positiva muito forte entre o peso e a altura dos alunos, pelo que a alturas elevadas estão associados pesos elevados e a alturas baixas estão associadas pesos baixos.

(SPSS) Analyse → Correlate… → Bivariate
(Variables: Altura, Peso; Correlation Coefficients: ☑ Pearson; Options → ☑ Cross-product deviations and covariances)

Correlations

		Altura (m)	Peso (kg)
Altura (m)	Pearson Correlation	1	,968**
	Sig. (2-tailed)		,000
	Sum of Squares and Cross-products	,039	4,625
	Covariance	,004	,514
	N	10	10
Peso (kg)	Pearson Correlation	,968**	1
	Sig. (2-tailed)	,000	
	Sum of Squares and Cross-products	4,625	578,900
	Covariance	,514	64,322
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

		Altura (pol)	Peso (lb)
Altura (pol)	Pearson Correlation	1	,967**
	Sig. (2-tailed)		,000
	Sum of Squares and Cross-products	61,101	401,245
	Covariance	6,789	44,583
	N	10	10
Peso (lb)	Pearson Correlation	,967**	1
	Sig. (2-tailed)	,000	
	Sum of Squares and Cross-products	401,245	2815,405
	Covariance	44,583	312,823
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

2. Foi pedido a uma empresa de estudos de mercado que avaliasse a que se devia o sucesso de alguns detergentes de lavar a roupa. Para tal, pediu-se a algumas das donas de casa a opinião sobre 10 detergentes (1 – muito má, 2 – má, 3 – média, 4 – boa, 5 – muito boa) e registou-se também a quantidade vendida, em toneladas por ano, desses detergentes. Na Tabela 2.15 apresentam-se os resultados obtidos. Calcule o coeficiente de correlação de Spearman e interprete.

Tabela 2.15: *Opinião sobre diversos detergentes e a quantidade vendida.*

Detergente	A	B	C	D	E	F	G	H	I	J
Opinião	3	5	5	5	2	2	2	1	4	3
Quant. vendida	9,5	9,7	9,8	9,9	9,2	9,3	9,1	9,0	9,6	9,4

Resolução:

Coeficiente de correlação de Spearman: (com base nos valores obtidos na Tabela 2.16)

$$r_S = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2 = 1 - \frac{6}{10(10^2-1)} \times 4{,}5 = 0{,}9727.$$

Tabela 2.16: *Determinação das ordenações*

	Detergente	A	B	C	D	E	F	G	H	I	J
Ordenações	Opinião	5,5	9	9	9	3	3	3	1	7	5,5
	Quant. vendida	6	8	9	10	3	4	2	1	7	5
	d_i	-0,5	1	0	-1	0	-1	1	0	0	0,5
	d_i^2	0,25	1	0	1	0	1	1	0	0	0,25

Portanto, existe associação positiva muito forte entre as ordenações da opinião sobre o detergente e a quantidade vendida, ou seja, os detergentes mais vendidos são os que têm melhor opinião junto das donas de casa.

(SPSS) Analyse → Correlate… → Bivariate

(Variables: Opiniao, Quantidade; Correlation Coefficients: ☑ Spearman)

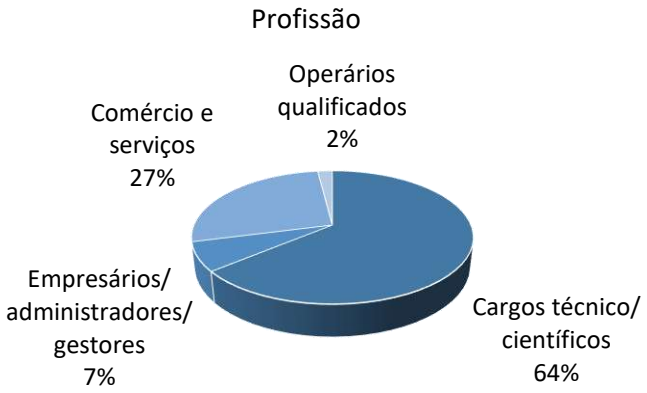
Correlations

			Opinião	Quantidade vendida
Spearman's rho	Opinião	Correlation Coefficient	1,000	,972**
		Sig. (2-tailed)	.	,000
		N	10	10
	Quantidade vendida	Correlation Coefficient	,972**	1,000
		Sig. (2-tailed)	,000	.
		N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

2.3 Exercícios propostos

1. Na revista Visão de 12 de setembro de 2002 é apresentado um artigo com o título "O Vício da Adrenalina". Neste artigo, é descrito o perfil do português radical em termos de profissão, sexo, residência e idade. Relativamente à profissão, consta a seguinte informação:



(Fonte: Contributos do INATEL para o Desporto-Aventura em Portugal – A Realidade dos Desportos Aventura, 2001.)

Com base numa amostra constituída por 1000 portugueses *radicais* inquiridos acerca da respetiva profissão, resolva as alíneas seguintes:

a) Defina e classifique a variável aleatória em estudo.
b) Construa a tabela de frequências.
c) Qual é a frequência absoluta dos portugueses radicais que ocupam cargos técnicos/científicos?
d) O que pode concluir deste estudo?

2. Realizou-se um estudo sobre a opinião dos alunos acerca da qualidade das refeições que lhes foram servidas numa determinada cantina. Os resultados obtidos foram:

Qualidade das refeições	N.º de alunos
Deficiente	1
Normal	9
Boa	27
Muito boa	13
	50

a) Defina e classifique a variável em estudo.
b) Diga o que representa o valor 50.
c) Qual é a percentagem de alunos que considera a qualidade das refeições "boa"?
d) Represente graficamente a distribuição.
e) Calcule as medidas de tendência central adequadas para esta distribuição.
f) Numa frase simples, procure explicar qual é a opinião destes alunos sobre a qualidade das refeições servidas na referida cantina.

3. Considere os resultados finais de Estatística de 20 estudantes de uma Universidade:

9	14	12	8	14	12	16	16	8	14
11	12	12	11	11	18	14	18	15	15

a) Os dados em estudo são de tipo qualitativo ou quantitativo?
b) Construa a tabela de frequências.
c) Represente graficamente a informação.
d) Calcule as medidas: média, mediana e moda.
e) Calcule a variância e o desvio padrão.
f) Calcule e interprete o coeficiente de variação.
g) Calcule a amplitude da amostra e a amplitude interquartil.
h) Calcule o valor do percentil 48 e do 8º decil.
i) Represente os dados numa caixa de bigodes.
j) Estude a distribuição quanto à assimetria e achatamento.
k) Elabore um pequeno texto que integre toda a informação anterior, sem esquecer de referir o que pode dizer sobre a existência de valores atípicos.

4. Foi feito um inquérito a novo grupo de utentes (40) de um posto de saúde para determinar quantas consultas requereram durante o primeiro ano de utilização desse posto de saúde. Obtiveram-se os seguintes resultados:

1	4	1	2	2	3	3	2	1	2	5	1	2	4	2	1	3	1	0	1
3	2	3	1	0	1	2	7	4	3	2	1	1	3	1	0	4	2	3	5

a) Construa a tabela de frequências.
b) Construa um gráfico para as frequências absolutas.
c) Represente graficamente as frequências relativas acumuladas.

5. Dada a seguinte distribuição do número de avarias nos elevadores em 200 edifícios públicos:

N.º de avarias	0	1	2	3	4	5
Frequências	53	68	44	17	16	2

a) Faça a representação gráfica das frequências relativas e das frequências relativas acumuladas.
b) Calcule a média, o desvio padrão e a moda.

6. 100 famílias do país A e 150 do país B foram classificadas segundo o número de filhos, tendo-se obtido os seguintes valores:

n.º de filhos	0	1	2	3	4	5	6	7	8
n.º de famílias do país A	11	13	20	25	14	10	4	2	1
n.º de famílias do país B	20	32	31	30	10	10	9	8	0

a) Defina e classifique a(s) variáveis(s) em estudo.
b) No país A, as famílias têm em média quantos filhos?
c) Complete as seguintes frases:
   i. "10% das famílias do país A têm no mínimo ... filhos."
   ii. "10% das famílias do país A têm no máximo ... filhos."
d) Determine o valor do coeficiente de variação para o país A. Interprete-o.
e) Represente graficamente as frequências observadas no país A.
f) Com base nas medidas de localização de tendência central pode-se considerar que as famílias do país A têm mais filhos do que as do país B? Justifique.
g) O que pode dizer quanto à assimetria da distribuição do número de filhos por família no país B. Interprete o resultado obtido.

7. Os dados que se seguem referem-se ao comprimento (em cm) de um grupo de bebés prematuros (idade gestacional inferior a 36 semanas) nascidos durante um mês numa maternidade.

29,9	40,2	37,8	19,7	30,0	29,7	19,4	39,2	24,7	20,4
19,1	34,7	33,5	18,3	19,4	27,3	38,2	16,2	36,8	33,1
41,4	13,6	32,2	24,3	19,1	37,4	23,8	33,3	31,6	20,1
17,2	13,3	37,7	12,6	39,6	24,6	18,6	18,0	33,7	38,2

a) Ordene os dados e calcule a média, mediana, desvio padrão, quartis e o quantil de ordem 2/3. Encontre um valor tal que 70% dos bebés observados tenham comprimentos superiores a esse valor.
b) Faça um agrupamento dos dados em classes, de forma conveniente, e represente-os graficamente através de um histograma.
c) Calcule a média, a mediana e variância para os dados agrupados. Compare estes valores aproximados com os correspondentes valores exatos obtidos em (a).

8. Com base numa amostra constituída por rendimentos anuais coletáveis, em unidades monetárias (u. m.), de 1000 famílias residentes numa determinada cidade, obteve-se a seguinte distribuição de frequências:

Rendimentos coletáveis (u. m.)	Frequências absolutas
[0; 1500[	6
[1500; 3000[	102
[3000; 4500[	134
[4500; 7500[	293
[7500; 15000[	364
[15000; 30000[	101

a) Represente o histograma e esboce o polígono de frequências relativas acumuladas.
b) Calcule a média, mediana, a moda e o terceiro quartil. Interprete os valores obtidos.
c) Calcule o desvio padrão e o coeficiente de variação.
d) Comente a proposição: "Mais de metade das famílias têm rendimento coletável superior à média".

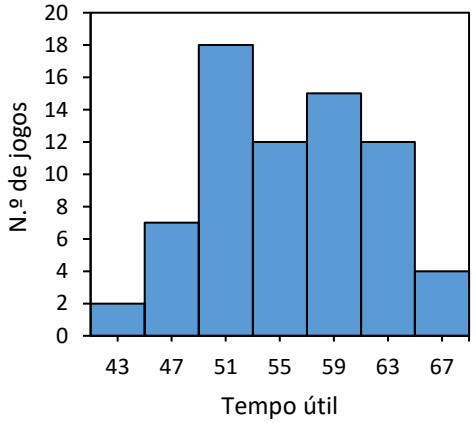
e) Complete as seguintes afirmações:
   i. "90% das famílias têm um rendimento coletável inferior a … u. m.."
   ii. "…% das famílias têm um rendimento coletável superior a 8000 u. m.."

9. A tabela seguinte representa os tempos que 100 utentes de um serviço de urgência demoraram a ser atendidos:

Tempo (em min)	5 a 9	10 a 14	15 a 19	20 a 24	25 a 29	30 a 34	35 a 39
Frequências	1	10	37	36	13	2	1

a) Calcule a média, a mediana e a moda. Compare os valores obtidos.
b) Diga, sem efetuar mais cálculos, qual o valor do 2º quartil e interprete o valor obtido.
c) Calcule o desvio padrão e a variância.
d) Determine os graus de assimetria de Bowley e de Pearson.

10.O futebol é um jogo bastante popular em diversos países. Seguidamente apresentam-se os resultados obtidos no SPSS, em 70 jogos de futebol, relativos ao tempo útil de jogo.



Statistics
Tempo útil

N	Valid	70
	Missing	0
Mean		55,551
Median		55,065
Std. Deviation		6,0215
Skewness		0,079
Std. Error of Skewness		0,287
Kurtosis		-0,729
Std. Error of Kurtosis		0,566
Minimum		42,13
Maximum		68,25
Percentiles	25	50,4225
	50	55,0650
	75	60,3875

Com base na informação disponibilizada:
a) Construa a tabela de frequências.
b) Desenhe a caixa de bigodes correspondente.
c) Complete a seguinte afirmação: "75% dos jogos tiveram tempo útil de jogo superior a ..."
d) Qual a tempo útil médio por jogo? E o mais habitual?
e) Considera que a média é representativa dos dados? Justifique.
f) "Registou-se uma maior frequência de jogos com tempo útil de jogo elevado do que com tempo útil de jogo baixo". Concorda com a afirmação? Justifique.
g) Estude a distribuição quanto ao achatamento e interprete o seu significado.

11.Pretende-se analisar o número de dias internamento médio das parturientes numa unidade hospitalar. Utilizando um software estatístico, foi feita uma tabela de frequências:

		Frequency	Percent
Valid	2	18	45,0
	3	4	10,0
	4	5	12,5
	5	7	17,5
	6	6	15,0
	Total	40	100,0

a) Complete a tabela de frequências.
b) Construa uma caixa de bigodes.
c) Calcule a média, moda, mediana e variância.
d) Com base no gráfico adequado, utilizando valores aproximados, complete as seguintes frases:
   i. 75% dos doentes estiveram internados até aproximadamente … dias.
   ii. 50 % dos internamentos foram de … a … dias.

12.Inserido num estudo da avaliação da obesidade em Portugal, mediu-se o perímetro torácico (em cm) de 210 indivíduos, tendo-se posteriormente agrupado os dados na seguinte tabela:

Classes	[75, 80[	[80, 85[	[85, 90[	[90, 95[	[95, 100[	[100, 105[	[105, 110[	[110, 115[
n_i	11	43	77	50	14	8	5	2

a) Calcule a média, moda e mediana.
b) Determine o desvio padrão.
c) Qual a proporção de indivíduos que têm um perímetro torácico superior ou igual a 100 cm?

13. O Centro de Geofísica de Évora disponibiliza na internet a informação recolhida por algumas estações meteorológicas que estão sob a sua responsabilidade. No caso concreto da estação meteorológica da Mitra, é possível aceder à informação diária sobre a temperatura do ar, humidade relativa, radiação global, velocidade do vento e precipitação.
No quadro seguinte apresentam-se algumas estatísticas descritivas, obtidas no SPSS, sobre a humidade relativa (em %) e temperatura máxima (em °C) observadas em outubro, de um determinado ano, na estação meteorológico da Mitra:

Statistics

		Humidade	Temperatura
N	Valid	31	31
	Missing	0	0
Mean		53,224	23,252
Median		55,273	21,997
Variance		169,278	32,464
Skewness		-0,867	0,948
Std. Error of Skewness		0,421	0,421
Kurtosis		0,172	-0,4108
Std. Error of Kurtosis		0,821	0,821
Minimum		22,187	15,895
Maximum		71,215	35,335
Percentiles	10	33,137	17,763
	25	45,246	18,804
	75	62,148	24,613
	90	69,092	32,748

Atendendo a que a covariância é –62,539, responda às seguintes questões, relativamente ao mês observado:
a) Defina e classifique a(s) característica(s) em estudo.
b) Qual foi o valor mínimo registado para a temperatura máxima?
c) Qual o valor médio de humidade relativa registado?
d) Complete as seguintes frases:
   i. “Em 50% dos dias a temperatura máxima foi inferior ou igual a ...°C.”
   ii. “Em 25% dos dias a temperatura máxima foi superior ou igual a ... °C.”
   iii. “Em 10% dos dias a humidade relativa foi no máximo ...°C.”
   iv. “A diferença entre o valor mais elevado e mais baixo registado para a temperatura máxima foi de ... °C.”

e) Calcule o coeficiente de dispersão para a temperatura máxima e humidade relativa. Interprete os valores obtidos.
f) "Registou-se uma maior frequência de dias com temperaturas máximas elevadas do que com temperaturas máximas baixas". Concorda com a afirmação? Justifique.
g) Calcule e interprete o coeficiente de correlação linear de Pearson entre a humidade relativa e a temperatura máxima.

14. No quadro seguinte, indicam-se os preços (X) dum bem alimentar (em unidades monetárias) praticados durante 12 meses consecutivos e as quantidades vendidas (Y).

x	110	90	80	76	74	71	70	65	63	60	55	50
y	55	70	90	100	90	105	80	110	125	115	130	131

a) Represente graficamente a informação disponibilizada.
b) Através da análise gráfica, parece-lhe existir relação linear entre as duas variáveis?
c) Calcule e interprete o valor do coeficiente de correlação linear de Pearson.

15. O departamento de qualidade de um determinado hospital, pretende fazer um estudo sobre o tempo que os utentes encaminhados para cirurgia pelos seus médicos de família demoram a ser efetivamente operados. Estes utentes são primeiro chamados a comparecer a uma consulta de referência no hospital (X dias após o encaminhamento) e só posteriormente são operados (Y dias após a consulta de referência). Escolheram-se aleatoriamente 15 utentes nestas condições, tendo sido reportados os seguintes tempos de espera:

x	69	76	51	34	62	13	40	7	64	41	64	26	40	44	48
y	28	64	7	26	38	18	40	20	44	32	31	32	36	25	73

Considera que estes tempos estão correlacionados? Calcule e interprete o valor do coeficiente de correlação mais indicado.

16.Pretende-se avaliar a relação entre o tempo de resolução de *puzzles* e a aptidão para o raciocínio matemático. Para tal, pediu-se a um grupo de 10 alunos do 1º ciclo para resolver um determinado puzzle. Na tabela seguinte apresentam-se os resultados obtidos para cada um dos alunos relativamente ao tempo de resolução do *puzzle* e a nota obtida em matemática.

Aluno	Tempo	Nota
1	10	Muito bom
2	15	Bom
3	40	Muito mau
4	30	Mau
5	20	Satisfaz
6	35	Mau
7	13	Bom
8	25	Satisfaz
9	9	Muito bom
10	30	Mau

Calcule o coeficiente de correlação de Spearman e interprete o valor obtido.

3 Introdução às probabilidades

A **estatística descritiva**, tal como o nome indica, permite descrever um conjunto de dados. Muitas vezes esse conjunto é apenas uma parte de um todo, pelo que existe alguma incerteza quando se pretende extrair conclusões para o todo, i. e., inferir ou extrapolar. Neste capítulo serão apresentados os conceitos base da **teoria das probabilidades**, área fundamental para o desenvolvimento da inferência estatística e também responsável pela modelação de fenómenos cujo comportamento depende duma componente aleatória.

3.1 Conceitos da teoria das probabilidades

Designa-se por **fenómenos aleatórios** os fenómenos influenciados pelo acaso, no sentido de que não é possível prever de forma determinística o seu futuro, a partir do passado.

Uma **experiência** diz-se **aleatória** se:

- O conjunto dos resultados possíveis é conhecido antecipadamente;
- O resultado da experiência não pode ser previsto com exatidão;
- É possível repetir a experiência em condições similares;
- Existe regularidade quando a experiência é repetida muitas vezes.

O **espaço de resultados**, $\boldsymbol{\Omega}$, é o conjunto (não vazio) de todos os resultados possíveis numa experiência aleatória. Diz-se:

- **Discreto** – quando o espaço de resultados é finito ou infinito numerável;
- **Contínuo** – quando o espaço de resultados é infinito não numerável.

Exemplos:

Experiência E1: Observação da face no lançamento de um dado.
Ω_1: $\{1, 2, 3, 4, 5, 6\}$.

Experiência E2: Observação do número de avarias registadas num elevador durante a sua vida útil.
Ω_2: $\{1, 2, 3, 4,..., n,...\}$.

Experiência E3: Observação do intervalo de tempo entre chegadas sucessivas de dois alunos ao bar da universidade.
Ω_3: $[0; +\infty[$.

Um **acontecimento** é todo o subconjunto do espaço de resultados. Os acontecimentos podem ser:

- **Elementares** – quando o subconjunto é singular (com um só elemento);
- **Compostos** – quando o subconjunto não é singular.

Portanto, o espaço de resultados, Ω, é o conjunto de todos os acontecimentos elementares.

Diz-se que um acontecimento A se realizou quando o resultado da experiência aleatória, w, é um elemento de A, isto é, se $w \in A$.

Exemplo: (Experiência E1: Observação da face no lançamento de um dado)

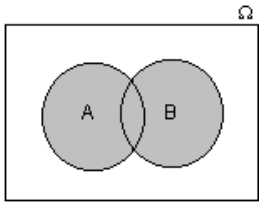
Acontecimentos:
- elementar: $A = \{\text{Saída da face } 2\} = \{2\}$;
- composto: $B = \{\text{Saída de face múltipla de } 3\} = \{3, 6\}$.

3.2 Álgebra dos acontecimentos

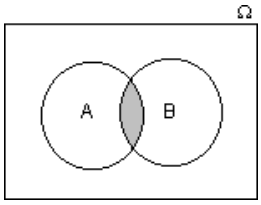
Sejam A e B dois acontecimentos, com $A \subseteq \Omega$ e $B \subseteq \Omega$.

3.2.1 Definições

Designa-se por **acontecimento reunião**, $\boldsymbol{A \cup B}$, o acontecimento que ocorrerá se, e somente se, pelo menos um dos acontecimentos, A ou B, ocorrer.

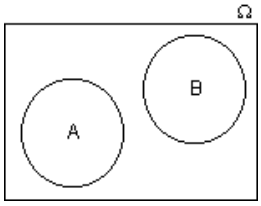


Designa-se por **acontecimento intersecção**, $\boldsymbol{A \cap B}$, o acontecimento que ocorrerá se, e somente se, ambos os acontecimentos A e B ocorrerem.

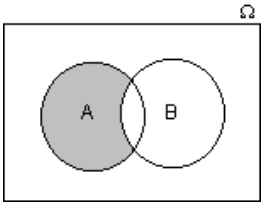


Designa-se por **acontecimento impossível**, $\emptyset$, o acontecimento que nunca ocorre qualquer que seja o resultado da experiência aleatória.

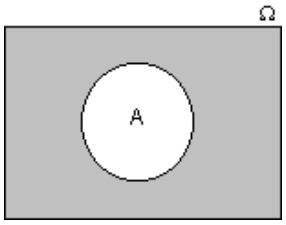
Diz-se que os **acontecimentos $\boldsymbol{A}$ e $\boldsymbol{B}$ são mutuamente exclusivos**, **disjuntos** ou **incompatíveis**, quando não possuem elementos comuns ($A \cap B = \emptyset$), ou seja, não ocorrem em simultâneo.



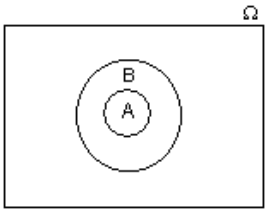
Designa-se por **acontecimento diferença**, $\boldsymbol{A - B}$ ou $\boldsymbol{A \backslash B}$, o acontecimento que ocorre quando ocorre A mas não ocorre B.



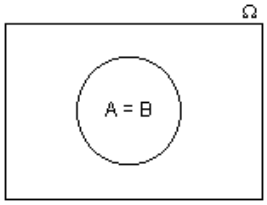
Designa-se por **acontecimento complementar**, $\overline{\boldsymbol{A}}$ ou $\boldsymbol{\Omega - A}$, o acontecimento que ocorre se, e somente se, A não ocorre e é constituído por todos os resultados do espaço Ω não favoráveis à ocorrência de A.



Designa-se por **implicação de acontecimentos**, $\boldsymbol{A \subset B}$, quando a realização do acontecimento A implica a realização do acontecimento B , i. e., se todo o elemento de A é elemento de B.



Designam-se por **acontecimentos idênticos**, $\boldsymbol{A = B}$, quando a realização de um acontecimento implica a realização do outro, i. e., $A \subseteq B$ e $B \subseteq A$.



3.2.2 Terminologia

Conjuntos	Acontecimentos
Ω	Acontecimento certo
$\emptyset$	Acontecimento impossível
$w \in \Omega$	Acontecimento elementar

3.2.3 Exemplo

Experiência E1: Observação da face no lançamento de um dado.

Espaço de resultados:

Ω_1: {1, 2, 3, 4, 5, 6}.

Acontecimentos:

$A = \{\text{Saída da face par}\} = \{2, 4, 6\}$

$B = \{\text{Saída de face múltipla de 3}\} = \{3, 6\}$

$C = \{\text{Saída da face 2}\} = \{2\}$

$D = \{\text{Saída de face múltipla de 2}\} = \{2, 4, 6\}$

Acontecimento reunião:

$A \cup B = \{$Saída da face par ou múltipla de 3$\} = \{2, 3, 4, 6\}$

Acontecimento intersecção:

$A \cap B = \{$Saída da face par e múltipla de 3$\} = \{6\}$

Acontecimento impossível:

$E = \{$Saída de face 10 no lançamento de um dado$\} = \emptyset$.

Acontecimentos mutuamente exclusivos:

B e C pois $B \cap C = \{$Saída de face múltipla de 3 e com face igual a 2$\} = \emptyset$.

Acontecimento diferença:

$A - B = \{$Saída da face par mas não múltipla de 3$\}$ = $\{2, 4\}$.

Acontecimento complementar:

$\overline{A} = \Omega \backslash A = \Omega - \mathrm{A} = \{$Saída de face não par$\}$ = $\{$Saída da face ímpar$\}$.

Implicação de acontecimentos:

$C \subset A = \{2\} \subset \{2, 4, 6\}$, i. e., a saída da face 2 implica a saída da face par.

Acontecimentos idênticos:

$A = D = \{2, 4, 6\}$.

3.2.4 Propriedades das operações

Propriedades	União	Intersecção
Comutativa	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Associativa	$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
Distributiva	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
Idempotência	$A \cup A = A$	$A \cap A = A$
Lei do Complementar	$A \cup \overline{A} = \Omega$	$A \cap \overline{A} = \emptyset$
Leis de De Morgan	$\overline{A \cup B} = \overline{A} \cap \overline{B}$	$\overline{A \cap B} = \overline{A} \cup \overline{B}$
Elemento Neutro	$A \cup \emptyset = \mathrm{A}$	$A \cap \Omega = A$
Elemento Absorvente	$A \cup \Omega = \Omega$	$A \cap \emptyset = \emptyset$

3.3 Definição de Probabilidade

O conceito de probabilidade pode ser definido de diferentes maneiras. De seguida apresentam-se apenas as definições mais usuais:

- Clássica ou *a priori* (Ω finito);
- Frequencista ou *a posteriori*;
- Subjetiva.

3.3.1 Definição clássica

"Probabilidade de um acontecimento é o quociente entre o número de casos favoráveis à ocorrência do acontecimento, $n(A)$, e o número de casos possíveis, n, todos os casos supostos igualmente prováveis."
Laplace, 1812.

$$P(A) = \frac{n(A)}{n}.$$

Exemplo: Num saco estão 40 bolas das quais 10 são brancas. Extrai-se, aleatoriamente, uma bola. Qual a probabilidade de a bola ser branca?

(Resposta: 10/40.)

Desvantagens:

- Nesta definição está implícita a hipótese de que todos os casos possíveis são igualmente prováveis, o que nem sempre se verifica;
- Está limitada à situação em que o espaço de resultados Ω é finito.
- As probabilidades são determinadas a priori, sem a realização da experiência.
- A determinação de $n(A)$ e de n é, por vezes, de elevada complexidade.

3.3.2 Definição frequencista

O **conceito frequencista de probabilidade** é utilizado em experiências aleatórias e independentes que, possuindo resultados que não são equiprováveis, são, no entanto, suscetíveis de repetição sob as mesmas condições. Seja $n(A)$ o número de vezes que o acontecimento A ocorre em n repetições de uma dada experiência, então

$$P(A) = \lim_{n\to\infty} \frac{n(A)}{n}.$$

A frequência relativa tende a estabilizar-se em torno de um valor que os frequencistas tomam como o valor aproximado da probabilidade $P(A)$.
Deste modo, os problemas associados com definições *a priori* são eliminados.

Exemplo: Na experiência que consiste no lançamento de uma moeda honesta, considere o acontecimento A = saída de coroa. Logo, $P(A) = 0{,}5$.

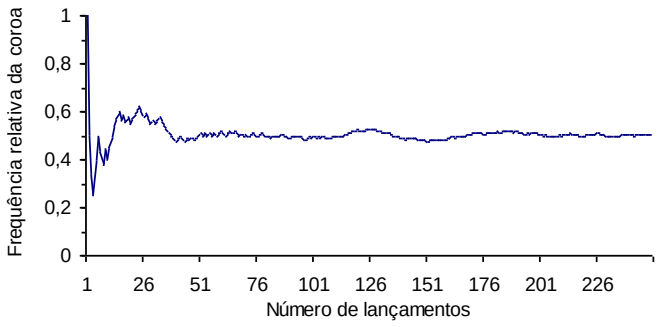


Figura 3.1: *Frequência de saída de coroas no lançamento de uma moeda honesta.*

Como se pode constatar pela Figura 3.1 quantas mais forem as repetições de lançamentos da moeda mais a proporção de vezes que saiu coroa se aproxima de 0,5.

Desvantagens:

- A experiência tem de ser repetida um elevado número de vezes nas mesmas condições, ou em condições semelhantes;
- Não há garantia que exista qualquer limite;
- A teoria é baseada na observação.

3.3.3 Definição subjetiva

Existem experiências aleatórias que não são suscetíveis de repetição sob condições idênticas e cujos resultados não são igualmente prováveis. Nestes casos é atribuída uma **probabilidade subjetiva** aos acontecimentos da experiência aleatória. As probabilidades são interpretadas como expressões do *grau de credibilidade* que cada indivíduo atribui à ocorrência dos acontecimentos.

Exemplo: Qual a probabilidade do atual governo se manter inalterado nos próximos 6 meses?

3.4 Axiomas da teoria das probabilidades

As probabilidades são definidas com base num conjunto de regras, ou axiomas, que devem satisfazer (**Axiomática simplificada de Kolmogorov**):

1º Axioma: Para qualquer acontecimento $A \subseteq \Omega$,

$$P(A) \geq 0.$$

2º Axioma: A probabilidade associada ao acontecimento certo Ω é

$$P(\Omega) = 1.$$

3º Axioma: Se dois acontecimentos A e B, definidos em Ω, forem mutuamente exclusivos, $A \cap B = \emptyset$, então

$$P(A \cup B) = P(A) + P(B).$$

Generalizando, se os acontecimentos $A_1, A_2, \dots, A_K$, definidos em Ω, forem mutuamente exclusivos, $A_i \cap A_j = \emptyset \ (\forall_{i \neq j:\, i,j=1,\dots,K})$, então

$$P(A_1 \cup A_2 \cup \dots \cup A_K) = P(A_1) + P(A_2) + \cdots + P(A_k) = \sum_{i=1}^{K} P(A_i).$$

3.5 Algumas propriedades matemáticas das probabilidades

A partir da Axiomática de Kolmogorov é possível deduzir um conjunto de propriedades matemáticas das probabilidades, que a seguir se apresentam, para os acontecimentos A e B definidos em Ω.

A probabilidade do acontecimento impossível é $P(\emptyset) = 0$.

Para qualquer acontecimento A, a probabilidade desse acontecimento satisfaz a relação,

$$0 \leq P(A) \leq 1.$$

Dado um acontecimento A, com probabilidade $P(A)$, a probabilidade do acontecimento contrário de A é

$$P(\overline{A}) = 1 - P(A).$$

Dados dois acontecimentos quaisquer A e B, a probabilidade do acontecimento diferença $B - A$ é dada por:

$$P(B - A) = P(B) - P(A \cap B).$$

Dados dois acontecimentos quaisquer A e B, a probabilidade da união é dada por:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Dados dois acontecimentos quaisquer A e B, a probabilidade da união verifica:

$$P(A \cup B) \leq P(A) + P(B).$$

Dados dois acontecimentos quaisquer A e B, se $A \subseteq B$ então

$$P(A) \leq P(B).$$

Sejam $A_1, A_2, \dots, A_K$ acontecimentos quaisquer definidos em Ω. Então

$$P\left(\bigcup_{i=1}^{K} A_i\right) = \sum_{i=1}^{K} P(A_i) - \sum_{i=1}^{K}\sum_{j=i+1}^{K} P(A_i \cap A_j) + \sum_{i=1}^{K}\sum_{j=i+1}^{K}\sum_{l=j+1}^{K} P(A_i \cap A_j \cap A_l) + \cdots + (-1)^{K+1} P\left(\bigcap_{i=1}^{K} A_i\right).$$

3.6 Probabilidade condicionada e independência

Sejam A e B dois acontecimentos, com $A \subseteq \Omega$ e $B \subseteq \Omega$.

3.6.1 Probabilidade condicionada

Definição: Dados os acontecimentos A e B a probabilidade de A ocorrer sabendo que B se realizou, ou a **probabilidade de A condicionada por B**, é definida por:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{se } P(B) > 0.$$

Invertendo a expressão acima, obtém-se a importante relação:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A),$$

se $P(A) > 0$ e $P(B) > 0$

3.6.2 Independência

Definição: Dois acontecimentos A e B são **independentes** se e só se:

$$P(A \cap B) = P(A)P(B),$$

com $P(A) \geq 0$ e $P(B) \geq 0$.

Se A e B são independentes então:

- $P(A|B) = P(A)$ se $P(B) > 0$;
- $P(B|A) = P(B)$ se $P(A) > 0$;

ou seja, o conhecimento da realização de B em nada afecta a probabilidade de se realizar A e vice-versa.

Observação: A independência entre os acontecimentos A e B implica a independência entre os acontecimentos A e $\overline{B}$, $\overline{A}$ e B, $\overline{A}$ e $\overline{B}$.

Por exemplo: $P\left(A \cap \overline{B}\right) = P(A - B) = P(A) - P(A \cap B) = P(A) - P(A)P(B)$
$= P(A) - \left(1 - P(B)\right) = P(A)P\left(\overline{B}\right).$

3.7 Teorema da probabilidade total e teorema de Bayes

3.7.1 Definição de partição

Os acontecimentos $A_1, A_2, \ldots, A_K$ dizem-se uma partição do espaço de resultados Ω se:

1. Forem mutuamente exclusivos, i. e.:

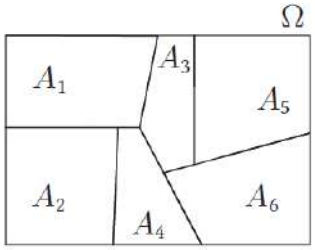
$$A_i \cap A_j = \emptyset \text{ qualquer que seja } i \neq j \text{ com } i, j = 1, 2, \ldots, K.$$

2. A união de todos os acontecimentos é o espaço de resultados, i. e.:

$$A_1 \cup A_2 \cup \ldots \cup A_K = \bigcup_{i=1}^{K} A_i = \Omega.$$

3. Todos os acontecimentos têm probabilidade não nula, i. e.:

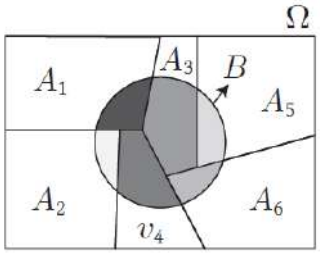
$$P(A_i) > 0, \text{ qualquer que seja } i = 1, 2, \ldots, K.$$



3.7.2 Teorema da probabilidade total

Teorema da probabilidade total: Se B é um acontecimento de Ω e $A_1, A_2, \ldots, A_K$ u ma partição de Ω então

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_K \cap B)$$
$$= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_K)P(A_K)$$
$$= \sum_{i=1}^{K} P(B|A_i)P(A_i).$$



3.7.3 Teorema de Bayes

Teorema de Bayes: Se B é um acontecimento de Ω, tal que $P(B) > 0$, e $A_1, A_2, \dots, A_K$ uma partição de Ω então

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + \cdots + P(B|A_K)P(A_K)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{K} P(B|A_i)P(A_i)}.$$

3.8 Exercícios resolvidos

1. Suponha que há três revistas, A, B e C, com as seguintes percentagens de leitura:
A – 9,8%; B – 22,9%; C – 12,1%; A e B – 5,1%; A e C – 3,7%; B e C – 6,0%; A, B e C – 2,4 %.
Calcule a probabilidade de uma pessoa escolhida ao acaso ser leitor:

a) de pelo menos uma das revistas.
b) da revista A e B mas não da revista C.
c) da revista A mas não das revistas B e C.

Resolução:

Acontecimentos:

$A \rightarrow$ A pessoa escolhida, ao acaso, é leitora da revista A;
$B \rightarrow$ A pessoa escolhida, ao acaso, é leitora da revista B;
$C \rightarrow$ A pessoa escolhida, ao acaso, é leitora da revista C.

a) $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
$= 0{,}098 + 0{,}229 + 0{,}121 - 0{,}051 - 0{,}037 - 0{,}06 + 0{,}024 = 0{,}324.$

b) $P\left(A \cap B \cap \overline{C}\right) = P\left((A \cap B) - C\right) = P(A \cap B) - P(A \cap B \cap C) = 0{,}051 - 0{,}024 = 0{,}027.$

c) $P\left(A \cap \overline{B} \cap \overline{C}\right) = P(A) - P(A \cap B) - P(A \cap C) + P(A \cap B \cap C) = 0{,}098 - 0{,}051 - 0{,}037 + 0{,}024 = 0{,}034.$

2. Numa determinada empresa 34% dos trabalhadores são do sexo masculino. Destes, 60% têm idade superior a 32 anos.

a) Selecionado ao acaso trabalhador do sexo masculino, qual a probabilidade de este ter no máximo 32 anos?
b) Sabendo que 38% dos trabalhadores são do sexo feminino e têm mais de 32 anos, calcule a probabilidade de um trabalhador selecionado ao acaso ter no máximo 32 anos.
c) Sabendo que o número total de trabalhadores na empresa é 250, complete o seguinte quadro:

Idade	Masculino	Feminino	Total
≤ 32			104
> 32			
Total			

Resolução:

Acontecimentos:

$M \rightarrow$ O trabalhador é do sexo masculino;
$F \rightarrow$ O trabalhador é do sexo feminino;
$T32 \rightarrow$ O trabalhador tem mais de 32 anos.

Sabe-se que:

- $P(M)$ = 0,34 $\Rightarrow P(F)$ = 0,66;
- $P(T32|M)$ = 0,60.

a) $P\left(\overline{T32}\middle|M\right) = 1 - P(T32|M) = 1 - 0{,}60 = 0{,}40.$

b) Sabe-se que $P(F \cap T32)$ = 0,38 e $P(M \cap T32) = P(T32|M)P(M) = 0{,}60 \times 0{,}34 = 0{,}204$.
Logo, $P(T32) = P(F \cap T32) + P(M \cap T32) = 0{,}38 + 0{,}204 = 0{,}584$.
Desta forma, $P\left(\overline{T32}\right) = 1 - P(T32) = 1 - 0{,}584 = 0{,}416$,

c) n = 250.
Para se poder preencher o quadro ainda é preciso calcular:
$P\left(\overline{T32} \cap M\right) = P(M) - P(M \cap T32)$ = 0,34 – 0,204 = 0,136;
$P\left(\overline{T32} \cap F\right) = P(F) - P(F \cap T32)$ = 0,66 – 0,38 = 0,28;
pois o resto já é conhecido das alíneas anteriores:

Idade	Masculino	Feminino	Total
≤ 32	0,136	0,28	0,416
> 32	0,204	0,38	0,584
Total	0,34	0,66	1

Multiplicando todas as células da tabela por 250 obtemos:

Idade	Masculino	Feminino	Total
≤ 32	34	70	104
> 32	51	95	146
Total	85	165	250

3. Numa fábrica, um certo tipo de chocolates é embalado em caixas e por uma das 3 linhas de produção diferentes: M_1, M_2 e M_3. Os registos mostram que uma pequena percentagem das caixas não é embalada em condições próprias para venda: 0,5% provêm de M_1, 0,8% de M_2 e 1% de M_3. Sabe-se que o volume diário de caixas embaladas por cada uma das linhas de produção é de 500, 100 e 2000 unidades, respetivamente.
Qual a probabilidade de uma caixa, escolhida ao acaso, não estar em condições para venda?
Sabendo que uma caixa não está em condições para venda, qual a probabilidade de ser proveniente da linha de produção M_2?

Resolução:

Acontecimentos:

$M_1 \rightarrow$ A caixa é embalada pela linha de produção M_1;
$M_2 \rightarrow$ A caixa é embalada pela linha de produção M_2;
$M_3 \rightarrow$ A caixa é embalada pela linha de produção M_3;
$D \rightarrow$ A caixa não está em condições para venda.

M_1, M_2 e M_3 são partição pois:

- As caixas são embaladas apenas nas linhas de produção M_1, M_2 ou M_3;
- Uma caixa não pode ser embalada em duas linhas de produção em simultâneo;
- Todas as linhas de produção embalam caixas.

Portanto, pode-se aplicar o Teorema da Probabilidade Total e o Teorema de Bayes para calcular as probabilidades pretendidas.

Na Tabela 3.1 apresenta-se um resumo das probabilidades conhecidas e o cálculo das restantes probabilidades necessárias para responder às alíneas do exercício.

Tabela 3.1: *Probabilidades individuais e condicionadas.*

Máquina	Produção	$P(M_i)$	$P(D\|M_i)$	$P(D\|M_i)P(M_i)$	$P(M_i\|D)$
M_1	500	500/2600 = 0,1923	0,005	0,001	0,1073
M_2	100	100/2600 = 0,0385	0,008	0,0003	0,0343
M_3	2000	2000/2600 = 0,7692	0,01	0,0077	0,8584
Total	2600	1		$P(D) = 0{,}009$	1

Na Figura 3.2 é feita a representação destas probabilidades num diagrama em árvore.

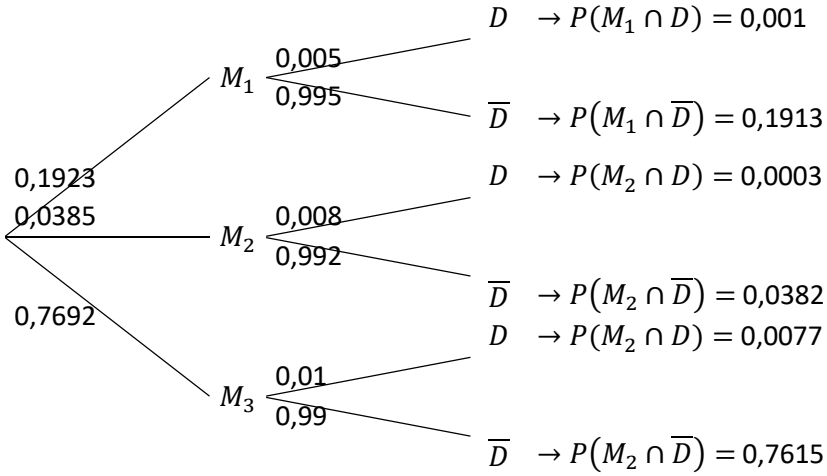


Figura 3.2: *Diagrama em árvore*

a) $P(D) = P(D|M_1)P(M_1) + P(D|M_2)P(M_2) + P(D|M_3)P(M_3) = 0{,}009.$

b) $P(M_2|D) = \dfrac{P(D|M_2)P(M_2)}{P(D)} = 0{,}0343.$

3.9 Análise Combinatória

A **Análise Combinatória** surge como um método auxiliar na contagem do número de casos favoráveis e do número de casos possíveis.

3.9.1 Arranjos e permutações

Os **arranjos simples**, ou **arranjos de n, k a k**, $^{n}A_{k}$, designam o número de amostras, sequências de k elementos, sem repetição de elementos, tiradas de um conjunto com n elementos:

$$^{n}A_{k} = \frac{n!}{(n-k)!}.$$

As **permutações simples**, ou **permutações de n**, P_n, designam o número de amostras, sequências de n elementos, sem repetição de elementos, tiradas de um conjunto com n elementos (caso particular de $^{n}A_{k}$ com $k = n$):

$$P_n = n!.$$

Os **arranjos completos**, ou **com repetição**, $^{n}A'_{k}$, designam o número de amostras, sequências de k elementos, tiradas de um conjunto com n elementos, admitindo que os elementos se podem repetir numa mesma sequência:

$$^{n}A'_{k} = n^{k}.$$

3.9.2 Combinações

As **combinações**, ou **combinações de n, k a k**, $^{n}C_{k}$, designam o número de subconjuntos com k elementos, que é possível formar a partir de um conjunto com n elementos:

$$^{n}C_{k} = \frac{n!}{k!\,(n-k)!}.$$

3.9.3 Exemplos

1. Com as letras A, B, C e D, quantas sequências de duas letras se podem formar:
   a) Podendo haver repetição de letras.
   b) Não podendo haver repetição de elementos.

Resolução:

a) Represente-se a situação descrita através de um esquema em árvore (Figura 3.3).

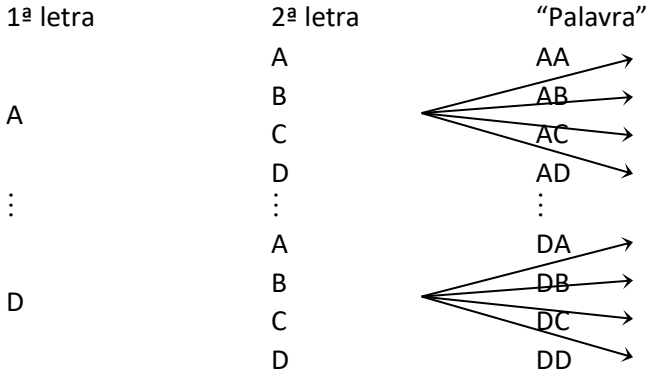


Figura 3.3: *Diagrama em árvore.*

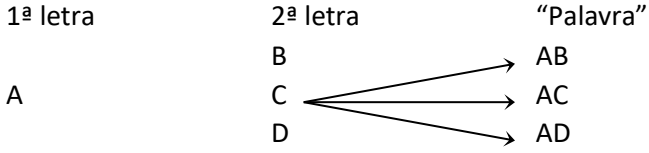
Para cada uma das 4 possibilidades para a 1ª letra tem-se 4 possibilidades para a 2ª letra, uma vez que se consideram diferentes as sequências AB e BA e aceitam-se sequências do tipo AA, BB, ...
Portanto, o número total de possibilidades é:

1ª letra 2ª letra

4 × 4 = 42 = $^{4}A'_{2}$ = 16 sequências distintas.

O número de sequências de 2 letras formadas com as 4 letras é 16.

b) Neste caso, tem-se 4 possibilidades para a 1ª letra, mas apenas 3 hipóteses para a 2ª letra, uma vez que fixada a 1ª letra, ela não poderá ocupar qualquer outra posição, ou seja, não se aceitam sequências do tipo AA, BB, ...
Na Figura 3.4 representa-se a situação descrita através de um esquema em árvore.



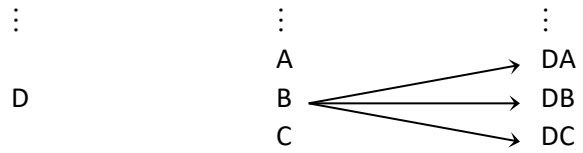


Figura 3.4: *Diagrama em árvore.*

Portanto, o número total de possibilidades é:

1ª letra 2ª letra

4 × 3 = $^{4}A_{2}$ = 12 sequências distintas.

O número de sequências que se podem formar com 2 letras distintas escolhidas entre as 4 letras A, B, C e D é 12.

2. No jogo Totoloto é necessário escolher 6 números de entre os 49 primeiros números naturais. Quantas apostas simples seriam necessárias fazer para haver a certeza de ganhar o primeiro prémio (acertar os 6 números)? E qual a probabilidade de ganhar esse prémio?

Resolução:

Trata-se de uma escolha sem reposição, uma vez que depois de escolhido um número ele não voltará a ser escolhido. Então as hipóteses de escolha são:

1º n.º 2º n.º 3ª n.º 4º n.º 5º n.º 6º n.º

49 × 48 × 47 × 46 × 45 × 44 = $^{49}A_{6}$ = 10.068.347.520 sequências possíveis.

De notar que depois de escolhido uma vez, um número não poderá voltar a ser escolhido. Por outro lado, as sequências 1, 2, 3, 4, 5, 6 ou 2, 1, 3, 4, 5, 6 ou 2, 3, 1, 4, 5, 6 ou ... correspondem à mesma escolha, ou seja, a ordem da escolha dos números não tem influência na contagem.

O número de sequências repetidas para cada conjunto de 6 números é

$$P_6 = 6! = 720.$$

Então, é preciso corrigir o resultado obtido com $^{49}A_{6}$ dividindo-o pelo número de repetições com cada conjunto de 6 números, obtendo-se:

$$\frac{^{49}A_{6}}{P_6} = {}^{49}C_{6} = 13.983.816 \text{ sequências distintas.}$$

Portanto, para haver a certeza de ganhar, é necessário fazer 13.983.816 apostas simples.

A probabilidade de sair o 1º prémio, jogando só uma chave simples, é:

$$\frac{1}{^{49}C_{6}} = \frac{1}{13.983.816} = 0{,}0000000715.$$

3. No jogo Euromilhões é necessário escolher 5 números de entre os 50 primeiros números naturais e 2 estrelas em 12 possíveis.

Quantas apostas simples seriam necessárias fazer para haver a certeza de ganhar o primeiro prémio (acertar os 5 números mais as 2 estrelas)? E qual a probabilidade de ganhar esse prémio?

Resolução:

O número de sequências distintas possíveis é dado por:

$$^{50}C_{5}\,{}^{12}C_{2} = 139.838.160.$$

Portanto, para haver a certeza de ganhar, é necessário fazer 139.838.160 apostas simples.

A probabilidade de sair o 1º prémio é:

$$\frac{1}{^{50}C_{5}\,{}^{12}C_{2}} = \frac{1}{139.838.160} = 0{,}00000000715.$$

4. Todos os professores que estão na sala de professores cumprimentaram-se apertando a mão. Sabendo que foram dados 45 apertos de mão, quantos professores estão na sala?

Resolução:

A ordem não influi na contagem e não pode haver repetição de elementos ⇒ Combinações.
Portanto, $^{n}C_{2} = 45 \Leftrightarrow \frac{n(n-1)}{2} = 45 \Longrightarrow n = 10$, ou seja, estão 10 professores a sala.

5. De quantos modos diferentes é possível dispor numa fila, para uma fotografia, 3 homens e 2 mulheres se:
   a) Se um dos homens, o mais alto, por exemplo, ficar no meio, e todos os restantes indistintamente em qualquer lugar?
   b) Se ficarem alternadamente homens e mulheres, nunca dois homens seguidos ou duas mulheres seguidas?
   c) Não considerando o sexo, de quantas formas distintas pode sentar estas pessoas em 5 lugares disponíveis?

Resolução:

a) As pessoas são diferentes umas das outras (a ordem influi) e não pode haver repetição de pessoas ⇒ Arranjos sem repetição.

	1ª lugar		2ª lugar		3ª lugar		4ª lugar		5ª lugar
Disposição das pessoas	–		–		H + alto		–		–
Nº de possibilidades	4	×	3	×	1	×	2	×	1

Podem-se dispor de $^{4}A_{4} = 24$ modos diferentes.

b) Uma vez que as pessoas são diferentes umas das outras, a ordem tem influência na contagem e não pode haver repetição de pessoas ⇒ Arranjos sem repetição.

	1ª lugar		2ª lugar		3ª lugar		4ª lugar		5ª lugar
Disposição das pessoas	H		M		H		M		H
Nº de possibilidades	3	×	2	×	2	×	1	×	1

Podem-se dispor de $^{3}A_{3}\ ^{2}A_{2} = 12$ modos diferentes.

c) Uma vez que as pessoas são diferentes umas das outras, a ordem tem influência na contagem e não pode haver repetição de pessoas ⇒ Arranjos sem repetição.

	1ª lugar		2ª lugar		3ª lugar		4ª lugar		5ª lugar
Nº de possibilidades	5	×	4	×	3	×	2	×	1

Podem-se sentar de $^{5}A_{5} = P_{5} = 5! = 120$ modos diferentes.

3.10 Exercícios propostos

1. Numa entrevista, um economista afirmou que considerava a "melhoria" da situação económica tão provável como a sua "estagnação". No entanto, encarava a "melhoria" como duas vezes mais provável do que a "quebra" da atividade económica.
   a) Que espaço de resultados está implícito nestas afirmações?
   b) Qual a probabilidade associada a cada resultado deste espaço?
   c) Que conceito de probabilidade está implícito neste problema?

2. Numa determinada empresa, o ordenado dos homens pode tomar os valores, de 1000 u.m., 1500 u.m., 2000 u.m. e 2500 u.m. As mulheres podem usufruir os seguintes ordenados: 500 u.m., 1000 u.m., 1500 u.m.

e 2000 u.m. Admitindo que a percentagem de homens a auferir cada um dos valores é a mesma, e o mesmo em relação às mulheres, e que os ordenados dos homens são independentes dos ordenados das mulheres, qual a probabilidade de um casal, escolhido aleatoriamente:

a) Ganhar mais de 2500 u.m.
b) Ganhar um múltiplo de 1000 u.m.
c) Ganhar entre 2000 e 3500 u.m. (inclusivé).

3. Na Britolândia existem no mercado três operadoras de telemóvel, A, B e C, com as seguintes percentagens de adesão:

$$P(A) = P(B) = P(C) = \frac{1}{4}, \quad P(A \cap B) = P(B \cap C) = 0, \quad P(A \cap C) = \frac{1}{8}.$$

Calcule a probabilidade de um indivíduo, escolhido ao acaso, ser aderente de pelo menos uma das operadoras.

4. Sejam A e B acontecimentos tais que $P(A) = 0{,}2$, $P(B) = p$ e $P(A \cup B) = 0{,}6$. Calcule p considerando A e B:

a) Mutuamente exclusivos.
b) Independentes.

5. Uma caixa contém 100 peças sendo 10 defeituosas. Considere-se a experiência aleatória que consiste em extrair sucessivamente 2 peças da caixa. Qual a probabilidade de se realizar o acontecimento A – a primeira peça é não defeituosa e a segunda é defeituosa?

6. Um teste para a deteção do vírus da SIDA foi aplicado a 5100 portadores e a 9900 não portadores deste vírus, tendo-se obtido os seguintes resultados:

Resultado do teste	Portador	Não portador
Positivo	4950	750
Negativo	150	9150

a) Calcule a probabilidade de um indivíduo escolhido ao acaso, de entre os submetidos ao teste:
b) Ter um resultado positivo no teste.
c) Ter um resultado positivo no teste e ser portador do vírus.
d) Não ser portador do vírus e ter um resultado negativo.
e) Ter um resultado positivo sabendo que não é portador do vírus.
f) Ter um resultado negativo sabendo que é portador do vírus.
g) Ser portador do vírus sabendo que o teste é positivo.
h) Não ser portador da doença sabendo que o teste deu negativo.
i) O resultado do teste é independente do facto do indivíduo ser portador do vírus?

7. Um estudante tem 3 exames. A probabilidade de ter nota positiva em cada um é de ½ e os resultados são independentes. Calcule a probabilidade de ter nota positiva:

a) Em pelo menos um exame.
b) Exatamente um exame.

8. Os sintomas febre, cansaço e dores no corpo, estão associadas em 60% dos casos às gripes e 40% às constipações. A automedicação é muito frequente nestas doenças, verificando-se que 40% das vezes os medicamentos ingeridos para o tratamento da gripe são os aconselhados para as constipações, e em 70% das situações os medicamentos utilizados para tratamento das constipações são os indicados para a gripe.

a) Qual a probabilidade de o medicamento ingerido ser realmente o indicado?
b) Sabendo que o medicamento era o apropriado para a doença, qual a probabilidade de o doente ter tido gripe?

9. O Pedro entrou agora na universidade e foi informado de que há 30% de possibilidade de vir a receber uma bolsa de estudo. No caso de a receber, a probabilidade de se licenciar é de 0,85, enquanto que no caso de não a obter, a probabilidade de se licenciar é de apenas 0,45.

a) Diga ao Pedro qual é a probabilidade de ele se licenciar.
b) Se, daqui a uns anos, encontrar o Pedro já licenciado, qual a probabilidade de que tenha recebido a bolsa de estudo?

10. Numa urbanização recente, um inquérito aos moradores revelou que 5% viviam em moradias, 20% em prédios em banda e os restantes em torres. Alguns desses moradores foram aí instalados através de programas de realojamento. Dos moradores que vivem em moradias 2% são realojados, o mesmo acontecendo com 3% dos que vivem em prédios em banda e com 10% dos que vivem em torres. Escolhido ao acaso um dos habitantes dessa urbanização, qual a probabilidade de:

a) Ele ter sido alvo do programa de realojamento?
b) Ele viver numa moradia sabendo que se trata de um realojado?

4 Variáveis aleatórias

Usualmente interessa determinar, no âmbito de uma experiência aleatória, a probabilidade de certos acontecimentos ocorrerem. Para tal é necessário quantificar os resultados dessa experiência, recorrendo às variáveis aleatórias.

4.1 Noção de variável aleatória

Muitas vezes o espaço de resultados, Ω, não é um conjunto numérico e o estudo a realizar é bastante facilitado quando se atribui um número real, x, ou conjunto ordenado de números reais, a cada elemento ω de Ω, i. e., $\omega \in \Omega$. Define-se, assim, uma função $X(\omega) = x$ que se designa por variável aleatória X.

Definição: Uma **variável aleatória** (v.a.), $\boldsymbol{X}$, é uma função que a cada resultado $\omega \in \Omega$ associa um valor numérico $X(\omega) \in \mathbb{R}$.

Na Figura 4.1 apresenta-se um esquema ilustrativo da definição proposta.

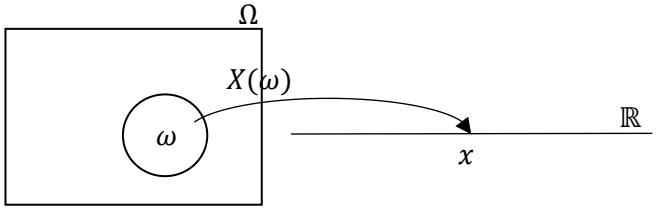


Figura 4.1: *Probabilidade utilizando a variável aleatória X.*

Portanto, ao conhecer-se as probabilidades dos vários elementos ω de Ω podem-se definir as probabilidades dos números reais x, ou seja, dos possíveis valores da variável aleatória X.

Se o conjunto de chegada de uma v. a. for constituído por elementos expressos numa escala nominal ou ordinal, então diz-se que a variável é **qualitativa**. Se os elementos forem expressos numa escala numérica, então diz-se que a variável é **quantitativa**.

Uma variável **quantitativa** classifica-se como discreta ou contínua conforme os elementos do contradomínio da aplicação que a define forem numeráveis ou não numeráveis.

Exemplo: Considere-se a experiência aleatória que consiste em lançar uma moeda ao ar três vezes e registar a face que ficou voltada para cima, face (F) ou coroa (C). O espaço de resultados para esta experiência é:
$\Omega = \{$(F, F, F); (F, F, C); (F, C, F); (C, F, F); (F, C, C); (C, F, C); (C, C, F); (C, C, C)$\}$
Definindo X – v. a. que representa o número de faces voltadas para cima em três lançamentos, então:

X(F, F, F) = 3	X(F, F, C) = 2	X(F, C, F) = 2	X(C, F, F) = 2
X(F, C, C) = 1	X(C, F, C) = 1	X(C, C, F) = 1	X(C, C, C) = 0

Logo, $x = X(\Omega) = 0, 1, 2, 3$.

Definição: Uma **v. a.** $(\boldsymbol{X}, \boldsymbol{Y})$ **bidimensional** é uma função que a cada resultado $\omega \in \Omega$ associa um valor numérico $X(\omega) \in \mathbb{R}^2$.

Exemplo: Considere-se a experiência aleatória que consiste em escolher aleatoriamente um aluno que tenha realizado, no ano passado, os exames de Estatística I e II.

Então a cada aluno é possível fazer corresponder um par de valores (x, y) onde X representa a nota no exame de Estatística I e Y representa a nota no exame de Estatística II, com $x = 0, 1, \ldots, 20$ e $y = 0, 1, \ldots, 20$. Portanto, a variável aleatória a estudar será $Z = (X, Y)$ – notas obtidas, pelo aluno, nos exames de Estatística I e II, com $Z(\Omega) = \{(x, y): x, y = 0, 1, \ldots, 20\}$.

4.2 Variáveis aleatórias unidimensionais

O objetivo agora é calcular probabilidades, não com base nos próprios acontecimentos, mas sim nos valores assumidos pela variável aleatória.

4.2.1 Variáveis aleatórias discretas

Uma v. a. X diz-se **discreta** se o número de valores possíveis para essa variável for finito ou infinito numerável.

4.2.1.1 Função de probabilidade

Definição: A **função (massa) de probabilidade**, $\boldsymbol{f(x)}$, da v. a. X designa a probabilidade dessa variável tomar cada um dos valores do seu domínio D, i. e.:

$$f(x) = \begin{cases} P(X = x), & \text{se } x \in D; \\ 0, & \text{se } x \notin D. \end{cases}$$

Propriedades de $\boldsymbol{f(x)}$: A função $f(x)$ tem que satisfazer as seguintes condições:

1. $0 \leq f(x) \leq 1$ qualquer que seja o valor de x;
2. $\sum_{x \in D} f(x) = 1.$

4.2.1.2 Função de distribuição

Definição: A **função de distribuição**, $\boldsymbol{F(x)}$, da v. a. X corresponde à acumulação da função de probabilidade para cada valor do domínio D da variável, i. e.:

$$F(x) = P(X \leq x) = \sum_{i \leq x} f(i).$$

Propriedades de $\boldsymbol{F(x)}$: A função $F(x)$ tem que satisfazer as seguintes condições:

1. $0 \leq F(x) \leq 1$, qualquer que seja o valor de x;
2. $F(x_1) \leq F(x_2)$, quaisquer que sejam x_1 e x_2 com $x_1 < x_2$, i. e., $F(x)$ é uma função monótona não decrescente;
3. $\lim_{x \to -\infty} F(x) = 0$ e $\lim_{x \to +\infty} F(x) = 1$;
4. $F(x)$ é contínua à direita;
5. $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$, quaisquer que sejam x_1 e x_2 com $x_1 < x_2$.

4.2.2 Variáveis aleatórias contínuas

Uma v. a. X diz-se **contínua** se o número de valores possíveis para essa variável não for numerável.

4.2.2.1 Função de densidade de probabilidade

Defina-se $f(x)$ como a **função densidade de probabilidade** da v.a. X.

Propriedades de $f(x)$: A função $f(x)$ tem que satisfazer as seguintes condições:

1. $f(x) \geq 0$, qualquer que seja o valor de x;
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Observação: Se X é v.a. contínua então $f(x) \neq P(X = x)$ e $P(X = x) = 0$.

4.2.2.2 Função de distribuição

Definição: A **função de distribuição**, $F(x)$, da v. a. X é definida por:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

Propriedades de $F(x)$: A função $F(x)$ tem que satisfazer as seguintes condições:

1. $0 \leq F(x) \leq 1$, qualquer que seja o valor de x;
2. $F(x_1) \leq F(x_2)$, quaisquer que sejam x_1 e x_2 com $x_1 < x_2$, i. e., $F(x)$ é uma função monótona não decrescente;
3. $\lim_{x \to -\infty} F(x) = 0$ e $\lim_{x \to +\infty} F(x) = 1$;
4. $F(x)$ é contínua à direita;
5. $F'(x) = \frac{\partial F(x)}{\partial x} = f(x)$;
6. $P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f(x)dx = F(x_2) - F(x_1)$, quaisquer que sejam x_1 e x_2 com $x_1 < x_2$.

4.2.3 Exercícios resolvidos

4.2.3.1 Variáveis aleatórias discretas

Considere o exemplo do lançamento da moeda, descrito na seção 4.1, onde X é a v. a. que representa o número de faces voltadas para cima em três lançamentos da moeda. Construa as funções de probabilidade e de distribuição associadas a esta variável aleatória. Represente-as graficamente.

Resolução:

Função de probabilidade: Para a variável aleatória X têm-se as seguintes probabilidades:

$$f(0) = P(X = 0) = P\big((C, C, C)\big) = \frac{1}{8};$$

$$f(1) = P(X = 1) = P\big((F, C, C) \cup (C, F, C) \cup (C, C, F)\big) = \frac{3}{8};$$

$$f(2) = P(X = 2) = P\big((F,F,C) \cup (F,C,F) \cup (C,F,F)\big) = \frac{3}{8};$$
$$f(3) = P(X = 3) = P\big((F,F,F)\big) = \frac{1}{8}.$$

Por conseguinte,

x	0	1	2	3
$f(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

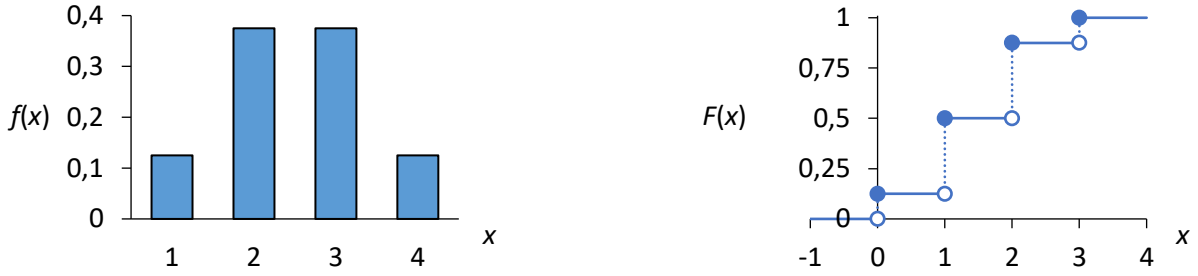
Função de distribuição:

$$F(0) = P(X \leq 0) = P(X = 0) = \frac{1}{8};$$
$$F(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{4}{8};$$
$$F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8};$$
$$F(3) = P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$
$$= \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1.$$

Portanto,

x	0	1	2	3
$F(x)$	$\frac{1}{8}$	$\frac{4}{8}$	$\frac{7}{8}$	1

Na Figura 4.2 é feita a representação gráfica das funções de probabilidade e de distribuição.



a) Função de probabilidade. *b) Função de distribuição.*

Figura 4.2: *Função de probabilidade e de distribuição.*

4.2.3.2 Variáveis aleatórias contínuas

Considere que o consumo semanal de um certo bem alimentar, em famílias de um determinado concelho, é uma v. a. X com a seguinte função densidade de probabilidade:

$$f(x) = \begin{cases} \frac{1}{9}x^2 & 0 < x < 3; \\ 0, & \text{caso contrário.} \end{cases}$$

a) Obtenha a função de distribuição.
b) Represente graficamente as funções densidade de probabilidade e de distribuição.
c) Calcule a probabilidade de, numa semana, o consumo do bem alimentar numa dada família ser superior a 1,5.
d) Calcule a probabilidade de, numa semana, o consumo do bem alimentar numa dada família se situar entre 1 e 2.

Resolução:

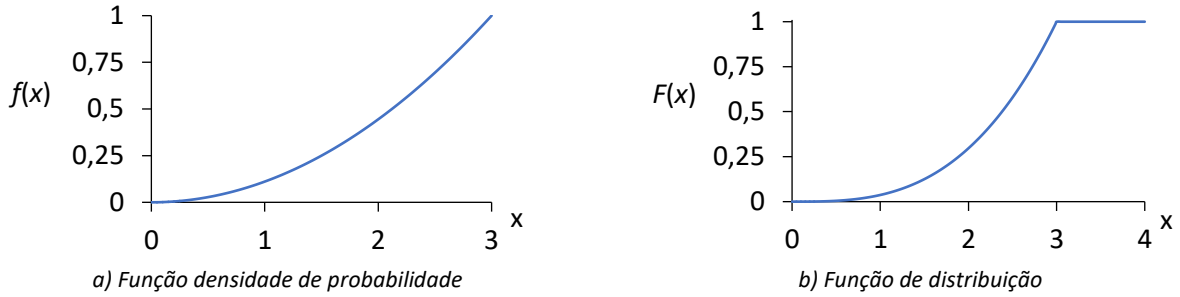
a) Função de distribuição:

$$F(x) = \int_{-\infty}^{x} f(u)du = \int_{0}^{x} \frac{1}{9}u^2 du = \frac{1}{9}\int_{0}^{x} u^2 du = \frac{1}{9}\left[\frac{u^3}{3}\right]_{u=0}^{u=x} = \frac{1}{9}\left(\frac{x^3}{3} - 0\right) = \frac{1}{27}x^3.$$

Portanto,

$$F(x) = \begin{cases} 0, & x \leq 0; \\ \frac{1}{27}x^3, & 0 < x < 3; \\ 1, & x \geq 3. \end{cases}$$

b) Na Figura 4.3 é feita a representação gráfica das funções de densidade e de distribuição.



a) Função densidade de probabilidade *b) Função de distribuição*

Figura 4.3: Funções de densidade de probabilidade e de distribuição.

c) $P(X > 1{,}5) = 1 - P(X \leq 1{,}5) = 1 - F(1{,}5) = 1 - \frac{1}{27} \times 1{,}5^3 = 0{,}875.$

d) $P(1 \leq X \leq 2) = F(2) - F(1) = \frac{1}{27} \times 2^3 - \frac{1}{27} = 0{,}2593.$

4.3 Variáveis aleatórias bidimensionais

4.3.1 Variáveis aleatórias bidimensionais discretas

Uma **v. a. (*X*, *Y*) bidimensional diz-se discreta** se e só se X e Y forem v. a. discretas.

4.3.1.1 Função de probabilidade conjunta

Definição: A **função de probabilidade conjunta**, $\boldsymbol{f(x,y)}$, da v. a. (X, Y) discreta designa a probabilidade dessa variável tomar cada um dos valores do seu domínio:

$$f(x, y) = P(X = x; Y = y).$$

Propriedades de $\boldsymbol{f(x,y)}$: A função $f(x, y)$ tem que satisfazer as seguintes condições:

1. $0 \leq f(x, y) \leq 1$, qualquer que seja o valor de x e y;
2. $\sum_{x \in D_x} \sum_{y \in D_y} f(x, y) = 1$.

4.3.1.2 Função de distribuição conjunta

Definição: A **função de distribuição conjunta**, $\boldsymbol{F(x,y)}$, da v. a. (X, Y) discreta é dada por:

$$F(x, y) = P(X \leq x; Y \leq y) = \sum_{-\infty}^{x} \sum_{-\infty}^{y} f(x, y).$$

Propriedades de $F(x,y)$: A função $F(x,y)$ tem que satisfazer as seguintes condições:

1. $0 \leq F(x,y) \leq 1$, qualquer que seja o valor de x e y;
2. $\lim\limits_{\substack{x\to-\infty \\ y\to-\infty}} F(x,y) = 0$ e $\lim\limits_{\substack{x\to+\infty \\ y\to+\infty}} F(x,y) = 1$;
3. $\lim\limits_{x\to-\infty} F(x,y) = 0$, para todo o y fixo;
4. $\lim\limits_{y\to-\infty} F(x,y) = 0$, para todo o x fixo;
5. $F(x_1,y_1) \leq F(x_2,y_2)$, quaisquer que sejam x_1, x_2, y_1 e y_2 com $x_1 < x_2$ e $y_1 < y_2$.

4.3.1.3 Função de probabilidade marginal

Definição: Considere-se a v. a. bidimensional (X,Y) discreta, com função de probabilidade conjunta $f(x,y)$.

A **função de probabilidade marginal**, $f_X(x)$, da v. a. X discreta é dada por:

$$f_X(x) = P(X = x;\ -\infty < Y < +\infty) = \sum_{y\in D_y} f(x,y).$$

A **função de probabilidade marginal**, $f_Y(y)$, da v. a. Y discreta é dada por:

$$f_Y(y) = P(-\infty < X < +\infty; Y = y) = \sum_{x\in D_x} f(x,y).$$

4.3.1.4 Função de probabilidade condicionada

Definição: Considere-se a v. a. bidimensional (X,Y) discreta, com função de probabilidade conjunta $f(x,y)$.

A **função de probabilidade de X condicionada pela realização do acontecimento** $\{Y = y\}$, com $P(Y = y) > 0$, é dada por:

$$f_{X|Y=y}(x) = P(X = x|Y = y) = \frac{P(X = x; Y = y)}{P(Y = y)} = \frac{f(x,y)}{f_Y(y)} \quad (y \text{ fixo}).$$

A **função de probabilidade de Y condicionada pela realização do acontecimento** $\{X = x\}$, com $P(X = x) > 0$, é dada por:

$$f_{Y|X=x}(y) = P(Y = y|X = x) = \frac{P(X = x; Y = y)}{P(X = x)} = \frac{f(x,y)}{f_X(x)} \quad (x \text{ fixo}).$$

4.3.2 Variáveis aleatórias bidimensionais contínuas

Uma **v. a. (X, Y) bidimensional diz-se contínua** se e só se X e Y forem v. a. contínuas.

4.3.2.1 Função de probabilidade conjunta

Definição: A **função densidade de probabilidade conjunta**, $f(x,y)$, da v. a. (X,Y) contínua tem que satisfazer as seguintes condições:

1. $f(x,y) \geq 0$, qualquer que seja o valor de x e y;
2. $\int_{D_x}\int_{D_y} f(x,y)dy\,dx = 1$.

4.3.2.2 Função de distribuição conjunta

Definição: A **função de distribuição conjunta**, $\boldsymbol{F(x,y)}$, da v. a. (X,Y) contínua é dada por:

$$F(x,y) = P(X \leq x; Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(x,y)dy\,dx.$$

Propriedades de $\boldsymbol{F(x,y)}$: A função $F(x,y)$ tem que satisfazer as seguintes condições:

1. $0 \leq F(x,y) \leq 1$, qualquer que seja o valor de x e y;
2. $\lim\limits_{\substack{x\to-\infty\\y\to-\infty}} F(x,y) = 0$ e $\lim\limits_{\substack{x\to+\infty\\y\to+\infty}} F(x,y) = 1$;
3. $\lim\limits_{x\to-\infty} F(x,y) = 0$, para todo o y fixo;
4. $\lim\limits_{y\to-\infty} F(x,y) = 0$, para todo o x fixo;
5. $\dfrac{\partial^2 F(x,\ y)}{\partial x \partial y} = f(x,y)$;
6. $F(x_1,y_1) \leq F(x_2,y_2)$, quaisquer que sejam x_1, x_2, y_1 e y_2 com $x_1 < x_2$ e $y_1 < y_2$.

4.3.2.3 Função de distribuição marginal

Definição: Considere-se a v. a. bidimensional (X,Y) contínua, com função densidade de probabilidade conjunta $f(x,y)$.

A **função de distribuição marginal**, $\boldsymbol{F_X(x)}$, da v. a. X contínua é dada por:

$$F_X(x) = P(X \leq x;\ -\infty < Y < +\infty) = \int_{-\infty}^{x}\int_{D_y} f(u,y)dy\,du.$$

A **função de distribuição marginal**, $\boldsymbol{F_Y(y)}$, da v. a. Y contínua é dada por:

$$F_Y(y) = P(\ -\infty < X < +\infty; Y \leq y) = \int_{-\infty}^{y}\int_{D_x} f(x,v)dx\,dv.$$

4.3.2.4 Função densidade de probabilidade marginal

Definição: Considere-se a v. a. bidimensional (X,Y) contínua, com função densidade de probabilidade conjunta $f(x,y)$.

A **função densidade de probabilidade marginal**, $\boldsymbol{f_X(x)}$, da v. a. X contínua é dada por:

$$f_X(x) = F_X'(x) = \frac{\partial F_X(x)}{\partial x} = \int_{D_y} f(x,y)dy.$$

A **função densidade de probabilidade marginal**, $\boldsymbol{f_Y(y)}$, da v. a. Y contínua é dada por:

$$f_Y(y) = F_Y'(y) = \frac{\partial F_Y(y)}{\partial y} = \int_{D_x} f(x,y)dx.$$

4.3.2.5 Função densidade de probabilidade condicionada

Definição: Considere-se a v. a. bidimensional (X, Y) contínua, com função densidade de probabilidade conjunta $f(x, y)$.

A **função densidade de probabilidade de X condicionada pela realização do acontecimento** $\{Y = y\}$, com $f_Y(y) > 0$, é dada por:

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)} \quad (y \text{ fixo}).$$

A **função densidade de probabilidade de Y condicionada pela realização do acontecimento** $\{X = x\}$, com $f_X(x) > 0$, é dada por:

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)} \quad (x \text{ fixo}).$$

4.3.3 Independência de variáveis aleatórias

Definição: Dada uma v. a. bidimensional (X, Y), diz-se que **X e Y são independentes** se:

$$f(x, y) = f_X(x) f_Y(y).$$

4.3.4 Exercícios resolvidos

4.3.4.1 Variáveis aleatórias bidimensionais discretas

Num pequeno grupo de casais empregados, o salário X do marido e o salário Y da respetiva esposa, em unidades monetárias (u.m.), têm a seguinte distribuição de probabilidade conjunta:

Salário da esposa (y)	Salário do marido (x)		
	1000	1500	2000
500	0,05	0,1	0,15
1000	0,1	0,2	0,1
1500	0,15	0,1	0,05

a) Determine as funções de probabilidade marginais dos salários.
b) Determine a função de probabilidade do salário do marido, quando o salário da esposa é de 1000 u.m.
c) Construa a função de probabilidade do salário da esposa, quando o salário do marido é de 1500 u.m.
d) Escolhido um casal ao acaso, qual a probabilidade:
   i. Do salário do marido ser inferior ao da esposa?
   ii. Do salário do marido ser inferior ou igual ao da esposa?
e) Analise a independência das variáveis.

Resolução:

a) Função de probabilidade marginal de X:

$$f_X(x) = P(X = x;\ -\infty < Y < +\infty),$$

i. e., probabilidade de o salário do marido ser igual a x.

- 1000 u.m.:

$$f_X(1000) = P(X = 1000; Y = 500) + P(X = 1000; Y = 1000) + P(X = 1000; Y = 1500)$$
$$= f(1000, 500) + f(1000, 1000) + f(1000, 1500) = 0{,}05 + 0{,}1 + 0{,}15 = 0{,}3.$$

- 1500 u.m.:

$$f_X(1500) = P(X = 1500; Y = 500) + P(X = 1500; Y = 1000) + P(X = 1500; Y = 1500)$$
$$= f(1500, 500) + f(1500, 1000) + f(1500, 1500) = 0{,}1 + 0{,}2 + 0{,}1 = 0{,}4.$$

- 2000 u.m.:

$$f_X(2000) = P(X = 2000; Y = 500) + P(X = 2000; Y = 1000) + P(X = 2000; Y = 1500)$$
$$= f(2000, 500) + f(2000, 1000) + f(2000, 1500) = 0{,}15 + 0{,}1 + 0{,}05 = 0{,}3.$$

Função de probabilidade marginal de Y:

$$f_Y(y) = P(-\infty < X < +\infty; Y = y),$$

i. e., probabilidade de o salário da esposa ser igual a y.

- 500 u.m.:

$$f_Y(500) = P(X = 1000; Y = 500) + P(X = 1500; Y = 500) + P(X = 2000; Y = 500)$$
$$= f(1000, 500) + f(1500, 500) + f(2000, 500) = 0{,}05 + 0{,}1 + 0{,}15 = 0{,}3.$$

- 1000 u.m.:

$$f_Y(1000) = P(X = 1000; Y = 1000) + P(X = 1500; Y = 1000) + P(X = 2000; Y = 1000)$$
$$= f(1000, 1000) + f(1500, 1000) + f(2000, 1000) = 0{,}1 + 0{,}2 + 0{,}1 = 0{,}4.$$

- 1500 u.m.:

$$f_Y(1500) = P(X = 1000; Y = 1500) + P(X = 1500; Y = 1500) + P(X = 2000; Y = 1500)$$
$$= f(1000, 1500) + f(1500{,}1\,500) + f(2000, 1500) = 0{,}15 + 0{,}1 + 0{,}05 = 0{,}3.$$

b) Função de probabilidade de X condicionada a $Y = 1000$:

$$f_{X|Y=1000}(x) = P(X = x|Y = 1000) = \frac{P(X = x, Y = 1000)}{P(Y = 1000)} = \frac{f(x, 1000)}{f_Y(1000)}.$$

- 1000 u.m.:

$$f_{X|Y=1000}(1000) = \frac{f(1000, 1000)}{f_Y(1000)} = \frac{0{,}1}{0{,}4} = 0{,}25.$$

- 1500 u.m.:

$$f_{X|Y=1500}(1500) = \frac{f(1500, 1000)}{f_Y(1000)} = \frac{0{,}2}{0{,}4} = 0{,}5.$$

- 2000 u.m.:

$$f_{X|Y=2000}(2000) = \frac{f(2000, 1000)}{f_Y(1000)} = \frac{0{,}1}{0{,}4} = 0{,}25.$$

c) Função de probabilidade de Y condicionada a $X = 1500$:

$$f_{Y|X=1500}(y) = P(Y = y|X = 1500) = \frac{P(X = 1500; Y = y)}{P(X = 1500)} = \frac{f(1500, y)}{f_X(1500)}. \ (x \text{ fixo})$$

- 500 u.m.:

$$f_{Y|X=1500}(500) = \frac{f(1500, 500)}{f_X(1500)} = \frac{0{,}1}{0{,}4} = 0{,}25.$$

- 1000 u.m.:

$$f_{Y|X=1500}(1000) = \frac{f(1500, 1000)}{f_X(1500)} = \frac{0{,}2}{0{,}4} = 0{,}5.$$

- 1500 u.m.:

$$f_{Y|X=1500}(1500) = \frac{f(1500, 1500)}{f_X(1500)} = \frac{0{,}1}{0{,}4} = 0{,}25.$$

d) i. $P(X < Y) = P(X = 1000;\ Y = 1500) = f(1000, 1500) = 0{,}15.$

ii. $P(X \leq Y) = P(X = 1000;\ Y = 1000) + P(X = 1000;\ Y = 1500) + P(X = 1500;\ Y = 1500)$
$= f(1000, 1000) + f(1000, 1500) + f(1500, 1500) = 0{,}1 + 0{,}15 + 0{,}1 = 0{,}35.$

e) X e Y são independentes se: $f(x,y) = f_X(x)f_Y(y)$, qualquer que seja o x e y.

Por ex., para $x = 1000$ e $y = 500$:

$$f_{XY}(1000;500) = 0{,}05;\ \ f_X(1000) = 0{,}3;\ \ f_Y(500) = 0{,}3.$$

Portanto, as variáveis X e Y não são independentes, pois $f(1000;500) \neq f_X(1000)f_Y(500)$

Em alternativa, X e Y são independentes se $f_{Y|X=x}(y) = f_Y(y)$ ou $f_{X|Y=y}(x) = f_X(x)$ para todos os valores $x \in D_x$ e $y \in D_y$. Nas alíneas a) e b) já se obtiveram $f_Y(y)$ e $f_{Y|X=1500}(y)$, sendo que $f_{Y|X=1500}(y) \neq f_Y(y)$ para $y = 500, 1000, 1500$. Portanto, as variáveis não são independentes.
Observação: basta falhar a igualdade para um dos valores de y ou y.

4.3.4.2 Variáveis aleatórias bidimensionais contínuas

Considere duas variáveis aleatórias X e Y, que representam a proporção de alunos aprovados nas disciplinas Estatística I e II, repectivamente, cuja função densidade conjunta é:

$$f(x,y) = kx(2 - x + y), \qquad 0 \leq x \leq 1, \qquad 0 \leq y \leq 1.$$

a) Qual o valor de k?
b) Qual a probabilidade de a proporção de alunos aprovados em Estatística I ser inferior à dos alunos aprovados em Estatística II?
c) Determine as funções densidade de probabilidade marginais.
d) Qual a probabilidade de serem aprovados mais de 50% dos alunos em Estatística I?
e) Determine $f_{X|Y}(x)$.
f) Analise a independência das variáveis.

Resolução:

a) Se $f(x,y)$ é função densidade de probabilidade então

$$\int_{D_x}\int_{D_y} f(x,y)dy\,dx = 1 \Leftrightarrow \int_0^1\int_0^1 kx(2-x+y)dy\,dx = 1$$

$$\Leftrightarrow k\int_0^1\left(2x - x^2 + x\int_0^1 y dy\right)dx = 1 \Leftrightarrow k\int_0^1\left(2x - x^2 + x\left[\frac{y^2}{2}\right]_{y=0}^{y=1}\right)dx = 1$$

$$\Leftrightarrow k\int_0^1\left(2x - x^2 + \frac{x}{2}\right)dx = 1 \Leftrightarrow k\left[x^2 - \frac{x^3}{3} + \frac{x^2}{4}\right]_{x=0}^{x=1} = 1 \Leftrightarrow k = \frac{12}{11}.$$

Portanto,

$$f(x,y) = \frac{12x}{11}(2 - x + y), \qquad 0 \leq x \leq 1, \qquad 0 \leq y \leq 1.$$

b) A probabilidade pretendida corresponde a

$$P(X < Y) = \int_0^1\int_0^y f(x,y)dx\,dy = \int_0^1\int_0^y \frac{12x}{11}(2-x+y)dx\,dy = \frac{12}{11}\int_0^1\int_0^y (2x - x^2 + xy)dx\,dy$$

$$= \frac{12}{11}\int_0^1\left(\left[x^2 - \frac{x^3}{3} + \frac{x^2}{2}y\right]_{x=0}^{x=y}\right)dy = \frac{12}{11}\int_0^1\left(y^2 - \frac{y^3}{3} + \frac{y^3}{2}\right)dy = \frac{12}{11}\left[\frac{y^3}{3} + \frac{y^4}{24}\right]_{y=0}^{y=1}$$

$$= 0{,}4091.$$

c) Função densidade de probabilidade marginal da v. a. X:

$$f_X(x) = \int_{D_y} f(x,y)dy = \int_0^1 \frac{12x}{11}(2-x+y)dy = \frac{12}{11}\left(2x - x^2 + x\left[\frac{y^2}{2}\right]_{y=0}^{y=1}\right) = \frac{12}{11}\left(\frac{5}{2}x - x^2\right)$$

$$= \frac{6x(5-2x)}{11}, \qquad 0 \leq x \leq 1.$$

Função densidade de probabilidade marginal da v. a. Y:

$$f_Y(y) = \int_{D_x} f(x,y)dx = \int_0^1 \frac{12x}{11}(2-x+y)dx = \frac{12}{11}\left[x^2 - \frac{x^3}{3} + \frac{x^2}{2}y\right]_{x=0}^{x=1} = \frac{12}{11}\left(\frac{2}{3} + \frac{1}{2}y\right)$$
$$= \frac{8+6y}{11}, \qquad 0 \le y \le 1.$$

d) A probabilidade pretendida corresponde a

$$P(X > 0{,}5) = \int_{0{,}5}^1 f_X(x)dx = \int_{0{,}5}^1 \frac{6x(5-2x)}{11}dx = \frac{6}{11}\left[\frac{5}{2}x^2 - \frac{2}{3}x^3\right]_{x=0{,}5}^{x=1}$$
$$= \frac{6}{11}\left(\frac{5}{2} - \frac{2}{3} - \frac{5}{2}0{,}5^2 + \frac{2}{3}0{,}5^3\right) = 0{,}7045.$$

e) Função densidade de probabilidade de X condicionada por Y:

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)} = \frac{\frac{12x}{11}(2-x+y)}{\frac{8+6y}{11}} = \frac{6x(2-x+y)}{4+3y}, \quad 0 \le x \le 1.$$

f) Nas alíneas c) e e) já se obtiveram $f_X(x)$ e $f_{X|Y=y}(x)$ sendo que $f_{X|Y=y}(x) \neq f_X(x)$, para pelo menos um valor de x. Portanto, as variáveis não são independentes.
Em alternativa,

$$f_{XY}(x;y) = \frac{12x}{11}(2-x+y);$$

$$f_X(x) = \frac{6x(5-2x)}{11}; \; f_Y(y) = \frac{8+6y}{11} \;\Rightarrow\; f_X(x)f_Y(y) = \frac{3x}{11}(20-8x+9y).$$

Portanto, as variáveis X e Y não são independentes, pois $f_{XY}(x;y) \neq f_X(x)f_Y(y)$.

4.4 Parâmetros de variáveis aleatórias

4.4.1 Média ou valor esperado

Definição: Seja X uma v. a. com função de probabilidade $f(x)$. O **valor esperado** de X (ou média de X), $\boldsymbol{E(X)}$ ou $\boldsymbol{\mu_X}$, quando existe, define-se por:

V. a. discreta	V. a. contínua
$E(X) = \mu_X = \sum_{x \in D_x} xf(x)$	$E(X) = \mu_X = \int_{D_x} xf(x)dx$

Propriedades do valor esperado: Sendo X e Y variáveis aleatórias e k uma constante real:

1. $E(k) = k$;
2. $E(kX) = kE(X)$;
3. $E(X+Y) = E(X) + E(Y)$;
4. $E(X-Y) = E(X) - E(Y)$;
5. $E(XY) = \begin{cases} E(X)E(Y), & \text{se } X \text{ e } Y \text{ independentes;} \\ E(X)E(Y) + Cov(X,Y), & \text{se } X \text{ e } Y \text{ não independentes.} \end{cases}$

Observação:

V. a. discreta	V. a. contínua
$E(XY) = \sum_{x \in D_x} \sum_{y \in D_y} xyf(x,y)$	$E(XY) = \int_{D_x} \int_{D_y} xyf(x,y)dydx$

Definição: Seja (X,Y) uma v. a. bidimensional sendo $f_{X|Y=y}(x)$ a função de probabilidade de X condicionada pela realização do acontecimento $\{Y = y\}$. O **valor esperado** de X (ou média de X) **condicionado**, $\boldsymbol{E(X|Y=y)}$ ou $\boldsymbol{\mu_{X|Y=y}}$, quando existe, define-se por:

V. a. discreta	V. a. contínua
$E(X\|Y=y) = \sum_{x \in D_x} xf_{X\|Y=y}(x)$	$E(X\|Y=y) = \int_{D_x} xf_{X\|Y=y}(x)dx$

O $E(X|Y=y)$ é uma função que depende do valor de y e, portanto, é uma v. a. cujo

$$E(E(X|Y=y)) = E(X).$$

Se as v.a. X e Y forem independentes, então $E(X|Y=y) = E(X)$.

4.4.2 Variância e desvio padrão

Definição: Seja X uma v. a. com função de probabilidade $f(x)$. A **variância** de X, $\boldsymbol{Var(X)}$ ou $\boldsymbol{\sigma_X^2}$, quando existe, define-se por:

$$Var(X) = \sigma_X^2 = E((X-\mu_X)^2) = E(X^2) - \big(E(X)\big)^2.$$

Portanto,

V. a. discreta	V. a. contínua
$\sigma_X^2 = \sum_{x \in D_x} (x-\mu_X)^2 f(x)$	$\sigma_X^2 = \int_{D_x} (x-\mu_X)^2 f(x)\,dx$
$= \sum_{x \in D_x} x^2 f(x) - \left(\sum_{x \in D_x} xf(x)\right)^2$	$= \int_{D_x} x^2 f(x)dx - \left(\int_{D_x} xf(x)dx\right)^2$

O **desvio padrão** de X, $\boldsymbol{\sigma_X}$, quando existe, define-se por: $\sigma_X = \sqrt{Var(X)}$.

Observação: (usando as propriedades do valor esperado)

$$\sigma_X^2 = E((X-\mu_X)^2) = E(X^2 - 2X\mu_X + \mu_X^2) = E(X^2) - E(-2X\mu_X) + E(\mu_X^2)$$
$$= E(X^2) - 2\mu_X E(X) + \mu_X^2,$$

como $\mu_X = E(X)$, então

$$= E(X^2) - 2\big(E(X)\big)^2 + \big(E(X)\big)^2$$
$$= E(X^2) - \big(E(X)\big)^2.$$

Propriedades da variância: Sendo X e Y v. a. e k uma constante real:

1. $Var(k) = 0$;
2. $Var(kX) = k^2 Var(X)$;
3. $Var(k + X) = Var(X)$;
4. $Var(X + Y) = \begin{cases} Var(X) + Var(Y), & \text{se } X \text{ e } Y \text{ independentes;} \\ Var(X) + Var(Y) + 2Cov(X,Y), & \text{se } X \text{ e } Y \text{ não independentes.} \end{cases}$
5. $Var(X - Y) = \begin{cases} Var(X) + Var(Y), & \text{se } X \text{ e } Y \text{ independentes;} \\ Var(X) + Var(Y) - 2Cov(X,Y), & \text{se } X \text{ e } Y \text{ não independentes.} \end{cases}$

Definição: Seja (X, Y) uma v. a. bidimensional. A **variância** de X **condicionada** pela realização do acontecimento $\{Y = y\}$, $\boldsymbol{Var(X|Y = y)}$ ou $\boldsymbol{\sigma^2_{X|Y=y}}$, quando existe, define-se por:

$$Var(X|Y = y) = E(X^2|Y = y) - \big(E(X|Y = y)\big)^2.$$

V. a. discreta	V. a. contínua
$\sigma^2_{X\|Y=y} = \sum_{x \in D_x} (x - \mu_X)^2 f_{X\|Y=y}(x)$	$\sigma^2_{X\|Y=y} = \int_{D_x} (x - \mu_X)^2 f_{X\|Y=y}(x)\, dx$

4.4.3 Momentos

Definem-se **momentos de ordem r em relação a um ponto V**, $\boldsymbol{\mu'_{r,V}}$, como:

$$\mu'_{r,V} = E((X - V)^r).$$

Portanto,

V. a. discreta	V. a. contínua
$\mu'_{r,V} = \sum_{x \in D_x} (x - V)^r f(x)$	$\mu'_{r,V} = \int_{D_x} (x - V)^r f(x) dx$

Casos particulares:

- $V = 0$: momentos ordinários ou em relação à origem: $\mu'_r = E(X^r)$;
- $V = \mu$: momentos centrados ou em relação à média: $\mu_r = E((X - \mu)^r)$.

4.4.4 Covariância

Definição: Seja (X, Y) uma v. a. bidimensional com função de probabilidade conjunta $f(x, y)$. A **covariância** entre X e Y, $\boldsymbol{Cov(X, Y)}$ ou $\boldsymbol{\sigma_{XY}}$, quando existe, define-se por:

$$Cov(X,Y) = \sigma_{XY} = E\big((X - \mu_X)(Y - \mu_Y)\big) = E(XY) - \mu_X \mu_Y = E(XY) - E(X)E(Y).$$

Portanto,

V. a. discreta	V. a. contínua
$\sigma_{XY} = \sum_{x \in D_x} \sum_{y \in D_y} xyf(x,y) - \mu_X \mu_Y$	$\sigma_{XY} = \int_{D_x} \int_{D_y} xyf(x,y) dy\, dx - \mu_X \mu_Y$

Observação: (usando as propriedades do valor esperado)

$$\begin{aligned}\sigma_{XY} &= E\big((X-\mu_X)(Y-\mu_Y)\big) = E(XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y)\\ &= E(XY) - E(X\mu_Y) - E(Y\mu_X) + E(\mu_X\mu_Y)\\ &= E(XY) - \mu_Y E(X) - \mu_X E(Y) + \mu_X\mu_Y,\\ &\quad \text{como } \mu_X = E(X) \text{ e } \mu_Y = E(Y)\\ &= E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y)\\ &= E(XY) - E(X)E(Y) = E(XY) - \mu_X\mu_Y.\end{aligned}$$

Propriedades da covariância: Sendo X e Y variáveis aleatórias e c e k constantes reais:

1. $Cov(cX, Y) = cCov(X, Y)$;
2. $Cov(X, kY) = kCov(X, Y)$;
3. $Cov(cX, kY) = ckCov(X, Y)$.

Teorema: Se as v.a. X e Y forem independentes, então $Cov(X, Y) = 0$.

4.4.5 Coeficiente de correlação linear

Definição: Seja (X, Y) uma v. a. bidimensional com função de probabilidade conjunta $f(x, y)$. A **correlação** linear entre X e Y, $\boldsymbol{Corr(X, Y)}$ ou $\boldsymbol{\rho_{XY}}$, quando existe, define-se por:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}, \quad -1 \le \rho_{XY} \le 1.$$

Propriedades do coeficiente de correlação linear:

- $\rho_{XY} = -1$, correlação linear *negativa* perfeita entre X e Y;
- $\rho_{XY} = 0$, *não existe* correlação linear entre X e Y;
- $\rho_{XY} = 1$, correlação linear *positiva* perfeita entre X e Y.

4.4.6 Exercícios resolvidos

4.4.6.1 Variável aleatória discreta

Retome o exercício da secção 4.2.3.1 e determine:

a) O número médio de lançamentos em que a face fica voltada para cima.
b) $E(4X + 2)$ e $E(X^2)$.
c) $Var(X)$ e $Var(2X + 5)$.

Resolução:

a) $E(X) = \sum_{x=0}^{3} xf(x) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 1{,}5.$

Em 3 lançamentos em média ficam 1,5 faces voltadas para cima.

b) $E(4X + 2) = E(4X) + E(2) = 4E(X) + 2 = 4 \times \frac{12}{8} + 2 = 8.$

c) $E(X^2) = \sum_{x=0}^{3} x^2 f(x) = 0^2 \times \frac{1}{8} + 1^2 \times \frac{3}{8} + 2^2 \times \frac{3}{8} + 3^2 \times \frac{1}{8} = 3.$

$$Var(X) = E(X^2) - \big(E(X)\big)^2 = \frac{24}{8} - \left(\frac{12}{8}\right)^2 = 0{,}75.$$
$$Var(2X+5) = Var(2X) = 2^2 Var(X) = 4 \times 0{,}75 = 3.$$

4.4.6.2 Variável aleatória contínua

Retome o exercício da secção 4.2.3.2 e determine:

a) O consumo médio semanal do bem alimentar, e uma determinada família.
b) O consumo mediano semanal do bem alimentar, e uma determinada família.
c) Calcule o terceiro quartil.
d) Calcule o valor esperado e o desvio padrão do consumo mensal (4 semanas), de uma dada família desse concelho. Admita a independência entre os consumos semanais

Resolução:

a) $$E(X) = \int_0^3 x f(x)dx = \frac{1}{9}\int_0^3 x^3 dx = \frac{1}{9}\left[\frac{x^4}{4}\right]_{x=0}^{x=3} = \frac{1}{9}\left(\frac{3^4}{4} - 0\right) = 2{,}25.$$

b) Sabemos que $F(\tilde{\mu}) = 0{,}5$, onde $\tilde{\mu}$ representa a mediana.
Portanto,
$$\frac{1}{27}\tilde{\mu}^3 = 0{,}5 \Leftrightarrow \tilde{\mu}^3 = 13{,}5 \Rightarrow \tilde{\mu} = \sqrt[3]{13{,}5} \Leftrightarrow \tilde{\mu} = 2{,}3811.$$

c) Sabemos que $F(Q_3) = 0{,}75$, onde Q_3 representa o terceiro quartil.
Portanto,
$$\frac{1}{27}Q_3^3 = 0{,}75 \Leftrightarrow Q_3^3 = 20{,}25 \Rightarrow Q_3 = \sqrt[3]{20{,}25} \Leftrightarrow Q_3 = 2{,}7257.$$

d) Seja X_i a v.a. que representa o consumo na semana i e M a v.a. que representa o consumo mensal.
O consumo mensal é dado por $M = X_1 + X_2 + X_3 + X_4$.
O consumo mensal esperado é:
$$\mu_M = E(M) = E(X_1 + X_2 + X_3 + X_4) = E(X_1) + E(X_2) + E(X_3) + E(X_4)$$
$$= E(X) + E(X) + E(X) + E(X) = 4 \times 2{,}25 = 9.$$
A variância do consumo mensal é:
$$\sigma_M^2 = Var(M) = Var(X_1 + X_2 + X_3 + X_4) = Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4),$$
porque a
$$Cov\big(X_i, X_j\big) = 0, i \neq j,$$
uma vez que os consumos semanais são independentes.
Como
$$Var(X_i) = Var(X) = E(X^2) - \big(E(X)\big)^2 = \int_0^3 x^2 f(x)dx - 2{,}25^2 = \frac{1}{9}\int_0^3 x^4 dx - 2{,}25^2$$
$$= \frac{1}{9}\left[\frac{x^5}{5}\right]_{x=0}^{x=3} - 2{,}25^2 = \frac{1}{9}\left(\frac{3^5}{5} - 0\right) - 2{,}25^2 = 0{,}3375,$$
então
$$\sigma_M^2 = Var(M) = 4 \times 0{,}3375 = 1{,}35$$
e o desvio padrão do consumo mensal é
$$\sigma_M = \sqrt{Var(M)} = 1{,}1619.$$

4.4.6.3 Variável aleatória bidimensional discreta

Retome o exemplo da secção 4.3.4.1 e determine:

a) O salário médio dos maridos e o salário médio das esposas.
b) $E(X^2)$.
c) $Var(X)$ e $Var(Y)$.
d) $E(Y|X = 1500)$.
e) O salário líquido do casal que é dado por $Z = 0{,}6X + 0{,}8Y - 100$. Determine o salário líquido médio e respetiva variância.
f) O valor do coeficiente de correlação linear de Pearson entre X e Y. Interprete.

Resolução:

a) $E(X) = \sum_{\substack{x=1000,1500,\\2000}} xf(x) = 1000 \times 0{,}3 + 1500 \times 0{,}4 + 2000 \times 0{,}3 = 1500.$

Em média, os maridos auferem 1500 u.m.

$E(Y) = \sum_{\substack{y=500,1000,\\1500}} yf(y) = 500 \times 0{,}3 + 1000 \times 0{,}4 + 1500 \times 0{,}3 = 1000.$

Em média as esposas auferem 1000 u.m.

b) $E(X^2) = \sum_{\substack{x=1000,1500,\\2000}} x^2 f(x) = 1000^2 \times 0{,}3 + 1500^2 \times 0{,}4 + 2000^2 \times 0{,}3 = 2400000.$

c) $Var(X) = E(X^2) - \big(E(X)\big)^2 = 2400000 - 1500^2 = 150000.$

Em alternativa,

$$Var(Y) = E\left(\big(Y - E(Y)\big)^2\right) = \sum_{\substack{y=500,1000,\\1500}} \big(y - E(Y)\big)^2 f(y)$$
$$= (500 - 1000)^2 \times 0{,}3 + (1000 - 1000)^2 \times 0{,}4 + (1500 - 1000)^2 \times 0{,}3 = 150000.$$

d) $E(Y|X = 1500) = \sum_{\substack{y=500,1000,\\1500}} yf_{Y|X=1500}(y) = 500 \times 0{,}25 + 1000 \times 0{,}5 + 1500 \times 0{,}25 = 1000.$

(Observação: $f_{Y|X=1500}$ já foi obtida anteriormente na secção 4.3.4.2)

Nos casais em que o marido aufere 1500 euros, em média as esposas auferem 1000 u.m.

e) $E(Z) = E(0{,}6X + 0{,}8Y - 100) = 0{,}6E(X) + 0{,}8E(Y) - 100 = 0{,}6 \times 1500 + 0{,}8 \times 1000 - 100$
$= 1600.$

Em média o casal aufere um rendimento líquido de 1600 u.m.

$$Var(Z) = Var(0{,}6X + 0{,}8Y - 100) = Var(0{,}6X) + Var(0{,}8Y) + 2Cov(0{,}6X; 0{,}8Y)$$
$$= 0{,}6^2 Var(X) + 0{,}8^2 Var(Y) + 2 \times 0{,}6 \times 0{,}8 Cov(X, Y)$$
$$= 0{,}6^2 \times 150000 + 0{,}8^2 \times 150000 + 2 \times 0{,}6 \times 0{,}8 \times (-50000) = 102000$$

pois

$E(XY) = \sum_{\substack{x=1000,\\1500,2000}} \sum_{\substack{y=500,\\1000,1500}} xyf(x,y) = 500 \times 1000 \times 0{,}05 + \cdots + 1500 \times 2000 \times 0{,}05 = 1450000,$

$Cov(X, Y) = E(XY) - E(X)E(Y) = 1450000 - 1500 \times 1000 = -50000.$

f) $\rho_{XY} = \dfrac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \dfrac{-50000}{\sqrt{150000 \times 150000}} = -0{,}3333 < 0.$

As variáveis X e Y não são independentes. Existe uma relação fraca do tipo negativo entre o salário do marido e da esposa, i. e., existe uma tendência para quanto mais elevados forem o salário dos maridos menores serão os salários das esposas e vice-versa.

4.4.6.4 Variável aleatória bidimensional contínua

Retome o exercício da secção 4.3.4.2 e determine:

a) A proporção média de alunos aprovados em Estatística I.
b) $Var(2X)$.
c) $Cov(3X, -Y)$.

Resolução:

a) $E(X) = \displaystyle\int_0^1 x f_X(x)dx = \int_0^1 x\frac{6x(5-2x)}{11}dx = \frac{6}{11}\int_0^1 (5x^2 - 2x^3)dx = \frac{6}{11}\left[\frac{5}{3}x^3 - \frac{1}{2}x^4\right]_0^1 = 0{,}6364.$

b) $Var(2X) = 2^2 Var(X) = 4\left(E(X^2) - \left(E(X)\right)^2\right).$

Como,

$$E(X^2) = \int_0^1 x^2 f_X(x)dx = \frac{6}{11}\int_0^1 (5x^3 - 2x^4)dx = \frac{6}{11}\left[\frac{5}{4}x^4 - \frac{2}{5}x^5\right]_0^1 = 0{,}4636,$$

então

$$Var(X) = 0{,}4636 - 0{,}6364^2 = 0{,}0586 \quad \text{e} \quad Var(2X) = 0{,}2345.$$

c) $Cov(3X, -Y) = 3 \times (-1)Cov(X,Y) = -3\sigma_{XY} = -3 \times \left(E(XY) - E(X)E(Y)\right).$

Como,

$$E(XY) = \int_0^1 \int_0^1 xyf(x,y)dy\,dx = \int_0^1 \int_0^1 xy\frac{12x}{11}(2 - x + y)dy\,dx$$
$$= \frac{12}{11}\int_0^1 x^2\left(\left[\frac{2-x}{2}y^2 + \frac{1}{3}y^3\right]_{y=0}^{y=1}\right)dx = \frac{12}{11}\left[\frac{4}{9}x^3 - \frac{1}{8}x^4\right]_{y=0}^{y=1} = 0{,}3485$$

e

$$E(Y) = \int_0^1 y f_y(y)dy = \int_0^1 y\frac{8+6y}{11}dy = \left[\frac{8}{22}y^2 + \frac{6}{33}y^3\right]_{y=0}^{y=1} = 0{,}5455,$$

então

$$Cov(X,Y) = 0{,}3485 - 0{,}6364 \times 0{,}5455 = 0{,}0013$$

e

$$Cov(3X, -Y) = -0{,}004.$$

4.5 Exercícios propostos

1. Uma máquina de jogos tem dois discos que funcionam independentemente um do outro. Cada disco tem 10 figuras: 4 maçãs, 3 bananas, 2 peras e 1 laranja. Uma pessoa paga 80 cêntimos e aciona a máquina. Se aparecerem 2 maçãs, ganha 40 cêntimos. Se aparecerem 2 bananas ganha 80 cêntimos. Se aparecerem 2 peras ganha 140 cêntimos e ganha 180 cêntimos se aparecerem 2 laranjas. Calcule a esperança de ganho numa única jogada.

2. O número de filhos com menos de 18 anos de idade em famílias afro-americanas nos Estados Unidos em 1990 é dado na seguinte tabela:

Número de crianças	0	1	2	3	4 ou mais
Proporção de famílias	0,42	0,24	0,19	0,10	0,05

Seja X a variável aleatória descrita na distribuição de probabilidade acima.

a) Qual a probabilidade de uma família afro-americana escolhida ao acaso ter no máximo um filho?
b) Qual a possibilidade de ter mais de 3 filhos?
c) Qual o número esperado de filhos numa família afro-americana?
d) Calcule o desvio padrão da variável X.

3. Admita que o número de revistas adquiridas por semana, X, pelos cidadãos de uma determinada cidade é descrito pela seguinte função de probabilidade:

x	0	1	2	3
$f(x)$	0,1	0,2	k	0,1

a) Determine o valor de k.
b) Calcule $P(1 < X < 3)$.
c) Qual a probabilidade de um cidadão adquirir pelo menos 2 revistas por semana?
d) Calcule a $P(X = 1|X \leq 2)$.
e) Em média quantas revistas são adquiridas por cidadão, por semana?
f) Determine $E(2X + 4)$.
g) Determine $Var(X)$ e $Var(2X + 4)$.

4. Admita que o número de exames, X, que um aluno realiza até ser aprovado na disciplina de Matemática segue a seguinte distribuição:

x	1	2	3	4 ou mais
$F(x)$	0,2	0,4	0,9	1

a) Determine a função de probabilidade $f(x)$ e represente-a graficamente;
b) Qual a probabilidade de um aluno realizar mais de 3 exames?
c) Calcule a probabilidade de um aluno escolhido ao acaso realizar no máximo 2 exames até obter a provação.
d) Calcule $P(1 < X \leq 2)$.
e) Sabendo que um aluno realizou no máximo 3 exames até ser aprovado na disciplina, qual a probabilidade de não ter realizado mais de 2?
f) Determine o menor valor de a de forma a ter $P(X \leq a) \geq 0{,}5$.
g) Em média, quantos exames realiza um aluno até obter aprovação na disciplina?
h) Calcule $Var(X)$, $E(2X)$ e $Var(3X)$.

5. Considere a experiência aleatória que consiste em lançar um dado equilibrado. Seja X o valor da face que fica voltada para cima.

a) Qual a função de probabilidade de X.
b) Determine $E(X)$ e $E(X^2)$.
c) Calcule $E((X + 3)^2)$ e $Var(3X - 2)$.
d) Sendo Y outra v.a. e sabendo que $Y = X/2 + 3$, determine $E(Y)$ e $Var(Y)$.

6. Um empreiteiro pretende levar a cabo um projeto de construção. Estima que os materiais custarão 25.000 Euros e o seu trabalho 900 Euros por dia. Se o projeto demorar X dias a ser completado, o custo de mão-de-obra total será de $900X$ Euros e o custo total do projecto (em Euros) será:

$$C = 25.000 + 900X.$$

O empreiteiro sabe que o projeto demorará entre 10 a 14 dias e que:

$$P\big((X = 10) \cup (X = 11)\big) = 0{,}4 \qquad P(X = 10) = P(X = 14)$$
$$P(X = 12) = 2P(X = 14) \qquad P(X = 11) = P(X = 13)$$

a) Indique a função de probabilidade da v. a. X.
b) Determine o custo total esperado do projeto (variável C).
c) Qual a probabilidade do projeto demorar no máximo 12 dias?

7. Uma cadeia de hipermercados vende, por semana, uma quantidade de peixe (em toneladas) que admitimos ser uma v. a. X com a seguinte função densidade de probabilidade:

$$f(x) = \begin{cases} \frac{1}{3}, & 0 < x < 3, \\ 0, & \text{caso contrário,} \end{cases}$$

a) Represente graficamente a função de densidade de probabilidade.
b) Descreva a função de distribuição de X e represente-a graficamente.
c) Calcule a $P(X \leq 2)$.
d) Calcule $E(X)$ e $Var(X)$.

8. Suponha que a necessidade diária de um dado medicamento num dado hospital pode ser descrita pela v. a. X com a seguinte função densidade de probabilidade:

$$f(x) = \begin{cases} cx^2, & 0 < x < 3, \\ 0, & \text{caso contrário.} \end{cases}$$

a) Determine o valor de c.
b) Calcule a $P(1 < X < 2)$.
c) Calcule $E(X)$ e $Var(3X)$.

9. Considere o vetor aleatório (X, Y) que representa o número de consultas anuais que um grupo de utentes requereu no serviço público (X) e no serviço privado (Y), catacterizado pela seguinte função de probabilidade conjunta.

	y		
x	2	3	4
1	1/12	1/6	0
2	1/6	0	1/3
3	1/12	1/6	0

a) Obtenha as funções de probabilidade marginais e as funções de distribuição marginais do número de consultas no serviço público e no serviço privado.
b) Determine a probabilidade de, ao escolher um utente ao acaso, este ter tido um produto do número de consultas par.
c) Qual a probabilidade de um utente selecionado ao acaso ter tido no máximo 2 consultas no serviço público e no máximo 3 no serviço privado?
d) Construa a função de probabilidade do número de consultas no serviço privado, apenas para os utentes que tiveram 2 consultas no público.

e) Calcule a probabilidade de, ao escolher ao acaso, um utente de entre os que tiveram 3 ou mais consultas no serviço privado, este ter tido 3 consultas no serviço público.
f) Determine a probabilidade de um utente, escolhido ao acaso, ter tido 3 consultas no serviço público.
g) Qual a probabilidade de, ao escolher um utente ao acaso, este ter tido 2 consultas no serviço privado, sabendo que teve menos de 6 consultas, ou seja,

$$P(X = 2|X + Y \leq 5)?$$

h) Será que o número de consultas no serviço público é independente do número de consultas no serviço privado? Justifique.

10.Considere-se uma turma constituída por 7 alunos, cujas alturas, X, e pesos, Y, se encontram caracterizados na tabela seguinte:

	y				
x	75	76	77	79	81
1,73	1/7	0	0	0	0
1,76	2/7	0	0	0	0
1,78	0	1/7	1/7	1/7	0
1,82	0	0	0	0	1/7

a) Qual a probabilidade de a altura ser igual a 1,76 m quando o peso é igual a 75 kg?
b) Existe relação entre a altura e o peso dos alunos? Justifique.
c) Qual o peso médio dos alunos?
d) Calcule a $Var(3X + 5)$.
e) Calcule $Cov(X, Y)$.

11. No último trimestre foram editados no mercado dois livros X e Y. O número de exemplares vendidos, em cada um dos meses do trimestre, de cada uma das obras, nas grandes livrarias do país tem a seguinte distribuição de probabilidade conjunta:

	x			
y	1000	1050	1100	$f(y)$
750	p	$2p$	$3p$	$6p$
1000	$2p$	$4p$	$2p$	$8p$
1250	$3p$	$2p$	p	$6p$
$f(x)$	$6p$	$8p$	$6p$	$20p$

a) Determine o valor de p.
b) Determine o número médio mensal de livros X vendidos.
c) Calcule a $Var(X)$?
d) Qual a proporção de livrarias em que se vendeu igual número de exemplares das duas obras?
e) Qual a probabilidade de uma livraria, escolhida ao acaso, ter vendido mais exemplares do livro Y do que do X?
f) Obtenha a função de probabilidade de X, para as livrarias em que se venderam 1250 exemplares do livro Y.
g) Seja $W = 6X + 8Y$ o lucro obtido pelos hipermercados com a venda dos livros, nesse trimestre. Calcule a média desta variável.
h) Existe alguma relação entre o número de exemplares vendidos dos livros X e Y? Justifique.
i) Calcule $Cov(X, Y)$ e o coeficiente de correlação.

12. Numa empresa de aluguer de aviões informam-nos que a procura diária de aviões de passageiros, X, e a procura diária de aviões de transporte rápido de correio, Y, constitui um par aleatório (X, Y) cuja função de probabilidade conjunta é dada por:

	y			
x	0	1	2	Total
0	0			0,25
1			0,05	0,35
2	0,1		0,1	p + 0,2
3	0	0,1		p
Total	0,2	0,5		

a) Complete a tabela, determinando o valor de p.
b) Existe alguma relação entre a procura diária de aviões de passageiros e a procura diária de aviões de transporte? Justifique.
c) Determine $E(Y)$ e $Var(Y)$.

13. Considere que as v. a. X e Y, independentes, têm a seguinte função de probabilidade:

x	1	2	3
$f(x)$	0,2	0,2	0,6

y	0	1	2
$f(y)$	0,2	0,4	0,4

a) Construa a distribuição conjunta de (X, Y).
b) Sendo $Z = 3X + Y$:
   i. Construa a função de probabilidade desta variável.
   ii. Calcule $E(Z)$ e $Var(Z)$.

14.Num dado hipermercado, as vendas mensais dos artigos A e B em unidades monetárias (u.m.) constituem um v.a. bidimensional (X, Y), cuja função densidade de probabilidade é

$$f(x,y) = \begin{cases} \dfrac{x+y}{16}, & 0 < x < 2,\ 2 < y < 4 \\ 0, & \text{restantes valores} \end{cases}$$

a) Calcule o volume médio mensal de vendas (em u.m.) do artigo A?
b) Em 10% dos meses, as vendas do artigo A foram superiores a quantas u.m.?
c) Estude a independência das variáveis.

15.Num dado hospital as necessidades diárias dos medicamentos X e Y podem ser descritas pela seguinte função densidade conjunta:

$$f(x,y) = \begin{cases} cxy, & 0 < x < 4,\ 1 < y < 5 \\ 0, & \text{caso contrário} \end{cases}$$

a) Determine o valor da constante c.
b) Calcule a probabilidade de num dia ser necessária uma quantidade entre 1 e 2 do medicamento X e entre 2 e 3 do medicamento Y?.
c) Qual a necessidade diária esperada do medicamento X? E do medicamento Y?
d) Pode afirmar que as variáveis são independentes?

5 Principais distribuições de probabilidade

Neste capítulo apresentam-se algumas das principais distribuições de probabilidade teóricas, discretas e contínuas. Algumas destas distribuições foram selecionadas por caracterizarem populações existentes na realidade e, outras distribuições, devido às suas propriedades servirem de base aos desenvolvimentos de resultados que se apresentam no capítulo seguinte.

5.1 Distribuições discretas

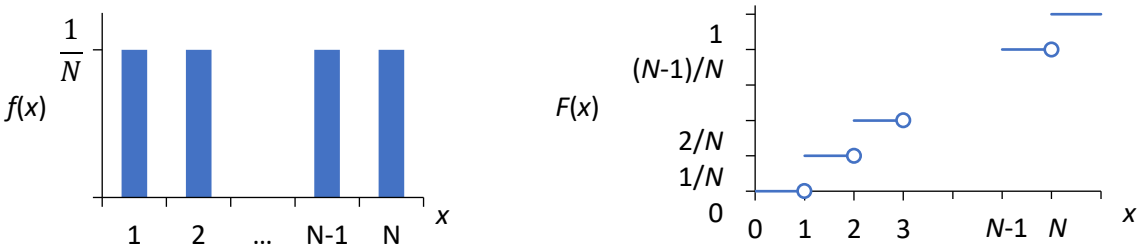
5.1.1 Distribuição Uniforme

Se os valores que a v. a. X pode assumir ocorrem com igual probabilidade então diz-se que X segue uma **distribuição Uniforme**.

Definição: A v. a. X segue uma **distribuição Uniforme discreta em N pontos**, $\boldsymbol{X \sim U\{1, 2, \dots, N\}}$, se a sua função de probabilidade é:

$$f(x) = P(X = x) = \frac{1}{N}, \qquad x = 1, 2, \dots, N.$$

O parâmetro caracterizador desta distribuição é N.



a) Função de probabilidade *b) Função de distribuição*

Figura 5.1: *Função de probabilidade e de distribuição da distribuição Uniforme discreta em N pontos.*

Para qualquer valor de N, esta distribuição tem uma forma muita característica sendo, por exemplo, sempre simétrica em torno da sua média (Figura 5.1)

Se $X \sim U\{1, 2, \dots, N\}$ então $\mu_X = E(X) = \frac{N+1}{2}$ e $\sigma_X^2 = Var(X) = \frac{N^2 - 1}{12}$.

5.1.2 Distribuição de Bernoulli e Binomial

5.1.2.1 Prova de Bernoulli

Definição: Uma experiência aleatória chama-se **prova de Bernoulli** se possuir as seguintes características:

- Tem apenas dois resultados possíveis, incompatíveis:
  - A que se designa por sucesso;
  - $\overline{A}$ designado por insucesso,
- $P(A) = p$ e $P(\overline{A}) = q = 1 - p$.

5.1.2.2 Distribuição de Bernoulli

Quando se realiza uma prova de Bernoulli e se considera X a v.a. que caracteriza a experiência tomando os valores:

- $x = 1$ quando o resultado da prova é um sucesso,
- $x = 0$ quando o resultado da prova é um insucesso,

então a **distribuição de Bernoulli** é o modelo adequado para descrever o comportamento probabilístico da v. a. X.

Definição: A v. a. discreta X, que designa o resultado da prova de Bernoulli, segue uma **distribuição Bernoulli**, i. e., $\boldsymbol{X \sim B(1; p)}$, se a sua função de probabilidade é:

$$f(x) = P(X = x) = p^x(1-p)^{1-x}, \qquad x = 0, 1 \text{ com } 0 < p < 1.$$

O parâmetro caracterizador desta distribuição é p.

Se $X \sim B(1; p)$ então $\mu_X = E(X) = p$ e $\sigma_X^2 = Var(X) = p(1-p) = pq$.

5.1.2.3 Distribuição Binomial

Quando se realizam n provas de Bernoulli independentes mantendo-se constante, de prova para prova, a probabilidade de sucesso e de insucesso, i. e., $P(A) = p$ e $P(\overline{A}) = q = 1 - p$, então a **distribuição de Binomial** é o modelo adequado para descrever o comportamento probabilístico da v. a. X que conta o número de sucessos em n provas realizadas.

Considere-se a seguinte sequência de n provas de Bernoulli independentes:

$$\overbrace{\underbrace{A\,A\ldots A}_{\substack{x \\ \text{sucessos}}}\,\underbrace{\overline{A}\,\overline{A}\ldots\overline{A}}_{\substack{n-x \\ \text{insucessos}}}}^{n \text{ provas}}.$$

A probabilidade desta sequência se verificar, dado que as provas são independentes, é:

$$P(AA\ldots A\overline{A}\,\overline{A}\ldots\overline{A}) = P(A)P(A)\ldots P(A)P(\overline{A})P(\overline{A})\ldots P(\overline{A}) = p^x(1-p)^{n-x}.$$

No entanto existem ${}^nC_x$ maneiras diferentes de se realizam x sucessos e $n - x$ insucessos sendo a probabilidade da sequência sempre a mesma. Portanto, a probabilidade de em n provas de Bernoulli obter exatamente x sucessos é dada por:

$$P(X = x) = {}^nC_x p^x(1-p)^{n-x}, \qquad x = 0, 1, \ldots, n.$$

Definição: A v. a. discreta X, que designa o número de sucessos em n provas de Bernoulli independentes, cada uma com probabilidade de sucesso p, segue uma **distribuição Binomial**, i. e., $\boldsymbol{X \sim B(n; p)}$, se a sua função de probabilidade é:

$$f(x) = P(X = x) = {}^nC_x p^x(1-p)^{n-x}, \qquad x = 0, 1, \ldots, n \text{ com } 0 < p < 1.$$

Os parâmetros caracterizadores desta distribuição são n e p.

Como se pode observar pela análise da Figura 5.2 a assimetria da distribuição Binomial depende do valor de p. Essa assimetria é atenuada com o aumento da dimensão da amostra (Figura 5.3).

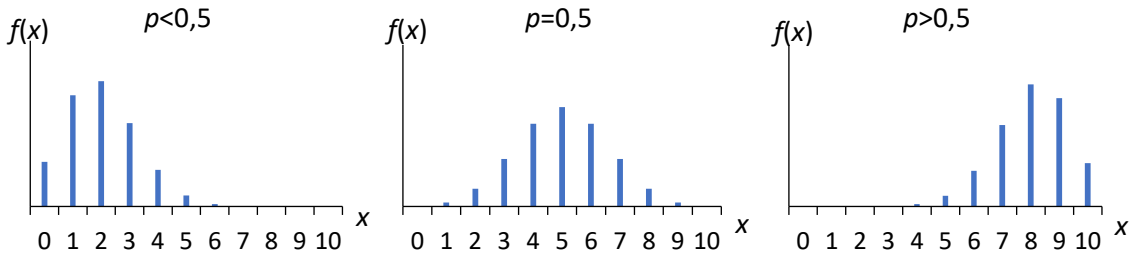


Figura 5.2: *Função de probabilidade da distribuição Binomial para diferentes valores de p e com $n = 10$.*

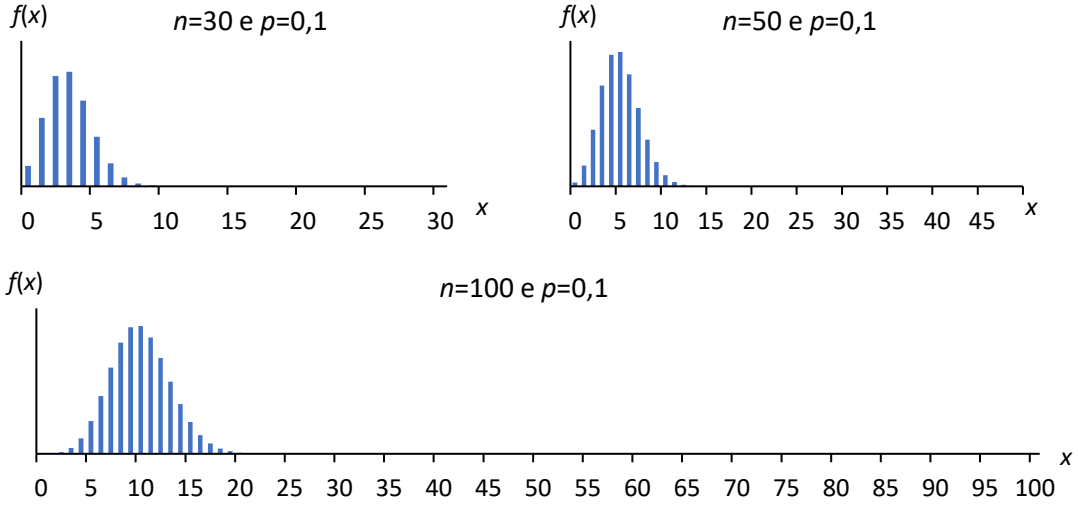


Figura 5.3: *Função de probabilidade da distribuição Binomial para diferentes valores de n e $p = 0{,}1$.*

Características relevantes:

- Quando $p = 0{,}5$ a distribuição Binomial é simétrica, qualquer que seja o n;
- Para $p > 0{,}5$, a distribuição é assimétrica negativa ou enviesada à direita;
- Para $p < 0{,}5$, a distribuição é assimétrica positiva ou enviesada à esquerda;
- Quanto mais afastado estiver p de 0,5 mais enviesada é a distribuição;
- Quanto maior for n, mais próxima da simetria estará a distribuição mesmo quando p é diferente de 0,5.

Se $X \sim B(n;p)$ então $\mu_X = E(X) = np$ e $\sigma_X^2 = Var(X) = np(1-p) = npq$.

Teorema da aditividade: Se $X_i, i = 1, 2, \ldots, K$, são v. a. independentes e $X_i \sim B(n_i; p)$ então

$$X_1 + X_2 + \cdots + X_K = \sum_{i=1}^{K} X_i \sim B\left(\sum_{i=1}^{K} n_i\,; p\right).$$

5.1.3 Distribuição Geométrica

A **distribuição Geométrica** é o modelo probabilístico adequado para descrever o comportamento probabilístico da v.a. X que conta o número de provas de Bernoulli independentes a realizar até obter a primeira prova com sucesso, mantendo-se constante, de prova para prova, a probabilidade de sucesso e insucesso.

Considere-se que foi necessário realizar n provas de Bernoulli até obter uma prova com sucesso. Uma vez que a experiência termina assim que se obtém uma prova com sucesso, em cada uma das primeiras $n-1$

provas obteve-se um insucesso. Deste modo, os resultados desta experiência são descritos pela seguinte sequência:

$$\overbrace{\underbrace{\bar{A}\ \bar{A}\ \dots\ \bar{A}}_{\substack{n-1\\ \text{insucessos}}}\ \underbrace{A}_{\substack{1\\ \text{sucesso}}}}^{n \text{ provas}}$$

Visto as provas serem independentes, a probabilidade desta sequência se verificar é:

$$P(\overline{A}\,\overline{A}\dots\overline{A}\cap A) = P(\overline{A})P(\overline{A})\dots P(\overline{A})P(A) = (1-p)^{n-1}p.$$

Portanto, a probabilidade de se terem que realizar x provas de Bernoulli independentes até obter a primeira prova com sucesso é dada por:

$$P(X=x) = (1-p)^{x-1}p, \qquad x = 1, 2, \dots$$

Definição: A v. a. discreta X, que designa o número de provas de Bernoulli independentes a realizar até obter o primeiro sucesso, sendo a p probabilidade de sucesso de cada prova, segue uma **distribuição Geométrica**, i. e., $\boldsymbol{X \sim Geom(p)}$, se a sua função de probabilidade é:

$$f(x) = P(X=x) = (1-p)^{x-1}p, \qquad x = 1, 2, \dots \text{ com } 0 < p < 1.$$

O parâmetro caracterizador desta distribuição é p.

A Figura 5.4 ilustra o comportamento genérico da função de probabilidade, sendo o decaimento mais acentuado quanto maior o valor de p.

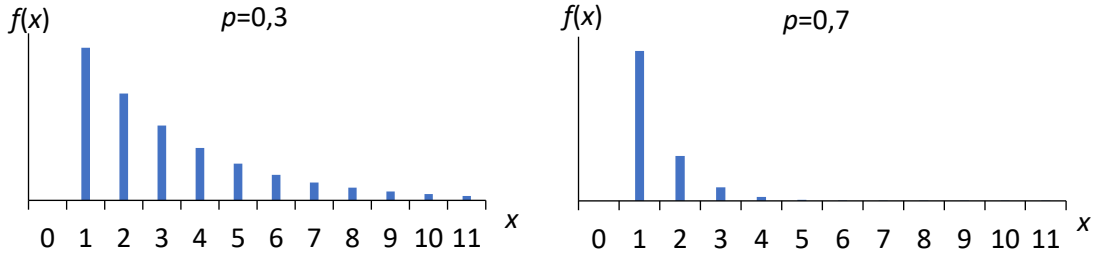


Figura 5.4: *Função de probabilidade da distribuição Geométrica para diferentes valores de p.*

Se $X \sim Geom(p)$ então $\mu_X = E(X) = \dfrac{1}{p}$ e $\sigma_X^2 = Var(X) = \dfrac{1-p}{p^2}$.

Propriedade (falta de memória): Se $X \sim Geom(p)$ então

$$P(X > x + a | X > a) = P(X > x), \qquad x > 0, a > 0.$$

Esta distribuição é também utilizada para descrever o comportamento probabilístico da v.a. $Y = X - 1$ que conta o número de provas de Bernoulli independentes realizadas com insucesso antes de obter o primeiro sucesso. Nesta situação, $Y \sim Geom(p)$ e a função de probabilidade é:

$$f(y) = P(Y=y) = (1-p)^{y}p, \qquad y = 0, 1, 2, \dots \text{com } 0 < p < 1.$$

Além disso,

$$\mu_Y = E(Y) = \frac{1-p}{p} \quad \text{e} \quad \sigma_Y^2 = Var(Y) = \frac{1-p}{p^2}.$$

5.1.4 Distribuição Binomial Negativa

A **distribuição Binomial Negativa** é o modelo probabilístico adequado para descrever o comportamento probabilístico da v.a. X que conta o número de provas de Bernoulli independentes a realizar até obter k provas com sucesso, mantendo-se constante, de prova para prova, a probabilidade de sucesso e insucesso. Considere-se que foi necessário realizar n provas de Bernoulli até obter k provas com sucesso. Como esta experiência termina quando se obtém k -ésima prova com sucesso, nas $n-1$ provas iniciais ocorreram $k-1$ provas com sucesso e $n-k$ provas com insucesso. Uma sequência possível de resultados é:

$$\overbrace{\underbrace{\bar{A}\,\bar{A}\ldots\bar{A}}_{\substack{n-k\\ \text{insucessos}}}\ \underbrace{A\,A\ldots\,A}_{\substack{k\\ \text{sucessos}}}}^{n \text{ provas}}$$

Visto as provas serem independentes, a probabilidade desta sequência se verificar é:

$$P(\overline{A}\,\overline{A}\ldots\overline{A}AA\ldots A) = P(\overline{A})P(\overline{A})\ldots P(\overline{A})P(A)P(A)\ldots P(A) = (1-p)^{n-k}p^k.$$

No entanto existem ${}^{n-1}C_{k-1}$ maneiras diferentes de se realizam $k-1$ sucessos em $n-1$ provas sendo a probabilidade da sequência sempre a mesma. Portanto, a probabilidade de se terem que realizar x provas de Bernoulli obter exatamente k sucessos é dada por:

$$P(X=x) = {}^{x-1}C_{k-1}(1-p)^{n-k}p^k, \qquad x = k, k+1, \ldots \text{ com } 0 < p < 1.$$

Definição: A v. a. discreta X, que designa o número de provas de Bernoulli independentes a realizar até obterem k sucessos, sendo a p probabilidade de sucesso de cada prova, segue uma **distribuição Binomial Negativa**, i. e., $\boldsymbol{X \sim BN(k;p)}$, se a sua função de probabilidade é:

$$f(x) = P(X=x) = {}^{x-1}C_{k-1}(1-p)^{x-k}p^k, \qquad x = k, k+1, \ldots \text{com } 0 < p < 1.$$

Os parâmetros caracterizadores desta distribuição são n e k.

A distribuição Binomial Negativa é sempre assimétrica positiva (Figura 5.5).

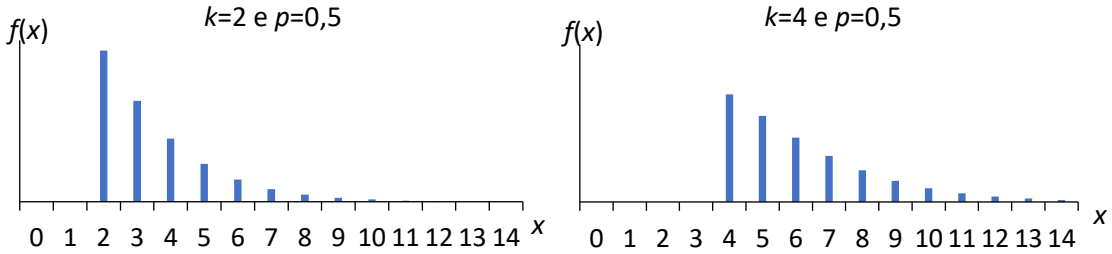


Figura 5.5: *Função de probabilidade da distribuição Binomial Negativa para diferentes valores de k.*

Se $X \sim BN(k;p)$ então $\mu_X = E(X) = \dfrac{k}{p}$ e $\sigma_X^2 = Var(X) = \dfrac{k(1-p)}{p^2}$.

Teorema da aditividade: Se $X_i, i = 1, 2, \ldots, K$, são v. a. independentes e $X_i \sim BN(k_i;p)$ então

$$X_1 + X_2 + \cdots + X_K = \sum_{i=1}^{K} X_i \sim BN\left(\sum_{i=1}^{K} k_i\,;p\right).$$

A distribuição geométrica é um caso particular da distribuição Binomial Negativa.

Se $X \sim BN(1;\ p)$ então $X \sim Geom(p)$.

5.1.5 Distribuição Multinomial

A **distribuição Multinomial** é a generalização da distribuição Binomial, em que se tem mais de dois resultados possíveis em cada experiência aleatória (prova).
Quando se realizam n experiências aleatórias independentes, existindo K resultados possíveis e incompatíveis para cada experiência, $A_i, i = 1, \dots, K$, mantendo-se constante de prova para prova a probabilidade de cada um desses resultados ocorrer,

$$P(A_i) = p_i \text{ e } P(\overline{A_i}) = 1 - p_i, \qquad i = 1, 2, \dots, K \text{ com } \sum_{i=1}^{K} P(A_i) = \sum_{i=1}^{K} p_i = 1,$$

então a distribuição Multinomial é o modelo adequado para descrever o comportamento probabilístico da v. a. $(X_1, X_2, \dots, X_K)$, onde X_i, representa o número de vezes que ocorre A_i, $i = 1, 2, \dots, K$, em n experiências realizadas.

Definição: A v. a. discreta multidimensional $(X_1, X_2, \dots, X_K)$, onde X_i o número de vezes que ocorre $A_i, i = 1, \dots, K$, em n provas independentes, segue uma **distribuição Multinomial**, i. e., $(\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_K}) \sim \boldsymbol{M(n; p_1; \dots; p_K)}$, se a sua função de probabilidade conjunta é:

$$f(x_1, x_2, \dots, x_K) = P(X_1 = x_1; X_2 = x_2; \dots; X_K = x_K) = \frac{n}{x_1!\, x_2! \dots x_K!} p_1^{x_1} p_2^{x_2} \dots p_K^{x_K},$$

para $x_i \geq 0$ e $0 < p_i < 1, i = 1, 2, \dots, K$, com $x_1 + \dots + x_K = n$ e $p_1 + \dots + p_K = 1$.

Os parâmetros caracterizadores desta distribuição são n e p_i $(i = 1, 2, \dots, K-1)$.

De notar que a K-ésima variável é definida à custa das restantes, i.e.:

$$x_K = n - x_1 - \dots - x_{K-1} \text{ e } p_K = 1 - p_1 - \dots - p_{K-1},$$

ou seja, x_K e p_K são dependentes.

Se $(X_1, X_2, \dots, X_K) \sim M(n; p_1; p_2; \dots; p_K)$ então, para $i = 1, 2, \dots, K$:

$$\mu_{X_i} = E(X_i) = np_i \text{ e } \sigma_{X_i}^2 = Var(X_i) = np_i(1 - p_i) = np_i q_i$$

$$\sigma_{X_i X_j} = Cov(X_i, X_j) = -np_i p_j, i \neq j.$$

5.1.6 Distribuição Hipergeométrica

A **distribuição Hipergeométrica** é o modelo probabilístico adequado para descrever os processos em que se colhem *sem reposição* amostras de n elementos de uma população com N elementos, dos quais Np possuem determinado atributo (*sucesso*) e $Nq = N(1 - p)$ não possuem esse atributo (*insucesso*).
Existem ${}^{N}C_n$ maneiras diferentes de recolher, sem reposição, uma amostra de n elementos de uma população com N elementos. Existem ${}^{Np}C_x$ maneiras diferentes de recolher x sucessos num total de Np sucessos e ${}^{Nq}C_{n-x}$ maneiras diferentes de obter $n - x$ insucessos em Nq insucessos. Portanto, a

probabilidade de se registarem x sucessos em n extracções sem reposição, e consequentemente $n-x$ insucessos, é dada por:

$$P(X=x)=\frac{{}^{Np}C_x\ {}^{Nq}C_{n-x}}{{}^{N}C_n},\qquad x=0,1,\ldots,n.$$

Definição: A v. a. discreta X, que designa o número de sucessos ocorridos em n extracções sem reposição, de uma população com N elementos com Np sucessos, segue uma **distribuição Hipergeométrica**, i. e., $\boldsymbol{X\sim H(N;n;p)}$, se a sua função de probabilidade é:

$$f(x)=P(X=x)=\frac{{}^{Np}C_x\ {}^{Nq}C_{n-x}}{{}^{N}C_n},\qquad x=0,1,\ldots,n \text{ com } 0<p<1 \text{ e } p+q=1.$$

Os parâmetros caracterizadores desta distribuição são N, n e p.

Se $X\sim H(N;n;p)$ então $\mu_X=E(X)=np$ e $\sigma_X^2=Var(X)=np(1-p)\dfrac{N-n}{N-1}$.

Aproximação da distribuição Hipergeométrica à Binomial:

Quando N é grande comparado com n, a diferença entre as distribuições Hipergeométrica e Binomial é esbatida. Desta forma, quando $n<0{,}1N$ aplica-se a distribuição Binomial por facilidade de cálculo pois esta oferece uma boa aproximação da distribuição Hipergeométrica:

Se $X\sim H(N;n;p)$ e $n<0{,}1N$ então $X\overset{\circ}{\sim}B(n;p)$.

5.1.7 Distribuição Poisson

A **distribuição Poisson** é o modelo probabilístico adequado para descrever os fenómenos em que os acontecimentos se repetem no tempo (ou no espaço).

Para que a v. a. X, que designa o número de ocorrências num determinado intervalo de tempo, siga uma distribuição de Poisson tem que verificar as seguintes condições (Murteira *et al.*, 2007):

- O número de ocorrências em intervalos disjuntos (não sobrepostos) são independentes entre si;
- A probabilidade de se registar uma ocorrência num qualquer intervalo de amplitude muito pequena é aproximadamente proporcional à dimensão do intervalo;
- A probabilidade de um certo número de ocorrências se verificar é a mesma para intervalos com a mesma amplitude, i. e., esta probabilidade depende apenas da amplitude do intervalo e não da posição em que se situa esse intervalo;
- A probabilidade de se verificarem duas ou mais ocorrências num período muito pequeno é aproximadamente igual a zero.

Alguns exemplos de fenómenos que se adequam a uma distribuição de Poisson:

- Número de chamadas telefónicas que chegam, em certo período de tempo, a uma central telefónica;
- Número de avarias que ocorrem numa máquina, num certo intervalo de tempo;
- Número de doentes que chegam a determinado hospital, por unidade de tempo.

Definição: A v. a. discreta X, que designa o número de ocorrências num determinado intervalo de tempo, segue uma **distribuição Poisson**, i.e., $\boldsymbol{X \sim P(\lambda)}$, se a sua função de probabilidade é:

$$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots \text{com } \lambda > 0.$$

O parâmetro caracterizador desta distribuição é λ.

A assimetria da distribuição depende do valor de λ. Para valores grandes de λ a distribuição tende a ser simétrica, enquanto que para valores pequenos de λ esta distribuição é enviesada à esquerda, i. e., é assimétrica positiva (Figura 5.6).

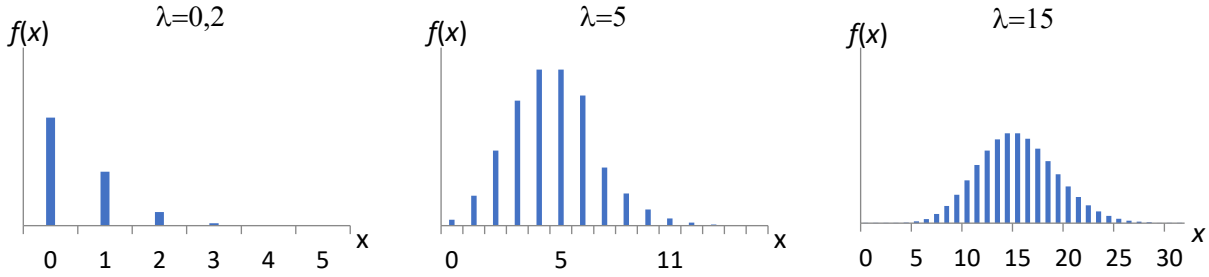


Figura 5.6: *Função de probabilidade da distribuição Poisson para diferentes valores de λ.*

Se $X \sim P(\lambda)$ então $\mu_X = E(X) = \lambda$ e $\sigma_X^2 = Var(X) = \lambda$.

Teorema da aditividade: Se $X_i, i = 1, 2, \ldots, K$, são v. a. independentes e $X_i \sim P(\lambda_i)$ então

$$X_1 + X_2 + \cdots + X_K = \sum_{i=1}^{K} X_i \sim P\left(\sum_{i=1}^{K} \lambda_i\right).$$

Aproximação da distribuição Binomial à Poisson:

A distribuição Binomial converge para a distribuição Poisson, quando $n \to \infty$ e $p \to 0$, mantendo-se constante $\lambda = np$.

Teorema: Se $X \sim B(n; p)$ com $n \to +\infty$ e $p \to 0$ então $X \overset{\circ}{\sim} P(np)$.

Observação: Na prática utiliza-se $n > 20$ e $p \leq 0{,}05$.

A aproximação é tanto melhor quanto maior o valor de n e menor o valor de p (Figura 5.7).

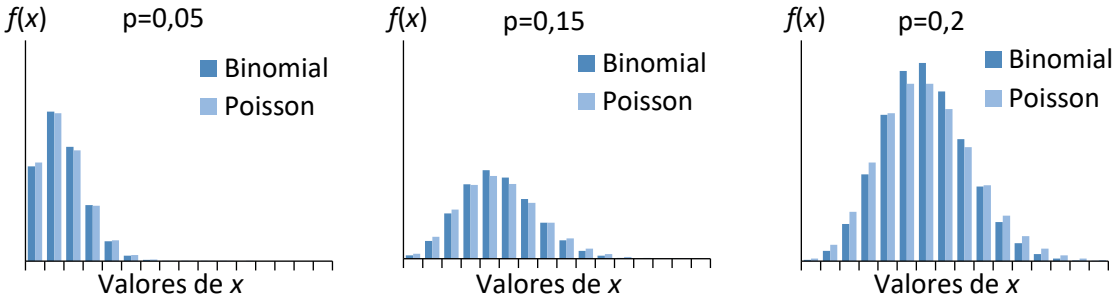


Figura 5.7: *Aproximação da distribuição $B(n = 30; p)$ pela Poisson para diferentes valores de p.*

5.1.8 Exercícios resolvidos

5.1.8.1 Distribuição Uniforme discreta

Considere-se a experiência aleatória que consiste no lançamento de um dado.
Seja X a v. a. que representa o valor da face voltada para cima.

a) Descreva a função de probabilidade da v. a. X.
b) Determine o valor esperado e a variância de X.
c) Qual a probabilidade de sair um número par no dado?
d) Determine a probabilidade de sair um número superior a 3 no lançamento.
e) Num lançamento, qual a probabilidade de sair um número inferior a 2?

Resolução:

a) Função de probabilidade:

$$f(x) = P(X = x) = \frac{1}{6}, \qquad x = 1, 2, 3, 4, 5, 6.$$

Portanto, $X \sim U\{1, 2, \dots, 6\}$.

b) $E(X) = \frac{6+1}{2} = 3{,}5.$

$Var(X) = \frac{6^2 - 1}{12} = 2{,}9167.$

c) $P(X = 2 \cup X = 4 \cup X = 6) = P(X = 2) + P(X = 4) + P(X = 6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$

d) $P(X > 3) = P(X = 4) + P(X = 5) + P(X = 6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$

e) $P(X \leq 2) = P(X = 1) + P(X = 2) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$

5.1.8.2 Distribuição Binomial

Com a ingestão de um determinado fármaco para o tratamento da depressão, 20% dos doentes referem sentir efeitos secundários durante a 1ª semana de tratamento. Recolheu-se uma amostra aleatória de 15 doentes.
Seja X a v. a. que representa o número de doentes que sentem os referidos efeitos secundários numa amostra de 15 doentes.

a) Identifique a distribuição da v. a. X.
b) Em média quantos doentes sentiram efeitos secundários?
c) Calcule a variância da v. a. X.
d) Qual a probabilidade de exatamente 5 doentes sentirem efeitos secundários?
e) Calcule a probabilidade de 2 doentes sentirem efeitos secundários.
f) Determine a probabilidade de pelo menos 1 e menos de 4 doentes sentirem efeitos secundários.
g) Recolheu-se outra amostra aleatória de 10 doentes independente da primeira. Qual a probabilidade de no conjunto das duas amostras 5 doentes sentirem efeitos secundários?

Resolução:

a) Os $n = 15$ doentes são independentes.
Só existem 2 resultados possíveis por tratamento: sente efeitos secundários (E) ou não sente efeitos secundários $(\overline{E})$, sendo

$P(E) = 0{,}2 = p$ e $P(\overline{E}) = 1 - 0{,}2 = 0{,}8 = 1 - p = q$.
Logo, $X \sim B(n = 15; p = 0{,}2)$.
Função de probabilidade:

$$f(x) = P(X = x) = {}^{15}C_x 0{,}2^x 0{,}8^{15-x}, \qquad x = 0, 1, \dots, 15.$$

b) $E(X) = 15 \times 0{,}2 = 3$, ou seja, em média 3 sentem efeitos secundários

c) $Var(X) = 15 \times 0{,}2 \times 0{,}8 = 2{,}4$.

d) $P(X = 5) = {}^{15}C_5 0{,}2^5 0{,}8^{15-5} = 0{,}1032$.

e) $P(X > 2) = P(X = 3) + P(X = 4) + \cdots + P(X = 15)$, ou

$$= 1 - P(X \leq 2) = 1 - \big(P(X = 0) + P(X = 1) + P(X = 2)\big)$$
$$= 1 - \left({}^{15}C_0 0{,}2^0 0{,}8^{15-0} + {}^{15}C_1 0{,}2^1 0{,}8^{15-1} + {}^{15}C_2 0{,}2^2 0{,}8^{15-2}\right) = 0{,}602.$$

f) $P(1 \leq X < 4) = P(X = 1) + P(X = 2) + P(X = 3)$

$$= {}^{15}C_1 0{,}2^1 0{,}8^{15-1} + {}^{15}C_2 0{,}2^2 0{,}8^{15-2} + {}^{15}C_3 0{,}2^3 0{,}8^{15-3} = 0{,}6130.$$

g) Seja Y a v.a. que representa o número de doentes que sentem os referidos efeitos secundários numa amostra de 10 doentes.

$$Y \sim B(n = 10; p = 0{,}2).$$

Seja $W = X + Y$ a v.a. que representa o número de doentes que sentem os referidos efeitos secundários numa amostra de 15 + 10 (= 25) doentes.
Pelo teorema da aditividade da distribuição Binomial, $W \sim B(n = 25; p = 0{,}2)$.
Portanto, $P(W = 5) = {}^{25}C_5 0{,}2^5 0{,}8^{25-5} = 0{,}1960$.

5.1.8.3 Distribuição Geométrica

Com a ingestão de um determinado fármaco para o tratamento da depressão, 20% dos doentes referem sentir efeitos secundários durante a 1ª semana de tratamento.

a) Quantos doentes espera ter que inquirir até encontrar um que refira ter sentido os efeitos secundários?
b) Qual a probabilidade de ter que inquirir 5 doentes até encontrar um que tenha sentido os efeitos secundários?
c) Determine a probabilidade de ter que inquirir mais de 3 doentes até encontrar um que tenha sentido os efeitos secundários.
d) Sabendo que vão inquirir mais de 4 doentes que não tiveram efeitos secundários, qual a probabilidade de no total se terem que inquirir mais de 7 doentes até encontrar um que não sentiu efeitos secundários?

Resolução:

Seja X a v. a. que representa o número de doentes a inquirir até encontrar um que refira ter sentido os efeitos secundários.
Só existem 2 resultados possíveis por tratamento: sente efeitos secundários (E) ou não sente efeitos secundários $(\overline{E})$, sendo
$P(E) = 0{,}2 = p$ e $P(\overline{E}) = 1 - 0{,}2 = 0{,}8 = 1 - p = q$.
Logo, $X \sim Geom(p = 0{,}2)$.
Função de probabilidade:

$$f(x) = P(X = x) = 0{,}8^{15-x} 0{,}2, \qquad x = 1, 2, \dots$$

a) $E(X) = \frac{1}{p} = \frac{1}{0{,}2} = 5$ (número médio de doentes a inquirir).

b) $P(X = 5) = (1 - 0{,}2)^{5-1} \times 0{,}2 = 0{,}0819.$

c) $P(X > 3) = 1 - P(X \leq 3) = 1 - \big(P(X = 1) + P(X = 2) + P(X = 3)\big) =$
$= 1 - ((1 - 0{,}2)^{1-1} \times 0{,}2 + (1 - 0{,}2)^{2-1} \times 0{,}2 + (1 - 0{,}2)^{3-1} \times 0{,}2) = 0{,}512.$

b) $P(X > 7|X > 4)$
Recorrendo à propriedade da falta de memória da distribuição Geométrica,
$P(X > 7|X > 4) = P(X > 7 - 4) = P(X > 3) = 0{,}512.$

5.1.8.4 Distribuição Binomial Negativa

Com a ingestão de um determinado fármaco para o tratamento da depressão, 20% dos doentes referem sentir efeitos secundários durante a 1ª semana de tratamento.

a) Quantos doentes espera ter que inquirir até encontrar 3 que refiram ter sentido os efeitos secundários?
b) Qual a probabilidade de ter que inquirir 5 doentes até encontrar 3 que tenha sentido os efeitos secundários?

Resolução:

Seja X a v. a. que representa o número de doentes a inquirir até encontrar 3 que refiram ter sentido os efeitos secundários.
Só existem 2 resultados possíveis por tratamento: sente efeitos secundários (E) ou não sente efeitos secundários $\left(\overline{E}\right)$, sendo

$$P(E) = 0{,}2 = p \quad \text{e} \quad P\left(\overline{E}\right) = 1 - 0{,}2 = 0{,}8 = 1 - p = q.$$

Logo, $X \sim BN(k = 3; p = 0{,}2)$.
Função de probabilidade:

$$f(x) = P(X = x) = {}^{x-1}C_2 0{,}8^{x-3} 0{,}2^k, \qquad x = 3, 4, \ldots$$

a) $E(X) = \frac{k}{p} = \frac{3}{0{,}2} = 15$ (número médio de doentes a inquirir).

b) $P(X = 5) = {}^{5-1}C_{3-1}(1 - 0{,}2)^{5-3} 0{,}2^3 = 0{,}0307.$

5.1.8.5 Distribuição Multinomial

Num serviço hospitalar existem 100 de comprimidos de igual aparência dos quais 40 são analgésicos, 35 são para o controlo da hipertensão e 25 para o controlo da diabetes.
Considere a experiência aleatória que consiste em retirar, com reposição, uma amostra de 15 comprimidos.

a) Identifique a distribuição da v. a. em estudo.
b) Em média quantos comprimidos analgésicos espera ter na amostra? E comprimidos para o controlo da hipertensão? E para o controlo da diabetes?
c) Determine as variâncias.
d) Qual a probabilidade de ter exatamente 5 comprimidos analgésicos e 3 para controlo da hipertensão na amostra?
e) Calcule a probabilidade de retirar no mínimo 9 comprimidos analgésicos e exatamente 5 comprimidos para controlo da diabetes.
f) Qual a probabilidade de a amostra conter pelo menos 14 comprimidos para controlo da hipertensão?

Resolução:

a) Sejam:
- X_1 a v. a. que representa o número de comprimidos analgésicos nos 15 retirados;
- X_2 a v. a. que representa o número de comprimidos para controlo da hipertensão nos 15 retirados;
- X_3 a v. a. que representa o número de comprimidos para controlo da diabetes nos 15 retirados.

Neste caso tem-se:

$$N = 100;\ p_1 = \frac{40}{100} = 0{,}4;\ p_2 = \frac{35}{100} = 0{,}35;\ p_3 = \frac{25}{100} = 0{,}25;\ n = 15.$$

Logo, $(X_1, X_2, X_3) \sim M(n = 15;\ p_1 = 0{,}4;\ p_2 = 0{,}35;\ p_3 = 0{,}25)$.

Função de probabilidade:

$$f(x_1, x_2, x_3) = P(X_1 = x_1; X_2 = x_2; X_3 = x_3) = \frac{15!}{x_1!\, x_2!\, x_3!} 0{,}40^{x_1} 0{,}35^{x_2} 0{,}25^{x_3},$$

para $x_1 + x_2 + x_3 = 15$ e $x_1, x_2, x_3 = 0, 1, \ldots, 15$.

b) $E(X_1) = 15 \times 0{,}4 = 6$ (n.º médio de comprimidos analgésicos).
$E(X_2) = 15 \times 0{,}35 = 5{,}25$ (n.º médio de comprimidos para a hipertensão).
$E(X_3) = 15 \times 0{,}25 = 3{,}75$ (n.º médio de comprimidos para a diabetes).

c) $Var(X_1) = 15 \times 0{,}4 \times 0{,}6 = 3{,}6$.
$Var(X_2) = 15 \times 0{,}35 \times 0{,}65 = 3{,}4125$.
$Var(X_3) = 15 \times 0{,}25 \times 0{,}75 = 2{,}8125$.

d) $P(X_1 = 5; X_2 = 3) = P(X_1 = 5; X_2 = 3; X_3 = 7) = \frac{15!}{5!3!7!} 0{,}4^5 \times 0{,}35^3 \times 0{,}25^7 = 0{,}0097$.

e) $P(X_1 \geq 9; X_3 = 5) = P(X_1 = 9; X_2 = 1; X_3 = 5) + P(X_1 = 10; X_2 = 0; X_3 = 5)$
$$= \frac{15!}{9!\,1!\,5!} 0{,}4^9 \times 0{,}35^1 \times 0{,}25^5 + \frac{15!}{10!\,0!\,5!} 0{,}4^{10} \times 0{,}35^0 \times 0{,}25^5 = 0{,}003.$$

f) $P(X_2 \geq 14) = P(X_1 = 1; X_2 = 14; X_3 = 0) + P(X_1 = 0; X_2 = 14; X_3 = 1)$
$$+P(X_1 = 0; X_2 = 15; X_3 = 0)$$
$$= \frac{15!}{1!\,14!} 0{,}4^1 \times 0{,}35^{14} + \frac{15!}{14!\,1!} \times 0{,}35^{14} \times 0{,}25^1 + \frac{15!}{15!} \times 0{,}35^{15} \approx 0.$$

5.1.8.6 Distribuição Hipergeométrica

Suponha que, de 100 candidatos a um emprego numa empresa de telecomunicações, apenas 45 têm as qualificações pretendidas. Considere que foram selecionados ao acaso, e sem reposição, 20 candidatos para uma entrevista piloto.

a) Identifique a v. a. em estudo e respetiva distribuição.
b) Em média quantos dos candidatos selecionados terão as qualificações pretendidas? Determine a variância.
c) Calcule a probabilidade de, no grupo selecionado:
   i. Exatamente 5 terem as qualificações pretendidas.
   ii. No mínimo 2 terem as qualificações necessárias.
   iii. Menos de 3 candidatos terem as qualificações exigidas.

Resolução:

a) Seja X a v. a. que representa o número de candidatos com as qualificações pretendidas, em 20 seleccionados ao acaso sem reposição.

Neste caso tem-se $N = 100$ candidatos, $Np = 45$ candidatos com as qualificações pretendidas (ou seja, $p = 45/100 = 0{,}45$), $Nq = 55$ candidatos sem as qualificações pretendidas e $n = 20$ candidatos selecionados.
Logo, $X \sim H(N = 100; n = 20; p = 0{,}45)$.
Função de probabilidade:

$$f(x) = P(X = x) = \frac{{}^{45}C_x \, {}^{55}C_{20-x}}{{}^{100}C_{20}}, \qquad x = 0{,}1, \ldots, 20.$$

b) $E(X) = np = 20 \times \frac{45}{100} = 9$ (número médio de candidatos com qualificações).
$Var(X) = np(1-p)\frac{N-n}{N-1} = 20 \times \frac{45}{100} \times \frac{55}{100} \times \frac{100-20}{100-1} = 4.$

c) i. $P(X = 5) = \frac{{}^{45}C_5 \, {}^{55}C_{15}}{{}^{100}C_{20}} = 0{,}0271.$

ii. $P(X \geq 2) = P(X = 2) + P(X = 3) + \cdots + P(X = 20)$, ou

$$= 1 - \big(P(X = 0) + P(X = 1)\big) = 1 - \left(\frac{{}^{45}C_0 \, {}^{55}C_{20}}{{}^{100}C_{20}} + \frac{{}^{45}C_1 \, {}^{55}C_{19}}{{}^{100}C_{20}}\right) \approx 1.$$

iii. $P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2) = \frac{{}^{45}C_0 \, {}^{55}C_{20}}{{}^{100}C_{20}} + \frac{{}^{45}C_1 \, {}^{55}C_{19}}{{}^{100}C_{20}} + \frac{{}^{45}C_2 \, {}^{55}C_{18}}{{}^{100}C_{20}} = 0{,}0003.$

5.1.8.7 Distribuição Poisson

Seja X a v. a. que representa o número de automóveis que entram numa autoestrada (AE) num período de 30 segundos. Sabe-se que X é uma v. a. de Poisson com desvio padrão a 3.

a) Descreva a função de probabilidade da v. a. em estudo.
b) Em média quantos automóveis entram na AE num período de 30 segundos? Calcule a variância.
c) Qual a probabilidade de entrarem no mínimo 2 automóveis na AE num período de 30 segundos?
d) Determine a probabilidade de entrarem no máximo 3 automóveis na AE num minuto.
e) Calcule a probabilidade de entrarem mais de 2 e menos de 5 automóveis na AE num período de 15 segundos.

Resolução:

a) $X \sim P(\lambda = 3^2 = 9)$.
Função de probabilidade:

$$f(x) = P(X = x) = \frac{e^{-9}9^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

b) $E(X) = \lambda = 9$ (n.º médio de automóveis que entram na AE por cada 30 segundos)
$Var(X) = \lambda = 9.$

c) $P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + \cdots$ ou

$$= 1 - \big(P(X = 0) + P(X = 1)\big) = 1 - \left(\frac{e^{-9}9^0}{0!} + \frac{e^{-9}9^1}{1!}\right) = 0{,}9998.$$

d) Seja X' a v.a. que representa o n.º de automóveis que entram na AE num período de 1 minuto.
Logo, $X' \sim P(\lambda' = 2\lambda = 18)$ pois 1 minuto = 2 × 30 segundos
$P(X' \leq 3) = P(X' = 0) + P(X' = 1) + P(X' = 2) + P(X' = 3)$

$$= \frac{e^{-18}18}{0!} + \frac{e^{-18}18^1}{1!} + \frac{e^{-18}18^2}{2!} + \frac{e^{-18}18^3}{3!} \approx 0.$$

e) Seja X'' a v.a. que representa o n.º de automóveis que entram na AE num período de 15 segundos
Logo, $X'' \sim P\left(\lambda'' = \frac{\lambda}{2} = 4{,}5\right)$ pois 15 segundos = 30 segundos ÷ 2.

$$P(2 < X'' < 5) = P(X'' = 3) + P(X'' = 4) = \frac{e^{-4{,}5}4{,}5^3}{3!} + \frac{e^{-4{,}5}4{,}5^4}{4!} = 0{,}3585.$$

5.2 Distribuições contínuas

5.2.1 Distribuição Uniforme

Se os valores da v. a. X podem ocorrer dentro dum certo intervalo limitado (a, b), e se quaisquer dois sub-intervalos de igual amplitude têm a mesma probabilidade, então diz-se que X segue uma distribuição uniforme.

Definição: A v. a. contínua X segue uma **distribuição Uniforme no intervalo** $(\boldsymbol{a}, \boldsymbol{b})$, i.e. $\boldsymbol{X \sim U(a;\ b)}$, se a sua função densidade de probabilidade é:

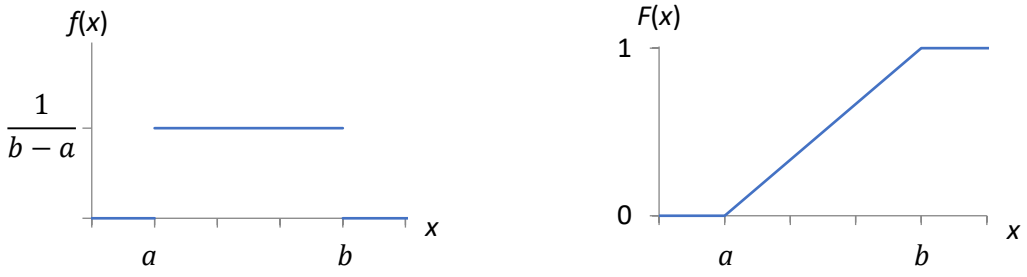
$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{a} < x < b \\ 0, & \text{outros valores} \end{cases}, \text{ com } -\infty < a < b < +\infty.$$

Os parâmetros caracterizadores desta distribuição são a e b.

A função de distribuição é dada por:

$$F(X) = P(X \leq x) = \begin{cases} 0, & x \leq a, \\ \frac{x-a}{b-a}, & a < x < b \\ 1, & x \geq b \end{cases}$$

Para qualquer valor de a e b, a distribuição uniforme contínua é sempre simétrica em torno da sua média (Figura 5.8).



a) Função densidade de probabilidade *b) Função de distribuição*

Figura 5.8: *Função densidade de probabilidade e de distribuição da distribuição Uniforme*$(a;\ b)$.

Se $X \sim U(a;\ b)$ então $\mu_X = E(X) = \frac{a+b}{2}$ e $\sigma_X^2 = Var(X) = \frac{(b-a)^2}{12}$.

5.2.2 Distribuição Normal

A **distribuição Normal**, ou **Gaussiana**, é uma das mais importantes, senão a mais importante distribuição contínua, sendo bastante utilizada tanto para aplicações práticas como para estudos teóricos. Muitas características da população existentes na realidade são bem representadas por esta distribuição que também goza de importantes propriedades.

Definição: A v. a. contínua X segue uma **distribuição Normal com média μ e desvio padrão σ**, i. e., $X \sim N(\mu;\ \sigma)$, se a sua função densidade de probabilidade é:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < +\infty \text{ com } -\infty < \mu < +\infty \text{ e } 0 < \sigma < +\infty.$$

Os parâmetros caracterizadores da desta distribuição são μ e σ.

Principais características:

- A função densidade de probabilidade de uma v. a. com a distribuição Normal tem a forma de sino, é simétrica em torno de μ e tem pontos de inflexão em $x = \mu \pm \sigma$.
- A média μ localiza o centro da distribuição e σ mede a variabilidade de X em torno de μ (Figura 5.9 e Figura 5.10).

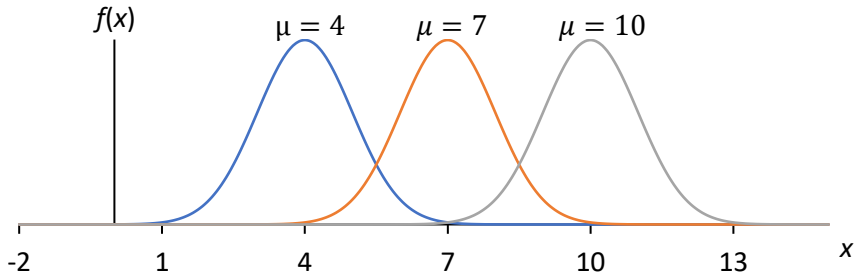


Figura 5.9: *Função densidade de probabilidade da distribuição Normal para diferentes valores de μ e σ = 1.*

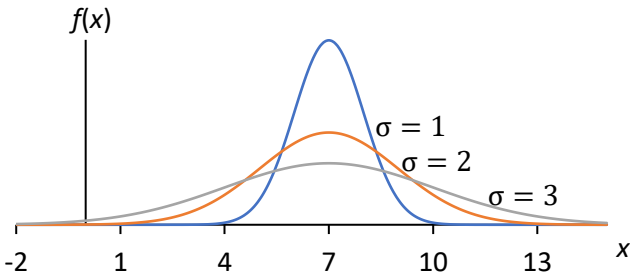


Figura 5.10: *Função densidade de probabilidade da distribuição Normal para diferentes valores de σ e μ = 7.*

Se $X \sim N(\mu;\ \sigma)$ então $\mu_X = E(X) = \mu$ e $\sigma_X^2 = Var(X) = \sigma$.

Resultados importantes:

- Se $X \sim N(\mu;\ \sigma)$ então a v. a. estandardizada $Z = \frac{X-\mu}{\sigma} \sim N(0;1)$.
- A função de distribuição $F(z)$ da v. a. $Z \sim N(0;1)$ é representada por $\Phi(z)$ e está tabulada (Anexo A).
- $\Phi(-z) = 1 - \Phi(z)$.
- $\Phi(z) = \alpha \Rightarrow z = \Phi^{-1}(\alpha)$.

Usualmente utiliza-se a notação z_α para representar o quantil de probabilidade α de uma v. a. $X \sim N(0;1)$, i.e., $P(Z \leq z_\alpha) = \alpha$ o que é equivalente a $z_\alpha = \Phi^{-1}(\alpha)$.

Teorema da aditividade: Se $X_i, i = 1, 2, \ldots, K$, são v. a. independentes e $X_i \sim N(\mu_i;\ \sigma_i)$ então

$$\sum_{i=1}^{K} a_i X_i \sim N\left(\sum_{i=1}^{K} a_i \mu_i\,;\ \sqrt{\sum_{i=1}^{K} a_i^2 \sigma_i^2}\right).$$

Corolário: Se $X_i, i = 1, 2, \ldots, n$, são v. a. independentes e $X_i \sim N(\mu;\ \sigma)$, então:

$$S_n = \sum_{i=1}^{n} X_i \sim N\left(n\mu; \sigma\sqrt{n}\right);$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right).$$

5.2.3 Distribuição Exponencial

A **distribuição Exponencial** está relacionada com o processo de **Poisson**. Demonstra-se que num processo de Poisson, o tempo de espera até à ocorrência do primeiro sucesso é uma v. a. que segue uma distribuição Exponencial. Esta distribuição é também adequada para descrever o tempo até à ocorrência do próximo sucesso ou o tempo entre dois sucessos consecutivos.

Definição: A v. a. contínua X segue uma **distribuição Exponencial**, i. e., $\boldsymbol{X \sim Exp(\lambda)}$, se a sua função densidade de probabilidade é:

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0 \text{ com } \lambda > 0.$$

O parâmetro caracterizador desta distribuição é λ.

A função densidade de probabilidade da Exponencial encontra-se representada na Figura 5.11.

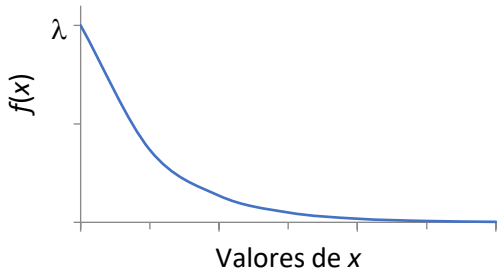


Figura 5.11: *Função densidade de probabilidade da distribuição Exponencial.*

A função de distribuição é dada por:

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x}, \quad x \geq 0 \text{ com } \lambda > 0.$$

Se $X \sim Exp(\lambda)$, então $\mu_X = E(X) = \frac{1}{\lambda}$ e $\sigma_X^2 = Var(X) = \frac{1}{\lambda^2}$.

Propriedade (falta de memória): Se $X \sim Exp(\lambda)$ então

$$P(X > x + a | X > a) = P(X > x), \qquad x > 0, \qquad a > 0.$$

5.2.4 Distribuição Qui-quadrado

Definição: A v. a. contínua X segue uma **distribuição Qui-Quadrado com n graus de liberdade**, i. e., $X \sim \chi^2_n$, se a sua função densidade de probabilidade é:

$$f(x) = \frac{1}{2^{\frac{n}{2}}\,\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \; x > 0, \; n > 0,$$

onde $\Gamma(\alpha)$ é a função Gama, definida por $\Gamma(\alpha) = \int_0^{+\infty} e^{-x} x^{\alpha-1} dx$.

O parâmetro caracterizador desta distribuição é n.

Principais características:

- A v. a. só toma valores positivos;
- É uma função não simétrica.

A forma da distribuição depende dos graus de liberdade (Figura 5.12), tornando-se menos assimétrica com o aumento do número de graus de liberdade. Esta distribuição está tabelada (Anexo B).

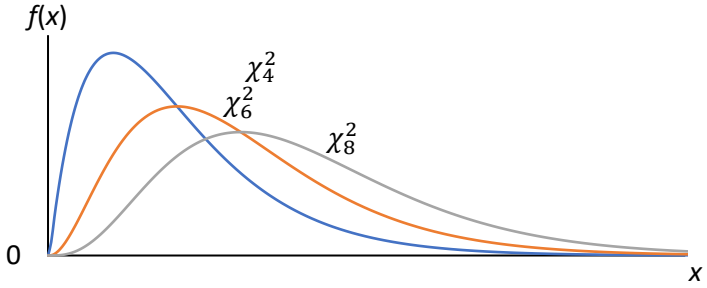


Figura 5.12: *Função densidade de probabilidade distribuição Qui-Quadrado para diferentes graus de liberdade.*

Se $X \sim \chi^2_n$, então $\mu_X = E(X) = n$ e $\sigma^2_X = Var(X) = 2n$.

Usualmente utiliza-se a notação $\chi^2_{n;\,\alpha}$ para representar o quantil de probabilidade α de uma v. a. $X \sim \chi^2_n$. Portanto, $\chi^2_{n;\,\alpha}$ corresponde ao menor valor k tal que $P(X \leq k) = \alpha$.

Teorema: Se $X \sim N(\mu;\ \sigma)$, então

$$Z^2 = \left(\frac{X - \mu}{\sigma}\right)^2 \sim \chi^2_1.$$

Corolário: Se $X_i, i = 1, 2, \ldots, K$, são v. a. independentes e $X_i \sim N(\mu;\ \sigma)$, então

$$\sum_{i=1}^{K} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_K.$$

Teorema da aditividade: Se $X_i, i = 1, 2, \dots, K$, são v. a. independentes e $X_i \sim \chi^2_{K_i}$, então

$$\sum_{i=1}^{K} X_i \sim \chi^2_m, \text{ com } m = \sum_{i=1}^{K} K_i.$$

5.2.5 Distribuição Gama

As **distribuições Exponencial** e **Qui-quadrado** são casos particulares de uma distribuição mais geral, a **distribuição Gama**.

Definição: A v. a. contínua X segue uma distribuição Gama com parâmetros $\boldsymbol{\alpha}$ e $\boldsymbol{\lambda}$, i. e., $\boldsymbol{X \sim G(\alpha;\ \lambda)}$, se a sua função densidade de probabilidade é:

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x},\ x > 0,\ n > 0,$$

onde $\Gamma(\alpha)$ é a função Gama (definida na secção 5.2.4).

O parâmetros caracterizadores desta distribuição são α e λ.

Casos particulares:

- $X \sim \chi^2_n \iff X \sim G\left(\alpha = \frac{n}{2};\ \lambda = \frac{1}{2}\right)$;
- $X \sim Exp(\lambda) \iff X \sim G(\alpha = 1;\ \lambda)$.

O aspeto da distribuição depende do valor dos parâmetros (Figura 5.13).

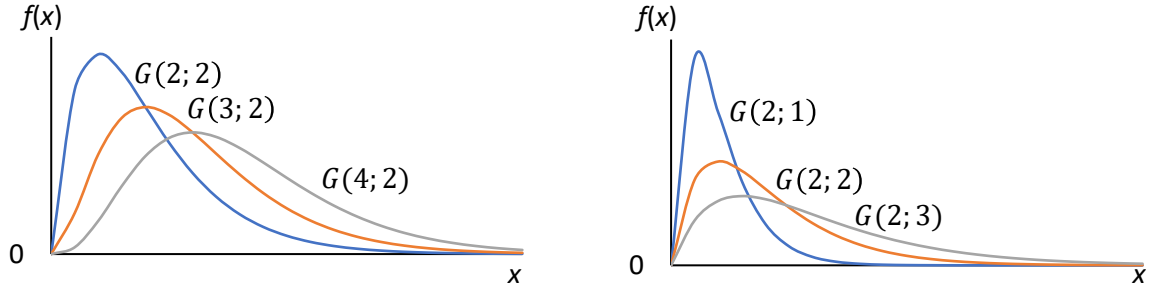


Figura 5.13: *Função densidade de probabilidade da distribuição Gama para diferentes valores de α e λ.*

Se $X \sim G(\alpha;\ \lambda)$, então $\mu_X = E(X) = \frac{\alpha}{\lambda}$ e $\sigma^2_X = Var(X) = \frac{\alpha}{\lambda^2}$.

Teorema da aditividade: Se $X_i, i = 1, 2, \dots, K$, são v. a. independentes e $X_i \sim G(\alpha_i;\ \lambda)$, então

$$\sum_{i=1}^{K} X_i \sim G(\alpha;\ \lambda), \text{ com } \alpha = \sum_{i=1}^{K} \alpha_i.$$

A distribuição Gama pode ser como uma generalização da distribuição Exponencial para descrever a v.a. X que representa **o tempo de espera até à ocorrência do n-ésimo sucesso**. A variável X resulta da soma dos

tempos de espera entre as várias ocorrências sucessivas (X_i) até à ocorrência pretendida. Deste modo, pelo teorema da aditividade como $X_i, i = 1, \dots, n$, são v. a. independentes e $X_i \sim Exp(\lambda) \Leftrightarrow X_i \sim G(1;\ \lambda)$, então

$$X = \sum_{i=1}^{n} X_i \sim G(n;\ \lambda).$$

5.2.6 Distribuição *t*-Student

Definição: A v. a. contínua X segue uma **distribuição *t*-Student** com $\boldsymbol{n}$ graus de liberdade, i.e., $\boldsymbol{X \sim t_n}$, se a sua função densidade de probabilidade é:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < x < +\infty,\ n > 0,$$

onde $\Gamma(\alpha)$ é a função Gama.

O parâmetro caracterizador desta distribuição é n.

Principais características:

- É simétrica em relação ao eixo $x = 0$;
- A distribuição *t*-Student tende para a distribuição Normal à medida que n aumenta.

A distribuição *t*-Student é mais pontiaguda e tem caudas mais pesadas do que a distribuição Normal (Figura 5.14). Esta distribuição está tabelada (Anexo C).

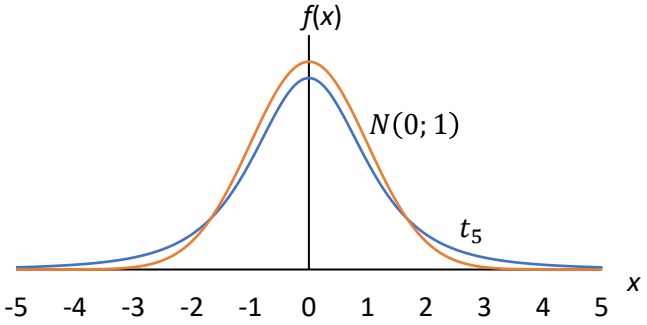


Figura 5.14: *Comparação da função densidade de probabilidade da distribuição t-Student (com $n = 5$) com a distribuição $N(0;1)$.*

Se $X \sim t_n$, então $\mu_X = E(X) = 0$ e $\sigma_X^2 = Var(X) = \frac{n}{n-2}, n > 2$.

Usualmente utiliza-se a notação $t_{n;\,\alpha}$ para representar o quantil de probabilidade α de uma v. a. $X \sim t_n$. Portanto, $t_{n;\,\alpha}$ corresponde ao menor valor k tal que $P(X \leq k) = \alpha$.

Teorema: Se X e Y forem v. a. independentes e $X \sim N(\mu;\ \sigma)$ e $Y \sim \chi^2_n$, então

$$T = \frac{\left(\frac{X-\mu}{\sigma}\right)}{\sqrt{\frac{Y}{n}}} \sim t_n.$$

5.2.7 Distribuição F de Fisher-Snedecor

Definição: A v. a. contínua X segue uma **distribuição F de Fisher-Snedecor** (usualmente denominada por distribuição F) **com m e n graus de liberdade**, i.e., $X \sim F_{m;n}$, se a sua função densidade de probabilidade é:

$$f(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{x^{\frac{m-2}{2}}}{\left(1+\frac{m}{n}x\right)^{\frac{m+n}{2}}}, \; x > 0, \; m > 0, n > 0.$$

Os parâmetros caracterizadores desta distribuição são m e n.

Principais características:

- A v. a. só toma valores positivos;
- É uma função não simétrica.

A forma da distribuição F depende dos valores dos parâmetros m e n (Figura 5.15). Esta distribuição está tabelada (Anexo A4).

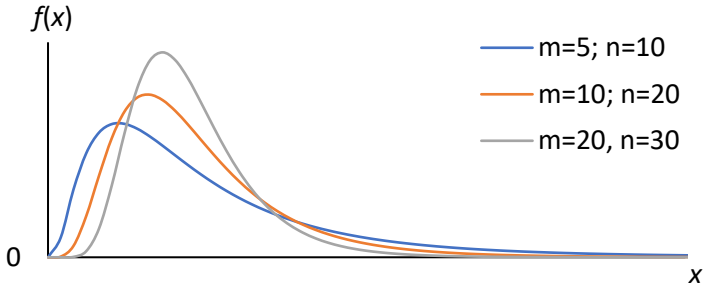


Figura 5.15: *Função densidade de probabilidade da distribuição F para diferentes valores de m e n.*

Se $X \sim F_{m;\,n}$ então $\mu_X = E(X) = \frac{n}{n-2}, n > 2,$ e $\sigma_X^2 = Var(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(m-4)}, n > 4.$

Usualmente utiliza-se a notação $F_{m;\,n;\,\alpha}$ para representar o quantil de probabilidade α de uma v. a. $X \sim F_{m;n}$. Portanto, $F_{m;\,n;\,\alpha}$ corresponde ao menor valor k tal que $P(X \leq k) = \alpha$.

Teorema: Se $X \sim F_{m;n}$, então

$$\frac{1}{X} \sim F_{n;\,m} \quad \text{e} \quad F_X^{-1}(\alpha) = \frac{1}{F_{\frac{1}{X}}^{-1}(1-\alpha)} \Leftrightarrow F_{m;\,n;\,\alpha} = \frac{1}{F_{n;\,m;\,1-\alpha}}.$$

Teorema: Se $X \sim t_n$, então $X^2 \sim F_{1;n}$.

Teorema: Se X e Y forem v. a. independentes e $X \sim \chi_m^2$ e $Y \sim \chi_n^2$, então

$$F = \frac{\frac{X}{m}}{\frac{Y}{n}} \sim F_{m;n}.$$

5.2.8 Exercícios resolvidos

5.2.8.1 Distribuição Uniforme

Considere que o tempo de viagem de autocarro entre Lisboa e Évora segue uma distribuição Uniforme entre 100 a 120 minutos.

a) Identifique a v. a. em estudo e a sua distribuição densidade de probabilidade.
b) Qual a duração média das viagens? Calcule a variância.
c) Qual a probabilidade de uma viagem demorar mais de 115 minutos?
d) Qual a proporção de viagens de demoram menos de 110 minutos?
e) Calcule a probabilidade de uma viagem durar entre 105 e 115 minutos?

Resolução:

a) Seja X a v.a. que representa o tempo de viagem de autocarro entre Lisboa e Évora, com $X \sim U(100;\ 120)$. Função densidade de probabilidade:

$$f(x) = \frac{1}{120 - 100} = 0{,}05, \qquad 100 < x < 120.$$

b) $E(X) = \frac{100 + 120}{2} = 110$ (duração média das viagens).

$Var(X) = \frac{(120 - 100)^2}{12} = 33{,}3333.$

c) Como a função de distribuição é:

$$F(x) = P(X \leq x) = \frac{x - 100}{20}, \qquad 100 < x < 120,$$

então

$P(X > 115) = 1 - P(X \leq 115) = 1 - \frac{115 - 100}{20} = 0{,}25.$

d) $P(X < 110) = \frac{110 - 100}{20} = 0{,}5.$

e) $P(105 < X < 115) = \frac{115 - 105}{20} = 0{,}5.$

5.2.8.2 Distribuição Normal

Uma empresa, monopolista no mercado de determinado produto, tem produção constante de 90 toneladas (ton.) por dia. Sabe-se que a procura diária é uma v. a. com distribuição Normal, com média $\mu = 80$ toneladas e desvio padrão $\sigma = 10$ toneladas.

a) Qual a procura média diária deste produto? Calcule a variância.
b) Determine a probabilidade de a procura ser inferior a 78 toneladas.
c) Calcule a probabilidade de ocorrer procura excedentária?
d) Qual a percentagem de dias em que a procura se situa entre 77 e 82 toneladas?
e) Qual deve ser a produção mínima diária para que a probabilidade de ocorrer procura excedentária seja 0,025.
f) Qual a probabilidade de em 9 dias, selecionados ao acaso, a procura total ser superior a 780 ton.?
g) Selecionados ao acaso 4 dias, determine a probabilidade de, em média, a procura ser no máximo 75 ton. por dia?

Resolução:

a) Seja X a v. a. que representa a procura diária, em toneladas, do produto, com $X \sim N(\mu = 80;\ \sigma = 10)$.
$E(X) = \mu = 80$ toneladas (procura média).
$Var(X) = \sigma^2 = 10^2 = 100$.

b) $P(X < 78) = P\left(\frac{X-\mu}{\sigma} < \frac{78-80}{10}\right) = P(Z < -0{,}2) = \Phi(-0{,}2) = 1 - \Phi(0{,}2) = 1 - 0{,}5793 = 0{,}4207.$

c) $P(X > 90) = 1 - P(X \leq 90) = 1 - P\left(Z \leq \frac{90-80}{10}\right) = 1 - \Phi(1) = 1 - 0{,}8413 = 0{,}1587.$

d) $P(77 \leq X \leq 82) = P\left(\frac{77-80}{10} \leq Z \leq \frac{82-80}{10}\right) = P(-0{,}3 \leq Z \leq 0{,}2) = \Phi(0{,}2) - \Phi(-0{,}3)$
$= \Phi(0{,}2) - \big(1 - \Phi(0{,}3)\big) = 0{,}5793 - (1 - 0{,}6179) = 0{,}1972.$
Em 19,72% dos dias a procura situa-se entre 77 e 82 toneladas.

e) $P(X > k) = 0{,}025 \Leftrightarrow P(X \leq k) = 0{,}975 \Leftrightarrow P\left(Z \leq \frac{k-80}{10}\right) = 0{,}975 \Leftrightarrow \Phi\left(\frac{k-80}{10}\right) = 0{,}975$
como $\Phi(1{,}96) = 0{,}975$
$\Rightarrow \frac{k-80}{10} = 1{,}96 \Leftrightarrow k = 99{,}6.$
Portanto, a empresa deve produzir pelo menos 99,6 toneladas por dia.

f) Seja X_i a v. a. que representa a procura do produto, em toneladas, no dia i, com $X_i \sim N(\mu = 80;\ \sigma = 10)$.
Seja S_9 a v. a. que representa a procura total, em toneladas, em 9 dias selecionados ao acaso (i.e., X_i são independentes, $i = 1, \ldots, 9$), com $S_9 = X_1 + X_2 + \cdots + X_9 \sim N\big(n\mu = 720;\ \sigma\sqrt{n} = 30\big)$.
$P(S_4 > 780) = P\left(Z > \frac{780-720}{30}\right) = 1 - \Phi(2) = 1 - 0{,}9772 = 0{,}0228.$

g) Seja X_i a v. a. que representa a procura do produto, em toneladas, no dia i, com $X_i \sim N(\mu = 80;\ \sigma = 10)$.
Seja $\overline{X}$ a v. a. que representa a procura total, em toneladas, em 4 dias selecionados ao acaso (i.e., X_i são independentes, $i = 1, \ldots, 4$), com
$$\overline{X} = \frac{X_1 + X_2 + X_3 + X_4}{4} \sim N\left(\mu = 80;\ \frac{\sigma}{\sqrt{n}} = 5\right).$$
$P\big(\overline{X} \leq 75\big) = P\left(Z \leq \frac{75-80}{5}\right) = \Phi(-1) = 1 - \Phi(1) = 1 - 0{,}8413 = 0{,}1587.$

5.2.8.3 Distribuição Exponencial

Seja X a v. a. que representa o tempo, em segundos, decorrido entre a entrada consecutiva de 2 automóveis numa autoestrada (AE). Admita que esta v. a. segue uma distribuição Exponencial com valor médio 15 segundos.

a) Descreva as funções densidade de probabilidade e de distribuição associadas a esta v. a.
b) Qual o valor esperado e a variância de X.
c) Calcule a probabilidade do tempo que medeia a entrada consecutiva de 2 automóveis na AE ser superior a 50 segundos.
d) Determine a probabilidade do intervalo de tempo que separa a entrada consecutiva de 2 automóveis na AE ser inferior a 1 minuto.
e) Calcule $P(X > \sigma_X)$.
f) Determine a probabilidade do tempo que ocorre entre a entrada consecutiva de 2 automóveis na AE estar entre 10 e 20 minutos.

Resolução:

Sabemos que

$$X \sim Exp\left(\lambda = \frac{1}{15}\right).$$

a) Função densidade de probabilidade:

$$f(x) = \frac{1}{15} e^{-\frac{x}{15}}, \qquad x > 0.$$

Função de distribuição:

$$F(x) = P(X \leq x) = 1 - e^{-\frac{1}{15}x}, \ x > 0.$$

b) $E(X) = \frac{1}{\frac{1}{15}} = 15$ segundos.

$Var(X) = \frac{1}{\left(\frac{1}{15}\right)^2} = 15^2 = 225.$

c) $P(X > 50) = 1 - P(X \leq 50) = 1 - F(50) = 1 - \left(1 - e^{-\frac{1}{15}50}\right) = 0{,}0357.$

d) Dado que 1 minuto = 60 segundos

$P(X < 60) = F(60) = 1 - e^{-\frac{1}{15}60} = 0{,}9817.$

e) Como $\sigma_X = \sqrt{Var(X)} = 15$ segundos,

$P(X > \sigma_X) = P(X > 15) = 1 - P(X \leq 15) = 1 - F(15) = 1 - \left(1 - e^{-\frac{1}{15}15}\right) = 0{,}3679.$

f) $P(10 < X < 20) = F(20) - F(10) = 1 - e^{-\frac{1}{15}20} - \left(1 - e^{-\frac{1}{15}10}\right) = 0{,}2498.$

5.3 Exercícios propostos

1. No âmbito de um estudo sobre a criação de um amplo espaço verde numa cidade de média dimensão, constatou-se que 80% dos cidadãos concordavam com a medida. Foram inquiridos 10 cidadãos.

a) Defina a função de probabilidade da variável aleatória X que representa o número de cidadãos que concordam com a criação do espaço verde, em n inquiridos.
b) Calcule a probabilidade de:
   i. Nenhum concordar.
   ii. Pelo menos 2 concordarem.
   iii. Entre 7 a 9 concordarem.
c) Calcule o valor esperado e a variância da distribuição definida em a).

2. A percentagem de alunos de Psicologia que procuram resolver os exercícios das aulas práticas de Estatística é de 15%.

a) Qual a probabilidade de numa turma com 20 alunos, pelo menos 2 alunos terem tentado resolver os exercícios?
b) Numa turma de 40 alunos, em média quantos deles tentaram resolver os exercícios?
c) Numa turma de 100 alunos, qual a probabilidade de (recorra a um *software*):
   i. No mínimo 20 alunos terem tentado resolver os exercícios?
   ii. Mais de 15 e no máximo 25 alunos terem tentado resolver os exercícios?

3. Numa experiência biológica, para a qual a escolha das cobaias é bastante dispendiosa, verifica-se, que a experiência é bem sucedida em 40% dos casos.
   a) Se tiver 10 cobaias, qual a probabilidade de ter pelo menos duas experiências bem sucedidas?
   b) Quantas cobaias são necessárias para que o número esperado de sucessos seja 24? Justifique a resposta.
   c) Quantas cobaias serão necessárias para garantir que a probabilidade de obter pelo menos uma experiência com sucesso não seja inferior a 0,95?
   d) Quantas cobaias espera ter que usar até obter uma experiência com sucesso?
   e) Qual a probabilidade de ter que realizar 3 experiências até obter uma com sucesso?
   f) Quantas experiências espera ter que realizar até obter 10 bem-sucedidas?
   g) Calcule a probabilidade de só necessitar realizar 11 experiências até obter 10 bem sucedidas.

4. O responsável do *Lugar da Música* estima que 60% dos seus clientes preferem fazer *download* de música ligeira, 30% clássica e os restantes preferem comprar música etnográfica. Em 10 clientes, qual a probabilidade de:
   a) Três clientes terem feito um *download* de música ligeira, dois um *download* de música clássica e os restantes um *download* de música etnográfica?
   b) Haver um cliente interessado em fazer *download* de música etnográfica e pelo menos sete em música ligeira?

5. O responsável de crédito duma instituição financeira, ao analisar os relatórios dos vários departamentos regionais, verificou que dos 12 novos clientes em Évora, 2 não tinham satisfeito os seus compromissos e 4 tinham pedido a renegociação das condições de crédito.
Pela experiência sabe que, relativamente aos novos clientes, a não satisfação dos compromissos e o pedido de renegociação das condições de crédito ocorrem, respetivamente, em 1% e 5% dos casos.
Acha que o responsável de crédito da instituição tem razões para estranhar a informação do departamento regional de Évora? Justifique.

6. Uma empresa possui em *stock* 1000 lâmpadas das quais 5%, em média, estão fundidas. A fim de atender a uma encomenda de um cliente, a empresa preparou um lote de 100 lâmpadas. Calcule a probabilidade de nesse lote irem 5 lâmpadas fundidas (note que se trata de uma extração sem reposição).

7. O cliente do exercício anterior ao receber as 100 lâmpadas resolve testar algumas. Calcule a probabilidade de num grupo de dez lâmpadas aparecer uma fundida se:
   a) As extrações forem sem reposição;
   b) As extrações forem com reposição.

8. Das 100 crianças que frequentam o infantário *Os Traquinas* 15 são bebés com menos de 1 ano de idade. São selecionadas, ao acaso, 10 crianças para irem assistir às gravações do *Batatoon*.
   a) Qual a probabilidade de nesse grupo não irem bebés com menos de 1 ano?
   b) Qual o valor esperado e a variância da variável aleatória em estudo?

9. A observação de atos de agressividade numa tribo de chimpanzés-pigmeus leva a crer que eles ocorrem de acordo com um processo de Poisson com taxa 7/hora.
   a) O que é a taxa de um processo de Poisson?
   b) Em termos gerais, quantos atos de agressividade ocorrem durante um dia?
   c) Qual é a probabilidade de numa hora ocorrerem menos do que três atos de agressividade?

10.Suponha que um livro de 585 páginas contém 43 erros tipográficos. Se esses erros estiverem aleatoriamente distribuídos pelo livro, qual a probabilidade de:

a) Uma página qualquer não ter erros;
b) Onze páginas escolhidas ao acaso não terem erros.

11.O serviço de atendimento de uma instituição de saúde recebe, em média, 12 reclamações em cada período de 8 horas. Qual a probabilidade de, num período de uma hora:

a) Não ser recebida nenhuma reclamação;
b) Serem recebidas pelo menos duas reclamações.

12.O verdadeiro peso de sacos de quilo de arroz é aleatório e apresenta uma densidade de probabilidade uniformemente distribuída entre 0,8 e 1,05 Kg.

a) Indique a função de densidade de probabilidade em questão.
b) Qual a probabilidade de um saco de arroz pesar menos de um quilo?
c) Qual o peso médio dos referidos sacos?

13.Num certo hospital, a temperatura na sala de espera das urgências tem distribuição uniforme entre 20 e 24 graus.

a) Construa a função de distribuição adequada.
b) Calcule a temperatura média da sala e a variância.
c) Qual a probabilidade de a temperatura da sala estar entre 21 e 23 graus?

14.Seja X a variável aleatória que representa a altura, em cm, dos alunos do sexo masculino de uma certa universidade. Sabe-se que $X \sim N(160;\ 10)$. Sem efectuar os cálculos, diga qual das seguintes afirmações está correta e justifique:

a) $P(X < 150) < 0{,}5$;
b) $P(X > 150) < 0{,}5$;
c) $P(X > 170) > 0{,}5$;
d) $P(X < 170) < 0{,}5$.

15.Admita-se que a carga de rutura de certo tipo de fio de aço se distribui normalmente com valor médio de 300 (quilos) e variância de 36 (quilos^2). Qual a probabilidade de que se parta um fio desse tipo, tomado ao acaso, se o sujeitar a uma carga superior ou igual a 282 quilos?

16.O tempo, em minutos, que os alunos do curso de Sociologia demoram a resolver todas as alíneas do exame da disciplina de Estatística pode considerar-se que segue uma distribuição Normal com média 110 minutos e desvio padrão 15 minutos.

a) Sabendo que o professor determina que a duração máxima para resolver o exame é de 2 horas (120 minutos), qual a probabilidade de um aluno não acabar de resolver o exame?
b) Numa das salas de exame estão 50 alunos, quantos espera que acabem de resolver o exame?
c) Qual a probabilidade de um aluno demorar menos de 105 minutos a resolver o exame?
d) Determine a probabilidade de um aluno demorar entre 100 e 110 minutos a resolver o exame.
e) Complete: “29,12% dos alunos demoraram mais de .... minutos a resolver o exame”.
f) Complete: “75% dos alunos demoraram no máximo .... minutos a resolver o exame”.

17. Sejam X e Y duas v. a. que representam as temperaturas médias do ar (em °C) registadas, respetivamente, pelas estações meteorológicas A e B. Sabe-se que $X \sim N(15;\ 2)$ e $Y \sim N(16;\ 2)$.

a) Diga, justificando, qual das seguintes afirmações está correta:
   i. $P(X < 14) < P(Y > 15)$
   ii. $P(X < 14) = P(Y > 15)$

iii. $P(X < 14) > P(Y > 15)$

b) Complete: "Em 5% dos dias a temperatura média, registada na estação meteorológica A, é superior a ...°C".

18. A distribuição dos rendimentos familiares num determinado bairro com 5000 famílias segue uma lei Normal, com $\mu = 180$ u. m. e $\sigma = 5$ u. m.

a) Calcule a probabilidade de uma família:
   i. Ter um rendimento superior a 190 u. m.
   ii. Não auferir mais de 163 u. m.
   iii. Auferir entre 175 e 188 u. m.
b) Qual o rendimento máximo auferido pelo grupo das 500 famílias de menores rendimentos;
c) Qual o rendimento mínimo das 500 famílias com maiores rendimentos?

19.O tempo de execução de determinada tarefa é uma variável aleatória com distribuição Normal, com $\mu = 72$ minutos e $\sigma = 12$ minutos.

a) Calcule a probabilidade da tarefa:
   i. Levar mais de 93 minutos a ser executada.
   ii. Não demorar mais de 65 minutos.
   iii. Gastar entre 63 e 78 minutos.
b) Determine os menores valores a e b tais que:
   i. $P(X > a) = 0{,}2525$.
   ii. $P(X < b) = 0{,}0054$.

20.Sendo X a variável aleatória que representa o tempo, em minutos, que um dado medicamento demora a ser administrado num paciente, que segue a distribuição $N(23;\ 4)$, determine:

a) A probabilidade do medicamento demorar menos de 21 minutos a ser administrado.
b) O menor valor k tal que $P(k < X < 21) = 0{,}0819$.

21.Seja X uma variável aleatória com distribuição Normal de valor médio 10 kg e desvio padrão de 2 kg, que representa a quantidade de lixo orgânico produzido semanalmente por cada agregado familiar numa determinada cidade. O departamento de gestão da recolha do lixo da Câmara Municipal dessa cidade, considera que a quantidade expectável de lixo produzido por agregado familiar se situa entre os 8 e os 12 kg.

a) Qual a probabilidade de que um agregado familiar, escolhido ao acaso, produza semanalmente uma quantidade de lixo expectável?
b) Qual a probabilidade de que em 10 agregados familiares, escolhidos aleatoriamente, pelo menos 2 produzam uma quantidade fora dos parâmetros esperados?

22.Considere somente os clientes do banco X que possuem o produto de poupança PPX. Seja X a v.a. que representa o saldo diário da conta à ordem desses clientes. Sabe-se que X segue uma distribuição Normal com média 700 Euros e desvio padrão 300 Euros.

a) Calcule $E(X^2)$.
b) Selecionado ao acaso um cliente de entre os que possuem um saldo à ordem superior a 850 Euros, qual a probabilidade desse cliente ter um saldo à ordem superior a 900 Euros?
c) Assumindo a independência entre os clientes:
   i. Num grupo de 10 clientes, qual a probabilidade de o saldo diário total destes clientes ser superior a 7000 Euros?
   ii. Qual a probabilidade de, em 20 clientes selecionados ao acaso, 5 terem um saldo à ordem superior a 850 Euros?

iii. Determine a probabilidade de ter que selecionar 5 clientes até obter 1 com um saldo à ordem superior a 850 Euros?
iv. Calcule o número esperado de clientes a selecionar até obter 3 com um saldo à ordem superior a 850 Euros?

23.Os responsáveis de uma empresa produtora de açúcar, face a uma diminuição registada nas vendas do produto, encararam a hipótese de fechar uma das suas sucursais. De acordo com estudos efetuados, para não encerrar a dita sucursal, a empresa necessita que a procura semanal (7 dias) seja superior a 65.000 kg. Sabendo-se que a procura média diária de açúcar se distribui normalmente, com média 10.000 kg e desvio padrão 1.000 kg, determine a probabilidade da sucursal encerrar.

24.Considere as variáveis aleatórias (v. a.) independentes X_1 e X_2, sendo

$$X_1 \sim N(\mu_1 = 60;\ \sigma_1 = 15) \text{ e } X_2 \sim N(\mu_2 = 20;\ \sigma_2 = 4).$$

Sabendo que a v. a. $Y = X_1 - X_2$ segue uma distribuição $N(\mu_Y;\ \sigma_Y)$, determine os respectivos parâmetros μ_Y e σ_Y.

25.Considere que o peso dos bacalhaus abatidos diariamente na unidade A, da empresa de aquicultura *AquiBacalhaus*, segue uma distribuição Normal com média 2,5 kg e desvio padrão 1 kg. O ganho com a venda de cada bacalhau com peso superior a 3 kg é de 5 euros, com peso entre os 1,5 kg e 3 kg é de 4 euros e os restantes a 3 euros.

a) Qual a probabilidade de um bacalhau pesar mais de 3 kg?
b) 10% dos bacalhaus pesam no máximo quantos kg?
c) Seja G a v.a. que representa o ganho com a venda de um bacalhau. Qual a função de probabilidade da v.a. G?
d) Qual o ganho esperado num dia em que são abatidos 1000 bacalhaus?
e) Sabendo que $E(G^2) = 17{,}67$ calcule $Var(0{,}8G - 0{,}1)$.
f) Num dia em que são abatidos 500 bacalhaus, qual a probabilidade do peso médio dos bacalhaus ser superior a 2,6 kg?
g) Se reduzir o número de animais abatidos o que acontece à probabilidade calculada na alínea anterior? Justifique sem efetuar o cálculo da probabilidade.

26.No fabrico em série de determinado tipo de chaves surgem aproximadamente 4% de defeituosas. Determine a probabilidade aproximada de uma amostra de 300 chaves conter não menos de 8 nem mais de 12 chaves defeituosas.

27.O número de alunos que chegam, por hora, à secção de fotocópias da *Casa das Folhas* em Évora é uma v. a. X com distribuição Poisson com variância 5.

a) Determine o número de alunos que chegam, por hora, à *Casa das Folhas*.
b) Qual a probabilidade de chegarem no máximo 2 alunos, por hora, à *Casa das Folhas*.
c) Sabendo que a *Casa das Folhas* funciona 8 horas por dia, qual a probabilidade aproximada de, num dia, chegarem exatamente 20 alunos?

28.O tempo de atendimento de um aluno na secção de fotocópias da *Casa das Folhas*, em Évora, é uma v. a. T com distribuição Exponencial de valor médio igual a 10 minutos.

a) Descreva a função de distribuição da v. a. T.
b) Qual a probabilidade de lhe tirarem as fotocópias em menos de 15 minutos?

29.Suponha que o tempo T decorrido entre a entrada consecutiva de 2 automóveis numa auto-estrada, segue uma distribuição Exponencial com valor médio de 15 segundos.

a) Defina a função densidade de probabilidade da variável aleatória T.
b) Determine a $P(T > \sigma_T)$.

30.Suponha que o tempo de duração (X) de determinadas lâmpadas, é uma v. a. com distribuição exponencial cuja função de distribuição de probabilidade é dada por:

$$F(x) = \begin{cases} 1 - e^{-\frac{x}{500}}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

Calcule a probabilidade do tempo de duração de uma lâmpada:

a) Estar entre 100 e 200 horas.
b) Ser inferior a 300 horas.

31.Seja X o tempo espera (em minutos) para ser atendido num dado restaurante à hora do almoço. Admita que $X \sim \chi^2_{23}$.

a) Determine a probabilidade de ter que esperar entre 14,85 e 32,01 minutos.
b) Obtenha os menores valores de a e b tais que $P(a < X < b) = 0{,}95$ e $P(X < a) = 0{,}025$.
c) Qual a média e a variância de X.
d) Determine $\chi^2_{23;\,0{,}05}$ e $\chi^2_{23;\,0{,}95}$. Interprete.

32.Seja T a v.a. que representa a temperatura num certo frigorífico, sabendo-se que $T \sim t_{16}$, obtenha:

a) A probabilidade da temperatura ser de pelo menor 2,12 °C.
b) O menor valor k tal que $P(T < k) = 0{,}10$.
c) $t_{16;\,0{,}99}$ e interprete.

33.Considere que a v. a. F, que a cotação de fecho (em u.m.) das ações X transacionadas numa dada bolsa de valores tem distribuição $F_{10;\,6}$. Calcule:

a) A probabilidade de ao selecionar, ao acaso, um dia a cotação de fecho ser inferior a 7,87 u.m.
b) O menor valor k tal que $P(F > k) = 0{,}90$.
c) $f_{10;6;\,0{,}05}$ e interprete.

6 Distribuições por amostragem

Como já foi referido, nem sempre é possível e viável estudar com exatidão toda a população, pelo que se retira uma amostra dessa população e, com base na informação amostral, infere-se, quando possível, para a população. A base dessa inferência assenta no tipo de amostragem probabilística que se realizou e nas distribuições de probabilidade, cuja junção resulta nas distribuições por amostragem. Ao longo deste capítulo assume-se que a amostra é aleatória.

Definição: A v. a. $(X_1, X_2, \dots, X_n)$ diz-se uma **amostra aleatória** (a. a.) retirada de uma determinada população, se a sua função (densidade) de probabilidade conjunta for dada por:

$$f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2)\dots f(x_n) = \prod_{i=1}^{n} f(x_i),$$

ou seja, $X_1, X_2, \dots, X_n$ são independentes e identicamente distribuídas (i. i. d.).

6.1 Teorema do limite central

Teorema do Limite Central (T. L. C.): Seja $X_1, X_2, \dots, X_n$, uma a. a. de dimensão n, com $E(X_i) = \mu$ e $Var(X_i) = \sigma^2$ para $i = 1, 2, \dots, n$. Considere-se $S_n = X_1 + X_2 + \cdots + X_n$. Para valores grandes de n tem-se que:

$$\frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \overset{\circ}{\sim} N(0;1).$$

Observação: Não existe consenso sobre o que se considera uma amostra grande: por exemplo, alguns autores consideram $n > 30$, outros $n > 50$ e alguns são ainda mais exigentes.

Corolário: Seja $X_1, X_2, \dots, X_n$, uma a. a. de dimensão n, com $E(X_i) = \mu$ e $Var(X_i) = \sigma^2$ para $i = 1, 2, \dots, n$. Considere-se

$$\overline{X} = \frac{S_n}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

Para valores grandes de n tem-se que:

$$\frac{\overline{X} - E(\overline{X})}{\sqrt{Var(\overline{X})}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \overset{\circ}{\sim} N(0;1).$$

6.1.1.1 Correção de continuidade

A variante do T. L. C. para distribuições discretas introduz a correção de continuidade que permite aproximar uma distribuição discreta por uma distribuição contínua, neste caso a distribuição Normal.

Correção de continuidade: Seja X uma v. a. discreta, com variação Δ (i.e., X pode tomar os valores $0, \Delta, 2\Delta, 3\Delta, \dots$), com $E(X) = \mu_X$ e $Var(X) = \sigma_X^2$. Então a correção de continuidade é efetuada da seguinte forma:

$$P(X = k) \approx P\left(k - \frac{\Delta}{2} \leq X \leq k + \frac{\Delta}{2}\right).$$

Note-se que nas distribuições usuais, por exemplo, Binomial e Poisson, $\Delta = 1$, pois tratam-se de distribuições de contagem logo

$$P(X = k) \approx P(k - 0{,}5 \leq X < k + 0{,}5).$$

Com a aplicação das propriedades da distribuição Normal, podem generalizar-se as seguintes regras:

- $P(X \leq k) \approx P\left(Z \leq \dfrac{k + 0{,}5 - \mu_X}{\sigma_X}\right) = \Phi\left(\dfrac{k + 0{,}5 - \mu_X}{\sigma_X}\right)$;
- $P(X < k) \approx P\left(Z < \dfrac{k - 0{,}5 - \mu_X}{\sigma_X}\right) = \Phi\left(\dfrac{k - 0{,}5 - \mu_X}{\sigma_X}\right)$;
- $P(X \geq k) \approx P\left(Z \geq \dfrac{k - 0{,}5 - \mu_X}{\sigma_X}\right) = 1 - \Phi\left(\dfrac{k - 0{,}5 - \mu_X}{\sigma_X}\right)$;
- $P(X > k) \approx P\left(Z > \dfrac{k + 0{,}5 - \mu_X}{\sigma_X}\right) = 1 - \Phi\left(\dfrac{k + 0{,}5 - \mu_X}{\sigma_X}\right)$.

6.1.1.2 Aproximação da distribuição Binomial pela distribuição Normal

Se $X \sim B(n; p)$, com $n > 50$ e $0{,}1 < p < 0{,}9$, então

$$X \overset{\circ}{\sim} N\left(\mu = np;\ \sigma = \sqrt{npq}\right), \text{ ou seja, } Z = \frac{X - np}{\sqrt{npq}} \overset{\circ}{\sim} N(0;\ 1).$$

Na Figura 6.1 ilustra-se, com um exemplo, como se processa a aproximação da distribuição Binomial pela distribuição Normal.

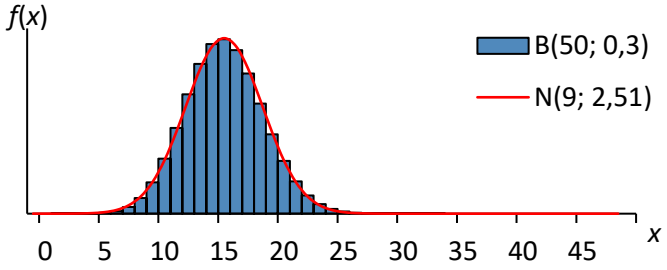


Figura 6.1: *Aproximação da distribuição $B(n = 50; p = 0{,}3)$ pela distribuição $N(\mu = 15;\ \sigma = 3{,}2404)$.*

6.1.1.3 Aproximação da distribuição Poisson pela distribuição Normal

Se $X \sim P(\lambda)$, com $\lambda > 20$ então $X \overset{\circ}{\sim} N\left(\mu = \lambda;\ \sigma = \sqrt{\lambda}\right)$, ou seja, $Z = \frac{X - \lambda}{\sqrt{\lambda}} \overset{\circ}{\sim} N(0;\ 1)$.

Na Figura 6.2 ilustra-se, com um exemplo, como se processa a aproximação da distribuição Poisson pela distribuição Normal.

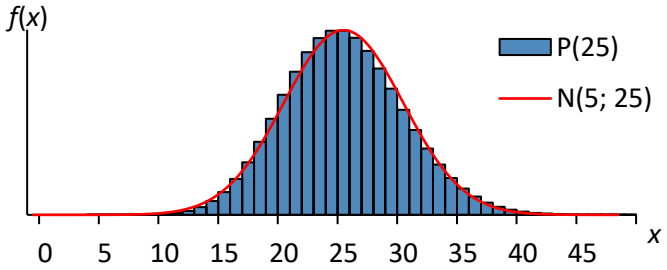


Figura 6.2: *Aproximação da distribuição $P(\lambda = 25)$ pela distribuição $N(\mu = 25;\ \sigma = 5)$.*

6.2 Distribuição da média amostral

Seja $X_1, X_2, \dots, X_n$, uma a. a. duma dada população com média μ e variância σ^2. A **média amostral** é definida por:

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}.$$

Propriedades:

- $\mu_{\overline{X}} = E(\overline{X}) = \mu$;
- $\sigma^2_{\overline{X}} = Var(\overline{X}) = \frac{1}{n}\sigma^2$.

Fator de correção de população finita:

Se a dimensão N da população é conhecida e a amostra foi recolhida com base numa amostragem sem reposição tendo uma dimensão $n \leq 0{,}05N$, então:

$$\sigma^2_{\overline{X}} = \frac{1}{n}\sigma^2\left(\frac{N-n}{N-1}\right).$$

6.2.1 Quando a variância é conhecida

Se a distribuição da população é Normal com desvio padrão σ conhecido, então $\overline{X} \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$, ou seja,

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1).$$

Se a distribuição da população não for Normal, mas a amostra é de grande dimensão então, pelo corolário do T. L. C., vem

$$Z \overset{\circ}{\sim} N(0;1).$$

6.2.2 Quando a variância é desconhecida

Se a população é Normal mas o desvio padrão σ é desconhecido, e não rejeitando a hipótese de independência das distribuições por amostragem da média e da variância da amostra, então tem-se que:

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}.$$

Se a distribuição da população não for Normal, mas a amostra for de grande dimensão então, por extensão do corolário do T. L. C.,

$$\overline{X} \overset{\circ}{\sim} N\left(\mu; \frac{S}{\sqrt{n}}\right), \text{ ou seja, } Z = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \overset{\circ}{\sim} N(0;1).$$

Observação: Para valores elevados de n, a distribuição *t*-Student toma valores muito próximos aos da $N(0;1)$. Existem alguns programas estatísticos (por exemplo, SPSS) que, nestas condições, não utilizam a distribuição Normal mas a *t*-Student.

6.3 Distribuição da diferença de médias amostrais

Sejam $X_{11}, X_{12}, \dots, X_{1n_1}$ e $X_{21}, X_{22}, \dots, X_{2n_2}$ duas a. a. independentes, de dimensão n_1 e n_2 retiradas de duas populações com médias μ_1 e μ_2 e desvios padrão σ_1 e σ_2, respetivamente, e

$$\overline{X}_1 = \sum_{i=1}^{n_1} \frac{X_{1i}}{n_1} \text{ e } \overline{X}_2 = \sum_{i=1}^{n_2} \frac{X_{2i}}{n_2}.$$

6.3.1 Quando as variâncias são conhecidas

Se as populações forem Normais, sendo $X_1 \sim N(\mu_1; \sigma_1)$ e $X_2 \sim N(\mu_2; \sigma_2)$, com σ_1 e σ_2 conhecidos, então, pelo Teorema da aditividade da distribuição Normal, tem-se que:

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0; 1).$$

Se as populações não forem Normais, mas as amostras forem de grande dimensão então, pelo corolário do T. L. C., vem $Z \overset{\circ}{\sim} N(0; 1)$, considerando o já anteriormente exposto em situação análoga.

6.3.2 Quando as variâncias são desconhecidas mas iguais

Se as populações forem Normais, sendo $X_1 \sim N(\mu_1; \sigma_1)$ e $X_2 \sim N(\mu_2; \sigma_2)$, com σ_1 e σ_2 desconhecidos mas iguais $(\sigma_1 = \sigma_2)$, então demonstra-se que:

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}.$$

Se as populações não forem Normais, mas as amostras forem de grande dimensão então, por extensão do T. L. C., a expressão anterior segue aproximadamente uma $N(0; 1)$.

6.3.3 Quando as variâncias são desconhecidas mas diferentes

Se as populações forem Normais, sendo $X_1 \sim N(\mu_1; \sigma_1)$ e $X_2 \sim N(\mu_2; \sigma_2)$, com σ_1 e σ_2 desconhecidos e diferentes $(\sigma_1 \neq \sigma_2)$, então, pela aproximação de Welch tem-se que (Murteira *et al.*, 2007):

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{\circ}{\sim} t_v, \text{ onde } v = \left[\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2}\right].$$

sendo $[r]$ a parte inteira de r, ou seja, arredonda-se por defeito o valor obtido.

Também neste caso, se as populações não forem Normais, mas as amostras forem de grande dimensão então, por extensão do T. L. C., a expressão anterior segue aproximadamente uma $N(0; 1)$, sendo válidas as observações anteriores.

6.4 Distribuição da proporção amostral

Seja $X_1, X_2, \dots, X_n$, uma a. a. duma população Bernoulli com probabilidade de sucesso p, em que X_i toma o valor 1 se for um sucesso e o valor 0 se for um insucesso. A **proporção amostral** de sucessos é dada por:

$$\overline{P} = \sum_{i=1}^{n} \frac{X_i}{n}.$$

Propriedades:

- $\mu_{\overline{P}} = E(\overline{P}) = p.$
- $\sigma^2_{\overline{P}} = Var(\overline{P}) = \frac{p(1-p)}{n}.$

Se a dimensão da amostra for pequena então sabe-se que,

$$P(X = x) = {}^{n}C_x p^x (1-p)^{n-x}, \qquad x = 0, 1, \dots, n, \text{ com } 0 < p < 1,$$

pelo que

$$P(\overline{P} = a) = P(X = na) = {}^{n}C_{na} p^{na} (1-p)^{n-na}, \qquad a = 0, \frac{1}{n}, \frac{2}{n} \dots, \frac{n}{n}.$$

Se a dimensão da amostra for grande então, pelo Teorema de Moivre-Laplace (i.e., variante do T. L. C. para a distribuição Binomial),

$$\overline{P} \overset{\circ}{\sim} N\left(p; \sqrt{\frac{p(1-p)}{n}}\right), \text{ ou seja, } Z = \frac{\overline{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\circ}{\sim} N(0; 1).$$

Observação: Esta aproximação tem sido muitas vezes criticada, pois tem mostrado resultados consideravelmente insatisfatórios e, apenas, continua a ser usada dada a sua simplicidade e implementação nos diversos programas estatísticos. Recomenda-se especial atenção se a proporção for tendencialmente pequena ou grande, i.e., valores próximos de 0 ou de 1 (Pires e Amado, 2008).

6.5 Distribuição da diferença de proporções amostrais

Sejam $X_{11}, X_{12}, \dots, X_{1n_1}$ e $X_{21}, X_{22}, \dots, X_{2n_2}$ duas a. a. independentes, de dimensão n_1 e n_2 retiradas de duas populações Bernoulli, com parâmetros p_1 e p_2, respetivamente, em que X_{ij} toma o valor 1 se for um sucesso e o valor 0 se for um insucesso, e

$$\overline{P}_1 = \sum_{i=1}^{n_1} \frac{X_{1i}}{n_1} \text{ e } \overline{P}_2 = \sum_{i=1}^{n_2} \frac{X_{2i}}{n_2}.$$

Se as dimensões das amostras forem grandes, então pelo Teorema de Moivre Laplace tem-se:

$$\overline{P}_1 \overset{\circ}{\sim} N\left(p_1; \sqrt{\frac{p_1(1-p_1)}{n_1}}\right) \quad \text{e} \quad \overline{P}_2 \overset{\circ}{\sim} N\left(p_2; \sqrt{\frac{p_2(1-p_2)}{n_2}}\right),$$

donde pelo Teorema da aditividade da distribuição Normal vem:

$$\overline{P}_1 - \overline{P}_2 \overset{\circ}{\sim} N\left(p_1 - p_2; \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right),$$

ou seja,

$$Z = \frac{\left(\overline{P}_1 - \overline{P}_2\right) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \overset{\circ}{\sim} N(0;1).$$

6.6 Distribuição da variância amostral

Seja $X_1, X_2, \ldots, X_n$, uma a. a. duma dada população com variância σ^2. A **variância amostral** é definida por:

$$S^2 = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n-1}.$$

Propriedades:

- $\mu_{S^2} = E(S^2) = \sigma^2$.
- Se a distribuição da população for Normal então $\sigma^2_{S^2} = Var(S^2) = \frac{2\sigma^4}{n-1}$.

Se a distribuição da população for Normal, com variância σ^2, então:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

6.7 Distribuição do quociente de variâncias amostrais

Sejam $X_{11}, X_{12}, \ldots, X_{1n_1}$ e $X_{21}, X_{22}, \ldots, X_{2n_2}$ duas a. a. independentes, de dimensão n_1 e n_2 retiradas de duas populações Normais, sendo $X_1 \sim N(\mu_1;\ \sigma_1)$ e $X_2 \sim N(\mu_2;\ \sigma_2)$, respetivamente, e

$$S_1^2 = \sum_{i=1}^{n_1} \frac{\left(X_{1i} - \overline{X}_1\right)^2}{n_1 - 1} \text{ e } S_2^2 = \sum_{i=1}^{n_2} \frac{\left(X_{2i} - \overline{X}_2\right)^2}{n_2 - 1},$$

então, por um teorema da distribuição F, tem-se que:

$$F = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F_{n_1-1;\, n_2-1}.$$

6.8 Quadros resumo

A Tabela 6.1 e a Tabela 6.2 apresentam algumas aproximações apenas para simplificar os conceitos e a sua utilização, mas tendo em consideração as ressalvas feitas anteriormente, que se sustentam na proximidade das distribuições t_n e $N(0;1)$, para valores elevados de n.

Tabela 6.1: *Quadro resumo das distribuições amostrais (1 população).*

Estatística	σ^2 conhecido?	Tipo de população	Distribuição por amostragem
$\overline{X}$	Sim	Normal (ou qualquer se n grande[†])	$Z = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} \sim N(0;1)$
	Não	Normal (ou qualquer se n grande[†])	$T = \dfrac{\overline{X} - \mu}{\dfrac{S}{\sqrt{n}}} \sim t_{n-1}$
$\overline{P}$ (n grande)	—	Bernoulli	$Z = \dfrac{\overline{P} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \overset{\circ}{\sim} N(0;1)$
S^2	—	Normal	$\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

Tabela 6.2: *Quadro resumo das distribuições amostrais (2 populações).*

Estatística	σ_1^2 e σ_2^2 conhecidos?	Tipo de populações	Distribuição por amostragem
$\overline{X}_1 - \overline{X}_2$	Sim	Normais (ou quaisquer se n_1 e n_2 grandes[†])	$Z = \dfrac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0;1).$
	Não ($\sigma_1^2 = \sigma_2^2$)	Normais (ou quaisquer se n_1 e n_2 grandes[†])	$T = \dfrac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n_1+n_2-2}$
	Não ($\sigma_1^2 \neq \sigma_2^2$)	Normais (ou quaisquer se n_1 e n_2 grandes[†])	$T = \dfrac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \overset{\circ}{\sim} t_v$, onde $v = \left[\dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{1}{n_1-1}\left(\dfrac{S_1^2}{n_1}\right)^2 + \dfrac{1}{n_2-1}\left(\dfrac{S_2^2}{n_2}\right)^2}\right]$
$\overline{P}_1 - \overline{P}_2$ (n_1 e n_2 grandes)	—	Bernoulli	$Z = \dfrac{\left(\overline{P}_1 - \overline{P}_2\right) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \overset{\circ}{\sim} N(0;1)$
$\dfrac{S_1^2}{S_2^2}$	—	Normais	$F = \dfrac{S_1^2}{S_2^2}\dfrac{\sigma_2^2}{\sigma_1^2} \sim F_{n_1-1;\, n_2-1}$

[†] Neste caso a distribuição por amostragem é aproximadamente $N(0;1)$.

6.9 Exercícios resolvidos

6.9.1 Teorema do Limite Central

1. Considere que o tempo de viagem de autocarro entre Lisboa e Évora segue uma distribuição Uniforme entre 100 a 120 minutos. Numa amostra aleatória de 30 viagens, qual a probabilidade de a média dos tempos de viagem ser inferior a 112 minutos?

Resolução:

Seja X_i a v.a. que representa o tempo da viagem i de autocarro entre Lisboa e Évora, com $X_i \sim U(100;\ 120)$. Logo,

$$\mu = E(X_i) = \frac{100+120}{2} = 110;$$

$$\sigma^2 = Var(X_i) = \frac{(120-100)^2}{12} = \frac{100}{3}.$$

Seja $\overline{X} = \frac{1}{30}(X_1 + \cdots + X_{30})$ a v.a. que representa a média dos 30 tempos de viagem de autocarro entre Lisboa e Évora.
Pelo corolário do T. L. C.,

$$Z = \frac{\overline{X} - 110}{\sqrt{\frac{100}{3\times 30}}} \overset{\circ}{\sim} N(0;1).$$

Logo,

$$P\left(\overline{X} < 112\right) \approx \mathrm{P}\left(Z < \frac{112-110}{\sqrt{\frac{100}{3\times 30}}}\right) = \Phi(1{,}80) = 0{,}9641.$$

2. Seja X a v. a. que representa o tempo decorrido entre a entrada consecutiva de 2 automóveis numa autoestrada (AE), que segue uma distribuição Exponencial com valor médio 15 segundos. Numa amostra aleatória de 100 entradas consecutivas, qual a probabilidade de a soma dos tempos ser superior a 1350 segundos?

Resolução:

Seja X_i a v.a. que representa o tempo decorrido entre a i-ésima entrada consecutiva de 2 automóveis numa auto-estrada (AE), com $X_i \sim\ Exp\left(\lambda = \frac{1}{15}\right)$. Logo,

$$\mu = E(X_i) = \frac{1}{\frac{1}{15}} = 15 \text{ segundos},$$

$$\sigma^2 = Var(X_i) = \frac{1}{\left(\frac{1}{15}\right)^2} = 15^2 = 225.$$

Seja $S_{100} = X_1 + \cdots + X_{100}$ a v.a. que representa a soma dos 100 decorrido entre a i-ésima entrada consecutiva de 2 automóveis numa autoestrada (AE).
Pelo T. L. C.,

$$Z = \frac{S_{100} - 100\times 15}{\sqrt{100\times 225}} \overset{\circ}{\sim} N(0;1).$$

Logo,

$$P(S_{100} > 1350) = 1 - P(S_{100} \leq 1350) \approx 1 - P\left(Z < \frac{1350 - 1500}{\sqrt{100 \times 225}}\right) = 1 - \Phi(-1) = \Phi(1) = 0{,}8413.$$

3. Com a ingestão de um determinado fármaco para o tratamento da depressão, 20% dos doentes referem sentir efeitos secundários durante a 1ª semana de tratamento. Seja X a v. a. que representa o número de doentes que sentem os referidos efeitos secundários numa amostra de n doentes. Numa amostra de 100 doentes, calcule a probabilidade de:

a) Pelo menos 15 terem sentido os efeitos secundários.
b) Mais de 16 e menos de 45 (inclusive) terem sentido efeitos secundários.

Resolução:

Sabe-se que $X \sim B(n = 100; p = 0{,}2)$.
Como $n > 50$ e $0{,}1 < p < 0{,}9$ podemos usar a aproximação à distribuição Normal e

$$X \overset{\circ}{\sim} N\left(\mu = np = 20;\ \sigma = \sqrt{npq} = 4\right).$$

a) $P(X \geq 15) = 1 - P(X < 15) \approx 1 - P\left(Z < \frac{15 - 0{,}5 - 20}{4}\right) = 1 - \Phi(-1{,}38) = \Phi(1{,}38) = 0{,}9162.$

b) $P(16 < X < 45) \approx P\left(\frac{16 + 0{,}5 - 20}{4} < Z < \frac{45 - 0{,}5 - 20}{4}\right) = P(-0{,}88 < Z < 6{,}13)$
$= \Phi(6{,}13) - \Phi(-0{,}88) = \Phi(6{,}13) - \left(1 - \Phi(0{,}88)\right) \approx 1 - (1 - 0{,}8106) = 0{,}8106.$

4. Seja X a v.a. que representa o número de automóveis que entram numa autoestrada (AE) num período de 30 segundos, com $X \sim P(\lambda = 9)$. Num período de 5 minutos, qual a probabilidade de:

a) Entrarem mais de 59 automóveis na AE?
b) Entrarem no mínimo 55 e no máximo 80 automóveis na AE?

Resolução:

Seja X' a v. a. que representa o número de automóveis que entram numa AE num período de 5 minutos (= 300 segundos = 30 × 10 segundos), com $X' \sim P(\lambda' = 10\lambda = 90)$.

Como $\lambda > 20$, podemos usar a aproximação à distribuição Normal e

$$X' \overset{\circ}{\sim} N\left(\mu = \lambda' = 90;\ \sigma = \sqrt{\lambda'} = \sqrt{90}\right).$$

a) $P(X' > 59) = 1 - P(X' \leq 59) \approx 1 - P\left(Z \leq \frac{59 + 0{,}5 - 90}{\sqrt{90}}\right) = 1 - P(Z \leq -3{,}21) = 1 - \Phi(-3{,}21)$
$= \Phi(3{,}21) = 0{,}9993.$

b) $P(55 \leq X' \leq 80) \approx P\left(\frac{55 - 0{,}5 - 90}{\sqrt{90}} \leq Z \leq \frac{80 + 0{,}5 - 90}{\sqrt{90}}\right) = P(-3{,}74 \leq Z \leq -1{,}00)$
$= \Phi(-1{,}00) - \Phi(-3{,}74) = \left(1 - \Phi(1{,}00)\right) - \left(1 - \Phi(3{,}74)\right) = \Phi(3{,}74) - \Phi(1{,}00)$
$\approx 0{,}9999 - 0{,}8413 = 0{,}1586.$

6.9.2 Distribuição da média amostral

6.9.2.1 Quando a variância é conhecida

Um fabricante de automóveis defende que o novo modelo que vai ser lançado no próximo mês gasta em média 9,7 litros aos 100 km, em circuito urbano, com desvio padrão de 1 litro. Admita que o consumo segue uma distribuição Normal.

a) Qual a probabilidade de numa amostra aleatória de 20 automóveis o gasto médio ser superior a 10 litros.
b) Qual a deverá ser a dimensão da amostra para obter, com pelo menos 90% probabilidade, um gasto médio inferior a 10 litros.

Resolução:

Seja X a v.a. que representa o número de litros consumidos pelo automóvel aos 100 km, em circuito urbano, com $X \sim N(\mu = 9{,}7;\ \sigma = 1)$.

a) $n = 20$.

$$\left.\begin{array}{l} X \text{ dist. Normal} \\ \sigma \text{ conhecido} \end{array}\right| \Rightarrow Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1).$$

$$P\left(\overline{X} > 10\right) = 1 - P\left(\overline{X} \leq 10\right) = 1 - P\left(Z \leq \frac{10 - 9{,}7}{\frac{1}{\sqrt{20}}}\right) = 1 - P(Z \leq 1{,}34) = 1 - \Phi(1{,}34)$$
$$= 1 - 0{,}9099 = 0{,}0901.$$

b) $n = ?$

$$P\left(\overline{X} < 10\right) \geq 0{,}9 \Leftrightarrow P\left(Z < \frac{10 - 9{,}7}{\frac{1}{\sqrt{n}}}\right) \geq 0{,}9 \Leftrightarrow \Phi\left(0{,}3\sqrt{n}\right) \geq 0{,}9$$

como $\Phi(1{,}282) = 0{,}9$

$$\Leftrightarrow 0{,}3\sqrt{n} \geq 1{,}282 \Leftrightarrow \sqrt{n} \geq 4{,}2733 \Rightarrow n \geq 4{,}2733^2 = 18{,}3 \Rightarrow n \geq 19.$$

6.9.2.2 Quando a variância é desconhecida

Um fabricante de automóveis defende que o novo modelo que vai ser lançado no próximo mês gasta em média 9,7 litros aos 100 km, em circuito urbano, e o desvio padrão é desconhecido.
Através de um esquema de amostragem estimou-se tal desvio padrão como sendo $s = 1$ litro. Admitindo que o consumo segue uma distribuição Normal, qual a probabilidade de, numa amostra aleatória de 20 automóveis, o consumo médio amostral ser superior a 10 litros? E inferior a 8,9 litros?

Resolução:

Seja X a v.a. que representa o número de litros consumidos pelo automóvel aos 100 km, em circuito urbano, com $X \sim N(\mu = 9{,}7;\ \sigma = ?)$.
$n = 20$.

$$\left.\begin{array}{l} X \text{ dist. Normal} \\ \sigma \text{ desconhecido} \end{array}\right| \Rightarrow T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1=19}.$$

$$P(\overline{X} > 10) = 1 - P(\overline{X} \leq 10) = 1 - P\left(T \leq \frac{10 - 9{,}7}{\frac{1}{\sqrt{20}}}\right) = 1 - P(T \leq 1{,}342) \approx 1 - 0{,}9 = 0{,}1.$$

$$P(\overline{X} < 8{,}9) = P\left(T < \frac{8{,}9 - 9{,}7}{\frac{1}{\sqrt{20}}}\right) = P(T < -3{,}578) = 1 - P(T < 3{,}578) \approx 1 - 0{,}999 = 0{,}001.$$

6.9.3 Distribuição da diferença de médias amostrais

6.9.3.1 Quando as variâncias são conhecidas

Uma dada empresa farmacêutica lançou no mercado um novo medicamento, para dormir, que tem estado a ser utilizado nos hospitais. Constatou-se que os doentes não sujeitos a este medicamento em média dormiam 7,5 horas, com desvio padrão de 1,4 horas, ao passo que os doentes aos quais se administrou este medicamento dormiam em média 8 horas com desvio padrão de 2 horas.
Num determinado hospital observaram-se 31 doentes não sujeitos ao referido medicamento e 61 sob a referida medicação. Qual a probabilidade de os doentes do primeiro grupo observado dormirem em média mais do que os do segundo grupo? Assuma a normalidade das distribuições.

Resolução:

Sejam:

- X_1 a v.a. que representa o número de horas de sono dos doentes não sujeitos ao medicamento,
- X_2 a v.a. que representa o número de horas de sono dos doentes sujeitos ao medicamento,

com

$$X_1 \sim N(\mu_1 = 7{,}5; \sigma_1 = 1{,}4) \text{ e } X_2 \sim N(\mu_2 = 8; \sigma_2 = 2).$$

$$\left.\begin{array}{l} X_1 \text{ e } X_2 \text{ dist. Normal} \\ \sigma_1\ (= 1{,}4) \text{ e } \sigma_2\ (= 2) \text{ conhecidos} \end{array}\right| \Rightarrow Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0; 1).$$

$$P(\overline{X}_1 > \overline{X}_2) = P(\overline{X}_1 - \overline{X}_2 > 0) = 1 - P(\overline{X}_1 - \overline{X}_2 \leq 0) = 1 - P\left(Z \leq \frac{0 - (7{,}5 - 8)}{\sqrt{\frac{1{,}4^2}{31} + \frac{2^2}{61}}}\right) = 1 - \Phi(1{,}39)$$

$$= 1 - 0{,}9177 = 0{,}0823.$$

6.9.3.2 Quando as variâncias são desconhecidas

Uma dada empresa farmacêutica lançou no mercado um novo medicamento, para dormir, que tem estado a ser utilizado nos hospitais. Constatou-se que os doentes não sujeitos a este medicamento em média dormiam 7,5 horas, enquanto os doentes aos quais se administrou este medicamento dormiam em média 8 horas.
Num determinado hospital observaram-se n_1 doentes não sujeitos ao referido medicamento e n_2 sob a referida medicação tendo-se obtido, respetivamente, os seguintes desvios-padrão: 1,4 horas e 2 horas. Determine a probabilidade de os doentes do primeiro grupo dormirem em média menos do que os do

segundo grupo, quando $n_1 = 20$ e $n_2 = 27$, quando se verifica a normalidade das distribuições e se considera:

a) A igualdade das variâncias populacionais.

b) A desigualdade das variâncias populacionais.

Resolução:

Sejam:

- X_1 a v.a. que representa o número de horas de sono dos doentes não sujeitos ao medicamento,
- X_2 a v.a. que representa o número de horas de sono dos doentes sujeitos ao medicamento,

Com $X_1 \sim N(\mu_1 = 7{,}5; \sigma_1 = ?)$ e $X_2 \sim N(\mu_2 = 8; \sigma_2 = ?)$.

$n_1 = 20;\ \ s_1 = 20;\ \ n_2 = 27$ e $s_2 = 2$.

a)

X_1 e X_2 dist. Normal
σ_1 e σ_2 desconhecidos, mas iguais

$$\Longrightarrow T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2=45}.$$

$$P(\overline{X}_1 < \overline{X}_2) = P(\overline{X}_1 - \overline{X}_2 < 0) = P\left(T < \frac{0-(7{,}5-8)}{\sqrt{\frac{(20-1)1{,}4^2 + (27-1)2^2}{20+27-2}}\sqrt{\frac{1}{20}+\frac{1}{27}}}\right) = P(T < 0{,}957)$$

$$\approx 0{,}83.$$

b)

X_1 e X_2 dist. Normal
σ_1 e σ_2 desconhecidos e diferentes

$$\Longrightarrow T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \overset{\circ}{\sim} t_{v=44},$$

pois

$$v = \left\lfloor \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \right\rfloor = \left\lfloor \frac{\left(\frac{1,4^2}{20}+\frac{2^2}{27}\right)^2}{\frac{1}{20-1}\left(\frac{1,4^2}{20}\right)^2 + \frac{1}{27-1}\left(\frac{2^2}{27}\right)^2} \right\rfloor = [44{,}9] = 44.$$

$$P(\overline{X}_1 < \overline{X}_2) = P(\overline{X}_1 - \overline{X}_2 < 0) = P\left(T < \frac{0-(7{,}5-8)}{\sqrt{\frac{1,4^2}{20}+\frac{2^2}{27}}}\right) = P(T < 1{,}008) \approx 0{,}84.$$

6.9.4 Distribuição da proporção amostral

Numa dada Repartição de Finanças, sabe-se que 70% dos contribuintes pagam o IUC (Imposto Único de Circulação) dentro do prazo.

a) Qual a probabilidade de numa amostra de 35 IUC, pelo menos 65% terem sido pagos dentro do prazo?

b) Para a probabilidade da alínea anterior ser no mínimo 80%, qual deve ser a dimensão mínima da amostra a recolher (admita que a amostra será de grande dimensão)?

Resolução:

Sejam:

- X_i a v. a. que designa se o contribuinte i paga o IUC dentro do prazo, $i = 1, \dots, n$,
- $\overline{P}$ a v.a. que representa a proporção de contribuintes que pagam o IUC dentro do prazo, em n contribuintes.

$p = 0{,}7$.

a)
$$\left.\begin{array}{l} X \text{ distribuição Bernoulli} \\ n\ (=35) \text{ grande} \end{array}\right| \Longrightarrow Z = \frac{\overline{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\circ}{\sim} N(0;1).$$

$$P\left(\overline{P} \geq 0{,}65\right) = 1 - P\left(\overline{P} < 0{,}65\right) = 1 - P\left(Z < \frac{0{,}65 - 0{,}7}{\sqrt{\frac{0{,}7 \times 0{,}3}{35}}}\right) = 1 - \Phi(-0{,}65) = \Phi(0{,}65) = 0{,}7422.$$

b) $n = ?$

$$P\left(\overline{P} \geq 0{,}65\right) \geq 0{,}8 \Leftrightarrow 1 - P\left(\overline{P} < 0{,}65\right) \geq 0{,}8 \Leftrightarrow P\left(Z < \frac{0{,}65 - 0{,}7}{\sqrt{\frac{0{,}7 \times 0{,}3}{n}}}\right) \leq 0{,}2 \Leftrightarrow \Phi\left(-0{,}11\sqrt{n}\right) \leq 0{,}2$$

$$\Leftrightarrow \Phi\left(0{,}11\sqrt{n}\right) \geq 0{,}8$$

Como $\Phi(0{,}84) \approx 0{,}8$

$$\Leftrightarrow 0{,}11\sqrt{n} \geq 0{,}84 \Leftrightarrow \sqrt{n} \geq 7{,}6364 \Longrightarrow n \geq 58{,}3 \Longrightarrow n \geq 59.$$

6.9.5 Distribuição da diferença de proporções amostrais

A proporção de clientes que optaram pela marca de telemóveis *Noko* na loja *TeleMN* foi 0,35 e na loja *Optcel* foi 0,29. Calcule a probabilidade de, recolhendo uma amostra de 200 clientes na primeira loja e de 150 clientes na segunda, a proporção amostral de clientes que optaram pela marca *Noko* na loja *TeleMN* ser superior à da loja *Optcel*.

Resolução:

Sejam:

- X_{1i} a v. a. que designa se o cliente i optou pela marca *Noko* na loja *TeleMN*, $i = 1, \dots, n_1$.
- X_{2i} a v. a. que designa se o cliente i optou pela marca *Noko* na loja *Optcel*, , $i = 1, \dots, n_2$.
- $\overline{P}_1$ a v. a. que representa a proporção de clientes que optaram pela marca *Noko* na loja *TeleMN*, em n_1 clientes
- $\overline{P}_2$ a v. a. que representa a proporção de clientes que optaram pela marca *Noko* na loja *Optcel*, em n_2 clientes.

$p_1 = 0{,}35$ e $p_2 = 0{,}29$.

$$\left.\begin{array}{l} X_1 \text{ e } X_2 \text{ dist. Bernoulli} \\ n_1\ (=200) \text{ e } n_2\ (=150) \text{ grandes} \end{array}\right| \Longrightarrow Z = \frac{\left(\overline{P}_1 - \overline{P}_2\right) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \overset{\circ}{\sim} N(0;1).$$

$$P\left(\overline{P}_1 > \overline{P}_2\right) = P\left(\overline{P}_1 - \overline{P}_2 > 0\right) = 1 - P\left(\overline{P}_1 - \overline{P}_2 \leq 0\right) = 1 - P\left(Z \leq \frac{0 - (0{,}35 - 0{,}29)}{\sqrt{\frac{0{,}35(1-0{,}35)}{200} + \frac{0{,}29(1-0{,}29)}{150}}}\right)$$

$$= 1 - \Phi(-1{,}2) = 1 - \big(1 - \Phi(1{,}2)\big) = \Phi(1{,}2) = 0{,}8849.$$

6.9.6 Distribuição da variância amostral

Uma determinada empresa farmacêutica lançou no mercado um novo medicamento, para dormir, que tem estado a ser utilizado nos hospitais. Constatou-se que os doentes sujeitos a este medicamento em média dormiam 8 horas, sendo o desvio padrão de 2 horas, e que a distribuição do número de horas de sono podia ser considerada Normal.
Qual a probabilidade de, numa amostra aleatória de 31 doentes sujeitos ao referido medicamento:

a) A variância amostral ser superior a 5 horas?
b) A variância amostral ser inferior a 2,25 horas?

Resolução:

Seja X a v.a. que representa o número de horas de sono dos doentes sujeitos ao medicamento, com $X \sim N(\mu = 8;\ \sigma = 2)$.

$$\left.\begin{array}{l} X \text{ dist. Normal} \\ n = 31 \end{array}\right| \Longrightarrow \chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1=30}.$$

a) $P(S^2 > 5) = 1 - P(S^2 \leq 5) = 1 - P\left(\chi^2 \leq \frac{(31-1)\times 5}{2^2}\right) = 1 - P(\chi^2 \leq 37{,}5) \approx 1 - 0{,}84 = 0{,}16.$

b) $P(S^2 < 2{,}25) = 1 - P\left(\chi^2 < \frac{(31-1)\times 2{,}25}{2^2}\right) = P(\chi^2 \leq 16{,}875) \approx 0{,}026.$

6.9.7 Distribuição do quociente de variâncias amostrais

Uma determinada empresa farmacêutica lançou no mercado um novo medicamento, para dormir, que tem estado a ser utilizado nos hospitais. Constatou-se que os doentes não sujeitos a este medicamento em média dormiam 7,5 horas sendo o desvio padrão de 1,4 horas, enquanto os doentes aos quais se administrou este medicamento dormiam em média 8 horas sendo o desvio padrão de 2 horas.
Num determinado hospital observaram-se 31 doentes não sujeitos ao referido medicamento e 61 sujeitos ao medicamento. Qual a probabilidade da variância amostral do primeiro grupo ser inferior à do segundo grupo? Admita a normalidade dos dados.

Resolução:

Sejam:

- X_1 a v.a. que representa o número de horas de sono dos doentes não sujeitos ao medicamento,
- X_2 a v.a. que representa o número de horas de sono dos doentes sujeitos ao medicamento,

com $X_1 \sim N(\mu_1 = 7{,}5;\ \sigma_1 = 1{,}4)$ e $X_2 \sim N(\mu_2 = 8;\ \sigma_2 = 2)$.

$$\left.\begin{array}{l} X_1 \text{ e } X_2 \text{ dist. Normal} \\ n_1 = 31 \text{ e } n_2 = 61 \end{array}\right| \Longrightarrow F = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F_{n_1-1;\, n_2-1\,=\,30;60}.$$

$$P(S_1^2 < S_2^2) = P\left(\frac{S_1^2}{S_2^2} < 1\right) = P\left(F < 1 \times \frac{2^2}{1{,}4^2}\right) = P(F < 2{,}04) \approx 0{,}99.$$

6.10 Exercícios propostos

1. O tempo de atendimento de um aluno na secção de fotocópias da *Casa das Folhas*, em Évora, é uma v. a. T com distribuição Exponencial de valor médio igual a 10 minutos. Numa amostra aleatória de 45 alunos, qual a probabilidade da média do tempo de atendimento por aluno ser inferior a 9 minutos?

2. Admita que o peso dos robalos criados em aquicultura aos 14 meses é descrito por uma distribuição Uniforme no intervalo 300 e 400 gr. Numa amostra aleatória de 56 robalos qual a probabilidade do peso total destes peixes ser superior a 20 kg?

3. Um inquérito conduzido há 5 anos revelou que 30% dos adultos eram consumidores regulares de bebidas alcoólicas. Supondo que o resultado se mantém atualmente, calcule a probabilidade aproximada de numa amostra aleatória de 1000 adultos o número de consumidores ser:
   a) Inferior a 280?
   b) Superior a 316?

4. Num estudo recentemente efetuado em várias maternidades verificou-se que, em média, 25% dos nascimentos ocorrem antes da data prevista pelo médico. Determine um valor aproximado da probabilidade de que pelo menos 35 de 100 grávidas, escolhidas ao acaso, ocorram antes da data prevista.

5. O número de avarias que uma máquina tem por dia é uma variável aleatória com distribuição Poisson de média 0,2. Calcule a probabilidade aproximada de a referida máquina ter durante um ano (365 dias) exatamente 75 avarias.

6. Numa determinada Universidade, a altura dos alunos do sexo masculino segue uma distribuição Normal com média 165 cm e desvio padrão 7,5 cm.
   a) Qual a probabilidade de, numa amostra aleatória de 10 alunos do sexo masculino, a altura média amostral:
      i. Ser no mínimo 170 cm?
      ii. Estar entre 150 cm e 160 cm?
   b) Se aumentar a dimensão da amostra da alínea anterior, o que espera que aconteça às probabilidades calculadas?
   c) Calcule a dimensão da amostra a recolher, de forma a que a média da amostra não difira da média populacional mais do que 1 cm com probabilidade superior a 0,9.
   d) Se na alínea anterior aumentar o valor da probabilidade o que acontece à dimensão da amostra? Justifique.
   e) Admita agora que σ é desconhecido e que o desvio padrão amostral é de 7,5 cm. Calcule a probabilidade pedida em a).
   f) Numa amostra de 20 alunos:
      i. Qual a probabilidade da variância amostral ser inferior a 49 cm?
      ii. A variância amostral é superior a ... com 1% de probabilidade.

7. Num determinado país a temperatura média diária em °C, registada nos meses de verão, pode-se considerar que segue uma distribuição Normal com média 30°C e desvio padrão 3°C.
   a) Numa amostra aleatória de 28 dias:
      i. Qual a probabilidade de a temperatura média destes dias ser inferior a 31°C?
      ii. Determine a probabilidade de a temperatura média destes dias estar entre os 29°C e os 32°C.
      iii. A probabilidade de, nestes dias, a temperatura média ser superior a ...°C é 90%.

b) Se aumentar a dimensão da amostra o que acontece à probabilidade calculada nas alíneas i) e ii)? Justifique.
c) Se diminuir a probabilidade da alínea iii) o que espera que aconteça à temperatura média?
d) Admita agora que σ é desconhecido e que o desvio padrão amostral é de 3°C. Resolva novamente as alíneas i) e ii).
e) Calcule a dimensão da amostra a recolher, para que a média amostral não difira da média populacional mais do que 0,5°C com probabilidade superior a 0,95.
f) Se na alínea anterior aumentar o valor da probabilidade o que acontece à dimensão da amostra? Justifique.

8. Num estudo realizado em 1990 pelo *National Center for Health Statistics*, 19% dos jovens com 18 anos afirmaram que não tinham ouvido falar do vírus HIV-SIDA.
a) Determine a probabilidade de, numa amostra aleatória com 175 jovens desta população, a percentagem de jovens que não tenham ouvido falar do vírus HIV-SIDA:
   i. Seja pelo menos 25%.
   ii. Seja no máximo 35%.
   iii. Esteja entre 30% e 40%.
b) Qual deve ser a dimensão da amostra de forma a obter, com 90% de probabilidade, no máximo 25% de jovens que não tenham ouvido falar deste vírus?

9. Num processo eleitoral, um determinado candidato obteve 46% dos votos. Determine a probabilidade de, numa secção de voto (com votantes ocasionais), ter havido uma maioria absoluta de votos a favor desse candidato, sabendo que o número de votantes foi:
a) 200.
b) 1000.
c) Comente a diferença obtida entre os resultados das duas alíneas anteriores.

10. Uma repartição de finanças tem dois funcionários a receber declarações de IRS, o Sr. Vagaroso e o Sr. Caracol. Sabe-se que o tempo de atendimento do Sr. Vagaroso é em média de 21 minutos e o do Sr. Caracol de 29 minutos, sendo admissível a normalidade dos dados.
Observaram-se aleatoriamente os tempos de atendimento de 16 utentes pelo Sr. Vagaroso e de 21 utentes pelo Sr. Caracol.
a) Qual a probabilidade do tempo médio amostral de atendimento do Sr. Caracol ser superior ao do Sr. Vagaroso sabendo que:
   i. Os desvios padrão populacionais são conhecidos: $\sigma_1 = 20$ min. e $\sigma_2 = 10$ min?
   ii. Os desvios padrão amostrais observados foram 20 e 10 minutos, respetivamente? (assuma a homogeneidade das variâncias populacionais)
b) Pronuncie-se quanto às diferenças obtidas na alínea a).
c) Admitindo que conhece a informação da alínea a) i), qual a probabilidade da variância amostral no tempo de atendimento do Sr. Caracol ser inferior à do Sr. Vagaroso?

11. Realizou-se um estudo exaustivo em dois aviários, *Nitrofrango* e *Frangosao*, tendo-se verificado que o tempo de desenvolvimento de um frango era em média de 38 dias no primeiro aviário e 42 dias no segundo aviário, sendo ainda admissível a normalidade das distribuições.
Recolheu-se uma amostra aleatória de 25 frangos no aviário *Nitrofrango* e de 20 frangos no aviário *Frangosao*.
a) Qual a probabilidade do tempo médio amostral de desenvolvimento dos frangos ser superior no aviário *Nitrofrango* do que no *Frangosao* sabendo que:

i. Os desvios padrão populacionais são conhecidos e iguais a 6 e 5 dias, respetivamente?
ii. Os desvios padrão amostrais observados foram 6 e 5 dias, respetivamente? (assuma a heterogeneidade das variâncias populacionais).

b) Pronuncie-se quanto às diferenças obtidas em a).

12. Num estudo comparativo do insucesso escolar entre alunos cujos pais são divorciados e alunos com vida familiar considerada como "regular", verificou-se uma taxa de insucesso escolar de 40% e 35%, respetivamente.
Selecionaram-se, ao acaso, dois grupos de estudantes tendo-se observado 35 alunos com pais divorciados e 80 estudantes de família "regular". Qual a probabilidade de a proporção de alunos com insucesso escolar ser maior no grupo dos estudantes com pais divorciados do que no grupo dos estudantes com uma família "regular"?

13. Numa determinada disciplina verifica-se que dos alunos que optam apenas pelo regime de frequências 45% são aprovados, e dos que optam pelo exame final 40% são aprovados.
Selecionaram-se ao acaso 45 alunos que optaram apenas pelo regime de frequências e 64 que optaram só pelo exame final. Qual a probabilidade de a proporção de alunos aprovados no exame final ser superior à proporção de alunos que optou pelo regime de frequências?

7 Estimação

Tal como foi definido na introdução, existe agora o objetivo adicional de caracterizar a população a partir da qual foi retirada a amostra, procurando, nomeadamente, estimar os parâmetros desta população.
Existem dois processos de estimação paramétrica:

- **Estimação pontual:** produção de um valor, que se pretende que seja o melhor, para um determinado parâmetro da população, com base na informação amostral;
- **Estimação intervalar:** construção de um intervalo que, com certo grau de certeza previamente estipulado, contenha o verdadeiro valor do parâmetro da população.

7.1 Estimação pontual

Definição: Um estimador dum parâmetro da população é uma variável aleatória (v. a.) que depende da informação amostral e cujas realizações fornecem aproximações para o parâmetro desconhecido. A um valor específico assumido por este estimador para uma amostra em concreto chama-se **estimativa**.

7.1.1 Propriedades dos estimadores

Principais propriedades desejáveis nos estimadores:

- *Não enviesamento* – em termos médios, o estimador atinge o valor real do parâmetro;
- *Eficiência* – o estimador é mais eficiente quanto menor for a sua variância;
- *Suficiência* – propriedade de retirar da amostra toda a informação relevante sobre o parâmetro;
- *Consistência* – para n grande, o estimador deve ser aproximadamente igual ao parâmetro.

Na Figura 7.1 e Figura 7.2 pretende-se ilustrar, através de um alvo, o significado das propriedades enviesamento, precisão e consistência.

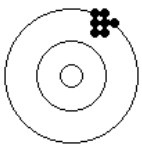
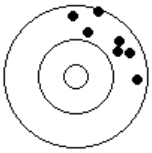
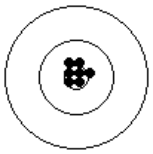
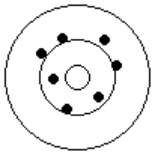
	Eficiente	Não Eficiente
Enviesado		
Não enviesado		

Figura 7.1: *Ilustração das propriedades de enviesamento e precisão.*
(Adaptado de http://techniques.geog.ox.ac.uk/mod_2/glossary/samp.html em 2/5/2005)

Definição: Um estimador $\hat{\theta}$ diz-se **não enviesado** ou **centrado** do parâmetro θ se:

$$E(\hat{\theta}) = \theta.$$

Definição: Seja $\hat{\theta}$ um estimador do parâmetro θ. O **enviesamento** (Env) de $\hat{\theta}$ é dado por:

$$Env(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

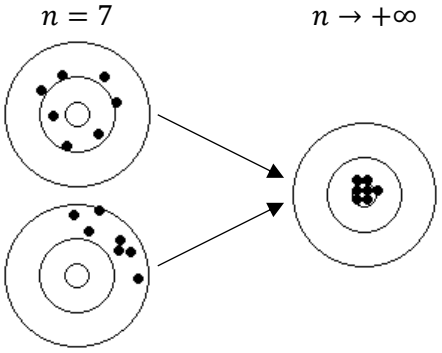


Figura 7.2: Ilustração da propriedade consistência.

Definição: Sejam $\hat{\theta}$ e $\tilde{\theta}$ dois estimadores centrados do parâmetro θ, baseados no mesmo número de observações. Então:

- $\hat{\theta}$ diz-se **mais eficiente** do que $\tilde{\theta}$ se: $Var(\hat{\theta}) < Var(\tilde{\theta})$
- A **eficiência relativa** do primeiro estimador relativamente ao segundo é dada por:

$$\text{Eficiência relativa} = \frac{Var(\hat{\theta})}{Var(\tilde{\theta})}.$$

Para comparar dois estimadores enviesados utiliza-se o critério do erro quadrático médio.

Definição: Seja $\hat{\theta}$ um estimador do parâmetro θ. O **erro quadrático médio** (EQM) de $\hat{\theta}$ é dado por:

$$EQM(\hat{\theta}) = Var(\hat{\theta}) + \left(Env(\hat{\theta})\right)^2.$$

Observação: Se $\hat{\theta}$ um estimador centrado para o parâmetro θ, então $Env(\hat{\theta}) = 0$ e, por conseguinte, $EQM(\hat{\theta}) = Var(\hat{\theta})$.

Definição: Sejam $\hat{\theta}$ e $\tilde{\theta}$ dois estimadores. Diz-se que $\hat{\theta}$ é "melhor" que $\tilde{\theta}$ se:

$$EQM(\hat{\theta}) < EQM(\tilde{\theta}).$$

Definição: Um estimador $\hat{\theta}$ diz-se **suficiente** para parâmetro θ se distribuição condicional da amostra $(X_1, X_2, \dots, X_n)$, dado o valor observado $\hat{\theta} = t$, não depende de θ.

Um estimador diz-se suficiente se extrai da amostra toda a informação que esta contém sobre o parâmetro θ de tal maneira que, dado o valor observado $\hat{\theta} = t$, o conhecimento dos valores observados para os elementos da amostra nada acrescenta sobre θ. Os estimadores suficientes gozam da propriedade de retirar da amostra toda a informação relevante sobre o parâmetro (Murteira *et al.*, 2007).

Definição: Um estimador $\hat{\theta}$ diz-se **consistente** quando para qualquer valor positivo δ, se verifica a condição:

$$\lim_{n\to\infty} P\left[|\hat{\theta} - \theta| < \delta\right] = 1.$$

Teorema: As condições:

- $\lim_{n\to+\infty} E(\hat{\theta}) = \theta$;
- $\lim_{n\to+\infty} Var(\hat{\theta}) = 0$.

são suficientes para que $\hat{\theta}$ seja estimador **consistente**.

7.1.2 Métodos de estimação

De entre os métodos de estimação pontuais mais usuais destacam-se:

- *Método dos momentos* – os estimadores obtêm-se por substituição dos momentos da amostra nas expressões que representam os momentos na população. Em condições muito gerais são (Murteira *et al.*, 2007), os estimadores obtidos são consistentes e possuem distribuição Normal se a amostra é muito grande.
- *Método dos mínimos quadrados* – são usualmente utilizados no âmbito da regressão linear.
- *Método da máxima verosimilhança* – é provavelmente o método mais importante. Geralmente, os estimadores de máxima verosimilhança gozam das propriedades desejáveis num bom estimador: são os mais eficientes, e consistentes. Embora, usualmente, não sejam centrados costumam ser assimptoticamente não enviesados. Além disso, possuem distribuição assimptótica Normal e gozam da propriedade da invariância, i.e., "Se $g(\theta)$ é função biunívoca de θ, então $g(\hat{\theta})$ é estimador de máxima verosimilhança de $g(\theta)$" (Murteira *et al.*, 2007).

Uma forma de obter uma estimativa pontual dum parâmetro da população (por exemplo: μ, σ, σ^2 e p) é retirar uma amostra aleatória (a. a.) representativa dessa população e calcular o valor da estatística correspondente (por exemplo: $\overline{x}$, s, s^2 e $\overline{p}$).

7.1.2.1 Método dos momentos

Este método de estimação é um dos mais simples e mais antigo para obter estimadores de um ou mais parâmetros de uma distribuição. A ideia base é utilizar os momentos da amostra para estimar os correspondentes momentos da população, e, a partir daí, estimar os parâmetros de interesse (Murteira *et al.*, 2007). Seja $X_1, X_2, \dots, X_n$ uma a. a. duma dada população com função (densidade) de probabilidade, f. (d.) p., $f(x\,;\theta_1, \theta_2, \dots, \theta_k)$ que depende de k parâmetros. Admitindo que existem os momentos ordinários, μ'_r, da população X, estes são função dos k parâmetros,

$$\mu'_r = E(X^r) \begin{cases} \sum_{i=1}^{N} x^r f(x\,;\theta_1, \theta_2, \dots, \theta_k)\,, & \text{para distribuições discretas;} \\ \int_{-\infty}^{+\infty} x^r f(x\,;\theta_1, \theta_2, \dots, \theta_k)dx\,, & \text{para distribuições contínuas.} \end{cases}$$

Os correspondentes momentos amostrais são dados por

$$m'_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r.$$

O **método dos momentos** consiste em considerar que os estimadores dos momentos ordinários são dados pelos momentos ordinários amostrais, ou seja,

$$\hat{\mu}'_r = m'_r, i = 1, \dots, k.$$

7.1.2.2 Método da máxima verosimilhança

Este método só pode ser aplicado se a distribuição da população for conhecida.

Definição: Seja $X_1, X_2, \dots, X_n$ uma a. a. duma dada população com função (densidade) de probabilidade, f. (d.) p., $f(x\,; \theta_1, \theta_2, \dots, \theta_k) = f(x;\ \boldsymbol{\theta})$. Então a f. (d.) p. conjunta das variáveis que constituem a amostra é dada por:

$$f(x_1, x_2, \dots, x_n;\ \boldsymbol{\theta}) = f(x_1;\ \boldsymbol{\theta}) f(x_2;\ \boldsymbol{\theta}) \dots f(x_n;\ \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_n;\ \boldsymbol{\theta}).$$

Considerando uma amostra em concreto, designa-se por **função de verosimilhança** a função de $\boldsymbol{\theta}$ e da amostra tal que:

$$\mathcal{L}(.) = \mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}; x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f(x_n;\ \boldsymbol{\theta}).$$

O **método da máxima verosimilhança** consiste em encontrar o estimador $\widehat{\boldsymbol{\theta}}$ que maximiza o valor da função de verosimilhança para uma determinada amostra, ou seja, o valor de $\boldsymbol{\theta}$ que torna aquela amostra concreta mais provável, i. e., mais verosímil.

Frequentemente, o estimador o estimador de máxima verosimilhança pode ser por derivação, ou seja:

1. Determinar a função de verosimilhança $\mathcal{L}(\boldsymbol{\theta})$;
2. Se necessário, aplicar a transformação logarítmica à função de verosimilhança, $\ln\big(\mathcal{L}(\boldsymbol{\theta})\big)$. Geralmente, esta transformação torna o problema da maximização mais simples.
3. Determinar os pontos onde a 1ª derivada da função $\mathcal{L}(\boldsymbol{\theta})$, ou $\ln\big(\mathcal{L}(\boldsymbol{\theta})\big)$, em ordem a cada um dos θ_i se anula (condição de 1ª ordem):

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i} = 0 \ \text{ ou } \ \frac{\partial \ln\big(\mathcal{L}(\boldsymbol{\theta})\big)}{\partial \theta_i} = 0.$$

4. Verificar se a 2ª derivada da função $\mathcal{L}(\boldsymbol{\theta})$, ou $\ln\big(\mathcal{L}(\boldsymbol{\theta})\big)$ em ordem a θ_i é negativa (condição de 2ª ordem):

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_i^2} < 0 \ \text{ ou } \ \frac{\partial^2 \ln\big(\mathcal{L}(\boldsymbol{\theta})\big)}{\partial \theta_i^2} < 0.$$

7.1.3 Exercícios resolvidos

1. Suponha que uma variável aleatória X representa o número de avarias de um dispositivo durante um período de tempo e que obedece a uma lei de Poisson de parâmetro λ desconhecido. Para este parâmetro foram sugeridos dois estimadores:

$$\hat{\lambda} = \frac{X_1 + \dots + X_n}{n} \ \text{ e } \ \tilde{\lambda} = \frac{X_1 + X_n}{2}.$$

a) Compare-os quanto ao enviesamento.
b) Deduza a variância para cada um deles.
c) Qual dos dois estimadores é mais eficiente? Justifique a sua escolha.

d) Estude os dois estimadores quanto à consistência.

Resolução:

Se $X \sim P(\lambda)$, então $E(X) = Var(X) = \lambda$.
Se $X_1, X_2, \dots, X_n$ é uma a. a. duma dada população $X \sim P(\lambda)$, então $X_i \sim P(\lambda), i = 1, \dots, n$.

a) $E(\hat{\lambda}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n}E(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\big(E(X_1) + E(X_2) + \cdots + E(X_n)\big)$

$= \frac{1}{n}(\lambda + \lambda + \cdots + \lambda) = \frac{1}{n}n\lambda = \lambda.$

$E(\tilde{\lambda}) = E\left(\frac{X_1 + X_n}{2}\right) = \frac{1}{2}\big(E(X_1) + E(X_n)\big) = \frac{1}{2}(\lambda + \lambda) = \lambda.$

Portanto, ambos os estimadores são centrados.

$b)\ Var(\hat{\lambda}) = Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)$

$= \frac{1}{n}Var(X_1 + X_2 + \cdots + X_n) \underset{\substack{\overline{X_i} \\ indep.}}{=} \frac{1}{n^2}(Var[X_1] + Var[X_2] + \cdots + Var[X_n])$

$= \frac{1}{n^2}(\lambda + \lambda + \cdots + \lambda) = \frac{1}{n^2}n\lambda = \frac{\lambda}{n}.$

$Var(\tilde{\lambda}) = Var\left(\frac{X_1 + X_n}{2}\right) \underset{\substack{\overline{X_i} \\ indep.}}{=} \frac{1}{2^2}\big(Var(X_1) + Var(X_n)\big) = \frac{1}{4}(\lambda + \lambda) = \frac{\lambda}{2}.$

c) Se $n = 1,\ Var(\hat{\lambda}) > Var(\tilde{\lambda}) \rightarrow \tilde{\lambda}$ é mais eficiente do que $\hat{\lambda}$;
Se $n = 2,\ Var(\hat{\lambda}) = Var(\tilde{\lambda}) \rightarrow \hat{\lambda}$ é tão eficiente como $\tilde{\lambda}$;
Se $n > 2,\ Var(\hat{\lambda}) < Var(\tilde{\lambda}) \rightarrow \tilde{\lambda}$ é mais eficiente do que $\hat{\lambda}$.

d) $\hat{\lambda}$ é um estimador consistente, pois $E(\hat{\lambda}) = \lambda$ e $Var(\hat{\lambda}) = \frac{\lambda}{n} \rightarrow 0$, quando $n \rightarrow \infty$.
$\tilde{\lambda}$ não é um estimador consistente: $E(\tilde{\lambda}) = \lambda$ mas $Var(\tilde{\lambda}) = \frac{\lambda}{2}$, seja qual for o valor de n.

2. Seja $X_1, X_2, \dots, X_n$ uma a. a. de uma distribuição Normal, $X \sim N(\mu;\ \sigma)$. Estime os parâmetros μ e σ pelo método:

a) Dos momentos.
b) Da máxima verosimilhança.

Resolução:

a) Sabe-se que:
- $\mu_1' = m_1' = E(X) = \mu$;
- $\mu_2' = m_2' = E(X^2) = Var(X) + \big(E(X)\big)^2 = \sigma^2 + \mu^2$.

Para obter os estimadores $\hat{\mu}$ e $\hat{\sigma}^2$ pelo método dos momentos, é preciso resolver o seguinte sistema de equações:

$$\begin{cases} \mu_1' = m_1' \\ \\ \mu_2' = m_2' \end{cases} \Leftrightarrow \begin{cases} \mu_1' = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \\ \mu_2' = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{cases} \Leftrightarrow \begin{cases} \mu = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \\ \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{cases} \Leftrightarrow \begin{cases} \mu = \overline{x} \\ \\ \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \end{cases}.$$

Portanto, os estimadores são:

$$\begin{cases} \hat{\mu} = \overline{X} \\ \hat{\sigma}^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \left(\dfrac{n-1}{n}\right) S^2 \end{cases}.$$

b) Função densidade de probabilidade (f. d. p.):

$$f(x;\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < \mu < +\infty, \qquad \sigma > 0.$$

Função de verosimilhança:

$$\mathcal{L}(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i\,; \mu\,, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma^2}\right)^2} = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$
$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Logaritmo da função de verosimilhança:

$$\ln\left(\mathcal{L}(\mu, \sigma^2)\right) = -\frac{n}{2}\left(\ln(2) + \ln(\pi) + \ln(\sigma^2)\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Condições de 1ª ordem:

$$\begin{cases} \dfrac{\partial \ln L\,(\mu, \sigma^2)}{\partial \mu} = 0 \\ \dfrac{\partial \ln L\,(\mu, \sigma^2)}{\partial \sigma^2} = 0 \end{cases} \Leftrightarrow \begin{cases} -\dfrac{1}{2\sigma^2}\left(-2\displaystyle\sum_{i=1}^{n} x_i + 2n\mu\right) = 0 \\ -\dfrac{n}{2}\dfrac{1}{\sigma^2} + \displaystyle\sum_{i=1}^{n}(x_i - \mu)^2 \dfrac{2}{4\sigma^4} = 0 \end{cases} \Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} x_i - n\mu = 0 \\ -n\sigma^2 + \displaystyle\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \mu = \displaystyle\sum_{i=1}^{n} \frac{x_i}{n} \\ \sigma^2 = \displaystyle\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n} \end{cases} \Leftrightarrow \begin{cases} \mu = \overline{x} \\ \sigma^2 = \displaystyle\sum_{i=1}^{n} \frac{(x_i - \overline{x})}{n} \end{cases} \Leftrightarrow \begin{cases} \mu = \overline{x} \\ \sigma^2 = \left(\dfrac{n-1}{n}\right) s^2 \end{cases}$$

Condições de 2ª ordem:

$$\begin{cases} \dfrac{\partial^2 \ln L\,(\mu, \sigma^2)}{\partial \mu^2} = -\dfrac{1}{2\sigma^2} 2n < 0 \\ \dfrac{\partial^2 \ln L\,(\mu, \sigma^2)}{\partial \sigma^4} = \dfrac{n}{2}\dfrac{1}{\sigma^4} - \displaystyle\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^6} < 0 \end{cases},$$

pois $n > 0, \sigma^2 > 0, \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^6} > \frac{n}{2}\frac{1}{\sigma^4}$.

Portanto, os estimadores de máxima verosimilhança obtidos foram:

$$\begin{cases} \hat{\mu} = \overline{X} = \displaystyle\sum_{i=1}^{n} \frac{X_i}{n} \\ \hat{\sigma}^2 = \displaystyle\sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n} = \frac{n-1}{n} S^2 \end{cases}.$$

3. Considere uma população com distribuição de Bernoulli, com parâmetro p, com $0 < p < 1$.
   a) Derive o estimador de máxima verosimilhança para o parâmetro p.
   b) Foi obtida uma amostra de dimensão $n = 3$, cujos valores observados foram (1, 1, 0).
      i. Esboce o gráfico da função de verosimilhança e interprete-o.
      ii. Forneça uma estimativa para p com base no método da máxima verosimilhança.

Resolução:

a) Função de probabilidade (f. p.):

$$f(x;p) = p^x(1-p)^{1-x}, \qquad x = 0, 1, \text{com } 0 < p < 1.$$

Função de verosimilhança:

$$\mathcal{L}(p) = \prod_{i=1}^{n} f(x_i;p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

Logaritmo da função de verosimilhança:

$$\ln\big(\mathcal{L}(p)\big) = \sum_{i=1}^{n} x_i \ln p + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p).$$

Condição de 1ª ordem:

$$\frac{\partial \ln\big(\mathcal{L}(p)\big)}{\partial p} = 0 \Leftrightarrow \sum_{i=1}^{n} x_i\left(\frac{1}{p}\right) + \left(n - \sum_{i=1}^{n} x_i\right)\left(\frac{-1}{1-p}\right) = 0 \Leftrightarrow (1-p)\sum_{i=1}^{n} x_i - p\left(n - \sum_{i=1}^{n} x_i\right) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} x_i - pn = 0 \Leftrightarrow p = \sum_{i=1}^{n} \frac{x_i}{n}.$$

Condição de 2ª ordem:

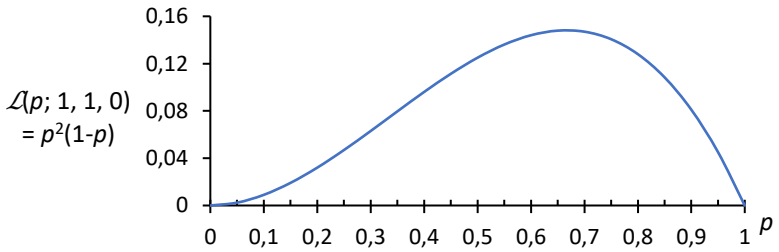
$$\frac{\partial^2 \ln L\,(p\,;x)}{\partial p^2} = -\frac{\sum_{i=1}^n x_i}{p^2} + \frac{(n - \sum_{i=1}^n x_i)(-1)}{(1-p)^2} = -\frac{\sum_{i=1}^n x_i}{p^2} - \frac{(n - \sum_{i=1}^n x_i)}{(1-p)^2} < 0,$$

pois $x_i \geq 0, p^2 > 0, n > 0, (1-p)^2 > 0$ e $n \geq \sum_{i=1}^n x_i$ pois $x_i = 0$ ou 1.
Portanto, o estimador de máxima verosimilhança é:

$$\hat{p} = \sum_{i=1}^{n} \frac{X_i}{n}.$$

b) i) Substituindo em $\mathcal{L}(p)$, (x_1, x_2, x_3) pelos valores observados na amostra, $(1, 1, 0)$, obtemos,

$$\mathcal{L}(p) = p^2(1-p).$$



A função de verosimilhança atinge o seu valor máximo quando o p se situa perto de 0,65, sendo este o valor de p mais provável que deu origem à observação desta amostra.

ii) $\hat{p} = \dfrac{2}{3} = 0{,}6667$

7.2 Estimação intervalar

Na estimação por intervalos, em vez de se propor apenas um valor concreto para certo parâmetro da população, constrói-se um intervalo de valores $(w_1;\ w_2)$ que, com um certo grau de certeza, previamente estipulado, contenha o verdadeiro valor do parâmetro.

Em muitos casos, o intervalo é da forma $\left(\hat{\theta}-\varepsilon;\ \hat{\theta}+\varepsilon\right)$, sendo $\hat{\theta}$ uma estimativa para o parâmetro de interesse θ, e ε é considerado uma medida de precisão ou medida do erro inerente à estimativa $\hat{\theta}$. Usualmente, ε é designado por erro de estimativa ou margem de erro (absoluta). Desta forma, este método de estimação incorpora a confiança que se pode atribuir às estimativas.

Definição: Seja θ um parâmetro desconhecido. Suponha-se que com base na informação amostral, pode-se encontrar as estatísticas $W_1(X_1, X_2, \dots, X_n)$ e $W_2(X_1, X_2, \dots, X_n)$, sendo $W_1(x_1, x_2, \dots, x_n) < W_2(x_1, x_2, \dots, x_n)$, tais que:

$$P(W_1(X_1, X_2, \dots, X_n) < \theta < W_2(X_1, X_2, \dots, X_n)) = 1-\alpha, \qquad 0 < \alpha < 1.$$

Sejam w_1 e w_2 realizações amostrais específicas de $W_1(X_1, X_2, \dots, X_n)$ e $W_2(X_1, X_2, \dots, X_n)$. Então o intervalo $(w_1; w_2)$ é chamado o **intervalo de confiança** (I. C.) a $100(1-\alpha)\%$ para θ, designando-se $100(1-\alpha)\%$ o **grau de confiança** do intervalo e α o **nível de significância**.

Se forem retiradas repetidamente um elevado número de amostras da população, o parâmetro θ deve estar contido em $100(1-\alpha)\%$ dos intervalos calculados da forma atrás descrita. Portanto, α representa o risco de o parâmetro procurado não estar no I. C. calculado.

O **erro de estimativa** corresponde ao erro máximo que, com a confiança especificada, se pode cometer na estimativa de θ. Nos I. C. centrados em torno da estimativa $\hat{\theta}$, o erro de estimativa corresponde à semiamplitude do I. C.

A amplitude do I. C. pode ser reduzida se:

- Aumentar a dimensão da amostra (n);
- Mantendo a dimensão da amostra, se diminuir o grau de confiança $(1-\alpha)$.

7.2.1 Método da variável fulcral

A especificação de um I. C. para um dado parâmetro θ implica o conhecimento simultâneo de:

- Um estimador para o parâmetro em causa;
- A distribuição de amostragem do estimador;
- Uma estimativa pontual para o parâmetro em causa.

Definição: Seja $X_1, X_2, \dots, X_n$ uma a. a. duma dada população com função (densidade) de probabilidade $f(x;\ \theta)$. Diz-se que a função das observações e do parâmetro de interesse θ, $W(x_1, x_2, \dots, x_n;\ \theta)$, é uma **variável fulcral** se a respetiva função (densidade) de probabilidade é conhecida e é independente de θ.

Obtenção do Intervalo de Confiança:

Escolher a variável fulcral, W, adequada para estimar o parâmetro pretendido;

1. Fixar o grau de confiança $(1-\alpha)$;
2. Procurar dois números no domínio de W, que por facilidade e conveniência se consideram ser os quantis de probabilidade $\frac{\alpha}{2}$ e $1-\frac{\alpha}{2}$, representados por $w_{\frac{\alpha}{2}}$ e $w_{1-\frac{\alpha}{2}}$:

   i. $P\left(W \leq w_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$

   ii. $P\left(W \geq w_{1-\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$ $\quad \Rightarrow P\left(w_{\frac{\alpha}{2}} < W < w_{1-\frac{\alpha}{2}}\right) = 1-\alpha$

   Observação: As condições i) e ii) permitem obter intervalos de amplitude mínima quando a distribuição de W é simétrica. Nos restantes casos torna a obtenção dos intervalos mais simples e conduz a intervalos com amplitude próxima da mínima.

3. Resolver a desigualdade $w_{\frac{\alpha}{2}} < W < w_{1-\frac{\alpha}{2}}$ de forma a obter,

$$W_1(X_1, X_2, \dots, X_n) < \theta < W_2(X_1, X_2, \dots, X_n).$$

4. Com base numa amostra concreta obter as realizações

$$w_1 = W_1(x_1, x_2, \dots, x_n) \text{ e } w_2 = W_2(x_1, x_2, \dots, x_n)$$

   de $W_1(X_1, X_2, \dots, X_n)$ e $W_2(X_1, X_2, \dots, X_n)$.

5. $(w_1;\ w_2)$ é o I. C. para θ com $100(1-\alpha)\%$ de confiança.

7.2.2 Intervalos de confiança para a média

Seja $X_1, X_2, \dots, X_n$ uma a. a. de dimensão n retirada de uma população com média μ e desvio padrão σ.

7.2.2.1 Quando a variância é conhecida

Se a distribuição da população é Normal e a variância σ^2 é conhecida, então a variável fulcral a utilizar na construção do intervalo de confiança é:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1).$$

Se a distribuição da população não for Normal, mas a amostra for de grande dimensão, então $Z \overset{\circ}{\sim} N(0;1)$. Em ambos os casos o I. C., de amplitude mínima, para μ é obtido da seguinte forma:

$$P\left(z_{\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow P\left(\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha,$$

como a distribuição é simétrica em torno de 0, tem-se que $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$, logo

$$\Leftrightarrow P\left(\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha,$$

sendo $z_{1-\frac{\alpha}{2}}$ o quantil de probabilidade $1-\frac{\alpha}{2}$ da distribuição $N(0;1)$.

Portanto, quando a população é Normal com variância conhecida, ou a população não é normal mas a variância é conhecida e a amostra é grande, o **I. C. para $\boldsymbol{\mu}$ com $\mathbf{100(1-\alpha)\%}$ de confiança** é dado por:

$$\left]\overline{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}};\ \overline{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right[.$$

7.2.2.2 Quando a variância é desconhecida

Se a distribuição da população é Normal com desvio padrão σ desconhecido, então a variável fulcral é:

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}.$$

O I. C. é obtido resolvendo:

$$P\left(t_{n-1;\,\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} < t_{n-1;\,1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\Leftrightarrow P\left(\overline{X} - t_{n-1;\,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < \mu < \overline{X} - t_{n-1;\,\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1 - \alpha,$$

como a distribuição é simétrica em torno de 0, tem-se que $t_{n-1;\,\frac{\alpha}{2}} = -t_{n-1;\,1-\frac{\alpha}{2}}$, logo

$$\Leftrightarrow P\left(\overline{X} - t_{n-1;\,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{n-1;\,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1 - \alpha,$$

sendo $t_{n-1,1-\frac{\alpha}{2}}$ o quantil de probabilidade $1-\frac{\alpha}{2}$ da distribuição t_{n-1}.

Portanto, quando a população é Normal com variância desconhecida, o **I. C. para $\boldsymbol{\mu}$ com $\mathbf{100(1-\alpha)\%}$ de confiança**, é dado por:

$$\left]\overline{X} - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}};\ \overline{X} + t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right[.$$

Se a distribuição da população não for Normal, mas a amostra for de grande dimensão então a variável fulcral é:

$$Z = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \overset{\circ}{\sim} N(0;1).$$

Quando a população não é normal mas a amostra é grande e a variância é desconhecida, o **I. C. aproximado para $\boldsymbol{\mu}$ com $\mathbf{100(1-\alpha)\%}$ de confiança**, é dado por:

$$\left]\overline{X} - z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}};\ \overline{X} + z_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right[.$$

Observação: Tal como já foi referido anteriormente no capítulo 6 (Distribuições amostrais), dada a proximidade entre as distribuições *t*-Student e $N(0;1)$, alguns programas estatísticos (por exemplo, SPSS) permitem apenas construir I. C. baseados na distribuição *t*-Student, quando a variância populacional é

desconhecida. Apesar de não ser teoricamente correto, não traz consequências práticas e simplifica a sua aplicação.

7.2.3 Intervalos de confiança para a diferença de médias

Sejam $X_{11}, X_{12}, \dots, X_{1n_1}$ e $X_{21}, X_{22}, \dots, X_{2n_2}$ duas a. a. independentes, de dimensão n_1 e n_2, retiradas de duas populações com médias μ_1 e μ_2 e desvios padrão σ_1 e σ_2, respetivamente.

7.2.3.1 Quando as variâncias são conhecidas

Se as populações forem Normais com desvios padrão σ_1 e σ_2 conhecidos, a variável fulcral a utilizar na construção do I. C. é:

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0;1).$$

Se as populações não forem Normais, mas as amostras forem de grande dimensão então $Z \overset{\circ}{\sim} N(0;1)$.
O I. C., de amplitude mínima, é obtido resolvendo:

$$P\left(z_{\frac{\alpha}{2}} < \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{1-\frac{\alpha}{2}} \right) = 1 - \alpha$$

$$\Leftrightarrow P\left(\left(\overline{X}_1 - \overline{X}_2\right) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \left(\overline{X}_1 - \overline{X}_2\right) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) = 1 - \alpha$$

como a distribuição é simétrica em torno de 0, tem-se que $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$, logo

$$\Leftrightarrow P\left(\left(\overline{X}_1 - \overline{X}_2\right) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \left(\overline{X}_1 - \overline{X}_2\right) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) = 1 - \alpha,$$

sendo $z_{1-\frac{\alpha}{2}}$ o quantil de probabilidade $1 - \frac{\alpha}{2}$ da distribuição $N(0;1)$.

Portanto, quando as populações são Normais com variâncias conhecidas, o **I. C. para $\boldsymbol{\mu_1 - \mu_2}$ com $\boldsymbol{100(1-\alpha)\%}$ de confiança** é dado por:

$$\left]\left(\overline{X}_1 - \overline{X}_2\right) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}};\ \left(\overline{X}_1 - \overline{X}_2\right) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right[$$

7.2.3.2 Quando as variâncias são desconhecidas e iguais

Se as populações forem Normais e σ_1 e σ_2 forem desconhecidos mas onde, para um determinado nível de confiança, a igualdade pode ser considerada ($\sigma_1 = \sigma_2$)†, então a variável fulcral é:

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

Neste caso o I. C., de amplitude mínima, é obtido resolvendo:

$$P\left(t_{n_1+n_2-2;\frac{\alpha}{2}} < \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < t_{n_1+n_2-2;\,1-\frac{\alpha}{2}}\right) = 1-\alpha,$$

considerando $S^* = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

$$\Leftrightarrow P\left(\left(\overline{X}_1 - \overline{X}_2\right) - t_{n_1+n_2-2;\,1-\frac{\alpha}{2}}\,S^* < \mu_1 - \mu_2 < \left(\overline{X}_1 - \overline{X}_2\right) - t_{n_1+n_2-2;\frac{\alpha}{2}}\,S^*\right) = 1-\alpha,$$

como a distribuição é simétrica em torno de 0, tem-se que $t_{n_1+n_2-2;\frac{\alpha}{2}} = -t_{n_1+n_2-2;\,1-\frac{\alpha}{2}}$, logo

$$\Leftrightarrow P\left(\left(\overline{X}_1 - \overline{X}_2\right) - t_{n_1+n_2-2;\,1-\frac{\alpha}{2}}S^* < \mu_1 - \mu_2 < \left(\overline{X}_1 - \overline{X}_2\right) + t_{n_1+n_2-2;\,1-\frac{\alpha}{2}}S^*\right) = 1-\alpha,$$

sendo $t_{n_1+n_2-2;\,1-\frac{\alpha}{2}}$ o quantil de probabilidade $1-\frac{\alpha}{2}$ da distribuição $t_{n_1+n_2-2}$.

Portanto, quando as populações são Normais com variâncias desconhecidas, mas iguais, o **I. C. para $\boldsymbol{\mu_1 - \mu_2}$ com $\boldsymbol{100(1-\alpha)\%}$ de confiança** é dado por:

$$\left]\left(\overline{X}_1 - \overline{X}_2\right) - t_{n_1+n_2-2;\,1-\frac{\alpha}{2}}S^*;\left(\overline{X}_1 - \overline{X}_2\right) + t_{n_1+n_2-2;\,1-\frac{\alpha}{2}}S^*\right[.$$

Se as populações não forem Normais, mas as amostras forem de grande dimensão então a variável fulcral anterior segue aproximadamente uma $N(0;1)$, e nesse caso no I. C. anterior em vez de $t_{n_1+n_2-2;\,1-\frac{\alpha}{2}}$ deve ser $z_{1-\frac{\alpha}{2}}$. Tal como foi referido anteriormente em situação análoga, alguns programas estatísticos apenas permitem construir o I. C. baseado na distribuição *t*-Student.

† Adiante abordar-se o intervalo de confiança para a razão de variâncias que permite avaliar o pressuposto de igualdade das variâncias.

7.2.3.3 Quando as variâncias são desconhecidas e diferentes

Se as populações forem Normais com σ_1 e σ_2 desconhecidos mas onde, para um determinado nível de confiança, existe evidência que são diferentes ($\sigma_1 \neq \sigma_2$)[†], então a variável fulcral é:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{\circ}{\sim} t_v, \text{ onde } v = \left[\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2} \right].$$

sendo $[r]$ a parte inteira de r, ou seja, arredondando-se por defeito o valor obtido.
Neste caso o I. C., de amplitude mínima, é obtido resolvendo:

$$P\left(t_{v;\frac{\alpha}{2}} < \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} < t_{v;1-\frac{\alpha}{2}} \right) = 1 - \alpha$$

$$\Leftrightarrow P\left((\overline{X}_1 - \overline{X}_2) - t_{v;1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) - t_{v;\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right) = 1 - \alpha,$$

como a distribuição é simétrica em torno de 0, tem-se que $t_{v;\frac{\alpha}{2}} = -t_{v;1-\frac{\alpha}{2}}$, logo

$$\Leftrightarrow P\left((\overline{X}_1 - \overline{X}_2) - t_{v;1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + t_{v;1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right) = 1 - \alpha,$$

sendo $t_{v;1-\frac{\alpha}{2}}$ o quantil de probabilidade $1 - \frac{\alpha}{2}$ da distribuição t_v.

Portanto, quando as populações são Normais com variâncias desconhecidas, mas iguais, o **I. C. para $\boldsymbol{\mu_1 - \mu_2}$ com $\boldsymbol{100(1-\alpha)\%}$ de confiança** é dado por:

$$\left] (\overline{X}_1 - \overline{X}_2) - t_{v;1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}; (\overline{X}_1 - \overline{X}_2) + t_{v;1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right[.$$

Se as populações não forem Normais, mas as amostras forem de grande dimensão então a variável fulcral anterior segue aproximadamente uma $N(0;1)$, e nesse caso no I. C. anterior em vez de $t_{v;1-\frac{\alpha}{2}}$ deve ser $z_{1-\frac{\alpha}{2}}$.

Tal como já foi referido anteriormente em situações análogas, alguns programas estatísticos apenas permitem construir o I. C. baseado na distribuição *t*-Student.

[†] Adiante abordar-se o intervalo de confiança para a razão de variâncias que permite avaliar o pressuposto de igualdade das variâncias.

7.2.4 Intervalos de confiança para a proporção

Seja $X_1, X_2, \dots, X_n$ uma a. a. de dimensão n retirada de uma população Bernoulli, com parâmetro p. Se a dimensão da amostra for suficientemente grande então, a variável fulcral a utilizar é:

$$Z = \frac{\overline{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{\circ}{\sim} N(0;1).$$

O I. C., de amplitude mínima, é obtido resolvendo:

$$P\left(z_{\frac{\alpha}{2}} < \frac{\overline{P} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{1-\frac{\alpha}{2}} \right) = 1 - \alpha$$

$$P\left(\overline{P} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} < p < \overline{P} - z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right) = 1 - \alpha,$$

como a distribuição é simétrica em torno de 0, tem-se que $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$, logo

$$\Leftrightarrow P\left(\overline{P} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} < p < \overline{P} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right) = 1 - \alpha,$$

sendo $z_{1-\frac{\alpha}{2}}$ o quantil de probabilidade $1 - \frac{\alpha}{2}$ da distribuição $N(0;1)$.

Uma vez que os limites de confiança dependem do parâmetro desconhecido p, para amostras de grande dimensão este pode ser substituído pelo seu estimador $\overline{P}$.

Portanto, quando a amostra é grande, o **I. C. para p com $100(1-\alpha)\%$ de confiança** é:

$$\left] \overline{P} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\overline{P}(1-\overline{P})}{n}};\ \overline{P} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\overline{P}(1-\overline{P})}{n}} \right[.$$

7.2.5 Intervalos de confiança para a diferença de proporções

Sejam $X_{11}, X_{12}, \dots, X_{1n_1}$ e $X_{21}, X_{22}, \dots, X_{2n_2}$ duas a. a. independentes, de dimensão n_1 e n_2 retiradas de duas populações Bernoulli, com parâmetros p_1 e p_2, respetivamente. Se as dimensões das amostras forem grandes, então a variável fulcral é:

$$Z = \frac{(\overline{P}_1 - \overline{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \overset{\circ}{\sim} N(0;1).$$

O I. C., de amplitude mínima, é obtido resolvendo:

$$P\left(z_{\frac{\alpha}{2}} < \frac{(\overline{P}_1 - \overline{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} < z_{1-\frac{\alpha}{2}} \right) = 1 - \alpha,$$

considerando

$$P^* = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2},$$

$$\Leftrightarrow P\left(\left(\overline{P}_1 - \overline{P}_2\right) - z_{1-\frac{\alpha}{2}}\sqrt{P^*} < p_1 - p_2 < \left(\overline{P}_1 - \overline{P}_2\right) - z_{\frac{\alpha}{2}}\sqrt{P^*}\right) = 1-\alpha,$$

como a distribuição é simétrica em torno de 0, tem-se que $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$, logo

$$\Leftrightarrow P\left(\left(\overline{P}_1 - \overline{P}_2\right) - z_{1-\frac{\alpha}{2}}\sqrt{P^*} < p_1 - p_2 < \left(\overline{P}_1 - \overline{P}_2\right) + z_{1-\frac{\alpha}{2}}\sqrt{P^*}\right) = 1-\alpha,$$

sendo $z_{1-\frac{\alpha}{2}}$ o quantil de probabilidade $1-\frac{\alpha}{2}$ da distribuição $N(0;1)$.

Uma vez que os limites de confiança dependem dos parâmetros desconhecidos p_1 e p_2, para amostras de grande dimensão estes podem ser substituídos pelos respetivos estimadores $\overline{P}_1$ e $\overline{P}_2$.

Portanto, quando as amostras são grandes, o **I. C. para $\boldsymbol{p_1 - p_2}$ com $\boldsymbol{100(1-\alpha)\%}$ de confiança** é dado por:

$$\left]\overline{P}_1 - \overline{P}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{P}_1(1-\overline{P}_1)}{n_1} + \frac{\overline{P}_2(1-\overline{P}_2)}{n_2}};\ \overline{P}_1 - \overline{P}_2 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{P}_1(1-\overline{P}_1)}{n_1} + \frac{\overline{P}_2(1-\overline{P}_2)}{n_2}}\right[.$$

7.2.6 Intervalos de confiança para a variância

Seja $X_1, X_2, \ldots, X_n$ uma a. a. de dimensão n retirada de uma população Normal com média μ e desvio padrão σ. A variável fulcral é:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

O I. C. é construído da seguinte forma:

$$P\left(\chi^2_{n-1;\frac{\alpha}{2}} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1;\,1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) = 1-\alpha,$$

sendo $\chi^2_{n-1,\frac{\alpha}{2}}$ e $\chi^2_{n-1,1-\frac{\alpha}{2}}$ os quantis de probabilidade $\frac{\alpha}{2}$ e $1-\frac{\alpha}{2}$, respetivamente, da distribuição χ^2_{n-1}.

Portanto, quando a população é Normal, o **I. C. para $\boldsymbol{\sigma^2}$ com $\boldsymbol{100(1-\alpha)\%}$ de confiança** é dado por:

$$\left]\frac{(n-1)S^2}{\chi^2_{n-1;\,1-\frac{\alpha}{2}}};\frac{(n-1)S^2}{\chi^2_{n-1;\frac{\alpha}{2}}}\right[.$$

7.2.7 Intervalos de confiança para a razão de variâncias

Sejam $X_{11}, X_{12}, \dots, X_{1n_1}$ e $X_{21}, X_{22}, \dots, X_{2n_2}$ duas a. a. independentes, de dimensão n_1 e n_2 retiradas de duas populações Normais com médias μ_1 e μ_2 e desvios padrão σ_1 e σ_2, respetivamente. A variável fulcral é:

$$F = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F_{n_1-1;\, n_2-1}.$$

O I. C. pode ser obtido da seguinte forma:

$$P\left(f_{n_1-1,n_2-1;\frac{\alpha}{2}} < \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} < f_{n_1-1,n_2-1\,;\,1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\frac{S_1^2}{S_2^2}\frac{1}{f_{n_1-1,n_2-1;1-\frac{\alpha}{2}}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{f_{n_1-1,n_2-1;\frac{\alpha}{2}}}\right) = 1-\alpha$$

Sendo $f_{n_1-1,n_2-1;\frac{\alpha}{2}}$ e $f_{n_1-1,n_2-1;1-\frac{\alpha}{2}}$ os quantis de probabilidade $\frac{\alpha}{2}$ e $1-\frac{\alpha}{2}$, respetivamente, da distribuição $F_{n_1-1;\, n_2-1}$.

Quando as populações são Normais, o **I. C. para $\frac{\sigma_1^2}{\sigma_2^2}$ com $100(1-\alpha)\%$ de confiança** é dado por:

$$\left]\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,1-\frac{\alpha}{2}}}; \frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,\frac{\alpha}{2}}}\right[.$$

Como $F_{n_1-1,n_2-1;\frac{\alpha}{2}} = \frac{1}{F_{n_2-1,n_1-1\,;\,1-\frac{\alpha}{2}}}$ este intervalo pode ser escrito da seguinte forma:

$$\left]\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1\,;\,1-\frac{\alpha}{2}}}; \frac{S_1^2}{S_2^2}F_{n_2-1,n_1-1\,;\,1-\frac{\alpha}{2}}\right[.$$

7.2.8 Intervalos para amostras emparelhadas

Considere-se agora o caso em que as duas amostras formam um par de observações $(X_{1i};\, X_{2i})$, $i = 1, \dots, n$, ou seja, trata-se de uma amostra emparelhada de dimensão n. Os n pares de observações são independentes e retirados de populações Normais com médias μ_1 e μ_2 e desvios padrão σ_1 e σ_2, respetivamente.
Para construir o intervalo de confiança pretendido calcular:

- $D_i = X_{1i} - X_{2i}, \qquad i = 1, \dots, n;$
- $\overline{D} = \overline{X}_1 - \overline{X}_2 = \sum_{i=1}^{n}\frac{X_{1i}}{n} - \sum_{i=1}^{n}\frac{Y_{2i}}{n} = \sum_{i=1}^{n}\frac{D_i}{n};$
- $S_D^2 = \sum_{i=1}^{n}\frac{(D_i - \bar{D})^2}{n-1}.$

As variáveis aleatórias D_i, $i = 1, \dots, n$, são independentes. Portanto, está-se no caso univariado, onde se podem aplicar os resultados anteriores sobre os Intervalos de confiança para a média (ver secção 7.2.2) e Intervalos de confiança para a variância (ver secção 7.2.6).

7.2.9 Intervalo de confiança para o coeficiente de correlação populacional

Seja ρ o coeficiente de correlação de Pearson de uma população Normal bivariada. A distribuição do **coeficiente de correlação linear de Pearson amostral**, $\boldsymbol{R}$, não é Normal, pois depende não só do sinal como também da magnitude do coeficiente. Quando o coeficiente se afasta de zero, a distribuição é muito assimétrica, positiva ou negativa conforme o sinal do coeficiente. Quando o coeficiente se aproxima de zero, a distribuição é simétrica, e para grandes amostras a distribuição é aproximadamente Normal.
Para ultrapassar esta dificuldade, Fisher demonstrou que a transformação do coeficiente de correlação R em Z_R produziria uma variável normalmente distribuída com média zero:

$$Z_R = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$$

O desvio padrão de Z_R é dado por

$$S_{Z_R} = \frac{1}{\sqrt{n-3}},$$

onde n é a dimensão da amostra. Então, a variável fulcral é:

$$Z = \frac{Z_R - Z_\rho}{\frac{1}{\sqrt{n-3}}} \sim N(0;1).$$

Portanto, o **I. C. para $\boldsymbol{Z_\rho}$ com $\boldsymbol{100(1-\alpha)\%}$ de confiança** é dado por:

$$\left]Z_R - z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}; Z_R + z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}\right[, \text{ com } Z_R = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right).$$

Para transformar o intervalo obtido para Z_ρ num intervalo para ρ transformam-se os limites do intervalo de Z_ρ através da expressão

$$\rho = \frac{e^{2Z_\rho}-1}{e^{2Z_\rho}+1}.$$

O **I. C. para $\boldsymbol{\rho}$ com $\boldsymbol{100(1-\alpha)\%}$ de confiança** é dado por:

$$\left]\frac{e^{2Z_{\rho_{inf}}}-1}{e^{2Z_{\rho_{inf}}}+1}; \frac{e^{2Z_{\rho_{sup}}}-1}{e^{2Z_{\rho_{sup}}}+1}\right[,$$

onde $Z_{\rho_{inf}} = Z_R - z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}$; $Z_{\rho_{sup}} = Z_R + z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}$ e $Z_R = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$.

De salientar que o I. C. para Z_ρ está centrado em torno de Z_R, enquanto que o I. C. resultante para ρ não está centrado em torno de R.

7.2.10 Quadros resumo

A Tabela 7.1 e a **Tabela 7.2** apresentam algumas aproximações apenas para simplificar os conceitos e a sua utilização, mas tendo em consideração as ressalvas feitas anteriormente e que se sustentam na proximidade das distribuições t_n e $N(0;1)$, para valores elevados de n.

Tabela 7.1: *Quadro resumo dos intervalos de confiança (1 população).*

Parâmetro	σ^2 conhecido?	Tipo de população	Intervalo de confiança $(1-\alpha)100\%$
μ	Sim	Normal (ou qualquer se n grande)	$\left]\overline{X} \pm z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right[$
	Não	Normal (ou qualquer se n grande[†])	$\left]\overline{X} \pm t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right[$
p (n grande)	—	Bernoulli	$\left]\overline{P} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}}\right[$
σ^2	—	Normal	$\left]\frac{(n-1)S^2}{\chi^2_{n-1\,;\,1-\frac{\alpha}{2}}};\frac{(n-1)S^2}{\chi^2_{n-1\,;\,\frac{\alpha}{2}}}\right[$

Tabela 7.2: *Quadro resumo dos intervalos de confiança (2 populações).*

Parâmetro	σ_1^2 e σ_2^2 conhecidos?	Tipo de populações	Intervalo de Confiança $(1-\alpha)100\%$
$\mu_1-\mu_2$	Sim	Normais (ou quaisquer se n_1 e n_2 grandes)	$\left]\left(\overline{X}_1-\overline{X}_2\right) \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right[$
	Não ($\sigma_1^2=\sigma_2^2$)	Normais (ou quaisquer se n_1 e n_2 grandes[†])	$\left]\left(\overline{X}_1-\overline{X}_2\right) \pm t_{n_1+n_2-2\,;\,1-\frac{\alpha}{2}}S^*\right[$ com $S^*=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
	Não ($\sigma_1^2\neq\sigma_2^2$)	Normais (ou quaisquer se n_1 e n_2 grandes[†])	$\left]\left(\overline{X}_1-\overline{X}_2\right) \pm t_{v\,;\,1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right[$ com $v=\left\lfloor\frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}\right\rfloor$
p_1-p_2 (n_1 e n_2 grandes)	—	Bernoulli	$\left]\left(\overline{P}_1-\overline{P}_2\right) \pm z_{1-\frac{\alpha}{2}}\sqrt{\overline{P}^*}\right[$ com $\overline{P}^*=\frac{\overline{P}_1(1-\overline{P}_1)}{n_1}+\frac{\overline{P}_2(1-\overline{P}_2)}{n_2}$

[†] Neste caso em vez da distribuição *t*-Student deve ser usado o quantil da distribuição $N(0;1)$.

Erro! A origem da referência não foi encontrada. *(continuação)*

Parâmetro	σ_1^2 e σ_2^2 conhecidos?	Tipo de populações	Intervalo de Confiança $(1-\alpha)100\%$
$\frac{\sigma_2^2}{\sigma_1^2}$	—	Normais	$\left]\frac{S_1^2}{S_2^2}\frac{1}{f_{n_1-1,n_2-1;\,1-\frac{\alpha}{2}}};\frac{S_1^2}{S_2^2}f_{n_2-1,n_1-1;\,1-\frac{\alpha}{2}}\right[$
ρ	—	Normais	$\left]\frac{e^{2Z_{\rho_{inf}}}-1}{e^{2Z_{\rho_{inf}}}+1};\frac{e^{2Z_{\rho_{sup}}}-1}{e^{2Z_{\rho_{sup}}}+1}\right[$, com $Z_{\rho_{inf}}=Z_R-z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}$; $Z_{\rho_{sup}}=Z_R+z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}$; $Z_R=\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$.

7.2.11 Exercícios resolvidos

7.2.11.1 Intervalo de confiança para a média

7.2.11.1.1 Quando a variância é conhecida

Um certo nutricionista concebeu um novo programa dietético para as pessoas que têm uma vida profissional muito sedentária. Com este novo programa ele afirma que a perda média de peso, ao fim de um mês, é de 6,1 kg.
Para testar a veracidade de tal afirmação, aplicou-se este programa a 50 telefonistas tendo-se verificado uma perda média de peso de 3,4 kg.
Admitindo que a perda de peso segue uma distribuição Normal com um desvio padrão de 1,8 kg:

a) Com 95% de confiança, concorda com a afirmação do nutricionista relativamente à perda média de peso?
b) Sem efetuar cálculos, qual a sua opinião, se considerar um grau de confiança de 90%? E se o grau de confiança fosse de 99%?

Resolução:

Seja X a v.a. que representa a perda de peso, em kg, ao fim de um mês com o programa dietético, com $X \sim N(\mu = ?;\ \sigma = 1{,}8)$.

$n = 50$ e $\overline{x} = 3{,}4$.

Afirmação do nutricionista: $\mu = 6{,}1$.

a) I. C a 95% para μ é dado por:

$$\left]\overline{X} - z_{0{,}975}\frac{\sigma}{\sqrt{n}};\ \overline{X} + z_{0{,}975}\frac{\sigma}{\sqrt{n}}\right[.$$

Substituindo pelos valores conhecidos, sendo $z_{0{,}975} = 1{,}96$, obtém-se

$$\left]3{,}4 - 1{,}96\frac{1{,}8}{\sqrt{50}};\ 3{,}4 + 1{,}96\frac{1{,}8}{\sqrt{50}}\right[=]2{,}901;\ 3{,}899[.$$

Portanto, face aos resultados obtidos (6,1 não está contido no I. C. a 95%), não se pode concordar com a afirmação do nutricionista, pois com 95% de probabilidade a perda média de peso situa-se entre 2,09 kg e 3,90 kg.

b) Quando se diminui o grau de confiança do intervalo, mantendo os restantes fatores constantes, a amplitude do intervalo diminui. Portanto, o intervalo de confiança a 90% estaria contido no I. C. a 95%, donde a opinião seria a mesma pelos mesmos motivos.
Se pelo contrário aumentasse o grau de confiança, mantendo os restantes fatores constantes, a amplitude do I. C. também aumentaria, pelo que seria necessário efetuar os cálculos para verificar se com esse aumento o I. C. passaria a englobar o valor 6,1, e só depois se poderia dar a opinião.

7.2.11.1.2 Quando a variância é desconhecida

Junto do rio A situa-se uma empresa fabricante de transformadores elétricos. Na sua laboração é usado um agente químico (PCB) altamente danoso quando libertado para o meio ambiente. Um organismo fiscalizador do ambiente dispõe de duas técnicas para determinar os níveis de PCB nos peixes dos rios. Uma amostra de 10 peixes capturados no rio A foi analisada por uma das técnicas, tendo-se obtido as seguintes concentrações (em ppm) de PCB:

11,5 10,8 11,6 9,4 12,4 11,4 12,2 11,0 10,6 10,8

Construa um intervalo de confiança a 99% para o nível médio de concentração de PCB nos peixes capturados no rio A.

Resolução:

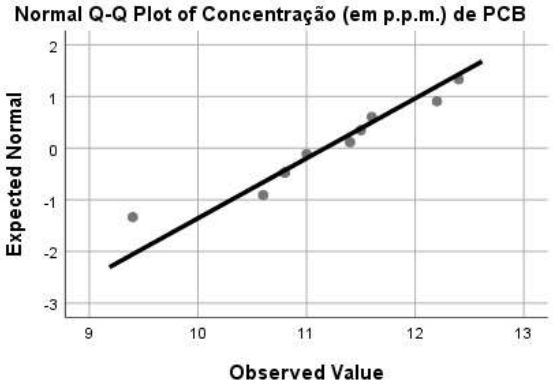
Seja X a v.a. que representa as concentrações (em ppm.) de PCB.
$n = 10, \quad \overline{x} = 11{,}17$ e $s = 0{,}8616$

Sabemos que se $X \sim N(\mu;\ \sigma)$, o I. C a 99% para μ é dado por:

$$\left]\overline{X} - t_{n-1;0,995}\frac{S}{\sqrt{n}};\ \overline{X} + t_{n-1;0,995}\frac{S}{\sqrt{n}}\right[.$$

Como nada é referido sobre a distribuição de X, vamos recorrer ao gráfico quantil-quantil para testar se os dados são provenientes de uma distribuição Normal.

(SPSS) Analyze → Descriptive Statistics → Explore…
(Dependent list: PCB; Display: ⊙ Plots;
Plots… → ☑ Normality plots with tests)



Uma vez que os pontos se posicionam, aproximadamente, sobre a reta, podemos considerar que os dados são provenientes de uma população com distribuição Normal e assim usar a expressão do IC acima.
Substituindo pelos valores conhecidos, sendo $t_{9;0,995} = 3{,}25$, obtém-se

$$\left]11{,}17 - 3{,}25\frac{0{,}8616}{\sqrt{10}};\ 11{,}17 + 3{,}25\frac{0{,}8616}{\sqrt{10}}\right[=]10{,}285; 12{,}055[.$$

Portanto, com 99% de confiança a concentração média (em ppm) de PCB situa-se entre 10,285 e 12,055 ppm.

(SPSS) Analyse → Compare Means → One-Sample T Test…
(Teste Variable: PCB; Options → Confidence Interval: 99)

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Concentração (em p.p.m.) de PCB	10	11,1700	0,8616	0,2725

One-Sample Test

	Test Value = 0					
					99% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Concentração (em p.p.m.) de PCB	40,997	9	,000	11,1700	10,2846	12,0554

7.2.11.2 Intervalo de confiança para a diferença de médias

7.2.11.2.1 Quando as variâncias são conhecidas

Havendo indícios de que o esquema de avaliação e as classificações finais atribuídas diferem fortemente entre duas escolas, decidiu-se comprovar estatisticamente esta hipótese. Os desvios-padrão são conhecidos sendo 2,1 valores na escola A e 1,8 valores na escola B. Assim, retirou-se uma amostra de testes de alunos em cada uma das escolas que levaram aos seguintes resultados:

Escola	n_i	$\overline{x}_i$
A	41	12,9
B	31	14,7

Recorrendo a um intervalo de confiança a 90%, diga se há diferenças entre as classificações médias das escolas A e B. Justifique.

Resolução:

Sejam:

- X_1 a v.a. que representa a classificação final dos alunos na escola A,
- X_2 a v.a. que representa a classificação final dos alunos na escola B,

com $\sigma_1 = 2{,}1$ e $\sigma_2 = 1{,}8$.

Como as amostras são grandes, o I. C. para $\mu_1 - \mu_2$ a 90% é dado por:

$$\left]\left(\overline{X}_1 - \overline{X}_2\right) - z_{0{,}95}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}};\left(\overline{X}_1 - \overline{X}_2\right) + z_{0{,}95}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right[.$$

Substituindo pelos valores conhecidos, sendo $z_{0{,}95} = 1{,}645$, obtém-se

$$\left](12{,}9 - 14{,}7) \pm 1{,}645\sqrt{\frac{2{,}1^2}{41} + \frac{1{,}8^2}{31}}\right[=]-2{,}558;\ -1{,}043[.$$

Com 90% de confiança, existe evidência de que as classificações médias são diferentes (0 não está no intervalo). Como ambos os limites do intervalo são negativos então significa que $\mu_1 < \mu_2$, ou seja, a classificação média é superior na escola B do que na escola A[†].

7.2.11.2.2 Quando as variâncias são desconhecidas e iguais

Um determinado método de análise permite determinar o conteúdo de enxofre no petróleo bruto. Os ensaios efetuados em 10 e 8 amostras aleatórias de 1 kg de petróleo bruto, provenientes de furos pertencentes respetivamente aos campos A e B, revelaram os seguintes resultados (em gramas):

Campo A:	111	114	105	112	107	109	112	110	110	106
Campo B:	109	103	101	105	106	108	110	104		

Construa um intervalo de confiança a 90% para a diferença entre os valores esperados da quantidade de enxofre por quilograma de petróleo proveniente de cada campo, considerando que populações são Normais, com variâncias desconhecidas mas iguais.

Resolução:

Sejam:

- X_1 a v.a. que representa a quantidade de enxofre por quilograma de petróleo do campo A,
- X_2 a v.a. que representa a quantidade de enxofre por quilograma de petróleo do campo B,

com $X_1 \sim N(\mu_1 =?\,; \sigma_1 =?)$ e $X_2 \sim N(\mu_2 = ?; \sigma_2 =?)$, mas $\sigma_1 = \sigma_2$.

$n_1 = 10, \quad \overline{x}_1 = 109{,}6 \quad \text{e} \quad s_1 = 2{,}875,$

$n_2 = 8, \quad \overline{x}_2 = 105{,}75 \quad \text{e} \quad s_2 = 3{,}105.$

O I. C. para $\mu_1 - \mu_2$ a 90% é dado por:

$$\left]\left(\overline{X}_1 - \overline{X}_2\right) - t_{n_1+n_2-2;\,0{,}95}S^*\,;\left(\overline{X}_1 - \overline{X}_2\right) + t_{n_1+n_2-2;\,0{,}95}S^*\right[,$$

$$\text{com } S^* = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

Substituindo pelos valores conhecidos,

$$s^* = \sqrt{\frac{(10-1)2{,}875^2 + (8-1)3{,}105^2}{10+8-2}}\sqrt{\frac{1}{10}+\frac{1}{8}} = 1{,}413$$

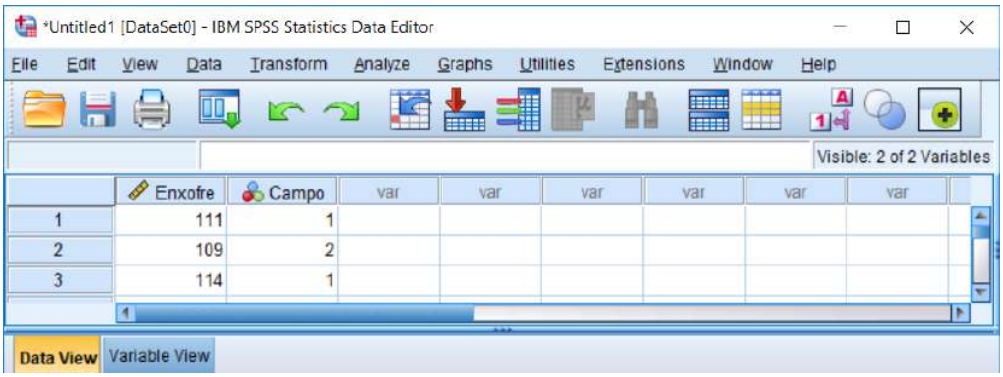
e como $t_{16\,;\,0{,}95} = 1{,}746$, obtém-se

$$\left](109{,}6 - 105{,}75) \pm 1{,}746 \times 1{,}413;\,[= \right]1{,}384;\; 6{,}316[.$$

Com 90% de confiança, existe evidência de que o teor médio de enxofre nos campos A e B é diferente (0 não está no intervalo). Uma vez que ambos os limites são positivos, então significa que $\mu_1 > \mu_2$, ou seja, o conteúdo médio de enxofre por quilograma de petróleo extraído do campo A é superior ao registado no campo B[†].

[†] De notar que o grau de confiança estabelecido apenas é válido para a igualdade e não para a desigualdade (< ou >). Esta observação é válida para interpretações em análises futuras similares.

(SPSS)



(SPSS) Analyze → Compare Means → Independent-Samples T Test...
(Test Variable: Enxofre; Grouping Variable: Campo;
Define Groups → Use specified values → Group 1: 1; Group 2: 2;
Options → Confidence Interval: 90)

Group Statistics

	Campo	N	Mean	Std. Deviation	Std. Error Mean
Enxofre por kg de petróleo	A	10	109,60	2,875	,909
	B	8	105,75	3,105	1,098

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									90% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Enxofre por kg de petróleo	Equal variances assumed	,086	,773	2,725	16	,015	3,850	1,413	1,384	6,316
	Equal variances not assumed			2,701	14,6	,017	3,850	1,425	1,346	6,354

Pela análise do *output* fornecido pelo SPSS verifica-se que não existe evidência da diferença das variâncias populacionais (Teste de Levene[†] para a igualdade das variâncias, $valor\ p = 0{,}773$). Desta forma o intervalo de confiança lê-se na primeira linha de resultados do quadro.

7.2.11.2.3 Quando as variâncias são desconhecidas e diferentes

Para um estudo sobre a caracterização da altura da população portuguesa, foi recolhida uma amostra de 1861 pessoas, com as seguintes características:

Group Statistics

	Sexo	N	Mean	Std. Deviation
Altura	Masculino	853	168,46	7,617
	Feminino	1007	158,48	6,652

Supondo a normalidade das distribuições e assumindo que as variâncias populacionais são desconhecidas e diferentes, verifique se se pode considerar que as alturas médias dos homens e das mulheres são iguais, com 95% de confiança.

[†] Ver testes de hipóteses, capítulo 8.

Resolução:

Sejam:

- X_1 a v.a. que representa a altura dos indivíduos do sexo masculino,
- X_2 a v.a. que representa a altura dos indivíduos do sexo feminino,

com $X_1 \sim N(\mu_1 =?\,; \sigma_1 =?)$ e $X_2 \sim N(\mu_2 = ?; \sigma_2 =?)$, mas $\sigma_1^2 \neq \sigma_2^2$.

$n_1 = 853, \quad \overline{x}_1 = 168{,}46 \quad$ e $\quad s_1 = 7{,}617,$
$n_2 = 1007, \quad \overline{x}_2 = 158{,}48 \quad$ e $\quad s_2 = 6{,}652.$

O I. C. para $\mu_1 - \mu_2$ a 95% é dado por:

$$\left]\left(\overline{X}_1 - \overline{X}_2\right) - t_{v;\,0{,}975}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}; \left(\overline{X}_1 - \overline{X}_2\right) + t_{v;\,0{,}975}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right[$$

$$\text{com } v = \left[\frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}\right].$$

Substituindo pelos valores conhecidos,

$$v = \left[\frac{\left(\frac{7{,}617^2}{853}+\frac{6{,}652^2}{1007}\right)^2}{\frac{1}{853-1}\left(\frac{7{,}617^2}{853}\right)^2 + \frac{1}{1007-1}\left(\frac{6{,}652^2}{1007}\right)^2}\right] = [1705{,}6] = 1705,$$

e como $t_{1705;\,0{,}975} = 1{,}96$, obtém-se

$$\left](168{,}46 - 158{,}48) \pm 1{,}96\sqrt{\frac{7{,}617^2}{853}+\frac{6{,}652^2}{1007}}\right[=]9{,}32;\ 10{,}64[.$$

Com 90% de confiança, existe diferença significativa entre as médias das alturas dos homens e das mulheres (0 não está contido do I. C. a 95%). Como ambos os limites do intervalo são positivos então significa que $\mu_H > \mu_M$, ou seja, a altura média dos homens é superior à altura média das mulheres.

(SPSS) Analyze → Compare Means → Independent-Samples T Test...
(Test Variable: altura; Grouping Variable: sexo;
Define Groups → Use specified values → Group 1: 1; Group 2: 2;
Options → Confidence Interval: 95)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Altura	Equal variances assumed	10,707	,001	30,150	1858	,000	9,976	,331	9,327	10,625
	Equal variances not assumed			29,816	1705	,000	9,976	,335	9,320	10,633

Recorrendo a este resultado do SPSS (não reprodutível pois estes dados não foram disponibilizados), pode-se verificar que existe evidência da diferença das variâncias populacionais (Teste de Levene[†] para igualdade de variâncias, $valor\text{-}p = 0{,}001 \leq \alpha$). Interpretando então a linha inferior do quadro (onde não se assume a igualdade das variâncias), o intervalo de confiança a 95% é]9,320; 10,633[que não contem o valor 0. Logo existe evidência de que as alturas médias dos homens e das mulheres são diferentes. Como os valores dos extremos dos intervalos são positivos então significa que $\mu_H > \mu_M$, ou seja, a altura média dos homens é superior à altura média das mulheres.

7.2.11.3 Intervalo de confiança para a proporção

Numa certa cidade A recolheu-se uma amostra aleatória de 150 homens tendo 54 afirmado que viam o telejornal todos os dias.

a) Com 90% de confiança, será que se pode considerar que a proporção de homens, daquela cidade, que veem o telejornal todos os dias é de 40%.
b) Mantendo-se o resto constante, qual deveria ser a dimensão da amostra de forma a que o erro de estimativa do intervalo de confiança não ultrapasse 5%?

Resolução:

Sejam:

- X_i a v. a. que designa se o i-ésimo homem afirmou ver o telejornal,
- $\overline{P}$ a v. a. que representa a proporção de homens que afirmaram ver o telejornal, em n homens.

$n = 150$ e $\overline{p} = \frac{54}{150} = 0{,}36$.

a) Afirmação: $p = 0{,}4$.
I. C. a 90% para p é dado por:

$$\left]\overline{P} - z_{0{,}95}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}};\ \overline{P} + z_{0{,}95}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}}\right[.$$

Substituindo pelos valores conhecidos, com $z_{0{,}95} = 1{,}645$, obtém-se

$$\left]0{,}36 - 1{,}645\sqrt{\frac{0{,}36(1-0{,}36)}{150}};\, 0\,,36 + 1{,}645\sqrt{\frac{0{,}36(1-0{,}36)}{150}}\right[=]0{,}3464;\, 0{,}5077[.$$

Deste modo, face aos resultados obtidos (0,4 está contido do I. C. a 90%) não é de rejeitar a hipótese de que a proporção de homens que vê o telejornal todos os dias é de 40%, pois com 90% de confiança a percentagem de homens que vê diariamente o telejornal situa-se entre 34,64% e 50,77%.

b) Erro de estimativa $\leq 0{,}05$, então $n = ?$
O erro de estimativa associado ao I. C. a 90% para p é:

$$z_{0{,}95}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}},$$

que se pretende que seja inferior ou igual a 0,5. Portanto,

$$z_{0{,}95}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \leq 0{,}05 \Leftrightarrow 1{,}645\sqrt{\frac{0{,}36(1-0{,}36)}{n}} \leq 0{,}05$$

[†] Ver testes de hipóteses, capítulo 8.

$$\Leftrightarrow \sqrt{n} \geq \frac{1{,}645\sqrt{0{,}36(1-0{,}36)}}{0{,}05} \Leftrightarrow \sqrt{n} \geq 15{,}792$$

$$\Rightarrow n \geq 15{,}792^2 \Leftrightarrow n \geq 249{,}4 \Rightarrow n \geq 250$$

Desta forma, a dimensão mínima da amostra que garante que o erro de estimativa do I. C. a 90% é no máximo de 5% é de 250 homens.

7.2.11.4 Intervalo de confiança para a diferença de proporções

Numa certa cidade A recolheu-se uma amostra aleatória de 150 homens tendo 54 afirmado que viam o telejornal todos os dias. Numa outra cidade do país, cidade B, 80 dos 200 homens selecionados aleatoriamente responderam afirmativamente.
Com 95% de confiança, será de admitir que a proporção de homens que vê o telejornal todos os dias é igual nas duas cidades?

Resolução:

Sejam:

- X_{1i} a v. a. que designa se o i-ésimo homem, da cidade A, afirmou ver o telejornal, $i = 1, \dots, n_1$,
- X_{2i} a v. a. que designa se o i-ésimo homem, da cidade B, afirmou ver o telejornal, $i = 1, \dots, n_2$,
- $\overline{P}_1$ a v. a. que representa a proporção de homens, da cidade A, que afirmaram ver o telejornal, em n_1 homens,
- $\overline{P}_2$ a v. a. que representa a proporção de homens, da cidade B, que afirmaram ver o telejornal, em n_2 homens.

$n_1 = 150;\ \overline{p}_1 = \frac{54}{150} = 0{,}36;\ n_2 = 200$ e $\overline{p}_2 = \frac{80}{200} = 0{,}4.$

O I. C. a 95% para $p_1 - p_2$ é dado por:

$$\left]\overline{P}_1 - \overline{P}_2 - z_{0{,}975}\sqrt{\frac{\overline{P}_1(1-\overline{P}_1)}{n_1} + \frac{\overline{P}_2(1-\overline{P}_2)}{n_2}}\,\overline{P}_1 - \overline{P}_2 + z_{0{,}975}\sqrt{\frac{\overline{P}_1(1-\overline{P}_1)}{n_1} + \frac{\overline{P}_2(1-\overline{P}_2)}{n_2}}\right[.$$

Substituindo pelos valores conhecidos, sendo $z_{0{,}975} = 1{,}96$, obtém-se:

$$\left](0{,}36 - 0{,}4) \pm 1{,}96\sqrt{\frac{0{,}36(1-0{,}36)}{150} + \frac{0{,}4(1-0{,}4)}{200}}\right[=]-0{,}143; 0{,}063[.$$

Portanto, com 95% de probabilidade a diferença entre a percentagem de homens da cidade A e cidade B que veem o telejornal diariamente está entre -14,3% e 6,3%. Como o 0 está contido no intervalo não é de excluir a hipótese de que a percentagem de homens é idêntica nas duas cidades, com a referida confiança.

7.2.11.5 Intervalo de confiança para a variância

Um certo nutricionista concebeu um novo programa dietético para as pessoas que têm uma vida profissional muito sedentária. Aplicou-se este programa a 51 telefonistas tendo-se verificado uma perda média de peso de 3,4 kg e um desvio padrão de 1,8 kg.
Admitindo que a perda de peso segue uma distribuição Normal, estime com 95% de confiança a variabilidade do número de quilogramas perdidos com a aplicação do programa dietético.

Resolução:

Seja X a v.a. que representa o número de quilogramas perdidos com a aplicação do programa dietético, com $X \sim N(\mu = ?\,;\, \sigma = ?\,)$.

$n = 51,\ \overline{x} = 3{,}4 \quad \text{e} \quad s = 1{,}8.$

O I. C. a 95% para σ^2 é dado por:

$$\left]\frac{(n-1)S^2}{\chi^2_{n-1;\,0{,}975}}; \frac{(n-1)S^2}{\chi^2_{n-1;\,0{,}025}}\right[.$$

Substituindo pelos valores conhecidos, sendo $\chi^2_{50;0{,}975} = 71{,}42$ e $\chi^2_{50\,;\,0{,}025} = 32{,}36$, obtém-se:

$$\left]\frac{(51-1)1{,}8^2}{71{,}42}; \frac{(51-1)1{,}8^2}{32{,}36}\right[=]2{,}268; 5{,}006[.$$

Em suma, com 95% de probabilidade, a variância do número de quilos perdidos com o programa dietético estará compreendida entre 2,27 kg e 5,01 kg.

7.2.11.6 Intervalo de confiança para a razão de variâncias

Para substituir uma máquina antiquada encaram-se 2 alternativas: o equipamento A ou equipamento B. Dado tratar-se de uma decisão que envolve custos consideráveis uma vez que os equipamentos são bastante dispendiosos, resolveu-se testar os dois equipamentos durante um período experimental.
No final do período experimental selecionaram-se 31 e 61 peças da produção dos equipamentos A e B, respetivamente, tendo-se registado os seguintes valores relativamente à característica de interesse na avaliação da qualidade do trabalho das máquinas:

$$\sum_{i=1}^{31} x_{iA} = 43{,}4; \qquad \sum_{1=1}^{31} x_{iA}^2 = 123{,}76; \qquad \sum_{i=1}^{61} x_{iB} = 91{,}5; \qquad \sum_{i=1}^{61} x_{iB}^2 = 269{,}25$$

Utilizando um intervalo de confiança a 95% diga se há razões para crer que com a máquina A se consegue uma menor variabilidade da característica de avaliação do que com a máquina B. Admita a normalidade das distribuições.

Resolução:

Sejam:

- X_1 a v.a. que representa o valor da característica de interesse no equipamento A,
- X_2 a v.a. que representa o valor da característica de interesse no equipamento B,

Com $X_1 \sim N(\mu_1 = ?\,;\, \sigma_1 = ?\,)$ e $X_2 \sim N(\mu_2 = ?;\, \sigma_2 = ?\,)$.

$n_1 = 31$ e $n_2 = 61$.

$$\overline{x}_1 = \sum_{i=1}^{31} \frac{x_{1i}}{31} = \frac{43{,}4}{31} = 1{,}4; \qquad s_1^2 = \frac{1}{30}\left(\sum_{i=1}^{31} x_{1i}^2 - 31 \times \overline{x}_1^2\right) = \frac{123{,}76 - 31 \times 1{,}4^2}{30} = 2{,}1;$$

$$\overline{x}_2 = \sum_{i=1}^{61} \frac{x_{2i}}{61} = \frac{91{,}5}{61} = 1{,}5; \qquad s_2^2 = \frac{1}{60}\left(\sum_{i=1}^{61} x_{2i}^2 - 61 \times \overline{x}_2^2\right) = \frac{269{,}25 - 61 \times 1{,}5^2}{60} = 2{,}2.$$

O I. C. a 95% para $\frac{\sigma_1^2}{\sigma_2^2}$ é dado por:

$$\left]\frac{S_1^2}{S_2^2}\frac{1}{f_{n_1-1;n_2-1\,;\,1-\frac{\alpha}{2}}}; \frac{S_1^2}{S_2^2} f_{n_2-1;n_1-1\,;\,1-\frac{\alpha}{2}}\right[.$$

Substituindo pelos valores conhecidos, sendo $f_{30;\,60;\,0,975} = 1,82$ e $f_{60;\,30;\,0,975} = 1,94$, obtém-se:

$$\left]\frac{2,1}{2,2}\times\frac{1}{1,82};\ \frac{2,1}{2,2}\times 1,94\right[=]0,526;\ 1,852[.$$

Com 95% confiança não há razões para crer que exista diferença na variabilidade da característica de avaliação obtida com as duas máquinas (a igualdade das variâncias), uma vez que o valor 1 está presente no intervalo.

7.2.11.7 Intervalo de confiança para amostras emparelhadas

Pretendendo-se comparar a eficácia de uma ação de formação sobre a utilização dos melhores métodos contracetivos, apresentam-se os resultados obtidos em 12 casais (antes e depois de frequentarem a ação de formação).
As maiores pontuações correspondem à utilização dos métodos contracetivos a cada casal.

Casal n.º	1	2	3	4	5	6	7	8	9	10	11	12
Antes (x_i)	14	21	33	29	34	26	21	15	16	20	29	18
Depois (y_i)	19	21	41	26	40	33	28	27	24	25	27	26

Resolução:

Estamos perante duas amostras emparelhadas.
Sejam:

- X_1 a v.a. que representa o resultado dos casais antes da ação de formação,
- X_2 a v.a. que representa o resultado dos casais depois da ação de formação.

Construção da variável $D_i = X_{1i} - X_{2i}$.

Casal (i)	1	2	3	4	5	6	7	8	9	10	11	12
Antes (x_{1i})	14	21	33	29	34	26	21	15	16	20	29	18
Depois (x_{2i})	19	21	41	26	40	33	28	27	24	25	27	26
$d_i = x_{1i} - x_{2i}$	-5	0	-8	3	-6	-7	-7	-12	-8	-5	2	-8

$n = 12;\ \overline{d} = 5,083$ e $s_d = 4,502$.

Assumindo a normalidade, o I. C. para μ_D a 95% é obtido através de:

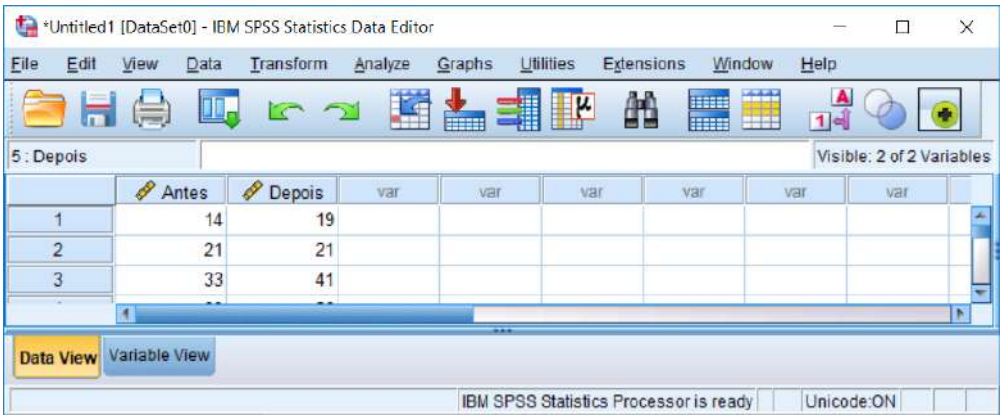
$$\left]\overline{X}_D - t_{11;\,0,975}\frac{S_D}{\sqrt{n}};\ \overline{X}_D + t_{11;\,0,975}\frac{S_D}{\sqrt{n}}\right[.$$

Substituindo pelos valores conhecidos, sendo $t_{11;\,0,975} = 2,201$, obtém-se:

$$\left]-5,083 \pm 2,201\frac{4,502}{\sqrt{12}}\right[=]-7,944;\ -2,223[.$$

A igualdade das pontuações não está incluída no intervalo de confiança obtido (0 não está no intervalo), logo com 95% de confiança numa das alturas (antes ou depois) a pontuação média é melhor. Como ambos os limites do intervalo são negativos então significa que $\mu_{X_1} - \mu_{X_2}$, ou seja, a pontuação média é superior depois de frequentarem a acção de formação. Logo, a acção de formação foi eficaz.

(SPSS)



(SPSS) Analyze → Compare Means → Paired-Samples T Test...
(Paired Variables → Variable 1: Antes; Variable 2: Depois;
Options → Confidence Interval: 90)

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Antes	23,00	12	6,994	2,019
	Depois	28,08	12	6,762	1,952

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Antes & Depois	12	,786	,002

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	90% Confidence Interval of the Difference Lower	Upper			
Pair 1	Antes - Depois	-5,083	4,502	1,300	-7,417	-2,750	-3,912	11	,002

7.2.11.8 Intervalo de confiança para o coeficiente de correlação

Realizou-se uma sondagem junto de 52 alunos do curso de Engenharia Civil, escolhidos aleatoriamente, sobre as notas obtidas nas disciplinas de Matemática e de Estatística. Nesta amostra, verificou-se que o coeficiente de correlação linear foi $r = 0{,}84$.

a) Determine um intervalo de confiança a 99% para o coeficiente de correlação populacional ρ.

b) Qual o grau de confiança associado ao seguinte intervalo de confiança $0{,}7557 < \rho < 0{,}8969$?

Resolução:

Sejam:

- X_1 a v.a. que representa a nota obtida pelo aluno na disciplina de Matemática,
- X_2 a v.a. que representa a nota obtida pelo aluno na disciplina de Estatística.

Vamos assumir que $(X_1; X_2)$ têm distribuição Normal bivariada.

a) $n = 52$.

O I. C. para ρ a 99% é obtido através de:

$$\left]\frac{e^{2Z_{\rho_{inf}}} - 1}{e^{2Z_{\rho_{inf}}} + 1}; \frac{e^{2Z_{\rho_{sup}}} - 1}{e^{2Z_{\rho_{sup}}} + 1}\right[$$

onde $Z_{\rho_{inf}} = Z_R - z_{0,995}\frac{1}{\sqrt{n-3}}$; $Z_{\rho_{sup}} = Z_R + z_{0,995}\frac{1}{\sqrt{n-3}}$ e $Z_R = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$.

Substituindo pelos valores conhecidos, em que $z_{0,995} = 2{,}576$ e

$$r = 0{,}84 \Rightarrow z_r = \frac{1}{2}ln\left(\frac{1+0{,}84}{1-0{,}84}\right) = 1{,}2212,$$

$$z_{r_{inf}} = 1{,}2212 - 2{,}576\frac{1}{\sqrt{52-3}} = 0{,}8532,$$

$$z_{r_{sup}} = 1{,}2212 + 2{,}576\frac{1}{\sqrt{52-3}} = 1{,}5892$$

obtém-se

$$]0{,}6927;\ 0{,}9200[.$$

Com 99% de confiança, o valor do coeficiente de correlação linear populacional situa-se entre 0,693 e 0,92.

b) $0{,}7557 < \rho < 0{,}8969 \Rightarrow 100(1-\alpha)\% = ?$

$$\begin{cases} \frac{e^{2z_{r_{inf}}}-1}{e^{2z_{r_{inf}}}+1} = 0{,}7557 \\ \frac{e^{2z_{r_{sup}}}-1}{e^{2z_{r_{sup}}}+1} = 0{,}8969 \end{cases} \Leftrightarrow \begin{cases} z_{r_{inf}} = 0{,}9862 \\ z_{r_{sup}} = 1{,}4562 \end{cases} \Leftrightarrow \begin{cases} z_R - z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}} = 0{,}9862 \\ z_R + z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}} = 1{,}4562 \end{cases}$$

$$\Leftrightarrow \begin{cases} 1{,}2212 - z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{52-3}} = 0{,}9862 \\ 1{,}2212 + z_{1-\frac{\alpha}{2}}\frac{1}{\sqrt{52-3}} = 1{,}4562 \end{cases} \Rightarrow z_{1-\frac{\alpha}{2}} = 1{,}645$$

Como $\Phi(1{,}645) = 0{,}95 \Leftrightarrow z_{0,95} = 1{,}645$, então

$$1 - \frac{\alpha}{2} = 0{,}95 \Leftrightarrow \alpha = 0{,}1.$$

Portanto, o grau de confiança associado ao intervalo $0{,}7557 < \rho < 0{,}8969$ é de 90%.

7.3 Exercícios propostos

1. Às 20:00 de 13 de junho de 2004 era possível ler a seguinte notícia na SIConline:
   "O PS é o vencedor das eleições europeias em Portugal, segundo a previsão SIC/Eurosondagem. Com 44,1 por cento a 47,9 por cento dos votos, os socialistas conseguem eleger 12 a 13 eurodeputados. A abstenção atingiu os 64 por cento. A coligação "Força Portugal" obteve 29,7 por cento a 33,5 por cento dos votos, valores que correspondem a 8 a 9 lugares no Parlamento Europeu. A CDU terá conseguido entre 10,1 por cento e 11,9 por cento e 2 a 3 deputados. Por sua vez, o Bloco de Esquerda teve 5,1 por cento a 6,9 por cento dos votos, que valem 1 eurodeputado. Os votos noutros partidos estão entre 2,8 por cento a 4,2 por cento. Os votos brancos/nulos são 1,5 por cento a 2,3 por cento, de acordo com a projecção....".

Dê um exemplo de uma estimativa:

a) Pontual;
b) Intervalar;

apresentadas nesta notícia.

2. Qual das seguintes expressões está correta? Justifique.
   a) $E(\mu) = \tilde{\mu}$. b) $E(\tilde{\mu}) = \mu$.

3. Qual das seguintes expressões está correta? Justifique.
   a) $P(\mu \leq x) = P(\tilde{\mu} \leq x)$. b) $P(\overline{X} \leq x) = P(\tilde{\mu} \leq x)$.

4. Distinga estimador de estimativa.

5. Considere uma população Normal, com média μ e desvio padrão σ, da qual se retirou uma amostra aleatória $X_1, X_2, \ldots, X_n$. Seja T a seguinte estatística:

$$T = \frac{\sum_{i=1}^{n-1} X_i + X_n}{n}.$$

   a) Qual a distribuição amostral de T e respectivos parâmetros.
   b) Diga se T é um estimador centrado para μ.

6. Considere uma população com média μ e os seguintes estimadores para μ, para amostras aleatórias de dimensão $n = 4$:

$$\hat{\mu}_1 = \frac{2X_1 + X_2 + X_3 + X_4}{5}; \qquad \hat{\mu}_2 = \frac{X_1 + X_3}{2}; \qquad \hat{\mu}_3 = \frac{X_1 + X_2 + X_3 + X_4}{4}.$$

   a) Compare os estimadores quanto ao não enviesamento e eficiência.
   b) Calcule a estimativa fornecida por cada um deles com base na seguinte amostra:

   7 8 10 11

7. São propostos os seguintes dois estimadores para a variância da população (σ^2):

$$S^{*2} = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n} \text{ e } S^2 = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n-1}.$$

   a) Estude-os quanto ao enviesamento?
   b) Calcule as estimativas fornecidas por cada um com base na seguinte amostra:

   32 27 32 28 31 37 25 36 30

Observação: $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - \sum_{i=1}^{n}(\bar{X} - \mu)^2$.

8. Obtenha o estimador de máxima verosimilhança para o parâmetro p, de uma população com distribuição Bernoulli, i. e., com f. p.:

$$f(x; p) = p^x(1-p)^{1-x}, \qquad x = 0, 1, \qquad 0 < p < 1.$$

9. Considere uma população com distribuição de Poisson, com parâmetro λ.
   a) Obtenha o estimador de λ pelo método dos momentos.
   b) Obtenha o estimador de máxima verosimilhança para o parâmetro λ.
   c) Recolheu-se uma amostra de dimensão $n = 5$, tendo-se observado os seguintes valores:

   10 14 9 6 12

      i. Esboce o gráfico da função de verosimilhança e interprete-o.
      ii. Forneça uma estimativa para λ com base no método da máxima verosimilhança.
      iii. Qual o valor da função de verosimilhança quando $\lambda = 11$?

10.Foram retiradas 25 peças da produção diária de uma máquina, encontrando-se, para uma certa medida, uma média de 5,2 mm. Sabe-se que as medições têm distribuição Normal. Construa intervalos de confiança para média populacional aos níveis de significância de 10%, 5% e 1%.

a) Com base num desvio padrão populacional de 1,2 mm.
b) Com base num desvio padrão amostral de 1,2 mm.
c) Justifique as diferenças obtidas.

11.Um profissional de saúde do Centro de Proteção Solar pretende estimar o tempo médio diário que os frequentadores habituais de uma determinada praia algarvia passam em banhos de sol. Para o efeito, recolheu uma amostra aleatória de 20 indivíduos tendo observado um tempo médio de exposição ao sol de 4 horas e um desvio padrão de 0,9 horas. Admita que o tempo despendido diariamente em banhos de sol pelos frequentadores habituais da referida praia segue uma distribuição Normal com desvio padrão 1 hora.

a) Dê uma estimativa pontual para o tempo médio diário despendido em banhos de sol pelos frequentadores habituais da referida praia.
b) Construa um intervalo de confiança a 90% para o tempo médio diário despendido em banhos de sol pelos frequentadores da praia.
c) Qual deve ser a dimensão mínima da amostra para o erro de estimativa, do intervalo de confiança da alínea anterior, ser inferior a 0,2 horas?
d) Qual deve ser a dimensão mínima da amostra para a amplitude do intervalo de confiança da alínea a) ser no máximo 0,5 horas?
e) Qual a dimensão da amostra utilizada para obter o seguinte intervalo de confiança a 95% para o tempo médio diário de exposição ao sol:]3,69; 4,31[?

12.Suponha que numa amostra de 100 válvulas de televisão fabricadas por certa companhia se obteve um valor médio igual a 1200 horas e um desvio padrão igual a 100 horas. Qual deveria ser o tamanho da amostra de forma que o erro da vida média estimada não ultrapassasse as 20 horas, com 99,73%?

13.O peso de componentes eletrónicos produzidos por determinada empresa, é uma variável aleatória que se supõe ter uma distribuição Normal. Pretendendo-se estudar a variabilidade dos pesos das referidas componentes, recolheu-se uma amostra de 11 elementos, cujos valores (em gramas) foram:

98 97 102 100 98 101 102 105 95 102 100

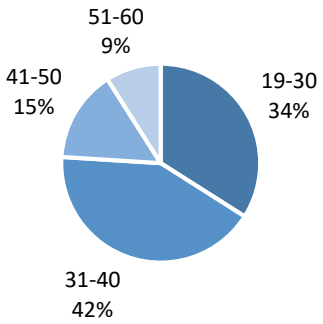
a) Apresente uma estimativa pontual para a variância do peso das componentes.
b) Construa um intervalo de confiança para a variância do peso, com um grau de confiança de 95%.

14.Os resultados seguintes foram obtidos a partir de medições do pH da água de um lago em diferentes períodos:

$$\sum_{i=1}^{10} x_i = 46{,}1\,; \qquad \sum_{i=1}^{10} x_i^2 = 220{,}81$$

a) Determine um intervalo de confiança para μ a 95%.
b) Determine um intervalo de confiança para σ^2 a 99%.

15.Na revista Visão de 12 de setembro de 2002 foi escrito um artigo sobre “O Vício da Adrenalina”. Neste artigo é apresentado o perfil do *português radical* em termos de profissão, sexo, residência e idade. Relativamente à idade apresentaram apenas o seguinte gráfico:



Fonte: Contributos do INATEL para o Desporto Aventura em Portugal – A Realidade dos Desportos Aventura, 2001.

Admita que foram inquiridos 1000 indivíduos.

a) Forneça uma estimativa pontual para a verdadeira proporção de portugueses radicais com idade superior ou igual a 51 anos.
b) Concorda com a seguinte afirmação: "Com 95% de confiança, a verdadeira percentagem de portugueses radicais com pelo menos 51 anos situa-se entre 6% e 12%."

16.Num trabalho realizado há já algum tempo concluiu-se que 62% dos passageiros que entram na estação A do metro têm como destino o centro da cidade. Esse valor tem vindo a ser utilizado em todos os estudos de transportes realizados desde então. Tendo surgido dúvidas sobre a atualidade daquele valor, pois crê-se que tem vindo a diminuir acompanhando o declínio do centro, realizou-se um inquérito naquela estação. Dos 240 passageiros inquiridos, 126 indicaram o centro como destino. Com base nestes resultados construa um intervalo de confiança a 90% para a percentagem de passageiros que entram na estação A e saem no centro da cidade, e verifique se os 62% ainda se mantêm consistentes com a realidade atual.

17.Num determinado período pré-eleitoral foi realizada uma sondagem com o objetivo de analisar a popularidade de cada um dos candidatos A e B num determinado distrito. Para tal foram inquiridos 780 residentes nesse distrito, manifestando-se 55% dos inquiridos a favor do candidato A.

a) Construa os intervalos de confiança a 90%, 95% e 98%, para a percentagem de residentes desse distrito que são a favor do candidato A.
b) Comente as diferenças obtidas para os 3 intervalos.
c) Suponha que a percentagem 55% resultou de uma amostra de 1020 elementos. Qual o intervalo de confiança a 95%? Comente.
d) Que tamanho deverá ter, no mínimo, a amostra caso se pretenda que a margem de erro, para o intervalo de confiança a 90% obtido na alínea a), não seja superior a 2,5%?

18.Pretende-se saber se as proporções de pinheiros afetados pelo Nemátodo são iguais em duas zonas A e B. Na zona A foi recolhida uma amostra de 150 pinheiros e verificou-se que 107 estavam afetados e na zona B recolheu-se uma amostra de 100 havendo 63 afetados. Que conclusão se pode tirar ao nível de significância de 0,05?

19.Selecionaram-se ao acaso alguns dias, tendo-se observado a temperatura máxima (em °C) no Ártico e no Antártico. Os valores obtidos apresentam-se na tabela seguinte:

Ártico	2,7	3,2	3,6	4,1	2,7	3,2	4,5	3,6	2,7
Antártico	4,1	4,5	3,6	2,7	3,6	3,2	4,1		

Admita que as distribuições das temperaturas máximas seguem a distribuição Normal.

a) Construa um intervalo de confiança a 95% para a diferença das médias das temperaturas máximas. Interprete.

b) Estime, com uma confiança de 99%, a média da temperatura máxima no Antártico.
c) Com 99% de confiança pode afirmar que a variabilidade nas temperaturas máximas é igual nos dois polos?

20.Num Hospital, escolheu-se ao acaso uma amostra de 7 doentes e verificou-se que dormiram por noite as seguintes horas:

7 5 8 8,5 6 7 8

Para testar a eficiência de um certo medicamento para dormir, sujeitou-se um outro grupo constituído por 5 doentes, à referida medicação. Constatou-se que dormiram as seguintes horas por noite:

9 8,5 9,5 10 8

a) Admitindo que a variável em estudo tem distribuição Normal na população, construa um intervalo de confiança de 99% para a diferença entre os tempos médios de sono entre os dois grupos.
b) Admita agora, que os 7 doentes observados inicialmente são sujeitos à mesma medicação e são registadas as horas de sono, pela mesma ordem:

9,8 8 8,5 8 9,5 6,5 7,5

Pode-se concluir que o uso do medicamento altera o número médio de horas de sono?

21.Com a finalidade de estimar a diferença de exposição média à radioatividade de trabalhadores de uma central nuclear nos anos de 1980 e 1989, foi efetuado um determinado estudo. Foram analisadas 2 amostras independentes, de 16 trabalhadores cada, para cada ano:

Medida	1980	1989
Média	0,94	0,62
Variância	0,040	0,028

a) Determine um intervalo de confiança para a razão de variâncias, ao nível de significância de 5%.
b) Será que existe diferença de exposição média à radioatividade entre 1980 e 1989?
c) Qual deveria ser a dimensão das amostras recolhidas (considerando amostras de igual dimensão e as variâncias amostrais como sendo populacionais) para não se ter um erro de estimativa superior a 0,05?

22.Selecionaram-se aleatoriamente dois grupos de alunos, respetivamente com 10 e 13 elementos, aos quais se transmitiu a mesma matéria, supostamente em condições idênticas, durante o mesmo período de tempo. Posteriormente foram avaliados tendo sido contados os tempos de resposta de cada elemento, relativamente às questões colocadas. Obtiveram-se os seguintes resultados, onde X e Y representam o tempo de resposta dos elementos do 1º e 2º grupo respetivamente:

$$\sum_{i=1}^{10} x_i = 383{,}2; \qquad \sum_{i=1}^{10} x_i^2 = 14957{,}374; \qquad \sum_{i=1}^{13} y_i = 401{,}3; \qquad \sum_{i=1}^{13} y_i^2 = 12669{,}2397.$$

Admita a normalidade dos tempos de resposta.

a) Com um grau de confiança de 95%, pode-se considerar que os tempos médios de resposta são idênticos nos dois grupos? (considere a igualdade das variâncias populacionais)
b) Determine os limites de confiança de 95% para a razão de variâncias. Interprete.
c) Suponha que se conhece σ_1 e σ_2 (considere que $\sigma_1 = s_1$ e $\sigma_2 = s_2$). Construa o intervalo de confiança a 95% para a diferença dos valores médios.
d) Comente as diferenças obtidas nos intervalos da alínea a) e c).

23.Um estudo comparativo do insucesso escolar entre alunos cujos pais são divorciados e alunos com vida familiar considerada como "regular", foram selecionados dois grupos de estudantes, A e B, tendo-se observado que, de entre os 31 alunos com pais divorciados (grupo A), 30% tinham insucesso escolar, enquanto que no grupo de 41 estudantes de família "regular" (grupo B), 26% estavam naquelas condições. Relativamente à idade, em anos, dos alunos obtiveram-se os seguintes valores com auxílio do SPSS:

Statistics

		Grupo A	Grupo B
N	Valid	31	41
	Missing	0	0
Mean		11,25	11,75
Std. Deviation		1,60	2,10

a) Forneça uma estimativa pontual para a idade média dos alunos cujos pais são divorciados.
b) Forneça uma estimativa pontual para o desvio padrão da idade dos alunos com vida familiar considerada como "regular".
c) Forneça uma estimativa pontual para a proporção de alunos com insucesso escolar cujos pais são divorciados.
d) Construa um intervalo de confiança a 95% para a idade média dos alunos cujos pais são divorciados. Interprete o intervalo obtido.
e) Qual deverá ser, no mínimo, a dimensão da amostra de forma a que a amplitude do intervalo anterior seja no máximo 1 ano?
f) Construa um intervalo de confiança a 95% para a proporção de alunos com insucesso escolar cujos pais são divorciados.
g) Determine a dimensão mínima da amostra de forma a que o erro de estimativa do intervalo anterior não ultrapasse 0,1?
h) Com 95% de certeza, pode-se afirmar que não existe diferença entre a idade média dos alunos dos dois grupos?
i) Complete: "Com 90% de confiança, a diferença entre a proporção de alunos aprovados nos dois grupos situa-se entre ... e ....".
j) Com base num intervalo de confiança a 90% pode-se considerar que a variabilidade das idades dos alunos dos dois grupos é idêntica? Que suposição adicional teve de fazer?
k) De que forma pode obter uma opinião diferente alterando o grau de confiança do intervalo anterior?

24.Como é habitual, sempre que se realizam eleições realizam-se sondagens com vista a projetar os resultados finais obtidos por cada partido.
Aquando da realização, em 2004, das eleições para o Parlamento Europeu, num determinado distrito inquiriram-se 1000 pessoas tendo o partido A obtido 44,5% das intenções de voto, enquanto nas eleições realizadas em 2019 das 2000 pessoas inquiridas apenas 42,2% manifestaram intenção de votar nesse partido.
a) Construa um intervalo de confiança a 99% para a proporção de votos no partido A em 2019.
b) Com base num intervalo de confiança a 95%, pode-se considerar que a percentagem de intenções de voto no partido A é idêntica nos dois anos?
c) Se alterar o grau de confiança, na alínea anterior, para 99%, mantém a sua opinião? Justifique sem efetuar os cálculos. E a 90%?

25.Numa amostra aleatória de 101 pessoas obteve-se uma correlação linear de 0,43, entre o grau de satisfação relativamente à eficácia de um determinado produto e o seu preço. Construa um intervalo de confiança a 99% para o coeficiente de correlação linear populacional.

26.Recolheu-se uma amostra aleatória de 101 residências, tendo-se registado o número de assoalhadas da residência e o peso (em kg) dos plásticos deitados no lixo, tendo-se observado uma correlação linear amostral de 0,74. Construa um intervalo de confiança a 99% para o coeficiente de correlação linear populacional. O que conclui?

27.Considere os dados da tabela seguinte, que são referentes a uma amostra de 33 cigarros para os quais se observou a quantidade de nicotina e de alcatrão contida em cada um deles.

Cigarro	Alcatrão	Nicotina	Cigarro	Alcatrão	Nicotina	Cigarro	Alcatrão	Nicotina
1	16	1,2	12	15	1,2	23	12	1
2	16	1,2	13	16	1,2	24	14	1
3	16	1	14	9	0,7	25	5	0,5
4	9	0,8	15	11	0,9	26	6	0,6
5	1	0,1	16	2	0,2	27	8	0,7
6	8	0,8	17	18	1,4	28	18	1,4
7	10	0,8	18	15	1,2	29	16	1,1
8	16	1	19	13	1,1	30	16	1,3
9	14	1	20	15	1	31	11	0,7
10	13	1	21	17	1,3	32	15	1,2
11	13	1,1	22	9	0,8	33	8	0,8

a) Construa um intervalo de confiança a 90% para o coeficiente de correlação linear populacional entre as variáveis em estudo.

b) Com base no intervalo obtido, considera que existe relação linear entre as variáveis?

8 Testes de hipóteses

Nos **testes de hipóteses** o objectivo é validar ou não determinadas afirmações sobre a população com base na informação amostral. No caso particular dos *testes de hipóteses paramétricos* a validação diz apenas respeito aos *parâmetros da população*.

Hipóteses a testar:

H_0: Hipótese nula (contém sempre uma igualdade) – o que se assume correto até prova em contrário;
H_1: Hipótese alternativa (contém sempre uma desigualdade >, <, $\neq$).

Os testes classificam-se em:

- *Unilateral direito* quando H_1 contiver a desigualdade >;
- *Unilateral esquerdo* quando H_1 contiver a desigualdade <;
- *Bilateral* se H_1 contiver a não igualdade $\neq$;

A hipótese H_0 é considerada verdadeira ao longo da realização do teste até ao momento em que haja evidência estatística clara apontando em sentido contrário.

Na prática, a definição das hipóteses unilaterais não tem sido uma matéria de consenso entre muitos autores, existindo algumas regras básicas que interessa referir:

- A igualdade está sempre em H_0 (isto é H_1 só pode conter os sinais $\neq$ ou < ou >);
- O que se pretende testar está em H_1;
- O que se aceita por defeito, sem prova, está em H_0.

Imagine a analogia com base num exemplo de um tribunal e considere as diferenças entre:

- H_0: Inocente *vs.* H_1: Culpado
  Significa que se procura testar a culpabilidade do indivíduo, mas se não houver uma forte evidência, ele será sempre considerado inocente. Só será preso se houver fortes evidências de crime.
- H_0: Culpado *vs.* H_1: Inocente
  Significa que se procura testar a inocência do indivíduo, mas se não houver uma forte evidência, ele será sempre considerado culpado. Em caso de dúvida é preso.

Decisão:

- Rejeitar H_0, ou
- Não rejeitar H_0.

Na Figura 8.1 apresenta-se um esquema onde se pretende ajudar a esquematizar a formulação das hipóteses assim como a terminologia das conclusões.

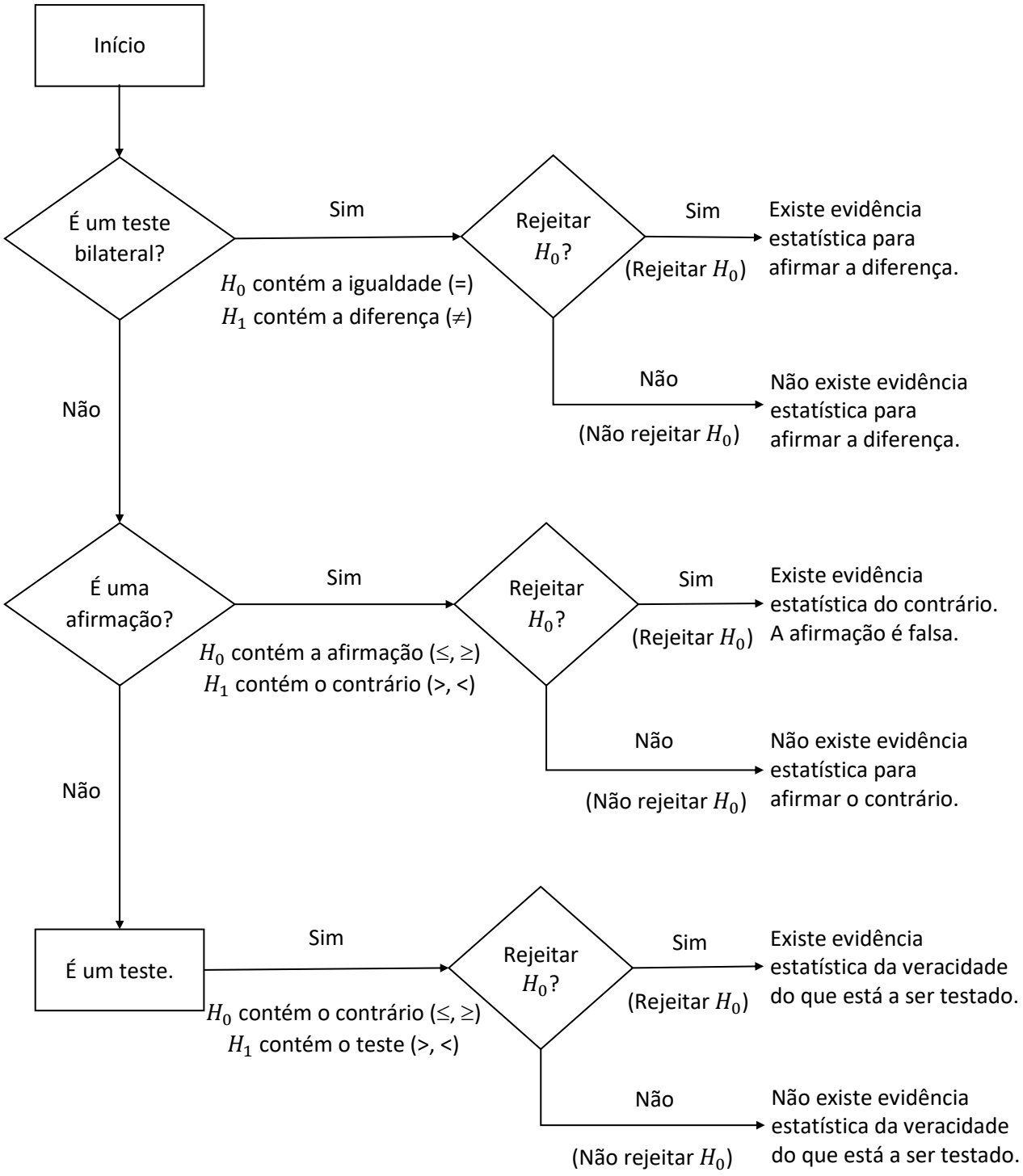


Figura 8.1: Formulação das hipóteses e terminologia das conclusões.

8.1 Metodologia

Na Figura 8.2 apresenta-se uma descrição das etapas a percorrer quando se realiza um teste de hipóteses.

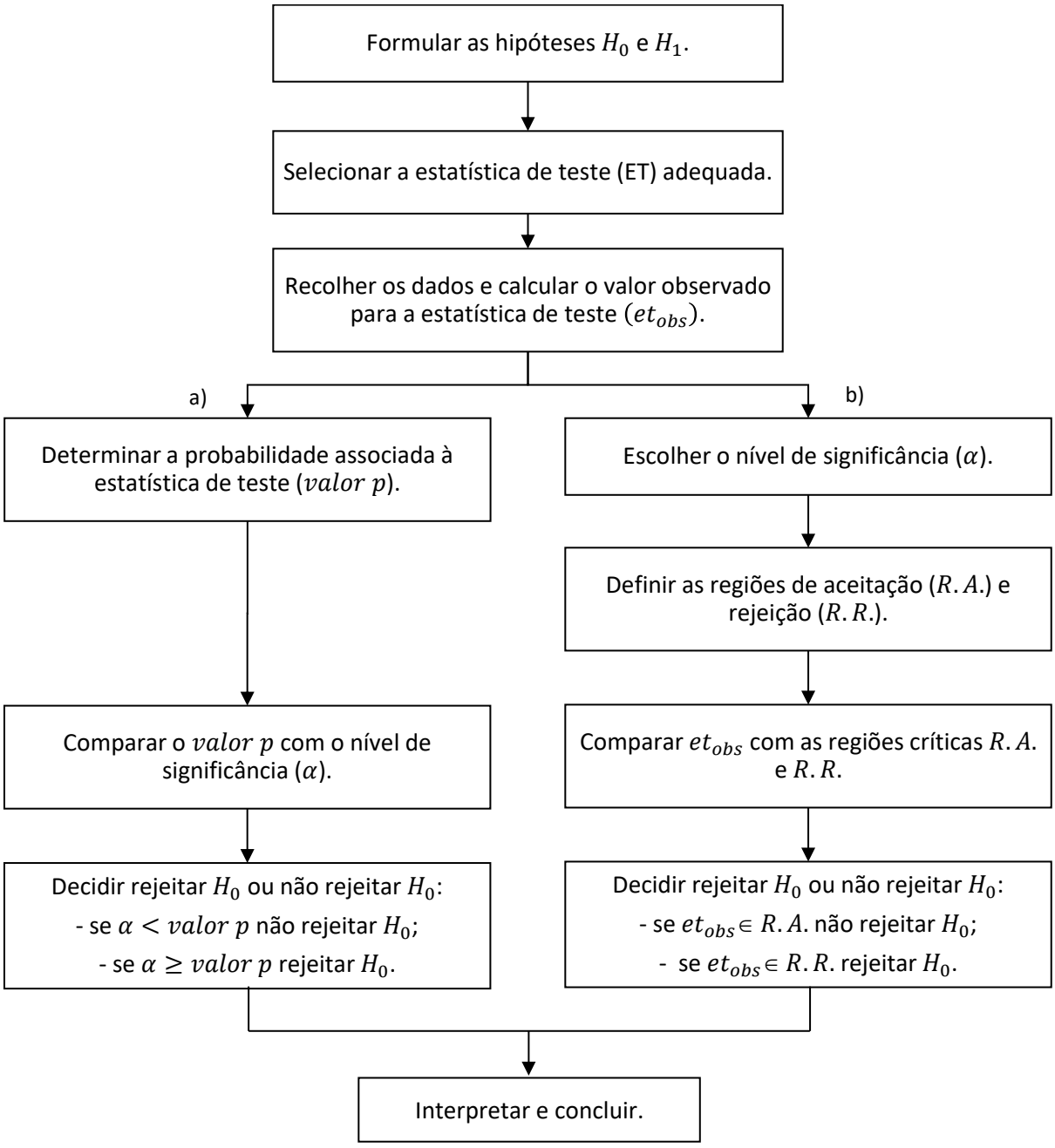


Figura 8.2: Metodologia.

Para uma melhor compreensão deste esquema descrevem-se sumariamente alguns conceitos utilizados:

- Estatística de teste (ET): é o equivalente da variável fulcral, apresentada no capítulo Intervalos de confiança, onde se considera a hipótese H_0 verdadeira;
- Região de aceitação $(R.A.)$: contém os valores admissíveis para o valor observado para a estatística de teste (et_{obs}), que sustentam a não evidência da hipótese H_1;
- Região de rejeição $(R.R.)$: contém os restantes valores possíveis para o valor observado para a estatística de teste, que sustentam a evidência da hipótese H_1.

As amplitudes das $R.A.$ e $R.R.$ dependem do nível de significância considerado, bem como as correspondentes conclusões.

De acordo com esta esquematização existem 2 alternativas de resolução:

a) Calcula-se o valor de probabilidade ($valor\ p$) associado à estatística de teste observada (et_{obs}) que posteriormente é comparado com o nível de significância (α) que se quer utilizar na decisão. Rejeita-se H_0 para valores de $\alpha \geq valor\ p$, caso contrário não se rejeita H_0.
b) Define-se, à partida, nível de significância que se quer utilizar na decisão, que serve de base à construção das regiões críticas e verifica-se a qual das regiões pertence o valor observado para a estatística de teste (et_{obs}). Rejeita-se H_0 quando $et_{obs} \notin R.A.$, caso contrário não se rejeita H_0.

A opção pela alternativa a) implica mais conhecimentos de probabilidades para a determinação do $valor\ p$, acessível e descrito no capítulo das Distribuições de Probabilidades, ou fornecido pelos programas de estatística (hoje em dia incontornáveis no ensino e na investigação). Este $valor\ p$ permite decidir (rejeitar ou não rejeitar H_0) para todos os níveis de significância, no contexto específico de cada estudo (amostra observada).

A opção b) pode ser considerada como a mais intuitiva e simples, mas mais limitada em termos de interpretação: apenas se obtém uma resposta para um determinado nível de significância, não sendo possível determinar, sem novos cálculos, a resposta para todos os níveis de significância.

8.2 Erros nos testes de hipóteses

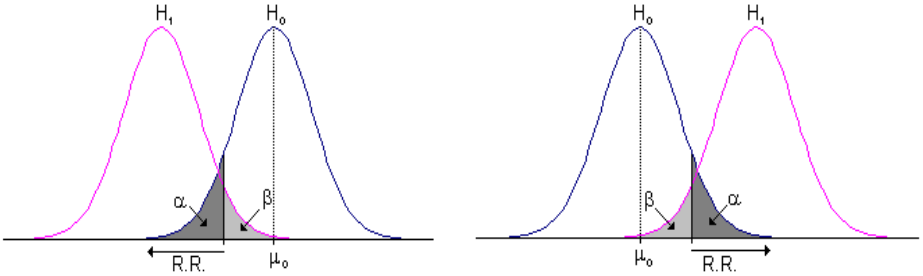
A uma decisão está sempre associada a um risco, o risco de errar. No caso particular dos testes de hipótese a tomada de decisão, para a população, é baseada na informação amostral pelo que se podem cometer erros. Uma das principais características dos testes de hipóteses é permitir controlar ou minimizar o risco associado às decisões erradas.

Riscos (erros) associados à tomada de decisão:

Na Tabela 8.1 esquematizam-se os tipos de erros e na Figura 8.3 é feita a representação gráfica da probabilidade associada a cada um deles.

Tabela 8.1: *Tipos de erros associados à decisão tomada.*

	Situação real	
Decisão	H_0 é verdadeira	H_0 é falsa
Rejeitar H_0	Decisão incorreta $P(\text{Rej. } H_0 \mid H_0 \text{ é verd.}) \leq \alpha$ Erro Tipo I	Decisão correta $P(\text{Rej. } H_0 \mid H_1 \text{ é verd.}) = 1 - \beta$
Não rejeitar H_0	Decisão correta $P(\text{Não rej. } H_0 \mid H_0 \text{ é verd.}) > 1 - \alpha$	Decisão incorreta $P(\text{Não rej. } H_0 \mid H_1 \text{ é verd.}) = \beta$ Erro de Tipo II



a) teste unilateral esquerdo ou bilateral. *b) teste unilateral direito ou bilateral*

Figura 8.3: *Probabilidades de um erro de Tipo I (sombreado escuro) e de um erro Tipo II (sombreado claro), num teste unilateral ou bilateral.*

Observação: A única forma de minimizar estes dois tipos de erro em simultâneo é aumentando a dimensão da amostra, n.

A abordagem dos testes de hipóteses para controlar os erros consiste em fixar a probabilidade associada ao erro de Tipo I, α, e minimizar a probabilidade associada ao erro de Tipo II, β.

A razão pela qual se atribui mais importância ao erro Tipo I deriva do seguinte ponto de vista: a possibilidade de rejeitar H_0 incorretamente é considerada mais grave, pois esta hipótese corresponde à que deve ser defendida, a menos que existam evidências fortes a apontarem em sentido contrário.

Exemplo: Coloque-se novamente o exemplo do funcionamento de um julgamento num tribunal, onde uma pessoa é inocente até se provar o contrário.

- H_0: A pessoa é inocente.
- H_1: A pessoa é culpada.
- Erro de Tipo I: A pessoa é condenada, mas está inocente.
- Erro de Tipo II: A pessoa é absolvida, mas é culpada.

O sistema judicial considera que é mais grave culpar um inocente do que absolver um culpado.

8.2.1 Erro Tipo I

Definição: O **erro de Tipo I**, ou de 1ª espécie, consiste em rejeitar a hipótese H_0 quando H_0 é verdadeira.

A probabilidade associada ao erro de Tipo I é:

$$P(\text{Rejeitar } H_0 \mid H_0 \text{ é verdadeira}) = P(\text{Rejeitar } H_0 \mid \theta = \theta_0) \leq \alpha.$$

8.2.2 Erro Tipo II

Definição: O **erro de Tipo II**, ou de 2ª espécie, consiste em não rejeitar a hipótese H_0 quando H_0 é falsa.

A probabilidade associada ao erro de Tipo II, onde $\theta = \theta_1$, é β e representa-se por:

$$\beta(\theta_1) = P(\text{Não rejeitar } H_0 | H_0 \text{ é falsa}) = P(\text{Não rejeitar } H_0 | H_1 \text{ é verdadeira})$$
$$= P(\text{Não rejeitar } H_0 | \theta = \theta_1).$$

O valor de $\beta(\theta_1)$ diminui à medida que o verdadeiro valor de $\theta = \theta_1$, se afasta de θ_0 (i.e., do valor indicado H_0), dado ser menos provável que não se detete o verdadeiro valor. Consequentemente, quanto mais próximo estiver θ_1 de θ_0 maior é o valor desta probabilidade.

8.2.3 Potência do teste

A potência de teste, $\pi = 1 - \beta$, permite medir a capacidade do teste para decidir acertadamente, quando a hipóteses H_0 é falsa.

Definição: Designa-se por **função potência de teste**, $\pi(\theta_1)$, a probabilidade associada à decisão de rejeição de H_0 quando esta é falsa, i. e.:

$$\pi(\theta_1) = P(\text{Rejeitar } H_0 | H_0 \text{ é falsa})$$
$$= P(\text{Rejeitar } H_0 | \theta = \theta_1)$$
$$= 1 - P(\text{Não rejeitar } H_0 | \theta = \theta_1)$$
$$= 1 - \beta(\theta_1).$$

Quanto mais perto estiver θ_1 de θ_0 menor é o valor da função potência, menos potente é o teste, uma vez que é menor a capacidade de distinguir os verdadeiros valores dos falsos. Por outro lado, quanto mais afastado estiver o verdadeiro valor de $\theta = \theta_1$, em relação a θ_0, mais capaz é o teste de tomar decisões corretas.

8.2.4 Valor *p*

Definição: O $\boldsymbol{valor\ p}$, (habitualmente denominado por $p\text{-}value$) é o menor nível de significância, α, a partir do qual se começa a rejeitar a hipótese H_0, i. e, se $\alpha \geq \text{valor } p$ então rejeitar H_0.

Na Figura 8.4 esquematiza-se a probabilidade associada ao $valor\ p$, tendo em conta o tipo de hipóteses formuladas.

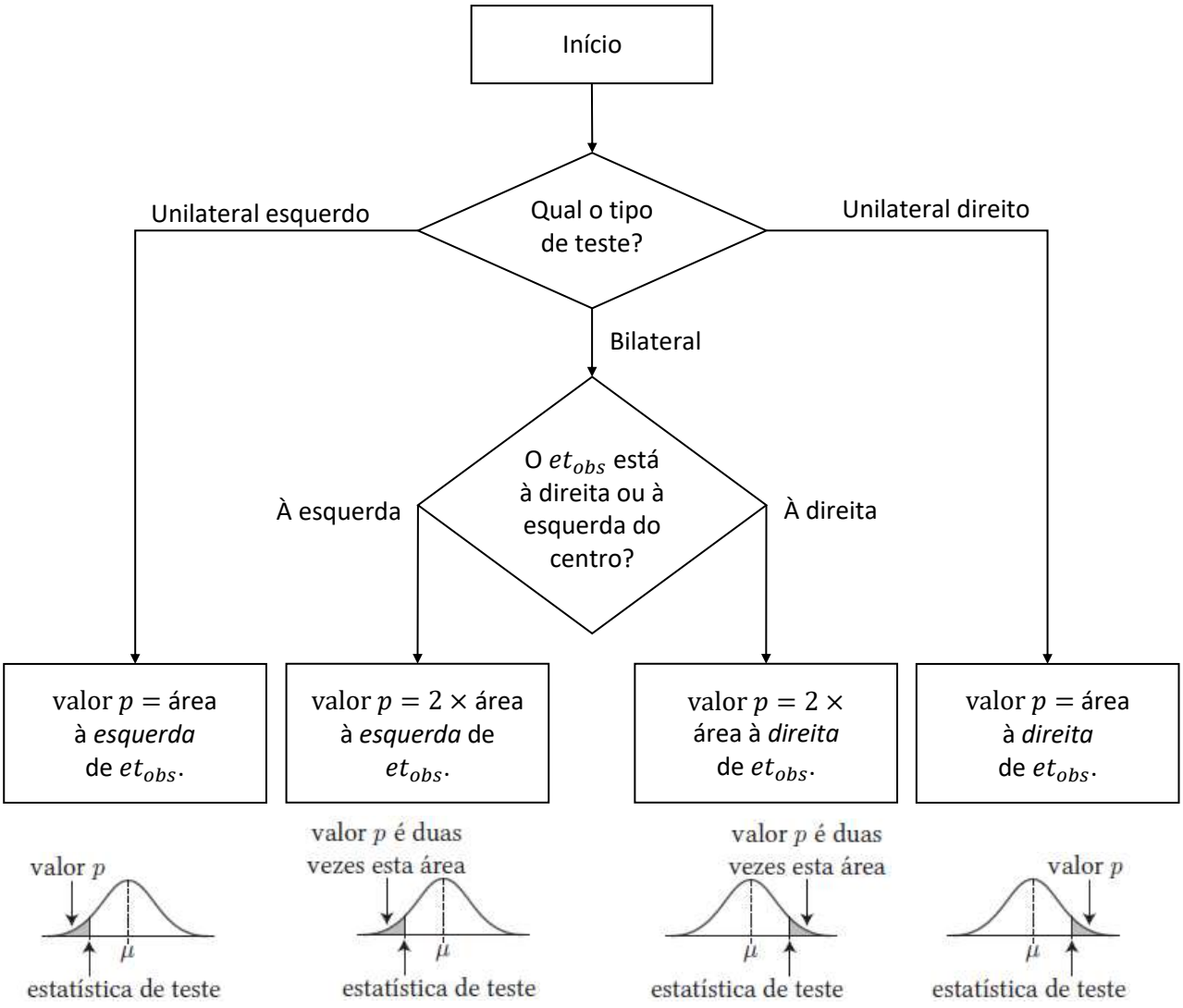


Figura 8.4: Determinação do valor p com base em distribuições simétricas (adaptado de Triola, 2017).

8.3 Teste de hipótese para a média

Considere uma população com média μ e desvio padrão σ da qual se extraiu aleatoriamente uma amostra de dimensão n, e para a qual que pretende realizar um teste de hipóteses para a média populacional μ.

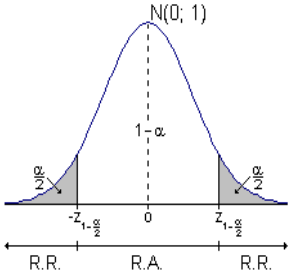
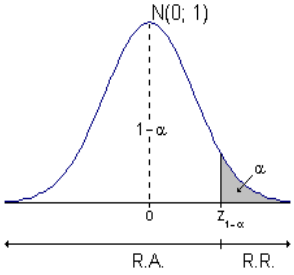
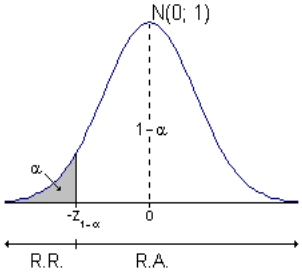
8.3.1 Variância conhecida

Quando a população é Normal com variância conhecida, a estatística de teste a utilizar é:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1),$$

onde μ_0 representa o valor de μ que se assume como sendo o verdadeiro até prova em contrário, ou seja, o valor que se assume para μ em H_0.

Se a distribuição da população não for Normal, mas a amostra for de grande dimensão, então a estatística de teste anterior $Z \overset{\circ}{\sim} N(0;1)$.

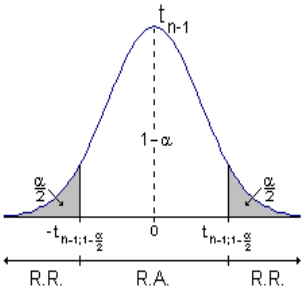
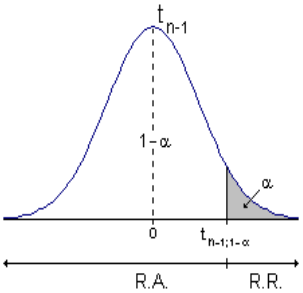
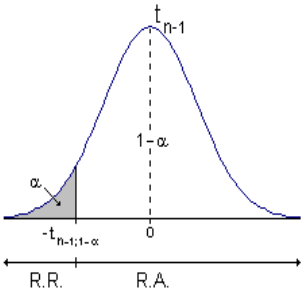
T. bilateral	T. unilateral direito	T. unilateral esquerdo
Hipóteses a testar: $H_0: \mu = \mu_0 \; vs \; H_1: \mu \neq \mu_0$	**Hipóteses a testar:** $H_0: \mu \leq \mu_0 \; vs \; H_1: \mu > \mu_0$	**Hipóteses a testar:** $H_0: \mu \geq \mu_0 \; vs \; H_1: \mu < \mu_0$
Regiões críticas:  $R.A.: \left]-z_{1-\frac{\alpha}{2}};\; z_{1-\frac{\alpha}{2}}\right[$ $R.R.: \left]-\infty;\; -z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}}; +\infty\right[$	**Regiões críticas:**  $R.A.:]-\infty;\; z_{1-\alpha}[$ $R.R.: [z_{1-\alpha}; +\infty[$	**Regiões críticas:**  $R.A.:]-z_{1-\alpha}; +\infty[$ $R.R.:]-\infty;\; -z_{1-\alpha}]$
Regra de decisão: Rejeitar H_0 quando $\lvert z_{obs}\rvert \geq z_{1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \geq z_{1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \leq -z_{1-\alpha}$
Cálculo do valor p: valor $p = 2 \times P(Z \geq \lvert z_{obs}\rvert)$ $= 2 \times \left(1 - \Phi(\lvert z_{obs}\rvert)\right)$	**Cálculo do valor p:** valor $p = P(Z \geq z_{obs})$ $= 1 - \Phi(z_{obs})$	**Cálculo do valor p:** valor $p = P(Z \leq z_{obs})$ $= \Phi(z_{obs})$

8.3.2 Variância desconhecida

Se a população for Normal com variância desconhecida, então a estatística de teste a utilizar é:

$$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1},$$

onde μ_0 representa o valor de μ que se assume como sendo o verdadeiro até prova em contrário, ou seja, o valor que se assume para μ em H_0.

T. bilateral	T. unilateral direito	T. unilateral esquerdo
Hipóteses a testar: $H_0: \mu = \mu_0 \; vs \; H_1: \mu \neq \mu_0$	**Hipóteses a testar:** $H_0: \mu \leq \mu_0 \; vs \; H_1: \mu > \mu_0$	**Hipóteses a testar:** $H_0: \mu \geq \mu_0 \; vs \; H_1: \mu < \mu_0$
Regiões críticas: 	**Regiões críticas:** 	**Regiões críticas:** 

$R.A.: \left]-t_{n-1;1-\frac{\alpha}{2}};\ t_{n-1;1-\frac{\alpha}{2}}\right[$ $R.R.: \left]-\infty;\ -t_{n-1;1-\frac{\alpha}{2}}\right] \cup \left[t_{n-1;1-\frac{\alpha}{2}};+\infty\right[$	$R.A.: \left]-\infty;\ t_{n-1;1-\alpha}\right[$ $R.R.: \left[t_{n-1;1-\alpha};+\infty\right[$	$R.A.: \left]-t_{n-1;1-\alpha};\ +\infty\right[$ $R.R.: \left]-\infty;\ -t_{n-1;1-\alpha}\right]$
Regra de decisão: Rejeitar H_0 quando $\lvert t_{obs}\rvert \geq t_{n-1;1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $t_{obs} \geq t_{n-1;1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $t_{obs} \leq -t_{n-1;1-\alpha}$
Cálculo do valor p: $valor\text{-}p = 2 \times P(T \geq \lvert t_{obs}\rvert)$	**Cálculo do valor p:** $valor\text{-}p = P(T \geq t_{obs})$	**Cálculo do valor p:** $valor\text{-}p = P(T \leq t_{obs})$

Quando a distribuição da população não é Normal, mas a amostra é de grande dimensão então a estatística de teste a usar é

$$Z = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \overset{\circ}{\sim} N(0;1),$$

e desta forma deve ser usada a distribuição Normal tanto no cálculo das regiões críticas como do valor p. Existem alguns programas estatísticos (por exemplo, SPSS) que, nestas condições, não utilizam a distribuição Normal mas a *t*-Student o que, apesar de não ser teoricamente correto, não traz consequências práticas e simplifica a sua aplicação, como já foi referido anteriormente em situações análogas.

8.4 Teste de hipótese para a diferença de 2 médias

Considere duas populações, com médias μ_1 e μ_2 e desvios padrão σ_1 e σ_2, das quais se extraíram aleatoriamente duas amostras independentes com dimensão n_1 e n_2, respectivamente. Pretende-se realizar um teste de hipóteses para comparar as duas médias populacionais μ_1 e μ_2, i. e., para a diferença entre as duas médias $\mu_1 - \mu_2$.

Nas secções seguintes considera-se que $(\mu_1 - \mu_2)_0 = \mu_0$.

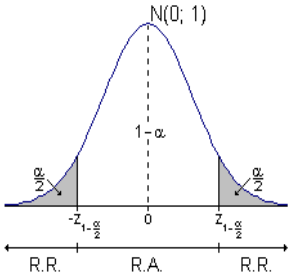
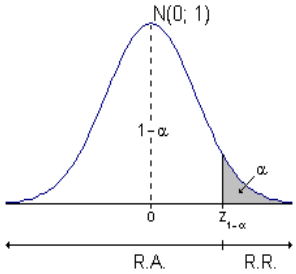
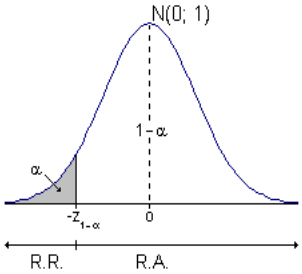
8.4.1 Quando as variâncias são conhecidas

Se as populações forem Normais com desvios-padrão σ_1 e σ_2, respectivamente, então a estatística de teste a utilizar é:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0;1),$$

onde $(\mu_1 - \mu_2)_0$ representa o valor de $\mu_1 - \mu_2$ que se assume como sendo o verdadeiro até prova em contrário, ou seja, o valor que se assume para $\mu_1 - \mu_2$ em H_0.

Se as populações não forem Normais, mas as amostras forem de grande dimensão, então a estatística de teste anterior $Z \overset{\circ}{\sim} N(0;1)$.

T. bilateral	T. unilateral direito	T. unilateral esquerdo
Hipóteses a testar: $H_0: \mu_1 - \mu_2 = \mu_0 \; vs$ $H_1: \mu_1 - \mu_2 \neq \mu_0$	**Hipóteses a testar:** $H_0: \mu_1 - \mu_2 \leq \mu_0 \; vs$ $H_1: \mu_1 - \mu_2 > \mu_0$	**Hipóteses a testar:** $H_0: \mu_1 - \mu_2 \geq \mu_0 \; vs$ $H_1: \mu_1 - \mu_2 < \mu_0$
Regiões críticas:  $R.A.: \left]-z_{1-\frac{\alpha}{2}};\; z_{1-\frac{\alpha}{2}}\right[$ $R.R.: \left]-\infty;\; -z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}}; +\infty\right[$	**Regiões críticas:**  $R.A.:]-\infty;\; z_{1-\alpha}[$ $R.R.: [z_{1-\alpha}; +\infty[$	**Regiões críticas:**  $R.A.:]-z_{1-\alpha}; +\infty[$ $R.R.:]-\infty;\; -z_{1-\alpha}]$
Regra de decisão: Rejeitar H_0 quando $\lvert z_{obs}\rvert \geq z_{1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \geq z_{1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \leq -z_{1-\alpha}$
Cálculo do valor p: $\text{valor } p = 2 \times P(Z \geq \lvert z_{obs}\rvert)$ $= 2 \times \left(1 - \Phi(\lvert z_{obs}\rvert)\right)$	**Cálculo do valor p:** $\text{valor } p = P(Z \geq z_{obs})$ $= 1 - \Phi(z_{obs})$	**Cálculo do valor p:** $\text{valor } p = P(Z \leq z_{obs})$ $= \Phi(z_{obs})$

8.4.2 Quando as variâncias são desconhecidas e iguais

Quando as duas amostras independentes são retiradas de duas populações com distribuição Normal com desvios-padrão σ_1 e σ_2, respectivamente, desconhecidos, mas cuja igualdade $(\sigma_1 = \sigma_2)$ pode ser considerada para um determinado nível de confiança[†], a estatística de teste a utilizar é:

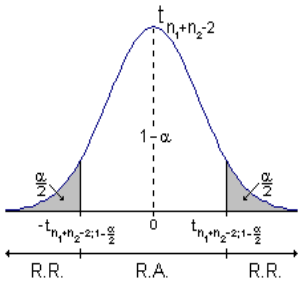
$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

onde $(\mu_1 - \mu_2)_0$ representa o valor que se assume para $\mu_1 - \mu_2$ em H_0.

T. bilateral	T. unilateral direito	T. unilateral esquerdo
Hipóteses a testar: $H_0: \mu_1 - \mu_2 = \mu_0 \; vs$ $H_1: \mu_1 - \mu_2 \neq \mu_0$	**Hipóteses a testar:** $H_0: \mu_1 - \mu_2 \leq \mu_0 \; vs$ $H_1: \mu_1 - \mu_2 > \mu_0$	**Hipóteses a testar:** $H_0: \mu_1 - \mu_2 \geq \mu_0 \; vs$ $H_1: \mu_1 - \mu_2 < \mu_0$

[†] Adiante abordar-se o teste de hipótese para a razão de variâncias que permite testar o pressuposto de igualdade das variâncias, podendo este também ser avaliado com recurso ao intervalo de confiança para a razão de variâncias.

Regiões críticas:



$R.A.: \left]-t_{n_1+n_2-2;1-\frac{\alpha}{2}};\ t_{n_1+n_2-2;1-\frac{\alpha}{2}}\right[$

$R.R.: \left]-\infty;\ -t_{n_1+n_2-2;1-\frac{\alpha}{2}}\right] \cup \left[t_{n_1+n_2-2;1-\frac{\alpha}{2}};+\infty\right[$

Regra de decisão:

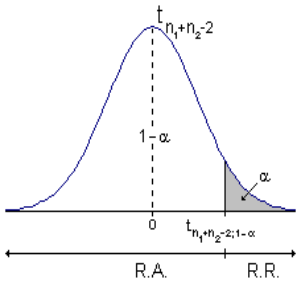
Rejeitar H_0 quando

$|t_{obs}| \geq t_{n_1+n_2-2;1-\frac{\alpha}{2}}$

Cálculo do valor p:

valor $p = 2 \times P(T \geq |t_{obs}|)$

Regiões críticas:



$R.A.: \left]-\infty;\ t_{n_1+n_2-2;1-\alpha}\right[$

$R.R.: \left[t_{n_1+n_2-2;1-\alpha};+\infty\right[$

Regra de decisão:

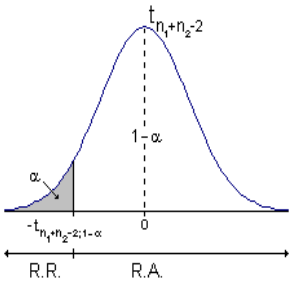
Rejeitar H_0 quando

$t_{obs} \geq t_{n_1+n_2-2;1-\alpha}$

Cálculo do valor p:

valor $p = P(T \geq t_{obs})$

Regiões críticas:



$R.A.: \left]-t_{n_1+n_2-2;1-\alpha};\ +\infty\right[$

$R.R.: \left]-\infty;\ -t_{n_1+n_2-2;1-\alpha}\right]$

Regra de decisão:

Rejeitar H_0 quando

$t_{obs} \leq -t_{n_1+n_2-2;1-\alpha}$

Cálculo do valor p:

valor $p = P(T \leq t_{obs})$

Se as populações não forem Normais, mas as amostras forem de grande dimensão então a estatística de teste anterior segue aproximadamente uma $N(0;1)$, sendo válidos os comentários anteriores.

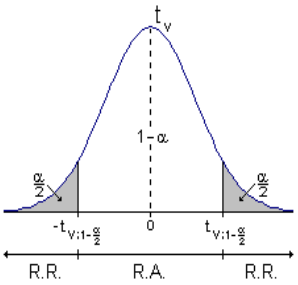
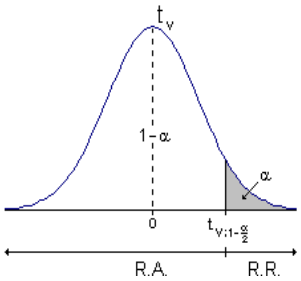
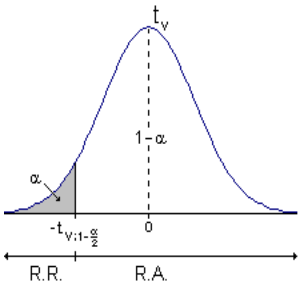
8.4.3 Quando as variâncias são desconhecidas e diferentes

Quando as duas amostras independentes são retiradas de duas populações com distribuição Normal com desvios-padrão σ_1 e σ_2, respectivamente, desconhecidos, mas onde existe evidência que são diferentes $(\sigma_1 \neq \sigma_2)$ para um determinado nível de confiança[†], a estatística de teste a utilizar é:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{\circ}{\sim} t_v, \text{ onde } v = \left[\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2}\right],$$

sendo $[r]$ a parte inteira de r (ou seja, arredondar o valor obtido por defeito) e $(\mu_1 - \mu_2)_0$ representa o valor que se assume para $\mu_1 - \mu_2$ em H_0.

[†] Adiante abordar-se o teste de hipótese para a razão de variâncias que permite testar o pressuposto de igualdade das variâncias, podendo também ser avaliado com recurso ao intervalo de confiança para a razão de variâncias.

T. bilateral	T. unilateral direito	T. unilateral esquerdo
Hipóteses a testar: $H_0: \mu_1 - \mu_2 = \mu_0 \; vs$ $H_1: \mu_1 - \mu_2 \neq \mu_0$	**Hipóteses a testar:** $H_0: \mu_1 - \mu_2 \leq \mu_0 \; vs$ $H_1: \mu_1 - \mu_2 > \mu_0$	**Hipóteses a testar:** $H_0: \mu_1 - \mu_2 \geq \mu_0 \; vs$ $H_1: \mu_1 - \mu_2 < \mu_0$
Regiões críticas:  $R.A.: \left]-t_{v;1-\frac{\alpha}{2}};\; t_{n-1;1-\frac{\alpha}{2}}\right[$ $R.R.: \left]-\infty;\; -t_{v;1-\frac{\alpha}{2}}\right] \cup \left[t_{v;1-\frac{\alpha}{2}};+\infty\right[$	**Regiões críticas:**  $R.A.: \left]-\infty;\; t_{v;\,1-\alpha}\right[$ $R.R.: \left[t_{v;\,1-\alpha};+\infty\right[$	**Regiões críticas:**  $R.A.: \left]-t_{v;1-\alpha};\; +\infty\right[$ $R.R.: \left]-\infty;\; -t_{v;1-\alpha}\right]$
Regra de decisão: Rejeitar H_0 quando $\lvert t_{obs}\rvert \geq t_{v;1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $t_{obs} \geq t_{v;1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $t_{obs} \leq -t_{v;1-\alpha}$
Cálculo do valor p: valor $p = 2 \times P(T \geq \lvert t_{obs}\rvert)$	**Cálculo do valor p:** valor $p = P(T \geq t_{obs})$	**Cálculo do valor p:** valor $p = P(T \leq t_{obs})$

Mais uma vez, se as populações não forem Normais, mas as amostras forem de grande dimensão então a estatística de teste anterior segue aproximadamente uma $N(0;1)$, sendo válidos os comentários anteriores.

8.5 Teste de hipótese para a proporção

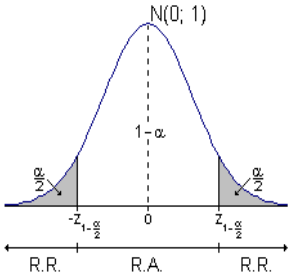
Considere uma população Bernoulli, com parâmetro p,da qual se retirou uma amostra aleatória suficientemente grande e para a qual se pretende realizar um teste de hipóteses para a proporção populacional p. A estatística de teste a utilizar é:

$$Z = \frac{\overline{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{\circ}{\sim} N(0;1),$$

onde p_0 representa o valor que se assume para p em H_0.

T. bilateral	T. unilateral direito	T. unilateral esquerdo
Hipóteses a testar: $H_0: p = p_0 \; vs \; H_1: p \neq p_0$	**Hipóteses a testar:** $H_0: p \leq p_0 \; vs \; H_1: p > p_0$	**Hipóteses a testar:** $H_0: p \geq p_0 \; vs \; H_1: p < p_0$

Regiões críticas:



$$R.A.: \left]-z_{1-\frac{\alpha}{2}};\ z_{1-\frac{\alpha}{2}}\right[$$
$$R.R.: \left]-\infty;\ -z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}};+\infty\right[$$

Regra de decisão:

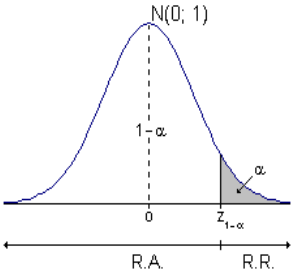
Rejeitar H_0 quando

$$|z_{obs}| \geq z_{1-\frac{\alpha}{2}}$$

Cálculo do valor p:

$$\text{valor } p = 2 \times P(Z \geq |z_{obs}|)$$
$$= 2 \times \left(1 - \Phi(|z_{obs}|)\right)$$

Regiões críticas:



$$R.A.:]-\infty;\ z_{1-\alpha}[$$
$$R.R.: [z_{1-\alpha};+\infty[$$

Regra de decisão:

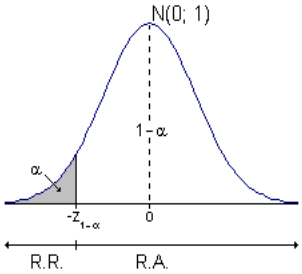
Rejeitar H_0 quando

$$z_{obs} \geq z_{1-\alpha}$$

Cálculo do valor p:

$$\text{valor } p = P(Z \geq z_{obs})$$
$$= 1 - \Phi(z_{obs})$$

Regiões críticas:



$$R.A.:]-z_{1-\alpha};+\infty[$$
$$R.R.:]-\infty;\ -z_{1-\alpha}]$$

Regra de decisão:

Rejeitar H_0 quando

$$z_{obs} \leq -z_{1-\alpha}$$

Cálculo do valor p:

$$\text{valor } p = P(Z \leq z_{obs})$$
$$= \Phi(z_{obs})$$

8.6 Teste de hipótese para a diferença de proporções

Considere duas populações Bernoulli, com parâmetros p_1 e p_2 das quais se extraíram aleatoriamente duas amostras independentes de grande dimensão n_1 e n_2, respectivamente. Pretende-se realizar um teste de hipóteses para comparar as duas proporções amostrais p_1 e p_2, i. e., para a diferença de proporções $p_1 - p_2$. Para este teste sabe-se que

$$Z = \frac{(\overline{P}_1 - \overline{P}_2) - (p_1 - p_2)_0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \overset{\circ}{\sim} N(0;1),$$

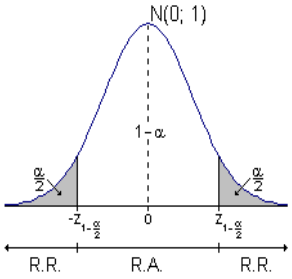
onde $(p_1 - p_2)_0$ representa o valor que se assume para $p_1 - p_2$ em H_0.

A expressão apresentada no denominador não é conhecida, apenas se conhecendo o valor da diferença $(p_1 - p_2)$ sob H_0. Como habitualmente este teste é realizado considerando $p_1 - p_2 = 0$, ou seja, $p_1 = p_2 = p$, e como esta proporção é desconhecida é substituída pelo seu estimador consistente. Desta forma, a estatística de teste a utilizar é:

$$Z = \frac{(\overline{P}_1 - \overline{P}_2) - (p_1 - p_2)_0}{\sqrt{\overline{P}^*\left(1 - \overline{P}^*\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{\circ}{\sim} N(0;1), \text{ onde } \overline{P}^* = \frac{n_1\overline{P}_1 + n_2\overline{P}_2}{n_1 + n_2}.$$

T. bilateral	**T. unilateral direito**	**T. unilateral esquerdo**
Hipóteses a testar:	**Hipóteses a testar:**	**Hipóteses a testar:**
$H_0: p_1 - p_2 = p_0\ vs$	$H_0: p_1 - p_2 \leq p_0\ vs$	$H_0: p_1 - p_2 \geq p_0\ vs$
$H_1: p_1 - p_2 \neq p_0$	$H_1: p_1 - p_2 > p_0$	$H_1: p_1 - p_2 < p_0$

Regiões críticas:



$$R.A.: \left]-z_{1-\frac{\alpha}{2}};\ z_{1-\frac{\alpha}{2}}\right[$$
$$R.R.: \left]-\infty;\ -z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}};+\infty\right[$$

Regra de decisão:

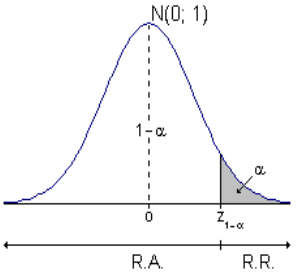
Rejeitar H_0 quando

$$|z_{obs}| \geq z_{1-\frac{\alpha}{2}}$$

Cálculo do valor p:

$$\text{valor } p = 2 \times P(Z \geq |z_{obs}|)$$
$$= 2 \times \left(1 - \Phi(|z_{obs}|)\right)$$

Regiões críticas:



$$R.A.:]-\infty;\ z_{1-\alpha}[$$
$$R.R.: [z_{1-\alpha};+\infty[$$

Regra de decisão:

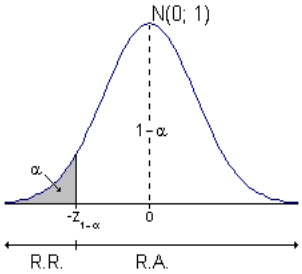
Rejeitar H_0 quando

$$z_{obs} \geq z_{1-\alpha}$$

Cálculo do valor p:

$$\text{valor } p = P(Z \geq z_{obs})$$
$$= 1 - \Phi(z_{obs})$$

Regiões críticas:



$$R.A.:]-z_{1-\alpha};+\infty[$$
$$R.R.:]-\infty;\ -z_{1-\alpha}]$$

Regra de decisão:

Rejeitar H_0 quando

$$z_{obs} \leq -z_{1-\alpha}$$

Cálculo do valor p:

$$\text{valor } p = P(Z \leq z_{obs})$$
$$= \Phi(z_{obs})$$

8.7 Teste de hipótese para a variância

Considere uma população Normal, com média μ e desvio padrão σ, da qual se extraiu aleatoriamente uma amostra de dimensão n, e para a qual se pretende realizar um teste de hipóteses para a variância populacional σ^2. A estatística de teste a utilizar é:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}.$$

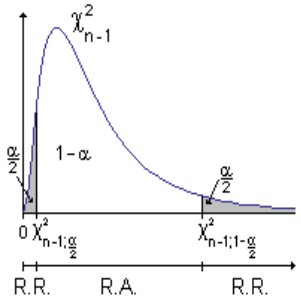
onde σ_0^2 representa o valor que se assume para σ^2 em H_0.

T. bilateral

Hipóteses a testar:

$$H_0: \sigma^2 = \sigma_0^2 \ vs\ H_1: \sigma^2 \neq \sigma_0^2$$

Regiões críticas:

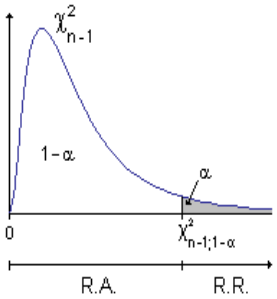


T. unilateral direito

Hipóteses a testar:

$$H_0: \sigma^2 \leq \sigma_0^2 \ vs\ H_1: \sigma^2 > \sigma_0^2$$

Regiões críticas:

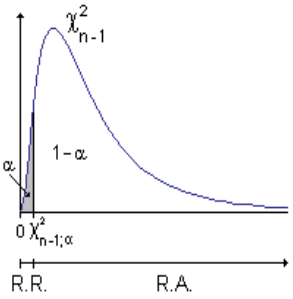


T. unilateral esquerdo

Hipóteses a testar:

$$H_0: \sigma^2 \geq \sigma_0^2 \ vs\ H_1: \sigma^2 < \sigma_0^2$$

Regiões críticas:



$R.A.: \left]\chi^2_{n-1;\,\frac{\alpha}{2}};\ \chi^2_{n-1;1-\frac{\alpha}{2}}\right[$ $R.R.: \left[0;\ \chi^2_{n-1;\,\frac{\alpha}{2}}\right] \cup \left[\chi^2_{n-1;\,1-\frac{\alpha}{2}};+\infty\right[$	$R.A.: \left[0;\ \chi^2_{n-1;\,1-\alpha}\right[$ $R.R.: \left[\chi^2_{n-1;\,1-\alpha};+\infty\right[$	$R.A.: \left]\chi^2_{n-1;\,\alpha};\ +\infty\right[$ $R.R.: \left[0;\ \chi^2_{n-1;\,\alpha}\right]$
Regra de decisão: Rejeitar H_0 quando $\chi^2_{obs} \leq \chi^2_{n-1;\,\frac{\alpha}{2}}$ ou $\chi^2_{obs} \geq \chi^2_{n-1;\,1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $\chi^2_{obs} \geq \chi^2_{n-1;\,1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $\chi^2_{obs} \leq \chi^2_{n-1;\,\alpha}$
Cálculo do valor p: $\text{valor } p = 2 \times min\{P(\chi^2 \leq \chi^2_{obs});\ P(\chi^2 \geq \chi^2_{obs})\}$	**Cálculo do valor p:** $\text{valor } p = P(\chi^2 \geq \chi^2_{obs}) = 1 - P(\chi^2 < \chi^2_{obs})$	**Cálculo do valor p:** $\text{valor } p = P(\chi^2 \leq \chi^2_{obs})$

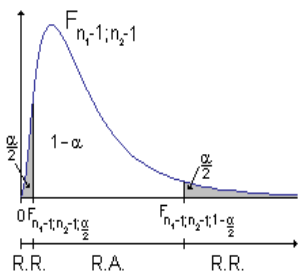
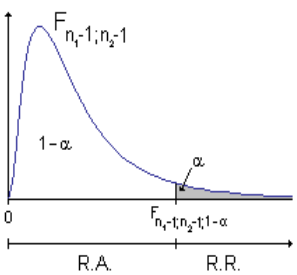
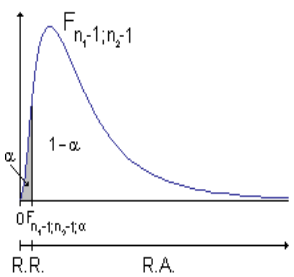
8.8 Teste de hipótese para o quociente de variâncias

Considere que duas populações Normais, com médias μ_1 e μ_2 e desvios padrão σ_1 e σ_2, das quais se extraíram aleatoriamente duas amostras independentes com dimensão n_1 e n_2, respectivamente. Pretende-se realizar um teste de hipóteses para comparar as variâncias populacionais σ_1^2 e σ_2^2, i. e., para o quociente de variâncias (σ_2^2/σ_1^2). A estatística de teste a utilizar é:

$$F = \frac{S_1^2}{S_2^2}\left(\frac{\sigma_2^2}{\sigma_1^2}\right)_0 \sim F_{n_1-1;\,n_2-1},$$

onde $(\sigma_2^2/\sigma_1^2)_0$ representa o valor que se assume para (σ_2^2/σ_1^2) em H_0.

Nesta secção considera-se apenas a situação em que $(\sigma_2^2/\sigma_1^2)_0 = 1$, o que equivale a comparar a igualdade das variâncias populacionais.

T. bilateral	**T. unilateral direito**	**T. unilateral esquerdo**
Hipóteses a testar: $H_0: \sigma_1^2 = \sigma_2^2\ vs\ H_1: \sigma_1^2 \neq \sigma_2^2 \Leftrightarrow$ $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1\ vs\ H_1: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$	**Hipóteses a testar:** $H_0: \sigma_1^2 \leq \sigma_0^2\ vs: \sigma_1^2 > \sigma_2^2 \Leftrightarrow$ $H_0: \frac{\sigma_1^2}{\sigma_2^2} \leq 1\ vs\ H_1: \frac{\sigma_1^2}{\sigma_2^2} > 1$	**Hipóteses a testar:** $H_0: \sigma_1^2 \geq \sigma_2^2\ vs\ H_1: \sigma_1^2 < \sigma_2^2 \Leftrightarrow$ $H_0: \frac{\sigma_1^2}{\sigma_2^2} \geq 1\ vs\ H_1: \frac{\sigma_1^2}{\sigma_2^2} < 1$
Regiões críticas: 	**Regiões críticas:** 	**Regiões críticas:** 

$R.A.: \left]f_{n_1-1;\, n_2-1;\, \frac{\alpha}{2}};\ f_{n_1-1;\, n_2-1;\, 1-\frac{\alpha}{2}}\right[$ $R.R.: \left[0;\ f_{n_1-1;\, n_2-1;\, \frac{\alpha}{2}}\right] \cup \left[f_{n_1-1;\, n_2-1;\, 1-\frac{\alpha}{2}};\ +\infty\right[$	$R.A.: \left[0;\ f_{n_1-1;\, n_2-1;\, 1-\alpha}\right[$ $R.R.: \left[f_{n_1-1;\, n_2-1;\, 1-\alpha};\ +\infty\right[$	$R.A.: \left]f_{n_1-1;\, n_2-1;\, \alpha};\ +\infty\right[$ $R.R.: \left[0;\ f_{n_1-1;\, n_2-1;\, \alpha}\right]$
Regra de decisão: Rejeitar H_0 quando $f_{obs} \leq f_{n_1-1;\, n_2-1;\, \frac{\alpha}{2}} = \frac{1}{f_{n_2-1;\, n_1-1;\, 1-\frac{\alpha}{2}}}$ ou $f_{obs} \geq f_{n_1-1;\, n_2-1;\, 1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $f_{obs} \geq f_{n_1-1;\, n_2-1;\, 1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $f_{obs} \leq f_{n_1-1;\, n_2-1;\, \alpha} = \frac{1}{f_{n_2-1;\, n_1-1;\, 1-\alpha}}$
Cálculo do valor p: valor $p = 2 \times min\{P(F \leq f_{obs});\ P(F \geq f_{obs})\}$	**Cálculo do valor p:** valor $p = P(F \geq f_{obs}) = 1 - P(F < f_{obs})$	**Cálculo do valor p:** valor $p = P(F \leq f_{obs})$

8.9 Teste de hipótese para amostras emparelhadas

Considere-se agora o caso em que as duas amostras formam um par de observações $(X_{1i};\ X_{2i})$, $i = 1, \dots, n$, ou seja, trata-se de uma amostra emparelhada. Os pares de observações são independentes e retirados de populações Normais, com médias μ_1 e μ_2 e desvios padrão σ_1 e σ_2, respectivamente. Neste caso, para testar a igualdade entre as médias populacionais, as hipóteses a formular são:

T. bilateral	**T. unilateral direito**	**T. unilateral esquerdo**
Hipóteses a testar: $H_0: \mu_1 = \mu_2\ vs\ H_1: \mu_1 \neq \mu_2$ $\Leftrightarrow H_0: \mu_1 - \mu_2 = 0\ vs$ $H_1: \mu_1 - \mu_2 \neq 0$	**Hipóteses a testar:** $H_0: \mu_1 \leq \mu_2\ vs\ H_1: \mu_1 > \mu_2$ $\Leftrightarrow H_0: \mu_1 - \mu_2 \leq 0\ vs$ $H_1: \mu_1 - \mu_2 > 0$	**Hipóteses a testar:** $H_0: \mu_1 \geq \mu_2\ vs\ H_1: \mu_1 < \mu_2$ $\Leftrightarrow H_0: \mu_1 - \mu_2 \geq 0\ vs$ $H_1: \mu_1 - \mu_2 < 0$

Para realizar o teste pretendido calcular:

- $D_i = X_{1i} - X_{2i}, \qquad i = 1, \dots, n;$
- $\overline{D} = \overline{X}_1 - \overline{X}_2 = \sum_{i=1}^{n} \frac{X_{1i}}{n} - \sum_{i=1}^{n} \frac{X_{2i}}{n} = \sum_{i=1}^{n} \frac{D_i}{n};$
- $S_D^2 = \sum_{i=1}^{n} \frac{(D_i - \bar{D})^2}{n-1}.$

Desta forma, as hipóteses a testar podem escritas da seguinte forma:

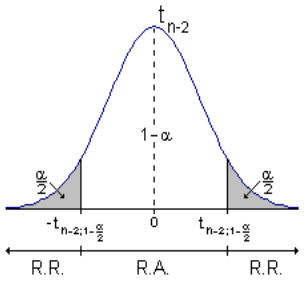
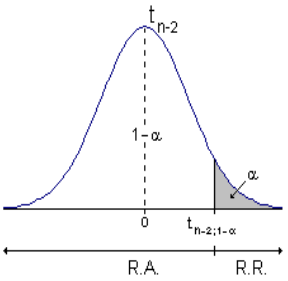
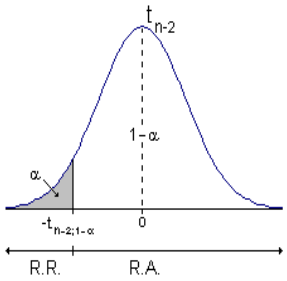
Hipóteses a testar: $H_0: \mu_D = 0\ vs\ H_1: \mu_D \neq 0$	**Hipóteses a testar:** $H_0: \mu_D \leq 0\ vs\ H_1: \mu_D > 0$	**Hipóteses a testar:** $H_0: \mu_D \geq 0\ vs\ H_1: \mu_D < 0$

Ou seja, está-se perante um teste de hipóteses para a média no caso em que a população segue uma distribuição Normal da qual se desconhece a sua variância. Esta situação já foi descrita nos capítulos anteriores (secção 8.3.2).

8.10 Teste de hipótese para o coeficiente de correlação

Considere uma amostra aleatória bidimensional $(X_i;\ Y_i)$ de dimensão n, ou seja, constituída por n pares de valores independentes e retirados de uma população Normal bivariada com correlação linear ρ. Suponha-se que se pretende saber se *existe correlação linear* entre essas duas variáveis numéricas, ou ainda se *existe correlação linear positiva ou negativa* entre elas. Neste caso, a estatística de teste a utilizar é:

$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2}.$$

T. bilateral	**T. unilateral direito**	**T. unilateral esquerdo**
Hipóteses a testar: $H_0: \rho = 0\ vs\ H_1: \rho \neq 0$	**Hipóteses a testar:** $H_0: \rho \leq 0\ vs\ H_1: \rho > 0$	**Hipóteses a testar:** $H_0: \rho \geq 0\ vs\ H_1: \rho < 0$
Regiões críticas:  $R.A.: \left]-t_{n-2;1-\frac{\alpha}{2}};\ t_{n-2;1-\frac{\alpha}{2}}\right[$ $R.R.: \left]-\infty;\ -t_{n-2;1-\frac{\alpha}{2}}\right] \cup \left[t_{n-2;1-\frac{\alpha}{2}};+\infty\right[$	**Regiões críticas:**  $R.A.: \left]-\infty;\ t_{n-2;1-\alpha}\right[$ $R.R.: \left[t_{n-2;1-\alpha};+\infty\right[$	**Regiões críticas:**  $R.A.: \left]-t_{n-2;1-\alpha};\ +\infty\right[$ $R.R.: \left]-\infty;\ -t_{n-2;1-\alpha}\right]$
Regra de decisão: Rejeitar H_0 quando $\lvert t_{obs}\rvert \geq t_{n-2;1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $t_{obs} \geq t_{n-2;1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $t_{obs} \leq -t_{n-2;1-\alpha}$
Cálculo do valor p: valor $p = 2 \times P(T \geq \lvert t_{obs}\rvert)$	**Cálculo do valor p:** valor $p = P(T \geq t_{obs})$	**Cálculo do valor p:** valor $p = P(T \leq t_{obs})$

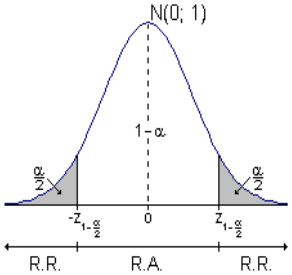
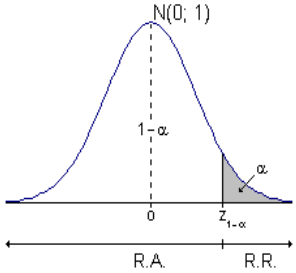
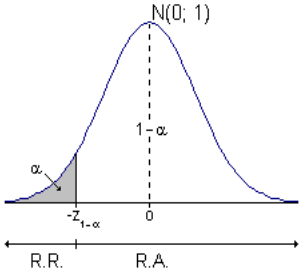
Caso se pretenda testar se o coeficiente de correlação populacional é ou não significativamente *diferente de um determinado valor* ρ_0, então a estatística de teste a utilizar é:

$$Z = \frac{Z_R - Z_{\rho_0}}{\frac{1}{\sqrt{n-3}}} \sim N(0;1),$$

uma vez que sempre que ρ se afasta de zero a distribuição amostral torna-se assimétrica. Z_{ρ_0} o valor de Z_ρ assumido em H_0.

Z_R e Z_{ρ_0} são obtidos através das expressões:

$$Z_R = \frac{1}{2} ln\left(\frac{1+R}{1-R}\right) \text{ e } Z_{\rho_0} = \frac{1}{2} ln\left(\frac{1+\rho_0}{1-\rho_0}\right).$$

T. bilateral	**T. unilateral direito**	**T. unilateral esquerdo**
Hipóteses a testar: $H_0: \rho = \rho_0 \; vs \; H_1: \rho \neq \rho_0$	**Hipóteses a testar:** $H_0: \rho \leq \rho_0 \; vs \; H_1: \rho > \rho_0$	**Hipóteses a testar:** $H_0: \rho \geq \rho_0 \; vs \; H_1: \rho < \rho_0$
Regiões críticas: 	**Regiões críticas:** 	**Regiões críticas:** 
$R.A.: \left]-z_{1-\frac{\alpha}{2}};\; z_{1-\frac{\alpha}{2}}\right[$ $R.R.: \left]-\infty;\; -z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}}; +\infty\right[$	$R.A.:]-\infty;\; z_{1-\alpha}[$ $R.R.: [z_{1-\alpha}; +\infty[$	$R.A.:]-z_{1-\alpha}; +\infty[$ $R.R.:]-\infty;\; -z_{1-\alpha}]$
Regra de decisão: Rejeitar H_0 quando $\lvert z_{obs}\rvert \geq z_{1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \geq z_{1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \leq -z_{1-\alpha}$
Cálculo do valor p: $\text{valor } p = 2 \times P(Z \geq \lvert z_{obs}\rvert)$ $= 2 \times \left(1 - \Phi(\lvert z_{obs}\rvert)\right)$	**Cálculo do valor p:** $\text{valor } p = P(Z \geq z_{obs})$ $= 1 - \Phi(z_{obs})$	**Cálculo do valor p:** $\text{valor } p = P(Z \leq z_{obs})$ $= \Phi(z_{obs})$

8.11 Determinação de valores-p unilaterais com base em valores-p bilaterais

Algunsprogramas estatísticos apenas calculam valores p associados a testes de hipóteses bilaterais (por exemplo: SPSS). Nas distribuições simétricas (caso da Normal e da *t*-Student) existe uma forma relativamente simples de deduzir os valores p unilaterais com base nos valores p bilaterais. Nas distribuições assimétricas (por exemplo: Qui-quadrado e F) a forma já não é tão intuitiva, não sendo aqui explorada.

Assim, nas estatísticas de teste que seguem distribuições simétricas, a dedução dos valores p unilaterais (valor p_{uni}) com base em valores p bilaterais (valor p_{bil}) pode ser apresentada da seguinte forma:

- Verificar se a hipótese H_1 gera uma proposição verdadeira ou falsa quando se substitui o(s) parâmetro(s) em teste pelas estatísticas calculadas com base na amostra.
    - Se gerar uma proposição verdadeira então:

$$\text{valor } p_{uni} = \frac{\text{valor } p_{bil}}{2}.$$

- Se gerar uma proposição falsa então:

$$\text{valor } p_{uni} = 1 - \frac{\text{valor } p_{bil}}{2}.$$

A explicação para estes passos pode ser muito facilmente e intuitivamente explicada com base em imagens das distribuições, imaginando onde estão as áreas de rejeição e onde está a ser projectado o valor observado da estatística de teste.

Exemplo 1 (1 amostra): Recolheu-se uma amostra com média 2,9 e efectuou-se o teste bilateral

$$H_0: \mu = 3 \; vs \; H_1: \mu \neq 3,$$

para o qual se obteve valor $p_{bil} = 0{,}024$.

Suponha-se agora que pretende testar se

$$H_0: \mu \leq 3 \; vs H_1: \mu > 3 \text{ (teste unilateral).}$$

Ao substituir μ por $\overline{x} = 2{,}9$ na hipótese H_1 dará $2{,}9 > 3$, o que é uma proposição falsa. Logo,

$$\text{valor } p_{uni} = 1 - \frac{0{,}024}{2} = 1 - 0{,}012 = 0{,}988.$$

Mas, caso se pretenda testar a hipótese

$$H_0: \mu \geq 3 \; vs \; H_1: \mu < 3,$$

como ao substituir μ por $\overline{x} = 2{,}9$ na hipótese H_1 dará $2{,}9 < 3$, o que é uma preposição verdadeira, então

$$\text{valor } p_{uni} = \frac{0{,}024}{2} = 0{,}012.$$

Exemplo 2 (duas amostras): Recolheram-se duas amostras de 2 populações independentes com médias de $\overline{x}_1 = 2{,}9$ e $\overline{x}_2 = 3{,}1$ e efectuou-se um teste bilateral à diferença das médias populacionais,

$$H_0: \mu_1 - \mu_2 = \mu_0 \; vs. \; H_1: \mu_1 - \mu_2 \neq \mu_0,$$

para o qual se obteve valor $p_{bil} = 0{,}04$.

Suponha-se agora que se pretende testar

$$H_0: \mu_1 - \mu_2 \leq \mu_0 \; vs. H_1: \mu_1 - \mu_2 > \mu_0.$$

Ao substituir μ_1 por $\overline{x}_1 = 2{,}9$ e μ_2 por $\overline{x}_2 = 3{,}1$ na hipótese H_1 dará $2{,}9 - 3{,}1 > 0$, o que é uma proposição falsa. Logo,

$$\text{valor } p_{uni} = 1 - \frac{0{,}04}{2} = 1 - 0{,}02 = 0{,}98.$$

Mas, caso as hipóteses a testar sejam

$$H_0: \mu_1 - \mu_2 \geq \mu_0 \; vs \; H_1: \mu_1 - \mu_2 < \mu_0,$$

como $2{,}9 - 3{,}1 > 0$, i.e., a proposição em H_1 é verdadeira, então

$$\text{valor } p_{uni} = \frac{0{,}04}{2} = 0{,}02.$$

8.12 Quadros resumo

A Tabela 8.2 e a Tabela 8.3 apresentam algumas aproximações apenas para simplificar os conceitos e a sua utilização, mas tendo em consideração as ressalvas feitas anteriormente e que se sustentam na proximidade das distribuições t-Student e $N(0;1)$, para valores elevados de n.

Tabela 8.2: *Quadro resumo dos testes de hipótese paramétricos (1 população).*

H_0	σ^2 conhecido?	Tipo de população	Estatística de Teste	H_1	Rejeitar H_0 se
$\mu = \mu_0$	Sim	Normal (ou qualquer se n grande[†])	$Z = \dfrac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1)$	$\mu < \mu_0$	$z_{obs} \leq -z_{1-\alpha}$
				$\mu \neq \mu_0$	$\lvert z_{obs} \rvert \geq z_{1-\frac{\alpha}{2}}$
				$\mu > \mu_0$	$z_{obs} \geq z_{1-\alpha}$
	Não	Normal (ou qualquer se n grande[†])	$T = \dfrac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$	$\mu < \mu_0$	$t_{obs} \leq -t_{n-1;1-\alpha}$
				$\mu \neq \mu_0$	$\lvert t_{obs} \rvert \geq t_{n-1;1-\frac{\alpha}{2}}$
				$\mu > \mu_0$	$t_{obs} \geq t_{n-1;1-\alpha}$
$p = p_0$ (n grande)	—	Bernoulli	$Z = \dfrac{\overline{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{\circ}{\sim} N(0;1)$	$p < p_0$	$z_{obs} \leq -z_{1-\alpha}$
				$p \neq p_0$	$\lvert z_{obs} \rvert \geq z_{1-\frac{\alpha}{2}}$
				$p > p_0$	$z_{obs} \geq z_{1-\alpha}$
$\sigma^2 = \sigma_0^2$	—	Normal	$\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$	$\sigma^2 < \sigma_0^2$	$\chi^2_{obs} \leq \chi^2_{n-1;\,\alpha}$
				$\sigma^2 \neq \sigma_0^2$	$\chi^2_{obs} \leq \chi^2_{n-1;\frac{\alpha}{2}}$ ou $\chi^2_{obs} \geq \chi^2_{n-1;\,1-\frac{\alpha}{2}}$
				$\sigma^2 > \sigma_0^2$	$\chi^2_{obs} \geq \chi^2_{n-1;\,1-\alpha}$

[†] Nesta situação a Estatística de Teste tem distribuição aproximadamente $N(0;1)$.

Tabela 8.3: Quadro resumo dos testes de hipótese paramétricos (2 populações)

H_0	σ_1^2 e σ_2^2 conhecidos?	Tipo de populações	Estatística de Teste	H_1	Rejeitar H_0 se
$\mu_1 - \mu_2 = \mu_0$	Sim	Normais (ou quaisquer se n_1 e n_2 grandes†)	$Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0;1)$	$\mu_1 - \mu_2 < \mu_0$	$z_{obs} \leq -z_{1-\alpha}$
				$\mu_1 - \mu_2 \neq \mu_0$	$\lvert z_{obs} \rvert \geq z_{1-\frac{\alpha}{2}}$
				$\mu_1 - \mu_2 > \mu_0$	$z_{obs} \geq z_{1-\alpha}$
	Não ($\sigma_1^2 = \sigma_2^2$)	Normais (ou quaisquer se n_1 e n_2 grandes†)	$T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n_1+n_2-2}$	$\mu_1 - \mu_2 < \mu_0$	$t_{obs} \leq -t_{n_1+n_2-2;1-\alpha}$
				$\mu_1 - \mu_2 \neq \mu_0$	$\lvert t_{obs} \rvert \geq t_{n_1+n_2-2;1-\frac{\alpha}{2}}$
				$\mu_1 - \mu_2 > \mu_0$	$t_{obs} \geq t_{n_1+n_2-2;1-\alpha}$
	Não ($\sigma_1^2 \neq \sigma_2^2$)	Normais (ou quaisquer se n_1 e n_2 grandes†)	$T = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \overset{\circ}{\sim} t_v,$ onde $v = \left\lfloor \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{1}{n_1-1}\left(\dfrac{S_1^2}{n_1}\right)^2 + \dfrac{1}{n_2-1}\left(\dfrac{S_2^2}{n_2}\right)^2} \right\rfloor$	$\mu_1 - \mu_2 < \mu_0$	$t_{obs} \leq -t_{v;1-\alpha}$
				$\mu_1 - \mu_2 \neq \mu_0$	$\lvert t_{obs} \rvert \geq t_{v;1-\frac{\alpha}{2}}$
				$\mu_1 - \mu_2 > \mu_0$	$t_{obs} \geq t_{v;1-\alpha}$

† Nesta situação a Estatística de Teste tem distribuição aproximadamente $N(0;1)$.

***Erro! A origem da referência não foi encontrada.**. (continuação)*

H_0	σ_1^2 e σ_2^2 conhecidos?	Tipo de populações	Estatística de Teste	H_1	Rejeitar H_0 se
$p_1 - p_2 = p_0$ (n_1 e n_2 grandes)	—	Bernoulli	$Z = \frac{(\overline{P}_1 - \overline{P}_2) - (p_1 - p_2)_0}{\sqrt{\overline{P}^* \left(1 - \overline{P}^*\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{\circ}{\sim} N(0;1)$, onde $\overline{P}^* = \frac{n_1 \overline{P}_1 + n_2 \overline{P}_2}{n_1 + n_2}$	$p_1 - p_2 < p_0$	$z_{obs} \leq -z_{1-\alpha}$
				$p_1 - p_2 \neq p_0$	$\lvert z_{obs} \rvert \geq z_{1-\frac{\alpha}{2}}$
				$p_1 - p_2 > p_0$	$z_{obs} \geq z_{1-\alpha}$
$\frac{\sigma_1^2}{\sigma_2^2} = \sigma_0^2$	—	Normais	$F = \frac{S_1^2}{S_2^2}\frac{1}{\sigma_0^2} \sim F_{n_1-1;\, n_2-1}$	$\frac{\sigma_1^2}{\sigma_2^2} < \sigma_0^2$	$f_{obs} \leq f_{n_1-1;\, n_2-1;\, \alpha}$
				$\frac{\sigma_1^2}{\sigma_2^2} \neq \sigma_0^2$	$f_{obs} \leq f_{n_1-1;\, n_2-1;\, \frac{\alpha}{2}}$ ou $f_{obs} \geq f_{n_1-1;\, n_2-1;\, 1-\frac{\alpha}{2}}$
				$\frac{\sigma_1^2}{\sigma_2^2} > \sigma_0^2$	$f_{obs} \geq f_{n_1-1;\, n_2-1;\, 1-\alpha}$
$\rho = 0$	—	Normais	$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2}$	$\rho < 0$	$t_{obs} \leq -t_{n-2;1-\alpha}$
				$\rho \neq 0$	$\lvert t_{obs} \rvert \geq t_{n-2;1-\frac{\alpha}{2}}$
				$\rho > 0$	$t_{obs} \geq t_{n-2;1-\alpha}$
$\rho = \rho_0$	—	Normais	$Z = \frac{Z_R - Z_{\rho_0}}{\frac{1}{\sqrt{n-3}}} \sim N(0;1)$, com $Z_R = \frac{1}{2} ln\left(\frac{1+R}{1-R}\right)$	$\rho < \rho_0$	$z_{obs} \leq -z_{1-\alpha}$
				$\rho \neq \rho_0$	$\lvert z_{obs} \rvert \geq z_{1-\frac{\alpha}{2}}$
				$\rho > \rho_0$	$z_{obs} \geq z_{1-\alpha}$

8.13 Exercícios resolvidos

8.13.1 Teste de hipótese para a média

8.13.1.1 Variância conhecida

O operador de telemóveis *Péssimus* pretende controlar o tempo médio de conversações telefónicas no período das 22h às 24h. Para tal seleccionou ao acaso 130 utentes tendo observado os seguintes tempos de conversação telefónica, em minutos, dos utentes no referido período:

0,3	0,3	0,4	0,4	0,4	0,5	0,5	0,5	0,5	0,6	0,6	0,6	0,6	0,6	0,7	0,7	0,7
0,7	0,7	0,7	0,8	0,8	0,8	0,8	0,8	0,8	0,8	0,8	0,9	0,9	0,9	0,9	0,9	0,9
0,9	0,9	0,9	0,9	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,1
1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,1	1,2	1,2	1,2	1,2
1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,2	1,3	1,3	1,3	1,3	1,3	1,3	1,3
1,3	1,3	1,3	1,3	1,3	1,3	1,4	1,4	1,4	1,4	1,4	1,4	1,4	1,4	1,4	1,4	1,5
1,5	1,5	1,5	1,5	1,5	1,5	1,5	1,5	1,6	1,6	1,6	1,6	1,6	1,6	1,6	1,7	1,7
1,7	1,7	1,7	1,8	1,8	1,8	1,9	1,9	2,0	2,0	2,1						

Suponha que o tempo de tais conversações segue uma distribuição Normal com desvio padrão de 1 minuto:

a) Teste ao nível de significância de 1% a hipótese de o tempo médio de conversação ser:
   i. Diferente de 1 minuto.
   ii. Superior a 1 minuto.
   iii. Inferior a 1 minuto.
b) Identifique o tipo de erro que pode estar associado a cada uma das decisões anteriores.
c) Para cada um dos testes da alínea a) calcule o nível de significância a partir do qual rejeita a hipótese nula.
d) Represente graficamente $\beta(\mu_1 = 0)$ e $\beta(\mu_1 = 1{,}3)$. Comente as diferenças obtidas.
e) Calcule a potência de teste para os testes efectuados na alínea a), considerando $\mu_1 = 1{,}3$.

Resolução:

Seja X a v. a. que representa o tempo, em minutos, de conversação telefónica, com $X \sim N(\mu;\ \sigma = 1)$.

$n = 130$ e $\overline{x} = \frac{1}{130}\sum_{i=1}^{130} x_i = \frac{150}{130} = 1{,}1538.$

a) $\alpha = 1\%$.
   i) $\mu \neq 1$?
   $H_0: \mu = 1\ vs\ H_1: \mu \neq 1$ (teste bilateral).
   Estatística de teste:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1)$$

   $z_{obs} = \frac{1{,}1538 - 1}{\frac{1}{\sqrt{130}}} = 1{,}7536.$
   Pela tabela $z_{1-\frac{\alpha}{2}} = z_{0{,}995} = 2{,}576$.
   Logo, $R.A.:]-2{,}576;\ 2{,}576[$ e $R.R.:]-\infty;\ -2{,}576] \cup [2{,}576;\ +\infty[$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 1%, não existe evidência estatística de que o tempo médio de conversação seja diferente de 1 minuto.

ii) $\mu > 1$?
$H_0: \mu \leq 1 \; vs \; H_1: \mu > 1$ (teste unilateral direito).
Estatística de teste:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1)$$

Como a estatística de teste é a mesma da alínea anterior, $z_{obs} = 1{,}7536$.
Pela tabela $z_{1-\alpha} = z_{0{,}99} = 2{,}326$. Logo, $R.A.:]-\infty;\ 2{,}326[$ e $R.R.: [2{,}326; +\infty[$.
Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 1%, não existe evidência estatística de que o tempo médio de conversação seja superior a 1 minuto.

iii) $\mu < 1$?
$H_0: \mu \geq 1 \; vs \; H_1: \mu < 1$ (teste unilateral esquerdo).
Estatística de teste:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1).$$

Como a estatística de teste é a mesma da alínea anterior, $z_{obs} = 1{,}7536$.
Pela tabela $z_{1-\alpha} = z_{0{,}99} = 2{,}326$. Logo, $R.A.:]-2{,}326; +\infty[$ e $R.R.:]-\infty;\ -2{,}326]$.
Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 1%, não existe evidência estatística de que o tempo médio de conversação seja inferior a 1 minuto.

b) Em todos os testes a decisão foi não rejeitar H_0. Logo, o erro que pode estar a ser cometido é o erro de Tipo II ou de 2ª espécie, i. e., não se rejeita H_0 mas na realidade H_0 é falsa.

c) $\text{valor } p = P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira}) = P(Z_{obs} \in R.R. | \mu = \mu_0)$.

i) $\text{valor } p = 2 \times P(Z \geq |z_{obs}|) = 2 \times P(Z \geq 1{,}7536) = 2 \times \big(1 - \Phi(1{,}7536)\big)$
$= 2 \times (1 - 0{,}9603) = 0{,}0794.$
Logo, a hipótese $H_0: \mu = 1$ é rejeitada para níveis de significância superiores ou iguais a 7,91%. Isto significa que para níveis de significância de 5% e 1% (referidos por serem os mais usuais), não existe evidência de que o tempo médio das conversações telefónicas no referido período seja diferente de 1 minuto. Para um nível de significância de 10 %, já se considera haver evidência de que este tempo médio é diferente de 1 minuto.

ii) $\text{valor } p = P(Z \geq z_{obs}) = P(Z \geq 1{,}7536) = 1 - \Phi(1{,}7536) = 1 - 0{,}9603 = 0{,}0397.$
Alternativa, com base no $valor\text{-}p$ bilateral calculado na alínea i): substituindo em H_1 μ por $\overline{x}$, $\overline{x} = 1{,}1538 > 1$ dá uma proposição verdadeira. Logo,

$$\text{valor } p_{uni} = \frac{0{,}0794}{2} = 0{,}0397.$$

A hipótese $H_0: \mu \leq 1$ é rejeitada para níveis de significância superiores ou iguais a 3,97%, i. e., para 1% de significância não existe evidência que o tempo médio de conversação seja superior a 1 minuto, mas para qualquer nível de significância superior a 3,97% essa conclusão já não se mantém.

iii) $\text{valor } p = P(Z \leq z_{obs}) = P(Z \leq 1{,}7536) = \Phi(1{,}7536) = 0{,}9603.$
Alternativa, com base no $valor\text{-}p$ bilateral calculado na alínea i): substituindo em H_1 μ por $\overline{x}$, $\overline{x} = 1{,}1538 < 1$ dá uma proposição falsa. Logo,

$$\text{valor } p_{uni} = 1 - \frac{0{,}0794}{2} = 1 - 0{,}0397 = 0{,}9603.$$

A hipótese H_0: $\mu \geq 1$ é rejeitada para níveis de significância superiores ou iguais a 96,03%. Como nunca se trabalha como níveis de significância desta ordem de grandeza, não existe evidência de que o tempo de conversação seja inferior a 1 minuto.

d) $\beta(\mu = \mu_1) = P(\text{Não rejeitar } H_0 | H_0 \text{ é falsa}) = P(\text{Não rejeitar } H_0 | \mu = \mu_1)$.

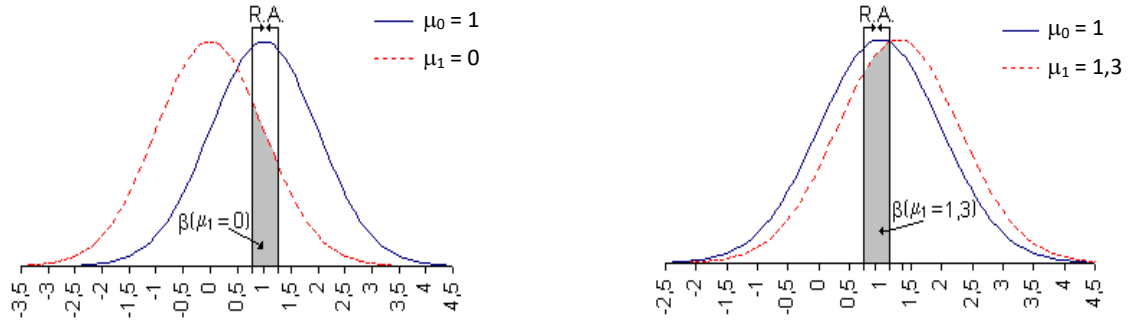
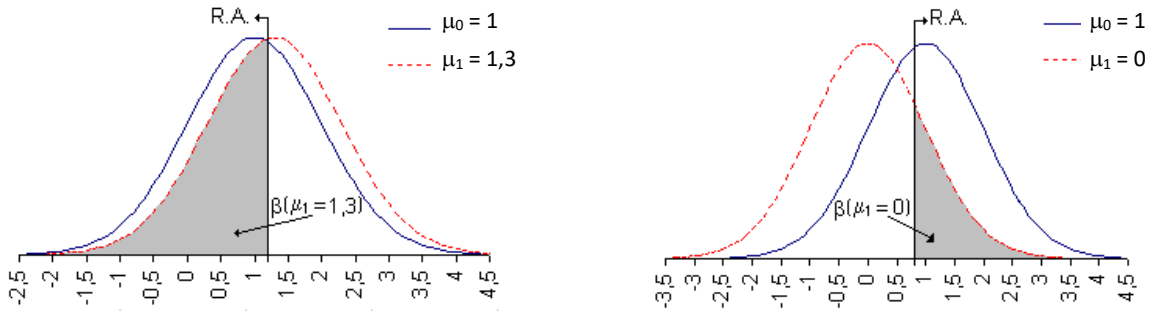


Figura 8.5: *Teste bilateral.*

Conforme se pode constatar pela Figura 8.5, se a verdadeira média for 1,3 o risco de não ser detectada é maior do que no caso em que a verdadeira média é 0, ou seja quanto mais afastado estiver μ_1 de μ_0 menor o valor de β.

No caso do teste unilateral direito apenas faz sentido $\beta(\mu_1 = 1{,}3)$, uma vez que se $\mu_1 = 0$ então toma-se uma decisão acertada (Figura 8.6 a).

No caso do teste unilateral esquerdo apenas faz sentido $\beta(\mu_1 = 0)$, uma vez que se $\mu_1 = 1{,}3$ então toma-se uma decisão acertada (Figura 8.6 b).



a) Teste unilateral direito. *b) Teste unilateral esquerdo.*

Figura 8.6: *Teste unilateral direito.*

e) $\pi(\mu_1 = 1{,}3) = P(\text{Rejeitar } H_0 | H_0 \text{ é falsa}) = P(\text{Rejeitar } H_0 | \mu = 1{,}3) = 1 - \beta(\mu_1 = 1{,}3)$.

i) No caso do *teste bilateral* rejeita-se H_0 quando $|z_{obs}| \geq z_{1-\frac{\alpha}{2}}$. Para $\alpha = 1\%$,

$$|Z_{obs}| \geq z_{0{,}995} \Leftrightarrow \left| \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \right| \geq z_{1-\frac{\alpha}{2}} \Leftrightarrow \begin{cases} \overline{X} \geq \mu_0 + z_{0{,}995} \frac{\sigma}{\sqrt{n}} \\ \text{ou} \\ \overline{X} \leq \mu_0 - z_{0{,}995} \frac{\sigma}{\sqrt{n}} \end{cases} \Leftrightarrow \begin{cases} \overline{X} \geq 1 + 2{,}576 \frac{1}{\sqrt{130}} \\ \text{ou} \\ \overline{X} \leq 1 - 2{,}576 \frac{1}{\sqrt{130}} \end{cases}$$

$$\Leftrightarrow \begin{cases} \overline{X} \geq 1{,}2259 \\ \text{ou} \\ \overline{X} \leq 0{,}7741 \end{cases}.$$

Logo,

$$\pi(\mu_1 = 1{,}3) = P\left(\left(\overline{X} \geq 1{,}2259 \cup \overline{X} \leq 0{,}7741\right)|\mu = 1{,}3\right)$$
$$= P\left(\overline{X} \geq 1{,}2259|\mu = 1{,}3\right) + P\left(\overline{X} \leq 0{,}7741|\mu = 1{,}3\right)$$
$$= P\left(\frac{\overline{X} - 1{,}3}{\frac{1}{\sqrt{130}}} \geq \frac{1{,}2259 - 1{,}3}{\frac{1}{\sqrt{130}}}\right) + P\left(\frac{\overline{X} - 1{,}3}{\frac{1}{\sqrt{130}}} \leq \frac{0{,}7741 - 1{,}3}{\frac{1}{\sqrt{130}}}\right)$$
$$= P(Z \geq -0{,}845) + P(Z \leq -5{,}996) = \left(1 - \Phi(-0{,}845)\right) + \Phi(-5{,}996)$$
$$\approx 1 - \left(1 - \Phi(0{,}845)\right) + 0 = 0{,}8009.$$

Portanto, a potência do teste é de 0,8009. Note-se que esta é a probabilidade de decidir algo correcto isto é, a probabilidade de rejeitar H_0 quando esta era falsa. Neste caso específico representa a probabilidade de decidir que o tempo médio de conversação foi diferente de 1 minuto, quando efectivamente o verdadeiro tempo médio de conversação foi de 1,3 minutos ($\mu_1 = 1{,}3$). Esta probabilidade considera-se elevada.

ii) No caso do teste unilateral direito rejeita-se H_0 quando $z_{obs} \geq z_{1-\alpha}$. Para $\alpha = 1\%$,

$$Z_{obs} \geq z_{0{,}99} \Leftrightarrow \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq z_{0{,}99} \Leftrightarrow \overline{X} \geq \mu_0 + z_{0{,}99}\frac{\sigma}{\sqrt{n}} \Leftrightarrow \overline{X} \geq 1 + 2{,}326\frac{1}{\sqrt{130}}$$
$$\Leftrightarrow \overline{X} \geq 1{,}204.$$

Logo,

$$\pi(\mu_1 = 1{,}3) = P\left(\overline{X} \geq 1{,}204|\mu = 1{,}3\right) = 1 - \Phi\left(\frac{1{,}204 - 1{,}3}{\frac{1}{\sqrt{130}}}\right) = 1 - \Phi(-1{,}095)$$
$$= 1 - \left(1 - \Phi(1{,}095)\right) = \Phi(1{,}095) = 0{,}8632.$$

Portanto, a potência do teste é de 0,8632, ou seja, se o verdadeiro tempo médio de conversação for de 1,3 minutos. Com este teste toma-se a decisão certa em 86,32% dos casos.

iii) No caso do teste unilateral esquerdo rejeita-se H_0 quando $z_{obs} \leq -z_{1-\alpha}$. Para $\alpha = 1\%$,

$$Z_{obs} \leq -z_{0{,}99} \Leftrightarrow \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq -z_{0{,}99} \Leftrightarrow \overline{X} \leq \mu_0 - z_{0{,}99}\frac{\sigma}{\sqrt{n}} \Leftrightarrow \overline{X} \leq 1 - 2{,}326\frac{1}{\sqrt{130}}$$
$$\Leftrightarrow \overline{X} \leq 0{,}796.$$

Logo,

$$\pi(\mu_1 = 1{,}3) = P\left(\overline{X} \leq 0{,}796|\mu = 1{,}3\right) = \Phi\left(\frac{0{,}796 - 1{,}3}{\frac{1}{\sqrt{130}}}\right) = \Phi(-5{,}746) \approx 0.$$

Portanto, a potência do teste é de aproximadamente 0, i. e., se o verdadeiro tempo médio de conversação for 1,3 minutos. Com este teste não é de todo expectável que se tome a decisão certa, uma vez que os dados indiciam indicações contrárias.

8.13.1.2 Variância desconhecida

Os dados seguintes representam os ganhos em peso, em quilogramas, nos primeiros 6 meses de vida de um grupo de crianças do sexo masculino escolhidas ao acaso.

4,1 4,5 3,6 2,8 3,6 3,2 4,1

Admita que se pode considerar que os ganhos em peso seguem uma distribuição Normal.

a) Poderá afirmar-se $(\alpha = 0{,}05)$ que o ganho médio em peso das crianças do sexo masculino é significativamente:
   i. Diferente de 3,1 kg?
   ii. Superior a 3,1 kg?
   iii. Inferior a 3,1 kg?

b) A partir de que nível de significância rejeita as hipóteses testadas anteriormente?

c) Para a alínea a) iii), calcule a potência de teste para $\mu_1 = 3$.

Resolução:

Seja X a v. a. que representa o ganho em peso, em kg, nos primeiros 6 meses de vida das crianças do sexo masculino, com $X \sim N(\mu;\ \sigma)$.

$n = 7;\ \ \overline{x} = 3{,}7;\ \ s^2 = 0{,}34$ e $s = 0{,}5831$.

a) $\alpha = 0{,}05$

i) $\mu \neq 3{,}1$?

$H_0: \mu = 3{,}1\ vs.\ H_1: \mu \neq 3{,}1$ (teste bilateral).

Estatística de teste:

$$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1=6}.$$

$$t_{obs} = \frac{3{,}7 - 3{,}1}{\frac{0{,}5831}{\sqrt{7}}} = 2{,}7225.$$

Pela tabela $t_{n-1;1-\frac{\alpha}{2}} = t_{6;\,0{,}975} = 2{,}447$.

Logo, $R.A.:]-2{,}447;\ 2{,}447[$ e $R.R.:]-\infty;\ -2{,}447] \cup [2{,}447;\ +\infty[$.

Como $t_{obs} \in R.R.$ rejeitar H_0. Portanto, ao nível de significância de 5%, existe evidência estatística de que o ganho médio em peso das crianças do sexo masculino é significativamente diferente de 3,1 kg.

(SPSS) Analyse → Compare Means → One-sample T Test…
(Test Variable: Peso; Test Value: 3,1)

T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Peso	7	3,700	,5831	,2204

One-Sample Test

	Test Value = 3.1					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Peso	2,722	6	,035	,6000	,061	1,139

ii) $\mu > 3{,}1$?

$H_0: \mu \leq 3{,}1\ vs\ H_1: \mu > 3{,}1$ (teste unilateral direito).

Estatística de teste:

$$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1=6}.$$

$$t_{obs} = \frac{3{,}7 - 3{,}1}{\frac{0{,}5831}{\sqrt{7}}} = 2{,}7225.$$

Pela tabela $t_{n-1;\,1-\alpha} = t_{6;\,0{,}95} = 2{,}326$. Logo, $R.A.:]-\infty;\ 1{,}943[$ e $R.R.: [1{,}943; +\infty[$.

Como $t_{obs} \in R.R.$ rejeitar H_0. Portanto, ao nível de significância de 5%, existe evidência estatística de que o ganho médio em peso das crianças do sexo masculino é significativamente superior a 3,1 kg.

iii) $\mu < 3{,}1$?

H_0: $\mu \geq 3{,}1$ vs H_1: $\mu < 3{,}1$ (teste unilateral esquerdo).

Estatística de teste:

$$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \sim t_{n-1=6}.$$

$$t_{obs} = \frac{3{,}7 - 3{,}1}{\frac{0{,}5831}{\sqrt{7}}} = 2{,}7225.$$

Pela tabela $t_{n-1;\,1-\alpha} = t_{6;\,0{,}95} = 2{,}326$.
Logo, $R.A.:]-2{,}326; +\infty[$ e $R.R.:]-\infty;\ -2{,}326]$.

Como $t_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que o ganho médio em peso das crianças do sexo masculino seja significativamente inferior a 3,1 kg.

b) valor $p = P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira}) = P(Z_{obs} \in R.R. | \mu = \mu_0)$.

i) $\text{valor } p = 2 \times P(T \geq |t_{obs}|) = 2 \times P(T \geq 2{,}7225) = 2 \times \big(1 - P(T < 2{,}7225)\big)$
$= 2 \times (1 - 0{,}9827) = 0{,}0345.$

A hipótese H_0: $\mu = 3{,}1$ é rejeitada para níveis de significância superiores ou iguais a 3,45%. Assim, como exemplos, para 1% de significância não existe evidência que o ganho seja diferente de 3,1 Kg, mas para 5 % já existe essa evidência.

ii) $\text{valor } p = P(T \geq t_{obs}) = P(T \geq 2{,}7225) = 1 - P(T < 2{,}7225) = 1 - 0{,}9827 = 0{,}0173.$
Alternativa, com base no valor p bilateral calculado na alínea i): substituindo em H_1 μ por $\overline{x}$, $\overline{x} = 3{,}7 > 3{,}1$ dá uma proposição verdadeira. Logo,

$$\text{valor } p_{uni} = \frac{0{,}0345}{2} = 0{,}0173.$$

A hipótese H_0: $\mu \leq 3{,}1$ é rejeitada para níveis de significância superiores ou iguais a 1,73%. Assim, como exemplos, para um nível de significância de 5% existe evidência de que o ganho médio foi superior a 3,1 Kg, mas para 1% não se pode manter a mesma decisão.

iii) $valor\text{-}p = P(T \leq t_{obs}) = P(T \leq 2{,}7225) = 0{,}9827.$
Alternativa, com base no $valor\text{-}p$ bilateral calculado na alínea i): substituindo em H_1 μ por $\overline{x}$, $\overline{x} = 3{,}7 > 3{,}1$ dá uma proposição falsa. Logo,

$$\text{valor } p_{uni} = 1 - \frac{0{,}0345}{2} = 1 - 0{,}0173 = 0{,}9827.$$

A hipótese H_0: $\mu \geq 3{,}1$ é rejeitada para níveis de significância superiores ou iguais a 98,27%. Para qualquer nível de significância sensato/usual (≤ 10%) não existe evidência de que o ganho médio foi inferior a 3,1 Kg.

c) $\pi(\mu_1 = 3) = P(\text{Rejeitar } H_0 | H_0 \text{ é falsa}) = P(\text{Rejeitar } H_0 | \mu = 1{,}3) = 1 - \beta(\mu_1 = 3)$.
No caso do *teste unilateral esquerdo* rejeita-se H_0 quando $t_{obs} \leq -t_{n-1;\, 1-\alpha}$. Para $\alpha = 5\%$,

$$T_{obs} \leq -t_{6;\, 0{,}95} \Leftrightarrow \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}} \leq -t_{6;\, 0{,}95} \Leftrightarrow \overline{X} \leq \mu_0 - t_{6;\, 0{,}95}\frac{S}{\sqrt{n}}$$

No caso desta amostra, $s = 0{,}5831$. Assim,

$$\overline{X} \leq 3{,}1 - 1{,}9432\frac{0{,}5831}{\sqrt{7}} \Leftrightarrow \overline{X} \leq 2{,}6718$$

Logo,

$$\pi(\mu_1 = 3) = P\big(\overline{X} \leq 2{,}6718 | \mu = 3\big) = P\left(T \leq \frac{2{,}672 - 3}{\frac{0{,}5831}{\sqrt{7}}}\right) = P(T \leq -1{,}489) = 1 - P(T < 1{,}489)$$
$$= 1 - 0{,}9065 = 0{,}0935.$$

Portanto, a potência do teste é de 0,0935, ou seja, é pouco provável que o teste detecte uma verdadeira diferença, isto é que se decida que o peso é inferior a 3,1 kg, mesmo que essa seja a realidade ($\mu_1 = 3$), pois os dados recolhidos indiciam o contrário.

8.13.2 Teste de hipótese para a diferença de médias

8.13.2.1 Quando as variâncias são conhecidas

Nos primeiros 6 meses de vida dois grupos aleatórios de crianças seguiram esquemas de alimentação diferentes: o grupo 1 seguiu o esquema A e o grupo 2 seguiu o esquema B. No quadro seguinte apresentam-se os ganhos em peso, em kg, dessas crianças.

Grupo 1	2,7	3,2	3,6	4,1	2,7	3,2	4,5	3,6	2,7
Grupo 2	4,1	4,5	3,6	2,7	3,6	3,2	4,1		

Sabe-se que as crianças dos dois grupos tinham, ao nascer, aproximadamente pesos iguais.
Admita que as distribuições dos pesos seguem a distribuição Normal com variâncias 0,36 e 0,32, respectivamente.

a) Ao nível de significância de 1%, poderá afirmar que o ganho médio em peso das crianças alimentadas segundo o esquema A é:
   i. Igual ao das crianças alimentadas segundo o esquema B?
   ii. Superior ao das crianças alimentadas segundo o esquema B?
   iii. Inferior ao das crianças alimentadas segundo o esquema B?

b) A partir de que nível de significância rejeita cada uma das hipóteses anteriores?

Resolução:

Sejam:

- X_1 a v.a. que representa o ganho em peso, em kg, das crianças alimentadas segundo o esquema A,
- X_2 a v.a. que representa o ganho em peso, em kg, das crianças alimentadas segundo o esquema B,

com $X_1 \sim N(\mu_1 = ?\,; \sigma_1 = \sqrt{0{,}36})$ e $X_2 \sim N(\mu_2 = ?; \sigma_2 = \sqrt{0{,}32})$.

$n_1 = 9, \quad \overline{x}_1 = 3{,}3367;$
$n_2 = 7, \quad \overline{x}_2 = 3{,}6857.$

a) $\alpha = 1\%$.

i) $\mu_1 = \mu_2$?

$H_0: \mu_1 = \mu_2 \; vs \; H_1: \mu_1 \neq \mu_2$
$\Leftrightarrow H_0: \mu_1 - \mu_2 = 0 \; vs \; H_1: \mu_1 - \mu_2 \neq 0$ (teste bilateral).
Estatística de teste:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0; 1).$$

$$z_{obs} = \frac{(3{,}3667 - 3{,}6857) - 0}{\sqrt{\frac{0{,}36}{9} + \frac{0{,}32}{7}}} = -1{,}0896.$$

Pela tabela $z_{1-\frac{\alpha}{2}} = z_{0{,}995} = 2{,}576$.

Logo, $R.A.:]-2{,}576;\ 2{,}576[$ e $R.R.:]-\infty;\ -2{,}576] \cup [2{,}576; +\infty[$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 1%, não existe evidência estatística de que o ganho médio em peso das crianças alimentadas segundo o esquema A seja diferente do das crianças alimentadas segundo o esquema B.

ii) $\mu_1 > \mu_2$?

$H_0: \mu_1 \leq \mu_2 \; vs. \; H_1: \mu_1 > \mu_2$
$\Leftrightarrow H_0: \mu_1 - \mu_2 \leq 0 \; vs. \; H_1: \mu_1 - \mu_2 > 0$ (teste unilateral direito).

Estatística de teste:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0; 1).$$

Como a estatística de teste é a mesma da alínea anterior, $z_{obs} = -1{,}0896$.

Pela tabela $z_{1-\alpha} = z_{0{,}99} = 2{,}326$. Logo, $R.A.:]-\infty;\ 2{,}326[$ e $R.R.: [2{,}326; +\infty[$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 1%, não existe evidência estatística de que o ganho médio em peso das crianças alimentadas segundo o esquema A seja superior ao das crianças alimentadas segundo o esquema B.

iii) $\mu_1 < \mu_2$?

$H_0: \mu_1 \geq \mu_2 \; vs \; H_1: \mu_1 < \mu_2$
$\Leftrightarrow H_0: \mu_1 - \mu_2 \geq 0 \; vs \; H_1: \mu_1 - \mu_2 < 0$ (teste unilateral esquerdo).

Estatística de teste: $Z = \dfrac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0;1)$.

Como a estatística de teste é a mesma da alínea anterior, $z_{obs} = -1{,}0896$.

Pela tabela $z_{1-\alpha} = z_{0{,}99} = 2{,}326$. Logo, $R.A.:]-2{,}326; +\infty[$ e $R.R.:]-\infty;\ -2{,}326]$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 1%, não existe evidência estatística de que o ganho médio em peso das crianças alimentadas segundo o esquema A seja inferior ao das crianças alimentadas segundo o esquema B.

b) $\text{valor } p = P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira}) = P(Z_{obs} \in R.R. | \mu = \mu_0)$.

i) $\text{valor } p = 2 \times P(Z \geq |z_{obs}|) = 2 \times P(Z \geq 1{,}0896) = 2 \times \left(1 - \Phi(1{,}0896)\right)$
$= 2 \times (1 - 0{,}8621) = 0{,}2758.$

A hipótese H_0: $\mu_1 = \mu_2$ é rejeitada para níveis de significância superiores ou iguais a 27,58%, logo não é rejeitada para qualquer nível de significância usual em investigação: não existe evidência estatística de que o ganho médio de peso das crianças alimentadas segundo o esquema A seja diferente do das crianças alimentadas segundo o esquema B.

ii) $\text{valor } p = P(Z \geq z_{obs}) = P(Z \geq -1{,}0896) = 1 - \Phi(-1{,}0896)$
$= 1 - \left(1 - \Phi(1{,}0896)\right) = \Phi(1{,}0896) = 0{,}8621.$

Alternativa, com base no valor p bilateral calculado na alínea i): substituindo em H_1 μ_1 e μ_2 por $\overline{x}_1$ e $\overline{x}_2$, respectivamente, $\overline{x}_1 - \overline{x}_2 = 3{,}3367 - 3{,}6857 > 0$ dá uma proposição falsa. Logo,

$$\text{valor } p_{uni} = 1 - \frac{0{,}2758}{2} = 0{,}8621.$$

A hipótese H_0: $\mu_1 \leq \mu_2$ é rejeitada para níveis de significância superiores ou iguais a 86,21%. Assim, não existe evidência estatística de que o ganho médio de peso das crianças alimentadas segundo o esquema A seja superior ao das crianças alimentadas segundo o esquema B.

iii) $valor\ p = P(Z \leq z_{obs}) = P(Z \leq -1{,}0896) = \Phi(-1{,}0896) = 1 - \Phi(1{,}0896)$
$= 1 - 0{,}8621 = 0{,}1379.$

Alternativa, com base no valor p bilateral calculado na alínea i): substituindo em H_1 μ_1 e μ_2 por $\overline{x}_1$ e $\overline{x}_2$, respectivamente, $\overline{x}_1 - \overline{x}_2 = 3{,}3367 - 3{,}6857 > 0$ dá uma proposição verdadeira. Logo,

$$\text{valor } p_{uni} = \frac{0{,}2758}{2} = 0{,}1379.$$

A hipótese H_0: $\mu_1 \geq \mu_2$ é rejeitada para níveis de significância superiores ou iguais a 13,79%: não existe evidência estatística de que o ganho médio de peso das crianças alimentadas segundo o esquema A seja inferior ao das crianças alimentadas segundo o esquema B.

8.13.2.2 Quando as variâncias são desconhecidas e iguais

Um determinado método de análise permite determinar o conteúdo de enxofre no petróleo bruto. Os ensaios efectuados em 10 e 8 amostras aleatórias de 1 kg de petróleo bruto, provenientes de furos pertencentes respectivamente aos campos A e B, revelaram os seguintes resultados (em gramas):

Campo A:	111	114	105	112	107	109	112	110	110	106
Campo B:	109	103	101	105	106	108	110	104		

a) Ao nível de significância de 10%, poderá afirmar que, em média, quantidade de enxofre por quilograma de petróleo do campo A é:
   i. Igual à do campo B?
   ii. Superior à do campo B?
   iii. Inferior à do campo B?

b) Calcule o $valor\ p$ associado a cada um dos testes anteriores.

Resolução:

Sejam:
- X_1 a v.a. que representa a quantidade de enxofre por quilograma de petróleo do campo A,
- X_2 a v.a. que representa a quantidade de enxofre por quilograma de petróleo do campo B.

Nada é referido sobre a distribuição de X_1 e X_2.

$n_1 = 10, \quad \overline{x}_1 = 109{,}6 \quad \text{e} \quad s_1 = 2{,}875,$
$n_2 = 8, \quad \overline{x}_2 = 105{,}75 \quad \text{e} \quad s_2 = 3{,}105.$

a) $\alpha = 10\%$.

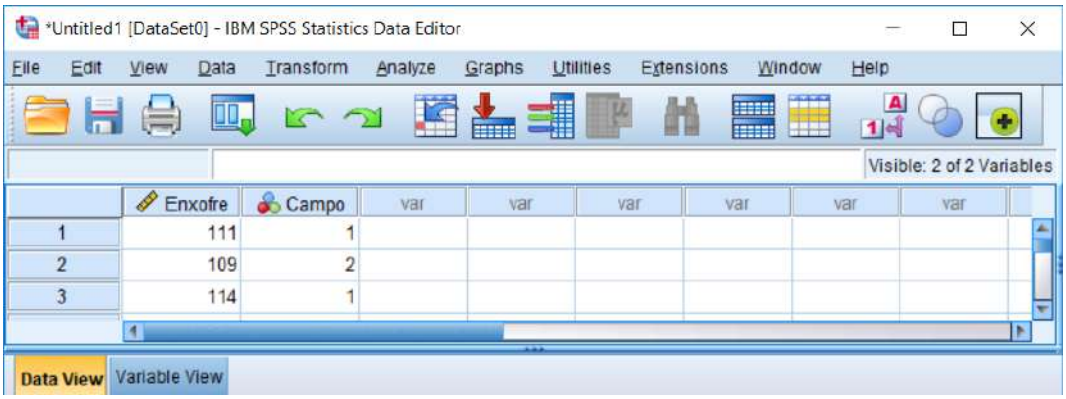
i) $\mu_1 = \mu_2$?

$H_0: \mu_1 = \mu_2\ vs\ H_1: \mu_1 \neq \mu_2$
$\Leftrightarrow H_0: \mu_1 - \mu_2 = 0\ vs\ H_1: \mu_1 - \mu_2 \neq 0$ (teste bilateral).

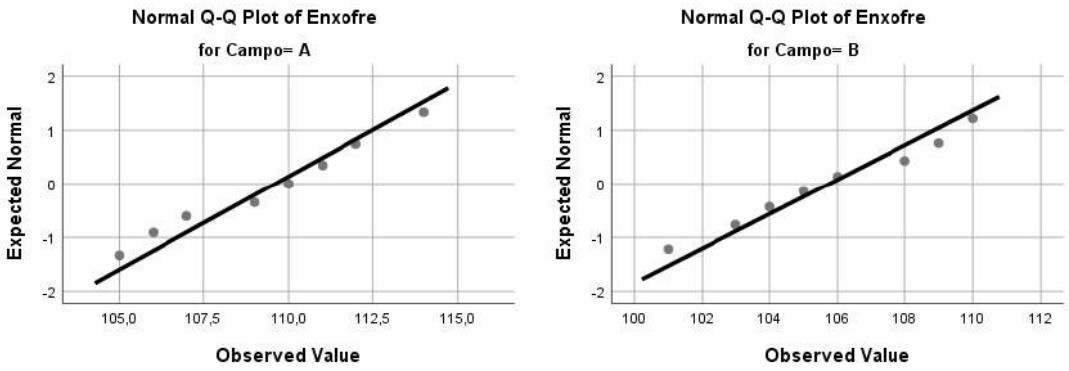
Para saber decidir qual a estatística de teste a utilizar, é preciso validar os pressupostos subjacentes:

- Normalidade:

  (SPSS)



  (SPSS) Analyze → Descriptive Statistics → Explore…
  (Dependent list: Enxofre; Factor List: Campo; Display: ⊙ Plots;
  Plots… → ☑ Normality plots with tests)



Em ambos os gráficos quantil-quantil os pontos posicionam-se sobre a recta, pelo que se pode considerar que ambos os conjuntos de dados são provenientes de populações com distribuição Normal.

- Igualdade das variâncias:

  O I. C. a 90% para $\frac{\sigma_1^2}{\sigma_2^2}$ é dado por:

$$\left]\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1;n_2-1\,;\,1-\frac{\alpha}{2}}};\frac{S_1^2}{S_2^2}F_{n_2-1;n_1-1\,;\,1-\frac{\alpha}{2}}\right[.$$

  Substituindo os valores, sendo $F_{9;\,7;0,95}=3{,}68$ e $F_{7;\,9;0,95}=3{,}29$, obtém-se:

$$\left]\frac{2{,}875^2}{3{,}105^2}\times\frac{1}{3{,}68};\frac{2{,}875^2}{3{,}105^2}\times 3{,}29\right[=]0{,}2332;\,2{,}8230[.$$

  Como 1 pertence ao intervalo obtido, ao nível de significância de 10% não há evidências de que σ_1^2 seja diferente de σ_2^2. Portanto, pode-se considerar que $\sigma_1^2=\sigma_2^2$.

Deste modo, a estatística de teste a usar é:

$$T=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)_0}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t_{n_1+n_2-2=10+8-2=16}.$$

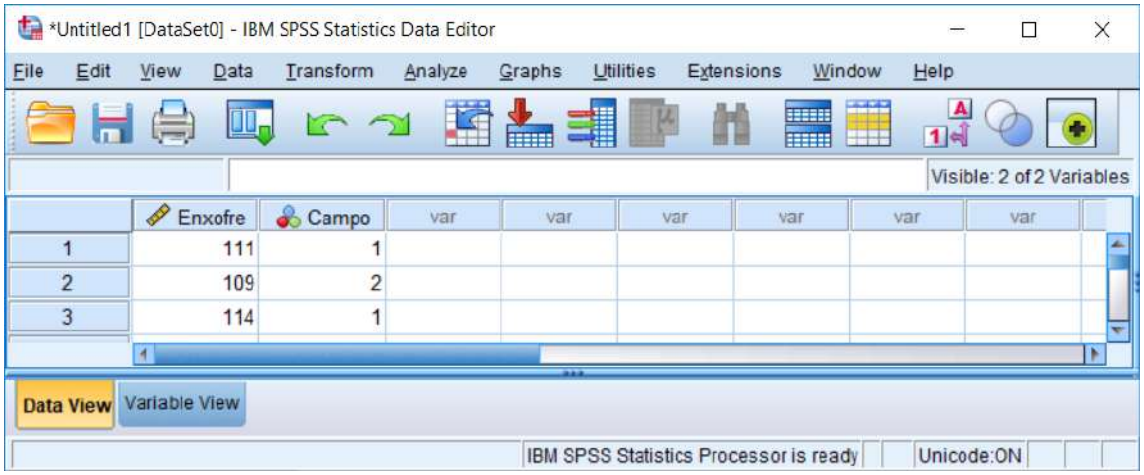
$$t_{obs}=\frac{(109{,}6-105{,}75)-0}{\sqrt{\frac{(10-1)2{,}875+(8-1)3{,}105}{10+8-2}}\sqrt{\frac{1}{10}+\frac{1}{8}}}=2{,}7255.$$

Pela tabela $t_{n_1+n_2-2;\,1-\frac{\alpha}{2}}=t_{16\,;\,0,95}=1{,}746$.

Logo, $R.A.:]-1{,}746;\ 1{,}746[$ e $R.R.:]-\infty;\ -1{,}746]\cup[1{,}746;+\infty[$.

Como $t_{obs}\in R.R.$ rejeitar H_0. Portanto, ao nível de significância de 10%, existe evidência estatística de que, em média, a quantidade de enxofre por quilograma de petróleo do campo A é diferente da do campo B.

(SPSS)



(SPSS) Analyse → Compare Means → Independent-Samples T Test…
(Test Variable: Enxofre; Grouping Variable: Campo;
Define Groups → Use specified values → Group 1: 1; Group 2: 2)

Group Statistics

	Campo	N	Mean	Std. Deviation	Std. Error Mean
Enxofre por kg de petróleo	A	10	109,60	2,875	,909
	B	8	105,75	3,105	1,098

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Enxofre por kg de petróleo	Equal variances assumed	,086	,773	2,725	16	,015	3,850	1,413	,855	6,845
	Equal variances not assumed			2,701	14,6	,017	3,850	1,425	,804	6,896

Pela linha superior do quadro (onde se assume a igualdade das variâncias) $t_{obs} = 2{,}725$ (coluna *t*), $v = 16$ (coluna *df*) e valor $p = 0{,}015$ (coluna *Sig.* (*2-tailed*)).

Observação: O SPSS aplica automaticamente o teste de Levene[†] para a igualdade de variâncias. Pelo valor p (valor $p = 0{,}773$), verifica-se que não há evidência de que as variâncias populacionais sejam diferentes.

ii) $\mu_1 > \mu_2$?

$H_0: \mu_1 \leq \mu_2 \; vs \; H_1: \mu_1 > \mu_2$
$\Leftrightarrow H_0: \mu_1 - \mu_2 \leq 0 \; vs \; H_1: \mu_1 - \mu_2 > 0$ (teste unilateral direito).

Estatística de teste:

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2=10+8-2=16}.$$

Como a estatística de teste é a mesma da alínea anterior, $t_{obs} = 2{,}7255$.

Pela tabela $t_{n_1+n_2-2;\,1-\alpha} = t_{16;\,0{,}90} = 1{,}337$.
Logo, $R.A.:]-\infty;\; 1{,}337[$ e $R.R.: [1{,}337; +\infty[$.

Como $t_{obs} \in R.R.$ rejeitar H_0. Portanto, ao nível de significância de 10%, existe evidência estatística de que, em média, a quantidade de enxofre por quilograma de petróleo do campo A é superior à do campo B.

iii) $\mu_1 < \mu_2$?

$H_0: \mu_1 \geq \mu_2 \; vs \; H_1: \mu_1 < \mu_2$
$\Leftrightarrow H_0: \mu_1 - \mu_2 \geq 0 \; vs \; H_1: \mu_1 - \mu_2 < 0$ (teste unilateral esquerdo).

[†] O teste de Levene é apresentado na secção 9.4.2.

Estatística de teste:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2=10+8-2=16}.$$

Como a estatística de teste é a mesma da alínea anterior, $t_{obs} = 2{,}7255$.

Pela tabela $t_{n_1+n_2-2;\,1-\alpha} = t_{16;\,0{,}90} = 1{,}337$.
Logo, $R.A.:]-1{,}337; +\infty[$ e $R.R.:]-\infty; -1{,}337]$.

Como $t_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 10%, não existe evidência estatística de que, em média, a quantidade de enxofre por quilograma de petróleo do campo A seja inferior à do campo B.

b) valor $p = P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira}) = P(Z_{obs} \in R.R. | \mu = \mu_0)$.

i) valor $p = 2 \times P(T \geq |t_{obs}|) = 2 \times P(T \geq 2{,}7255) = 2 \times \big(1 - P(T < 2{,}7255)\big)$
$= 2 \times (1 - 0{,}9925) = 0{,}015.$

A hipótese $H_0: \mu_1 = \mu_2$ é rejeitada para níveis de significância superiores ou iguais a 1,5%. Logo, para um nível de significância de 5% existe evidência de que o teor médio de enxofre no campo A é diferente do campo B, mas para 1% essa afirmação já não pode ser sustentada. Repare-se que o $valor\ p$ calculado é o valor indicado no quadro de resultados do SPSS anteriormente apresentado.

ii) valor $p = P(T \geq t_{obs}) = P(T \geq 2{,}7255) = 1 - P(T < 2{,}7255) = 1 - 0{,}9925 = 0{,}0075$.
Alternativa, com base no $valor\text{-}p$ bilateral calculado na alínea i): substituindo em H_1 μ_1 e μ_2 por $\overline{x}_1$ e $\overline{x}_2$, respectivamente, $\overline{x}_1 - \overline{x}_2 = 109{,}6 - 105{,}75 > 0$ dá uma proposição verdadeira. Logo,

$$valor\text{-}p_{uni} = \frac{0{,}015}{2} = 0{,}0075.$$

A hipótese $H_0: \mu_1 \leq \mu_2$ é rejeitada para níveis de significância superiores ou iguais a 0,75%. Assim, para qualquer nível de significância sensato/usual ($\leq 10\%$) existe evidência de que o teor médio de enxofre no campo A é superior ao do campo B.

iii) valor $p = P(T \leq t_{obs}) = P(T \leq 2{,}7255) = 0{,}9925$.
Alternativa, com base no $valor\text{-}p$ bilateral calculado na alínea i): substituindo em H_1 μ_1 e μ_2 por $\overline{x}_1$ e $\overline{x}_2$, respectivamente, $\overline{x}_1 - \overline{x}_2 = 109{,}6 - 105{,}75 > 0$ dá uma proposição falsa. Logo,

$$\text{valor } p_{uni} = 1 - \frac{0{,}015}{2} = 0{,}9925.$$

A hipótese $H_0: \mu_1 \geq \mu_2$ é rejeitada para níveis de significância superiores ou iguais a 99,25. Assim, existe não existe evidência de que o teor médio de enxofre no campo A seja inferior ao do campo B.

8.13.2.3 Quando as variâncias são desconhecidas e diferentes

Para um estudo sobre a caracterização da altura da população portuguesa, foi recolhida uma amostra de 1861 pessoas, com as seguintes características: (conjunto de dados semelhante ao disponibilizado no Capítulo 12, mas com uma amostra de maior dimensão):

Group Statistics

	Sexo	N	Mean	Std. Deviation
Altura	Masculino	853	168,46	7,617
	Feminino	1007	158,48	6,652

Supondo a Normalidade das distribuições e assumindo que as variâncias populacionais são desconhecidas e diferentes, verifique se se pode considerar que as alturas médias dos homens e das mulheres são iguais, com 95% de confiança.

a) Suspeita que em média a altura dos homens não é igual à das mulheres. Teste esta hipótese ao nível de significância de 5%.
b) Calcule o valor p associado ao teste da alínea anterior.
c) Teste a hipótese de a média da altura dos homens ser superior à das mulheres, ao nível de significância de 0,5%?
d) Determine o valor p associado ao teste anterior.
e) Ao nível de significância de 2,5%, pode-se afirmar que em média a altura dos homens é superior à das mulheres?
f) A partir de que nível de significância é rejeitada a hipótese do teste anterior?

Resolução:

Sejam:

- X_1 a v.a. que representa a altura dos homens,
- X_2 a v.a. que representa a altura das mulheres,

com $X_1 \sim N(\mu_1 = ?\,; \sigma_1 = ?)$ e $X_2 \sim N(\mu_2 = ?; \sigma_2 = ?)$, mas $\sigma_1^2 \neq \sigma_2^2$.

$n_1 = 853, \quad \overline{x}_1 = 168{,}46 \quad$ e $\quad s_1 = 7{,}617,$
$n_2 = 1007, \quad \overline{x}_2 = 158{,}48 \quad$ e $\quad s_2 = 6{,}652.$

a) $\alpha = 5\%$, $\mu_1 \neq \mu_2$?

$H_0: \mu_1 = \mu_2 \; vs H_1: \mu_1 \neq \mu_2$
$\Leftrightarrow H_0: \mu_1 - \mu_2 = 0 \; vs \; H_1: \mu_1 - \mu_2 \neq 0$ (teste bilateral).

Estatística de teste:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{\circ}{\sim} t_v, \text{ onde } v = \left[\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2} \right].$$

$$t_{obs} = \frac{(168{,}46 - 158{,}48) - 0}{\sqrt{\frac{7{,}617^2}{853} + \frac{6{,}652^2}{1007}}} = 29{,}816.$$

$$v = \left[\frac{\left(\frac{7{,}617^2}{853} + \frac{6{,}652^2}{1007}\right)^2}{\frac{1}{853 - 1}\left(\frac{7{,}617^2}{853}\right)^2 + \frac{1}{1007 - 1}\left(\frac{6{,}652^2}{1007}\right)^2} \right] = [1705{,}6] = 1705.$$

Pela tabela $t_{v;\, 1 - \frac{\alpha}{2}} = t_{1705\,;\, 0{,}975} = 1{,}96.$

Logo, $R.A.:]-1{,}96;\; 1{,}96[$ e $R.R.:]-\infty;\; -1{,}96] \cup [1{,}96; +\infty[$.

Como $t_{obs} \in R.R.$ rejeitar H_0. Portanto, ao nível de significância de 5%, existe evidência estatística de existe diferença significativa entre as médias das alturas dos homens e das mulheres.

(SPSS) Analyse → Compare Means → Independent-Samples T Test...
(Test Variable: altura; Grouping Variable: sexo;
Define Groups → Use specified values → Group 1: 1; Group 2: 2;
Options → Confidence Interval: 95)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Altura	Equal variances assumed	10,707	,001	30,150	1858	,000	9,976	,331	9,327	10,625
	Equal variances not assumed			29,816	1705	,000	9,976	,335	9,320	10,633

Interpretando a linha inferior do quadro (onde não se assume a igualdade das variâncias), $t_{obs} = 29816$ (coluna *t*), $v = 1705$ (coluna *df*) e $valor\ p < 0{,}001$ (coluna *Sig.* (*2-tailed*)).

Observação: Caso nada fosse referido sobre a possibilidade de se considerar a igualdade das variâncias, pelo valor p do teste de Levene† para a igualdade de variâncias (valor $p = 0{,}001 \leq \alpha$), verifica-se que há evidência de que as variâncias populacionais são diferentes.

b) $\text{valor } p = 2 \times P(T \geq |t_{obs}|) = 2 \times P(T \geq 29{,}816) = 2 \times \big(1 - P(T < 29{,}816)\big)$
$\approx 2 \times (1 - 1) = 0.$

Logo, aos níveis usuais de significância existe evidência de que a média das alturas dos homens difere das mulheres. Repare-se que o valor p calculado é o valor indicado no quadro de resultados do SPSS anteriormente apresentado.

c) $\alpha = 0{,}5\%$, $\mu_1 > \mu_2$?

$$H_0\colon \mu_1 \leq \mu_2 \ vs\ H_1\colon \mu_1 > \mu_2$$
$$\Leftrightarrow H_0\colon \mu_1 - \mu_2 \leq 0 \ vs\ H_1\colon \mu_1 - \mu_2 > 0 \text{ (teste unilateral direito).}$$

Estatística de teste:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{\circ}{\sim} t_v, \text{ onde } v = \left[\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2}\right].$$

Como a estatística de teste é a mesma da alínea anterior, $t_{obs} = 29{,}816$ e $v = 1705$.

Pela tabela $t_{v;\,1-\alpha} = t_{1705\,;\,0{,}995} = 2{,}576$.
Logo, $R.A.:]-\infty;\ 2{,}576[$ e $R.R.: [2{,}576;\ +\infty[$.

Como $t_{obs} \in R.R.$ rejeitar H_0. Portanto, ao nível de significância de 0,5%, existe evidência estatística de que, em média, os homens são mais altos do que as mulheres.

d) $\text{valor } p = P(T \geq t_{obs}) = P(T \geq 29{,}816\ ) = 1 - P(T < 29{,}816\ ) \approx 1 - 1 = 0$
Alternativa, com base no $valor\text{-}p$ bilateral calculado na alínea i): substituindo em H_1 μ_1 e μ_2 por $\overline{x}_1$ e $\overline{x}_2$, respectivamente, $\overline{x}_1 - \overline{x}_2 = 168{,}46 - 158{,}48 > 0$ dá uma proposição verdadeira. Logo,

† O teste de Levene é apresentado na secção 9.4.2.

$$\text{valor } p_{uni} \approx \frac{0}{2} = 0.$$

A hipótese $H_0: \mu_1 \leq \mu_2$ é rejeitada para níveis de significância de aproximadamente 0. Assim, para qualquer nível de significância sensato/usual ($\leq 10\%$) existe evidência de que em média os homens são mais altos do que as mulheres.

e) $\alpha = 2{,}5\%$, $\mu_1 < \mu_2$?

$$H_0: \mu_1 \geq \mu_2 \; vs \; H_1: \mu_1 < \mu_2$$
$$\Leftrightarrow H_0: \mu_1 - \mu_2 \geq 0 \; vs \; H_1: \mu_1 - \mu_2 < 0 \text{ (teste unilateral esquerdo).}$$

Estatística de teste:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{\circ}{\sim} t_v, \text{ onde } v = \left[\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)^2} \right].$$

Como a estatística de teste é a mesma da alínea anterior, $t_{obs} = 29{,}816$ e $v = 1705$.

Pela tabela $t_{v;\, 1-\alpha} = t_{1705\,;\, 0{,}975} = 1{,}96$.
Logo, $R.A.:]-1{,}96; +\infty[$ e $R.R.:]-\infty;\, -1{,}96]$.

Como $t_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 2,5%, não existe evidência estatística de que, em média, a altura dos homens seja inferior à das mulheres.

f) valor $p = P(T \leq t_{obs}) = P(T \leq 29{,}816\,) \approx 1$.
Alternativa, com base no $valor\text{-}p$ bilateral calculado na alínea i): substituindo em H_1 μ_1 e μ_2 por $\overline{x}_1$ e $\overline{x}_2$, respectivamente, $\overline{x}_1 - \overline{x}_2 = 168{,}46 - 158{,}48 > 0$ dá uma proposição falsa. Logo,

$$\text{valor } p_{uni} \approx 1 - \frac{0}{2} = 1.$$

A hipótese $H_0: \mu_1 \geq \mu_2$ é rejeitada para níveis de significância de aproximadamente 1. Assim, para qualquer nível de significância sensato/usual ($\leq 10\%$) não existe evidência de que em média a altura dos homens seja inferior à das mulheres.

8.13.3 Teste de hipótese para a proporção

Numa determinada cidade recolheu-se uma amostra aleatória de 150 homens tendo 54 afirmado que viam o telejornal todos os dias.

a) Teste a hipótese, ao nível de significância de 10%, da proporção de homens, daquela cidade, que veem o telejornal todos os dias ser:
   i. Diferente de 0,40?
   ii. Superior a 0,40?
   iii. Inferior a 0,40?

b) Para cada um dos testes anteriores calcule o respetivo $valor\ p$ e interprete.

Resolução:

Sejam:

- X_i a v. a. que designa se o i-ésimo homem afirmou ver o telejornal,
- $\overline{P}$ a v. a. que representa a proporção de homens que afirmaram ver o telejornal, em n homens.

$n = 150$ e $\overline{p} = \frac{54}{150} = 0{,}36$.

a) $\alpha = 10\%$.

i) $p \neq 0{,}4$?

$H_0: p = 0{,}4 \; vs \; H_1: p \neq 0{,}4$ (teste bilateral).

Estatística de teste:

$$Z = \frac{\overline{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{\circ}{\sim} N(0;1).$$

$$z_{obs} = \frac{0{,}36 - 0{,}4}{\sqrt{\frac{0{,}4(1-0{,}4)}{150}}} = -1.$$

Pela tabela $z_{1-\frac{\alpha}{2}} = z_{0{,}95} = 1{,}645$.

Logo, $R.A.:]-1{,}645;\ 1{,}645[$ e $R.R.:]-\infty;\ -1{,}645] \cup [1{,}645; +\infty[$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 10%, não existe evidência estatística de que a proporção de homens que afirmam ver o telejornal todos os dias naquela cidade é significativamente diferente de 0,4.

ii) $p > 0{,}4$?

$H_0: \mu \leq 0{,}4 \; vs \; H_1: \mu > 0{,}4$ (teste unilateral direito).

Estatística de teste:

$$Z = \frac{\overline{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{\circ}{\sim} N(0;1).$$

Como a estatística de teste é a mesma da alínea anterior, $z_{obs} = -1$.

Pela tabela $z_{1-\alpha} = z_{0{,}9} = 1{,}282$. Logo, $R.A.:]-\infty;\ 1{,}282[$ e $R.R.: [1{,}282; +\infty[$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 10%, não existe evidência estatística de que a proporção de homens que afirmam ver o telejornal todos os dias naquela cidade é significativamente superior a 0,4.

iii) $p < 0{,}4$?

$H_0: p \geq 0{,}4 \; vs \; H_1: p < 0{,}4$ (teste unilateral esquerdo).

Estatística de teste:

$$Z = \frac{\overline{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{\circ}{\sim} N(0;1).$$

Como a estatística de teste é a mesma da alínea anterior, $z_{obs} = -1$.

Pela tabela $z_{1-\alpha} = z_{0{,}9} = 1{,}282$. Logo, $R.A.:]-1{,}282; +\infty[$ e $R.R.:]-\infty;\ -1{,}282]$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 10%, não existe evidência estatística de que a proporção de homens que afirmam ver o telejornal todos os dias naquela cidade é significativamente inferior a 0,4.

b) valor $p = P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira}) = P(Z_{obs} \in R.R. | \mu = \mu_0)$.

i) valor $p = 2 \times P(Z \geq |z_{obs}|) = 2 \times P(Z \geq 1) = 2 \times \big(1 - \Phi(1)\big) = 2 \times (1 - 0{,}8413) = 0{,}3173$.

Logo, a hipótese H_0: $p = 0{,}4$ é rejeitada para níveis de significância superiores ou iguais a 31,73%, logo não existe evidência estatística que a proporção de homens que vê o telejornal todos os dias não seja de 40%.

ii) $\text{valor } p = P(Z \geq z_{obs}) = P(Z \geq -1) = 1 - P(Z < -1) = 1 - \Phi(-1) = \Phi(1) = 0{,}8413.$
Alternativa, com base no $\text{valor } p$ bilateral calculado na alínea i): substituindo em H_1 p por $\overline{p}$, $\overline{p} = 0{,}36 < 0{,}4$ dá uma proposição falsa. Logo,

$$\text{valor } p_{uni} = 1 - \frac{0{,}3173}{2} = 1 - 0{,}1587 = 0{,}8413.$$

A hipótese H_0: $p \leq 0{,}4$ é rejeitada para níveis de significância superiores ou iguais a 84,13%, logo não existe evidência estatística que a proporção de homens que vê o telejornal todos os dias seja superior a 40%.

iii) $\text{valor } p = P(Z \leq z_{obs}) = P(Z \leq -1) = \Phi(-1) = 1 - \Phi(1) = 1 - 0{,}8413 = 0{,}1587.$
Alternativa, com base no $\text{valor } p$ bilateral calculado na alínea i): substituindo em H_1 p por $\overline{p}$, $\overline{p} = 0{,}36 < 0{,}4$ dá uma proposição verdadeira. Logo,

$$valor\text{-}p_{uni} = \frac{0{,}3173}{2} = 0{,}1587.$$

A hipótese H_0: $p \geq 0{,}4$ é rejeitada para níveis de significância superiores ou iguais a 15,87%, logo não existe evidência estatística que a proporção de homens que vê o telejornal todos os dias seja inferior a 40%.

8.13.4 Teste de hipótese para a diferença de proporções

Realizou-se um estudo em duas cidades, A e B, sobre a percentagem de homens que viam o telejornal todos os dias. Na cidade A inquiriram-se aleatoriamente 150 homens tendo 54 afirmado que viam o telejornal todos os dias ao passo que na cidade B dos 200 inquiridos 80 fizeram tal afirmação.

a) Ao nível de significância de 5%, será de admitir que a proporção de homens que vê o telejornal todos os dias é:
   i. Diferente nas duas cidades?
   ii. Inferior na cidade B?
   iii. Superior na cidade B?

b) Para cada um dos testes anteriores calcule o nível de significância a partir do qual rejeita a hipótese nula.

Resolução:

Sejam:

- X_{1i} a v. a. que designa se o i-ésimo homem, da cidade A, afirmou ver o telejornal, $i = 1, \dots, n_1$,
- X_{2i} a v. a. que designa se o i-ésimo homem, da cidade B, afirmou ver o telejornal, $i = 1, \dots, n_2$,
- $\overline{P}_1$ a v. a. que representa a proporção de homens, da cidade A, que afirmaram ver o telejornal, em n_1 homens,
- $\overline{P}_2$ a v. a. que representa a proporção de homens, da cidade B, que afirmaram ver o telejornal, em n_2 homens.

$n_1 = 150;\ \overline{p}_1 = \frac{54}{150} = 0{,}36;\ n_2 = 200$ e $\overline{p}_2 = \frac{80}{200} = 0{,}4.$

a) $\alpha = 5\%.$

i) $p_1 \neq p_2$?

$H_0: p_1 = p_2 \; vs. H_1: p_1 \neq p_2$
$\Leftrightarrow H_0: \; p_1 - p_2 = 0 \; vs \; H_1: \; p_1 - p_2 \neq 0$ (teste bilateral).

Estatística de teste:

$$Z = \frac{\left(\overline{P}_1 - \overline{P}_2\right) - (p_1 - p_2)_0}{\sqrt{\overline{P}^*\left(1 - \overline{P}^*\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{\circ}{\sim} N(0;1), \text{ onde } \overline{P}^* = \frac{n_1\overline{P}_1 + n_2\overline{P}_2}{n_1 + n_2}.$$

$$\overline{p}^* = \frac{n_1\overline{p}_1 + n_2\overline{p}_2}{n_1 + n_2} = \frac{150 \times 0{,}36 + 200 \times 0{,}4}{150 + 200} = 0{,}3829,$$

$$z_{obs} = \frac{(0{,}36 - 0{,}4) - 0}{\sqrt{0{,}3829(1 - 0{,}3829)\left(\frac{1}{150} + \frac{1}{200}\right)}} = -0{,}7619.$$

Pela tabela $z_{1-\frac{\alpha}{2}} = z_{0{,}975} = 1{,}96$.

Logo, $R.A.:]-1{,}96;\; 1{,}96[$ e $R.R.:]-\infty;\; -1{,}96] \cup [1{,}96; +\infty[$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que a proporção de homens que afirmam ver o telejornal todos os dias é significativamente diferente nas duas cidades.

ii) $p_1 > p_2$?

$H_0: p_1 \leq p_2 \; vs. H_1: p_1 > p_2$
$\Leftrightarrow H_0: p_1 - p_2 \leq 0 \; vs \; H_1: p_1 - p_2 > 0$ (teste unilateral direito).

Estatística de teste:

$$Z = \frac{\left(\overline{P}_1 - \overline{P}_2\right) - (p_1 - p_2)_0}{\sqrt{\overline{P}^*\left(1 - \overline{P}^*\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{\circ}{\sim} N(0;1), \text{ onde } \overline{P}^* = \frac{n_1\overline{P}_1 + n_2\overline{P}_2}{n_1 + n_2}.$$

Como a estatística de teste é a mesma da alínea anterior, $z_{obs} = -0{,}7619$.

Pela tabela $z_{1-\alpha} = z_{0{,}95} = 1{,}645$. Logo, $R.A.:]-\infty;\; 1{,}645[$ e $R.R.: [1{,}645; +\infty[$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que a proporção de homens que afirmam ver o telejornal todos os dias seja significativamente inferior na cidade B.

iii) $p_1 < p_2$?

$H_0: p_1 \geq p_2 \; vs. H_1: p_1 < p_2$
$\Leftrightarrow H_0: p_1 - p_2 \geq 0 \; vs \; H_1: p_1 - p_2 < 0$ (teste unilateral esquerdo).

Estatística de teste:

$$Z = \frac{\left(\overline{P}_1 - \overline{P}_2\right) - (p_1 - p_2)_0}{\sqrt{\overline{P}^*\left(1 - \overline{P}^*\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{\circ}{\sim} N(0;1), \text{ onde } \overline{P}^* = \frac{n_1\overline{P}_1 + n_2\overline{P}_2}{n_1 + n_2}.$$

Como a estatística de teste é a mesma da alínea anterior, $z_{obs} = -0{,}7619$.

Pela tabela $z_{1-\alpha} = z_{0{,}95} = 1{,}645$. Logo, $R.A.:]-1{,}645; +\infty[$ e $R.R.:]-\infty;\; -1{,}645]$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que a proporção de homens que afirmam ver o telejornal todos os dias seja significativamente superior na cidade B.

b) i) $\text{valor } p = 2 \times P(Z \geq |z_{obs}|) = 2 \times P(Z \geq 0{,}7619) = 2 \times \big(1 - \Phi(0{,}7619)\big)$
$= 2 \times (1 - 0{,}7769) = 0{,}4461.$

A hipótese H_0: $p_1 - p_2 = 0$ é rejeitada para níveis de significância superiores ou iguais a 44,61%, logo não existe evidência de que as proporções sejam diferentes, para qualquer nível se significância usual (1%, 5%, 10%).

ii) $\text{valor } p = P(Z \geq z_{obs}) = P(Z \geq -0{,}7619) = 1 - \Phi(-0{,}7619) = \Phi(0{,}7619) = 0{,}7769.$
Alternativa, com base no $valor\text{-}p$ bilateral calculado na alínea i): substituindo em H_1 p_1 e p_2 por $\overline{p}_1$ e $\overline{p}_2$, respectivamente, $\overline{p}_1 - \overline{p}_2 = 0{,}36 - 0{,}4 < 0$ dá uma proposição falsa. Logo,

$$\text{valor } p_{uni} = 1 - \frac{0{,}4461}{2} = 0{,}7769.$$

A hipótese H_0: $p_1 - p_2 \leq 0$ é rejeitada para níveis de significância superiores ou iguais a 77,69%. logo não existe evidência de que a proporção na cidade A seja superior à verificada na cidade B.

iii) $\text{valor } p = P(Z \leq z_{obs}) = P(Z \leq -0{,}7619) = \Phi(-0{,}7619) = 1 - \Phi(0{,}7619)$
$= 1 - 0{,}7769 = 0{,}2231.$
Alternativa, com base no $valor\text{-}p$ bilateral calculado na alínea i): substituindo em H_1 p_1 e p_2 por $\overline{p}_1$ e $\overline{p}_2$, respectivamente, $\overline{p}_1 - \overline{p}_2 = 0{,}36 - 0{,}4 < 0$ dá uma proposição verdadeira. Logo,

$$\text{valor } p_{uni} = \frac{0{,}4461}{2} = 0{,}2231.$$

A hipótese H_0: $p_1 - p_2 \geq 0$ é rejeitada para níveis de significância superiores ou iguais a 22,31%, logo não existe evidência de que a proporção na cidade A seja inferior à verificada na cidade B.

8.13.5 Teste de hipótese para a variância

Pode considerar-se que o grau de acidez do azeite de determinada marca é uma variável aleatória Normal. Analisou-se uma amostra aleatória de 35 garrafas tendo-se obtido, para o grau de acidez, uma média de 1,3 graus e uma variância de 0,16.

a) Ao nível de significância de 5%, teste a hipótese de a variância populacional ser:
   i. Igual a 0,14.
   ii. Superior 0,14.
   iii. Inferior 0,14.
b) Calcule o $valor\text{-}p$ para cada um dos testes anteriores.
c) Para a alínea a) iii), calcule a potência de teste para $\sigma_1^2 = 0{,}1$.

Resolução:

Seja X a v. a. que representa o grau de acidez do azeite, com $X \sim N(\mu;\ \sigma = ?)$.

$n = 35;\ \ \overline{x} = 1{,}3$ e $s^2 = 0{,}16.$

a) $\alpha = 5\%$

i) $\sigma^2 = 0{,}14?$

H_0: $\sigma^2 = 0{,}14\ vs\ H_1$: $\sigma^2 \neq 0{,}14$ (teste bilateral).

Estatística de teste:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1=34}.$$

$\chi^2_{obs} = \frac{(35-1)0{,}16}{0{,}14} = 38{,}8571.$

Pela tabela, $\chi^2_{n-1;\frac{\alpha}{2}} = \chi^2_{34;\,0{,}025} = 19{,}81$ e $\chi^2_{n-1;1-\frac{\alpha}{2}} = \chi^2_{34;\,0{,}975} = 51{,}97$.

Logo, $R.A.:]19{,}81\ ;\ 51{,}97[$ e $R.R.: [0;\ 19{,}81\ ] \cup [51{,}97; +\infty[$.

Como $\chi^2_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que a variância populacional do grau de acidez seja diferente de 0,14.

ii) $\sigma^2 > 0{,}14$?

$H_0: \sigma^2 \leq 0{,}14\ vs\ H_1: \sigma^2 > 0{,}14$ (teste unilateral direito).

Estatística de teste:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1=34}.$$

Como a estatística de teste é a mesma da alínea anterior, $\chi^2_{obs} = 38{,}8571$.

Pela tabela, $\chi^2_{n-1;\,1-\alpha} = \chi^2_{34;\,0{,}95} = 48{,}60$. Logo, $R.A.: [0;\ 48{,}60[$ e $R.R.: [48{,}60; +\infty[$.

Como $\chi^2_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que a variância populacional do grau de acidez seja superior a 0,14.

iii) $\sigma^2 < 0{,}14$?

$H_0: \sigma^2 \geq 0{,}14\ vs\ H_1: \sigma^2 < 0{,}14$ (teste unilateral esquerdo).

Estatística de teste:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1=34}.$$

Como a estatística de teste é a mesma da alínea anterior, $\chi^2_{obs} = 38{,}8571$.

Pela tabela, $\chi^2_{n-1;\,\alpha} = \chi^2_{34;\,0{,}05} = 21{,}66$. Logo, $R.A.:]21{,}66;\ +\infty[$ $R.R.: [0;\ 21{,}66]$.

Como $\chi^2_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que a variância populacional do grau de acidez seja inferior a 0,14.

b) i) $\text{valor } p = 2 \times min\{P(\chi^2 \leq \chi^2_{obs}); P(\chi^2 \geq \chi^2_{obs})\}$
$= 2 \times min\{P(\chi^2 \leq 38{,}8571); P(\chi^2 \geq 38{,}8571)\}$
$= 2 \times min\{0{,}7399; 0{,}2601\} = 2 \times 0{,}2601 = 0{,}5201.$

A hipótese $H_0: \sigma^2 = 0{,}14$ é rejeitada para níveis de significância superiores ou iguais a 52,01%, logo não existe evidência estatística de que a variância do grau de acidez seja diferente de 0,14.

ii) $\text{valor } p = P(\chi^2 \geq \chi^2_{obs}) = 1 - P(\chi^2 < \chi^2_{obs}) = 1 - P(\chi^2 < 38{,}8571) = 1 - 0{,}7399 = 0{,}2601.$

A hipótese $H_0: \sigma^2 \leq 0{,}14$ é rejeitada para níveis de significância superiores ou iguais a 26,01%, logo não existe evidência estatística de que a variância do grau de acidez seja superior a 0,14, para níveis de significância usuais.

iii) valor $p = P(\chi^2 \leq \chi^2_{obs}) = P(\chi^2 \leq 38{,}8571) = 0{,}7399$.

A hipótese H_0: $\sigma^2 \geq 0{,}14$ é rejeitada para níveis de significância superiores ou iguais a 73,99%, logo não existe evidência estatística de que a variância do grau de acidez seja inferior a 0,14, para níveis de significância usuais.

c) $\pi(\sigma_1^2 = 0{,}1) = P(\text{Rejeitar } H_0 | H_0 \text{ é falsa}) = P(\text{Rejeitar } H_0 | \sigma^2 = 0{,}1)$

No caso do teste unilateral esquerdo rejeita-se H_0 quando $\chi^2_{obs} \leq \chi^2_{n-1;\,\alpha}$. Para $\alpha = 5\%$,

$$\chi^2_{obs} \leq \chi^2_{34;\,0{,}05} \Leftrightarrow \frac{(n-1)S^2}{\sigma_0^2} \leq \chi^2_{34;\,0{,}05} \Leftrightarrow S^2 \leq \chi^2_{34;\,0{,}05}\frac{\sigma_0^2}{(n-1)} \Leftrightarrow S^2 \leq 21{,}66\frac{0{,}14}{(35-1)}$$
$$\Leftrightarrow S^2 \leq 0{,}0892.$$

Logo,

$$\pi(\sigma_1^2 = 0{,}1) = P(S^2 \leq 0{,}089 | \sigma^2 = 0{,}1) = P\left(\chi^2 \leq \frac{34 \times 0{,}0892}{0{,}1}\right) = P(\chi^2 \leq 30{,}32) = 0{,}3513.$$

A potência do teste é de 0,3513, i. e., existe uma probabilidade moderada baixa deste teste detectar que a variância do grau de azeite é inferior a 0,14 quando efectivamente é igual a 0,1.

8.13.6 Teste de hipótese para o razão de variâncias

Nos primeiros 6 meses de vida dois grupos aleatórios de crianças seguiram esquemas de alimentação diferentes: o grupo 1 seguiu o esquema A e o grupo 2 seguiu o esquema B. No quadro seguinte apresentam-se os ganhos em peso, em kg, dessas crianças.

Grupo 1	2,7	3,2	3,6	4,1	2,7	3,2	4,5	3,6	2,7
Grupo 2	4,1	4,5	3,6	2,7	3,6	3,2	4,1		

Sabe-se que as crianças dos dois grupos tinham, ao nascer, aproximadamente pesos iguais.
Admita que as distribuições dos pesos seguem a distribuição Normal.

a) Com base num teste de hipóteses, ao nível de significância de 5%, pode concluir que a variabilidade é igual nos dois grupos?
b) Sem efectuar cálculos diga, justificando, qual a decisão que tomava no âmbito da alínea b), se considerasse um nível de significância de 1%?
c) Para o teste da alínea a) calcule o respectivo valor p e interprete.
d) Para a alínea a), calcule a potência de teste para $\left(\frac{\sigma_1^2}{\sigma_2^2}\right)_1$ = 0,072.

Resolução:

Sejam:

- X_1 a v.a. que representa o ganho em peso, em kg, das crianças alimentadas segundo o esquema A,
- X_2 a v.a. que representa o ganho em peso, em kg, das crianças alimentadas segundo o esquema B,

Com $X_1 \sim N(\mu_1 = ?\,;\, \sigma_1 = ?)$ e $X_2 \sim N(\mu_2 = ?;\, \sigma_2 = ?)$.

$n_1 = 9, \quad \overline{x}_1 = 3{,}3667 \quad \text{e} \quad s_1^2 = 0{,}4150,$
$n_2 = 7, \quad \overline{x}_2 = 3{,}6857 \quad \text{e} \quad s_2^2 = 0{,}3714.$

a) $\alpha = 5\%$, $\sigma_1^2 = \sigma_2^2$?

H_0: $\sigma_1^2 = \sigma_2^2$ vs H_1: $\sigma_1^2 \neq \sigma_2^2$

$\Leftrightarrow H_0$: $\frac{\sigma_1^2}{\sigma_2^2} = 1$ vs H_1: $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$ (teste bilateral).

Estatística de teste:

$$F = \frac{S_1^2}{S_2^2}\left(\frac{\sigma_2^2}{\sigma_1^2}\right)_0 \sim F_{n_1-1;\, n_2-1=8;\, 6}.$$

$f_{obs} = \frac{0{,}4150}{0{,}3714} \times 1 = 1{,}1173.$

Pela tabela, $f_{n_1-1;\, n_2-1;\, \frac{\alpha}{2}} = f_{8;\, 6;\, 0{,}25} = \frac{1}{f_{6;\, 8;\, 0{,}975}} = \frac{1}{4{,}65} = 0{,}215$ e $f_{n_1-1;\, n_2-1;\, 1-\frac{\alpha}{2}} = f_{8;\, 6;\, 0{,}975} = 5{,}6$.
Logo, $R.A.: \,]0{,}215;\ 5{,}6[$ e $R.R.: [0;\ 0{,}215] \cup [5{,}6; +\infty[$

Como $f_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que a variabilidade nos pesos seja significativamente diferente nos dois grupos.

b) Se $\alpha = 1\%$ a decisão tomada no teste anterior era a mesma, ou seja, não rejeitar H_0. Esta situação é originada pelo facto de quando se diminui o nível de significância também se diminui a $R.R.$ e consequentemente a $R.A.$ aumenta. Portanto, se quando α = 5% f_{obs} está na $R.A.$ então α = 1% a situação mantém-se.

c) $$\begin{aligned} \text{valor } p &= 2 \times min\{P(F \le f_{obs});\, P(F \ge f_{obs})\} \\ &= 2 \times min\{P(F \le 1{,}1173);\, P(F \ge 1{,}1173)\} \\ &= 2 \times min\{0{,}5409;\, 0{,}4591\} \\ &= 2 \times 0{,}4591 = 0{,}9181. \end{aligned}$$

A hipótese H_0: $\sigma_1^2 = \sigma_2^2$ é rejeitada para níveis de significância superiores ou iguais a 91,81%, indicando que não existe evidência de que as variâncias sejam diferentes.

d) $\pi\left(\left(\frac{\sigma_1^2}{\sigma_2^2}\right)_1 = 0{,}072\right) = P\left(\text{Rejeitar } H_0 \middle| \frac{\sigma_1^2}{\sigma_2^2} = 0{,}072\right) = 1 - \beta\left(\left(\frac{\sigma_1^2}{\sigma_2^2}\right)_1 = 0{,}072\right)$

No caso do *teste bilateral* rejeita-se H_0 se $f_{obs} \le f_{n_1-1;\, n_2-1;\, \frac{\alpha}{2}}$ ou $f_{obs} \ge f_{n_1-1;\, n_2-1;\, 1-\frac{\alpha}{2}}$. Para $\alpha = 5\%$,

$$\begin{cases} F_{obs} \le f_{8;\, 6;\, 0{,}25} \\ \text{ou} \\ F_{obs} \ge f_{8;\, 6;\, 0{,}975} \end{cases} \Leftrightarrow \begin{cases} \frac{S_1^2}{S_2^2}\left(\frac{\sigma_2^2}{\sigma_1^2}\right)_0 \le f_{8;\, 6;\, 0{,}25} \\ \text{ou} \\ \frac{S_1^2}{S_2^2}\left(\frac{\sigma_2^2}{\sigma_1^2}\right)_0 \ge f_{8;\, 6;\, 0{,}975} \end{cases} \Leftrightarrow \begin{cases} \frac{S_1^2}{S_2^2} \le f_{8;\, 6;\, 0{,}25}\left(\frac{\sigma_1^2}{\sigma_2^2}\right)_0 \\ \text{ou} \\ \frac{S_1^2}{S_2^2} \ge f_{8;\, 6;\, 0{,}975}\left(\frac{\sigma_1^2}{\sigma_2^2}\right)_0 \end{cases} \Leftrightarrow \begin{cases} \frac{S_1^2}{S_2^2} \le 0{,}2150 \times 1 \\ \text{ou} \\ \frac{S_1^2}{S_2^2} \ge 5{,}6 \times 1 \end{cases}$$

Logo,

$$\begin{aligned} \pi\left(\left(\frac{\sigma_1^2}{\sigma_2^2}\right)_1 = 0{,}072\right) &= P\left(\frac{S_1^2}{S_2^2} \le 0{,}2150 \middle| \frac{\sigma_1^2}{\sigma_2^2} = 0{,}072\right) + P\left(\frac{S_1^2}{S_2^2} \ge 5{,}60 \middle| \frac{\sigma_1^2}{\sigma_2^2} = 0{,}072\right) \\ &= P\left(\frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \le \frac{0{,}215}{0{,}072}\right) + P\left(\frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2} \ge \frac{5{,}60}{0{,}072}\right) \\ &= P(F \le 2{,}986) + P(F \ge 77{,}778) \\ &= P(F \le 2{,}986) + \big(1 - P(F < 77{,}778)\big) \\ &\approx 0{,}9002 + (1 - 1) = 0{,}9002. \end{aligned}$$

A potência do teste é de 0,9002, i. e., se o verdadeiro quociente entre as variáveis for de 0,072, com 90,02% de certeza tomaremos a decisão certa com este teste.

8.13.7 Teste de hipótese para amostras emparelhadas

Um professor de estatística seleccionou aleatoriamente um grupo de 10 alunos, aprovados na disciplina pelo regime de frequências, tendo registado as suas notas nas frequências:

Aluno	1	2	3	4	5	6	7	8	9	10
1ª freq.	10	11	9	10	18	15	16	13	11	10
2ª freq.	9	12	12	12	18	14	18	12	13	10

Ao nível de significância de 5% pode afirmar que as notas médias dos alunos na 1ª frequência são superiores às obtidas na 2ª frequência? Assuma a Normalidade das notas.

Resolução:

Sejam:

- X_1 a v.a. que representa a nota do aluno na 1ª frequência,
- X_2 a v.a. que representa a nota do aluno na 2ª frequência,

com $X_1 \sim N(\mu_1 = ?\,; \sigma_1 = ?)$ e $X_2 \sim N(\mu_2 = ?; \sigma_2 = ?)$.

$\alpha = 1\%, \mu_1 > \mu_2$?

$$H_0: \mu_1 \leq \mu_2 \; vs. \; H_1: \mu_1 > \mu_2$$
$$\Leftrightarrow H_0: \mu_1 - \mu_2 \leq 0 \; vs. \; H_1: \mu_1 - \mu_2 > 0$$

como as amostras são emparelhadas

$$\Leftrightarrow H_0: \mu_D \leq 0 \; vs. \; H_1: \mu_D > 0 \text{ (teste unilateral direito).}$$

Estatística de teste:

$$T = \frac{\overline{D} - \mu_0}{\frac{S_D}{\sqrt{n}}} \sim t_{n-1}.$$

Aluno (i)	1	2	3	4	5	6	7	8	9	10
1ª freq. (x_{1i})	10	11	9	10	18	15	16	13	11	10
2ª freq. (x_{2i})	9	12	12	12	18	14	18	12	13	10
$d_i = x_{1i} - x_{2i}$	1	-1	-3	-2	0	1	-2	1	-2	0

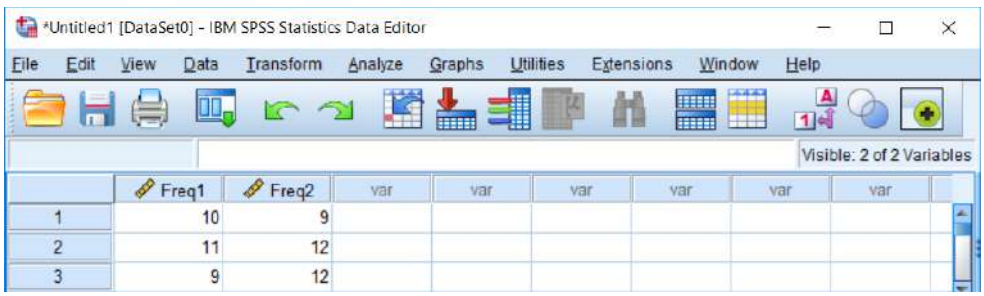
$n = 10; \;\; \overline{d} = -0{,}7$ e $s_d = 1{,}4944.$

$$t_{obs} = \frac{-0{,}7 - 0}{\frac{1{,}4944}{\sqrt{10}}} = -1{,}481.$$

Pela tabela $t_{n-1;\,1-\alpha} = t_{9;\,0{,}95} = 1{,}833$. Logo, $R.A.:]-\infty;\; 1{,}833[$ e $R.R.: [1{,}833; +\infty[$.

Como $t_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que as notas médias obtidas pelos alunos na 1ª frequência sejam superiores às da 2ª frequência.

(SPSS)



(SPSS) Analyse → Compare Means → Paired-Samples T Test...
(Paired Variables → Variable 1: Freq1; Variable 2: Freq2)

T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Freq1	12,30	10	3,057	,967
	Freq2	13,00	10	2,981	,943

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Freq1 & Freq2	10	,878	,001

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Freq1 - Freq2	-,700	1,494	,473	-1,769	,369	-1,481	9	,173

8.13.8 Teste de hipótese para o coeficiente de correlação

Um agricultor pretende estimar a produção total do seu pomar a partir do seu "palpite" sobre o peso da fruta de cada uma das laranjeiras. Para poder utilizar essa estimativa tem de existir correlação positiva entre o "palpite" (X) e o peso real (Y). Colheram-se e pesaram-se as laranjas de uma amostra aleatória de 25 laranjeiras desse pomar tendo-se obtido os seguintes resultados:

$$\sum_{i=1}^{25} x_i = 700; \qquad \sum_{i=1}^{25} x_i^2 = 19994; \qquad \sum_{i=1}^{25} y_i = 660; \qquad \sum_{i=1}^{25} y_i^2 = 18200; \qquad \sum_{i=1}^{25} x_i y_i = 18950$$

Admita que X e Y seguem uma distribuição Normal.

a) Considerando um nível de significância de 5%, diga se é de utilizar aquela estimativa.
b) Uma vez que a estimativa será tanto melhor quanto maior for a correlação entre X e Y, teste, ao nível de significância de 5%, a hipótese de o coeficiente de correlação ser no mínimo 0,9.

Resolução:

Sejam:

- X a v. a. que representa o "peso-palpite" da laranjeira,
- Y a v. a. que representa o peso real da laranjeira,

com $X \sim N(\mu_X = ?;\ \sigma_X = ?)$ e $Y \sim N(\mu_Y = ?;\ \sigma_Y = ?)$.

$n = 25$.

$$r = \frac{\sum_{i=1}^{25} x_i y_i - \frac{1}{n}\sum_{i=1}^{25} x_i \sum_{i=1}^{25} y_i}{\sqrt{\left(\sum_{i=1}^{25} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)\left(\sum_{i=1}^{25} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

$$= \frac{18950 - \frac{1}{25} \times 700 \times 600}{\sqrt{\left(19994 - \frac{1}{25} \times 700^2\right)\left(18200 - \frac{1}{25} \times 660^2\right)}} = 0{,}85$$

a) $\alpha = 5\%, \rho > 0$ (existe correlação positiva)?

$H_0: \rho \leq 0 \; vs \; H_1: \rho > 0$ (teste unilateral direito)

Estatística de teste:

$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2=23}.$$

$$t_{obs} = \frac{0{,}85}{\sqrt{\frac{1-0{,}85^2}{25-2}}} = 7{,}7384.$$

Pela tabela, $t_{n-2;1-\alpha} = t_{23;0,95} = 1{,}714$. Logo, $R.A.:]-\infty;\ 1{,}714[$ e $R.R.: [1{,}714; +\infty[$.

Como $t_{obs} \in R.R.$ rejeitar H_0. Portanto, ao nível de significância de 5%, existe evidência estatística de que existe correlação linear entre o "peso-palpite" e o peso real.

Portanto, a estimativa do peso poderá ser utilizada pelo agricultor.

b) $\alpha = 5\%, \rho \geq 0{,}9$?

$H_0: \rho \geq 0{,}9 \; vs \; H_1: \rho < 0{,}9$ (teste unilateral esquerdo)

Estatística de teste:

$$Z = \frac{Z_r - Z_{\rho_0}}{\frac{1}{\sqrt{n-3}}} \sim N(0;1).$$

$r = 0{,}85 \Rightarrow Z_r = \frac{1}{2} ln\left(\frac{1+0{,}85}{1-0{,}85}\right) = 1{,}2562,$

$\rho_0 = 0{,}9 \Rightarrow Z_{\rho_0} = \frac{1}{2} ln\left(\frac{1+0{,}9}{1-0{,}9}\right) = 1{,}4722,$

$$z_{obs} = \frac{1{,}2562 - 1{,}4722}{\frac{1}{\sqrt{25-3}}} = -1{,}0134.$$

Pela tabela, $z_{1-\alpha} = z_{0,95} = 1{,}645$. Logo, $R.A.:]-1{,}645; +\infty[$ e $R.R.:]-\infty;\ -1{,}645]$.

Como $z_{obs} \in R.A.$ não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência estatística de que o coeficiente de correlação seja inferior a 0,9.

8.14 Exercícios propostos

1. Um funcionário das finanças disse ao seu superior que demorava, em média, 3,5 minutos a verificar as declarações de IRS com um desvio padrão de 1 minuto. Foram observados, aleatoriamente, tempos de verificação (em minutos) respeitantes a 51 declarações, tendo-se obtido os seguintes valores:

$$\sum_{i=1}^{51} x_i = 200{,}1; \qquad \sum_{i=1}^{51} (x_i - \bar{x})^2 = 763{,}48$$

a) Pronuncie-se a favor ou contra a afirmação do funcionário relativamente ao tempo médio, ao nível de significância de 5%.
b) Resolva a alínea anterior admitindo agora que a variância populacional é desconhecida.

2. Doentes com problemas renais podem ser tratados através da diálise, usando uma máquina que remove resíduos tóxicos do sangue, uma função normalmente desempenhada pelos rins. O deficiente funcionamento dos rins e a diálise pode causar outras alterações tais como retenção de fósforo, o que deve ser corrigido na dieta. Um estudo de nutrição de pacientes de hemodiálise registou o nível de fósforo no sangue (em miligramas de fósforo por decilitro de sangue) de vários pacientes em seis ocasiões. Os resultados de um paciente foram:

4,8 3,7 5,5 7,0 6,5 4,7

Supondo que $\sigma = 0{,}9$ mg/dl, e que o nível de fósforo no sangue segue uma distribuição Normal:

a) Construa um intervalo de confiança a 90% para o nível médio de fósforo no sangue.
b) Os valores médios de fósforo no sangue considerados Normais variam entre 2,6 e 4,8 mg/dl. Ao nível de significância de 5% existirá forte evidência de que aquele paciente apresenta um nível médio superior a 4,8?
c) Determine a potência do teste anterior contra a alternativa $\mu = 5{,}5$. Considere o nível de significância de 10%.

3. Os gestores de uma farmacêutica estão preocupados com a variabilidade que ocorre numa das etapas relacionada com o enchimento das embalagens de pomadas. Os gestores afirmam que o conteúdo destas segue uma distribuição Normal com média 150 ml.
Foi retirada uma amostra aleatória de 14 pomadas tendo-se observado o seu conteúdo, em ml:

135 141 142 142 142 143 143 147 149 149 150 150 152 159

a) Teste a veracidade da afirmação dos gestores aos níveis de significância de 1% e 10%, relativamente à média populacional.
b) Qual o tipo de erro associado a cada uma das decisões anteriores?
c) Como poderia realizar o teste da alínea a) recorrendo aos intervalos de confiança?
d) Sabendo que $\beta(\mu = \mu_1)$ representa a probabilidade de cometer um erro de tipo II quando a verdadeira média é igual a μ_1, indique, justificando *sem efectuar cálculos*, qual das afirmações está correcta, relativamente ao teste realizado:
   i. $\beta(\mu = 145) > \beta(\mu = 155)$;
   ii. $\beta(\mu = 145) = \beta(\mu = 155)$;
   iii. $\beta(\mu = 145) < \beta(\mu = 155)$.
e) Calcule a probabilidade de rejeitar indevidamente H_0.
f) A defesa do Consumidor recebeu vários protestos quanto ao conteúdo médio por embalagens, pelo que processou esta farmacêutica argumentando que o conteúdo médio seria inferior ao declarado pelos ditos gestores.
   i. Ao nível de significância de 10% a quem daria razão? E a 5% e 1%?
   ii. Qual o tipo de erro associado a cada uma das decisões anteriores?
   iii. Calcule a probabilidade de rejeitar indevidamente H_0.
   iv. *Sem efectuar cálculos*, qual das afirmações está correcta, relativamente ao teste realizado:
      i1) $\beta(\mu = 146) > \beta(\mu = 148)$;
      i2) $\beta(\mu = 146) = \beta(\mu = 148)$;
      i3) $\beta(\mu = 146) < \beta(\mu = 148)$.

4. Comente a seguinte proposição: *"A potência de um teste diminui à medida que o valor do parâmetro* alternativo *se aproxima do valor do parâmetro da hipótese nula"*.

5. Pode considerar-se que a variação da temperatura corporal é uma variável aleatória Normal.
Analisou-se uma amostra aleatória de 35 adultos, registou-se a sua temperatura corporal em 2 períodos de tempos distintos, tendo-se obtido uma variação média de 1,2 graus e uma variância 0,16.

a) Ao nível de significância de 8%, teste se a variação média expectável poderá ser considerada:
   i. igual a 1,3 graus.
   ii. inferior a 1,3 graus.
b) Calcule o $valor\text{-}p$ para cada um dos testes anteriores.

6. Duas amostras, A_1 com 40 lâmpadas e A_2 com 50 lâmpadas, foram obtidas de dois lotes muito numerosos, L_1 e L_2. As médias dos tempos de vida das lâmpadas observadas foram respetivamente de 1760 horas e 1800 horas.
Admita que o desvio padrão do tempo de vida de uma lâmpada é de 200 horas.
a) Proceda ao teste da hipótese: "os verdadeiros tempos de vida dos dois lotes são iguais". Considere o nível de significância de 1%.
b) Construa o intervalo de confiança a 95% para o valor médio do tempo de vida das lâmpadas do lote L_1.

7. Havendo indícios de que o esquema de avaliação e as classificações finais atribuídas diferem fortemente entre duas escolas, decidiu-se comprovar estatisticamente esta hipótese. Assim, retirou-se uma amostra de testes de alunos em cada uma das escolas tendo-se observado os seguintes resultados:

Escola A	11,9	12,3	12,4	14,1	14,3	15,1	15,5	15,5	16,0	16,2	16,2
	16,7	17,0	17,3	17,3	17,5	18,7	18,8	18,9	19,6	21,1	
Escola B	9,7	11,1	11,7	12,1	12,3	13,1	13,2	13,5	14,9	15,0	15,5
	17,2	17,7	18,0	20,4	20,5						

a) Com base numa análise gráfica pode considerar-se que os dados de cada uma das escolas são provenientes de uma população com distribuição Normal?
b) Teste, ao nível de significância de 5%, se a classificação média na escola A é igual a 15 valores.
c) Ao nível de significância de 1%, pode-se considerar que a classificação média na escola B é no máximo 14 valores?
d) Teste, ao nível de significância de 5%, a igualdade da variância das classificações nas duas escolas.
e) Para α = 10%, diga se a classificação média da escola A é igual à da escola B.
f) Uma comissão de avaliação independente suspeita que a classificação média da escola A é mais alta que na escola B.
   i. Ao nível de significância de 10%, concorda com a suspeita?
   ii. Mantém a sua decisão ao nível de significância de 5%? Justifique.
   iii. A partir de que nível de significância rejeita a hipótese nula do teste anterior?

8. Uma determinada empresa farmacêutica lançou no mercado um novo medicamento, para dormir, que tem estado a ser utilizado nos hospitais. Administrou-se este medicamento a um grupo de 51 doentes tendo-se observado que em média estes dormiram 7,5 horas sendo o desvio padrão de 2 horas.
a) Ao nível de significância de 10%, teste a hipótese de os doentes dormirem em média 8 horas.
b) A partir de que nível de significância rejeita a hipótese testada na alínea a)?
c) Um grupo de médicos suspeitando que o medicamento não está a ser eficaz, decidiu observar 41 doentes, em condições similares dos anteriores, não sujeitos ao referido medicamento. Para este grupo observou-se que em média dormiram 7 horas sendo o desvio padrão de 1,5 horas.
   i. Teste, ao nível de significância de 1%, a igualdade da variabilidade entre as horas de sono.
   ii. Concorda com a suspeita dos médicos ao nível de significância de 10%?

iii. Admita que as variâncias populacionais são conhecidas, $\sigma_1^2 = 2{,}25$ e $\sigma_2^2 = 4$, e que se realizou o teste através do Excel tendo-se obtido o output que se apresenta de seguida. Qual seria agora a sua opinião?

Teste z: duas amostras para médias

	grupo 1	*grupo 2*
Média	7,5	7
Variância conhecida	2,25	4
Observações	51	41
Hipótese de diferença de média	0	
z	1,3284	
P(Z<=z) uni-caudal	0,0920	
z crítico uni-caudal	1,2816	
P(Z<=z) bi-caudal	0,1841	
z crítico bi-caudal	1,6449	

9. Admita a Normalidade do consumo médio semanal de leite em famílias com 4 pessoas (2 adultos e 2 crianças) em duas zonas com diferentes características do país (Urbano e Rural). Efetuou-se uma experiência com o objetivo de comparar esses dois tipos de populações em termos consumos médios semanais de leite (em litros) tendo a amostragem dado os seguintes resultados:

Zona urbana	11,2	11,1	11,1	10,3	10,2	9,4	7,6	9,9	
Zona rural	10,7	10,3	10,8	12,5	10,7	10,3	11,0	7,8	12,2

a) Teste, ao nível de 5%, se consumo médio semanal de leite na população urbana é superior a 10 litros.
b) Calcule a potência do teste anterior (alínea a)) para $\mu_{urbano} = 11{,}5$ litros.
c) Ao nível de 5% teste a igualdade de variâncias do consumo semanal de leite nas duas populações.
d) Compare, ao nível de 1%, os consumos médios semanais das duas populações.

10. Com vista a avaliar um certo método de treino de atletas, foi realizada a seguinte experiência:
- foram escolhidos 5 atletas ao acaso **antes de submetidos ao treino** em causa;
- foi efetuado um teste a estes atletas, tendo-se obtido as seguintes marcas:

Atletas	A	B	C	D	E	$\bar{x}$	$\sum(x_i - \bar{x})^2$
Marcas	168	195	155	183	169	174	944

- **depois de submetidos ao treino** foram escolhidos outros 5 atletas que obtiveram as seguintes marcas no teste a que foram submetidos:

Atletas	F	G	H	I	J	$\bar{x}$	$\sum(x_i - \bar{x})^2$
Marcas	183	177	148	162	180	170	866

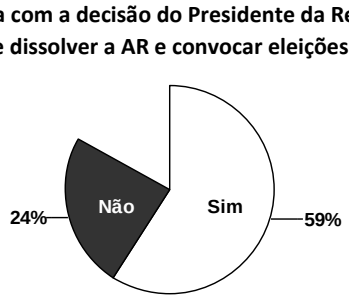
a) Obtenha o intervalo de confiança a 95% para o valor médio das marcas **antes** do treino.
b) Teste a igualdade de variâncias das marcas obtidas pelos atletas antes e depois do treino, ao nível de significância de 1%.
c) Obtenha o intervalo de confiança a 95% para a diferença dos valores médios das marcas **antes** e **depois** do treino.
d) Teste, ao nível de significância de 1%, se o treino introduz melhorias significativas na média das marcas obtidas, i.e., os atletas aumentam as suas marcas.

11. Supõe-se que menos de 60% dos alunos do curso de engenharia informática possui computador pessoal.
    a) Formule a hipótese nula e a hipótese alternativa.
    b) Numa amostra aleatória de 375 estudantes verificou-se que 217 possuíam computador. Poderá rejeitar H_0 ao nível de significância $\alpha = 0{,}05$? Qual o tipo de erro a que se sujeita? Explique as consequências práticas de cometer esse erro.

12. Pensa-se que a maioria dos engenheiros de minas que se formaram em 1980 estão atualmente colocados na indústria de minério de carvão.
    a) Formule a hipótese nula e a hipótese alternativa.
    b) Uma amostra aleatória de 50 desses profissionais foi selecionada e verificou-se que 29 trabalhavam na indústria de carvão. Será de rejeitar a hipótese nula?

13. O responsável pela requisição de material de escritório de uma empresa verificou que 50% da quantidade de esferográficas requisitadas mensalmente tinham o logótipo da empresa, mais caras do que as esferográficas Normais. Na verdade, elas eram realmente bonitas e de boa qualidade pelo que ele próprio já por diversas vezes as tinha oferecido a familiares e amigos. Suspeitando que o resto do pessoal da empresa fazia o mesmo e, atendendo a que tinha recebido ordens superiores para reduzir o mais possível os custos da empresa em material de escritório, fez passar uma circular pela empresa em que informava que, quem precisasse daquele tipo de esferográficas tinha de entregar a esferográfica antiga já gasta.
Passado algum tempo, num dos meses posteriores a ter tomado esta medida, verificou que em 1347 pedidos de esferográficas, 362 eram das ditas esferográficas. Teste, ao nível de significância de 5%, uma hipótese que lhe permita dizer se a medida implementada fez ou não baixar a quantidade de pedidos daquele tipo de esferográfica.

14. A 4 de Dezembro de 2004, o jornal Expresso publicou os resultados de uma sondagem realizada pela Eurosondagem para o jornal Expresso, SIC e Rádio Renascença, efetuada nos dias 1 e 2 de Dezembro. As entrevistas foram realizadas telefonicamente junto de pessoas com pelo menos 18 anos residindo em Portugal Continental. Foram efetuadas 1321 tentativas de entrevistas telefónicas, mas 265 recusaram responder. Foram validadas 1056 entrevistas. Admita para a escolha dos lares foi feita através de uma amostragem aleatória simples.
Relativamente à questão "Concorda com a decisão do Presidente da República de dissolver a Assembleia da República e convocar eleições?", foi publicada a seguinte informação:



    a) Quantas entrevistas telefónicas foram realizadas com sucesso?
    b) Dê uma estimativa pontual para a proporção de pessoas que concordaram com a decisão Sr. Presidente da República Jorge Sampaio.
    c) Ao nível de significância de 1% de confiança, pode-se considerar que a verdadeira percentagem de pessoas que concordam com a decisão do Presidente da República é 64%? Justifique.
    d) Como pode mudar a opinião anterior mantendo o grau de confiança e o valor da proporção amostral? Justifique *sem efetuar os cálculos*.

e) Na véspera, o jornal "Independente" publicou os resultados de uma sondagem realizada pelo IPOM – Instituto de Pesquisa de Opinião e Mercado – para este jornal. As entrevistas foram realizadas telefonicamente junto de pessoas com pelo menos 18 anos, recenseados e eleitores nos círculos eleitorais do Continente. Os resultados publicados referem-se a uma amostra de 578 entrevistas. Admita para a escolha dos lares foi feita através de uma amostragem aleatória simples.
Este jornal apresentou a seguinte informação:

Concorda com a decisão do Presidente de dissolver a Assembleia da República?	74,2% 19,2% 6,6%	Sim Não NS/NR

i. Ao nível de significância de 5%, pode-se considerar que a verdadeira percentagem de indivíduos que *não* concordam com a decisão do Presidente da República é superior na sondagem apresentada pelo jornal "Expresso" do que a apresentada pelo jornal "Independente"?
ii. Mantém a sua opinião ao nível de significância de 10%? Justifique *sem efetuar os cálculos*.
iii. A partir de que nível de significância rejeita a hipótese anterior?

15. Foi elaborado um determinado projeto paisagístico e, para estudar a sua favorabilidade, recolheu-se a opinião junto de 860 indivíduos tendo 387 manifestando-se favoráveis ao dito projeto. Duas semanas depois, numa amostra de 950 indivíduos obtiveram-se 456 respostas favoráveis ao projeto.

a) Teste, ao nível de significância de 1% pode-se considerar que, duas semanas depois, metade dos indivíduos era favorável ao projeto?
b) Ao nível de significância de 5%, pode-se considerar que houve um aumento na proporção de respostas favoráveis ao projeto duas semanas depois?

16. Nos distritos A e B realizou-se um inquérito, com vista a estudar o comportamento das mães com crianças com idade inferior a 1 ano. Inquiram-se 400 mães no distrito A e 350 no distrito B, tendo-se apurado que o número de crianças convenientemente acompanhadas pelos serviços médicos foi 300 no distrito A e 224 no distrito B.

a) Ao nível de significância de 1%, teste a hipótese de menos de 80% das crianças do distrito A terem sido convenientemente acompanhadas pelos serviços médicos.
b) A partir de que nível de significância rejeita a hipótese nula do teste anterior?
c) Averigue se a diferença encontrada nos dois distritos A e B é estatisticamente significativa (α = 5%).

17. Com o aproximar da época balnear, são várias as pessoas que iniciam programas de dieta para perder as gorduras acumuladas durante o inverno. Na *TVCompras* é anunciado um medicamento, à base de estratos naturais de plantas, cujo slogan é "Perca mais de 5 kg em 8 dias". Selecionaram-se aleatoriamente 24 pessoas às quais foi administrado este medicamento foi administrado. No quadro seguinte apresentam-se os seus pesos antes de tomarem o medicamento e 10 dias depois.

Sujeito	Antes	Depois	Sujeito	Antes	Depois	Sujeito	Antes	Depois
1	59	55	9	90	95	17	98	84
2	60	59	10	113	101	18	56	55
3	109	97	11	59	55	19	55	55
4	88	78	12	79	76	20	59	50
5	102	96	13	112	95	21	58	57
6	72	65	14	104	95	22	83	80
7	102	98	15	58	54	23	73	71
8	85	83	16	60	53	24	80	76

Admita que os pesos seguem uma distribuição Normal.

a) Ao nível de significância de 5%, pronuncie-se quanto à veracidade do slogan.
b) Ao nível de significância de 10%, considera que existe relação linear entre as duas variáveis?
c) Ao nível de significância de 10%, teste a hipótese de o coeficiente de correlação populacional ser 0,80.

18. Um investigador desenvolveu um novo programa de ensino individualizado que acredita aumentar o QI dum indivíduo. Escolheram-se aleatoriamente 8 alunos do ensino do 2º ciclo para participar em tal programa. Na tabela seguinte apresentam-se os seus QI registados, de forma equivalente, antes e depois do referido programa:

Aluno	1	2	3	4	5	6	7	8
Antes	81	89	90	97	108	111	118	124
Depois	89	88	94	96	118	111	121	121

ou seja,

$\overline{x} = 102{,}25;$ $\overline{y} = 104{,}75;$ $s_x = 15{,}2854;$ $s_y = 14{,}4593;$ $s_{xy} = 210{,}6429.$

Admita a Normalidade da população.

a) Concorda com o investigador ao nível de significância de 5%?
b) Ao nível de significância de 1%, considera que existe relação linear positiva entre as duas variáveis?
c) Teste, ao nível de significância de 5%, a hipótese de o coeficiente de correlação populacional ser no máximo 0,90.

9 Análise de variância - ANOVA

A Análise de Variância – ANOVA (**An**alysis **o**f **Va**riance) – foi introduzida pelo Sr. Ronald Fisher, com aplicações iniciais no domínio da agronomia e biologia, derivando daí grande parte da terminologia utilizada.
A ANOVA é uma técnica que permite analisar dados que são afetados por várias condições externas (fatores) que podem ou não operar em simultâneo.

Designa-se por **fator** a característica que permite distinguir os diferentes grupos. Desta forma, a cada grupo corresponde um **nível** do fator, tendo um fator K níveis.

Quando se pretende comparar mais de duas médias, poder-se-ia testar a igualdade entre todos os pares de médias através dos testes t, por exemplo, no entanto, esta não é a melhor solução uma vez que aumenta consideravelmente o erro de tipo I. A Análise de Variância constitui assim o procedimento adequado para comparar mais de duas médias.

Neste capítulo apenas se abordam os casos mais simples (1 e 2 fatores) da análise de variância que permitem comparar a igualdade de K médias populacionais e, quando justificável, a interação. Para o efeito, esta técnica baseia-se na comparação da variabilidade em torno de cada uma das médias amostrais e a variabilidade entre as médias amostrais, justificando o nome Análise de Variância. Aborda-se também somente o chamado modelo de efeitos fixos, ou seja, o caso em que os níveis do fator são fixos (estudam-se todos os níveis possíveis do fator).

Pressupostos:

- As amostras têm de ser aleatórias para se garantir a independência;
- As populações têm distribuição Normal;
- As populações têm a mesma variância (homocedasticidade).

Observações:

- A ANOVA é robusta ao pressuposto de Normalidade, desde que a distribuição populacional seja aproximadamente simétrica e mesocúrtica;
- A ANOVA é robusta a violações de homocedasticidade quando o número de observações em cada grupo é igual ou aproximadamente igual. Quando a igualdade de variâncias não pode ser assumida deve-se usar a estatística de Brown-Forsythe (1974a,b) ou a estatística de Welch (1947, 1951), em vez da Estatística F. Posteriormente, caso exista evidência de que as médias não são todas iguais, devem utilizar-se testes de comparação múltipla que não assumam a igualdade de variâncias.
- Neste livro, apenas será apresentado teoricamente o teste F e dois testes de comparação múltipla que assumem a igualdade de variâncias (Tuckey e Scheffé), sendo os restantes casos abordados com o apoio do SPSS.

9.1 Análise de variância simples

Aplica-se a análise de variância simples, ou a 1 fator, quando os valores amostrais estão separados em grupos segundo uma só característica (fator).

Objetivo: Testar a igualdade de três ou mais médias populacionais, isto é, testar se para um determinado fator a média é igual para todos os seus níveis.

Notação:

K número de níveis (grupos) do fator;

n_i número de observações no nível $i,\ i = 1, \dots, K$;

x_{ij} valor observado no nível i para o indivíduo $j,\ i = 1, \dots, K, j = 1, \dots, n_i$;

$n = \sum_{i=1}^{K} n_i$ número total de observações;

$\overline{\overline{X}} = \sum_{i=1}^{K}\sum_{j=1}^{n_i} \frac{X_{ij}}{n} = \sum_{i=1}^{K} \frac{n_i \overline{X}_i}{n}$ média global;

$\overline{X}_i = \sum_{j=1}^{n_i} \frac{X_{ij}}{n_j}$ média do nível i;

$S^2 = \sum_{i=1}^{K}\sum_{j=1}^{n_i} \frac{\left(X_{ij} - \overline{\overline{X}}\right)^2}{n-1} = \frac{1}{n-1}\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} X_{ij}^2 - n\overline{\overline{X}}^2\right)$ variância amostral global;

$S_i^2 = \sum_{j=1}^{n_i} \frac{\left(X_{ij} - \overline{X}_i\right)^2}{n_i - 1} = \frac{1}{n_i - 1}\left(\sum_{j=1}^{n_i} X_{ij}^2 - n_i \overline{X}_i^2\right)$ variância amostral do nível i;

$SQA = \sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(\overline{X}_i - \overline{\overline{X}}\right) = \sum_{i=1}^{K} n_i \left(\overline{X}_i - \overline{\overline{X}}\right)^2$ soma dos quadrados dos desvios entre os níveis do fator;

$SQE = \sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)^2 = \sum_{i=1}^{K} (n_i - 1) S_i^2$ soma dos quadrados dos desvios dentro dos níveis do fator;

$SQT = \sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(X_{ij} - \overline{\overline{X}}\right)^2 = (n-1)S^2$ soma dos quadrados dos desvios totais em torno da média global;

$MQA = \frac{SQA}{K-1}$ média dos quadrados entre os níveis;

$MQE = \frac{SQE}{n-K}$ média dos quadrados dentro dos níveis;

$MQT = \frac{SQT}{n-1}$ média dos quadrados totais.

Modelo matemático:

$$X_{ij} = \mu_i + \varepsilon_{ij} \Leftrightarrow X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad i = 1, \dots, K, \qquad j = 1, \dots, n_i,$$

onde:

$\mu_i = \mu + \alpha_i$ média populacional do nível i do fator;

μ média da população;

α_i efeito do fator;

ε_{ij} Resíduo.

Decomposição da variância:

Variação total = Variação explicada pelo fator independente + Variação devida ao erro,

i.e.,

$$SQT = SQA + SQE.$$

Hipóteses a testar:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_K$ vs $H_1: \exists_{i \neq j}: \mu_i \neq \mu_j$ para $i, j\ i = 1, \ldots, K$

(i.e., nem todas as médias μ_i são iguais).

Estatística de teste:

$$F = \frac{MQA}{MQE} \sim F_{K-1;\, n-K}.$$

Regra de decisão: Rejeitar H_0 se

$$f_{obs} \geq F_{K-1;\, n-K;\, 1-\alpha} \text{ (Teste unilateral direito),}$$

sendo $F_{K-1;\, n-K;\, 1-\alpha}$ o quantil de probabilidade $(1-\alpha)$ da distribuição $F_{k-1;\, n-K}$.

Tabela da análise de variância (ANOVA):

Fontes de Variação (F. V.)	Soma dos Quadrados (SQ)	Graus de Liberdade (g. l.)	Média dos Quadrados (MQ)	F
Entre os grupos (fator)	SQA	$K-1$	MQA	$f_{obs} = \frac{MQA}{MQE}$
Dentro dos grupos (residual)	SQE	$n-K$	MQE	
Total	SQT	$n-1$		

Observação: Para rejeitar H_0 basta que pelo menos 2 médias sejam diferentes.

9.2 Análise de variância dupla

Aplica-se a análise de variância dupla, ou a 2 fator, quando os valores amostrais estão separados em grupos segundo duas só características (fatores).

Designa-se por **célula** o conjunto de valores observados para o nível i do fator A e do nível j do fator B. Se existe mais do que uma observação por célula, as **repetições**, então estas são designadas por **réplicas**.

A principal desvantagem dos modelos com 1 só fator é não se poder considerar a interação entre os fatores. Além disso, os modelos com mais do que um fator são mais eficientes do que os modelos com um só fator.

Diz-se que existe **interação** quando para diferentes níveis de um fator a variável resposta não tem o mesmo efeito nos diferentes níveis do outro fator.

Graficamente, a interação é indicada pela falta de paralelismo entre as linhas (Figura 9.1).

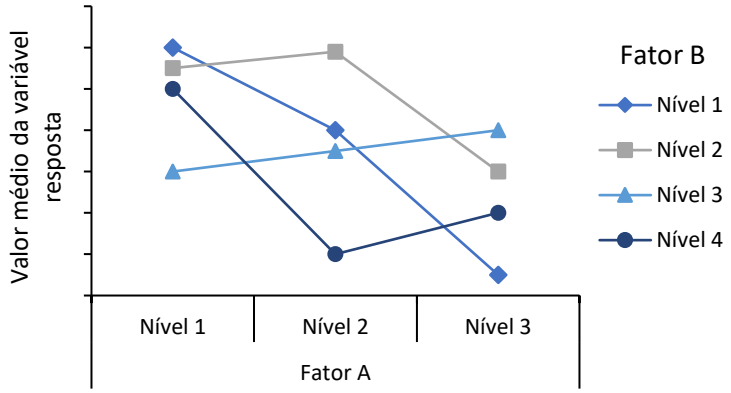


Figura 9.1: *Valor médio da variável resposta por nível do fator A (3 níveis) e por nível do fator B (4 níveis).*

Objetivo: Testar se existe interação entre os dois fatores e se, para cada um dos fatores, a média é igual para todos os seus níveis.

Observações:

- Quando não existem repetições, então não se testa a interação.
- Por simplicidade, apenas se vai estudar o caso em que existe igual número de observações por célula.

Notação:

a	número de níveis do fator A;
b	número de níveis do fator B;
$m = n_{ij}$	número de observações por célula, isto é, no nível i do fator A e no nível j do fator B (número de observações em todas as células é igual);
x_{ijk}	k-ésimo valor observado no nível i do fator A e no nível j do fator B;
$n = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij} = abm$	número total de observações;
$\overline{\overline{X}} = \overline{X}_{...} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m} \frac{X_{ijk}}{n}$	média global;
$\overline{X}_{i..} = \sum_{j=1}^{b}\sum_{k=1}^{m} \frac{X_{ijk}}{bm}$	média do nível i do fator A;
$\overline{X}_{.j.} = \sum_{i=1}^{a}\sum_{k=1}^{m} \frac{X_{ijk}}{am}$	média do nível j do fator B;
$\overline{X}_{ij.} = \sum_{k=1}^{m} \frac{X_{ijk}}{m}$	média do nível i do fator A e do nível j do fator B;
$SQA = bm\sum_{i=1}^{a}\left(\overline{X}_{i..} - \overline{\overline{X}}\right)^2$	soma dos quadrados dos desvios entre os níveis do fator A;

$SQB = am\sum_{j=1}^{b}\left(\overline{X}_{.j.} - \overline{\overline{X}}\right)^2$ soma dos quadrados dos desvios entre os níveis do fator B;

$SQAB = m\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{X}_{ij.} - \overline{X}_{i..} - \overline{X}_{.j.} + \overline{\overline{X}}\right)^2$ soma dos quadrados dos desvios dentro dos grupos;

$$SQE = \begin{cases} \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(X_{ijk} - \overline{X}_{ij.}\right)^2, & \text{se } m > 1 \\ \sum_{i=1}^{a}\sum_{j=1}^{b}\left(X_{ij1} - \overline{X}_{i..} - \overline{X}_{.j.} + \overline{\overline{X}}\right)^2, & \text{se } m = 1 \end{cases}$$ soma dos quadrados dos resíduos;

$SQT = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(X_{ijk} - \overline{\overline{X}}\right)^2$ soma dos quadrados dos desvios totais em torno da média global;

$MQA = \frac{SQA}{a-1}$ média dos quadrados do fator A;

$MQB = \frac{SQB}{b-1}$ média dos quadrados do fator B;

$MQAB = \frac{SQAB}{(a-1)(b-1)}$ média dos quadrados da interação entre os fatores A e B (só se calcula se $m > 1$);

$$MQE = \begin{cases} \frac{SQE}{ab(m-1)}, & \text{se } m > 1 \\ \frac{SQE}{(a-1)(b-1)}, & \text{se } m = 1 \end{cases}$$ média dos quadrados residual;

$MQT = \frac{SQT}{n-1}$ média dos quadrados totais.

Modelo matemático:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \qquad i = 1, \dots, a, \qquad j = 1, \dots, b, \qquad k = 1, \dots, m.$$

onde:

- μ média da população;
- α_i efeito do fator A;
- β_j efeito do fator B;
- γ_{ij} efeito da interação entre o fator A do fator B;
- ε_{ijk} resíduo.

Decomposição da variância:

Variação total	=	Variação explicada pelo fator A	+	Variação explicada pelo fator B	+	Variação explicada pela interação entre o fator A e fator B	+	Variação devida ao erro

i. e.,

$$SQT = SQA + SQB + SQAB + SQE$$

Hipóteses a testar:

- H_0^{AB}: Não existe interação entre o fator A e o fator B *vs*
  H_1^{AB}: Existe interação entre o fator A e o fator B.
- H_0^A: Todos os níveis do fator A têm igual média, i.e., $\mu_{1.} = \mu_{2.} = \cdots = \mu_{a.}$ *vs*
  H_1^A: Nem todos os níveis do fator A têm igual média.
- H_0^B: Todos os níveis do fator B têm igual média, i.e., $\mu_{.1} = \mu_{.2} = \cdots = \mu_{.b}$ *vs*
  H_1^B: Nem todos os níveis do fator B têm igual média.

Estatísticas de teste:

- $F_{AB} = \dfrac{MQAB}{MQE} \sim F_{(a-1)(b-1);\, ab(m-1)}$.
- $F_A = \dfrac{MQA}{MQE} \sim F_{a-1;\, ab(m-1)}$.
- $F_B = \dfrac{MQB}{MQE} \sim F_{b-1;\, ab(m-1)}$.

sendo $F_{m;\, n;\, 1-\alpha}$ o quantil de probabilidade $(1-\alpha)$ da distribuição $F_{m;\, n}$.

Regras de decisão:

- Rejeitar H_0^{AB} se $f_{AB_{obs}} \geq F_{(a-1)(b-1);\, ab(m-1);\, 1-\alpha}$ (Teste unilateral direito).
- Rejeitar H_0^A se $f_{A_{obs}} \geq F_{a-1;\, ab(m-1);\, 1-\alpha}$ (Teste unilateral direito).
- Rejeitar H_0^B se $f_{B_{obs}} \geq F_{b-1;\, ab(m-1);\, 1-\alpha}$ (Teste unilateral direito).

Observações:

- A hipótese de interação só se testa se $m > 1$.
- Quando se verifica que existe interação entre os fatores A e B, não se analisa o efeito isolado de cada um dos fatores, uma vez que o efeito que os dois fatores provocam na variável dependente é diferente do efeito provocado por cada um dos fatores isoladamente. No entanto, é possível averiguar quais as diferenças específicas, realizando testes de comparação múltipla para um fator fixando um nível específico do outro fator.
- Se $m = 1$ então na distribuição F, das estatísticas de teste F_A e F_B e respetivos pontos críticos nas regras de decisão, substituir $ab(m-1)$ por $(a-1)(b-1)$.

Tabela da análise de variância (ANOVA):

Fontes de Variação (F. V.)	Soma dos Quadrados (SQ)	Graus de Liberdade (g. l.)	Média dos Quadrados (MQ)	F
Fator A	SQA	$a-1$	MQA	$F_A = \dfrac{MQA}{MQE}$
Fator B	SQB	$b-1$	MQB	$F_B = \dfrac{MQB}{MQE}$
Interação AB	$SQAB$	$(a-1)(b-1)$	$MQAB$	$F_{AB} = \dfrac{MQAB}{MQE}$
Erro	SQE	$ab(m-1)$	MQE	
Total	SQT	$abm-1$		

Caso particular: Quando existe apenas uma observação por célula $(m = 1)$, não existe a linha da interação na tabela ANOVA e o $g.l._{Erro} = (a-1)(b-1)$.

9.3 Testes de comparação múltipla

A rejeição da hipótese nula do teste F da análise de variância apenas permite concluir a não igualdade entre as médias entre os K grupos. Para analisar quais as médias que diferem significativamente entre si é necessário aplicar um outro tipo de teste que compare cada par de médias.
Existe uma grande variedade de testes de comparação múltipla (Post-hoc). Neste capítulo apenas serão abordados:

- Teste HSD de Tukey;
- Teste de Scheffé.

Observações:

- O teste HSD de Tukey é mais preciso quando as amostras têm igual dimensão.
- O teste de Scheffé permite a utilização de amostras de dimensão diferente e é um método robusto relativamente aos pressupostos de Normalidade e igualdade das variâncias. Este teste tende a ser mais conservativo do que o teste HSD de Tukey.

Hipóteses a testar:

Nestes testes ensaia-se a igualdade entre todos os pares de médias, i. e., para todo o $i \neq j$ e $i, j = 1, 2, \dots, K$:

$$H_0: \mu_i = \mu_j \quad \text{vs} \quad H_1: \mu_i \neq \mu_j.$$

9.3.1 Teste HSD de Tukey

Estatística de teste:

$$W = \frac{\left|\overline{X}_i - \overline{X}_j\right|}{\sqrt{\frac{\text{MQE}}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim q_{K;\, n-K},$$

sendo $q_{K;\, n-K}$ a distribuição "Studentized Range" com K e $(n-K)$ graus de liberdade.

Observação: A distribuição Studentized Range encontra-se tabelada $\left(q_{K;\, n-K;\, 1-\alpha}\right)$ no Anexo E.

Regra de decisão: Rejeitar $H_0: \mu_i = \mu_j$ quando

$$\left|\overline{X}_i - \overline{X}_j\right| \geq q_{K;\, n-K;\, 1-\alpha} \sqrt{\frac{\text{MQE}}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)},$$

sendo $q_{K;\, n-K;\, 1-\alpha}$ o quantil de probabilidade $(1-\alpha)$ da distribuição "Studentized Range".

9.3.2 Teste de Scheffé

Estatística de teste:

$$F = \frac{\left|\overline{X}_i - \overline{X}_j\right|}{\sqrt{\text{MQE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim \sqrt{(k-1)F_{k-1;\, n-K}}$$

Regra de decisão: Rejeitar H_0: $\mu_i = \mu_j$ quando

$$\left|\overline{X}_i - \overline{X}_j\right| \geq \sqrt{(k-1)F_{K-1;\, n-K;\, 1-\alpha}\text{MQE}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)},$$

sendo $F_{K-1;\, n-K;\, 1-\alpha}$ o quantil de probabilidade $(1-\alpha)$ da distribuição $F_{k-1;\, n-K}$.

9.4 Teste à igualdade das K variâncias

Um dos pressupostos da análise de variância é o facto de as K populações serem Normais com igual variância. Uma forma de testar a veracidade da igualdade das variâncias, de distribuições Normais, é através do teste de Bartlett ou do teste de Levene.

Hipóteses a testar:

Nestes testes ensaia-se a igualdade entre todas as variâncias, i. e.,

$$H_0\colon \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2 \quad vs \quad H_1\colon \exists_{i\neq j}\colon \sigma_i^2 \neq \sigma_j^2 \text{ para } i, j\ i = 1, \ldots, K$$

(i.e., nem todas as variâncias σ_i^2 são iguais).

9.4.1 Teste de Bartlett

Estatística de teste:

$$\chi^2 = \frac{(n-K)\,ln(MQE) - \sum_{i=1}^{K}(n_i-1)ln\left(S_i^2\right)}{1+\frac{1}{3(K-1)}\left(\sum_{i=1}^{K}\frac{1}{n_i-1}-\frac{1}{n-K}\right)} \overset{\circ}{\sim} \chi^2_{K-1}.$$

Regra de decisão: Rejeitar H_0 quando

$$\chi^2_{obs} \geq \chi^2_{K-1;\, 1-\alpha}.$$

Observação: Este teste é muito sensível ao pressuposto da Normalidade, pelo que não deve ser utilizado quando existem dúvidas relativamente à Normalidade. Uma vez que a distribuição é apenas assintótica, só se deve aplicar o teste Bartlett quando existirem pelo menos 5 observações por grupo $(n_i \geq 5)$.

9.4.2 Teste de Levene

Este teste consiste em aplicar a análise de variância à uma nova variável Z_{ij}, que corresponde aos desvios absolutos entre a variável em estudo X_{ij} e a média, mediana ou média aparada, do respetivo grupo.

Estatística de teste:

$$F = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i\left(\overline{Z}_i - \overline{\overline{Z}}\right)^2}{\frac{1}{n-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(Z_{ij}-\overline{Z}_i\right)^2} \overset{\circ}{\sim} F_{K-1;\, n-K},$$

onde:

$\overline{\overline{Z}} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\frac{Z_{ij}}{n} = \sum_{i=1}^{K}\frac{n_i\overline{Z}_i}{n}$ média global;

$\overline{Z}_i = \sum_{j=1}^{n_i}\frac{Z_{ij}}{n_i}$ média do grupo i, $i = 1, 2, \ldots, K$.

podendo-se calcular Z_{ij} **por uma** das seguintes definições:

1. $Z_{ij} = \left|X_{ij} - \overline{X}_i\right|$, onde $\overline{X}_i$ é a média do i-ésimo grupo;
2. $Z_{ij} = \left|X_{ij} - \tilde{X}_i\right|$, onde $\tilde{X}_i$ é a mediana do i-ésimo grupo;
3. $Z_{ij} = \left|X_{ij} - \overline{X}_i^*\right|$, onde $\overline{X}_i^*$ é a média aparada do i-ésimo grupo.

Regra de decisão: Rejeitar H_0 quando

$$f_{obs} \geq F_{K-1;\, n-K;\, 1-\alpha},$$

sendo $F_{K-1;\, n-K;\, 1-\alpha}$ o quantil de probabilidade $(1-\alpha)$ da distribuição $F_{k-1;\, n-K}$.

Observações:

- De referir que, no artigo de 1960, Levene propôs apenas o uso da média.
- No caso de grupos com dimensões iguais, este teste possui a vantagem de ser um método robusto relativamente ao pressuposto de Normalidade. Portanto, se existir uma forte evidência de que os dados não provêm de uma distribuição Normal, então deve-se utilizar o teste de Levene em vez do teste de Bartlett.

9.5 Exercícios resolvidos

1. Um determinado departamento governamental está preocupado com os aumentos dos custos verificados no decurso de projetos de I&D (investigação e desenvolvimento) encomendados aos institutos A, B, C e D. Por este motivo, decidiu analisar os custos associados a diferentes projetos, calculando, para cada um deles, a razão entre o custo final incorrido e o custo inicialmente previsto (na altura da adjudicação). Para cada projeto, ambos os custos foram expressos numa base constante. No quadro que se segue apresentam-se os resultados obtidos:

Instituto	Rácio = Custo incorrido/Custo previsto					
A	1,0	0,8	1,9	1,1	2,7	
B	1,7	2,5	3,0	2,2	3,7	1,9
C	1,0	1,3	3,2	1,4	1,3	2,0
D	3,8	2,8	1,9	3,0	2,5	

a) O que pode dizer sobre a verificação do pressuposto de normalidade?
b) Teste a hipótese de homogeneidade das variâncias, ao nível de significância de 5%, utilizando o teste de:
   i. Bartlett
   ii. Levene
c) Construa a tabela ANOVA.
d) Averigue se os quatro institutos têm um comportamento médio global idêntico em relação ao agravamento dos custos, usando α = 1%. Mantém a sua decisão ao nível de significância de 5%?
e) Caso tenha concluído que o comportamento médio não era idêntico nos quatro institutos, no âmbito da alínea anterior, diga quais os institutos que diferem entre si considerando α = 10%, utilizando o teste:
   i. HSD de Tukey.
   ii. de Scheffé.

Resolução:

Seja X_{ij} a v.a. que representa o custo incorrido/custo previsto no instituto i, para o projeto j, com $i = 1(= A), 2\ (= B), 3(= C), 4(= D)$ e $j = 1, \dots, n_i$.

$K = 4$ (n.º de institutos);

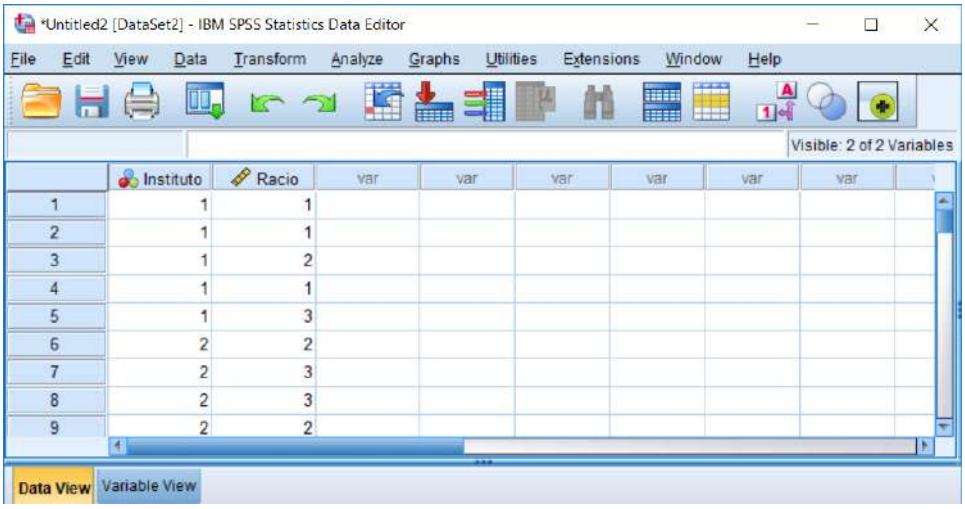
Instituto	A	B	C	D
n_i	5	6	6	5
$\overline{x}_i$	1,5	2,5	1,7	2,8
s_i^2	0,625	0,556	0,648	0,485

$$n = \sum_{i=1}^{4} n_i = 22,$$

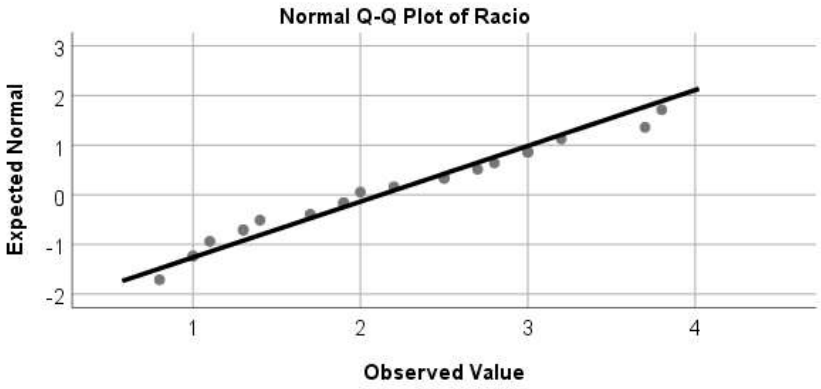
$$\overline{\overline{x}} = \sum_{i=1}^{4}\sum_{j=1}^{n_i} \frac{x_{ij}}{n} = \frac{1{,}0 + 0{,}8 + \cdots + 3{,}0 + 2{,}5}{22} = 2{,}1227.$$

a) A verificação do pressuposto de normalidade pode ser feita recorrendo ao gráfico QQ-plot.

(SPSS)



(SPSS) Analyse → Descriptive Satistics → Explore
(Dependent List: Racio; ⊙ Plots; Plots → ☑ Normality plots with tests)



Uma vez que os pontos de um modo geral se sobrepõem à reta, indicando a existência de uma correspondência direta entre os quantis observados e os quantis teóricos, podemos considerar que o pressuposto da normalidade é verificado.

b) $\alpha = 5\%, \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$?

H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_4^2$ vs H_1: Nem todas as variâncias σ_i^2 são iguais, $i = 1, \ldots, 4$.

i) Teste de Bartlett

Estatística de teste:

$$\chi^2 = \frac{(n-K)\, ln(MQE) - \sum_{i=1}^{K}(n_i - 1) ln\left(S_i^2\right)}{1 + \frac{1}{3(K-1)}\left(\sum_{i=1}^{K}\frac{1}{n_i - 1} - \frac{1}{n-K}\right)} \overset{\circ}{\sim} \chi^2_{K-1}.$$

$$SQE = \sum_{i=1}^{4}(n_i - 1)s_i^2 = 4 \times 0{,}625 + 5 \times 0{,}556 + 5 \times 0{,}648 + 4 \times 0{,}485 = 10{,}46$$

$$MQE = \frac{\text{SQE}}{n-K} = \frac{10{,}460}{22-4} = 0{,}5811$$

$$\chi^2_{obs} = \frac{(22-4)\, ln(0{,}5811) - (4\, ln(0{,}625) + \cdots + 4\, ln(0{,}485))}{1 + \frac{1}{3(4-1)}\left(\left(\frac{1}{4} + \frac{1}{5} + \frac{1}{5} + \frac{1}{4}\right) - \frac{1}{22-4}\right)} = 0{,}1081.$$

Como $\chi^2_{obs} < \chi^2_{3;\,0{,}95} = 7{,}815$ não rejeitar H_0.

Ao nível de significância de 5%, não existe evidência estatística para afirmar que as variâncias dos grupos são diferentes. Pode-se, por isso considerar que é verificado o pressuposto da homocedasticidade (desde que o pressuposto da Normalidade se verifique).

ii) Teste de Levene

Estatística de teste:

$$F = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i \left(\overline{Z}_i - \overline{\overline{Z}}\right)^2}{\frac{1}{n-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(Z_{ij} - \overline{Z}_i\right)^2} \overset{\circ}{\sim} F_{K-1;\, n-K}.$$

Considere-se o caso em que $Z_{ij} = \left|X_{ij} - \overline{X}_i\right|$. No quadro seguinte apresentam-se os valores z_{ij} obtidos.

Instituto	z_{ij}						$\overline{z}_i$
A	0,5	0,7	0,4	0,4	1,2		0,64
B	0,8	0	0,5	0,3	1,2	0,6	0,567
C	0,7	0,4	1,5	0,3	0,4	0,3	0,6
D	1	0	0,9	0,2	0,3		0,48

$$\overline{\overline{z}} = \sum_{i=1}^{4}\sum_{j=1}^{n_i}\frac{z_{ij}}{n} = \frac{0{,}5 + 0{,}7 + \cdots + 0{,}2 + 0{,}6}{22} = 0{,}573$$

$$\sum_{i=1}^{4} n_i \left(\overline{z}_i - \overline{\overline{z}}\right)^2 = 5 \times (0{,}64 - 0{,}573)^2 + 6 \times (0{,}567 - 0{,}573)^2 + 6 \times (0{,}6 - 0{,}573)^2$$
$$+ 5 \times (0{,}48 - 0{,}573)^2 = 0{,}070$$

$$\sum_{i=1}^{4}\sum_{j=1}^{n_i}\left(z_{ij} - \overline{z}_i\right)^2 = (0{,}5 - 0{,}64)^2 + \cdots + (1{,}2 - 0{,}64)^2 + (0{,}8 - 0{,}567)^2 + \cdots + (0{,}6 - 0{,}567)^2$$
$$+(0{,}7 - 0{,}6)^2 + \cdots + (0{,}3 - 0{,}6)^2 + (1 - 0{,}48)^2 + \cdots + (0{,}3 - 0{,}48)^2 = 3{,}173.$$

$$f_{obs} = \frac{\frac{0{,}070}{4-1}}{\frac{3{,}173}{22-4}} = 0{,}133.$$

Como $f_{obs} < F_{3;\,18;\,0{,}95} = 3{,}16$, logo não rejeitar H_0.

Ao nível de significância de 5%, não existe evidência estatística para afirmar que as variâncias dos grupos são diferentes. Pode-se por isso considerar que é verificado o pressuposto da homocedasticidade.

(SPSS) Analyse → Compare Means → One-Way ANOVA
(Dependent variable: Racio; Factor: Instituto; Options → Statistics: ☑ homogeneity of variance test)

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Racio	Based on Mean	,133	3	18	,939
	Based on Median	,030	3	18	,993
	Based on Median and with adjusted df	,030	3	15,393	,993
	Based on trimmed mean	,111	3	18	,953

O valor $p = 0{,}939$, logo não se rejeita H_0 para os níveis de significância inferiores a 93,9%. Assim, pode-se considerar que é verificado o pressuposto da homocedasticidade.

c) $SQE = 10{,}46$ (pela alínea a i);

$$SQA = \sum_{i=1}^{4} n_i \left(\overline{x}_i - \overline{\overline{x}}\right)^2$$
$$= 5 \times (1{,}5 - 2{,}123)^2 + 6 \times (2{,}5 - 2{,}123)^2 + \cdots + 5 \times (2{,}8 - 2{,}123)^2 = 6{,}159;$$

$$SQT = SQA + SQE = 6{,}159 + 10{,}46 = 16{,}619;$$

$$MQA = \frac{SQA}{K-1} = \frac{6{,}159}{3} = 2{,}053;$$

$MQE = 0{,}581$ (pela alínea a i);

$$f_{obs} = \frac{MQA}{MQE} = \frac{2{,}053}{0{,}581} = 3{,}533.$$

Tabela ANOVA

Fonte de variação	SQ	gl	MQ	F
Entre grupos	6,159	3	2,053	3,533
Dentro de grupos	10,460	18	0,581	
Total	16,619	21		

(SPSS) Analyse → Compare Means → One-Way ANOVA
(Dependent variable: Racio; Factor: Instituto)

ANOVA
Racio

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6,159	3	2,053	3,533	,036
Within Groups	10,460	18	,581		
Total	16,619	21			

d) $\alpha = 1\%$ e 5%, $\mu_1 = \mu_2 = \mu_3 = \mu_4$?

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ vs H_1: Nem todas as médias μ_i são iguais, $i = 1, \dots, 4$.

Estatística de teste:

$$F = \frac{MQA}{MQE} \sim F_{K-1;\, n-K}.$$

$f_{obs} = 3{,}533$ (pela alínea b)

Como $f_{obs} < F_{3;\,18;\,0{,}99} = 5{,}09$ não se rejeita H_0 para $\alpha = 1\%$. Mas, rejeita-se para $\alpha = 5\%$ pois $f_{obs} \geq F_{3;\,18;\,0{,}95} = 3{,}16$.

Ao nível de significância de 1%, não existe evidência estatística para afirmar que o comportamento médio global, em relação ao agravamento dos custos, não é idêntico nos quatro institutos.

Ao nível de significância de 5% não mantenho a decisão, pois neste caso rejeita-se a hipótese H_0. Portanto, quando $\alpha = 5\%$, existe evidência estatística para afirmar que o comportamento médio global, em relação ao agravamento dos custos, não é idêntico nos quatro institutos.

Pelo $\text{valor}\, p = 0{,}036$ (apresentado na tabela ANOVA) verifica-se que para níveis de significância inferiores a 3,6% (por exemplo 1%) não se rejeita H_0 e para níveis superiores a 3,6% (exemplo de 5% e 10%) já se rejeita H_0. Portanto, existe evidência de que as médias dos rácios não são todas iguais.

e) $\alpha = 10\%$, $\mu_i = \mu_j$?

O objetivo é comparar todos os pares de médias.
H_0: $\mu_i = \mu_j$ vs H_1: $\mu_i \neq \mu_j$, para todo o $i \neq j$ e $i, j = 1, 2, \dots, 4$.

i) Teste HSD de Tukey
Sabe-se que rejeita-se H_0 quando:

$$|\overline{X}_i - \overline{X}_j| \geq q_{K;\, n-K;\, 1-\alpha} \sqrt{\frac{\text{MQE}}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = q_{4;\,18;\,0{,}90} \sqrt{\frac{0{,}581}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} = 3{,}487 \sqrt{\frac{0{,}581}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}.$$

Ora,

- $|\overline{x}_1 - \overline{x}_2| = |1{,}5 - 2{,}5| = 1 < 3{,}487 \sqrt{\frac{0{,}581}{2}\left(\frac{1}{5} + \frac{1}{6}\right)} = 1{,}138,$ logo não rejeitar H_0: $\mu_1 = \mu_2$.
- $|\overline{x}_1 - \overline{x}_3| = |1{,}5 - 1{,}7| = 0{,}2 < 3{,}487 \sqrt{\frac{0{,}581}{2}\left(\frac{1}{5} + \frac{1}{6}\right)} = 1{,}138,$ logo não rejeitar H_0: $\mu_1 = \mu_3$.
- $|\overline{x}_1 - \overline{x}_4| = |1{,}5 - 2{,}8| = 1{,}3 \geq 3{,}487 \sqrt{\frac{0{,}581}{2}\left(\frac{1}{5} + \frac{1}{5}\right)} = 1{,}189,$ logo rejeitar H_0: $\mu_1 = \mu_4$.
- $|\overline{x}_2 - \overline{x}_3| = |2{,}5 - 1{,}7| = 0{,}8 < 3{,}487 \sqrt{\frac{0{,}581}{2}\left(\frac{1}{6} + \frac{1}{6}\right)} = 1{,}085,$ logo não rejeitar H_0: $\mu_2 = \mu_3$.
- $|\overline{x}_2 - \overline{x}_4| = |2{,}5 - 2{,}8| = 0{,}3 < 3{,}487 \sqrt{\frac{0{,}581}{2}\left(\frac{1}{6} + \frac{1}{5}\right)} = 1{,}138,$ logo não rejeitar H_0: $\mu_2 = \mu_4$.
- $|\overline{x}_3 - \overline{x}_4| = |1{,}7 - 2{,}8| = 1{,}1 < 3{,}487 \sqrt{\frac{0{,}581}{2}\left(\frac{1}{6} + \frac{1}{5}\right)} = 1{,}138,$ logo não rejeitar H_0: $\mu_3 = \mu_4$.

Concluindo, através do teste HSD de Tukey, ao nível de significância de 10%, são detetadas diferenças entre os comportamentos médios globais, em relação ao agravamento dos custos, dos grupos, pois existe diferença significativa entre o comportamento médio global do instituto A e do instituto D. Pode-se assim considerar que existem dois grupos de institutos com comportamentos médios idênticos que são:

- Grupo 1: institutos A, B e C;
- Grupo 2: institutos B, C, D.

(SPSS) Analyse → Compare Means → One-Way ANOVA
(Dependent variable: Racio; Factor: Instituto; Post Hoc → Equal Variances Assumed: ☑ Tukey; Significance level: 0,1)

Multiple Comparisons
Dependent Variable: Racio
Tukey HSD

(I) Instituto	(J) Instituto	Mean Difference (I-J)	Std. Error	Sig.	90% Confidence Interval Lower Bound	90% Confidence Interval Upper Bound
1	2	-1,000	,462	,171	-2,14	,14
	3	-,200	,462	,972	-1,34	,94
	4	-1,300*	,482	,065	-2,49	-,11
2	1	1,000	,462	,171	-,14	2,14
	3	,800	,440	,298	-,29	1,89
	4	-,300	,462	,914	-1,44	,84
3	1	,200	,462	,972	-,94	1,34
	2	-,800	,440	,298	-1,89	,29
	4	-1,100	,462	,116	-2,24	,04
4	1	1,300*	,482	,065	,11	2,49
	2	,300	,462	,914	-,84	1,44
	3	1,100	,462	,116	-,04	2,24

*. The mean difference is significant at the 0.1 level.

Homogeneous Subsets
Racio
Tukey HSD[a,b]

Instituto	N	Subset for alpha = 0.1 1	2
1	5	1,50	
3	6	1,70	1,70
2	6	2,50	2,50
4	5		2,80
Sig.		,171	,116

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 5,455.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Os pares de médias significativamente diferentes encontram-se marcados com o símbolo * (1º quadro), de acordo com o nível de significância estabelecido (10%) como se pode comprovar pelo valor p apresentado. No 2º quadro são apresentados os grupos identificados.

ii) Teste de Scheffé

Sabe-se que rejeita-se H0 quando:

$$\left|\overline{X}_i - \overline{X}_j\right| \geq \sqrt{(k-1)F_{K-1;\,n-K;\,1-\alpha} MQE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = \sqrt{3F_{3;\,18;\,0,90} \times 0,5811\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

$$= \sqrt{3 \times 2,41 \times 0,5811\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = \sqrt{4,2014\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

Ora,

- $|\overline{x}_1 - \overline{x}_2| = |1{,}5 - 2{,}5| = 1 < \sqrt{4{,}2014\left(\frac{1}{5}+\frac{1}{6}\right)} = 1{,}24,$ logo não rejeitar $H_0: \mu_1 = \mu_2$.
- $|\overline{x}_1 - \overline{x}_3| = |1{,}5 - 1{,}7| = 0{,}2 < \sqrt{4{,}2014\left(\frac{1}{5}+\frac{1}{6}\right)} = 1{,}24,$ logo não rejeitar $H_0: \mu_1 = \mu_3$.
- $|\overline{x}_1 - \overline{x}_4| = |1{,}5 - 2{,}8| = 1{,}3 \geq \sqrt{4{,}2014\left(\frac{1}{5}+\frac{1}{5}\right)} = 1{,}3,$ logo rejeitar $H_0: \mu_1 = \mu_4$.
- $|\overline{x}_2 - \overline{x}_3| = |2{,}5 - 1{,}7| = 0{,}8 < \sqrt{4{,}2014\left(\frac{1}{6}+\frac{1}{6}\right)} = 1{,}18,$ logo não rejeitar $H_0: \mu_2 = \mu_3$.
- $|\overline{x}_2 - \overline{x}_4| = |2{,}5 - 2{,}8| = 0{,}3 < \sqrt{4{,}2014\left(\frac{1}{6}+\frac{1}{5}\right)} = 1{,}24,$ logo não rejeitar $H_0: \mu_2 = \mu_4$.
- $|\overline{x}_3 - \overline{x}_4| = |1{,}7 - 2{,}8| = 1{,}1 < \sqrt{4{,}2014\left(\frac{1}{6}+\frac{1}{5}\right)} = 1{,}24,$ logo não rejeitar $H_0: \mu_3 = \mu_4$.

Concluindo, através do teste de Scheffé, ao nível de significância de 10%, são detetadas diferenças entre os comportamentos médios globais, em relação ao agravamento dos custos, dos grupos. Conclui-se que diferença significativa entre o comportamento médio global do instituto A e do instituto D. Pode-se assim considerar que existem dois grupos de institutos com comportamentos médios idênticos que são:

- Grupo 1: institutos A, B e C;
- Grupo 2: institutos B, C, D.

(SPSS) Analyse → Compare Means → One-Way ANOVA
(Dependent variable: Racio; Factor: Instituto;
Post Hoc → Equal Variances Assumed: ☑ Scheffé; Significance level: 0,1)

Multiple Comparisons
Dependent Variable: Racio
Scheffe

(I) Instituto	(J) Instituto	Mean Difference (I-J)	Std. Error	Sig.	90% Confidence Interval Lower Bound	90% Confidence Interval Upper Bound
1	2	-1,000	,462	,233	-2,24	,24
	3	-,200	,462	,979	-1,44	1,04
	4	-1,300*	,482	,099	-2,60	,00
2	1	1,000	,462	,233	-,24	2,24
	3	,800	,440	,374	-,38	1,98
	4	-,300	,462	,934	-1,54	,94
3	1	,200	,462	,979	-1,04	1,44
	2	-,800	,440	,374	-1,98	,38
	4	-1,100	,462	,167	-2,34	,14
4	1	1,300*	,482	,099	,00	2,60
	2	,300	,462	,934	-,94	1,54
	3	1,100	,462	,167	-,14	2,34

*. The mean difference is significant at the 0.1 level.

Homogeneous Subsets
Racio
Scheffe[a,b]

Instituto	N	Subset for alpha = 0.1	
		1	2
1	5	1,50	
3	6	1,70	1,70
2	6	2,50	2,50
4	5		2,80
Sig.		,233	,167

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 5,455.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

De forma análoga ao exercício anterior, no 1º quadro o símbolo * identifica os pares de médias significativamente diferentes ao nível de significância estabelecido (10%). No 2º quadro são apresentados os grupos homogéneos identificados.

2. Considere a seguinte tabela ANOVA com dados que permitem testar a igualdade de K médias populacionais:

Fonte de variação	Soma dos quadrados	Graus de liberdade	Média dos quadrados	F
Entre os grupos	6,159	3		
Dentro dos grupos		20		
Total	16,619			

a) Formule as hipóteses a testar.
b) Calcule o valor da estatística de teste apropriada para testar a hipótese nula.
c) Complete a tabela ANOVA.
d) Para um nível de significância de 5%, qual deverá ser a regra de decisão? E a decisão?
e) Que condições deverão ser verificadas para que a aplicação da análise de variância seja adequada?

Resolução:

$g.l._A = K - 1 = 3 \Rightarrow K = 4.$

a) $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ vs H_1: Nem todas as médias μ_i são iguais, $i = 1, \dots, 4$.

b) $SQT = SQA + SQE = 16{,}619 - 6{,}159 = 10{,}46.$

$$MQE = \frac{SQE}{n-k} = \frac{10{,}46}{20} = 0{,}523;$$

$$MQA = \frac{SQA}{k-1} = \frac{6{,}159}{3} = 2{,}053;$$

$$f_{obs} = \frac{MQA}{MQE} = \frac{2{,}053}{0{,}523} = 3{,}9254.$$

c)

Fonte de variação	Soma dos quadrados	Graus de liberdade	Média dos quadrados	F
Entre os grupos	6,159	3	2,053	3,9254
Dentro dos grupos	10,460	20	0,523	
Total	16,619	23	0,7226	

$g.l._T = n - 1 = g.l._A + g.l._E = 3 + 20 = 23;$

$MQT = \frac{SQT}{g.l._T} = \frac{SQT}{n-1} = \frac{16{,}619}{23} = 0{,}7226.$

d) $\alpha = 5\%$.

Regra de decisão: Rejeitar H_0 quando $F_{obs} \geq F_{K-1;\, n-K;\, 1-\alpha} = F_{3;\, 20;\, 0{,}95} = 3{,}10$.

Como $f_{obs} = 3{,}9254 \geq 3{,}10$, rejeitar H_0.

Ao nível de significância de 5% existe evidência estatística para afirmar que as médias dos grupos não são todas iguais. Existe diferença entre pelo menos duas médias.

e) Pressupostos: Os $K = 4$ grupos devem ser independentes, as populações devem ser Normais com igual variância.

3. Numa escola de formação profissional, para comparar três métodos de ensino, constituíram-se 3 grupos de alunos e de cada um deles escolheu-se 1 aluno ao acaso. Seguidamente observou-se o tempo, em minutos, necessário para a resolução de um certo exercício, tendo-se obtido os seguintes resultados:

	Método de instrução		
	M1	M2	M3
Grupo 1	6	4	7
Grupo 2	5	8	6
Grupo 3	9	10	14

Assuma que são verificados os pressupostos para a aplicação da ANOVA.

a) Construa a tabela ANOVA.
b) Ao nível de significância de 10%, verifique se o tempo médio de resolução do exercício não é idêntico nos três grupos.
c) Teste, ao nível de significância de 1%, se existem diferenças significativas entre os métodos de instrução.

Resolução:

Seja X_{ijk} a v.a. que representa o tempo necessário para o elemento k do grupo i do método de instrução j, resolver um certo exercício, $i = 1, 2, 3$, $j = 1(= M1), 2(= M2), 3(= M3)$ e $k = 1$.

Dois fatores:

- Grupo – Fator A, que tem 3 níveis $(a = 3)$: grupo 1, grupo 2 e grupo 3;
- Método de instrução – Fator B que tem 3 níveis $(b = 3)$: M1, M2 e M3.

$m = 1$ (existe uma só observação por grupo e método de instrução, i. e., não há réplicas).
$n = a \times b \times m = 3 \times 3 \times 1 = 9$.

	Método (j)			
Grupo (i)	1	2	3	$\overline{x}_{i.}$
1	6	4	7	5,667
2	5	8	6	6,333
3	9	10	14	11
$\overline{x}_{.j}$	6,667	7,333	9	$7{,}667 = \overline{\overline{x}}$

a) Tabela ANOVA

Fonte de variação	SQ	gl	MQ	F
Linhas	50,667	2	25,333	6,909
Colunas	8,667	2	4,333	1,182
Erro	14,667	4	3,667	
Total	74	8		

$$SQA = bm\sum_{i=1}^{a}\left(\overline{x}_{i.} - \overline{\overline{x}}\right)^2 = 3 \times 1 \times ((5{,}667\text{-}7{,}667)^2 + (6{,}333\text{-}7{,}667)^2 + (11 - 7{,}667)^2) = 50{,}667$$

$$SQB = am\sum_{j=1}^{b}\left(\overline{x}_{.j} - \overline{\overline{x}}\right)^2 = 3 \times 1 \times ((6{,}667\text{-}7{,}667)^2 + (7{,}333 - 7{,}667)^2 + (9 - 7{,}667)^2) = 8{,}667$$

$$SQE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(x_{ijk} - \overline{x}_{ij}\right)^2 = (6 - 7{,}667)^2 + \cdots . + (14\text{-}7{,}667)^2 = 14{,}667$$

$$SQT = SQA + SQB + SQE = 74$$

$$MQA = \frac{SQA}{a-1} = \frac{50{,}667}{3-1} = 25{,}333$$

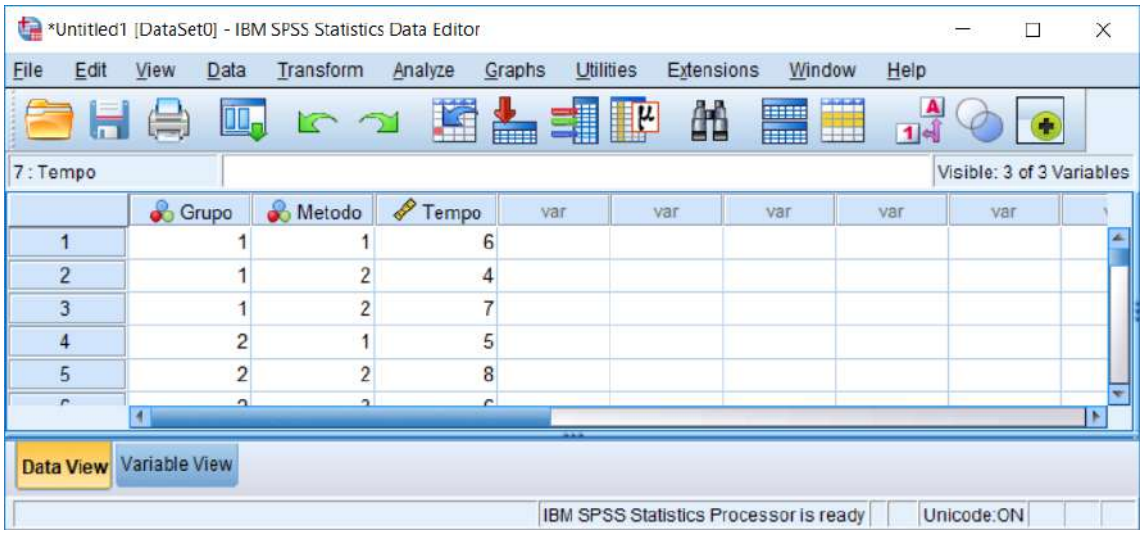
$$MQB = \frac{\text{SQB}}{b-1} = \frac{8{,}667}{3-1} = 4{,}333$$

$$MQE = \frac{SQE}{(a-1)(b-1)} = \frac{14{,}667}{(3-1)(3-1)} = 3{,}667$$

$$f_{A_{obs}} = \frac{MQA}{MQE} = 6{,}909$$

$$f_{B_{obs}} = \frac{MQB}{MQE} = 1{,}182$$

(SPSS)



(SPSS) Analyse → General Linear Model → Univariate
(Dependent variable: Tempo; Fixed Factor(s): Grupo, Método;
Model → Specify Model: ⊙ Build terms; Factors & Covariates: Grupo, Metodo; Build Terms(s): Main effects; Model: Grupo, Metodo; Sum of squares: Type III; ☑ Include intercept in model)

Tests of Between-Subjects Effects

Dependent Variable: Tempo

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	59,333[a]	4	14,833	4,045	,102
Intercept	529,000	1	529,000	144,273	,000
Grupo	50,667	2	25,333	6,909	,050
Método	8,667	2	4,333	1,182	,395
Error	14,667	4	3,667		
Total	603,000	9			
Corrected Total	74,000	8			

a. R Squared = ,802 (Adjusted R Squared = ,604)

Na tabela anterior devolvida pelo SPSS apenas se devem analisar as linhas Grupo, Metodo, Error e Corrected Total.

b) $\alpha = 10\%$, os grupos são idênticos: $\mu_{1.} = \mu_{2.} = \mu_{3.}$?

H_0^A: Todos os grupos têm igual média $(\mu_{1.} = \mu_{2.} = \mu_{3.})$ *vs* H_1^A: Nem todos os grupos têm igual média.

Estatística de teste:

$$F_A = \frac{MQA}{MQE} \sim F_{a-1;(a-1)(b-1)}.$$

Como $f_{A_{obs}} = \frac{MQA}{MQE} = 6{,}909 \geq F_{a-1;(a-1)(b-1);\,1-\alpha} = F_{2;\,4;\,0{,}90} = 4{,}32$ rejeitar H_0.

Portanto, ao nível de significância de 10%, existe evidência estatística para afirmar que o tempo médio de resolução dos exercícios não é igual nos três grupos de alunos.

Uma vez que se concluiu que existe diferença entre os métodos, é necessário aplicar os testes de comparação múltipla para identificar quais os grupos que diferem entre si. Vai apenas ser utilizado o teste HSD de Tukey visto que, quando as amostras têm igual dimensão, este é mais preciso.

Hipóteses a testar: H_0: $\mu_{i.} = \mu_{j.}$ vs H_1: $\mu_{i.} \neq \mu_{j.}$, para todo o $i \neq j$ e $i, j = 1, 2, 3$.

Sabe-se que se rejeita H_0 quando:

$$\left|\overline{X}_i - \overline{X}_j\right| \geq q_{\mathrm{a};(a-1)(b-1);1-\alpha}\sqrt{\frac{\mathrm{MQE}}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = q_{3\,;\,4\,;\,0{,}90}\sqrt{\frac{3{,}667}{2}\left(\frac{1}{3}+\frac{1}{3}\right)} = 3{,}98\sqrt{1{,}105} = 4{,}400$$

Ora,

- $|\overline{x}_{1.} - \overline{x}_{2.}| = |5{,}667 - 6{,}333| = 0{,}667 < 4{,}400$, logo não rejeitar H_0: $\mu_{1.} = \mu_{2.}$.
- $|\overline{x}_{1.} - \overline{x}_{3.}| = |5{,}667 - 11| = 5{,}333 \geq 4{,}400$, logo rejeitar H_0: $\mu_{1.} = \mu_{3.}$.
- $|\overline{x}_{2.} - \overline{x}_{3.}| = |6{,}333 - 11| = 4{,}667 \geq 4{,}400$, logo rejeitar H_0: $\mu_{2.} = \mu_{3.}$.

Através do teste HSD de Tukey, ao nível de significância de 10%, conclui-se que existe evidência estatística para afirmar que existe diferença significativa entre o grupo 3 e os restantes relativamente ao tempo médio necessário para a resolução do exercício. Assim, pode-se considerar que existem dois subgrupos de grupos de alunos com tempo médio de resolução do exercício idêntico que são:

- 1º subgrupo: grupos 1 e 2;
- 2º subgrupo: grupo 3.

(SPSS) Analyse → General Linear Model → Univariate
(Dependent variable: Tempo; Fixed Factor(s): Grupo, Método;
Model → Specify Model: ⊙ Build terms; Factors & Covariates: Grupo, Metodo; Build Terms(s): Main effects; Model: Grupo, Metodo;
Post-Hoc → Factors: Grupo; Post Hoc Tests for: Grupo; Equal Variances Assumed: ☑ Tukey;
Options → Significance level: 0,1)

Grupo
Multiple Comparisons
Dependent Variable: Tempo
Tukey HSD

(I) Grupo	(J) Grupo	Mean Difference (I-J)	Std. Error	Sig.	99% Confidence Interval Lower Bound	99% Confidence Interval Upper Bound
1	2	-,67	1,563	,907	-5,06	3,73
	3	-5,33*	1,563	,057	-9,73	-,94
2	1	,67	1,563	,907	-3,73	5,06
	3	-4,67*	1,563	,085	-9,06	-,27
3	1	5,33*	1,563	,057	,94	9,73
	2	4,67*	1,563	,085	,27	9,06

Based on observed means.
The error term is Mean Square(Error) = 3,667.
*. The mean difference is significant at the ,1 level.

Homogeneous Subsets
Tempo
Tukey HSD[a,b]

Grupo	N	Subset 1	Subset 2
1	3	5,67	
2	3	6,33	
3	3		11,00
Sig.		,907	1,000

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 3,667.
a. Uses Harmonic Mean Sample Size = 3,000.
b. Alpha = ,1.

Tal como nos exercícios anteriores, no 1º quadro o símbolo * identifica os pares de médias significativamente diferentes ao nível de significância estabelecido (10%). No 2º quadro são apresentados os grupos homogéneos identificados.

b) $\alpha = 1\%$, os métodos de instrução são idênticos: $\mu_{.1} = \mu_{.2} = \mu_{.3}$?

H_0^B: Todos os métodos de instrução têm igual média ($\mu_{.1} = \mu_{.2} = \mu_{.3}$) *vs*
H_1^B: Nem todos os métodos de instrução têm igual média.

Estatística de teste:

$$F_B = \frac{MQB}{MQE} \sim F_{b-1;(a-1)(b-1)}.$$

Como $f_{B_{obs}} = \frac{MQB}{MQE} = 1{,}1818 < F_{b-1;(a-1)(b-1);\,1-\alpha} = F_{2;\,4;\,0{,}99} = 18$ não rejeitar H_0.

Portanto, ao nível de significância de 1%, não existe evidência estatística para afirmar que o tempo médio de resolução dos exercícios não é igual segundo os três métodos de ensino.

4. A tabela que se segue apresenta o número de artigos produzidos por 4 operários em dois tipos de máquinas, I e II, ao longo de uma semana.

		Máquina									
		I					II				
Operário	A	15	18	17	20	12	14	16	18	17	15
	B	12	16	14	18	11	11	15	12	16	12
	C	14	17	18	16	13	12	14	16	14	11
	D	19	16	21	23	18	17	15	18	20	17

Análise este caso, exaustivamente, utilizando a técnica de Análise de Variâncias. O que pode concluir ao nível de significância de 5%?

Resolução:

Seja X_{ijk} a v.a. que representa o número de artigos produzidos na k-ésima vez pelo funcionário i na máquina j na k-ésima vez; $i = 1(= A), 2(= B), 3(= C), 4(= D),\ j = 1, 2$ e $k = 1, \dots, m$.

Fatores:

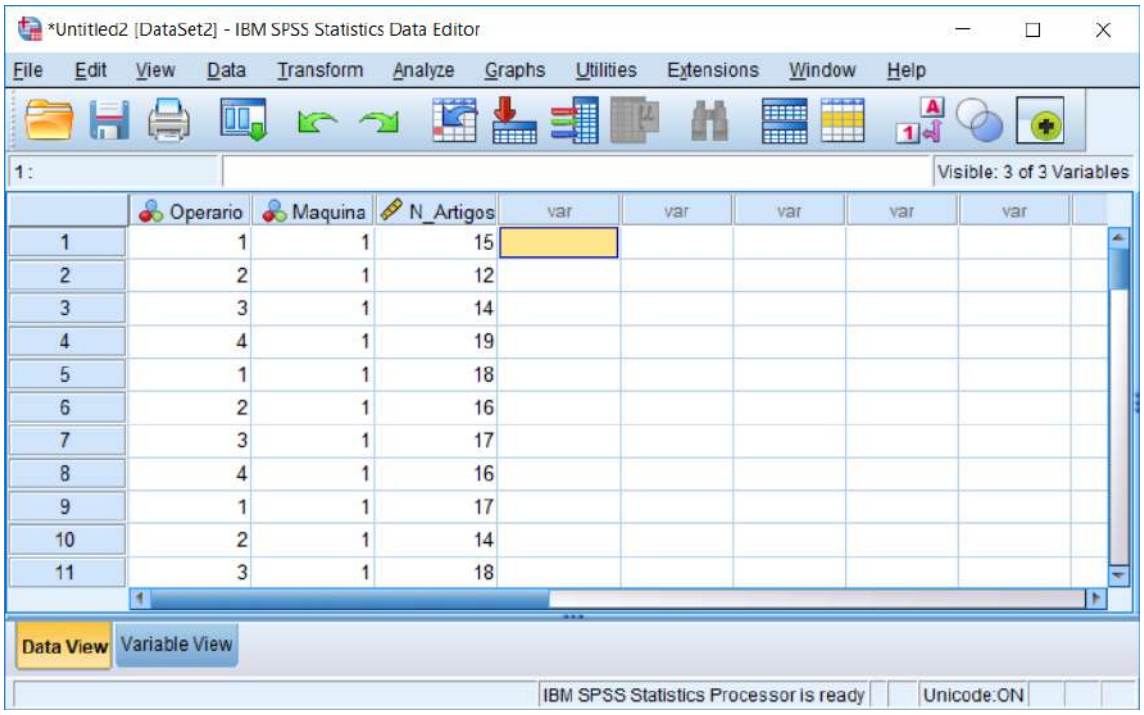
- Máquina – 2 níveis: I e II
- Operário – 4 níveis: A, B, C e D

$m = n_{ij} = 5$, i.e., há igual número de réplicas.

$n = \times\ b \times m = 4 \times 2 \times 5 = 40$.

A primeira etapa consiste na validação dos pressupostos subjacentes: normalidade e igualdade das variâncias. Dada a complexidade dos cálculos, ambos os pressupostos vão ser validados apenas com recurso ao SPSS.

(SPSS)

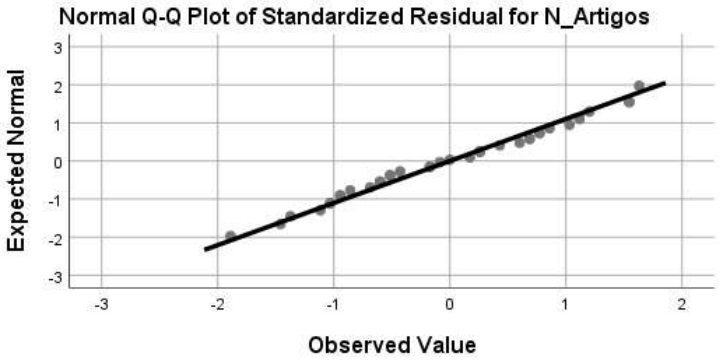


- Normalidade:
  Este pressuposto é avaliado verificando se os resíduos do modelo têm distribuição Normal.

  (SPSS) Analyse → General Linear Model → Univariate
  (Dependent variable: N_Artigos; Fixed Factor(s): Operario, Maquina;
  Model → Specify Model: ⊙ Full factorial; Sum of squares: Type III; ☑ Include intercept in model
  Save → Residuals: ☑ Standardized)

  Na janela de dados é criada uma nova variável (ZRE_1 com descritivo Standartizes Residuals).

  (SPSS) Analyse → Descritive statistics → Explore…
  (Dependent List: Standartized Residuals; ⊙ Plots;
  Plots → Specify Model: ☑ Normality plots with tests)



Pela análise do gráfico, verifica-se que os pontos apresentam o comportamento da reta pelo que se pode considerar que os resíduos têm distribuição Normal.

- Igualdade das variâncias, ou homogeneidade das variâncias ou homocedasticidade:

  H_0: As variâncias são todas iguais vs H_0: Nem todas as variâncias são iguais.

  O pressuposto da igualdade das variâncias (ou homogeneidade das variâncias ou homocedasticidade) vai ser avaliado recorrendo ao teste de Levene, por exemplo.

  (SPSS) Analyse → General Linear Model → Univariate
  (Dependent variable: N_Artigos; Fixed Factor(s): Operario, Maquina;
  Model → Specify Model: ⊙ Full factorial; Sum of squares: Type III; ☑ Include intercept in model;
  Options → ☑ Homogeneity tests; Significance level: 0,05)

Levene's Test of Equality of Error Variances[a,b]

		Levene Statistic	df1	df2	Sig.
N_Artigos	Based on Mean	,686	7	32	,683
	Based on Median	,397	7	32	,897
	Based on Median and with adjusted df	,397	7	26,907	,896
	Based on trimmed mean	,668	7	32	,698

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.
a. Dependent variable: N_Artigos
b. Design: Intercept + Operario + Maquina + Operario * Maquina

Com base na informação dada na linha da média, $f_{obs} = 0{,}686$ e valor $p = 0{,}683 > \alpha = 0{,}05$. Portanto, não rejeitar a hipótese de homogeneidade das variâncias.

Validados os pressupostos, estamos em condições de avançar para a análise de variância, sendo as hipóteses a testar ($\alpha = 5\%$):

- H_0^{AB}: Não existe interação entre o fator Máquina e o fator Operário *vs*
  H_1^{AB}: Existe interação entre o fator Máquina e o fator Operário;
- H_0^{A}: Todos os níveis do fator Máquina têm igual média *vs*
  H_1^{A}: Nem todos os níveis do fator Máquina têm igual média;
- H_0^{B}: Todos os níveis do fator Operário têm igual média *vs*
  H_1^{B}: Nem todos os níveis do fator Operário têm igual média.

Começa-se por verificar se existe interação. Se existir, o exercício acaba aqui, se não existir então é necessário testar as outras duas hipóteses.

	Máquina		
$\overline{x}_{ij}$	I	II	$\overline{x}_{i.}$
A	16,4	16	16,2
B	14,2	13,2	13,7
C	15,6	13,4	14,5
D	19,4	17,4	18,4
$\overline{x}_{.j}$	16,4	15	$15{,}7 = \overline{\overline{x}}$

- Efeito interação entre o fator Operário e o fator Máquina:

  H_0^{AB}: Não existe interação entre o fator Máquina e o fator Operário vs
  H_1^{AB}: Existe interação entre o fator Máquina e o fator Operário;

  Estatística de teste:

$$F_{AB} = \frac{MQAB}{MQE} \sim F_{(a-1)(b-1);\, ab(m-1)}.$$

$$MQAB = \frac{SQAB}{(a-1)(b-1)} = \frac{5{,}4}{(4-1)(2-1)} = 1{,}8;$$

$$SQAB = m\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{x}_{ij.} - \overline{x}_{i..} - \overline{x}_{.j.} + \overline{\overline{x}}\right)^2$$
$$= 5 \times ((16{,}4 - 16{,}2 - 16{,}4 + 15{,}7)^2 + (16 - 16{,}2 - 15 + 15{,}7)^2 + \cdots$$
$$+ (19{,}4 - 16{,}2 - 16{,}4 + 15{,}7)^2 + (17{,}4 - 16{,}2 - 15 + 15{,}7)^2) = 5{,}4;$$

$$MQE = \frac{SQE}{ab(m-1)} = \frac{173{,}6}{4 \times 2 \times (5-1)} = 5{,}425;$$

$$SQE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(X_{ijk} - \overline{x}_{ij.}\right)^2$$
$$= (15 - 16{,}4)^2 + \cdots + (12 - 16{,}4)^2$$
$$+(14 - 16)^2 + \cdots + (15 - 16)^2$$
$$+(12 - 14{,}2)^2 + \cdots + (11 - 14{,}2)^2$$
$$+(11 - 13{,}2)^2 + \cdots + (12 - 13{,}2)^2$$
$$+\cdots$$
$$+(14 - 15{,}6)^2 + \cdots + (13 - 15{,}6)^2$$
$$+(12 - 13{,}4)^2 + \cdots + (11 - 13{,}4)^2$$
$$+(19 - 19{,}4)^2 + \cdots + (18 - 19{,}4)^2$$
$$+(17 - 17{,}4)^2 + \cdots + (17 - 17{,}4)^2$$
$$= 173{,}6;$$

$f_{AB_{obs}} = \frac{MQAB}{MQE} = 0{,}332 < F_{(a-1)(b-1);\, ab(m-1);\, 1-\alpha} = F_{3;\, 32;\, 0{,}95} = 2{,}90$ logo não rejeitar H_0^{AB}.

Portanto, ao nível de significância de 5%, não existe evidência estatística para afirmar que existe interação entre os dois fatores.

Visto que não existe interação, a etapa seguinte é testar o efeito de cada um dos fatores isoladamente.

(SPSS) Analyse → General Linear Model → Univariate
(Dependent variable: N_Artigos; Fixed Factor(s): Operario, Maquina;
Model → Specify Model: ⊙ Full factorial; Sum of squares: Type III; ☑ Include intercept in model)

Tests of Between-Subjects Effects

Dependent Variable: N_Artigos

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	154,800[a]	7	22,114	4,076	,003
Intercept	9859,600	1	9859,600	1817,438	,000
Operario	129,800	3	43,267	7,975	,000
Maquina	19,600	1	19,600	3,613	,066
Operario * Maquina	5,400	3	1,800	,332	,802
Error	173,600	32	5,425		
Total	10188,000	40			
Corrected Total	328,400	39			

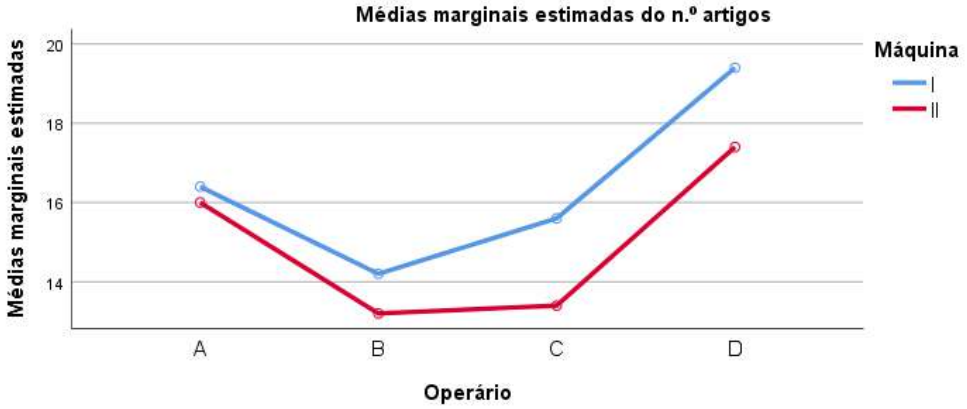
a. R Squared = ,471 (Adjusted R Squared = ,356)

Para o objetivo em causa, da tabela anterior devolvida pelo SPSS apenas se devem analisar as linhas:

- Operario – para testar o efeito do fator Operário,
- Maquina – para testar o efeito do fator Máquina,
- Operario * Maquina – para testar se existe interação entre os fatores Operário e Máquina,
- Error,
- Corrected Total.

A interação entre os fatores A e B pode ser apresentada recorrendo a um gráfico de linhas com a representação das médias $\overline{x}_{ij}$.

(SPSS) Analyse → General Linear Model → Univariate
(Dependent variable: N_Artigos; Fixed Factor(s): Operario, Maquina;
Model → Specify Model: ⊙ Full factorial; Sum of squares: Type III; ☑ Include intercept in model
Plots → Horizontal Axis: Operario; Separate Lines: Maquina; Add; Chart Type: ⊙ Line Chart)



Quando não existe interação as linhas devem apresentar um comportamento semelhante.

- Efeito do fator Operário:

H_0^A: Todos os Operário têm produção média igual vs
H_1^A: Nem todos os Operário têm produção média igual.

Estatística de teste:

$$F_A = \frac{MQA}{MQE} \sim F_{a-1;\, ab(m-1)}.$$

$$SQA = bm\sum_{i=1}^{a}\left(\overline{x}_{i.} - \overline{\overline{x}}\right)^2 = 2 \times 5((16{,}2 - 5{,}7)^2 + (13{,}7 - 15{,}7)^2 + \cdots + (18{,}4 - 5{,}7)^2) = 129{,}8;$$

$$MQA = \frac{SQA}{a-1} = \frac{129{,}8}{4-1} = 43{,}267;$$

$MQA = 5{,}425.$

Como $f_{A_{obs}} = \frac{MQA}{MQE} = 7{,}945 \geq F_{a-1;\, ab(m-1);\, 1-\alpha} = F_{3;\, 32;\, 0{,}95} = 2{,}9$ rejeitar H_0^A.

Ao nível de significância de 5%, existe evidência estatística para afirmar que nem todos os operários têm, em média, a mesma produção. Portanto, é necessário aplicar os testes de comparação múltipla para identificar quais os operários que diferem entre si. Como os n_i são todos iguais, vai ser utilizado o teste HSD de Tukey.

Hipóteses a testar: $H_0: \mu_{i.} = \mu_{j.}$ vs $H_1: \mu_{i.} \neq \mu_{j.}$, para todo o $i \neq j$ e $i, j = 1, 2, 3, 4$.

Sabe-se que se rejeita H_0 quando:

$$\left|\overline{X}_i - \overline{X}_j\right| \geq q_{a;ab(m-1);1-\alpha}\sqrt{\frac{MQE}{2}\left(\frac{1}{n_{i.}} + \frac{1}{n_{j.}}\right)} = q_{4;\, 32;\, 0{,}95}\sqrt{\frac{5{,}425}{2}\left(\frac{1}{10} + \frac{1}{10}\right)} = 3{,}83\sqrt{0{,}5425} = 2{,}821.$$

Ora,

- $|\overline{x}_{1.} - \overline{x}_{2.}| = |16{,}2 - 13{,}7| = 2{,}5 < 2{,}821$, logo não rejeitar $H_0: \mu_{1.} = \mu_{2.}$.
- $|\overline{x}_{1.} - \overline{x}_{3.}| = |16{,}2 - 14{,}5| = 1{,}7 < 2{,}821$, logo não rejeitar $H_0: \mu_{1.} = \mu_{3.}$.
- $|\overline{x}_{1.} - \overline{x}_{4.}| = |16{,}2 - 18{,}4| = 2{,}2 < 2{,}821$, logo não rejeitar $H_0: \mu_{1.} = \mu_{4.}$.
- $|\overline{x}_{2.} - \overline{x}_{3.}| = |13{,}7 - 14{,}5| = 0{,}8 < 2{,}821$, logo não rejeitar $H_0: \mu_{2.} = \mu_{3.}$.
- $|\overline{x}_{2.} - \overline{x}_{4.}| = |13{,}7 - 18{,}4| = 4{,}7 \geq 2{,}821$, logo rejeitar $H_0: \mu_{2.} = \mu_{4.}$.
- $|\overline{x}_{3.} - \overline{x}_{4.}| = |14{,}5 - 18{,}4| = 3{,}9 \geq 2{,}821$, logo rejeitar $H_0: \mu_{3.} = \mu_{4.}$.

Através do teste HSD de Tukey, ao nível de significância de 5%, conclui-se que existe evidência estatística para afirmar que existe diferença significativa no número médio de artigos produzidos pelo operário 4 e os operários 2 e 3. Assim, pode-se considerar que existem dois subgrupos de operários que são:

- 1º subgrupo (menor produção): Operários B, C e A;
- 2º subgrupo (maior produção): Operários A e D.

(SPSS) Analyse → General Linear Model → Univariate
(Dependent variable: N_Artigos; Fixed Factor(s): Operario, Maquina;
Model → Specify Model: ⊙ Full factorial; Sum of squares: Type III; ☑ Include intercept in model;
Post Hoc → Post Hoc Tests for: Operario; Equal Variances Assumed: ☑ Tukey; Options →
Significance level: 0,05)

Post Hoc Tests
Operario
Multiple Comparisons
Dependent Variable: N_Artigos
Tukey HSD

(I) Operario	(J) Operario	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
A	B	2,50	1,042	,097	-,32	5,32
	C	1,70	1,042	,376	-1,12	4,52
	D	-2,20	1,042	,171	-5,02	,62
B	A	-2,50	1,042	,097	-5,32	,32
	C	-,80	1,042	,868	-3,62	2,02
	D	-4,70*	1,042	,000	-7,52	-1,88
C	A	-1,70	1,042	,376	-4,52	1,12
	B	,80	1,042	,868	-2,02	3,62
	D	-3,90*	1,042	,004	-6,72	-1,08
D	A	2,20	1,042	,171	-,62	5,02
	B	4,70*	1,042	,000	1,88	7,52
	C	3,90*	1,042	,004	1,08	6,72

Based on observed means.
The error term is Mean Square(Error) = 5,425.
*. The mean difference is significant at the ,05 level.

Homogeneous Subsets

N_Artigos

Tukey HSD[a,b]

Operario	N	Subset 1	Subset 2
B	10	13,70	
C	10	14,50	
A	10	16,20	16,20
D	10		18,40
Sig.		,097	,171

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 5,425.
a. Uses Harmonic Mean Sample Size = 10,000.
b. Alpha = ,05.

Tal como nos exercícios anteriores, no 1º quadro estão representadas com * os pares de médias estatisticamente diferentes, ao nível de significância estabelecido (5%), e no 2º quadro os grupos homogéneos identificados.

- Efeito do fator Máquina:

  H_0^B: Todas as máquinas têm produção média igual vs
  H_1^B: Nem todas as máquinas têm produção média igual.

  Estatística de teste:

$$F_B = \frac{MQB}{MQE} \sim F_{b-1;\, ab(m-1)}.$$

$MQB = \frac{SQB}{b-1} = \frac{19{,}6}{2-1} = 19{,}6;$

$$SQB = am\sum_{j=1}^{b}\left(\overline{x}_{.j} - \overline{\overline{x}}\right)^2 = 4 \times 5 \times ((16{,}4 - 15{,}7)^2 + (15 - 15{,}7)^2) = 19{,}6;$$

$MQE = 5{,}425.$

Como $f_{B_{obs}} = \frac{MQB}{MQE} = 3{,}613 < F_{b-1;\, ab(m-1);\, 1-\alpha} = F_{1;\, 32;\, 0{,}95} = 4{,}15$, não rejeitar H_0^B.

Ao nível de significância de 5%, não existe evidência estatística para afirmar que a produção média por máquina seja diferente.

9.6 Exercícios propostos

1. De forma a controlar as quantidades calóricas das gorduras contidas nas refeições dos refeitórios de 4 escolas primárias, analisaram-se 6 refeições em cada uma das escolas, tendo-se observado o seguinte:

Escolas	Quantidades calóricas (Kj)					
1	145	140	119	138	148	143
2	135	131	129	143	145	130
3	127	117	123	130	131	126
4	146	151	138	154	145	152

Resolva as seguintes alíneas utilizando o programa SPSS:

a) Teste a hipótese de homogeneidade das variâncias, ao nível de significância de 5%.
b) Construa a tabela ANOVA.

c) Averigue se existe diferença entre as quantidades médias calóricas nas refeições das quatro escolas, usando $\alpha = 5\%$.
d) Mantém a sua decisão ao nível de significância de 1%?
e) Caso tenha concluído que a quantidade calórica não era idêntica nas quatro escolas no âmbito da alínea c), diga quais as escolas que diferem entre si considerando $\alpha = 5\%$.
f) Com base nos resultados da alínea anterior, qual a ou as escolas que oferecem refeições com quantidades médias calóricas significativamente menores? E maiores?
g) Qual a sua resposta se alterasse o nível de significância do teste da alínea e) para 10%?

2. Para verificar se a idade média de 1ª consulta de planeamento familiar era a mesma em 4 cidades distintas (A, B, C, D), recolheu-se uma amostra de utentes para cada uma das cidades tendo-se obtido os seguintes resultados:

Cidade	A	B	C	D
n_i	6	8	7	7
$\overline{x}_i$	19	18	22	21
s_i^2	1,3	1,7	4,3	2,3

Obteve-se também $SQT = 131{,}5$ e $MQT = 24{,}3$.

a) Ao nível de significância de 5% teste a hipótese pretendida.
b) Construa a tabela ANOVA, justificando todos os resultados.
c) Que condições deverão ser verificadas para que a aplicação da análise de variância seja adequada?

3. A fábrica *ReciclaPapel* produz sacos para hipermercados. O departamento técnico suspeita que a concentração de madeira de carvalho na polpa tem efeito sobre a resistência do papel. Para tal levou a cabo uma experiência aleatória considerando 4 níveis relevantes para o fator concentração tendo obtido os seguintes resultados (adaptado de Murteira *et al.*, 2007):

Concentração de carvalho (%)	Resistência do papel (medida em libras por polegada quadrada)					
5	7	8	15	11	9	10
10	12	17	13	18	19	15
15	14	18	19	17	16	18
20	19	25	22	23	18	20

Resolva as seguintes alíneas:

a) Teste a hipótese de homogeneidade das variâncias, ao nível de significância de 5%.
b) Construa a tabela ANOVA.
c) Averigue se existe diferença entre as resistências médias do papel para os diferentes de concentração, usando $\alpha = 1\%$.
d) Mantém a sua decisão ao nível de significância de 5%?
e) Caso tenha concluído que a resistência média não era idêntica nos 4 níveis de concentração, no âmbito da alínea c), diga quais os níveis que diferem entre si considerando $\alpha = 5\%$.
f) Com base nos resultados da alínea anterior, qual a concentração que origina uma resistência média do papel significativamente maior? E menor?
g) Qual a sua resposta se alterasse o nível de significância do teste da alínea e) para 10%?

4. Na disciplina de estatística utilizaram-se três métodos de ensino diferentes, com o objetivo de verificar qual deles originava melhores resultados. Observaram-se as classificações obtidas, no final do semestre, por três grupos de alunos selecionados ao acaso $(n_1 = 6, n_2 = 8, n_3 = a)$ submetidos a cada um dos métodos. Seguidamente realizou-se uma ANOVA, tendo-se obtido os seguintes resultados:

ANOVA

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	Estatística de Teste
Fator				
Erro	62,000	20		
Total	126,615			

a) Complete a tabela ANOVA, justificando todos os resultados.
b) Qual a dimensão da terceira amostra (n_3)?
c) Teste a hipótese de as classificações médias serem idênticas nos 3 grupos. ($\alpha = 5\%$)

5. A *Recicla* é uma empresa que recolhe o lixo urbano duma determinada região. O lixo recolhido é separado para reciclagem segundo o seu tipo: metal, papel, plástico e vidro. Dado que o equipamento a utilizar na reciclagem depende da quantidade de cada tipo de lixo, decidiu-se recolher várias amostras independentes dos vários tipos de lixo, tendo-se obtido os seguintes resultados:

	Metal	Papel	Plástico	Vidro
n_i	23	12	15	?
$\overline{x_i}$	2,2	9,4	1,9	3,8
s_i^2	1,4	2,2	1,1	1,8

E o seguinte quadro incompleto:

ANOVA

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	Estatística de Teste
Fator				
Erro				
Total	644,7136	57		

a) Construa a tabela ANOVA, justificando todos os resultados.
b) Qual a dimensão amostra de vidros recolhida (n_4)?
c) Quais as hipóteses a testar?
d) Teste a hipótese de a quantidade média de lixo ser igual nos 4 tipos. (use $\alpha = 5\%$.)
e) Quais os pressupostos da análise de variância? Foram todos validados antes da realização da alínea anterior? Se não, explique como o(s) poderia validar.

6. Uma empresa de audiometria pretende averiguar se existe diferença na duração, em segundos, dos anúncios de TV que passam nos intervalos da manhã, da tarde e da noite. Para o efeito recolheu uma amostra aleatória de 6 anúncios em cada um dos períodos em estudo, tendo observado os seguintes valores:

Manhã	30	25	30	60	90	60
Tarde	30	30	45	60	30	15
Noite	15	30	15	10	30	5

a) Quais os pressupostos a validar para poder aplicar a análise de variância?
b) Admita que os pressupostos enunciados na alínea anterior são verificados:
   i. Construa a tabela ANOVA.
   ii. Averigue se existe diferença entre as durações médias dos anúncios nos 3 períodos (considere $\alpha = 5\%$).
   iii. Caso tenha concluído que a duração média não era idêntica nos 3 períodos indique em qual dos períodos os anúncios têm menor duração ($\alpha = 5\%$).

7. A tabela que se segue apresenta as produções de 4 variedades de centeio cultivadas em lotes com três tipos diferentes de fertilizantes.

	Variedade de centeio			
Tipo de fertilizante	I	II	III	IV
A	4,5	6,4	7,2	6,7
B	8,8	7,8	9,6	7,0
C	5,9	6,8	5,7	5,2

Ao nível de significância de 1% existe diferença na produção:
a) Devida aos fertilizantes?
b) Devida às variedades de centeio?

8. Para comparar os três tipos de jogo praticados pelas equipas de futebol (ofensivo, defensivo ou misto), constituíram-se 4 grupos de equipas de futebol, da 1ª Liga, e de cada um deles escolheu-se ao acaso 1 equipa que praticasse cada um dos tipos de jogo. Seguidamente observou-se o n.º de jogos perdidos durante o campeonato, tendo-se obtido os seguintes resultados:

	Tipo de jogo		
Grupo	Ofensivo	Misto	Defensivo
A	3	2	1
B	4	3	2
C	5	4	3
D	2	3	2

Obteve-se também $SQ_{Grupo} = 5{,}667$, $SQE = 7{,}833$ e $MQ_{Tipo} = 0{,}167$. Admita que a Normalidade dos dados é verificada.
a) Construa a tabela ANOVA, justificando todos os resultados.
b) Teste, ao nível de significância de 5%, se existem diferenças significativas entre os tipos de jogo praticados.
c) Ao nível de significância de 5%, verifique se o n.º de jogos perdidos não é idêntico nos quatro grupos.

9. O diretor de uma revista de automóveis solicitou a 3 jornalistas (A, B e C) que efetuassem testes de consumo em estrada (E) e em cidade (C) de um determinado modelo, pedindo-lhes que conduzissem "normalmente". Ao formular o pedido desta forma o diretor procurava não só testar o modelo em causa como também caracterizar o comportamento dos jornalistas. Os resultados obtidos, em litros/100 kms, foram os seguintes:

Jornalista	Percurso			
	E		C	
A	6,6	6,2	9,1	10,2
	8,0	7,6	7,8	8,9
B	7,8	9,2	12,6	10,9
	8,5	8,9	11,6	13,3
C	10,1	9,7	12,6	10,7
	10,9	8,9	8,4	10,3

a) Quais as hipóteses a testar?
b) Teste as hipóteses anteriores, ao nível de significância de 5%.
c) Elabore um relatório sucinto com as conclusões a que chegou.

10. Num determinado curso, os alunos têm que fazer num ano letivo 5 disciplinas, distribuídas por 2 semestres.
Para investigar os fatores que influenciam os resultados finais, realizou-se uma experiência com 3 métodos de avaliação (avaliação contínua, frequência e exame final) e obteve-se a seguinte tabela de análise de variância:
ANOVA

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	Estatística de Teste
Semestre	56,88	1		4,67
Método				4,07
Interação	6,79			
Erro				
Total	308,74	17		

a) Qual o número de estudantes que participaram na experiência?
b) Complete a tabela.
c) Que condições tiveram de ser verificadas, para se poder aplicar os métodos de análise de variância a estes dados?
d) Que pode concluir quanto aos factos que mais influenciam os resultados ($\alpha = 1\%$)?

10 Testes não paramétricos

Os testes paramétricos são mais potentes do que os testes não paramétricos, ou seja, têm uma maior capacidade para detetar as diferenças realmente existentes. Portanto, os testes não paramétricos devem ser utilizados apenas como alternativa aos testes paramétricos (Tabela 10.1) quando:

- Não são satisfeitas as condições de aplicabilidade destes últimos, ou
- Quando as variáveis são do tipo ordinal.

Tabela 10.1: *Comparação dos testes paramétricos com os testes não paramétricos.*

Aplicação		Teste paramétrico	Teste não paramétrico
Localização	Uma amostra	Teste Z ou t para μ	Teste dos Sinais Teste de Wilcoxon
	Duas amostras emparelhadas	Teste Z ou t para μ_D	Teste dos Sinais Teste de Wilcoxon
	Duas amostras independentes	Teste Z ou t para $\mu_1 - \mu_2$	Teste Mann-Whitney-Wilcoxon
	3 ou mais amostras independentes	Análise de Variância (Teste F)	Teste de Kruskal-Wallis
Ajustamento	1 amostra		Teste de Ajustamento do Qui-quadrado Teste de Kolmogorov-Smirnov Teste de Shapiro-Wilk
Associação	2 amostras emparelhadas	Teste da correlação linear de Pearson	Teste da correlação ordinal de Spearman Teste de independência do Qui-quadrado
Homogeneidade			Teste de homogeneidade do Qui-quadrado

10.1 Testes de ajustamento

Existem diversos testes de ajustamento, apresentando-se neste livro apenas os testes do Qui-quadrado (o mais genérico e com menos restrições), de Kolmogorov-Smirnov e Shapiro-Wilk.
Os testes não paramétricos do Qui-quadrado (ajustamento e independência) são habitualmente utilizados na análise de variáveis agrupadas ou classificadas, ou seja, variáveis que dividem as observações em 2 ou mais categorias classes.

O teste de Kolmogorov-Smirnov é mais adequado para distribuições completamente especificadas, sendo também dos mais disponibilizados nos programas estatísticos.
O caso específico do teste de Shapiro-Wilk só pode ser utilizado para testar se os dados são provenientes de uma distribuição Normal.

Objetivo: Testar a "bondade" do ajustamento, isto é, saber até que ponto um conjunto de observações suporta a hipótese de ser uma amostra aleatória de uma população com uma determinada distribuição (a própria família é posta em causa).

10.1.1 Teste de ajustamento Qui-quadrado

O **teste ajustamento Qui-quadrado** pode ser aplicado a qualquer tipo de dados, embora na aplicação do teste os dados quantitativos sejam agrupados em classes, ou seja, sejam transformados em categóricos (qualitativos).

Ideia base:

- Construir K classes de valores de X: $A_1, A_2, \dots, A_K$.
- Recolher uma amostra aleatória e determinar as frequências absolutas simples observadas, O_i, para cada classe A_i, ou seja, o número de elementos da amostra que pertencem a A_i.
- Calcular a probabilidade, p_i^*, com base na distribuição teórica definida em H_0, de cada classe A_i conter elementos.
- Determinar as frequências absolutas estimadas, E_i, para cada classe A_i.
- Se a distribuição definida em H_0 se ajustar aos dados então as frequências observadas estarão próximas das estimadas.

Notação:

n	dimensão da amostra;
K	número de classes;
O_i	frequência observada para a classe A_i;
$p_i^* = P(X \in A_i \mid H_0)$	probabilidade dum elemento pertencer à classe A_i, sob H_0;
$E_i = np_i^*$	frequência estimada para a classe A_i, sob a hipótese H_0;
p	número de parâmetros estimados (os parâmetros são estimados recorrendo às suas estimativas obtidas com os dados da amostra).

Observação: Para que seja possível aplicar este teste devem verificar-se as seguintes condições:

1. Não mais de 20% das classes com $E_i < 5$;
2. Todas as classes com $E_i \geq 1$.

Quando estas regras não são cumpridas, pode-se proceder à agregação dessas classes com as adjacentes.

Hipóteses a testar:

H_0: X tem função (densidade) de probabilidade $f_0(x)$ *vs*
H_1: X não tem função (densidade) de probabilidade $f_0(x)$.

Estatística de teste:

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-p-1}.$$

Regra de decisão: Rejeitar H_0 se

$$\chi^2_{obs} \geq \chi^2_{K-p-1;\,1-\alpha} \text{ (Teste unilateral direito).}$$

Portanto, $R.A.$: $\left[0;\ \chi^2_{K-p-1;\,1-\alpha}\right[$ e $R.R.$: $\left[\chi^2_{K-p-1;\,1-\alpha}; +\infty\right[$.

Cálculo do valor p:

$$\text{valor } p = P\left(\chi^2 \geq \chi^2_{obs}\right) = 1 - P\left(\chi^2 < \chi^2_{obs}\right).$$

10.1.2 Teste de Kolmogorov-Smirnov

O **teste de Kolmogorov-Smirnov** pode ser aplicado a dados que estejam pelo menos na escala ordinal, i. e., exceto nominais, e a distribuição a testar deve estar completamente especificada.

Ideia base:

- Construir a função de distribuição empírica $F_n(x)$, i. e., calcular a frequência relativa acumulada para cada valor x observado na amostra.
- Para cada valor x observado na amostra, determinar o valor da função de distribuição $F_0(x)$ postulada em H_0.
- Se a distribuição definida em H_0 se ajustar aos dados, então a função de distribuição
- empírica (observada) deverá estar próxima da função de distribuição esperada.

Hipóteses a testar:

H_0: X tem função de distribuição $F_0(x)$ *vs*
H_1: X não tem função de distribuição $F_0(x)$.

Estatística de teste:
Maior diferença, em valor absoluto, registada entre a função de distribuição empírica e a função de distribuição definida em H_0, i. e.

$$D = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|.$$

Observação: A distribuição da estatística de teste de Kolmogorov-Smirnov encontra-se tabelada.

Regra de decisão: Rejeitar H_0 se

$$d_{obs} \geq d_{n;\,1-\alpha},$$

onde $d_{n;\,1-\alpha}$ é o quantil de probabilidade $(1-\alpha)$, obtido através da tabela apresentada no Anexo F om os pontos críticos para a estatística de Kolmogorov-Smirnov.

Portanto, $R.A.$: $\left[0;\ d_{n;\,1-\alpha}\right[$ e $R.R.$: $\left[d_{n;\,1-\alpha}; +\infty\right[$.

Observações:

- No caso de ajustamento a uma distribuição Normal de parâmetros desconhecidos, o teste de Kolmogorov-Smirnov deve incorporar a correção de Lilliefors (ver tabela com pontos críticos em Sheskin (2011)).
- Este teste deve ser utilizado em amostras grandes.

10.1.3 Teste de Shapiro-Wilk

O **teste de Shapiro-Wilk** é utilizado apenas para avaliar se os dados quantitativos (não agrupados em classes) se ajustam a uma distribuição Normal.

Hipóteses a testar:

H_0: X tem distribuição Normal *vs* H_1: X não tem distribuição Normal.

Estatística de teste:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{i:n})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

onde $x_{i:n}$ corresponde ao i-ésimo valor na amostra ordenada e os a_i são constantes geradas pelas médias, variâncias e covariâncias das estatísticas de ordem de uma amostra de tamanho n de uma distribuição Normal.

Regra de decisão: Rejeitar H_0 se valor $p \leq \alpha$.

10.2 Testes de associação

10.2.1 Teste de independência do Qui-quadrado

Objetivo: Testar a independência entre 2 variáveis, X e Y, que se encontram agrupadas em classes mutuamente exclusivas e exaustivas.

Ideia base:

- Construir L classes de valores de X: $X_1, X_2, \dots, X_L$.
- Construir C classes de valores de Y: $Y_1, Y_2, \dots, Y_C$.
- Determinar as frequências absolutas simples observadas, O_{ij}, para cada par de valores $(X;\ Y)$, ou seja, o número de elementos da amostra que pertencem a $(X_i;\ Y_j)$.
- Determinar as frequências absolutas estimadas,
$$E_{ij} = \frac{O_{i.}O_{.j}}{n}$$
para cada par $(X;\ Y)$ tendo em conta a condição de independência.

Se as variáveis forem independentes então as frequências observadas estarão próximas das estimadas, caso contrário não se verifica a hipótese de independência definida em H_0.

Tabela de contingência:

	Y_1	Y_2	…	Y_j	…	Y_J	Total (X)
X_1	O_{11}	O_{12}	…	O_{1j}	…	O_{1C}	$n_{1.}$
X_2	O_{21}	O_{22}	…	O_{2j}	…	O_{2C}	$n_{2.}$
…	…	…	…	…	…	…	…
X_i	O_{i1}	O_{i2}	…	O_{ij}	…	O_{iC}	$n_{i.}$
…	…	…	…	…	…	…	…
X_L	O_{L1}	O_{L2}	…	O_{Lj}	…	O_{LC}	$n_{L.}$
Total (Y)	$O_{.1}$	$O_{.2}$	…	$O_{.j}$	…	$O_{.C}$	n

Notação:

n número total de observações;

L número de categorias da variável X;

C número de categorias da variável Y;

O_{ij} frequência absoluta observada na célula $(i; j)$;

$O_{i.}$ frequência marginal observada na categoria i da variável X;

$O_{.j}$ frequência marginal observada na categoria j da variável Y;

E_{ij} frequência estimada para a célula $(i; j)$;

Hipóteses a testar:

H_0: As variáveis X e Y são independentes *vs* H_1:As variáveis X e Y não são independentes

Estatística de teste:

$$\chi^2 = \sum_{i=1}^{L}\sum_{j=1}^{C}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(L-1)(C-1)}.$$

Regra de decisão: Rejeitar H_0 se

$$\chi^2_{obs} \geq \chi^2_{(L-1)(C-1);\,1-\alpha} \text{ (Teste unilateral direito).}$$

Portanto, $R.A.$: $\left[0;\ \chi^2_{(L-1)(C-1);\,1-\alpha}\right[$ e $R.R.$: $\left[\chi^2_{(L-1)(C-1);\,1-\alpha}; +\infty\right[$.

Cálculo do $valor\ p$:

$$\text{valor } p = P\left(\chi^2 \geq \chi^2_{obs}\right) = 1 - P\left(\chi^2 < \chi^2_{obs}\right).$$

Observações:

- Este teste têm as mesmas condições de aplicabilidade que o teste de ajustamento do Qui-quadrado pois assenta na mesma base teórica. Deste modo devem verificar-se as seguintes condições; i) não mais de 20% das classes com $E_{ij} < 5$; ii) todas as classes com $E_{ij} \geq 1$. Quando estas regras não são cumpridas, pode-se proceder à agregação de classes adjacentes.
- Em tabelas 2×2 deve-se efetuar a *correção de Yates*, para melhorar a aproximação à distribuição χ^2, que consiste em considerar a seguinte estatística de teste:

$$\chi^2 = \sum_{i=1}^{L}\sum_{j=1}^{C}\frac{\left(\left|O_{ij} - E_{ij}\right| - 0{,}5\right)^2}{E_{ij}} = \frac{n(|O_{11}O_{22} - O_{12}O_{21}| - 0{,}5n)^2}{O_{1.}O_{2.}O_{.1}O_{.2}} \sim \chi^2_{(L-1)(C-1)=1}.$$

- Para tabelas 2×2 que não cumpram as condições de aplicabilidade (amostras pequenas) é calculado o teste Exato de Fisher, que considera uma distribuição Hipergeométrica (Skeskin, 2011).

10.2.2 Teste de correlação ordinal de Spearman

O teste de correlação ordinal de Spearman constitui a alternativa não paramétrica ao teste paramétrico para o coeficiente de correlação quando não existe a garantia da Normalidade da distribuição.

Objetivo: Testar se existe correlação entre duas variáveis X e Y.

Notação:

n número total de observações;

R_s coeficiente de correlação amostral de Spearman;

ρ_S coeficiente de correlação populacional.

10.2.2.1 Amostras muito pequenas

Estatística de teste: para amostras muito pequenas (usualmente $n < 10$) a estatística de teste a utilizar é:

$$R = R_s$$

Observação: A distribuição da estatística de teste de Spearman encontra-se tabelada (Anexo G).

T. bilateral	T. unilateral direito	T. unilateral esquerdo
Hipóteses a testar: $H_0: \rho_S = 0$ *vs* $H_1: \rho_S \neq 0$	**Hipóteses a testar:** $H_0: \rho_S \leq 0$ *vs* $H_1: \rho_S > 0$	**Hipóteses a testar:** $H_0: \rho_S \geq 0$ vs $H_1: \rho_S < 0$
Regiões críticas: $R.A.: \left]-r_{n;1-\frac{\alpha}{2}};\ r_{n;1-\frac{\alpha}{2}}\right[$ $R.R.: \left[-1;\ -r_{n;1-\frac{\alpha}{2}}\right] \cup \left[r_{n;1-\frac{\alpha}{2}}; 1\right]$	**Regiões críticas:** $R.A.: \left[-1;\ r_{n;1-\alpha}\right[$ $R.R.: \left[r_{n;1-\alpha}; 1\right]$	**Regiões críticas:** $R.A.: \left]-r_{n;1-\alpha}; 1\right]$ $R.R.: \left[-1;\ -r_{n;1-\alpha}\right]$
Regra de decisão: Rejeitar H_0 quando $\lvert r_{obs}\rvert \geq r_{n;1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $r_{obs} \geq r_{n;1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $r_{obs} \leq -r_{n;1-\alpha}$

10.2.2.2 Amostras não muito pequenas

Estatística de teste: para amostras com pelo menos 10 observações (valor de referência usual) a estatística de teste a utilizar é:

$$T = \frac{R_s}{\sqrt{\frac{1-R_s^2}{n-2}}} \sim t_{n-2}$$

T. bilateral	T. unilateral direito	T. unilateral esquerdo
Hipóteses a testar: $H_0: \rho_S = 0$ *vs* $H_1: \rho_S \neq 0$	**Hipóteses a testar:** $H_0: \rho_S \leq 0$ *vs* $H_1: \rho_S > 0$	**Hipóteses a testar:** $H_0: \rho_S \geq 0$ vs $H_1: \rho_S < 0$
Regiões críticas: $R.A.: \left]-t_{n-2;1-\frac{\alpha}{2}};\ t_{n-2;1-\frac{\alpha}{2}}\right[$ $R.R.: \left]-\infty;\ -t_{n-2;1-\frac{\alpha}{2}}\right] \cup \left[t_{n-2;1-\frac{\alpha}{2}}; +\infty\right[$	**Regiões críticas:** $R.A.: \left]-\infty;\ t_{n-2;1-\alpha}\right[$ $R.R.: \left[t_{n-2;1-\alpha}; +\infty\right[$	**Regiões críticas:** $R.A.: \left]-t_{n-2;1-\alpha};\ +\infty\right[$ $R.R.: \left]-\infty;\ -t_{n-2;1-\alpha}\right]$
Regra de decisão: Rejeitar H_0 quando $\lvert t_{obs}\rvert \geq t_{n-2;1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $t_{obs} \geq t_{n-2;1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $t_{obs} \leq -t_{n-2;1-\alpha}$
Cálculo do valor p: valor $p = 2 \times P(T \geq \lvert t_{obs}\rvert)$	**Cálculo do valor p:** valor $p = P(T \geq t_{obs})$	**Cálculo do valor p:** valor $p = P(T \leq t_{obs})$

10.3 Testes de localização

Quando não se verificam as condições de aplicabilidade dos testes paramétricos, os testes que a seguir se apresentam constituem a alternativa não paramétrica:

- **Teste dos Sinais e teste de Wilcoxon** – alternativa aos testes T e Z paramétricos para a média e diferença de médias populacionais no caso em que se tem uma ou duas amostras emparelhadas;
- **Teste de Mann-Whitney-Wilcoxon** – alternativa aos testes T e Z para a diferença de médias populacionais no caso em que se tem duas amostras independentes;
- **Teste de Kruskal-Wallis** – alternativa ao teste F da ANOVA.

De salientar que nestes testes o parâmetro de localização estudado é a *mediana*, representada por $\tilde{\mu}$, a qual constitui uma boa alternativa ao valor esperado, dado que:

- Não é influenciada por valores extremos;
- Nas distribuições simétricas a mediana é igual à média, i. e., $\tilde{\mu} = \mu$;
- Nas distribuições assimétricas a mediana está mais próxima do valor mais frequente.

10.3.1 Teste do Sinais

Designa-se por **teste do Sinais** porque utiliza apenas sinais positivos e negativos, em vez dos valores numéricos.

Notação:

n dimensão da amostra inicial;

n' dimensão da amostra corrigida (sem as observações "empate").

10.3.1.1 Uma amostra

Objetivo: Testar se a mediana da população é igual a um determinado valor $\tilde{\mu}_0$.

Ideia base: Para cada uma das observações x_i da amostra, calcular as diferenças $D_i = X_i - \tilde{\mu}_0$ e reter apenas o sinal do resultado final (+ ou –). Esta diferença será:

- *Positiva* quando o valor da observação for superior ao da mediana $\tilde{\mu}_0$;
- *Negativa* quando o valor da observação for inferior ao da mediana $\tilde{\mu}_0$;
- *Nula* quando o valor da observação for igual ao da mediana $\tilde{\mu}_0$ ("empate"), devendo-se neste caso eliminar este valor e corrigir a dimensão da amostra, n'.

Se a mediana da população for $\tilde{\mu}_0$ então, numa amostra aleatória, o número de sinais + será aproximadamente igual ao número de sinais –, ou seja a proporção de elementos da amostra com valor inferior ou igual a $\tilde{\mu}_0$ será 0,5 (i. e., $P(-) = 0{,}5$) e a proporção de elementos com valor superior ou igual à mediana será também 0,5 (i. e., $P(+) = 0{,}5$). Portanto, quando esta situação não se verifica é posta em causa a hipótese de a mediana da população ser $\tilde{\mu}_0$.

Hipóteses a testar:

T. bilateral	**T. unilateral direito**	**T. unilateral esquerdo**
$H_0: \tilde{\mu} = \tilde{\mu}_0$ *vs* $H_1: \tilde{\mu} \neq \tilde{\mu}_0$	$H_0: \tilde{\mu} \leq \tilde{\mu}_0$ *vs* $H_1: \tilde{\mu} > \tilde{\mu}_0$	$H_0: \tilde{\mu} \geq \tilde{\mu}_0$ vs $H_1: \tilde{\mu} < \tilde{\mu}_0$
$\Leftrightarrow H_0: P(+) = P(-)$ *vs* $H_1: P(+) \neq P(-)$	$\Leftrightarrow H_0: P(+) \leq P(-)$ *vs* $H_1: P(+) > P(-)$	$\Leftrightarrow H_0: P(+) \geq P(-)$ *vs* $H_1: P(+) < P(-)$
$\Leftrightarrow H_0: p = 0{,}5$ *vs* $H_1: p \neq 0{,}5$	$\Leftrightarrow H_0: p \leq 0{,}5$ *vs* $H_1: p > 0{,}5$	$\Leftrightarrow H_0: p \geq 0{,}5$ *vs* $H_1: p < 0{,}5$

Estatística de teste:

$$S^+ = \text{Número total de sinais} + \text{que ocorrem} \sim B(n'; 0{,}5).$$

10.3.1.1.1 Amostras pequenas

T. bilateral	**T. unilateral direito**	**T. unilateral esquerdo**
Regiões críticas: $R.A.: \left\{b_{\frac{\alpha}{2}} + 1, \dots, b_{1-\frac{\alpha}{2}} - 1\right\}$ $R.R.: \left\{0, \dots, b_{\frac{\alpha}{2}}, b_{1-\frac{\alpha}{2}}, \dots, n'\right\}$	**Regiões críticas:** $R.A.: \{0, 1, \dots, b_{1-\alpha} - 1\}$ $R.R.: \{b_{1-\alpha}, \dots, n'\}$	**Regiões críticas:** $R.A.: \{b_\alpha + 1, \dots, n'\}$ $R.R.: \{0, 1, \dots, b_\alpha\}$
Regra de decisão: Rejeitar H_0 quando $s^+_{obs} \leq b_{\frac{\alpha}{2}}$ ou $s^+_{obs} \geq b_{1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $s^+_{obs} \geq b_{1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $s^+_{obs} \leq b_\alpha$
Cálculo do valor p: $\text{valor } p = 2 \times$ $\times\, min\{P(S^+ \leq s^+_{obs}); P(S^+ \geq s^+_{obs})\}$	**Cálculo do valor p:** $\text{valor } p = P(S^+ \geq s^+_{obs})$	**Cálculo do valor p:** $\text{valor } p = P(S^+ \leq s^+_{obs})$

Sendo $b_{\frac{\alpha}{2}}, b_\alpha, B_{1-\alpha}$ e $b_{1-\frac{\alpha}{2}}$ os quantis de probabilidade $\frac{\alpha}{2}, \alpha, 1-\alpha$ e $1-\frac{\alpha}{2}$, respetivamente, obtidos através do cálculo directo:

- $b_{\frac{\alpha}{2}}$ é o maior inteiro tal que:

$$P\left(S^+ \leq b_{\frac{\alpha}{2}}\right) = \sum_{x=0}^{b_{\frac{\alpha}{2}}} {}^{n'}C_x 0{,}5^{n'} \leq \frac{\alpha}{2};$$

- b_α é o maior inteiro tal que:

$$P(S^+ \leq b_\alpha) = \sum_{x=0}^{b_\alpha} {}^{n'}C_x 0{,}5^{n'} \leq \alpha;$$

- $b_{1-\alpha}$ é o menor inteiro tal que:

$$P(S^+ \geq b_{1-\alpha}) = \sum_{x=b_{1-\alpha}}^{n'} {}^{n'}C_x 0{,}5^{n'} \leq \alpha \Leftrightarrow b_{1-\alpha} = n' - b_\alpha;$$

- $b_{1-\frac{\alpha}{2}}$ é o menor inteiro tal que:

$$P\left(S^+ \leq b_{1-\frac{\alpha}{2}}\right) = \sum_{x=b_{1-\frac{\alpha}{2}}}^{n'} {}^{n'}C_x 0{,}5^{n'} \leq \frac{\alpha}{2} \Leftrightarrow b_{1-\frac{\alpha}{2}} = n' - b_{\frac{\alpha}{2}}.$$

10.3.1.1.2 Amostras grandes

T. bilateral	T. unilateral direito	T. unilateral esquerdo
Regiões críticas:	**Regiões críticas:**	**Regiões críticas:**
$R.A.: \left]b'_{\frac{\alpha}{2}};\ b'_{1-\frac{\alpha}{2}}\right[$	$R.A.: [0;\ b'_{1-\alpha}[$	$R.A.:]b'_{\alpha};n']$
$R.R.: \left[0;\ b'_{\frac{\alpha}{2}}\right] \cup \left[b'_{1-\frac{\alpha}{2}};n'\right]$	$R.R.: [b'_{1-\alpha};n']$	$R.R.: [0;\ b'_{\alpha}]$
Regra de decisão:	**Regra de decisão:**	**Regra de decisão:**
Rejeitar H_0 quando	Rejeitar H_0 quando	Rejeitar H_0 quando
$s^+_{obs} \leq b'_{\frac{\alpha}{2}}$ ou $s^-_{obs} \geq b'_{1-\frac{\alpha}{2}}$	$s^+_{obs} \geq b'_{1-\alpha}$	$s^+_{obs} \leq b'_{\alpha}$

Sendo $b'_{\frac{\alpha}{2}}$, b'_{α}, $b'_{1-\alpha}$ e $b'_{1-\frac{\alpha}{2}}$ os quantis de probabilidade $\frac{\alpha}{2}$, α, $1-\alpha$ e $1-\frac{\alpha}{2}$, respetivamente, obtidos através da aproximação da distribuição Binomial à Normal, sem esquecer a correção de continuidade, da seguinte forma:

- $b'_{\frac{\alpha}{2}} = \frac{n'}{2} - 0{,}5 - z_{1-\frac{\alpha}{2}}\frac{\sqrt{n'}}{2}$
- $b'_{1-\alpha} = \frac{n'}{2} + 0{,}5 + z_{1-\alpha}\frac{\sqrt{n'}}{2}$
- $b'_{\alpha} = \frac{n'}{2} - 0{,}5 - z_{1-\alpha}\frac{\sqrt{n'}}{2}$
- $b'_{1-\frac{\alpha}{2}} = \frac{n'}{2} + 0{,}5 + z_{1-\frac{\alpha}{2}}\frac{\sqrt{n'}}{2}$

10.3.1.2 Duas amostras emparelhadas

Objetivo: Testar se duas amostras emparelhadas (dependentes) têm a mesma mediana.

Ideia base: Para cada par $(X_i; Y_i)$ de observações da amostra, calcular as diferenças $D_i = X_i - Y_i$ e reter apenas o sinal do resultado final (+ ou –). Esta diferença será:

- *Positiva* quando o valor da observação X_i for superior a Y_i;
- *Negativa* quando o valor da observação X_i for inferior a Y_i;
- *Nula* quando o valor da observação X_i for igual a Y_i ("empate"), devendo-se neste caso eliminar este valor e corrigir a dimensão da amostra, n'.

Se os dois conjuntos de dados tiverem a mesma mediana então o número de sinais + será aproximadamente igual ao número de sinais –. Portanto, quando esta situação não se verifica é posta em causa a hipótese da igualdade das medianas.

Hipóteses a testar:

T. bilateral	T. unilateral direito	T. unilateral esquerdo
$H_0: \tilde{\mu}_1 = \tilde{\mu}_2$ *vs* $H_1: \tilde{\mu}_1 \neq \tilde{\mu}_2$	$H_0: \tilde{\mu}_1 \leq \tilde{\mu}_2$ *vs* $H_1: \tilde{\mu}_1 > \tilde{\mu}_2$	$H_0: \tilde{\mu}_1 \geq \tilde{\mu}_2$ vs $H_1: \tilde{\mu}_1 < \tilde{\mu}_2$
$\Leftrightarrow H_0: P(+) = P(-)$ *vs* $H_1: P(+) \neq P(-)$	$\Leftrightarrow H_0: P(+) \leq P(-)$ *vs* $H_1: P(+) > P(-)$	$\Leftrightarrow H_0: P(+) \geq P(-)$ *vs* $H_1: P(+) < P(-)$
$\Leftrightarrow H_0: p = 0{,}5$ *vs* $H_1: p \neq 0{,}5$	$\Leftrightarrow H_0: p \leq 0{,}5$ *vs* $H_1: p > 0{,}5$	$\Leftrightarrow H_0: p \geq 0{,}5$ *vs* $H_1: p < 0{,}5$

Estatística de teste:

$$S^+ = \text{Número total de sinais} + \text{que ocorrem} \sim B(n'; 0{,}5)$$

Regra de decisão: É a mesma que no caso em que se tem apenas uma amostra, tendo em consideração o tipo de teste (bilateral, unilateral esquerdo ou unilateral direito) e a dimensão da amostra.

Observação: A aplicação do teste dos sinais a duas amostras emparelhadas corresponde à aplicação deste teste a uma amostra, cujas as observações são os D_i.

10.3.2 Teste de Wilcoxon

Enquanto que no teste dos Sinais perde-se a informação relativa à magnitude das diferenças, no **teste de Wilcoxon** para além dos sinais também se tem em conta essa magnitude.

Pressuposto: A distribuição da população (no caso de 1 amostra) ou da população das diferenças (no caso de 2 amostras emparelhadas) é simétrica. Nesta situação este teste é mais potente que o teste dos Sinais.

10.3.2.1 Uma amostra

Objetivo: Testar se a mediana da população é igual a um determinado valor $\tilde{\mu}_0$.

Ideia base: Para cada uma das observações X_i da amostra, calcular as diferenças $D_i = X_i - \tilde{\mu}_0$. Esta diferença será:

- *Positiva* quando o valor da observação for superior ao da mediana $\tilde{\mu}_0$;
- *Negativa* quando o valor da observação for inferior ao da mediana $\tilde{\mu}_0$;
- *Nula* quando o valor da observação for igual ao da mediana $\tilde{\mu}_0$ ("empate"), devendo-se neste caso eliminar este valor e corrigir a dimensão da amostra, n'.

Ordenar por ordem crescente os valores $|D_1|, |D_2|, \dots, |D_n|$ numerando-os e associando o sinal + se D_i é positivo ou um sinal – se D_i é negativo. Em situação de empates, ou seja, existem $|D_i|$ iguais, a ordem a atribuir a essas observações corresponde à média aritmética das suas ordens. De notar que a soma de todos os números de ordem é $n(n+1)/2$.

Se a mediana da população for $\tilde{\mu}_0$ então a soma das ordens, associadas aos D_i positivos, deverá ser idêntica à soma das ordens associadas aos D_i negativos, e simétrica em torno de $n(n+1)/4$. Quando esta situação não se verifica é posta em causa a hipótese de a mediana da população ser $\tilde{\mu}_0$.

Hipóteses a testar:

T. bilateral	T. unilateral direito	T. unilateral esquerdo
$H_0: \tilde{\mu} = \tilde{\mu}_0$ *vs* $H_1: \tilde{\mu} \neq \tilde{\mu}_0$	$H_0: \tilde{\mu} \leq \tilde{\mu}_0$ *vs* $H_1: \tilde{\mu} > \tilde{\mu}_0$	$H_0: \tilde{\mu} \geq \tilde{\mu}_0$ vs $H_1: \tilde{\mu} < \tilde{\mu}_0$

10.3.2.1.1 Amostras pequenas

Estatística de teste:

$$W = \text{Soma das ordens dos } |D_i| \text{ com sinal } +$$

T. bilateral	T. unilateral direito	T. unilateral esquerdo
Regiões críticas: $R.A.: \left]w_{n';\frac{\alpha}{2}};\ w_{n';1-\frac{\alpha}{2}}\right[$ $R.R.: \left[0; w_{n';\frac{\alpha}{2}}\right] \cup \left[w_{n';1-\frac{\alpha}{2}}; \frac{n'(n'+1)}{2}\right]$	**Regiões críticas:** $R.A.: [0;\ w_{n';1-\alpha}[$ $R.R.: \left[w_{n';1-\alpha}; \frac{n'(n'+1)}{2}\right]$	**Regiões críticas:** $R.A.: \left]w_{n';\alpha}; \frac{n'(n'+1)}{2}\right]$ $R.R.: [0;\ w_{n';\alpha}]$
Regra de decisão: Rejeitar H_0 quando $w_{obs} \leq w_{n';\frac{\alpha}{2}}$ ou $w_{obs} \geq w_{n';1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $w_{obs} \geq w_{n';1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $w_{obs} \leq w_{n';\alpha}$

Sendo $w_{n';\frac{\alpha}{2}}$, $w_{n';\alpha}$, $w_{n';1-\alpha}$ e $w_{n';1-\frac{\alpha}{2}}$ os quantis de probabilidade $\frac{\alpha}{2}$, α , $1-\alpha$ e $1-\frac{\alpha}{2}$, respetivamente, obtidos através da tabela com os ponto críticos para a estatística de Wilcoxon apresentada no Anexo H.

10.3.2.1.2 Amostras grandes

Estatística de teste:

$$Z = \frac{W - \frac{n'(n'+1)}{4}}{\sqrt{\frac{n'(n'+1)(2n'+1)}{24}}} \overset{\circ}{\sim} N(0;1),$$

onde $W =$ Soma das ordens dos $|D_i|$ com sinal $+$.

Quando existem empates (observações com a mesma ordem) pode-se aplicar uma correção no denominador da estatística de teste (ver Murteira *et al.*, 2007). Segundo estes autores a redução verificada não é muito significativa, pelo que geralmente não existe a preocupação de efetuar essa alteração.

T. bilateral	T. unilateral direito	T. unilateral esquerdo
Regiões críticas: $R.A.: \left]-z_{1-\frac{\alpha}{2}};\ z_{1-\frac{\alpha}{2}}\right[$ $R.R.: \left]-\infty;\ -z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}}; +\infty\right[$	**Regiões críticas:** $R.A.:]-\infty;\ z_{1-\alpha}[$ $R.R.: [z_{1-\alpha}; +\infty[$	**Regiões críticas:** $R.A.:]-z_{1-\alpha}; +\infty[$ $R.R.:]-\infty;\ -z_{1-\alpha}]$
Regra de decisão: Rejeitar H_0 quando $\lvert z_{obs}\rvert \geq z_{1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \geq z_{1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \leq -z_{1-\alpha}$
Cálculo do valor p: valor $p = 2 \times P(Z \geq \lvert z_{obs}\rvert) = 2 \times \left(1 - \Phi(\lvert z_{obs}\rvert)\right)$	**Cálculo do valor p:** valor $p = P(Z \geq z_{obs}) = 1 - \Phi(z_{obs})$	**Cálculo do valor p:** valor $p = P(Z \leq z_{obs}) = \Phi(z_{obs})$

10.3.2.2 Duas amostras emparelhadas

Objetivo: Testar se duas amostras emparelhadas (dependentes) têm a mesma mediana.

Ideia base: Para cada par $(X_i; Y_i)$ de observações da amostra, calcular as diferenças $D_i = X_i - Y_i$ e reter o sinal do resultado final (+ ou –). Esta diferença será:

- *Positiva* quando o valor da observação X_i for superior a Y_i;
- *Negativa* quando o valor da observação X_i for inferior a Y_i;
- *Nula* quando o valor da observação X_i for igual a Y_i ("empate"), devendo-se neste caso eliminar este valor e corrigir a dimensão da amostra, n'.

Seguidamente procede-se da mesma forma que no caso em que se dispõe apenas de uma amostra: ordenar por ordem crescente os valores $|D_1|, |D_2|, \dots, |D_n|$ numerando-os e associando o sinal + se D_i é positivo ou um sinal – se D_i é negativo. Em caso de empates, a ordem a atribuir a essas observações corresponde à média aritmética das suas ordens.

Hipóteses a testar:

T. bilateral	**T. unilateral direito**	**T. unilateral esquerdo**
$H_0: F_{X_1}(x) = F_{X_2}(x)$ *vs* $H_1: F_{X_1}(x) \neq F_{X_2}(x)$	$H_0: F_{X_1}(x) \leq F_{X_2}(x)$ *vs* $H_1: F_{X_1}(x) > F_{X_2}(x)$	$H_0: F_{X_1}(x) \geq F_{X_2}(x)$ *vs* $H_1: F_{X_1}(x) < F_{X_2}(x)$
$\Leftrightarrow H_0: \tilde{\mu}_1 = \tilde{\mu}_2$ *vs* $H_1: \tilde{\mu}_1 \neq \tilde{\mu}_2$	$\Leftrightarrow H_0: \tilde{\mu}_1 \leq \tilde{\mu}_2$ *vs* $H_1: \tilde{\mu}_1 > \tilde{\mu}_2$	$\Leftrightarrow H_0: \tilde{\mu}_1 \geq \tilde{\mu}_2$ vs $H_1: \tilde{\mu}_1 < \tilde{\mu}_2$

Estatística de teste:

- Para amostras pequenas:

$$W = \text{Soma das ordens dos } |D_i| \text{ com sinal } +.$$

- Para amostras grandes:

$$Z = \frac{W - \frac{n'(n'+1)}{4}}{\sqrt{\frac{n'(n'+1)(2n'+1)}{24}}} \overset{\circ}{\sim} N(0;1).$$

Quando existem empates (observações com a mesma ordem) pode-se aplicar uma correção no denominador da estatística de teste (ver Murteira *et al.*, 2007).

Regra de decisão: É a mesma que no caso em que se tem apenas uma amostra, tendo em consideração o tipo de teste (bilateral, unilateral esquerdo ou unilateral direito) e a dimensão das amostras.

Observação: A aplicação do teste de Wilcoxon a duas amostras emparelhadas corresponde à aplicação deste teste a uma amostra, cujas as observações são os D_i.

10.3.3 Teste de Mann-Whitney U

Este teste também é conhecido por teste de **Mann-Whitney-Wilcoxon** ou teste **Wilcoxon rank-sum**, que são versões equivalentes ao teste **Mann-Whitney U**.

Objetivo: Testar se duas amostras independentes têm a mesma mediana.

Ideia base: Recolhem-se duas amostras independentes com dimensão n_1 e n_2 respetivamente. Seguidamente ordenam-se todas as $n(=n_1+n_2)$ observações por ordem crescente atribuindo-se o número de ordem a cada uma delas. Soma-se todos os números de ordem da amostra menor, sendo R_1 o valor dessa soma. Se R_1 tomar valores pequenos então os valores das observações da amostra menor são predominantemente inferiores aos da amostra maior. Se R_1 tomar valores grandes então ocorre a situação inversa.

Observação: Em situação de empate aplicar a metodologia descrita no teste de Wilcoxon, ou seja, a ordem a atribuir a essas observações corresponde à média aritmética das suas ordens.

Hipóteses a testar:

T. bilateral	**T. unilateral direito**	**T. unilateral esquerdo**
$H_0: F_{X_1}(x) = F_{X_2}(x)$ *vs* $H_1: F_{X_1}(x) \neq F_{X_2}(x)$	$H_0: F_{X_1}(x) \leq F_{X_2}(x)$ *vs* $H_1: F_{X_1}(x) > F_{X_2}(x)$	$H_0: F_{X_1}(x) \geq F_{X_2}(x)$ *vs* $H_1: F_{X_1}(x) < F_{X_2}(x)$
$\Leftrightarrow H_0: \tilde{\mu}_1 = \tilde{\mu}_2$ *vs* $H_1: \tilde{\mu}_1 \neq \tilde{\mu}_2$	$\Leftrightarrow H_0: \tilde{\mu}_1 \leq \tilde{\mu}_2$ *vs* $H_1: \tilde{\mu}_1 > \tilde{\mu}_2$	$\Leftrightarrow H_0: \tilde{\mu}_1 \geq \tilde{\mu}_2$ vs $H_1: \tilde{\mu}_1 < \tilde{\mu}_2$

10.3.3.1.1 Amostras pequenas

Estatística de teste:

$$U = min\{U_1;\ U_2\},$$

com

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}, \qquad U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

e R_i a soma das ordens da amostra i, $i = 1, 2$.

Observação: No SPSS também é apresentada a estatística de teste W^*, proposta por Wilcoxon, e que corresponde a $W^* =$ Soma de todas as ordens das observações da amostra menor.

T. bilateral	**T. unilateral direito**	**T. unilateral esquerdo**
Regiões críticas:	**Regiões críticas:**	**Regiões críticas:**
$R.A.: \left\{u_{n_1;\,n_2;\,\frac{\alpha}{2}}+1, \ldots, u_{n_1;\,n_2;\,1-\frac{\alpha}{2}}-1\right\}$	$R.A.: \{0, \ldots, u_{n_1;\,n_2;\,1-\alpha}-1\}$	$R.A.: \{u_{n_1;\,n_2;\,\alpha}+1, \ldots, n_1n_2\}$
$R.R.: \left\{0, \ldots, u_{n_1;\,n_2;\,\frac{\alpha}{2}}, u_{n_1;\,n_2;\,1-\frac{\alpha}{2}}, \ldots, n_1n_2\right\}$	$R.R.: \{u_{n_1;\,n_2;\,1-\alpha}; n_1n_2\}$	$R.R.: \{0, \ldots, u_{n_1;\,n_2;\,\alpha}\}$
Regra de decisão:	**Regra de decisão:**	**Regra de decisão:**
Rejeitar H_0 quando $u_{obs} \leq u_{n_1;\,n_2;\,\frac{\alpha}{2}}$ ou $u_{obs} \geq u_{n_1;\,n_2;\,1-\frac{\alpha}{2}}$	Rejeitar H_0 quando $u_{obs} \geq u_{n_1;\,n_2;\,1-\alpha}$	Rejeitar H_0 quando $u_{obs} \leq u_{n_1;\,n_2;\,\alpha}$

Sendo $u_{n_1;\, n_2;\, \frac{\alpha}{2}}$, $u_{n_1;\, n_2;\, \alpha}$, $u_{n_1;\, n_2;\, 1-\alpha}$ e $u_{n_1;\, n_2;\, 1-\frac{\alpha}{2}}$ os quantis de probabilidade $\frac{\alpha}{2}$, α , $1-\alpha$ e $1-\frac{\alpha}{2}$, respetivamente, obtidos através da tabela com os ponto críticos para a estatística de Mann-Whitney U apresentada no Anexo I.

Observação: A distribuição da estatística U de Mann-Whitney é simétrica, logo $u_{n_1;\, n_2;\, 1-\alpha} = n_1 n_2 - u_{n_1;\, n_2;\, \alpha}$.

10.3.3.1.2 Amostras grandes

Estatística de teste:

$$Z = \frac{R_1 - \frac{n_2(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}} \overset{\circ}{\sim} N(0;1),$$

onde $R_1 =$ Soma de todas as ordens das observações da amostra menor.

Quando existem empates pode-se aplicar uma correção no denominador da estatística de teste (ver Guimarães e Cabral, 2010).

T. bilateral	**T. unilateral direito**	**T. unilateral esquerdo**
Regiões críticas: $R.A.: \left]-z_{1-\frac{\alpha}{2}};\ z_{1-\frac{\alpha}{2}}\right[$ $R.R.: \left]-\infty;\ -z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}};\ +\infty\right[$	**Regiões críticas:** $R.A.:]-\infty;\ z_{1-\alpha}[$ $R.R.: [z_{1-\alpha};\ +\infty[$	**Regiões críticas:** $R.A.:]-z_{1-\alpha};\ +\infty[$ $R.R.:]-\infty;\ -z_{1-\alpha}]$
Regra de decisão: Rejeitar H_0 quando $\lvert z_{obs}\rvert \geq z_{1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \geq z_{1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \leq -z_{1-\alpha}$
Cálculo do valor p: $\text{valor } p = 2 \times P(Z \geq \lvert z_{obs}\rvert)$ $= 2 \times \left(1 - \Phi(\lvert z_{obs}\rvert)\right)$	**Cálculo do valor p:** $\text{valor p} = P(Z \geq z_{obs})$ $= 1 - \Phi(z_{obs})$	**Cálculo do valor p:** $\text{valor } p = P(Z \leq z_{obs})$ $= \Phi(z_{obs})$

10.3.4 Teste de Kruskal-Wallis

O teste de Mann-Whitney U pode ser generalizado para mais de dois grupos através do teste de Kruskal-Wallis.

O teste de **Kruskal-Wallis** é a alternativa não paramétrica para a Análise de Variância Simples (ANOVA), devendo ser utilizado quando não são verificados os pressupostos da Normalidade ou da homogeneidade das variâncias, ou quando as variáveis são do tipo ordinal. Este teste é quase tão potente como o teste F da ANOVA.

Objetivo: Testar se existe diferença entre as distribuições dos K grupos., i. e., as amostras provêm de populações com a mesma distribuição, que é abordado como um teste aos parâmetros de localização central, usualmente à mediana (Pestana e Gageiro, 2014).

Pressupostos:

- Os dados têm de estar, pelo menos, na escala ordinal e assume-se que os dados são provenientes de uma população com distribuição contínua.
- As distribuições têm de ter a mesma forma.
- Os grupos têm de ter pelo menos 5 observações ($n_i \geq 5,\ i = 1, 2, \dots, K$).

Notação:

K	número de grupos;
n_i	número de observações no grupo $i, i = 1, \dots, K$;
$n = \sum_{i=1}^{K} n_i$	número total de observações;
R_i	soma das ordens das observações da amostra i.

Observação: Para calcular os R_i é necessário ordenar todas as n observações para depois se atribuir a cada uma delas o seu número de ordem.

Hipóteses a testar:

H_0: Os K grupos provêm da mesma população ou de populações idênticas *vs*
H_1: Nem todos os K grupos provêm da mesma população ou de populações idênticas

$\Leftrightarrow H_0: F_{X_1}(x) = F_{X_2}(x) = \cdots = F_{X_K}(x)$ *vs*
H_1: Há pelo menos um $F_{X_i}(x)$ que difere dos restantes, $i = 1, \dots, K$

$\Leftrightarrow H_0: \mu_1 = \mu_2 = \cdots = \mu_K$ *vs* H_1: Há pelo menos um μ_i que difere dos restantes, $i = 1, \dots, K$.

Estatística de teste:

$$\chi^2 = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1) \overset{\circ}{\sim} \chi^2_{K-1}$$

Quando exista um número excessivo de empates (observações com a mesma ordem) vários autores recomendam aplicar uma correção à estatística de teste (ver Skeskin, 2011).

Regra de decisão: Rejeitar H_0 se

$$\chi^2_{obs} \geq \chi^2_{K-1;\,1-\alpha} \text{ (Teste unilateral direito).}$$

Portanto, $R.A.$: $\left[0;\ \chi^2_{K-1;\,1-\alpha}\right[$ e $R.R.$: $\left[\chi^2_{K-1;\,1-\alpha};\ +\infty\right[$.

10.4 Teste à simetria

Hipóteses a testar:

T. bilateral	**T. unilateral direito**	**T. unilateral esquerdo**
H_0: A distribuição é simétrica *vs* H_1: A distribuição não é simétrica $\Leftrightarrow H_0: \gamma_1 = 0$ *vs* $H_1: \gamma_1 \neq 0$	H_0: A distribuição é simétrica ou assimétrica negativa *vs* H_1: A distribuição é assimétrica positiva $\Leftrightarrow H_0: \gamma_1 \leq 0$ *vs* $H_1: \gamma_1 > 0$	H_0: A distribuição é simétrica ou assimétrica positiva *vs* H_1: A distribuição é assimétrica negativa $\Leftrightarrow H_0: \gamma_1 \geq 0$ *vs* $H_1: \gamma_1 < 0$

Estatística de teste:

$$Z = \frac{g_a}{EP_{g_a}} \overset{\circ}{\sim} N(0;1),$$

com

$$EP_{g_a} = \sqrt{\frac{6\mathrm{n}(n-1)}{(n-2)(n+1)(n+3)}}$$

e g_a definido na secção 2.2.4.4.

Regra de decisão:

T. bilateral	T. unilateral direito	T. unilateral esquerdo
Regiões críticas: $R.A.: \left]-z_{1-\frac{\alpha}{2}};\ z_{1-\frac{\alpha}{2}}\right[$ $R.R.: \left]-\infty; z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}}; +\infty\right[$	**Regiões críticas:** $R.A.:]-\infty;\ z_{1-\alpha}[$ $R.R.: [z_{1-\alpha}; +\infty[$	**Regiões críticas:** $R.A.:]-z_{1-\alpha}; +\infty[$ $R.R.:]-\infty;\ -z_{1-\alpha}]$
Regra de decisão: Rejeitar H_0 quando $\lvert z_{obs}\rvert \geq z_{1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \geq z_{1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \leq -z_{1-\alpha}$
Cálculo do valor p: $\text{valor } p = 2 \times P(Z \geq \lvert z_{obs}\rvert)$ $= 2 \times \left(1 - \Phi(\lvert z_{obs}\rvert)\right)$	**Cálculo do valor p:** $\text{valor } p = P(Z \geq z_{obs})$ $= 1 - \Phi(z_{obs})$	**Cálculo do valor p:** $\text{valor } p = P(Z \leq z_{obs})$ $= \Phi(z_{obs})$

10.5 Teste ao achatamento

Hipóteses a testar:

T. bilateral	T. unilateral direito	T. unilateral esquerdo
H_0: A distribuição é mesocúrtica *vs* H_1: A distribuição não é mesocúrtica $\Leftrightarrow H_0: \gamma_2 = 0$ *vs* $H_1: \gamma_2 \neq 0$	H_0: A distribuição é mesocúrtica ou platicúrtica *vs* H_1: A distribuição é leptocúrtica $\Leftrightarrow H_0: \gamma_2 \leq 0$ *vs* $H_1: \gamma_2 > 0$	H_0: A distribuição é mesocúrtica ou leptocúrtica *vs* H_1: A distribuição é platicúrtica $\Leftrightarrow H_0: \gamma_2 \geq 0$ *vs* $H_1: \gamma_2 < 0$

Estatística de teste:

$$Z = \frac{k_a}{EP_{k_a}} \overset{\circ}{\sim} N(0;1),$$

com

$$EP_{k_a} = \sqrt{\frac{4(n^2-1)\left(EP_{g_a}\right)^2}{(n-3)(n+5)}}$$

e k_a definido na secção 2.2.5.3.

Regra de decisão:

T. bilateral	**T. unilateral direito**	**T. unilateral esquerdo**
Regiões críticas: $R.A.: \left]-z_{1-\frac{\alpha}{2}};\ z_{1-\frac{\alpha}{2}}\right[$ $R.R.: \left]-\infty; z_{1-\frac{\alpha}{2}}\right] \cup \left[z_{1-\frac{\alpha}{2}}; +\infty\right[$	**Regiões críticas:** $R.A.:]-\infty;\ z_{1-\alpha}[$ $R.R.: [z_{1-\alpha}; +\infty[$	**Regiões críticas:** $R.A.:]-z_{1-\alpha}; +\infty[$ $R.R.:]-\infty;\ -z_{1-\alpha}]$
Regra de decisão: Rejeitar H_0 quando $\lvert z_{obs}\rvert \geq z_{1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \geq z_{1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $z_{obs} \leq -z_{1-\alpha}$
Cálculo do valor p: $\text{valor } p = 2 \times P(Z \geq \lvert z_{obs}\rvert) = 2 \times \left(1 - \Phi(\lvert z_{obs}\rvert)\right)$	**Cálculo do valor p:** $\text{valor } p = P(Z \geq z_{obs}) = 1 - \Phi(z_{obs})$	**Cálculo do valor p:** $\text{valor } p = P(Z \leq z_{obs}) = \Phi(z_{obs})$

10.6 Quadro resumo

Na Tabela 10.2 apresentamos o resumo dos testes de hipóteses não paramétricos.

Tabela 10.2: *Quadro resumo dos testes de hipótese não paramétricos.*

Aplicação (n.º de amostras)	H_0	Teste	Dimensão da(s) amostra(s)	Estatística de Teste
Ajustamento (1)	$X \sim F_0(x)$	Ajustamento do χ^2	Qualquer	$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-p-1}$
		Kolmogorov-Smirnov	Qualquer	$D = \sup_{x \in \mathbb{R}} \lvert F_n(x) - F_0(x)\rvert$. (ver pontos críticos em tabela própria)
	X tem distribuição Normal	Shapiro-Wilks	Qualquer	$W = \frac{(\sum_{i=1}^{n} a_i x_{i:n})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
Associação (2 emparelhadas)	X e Y são independentes	Independência do χ^2	Qualquer	$\chi^2 = \sum_{i=1}^{L}\sum_{j=1}^{C} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(L-1)(C-1)}$ Para tabelas 2×2 (Correção de Yates): $\chi^2 = \frac{n(\lvert O_{11}O_{22} - O_{12}O_{21}\rvert - 0{,}5n)^2}{O_{1.}O_{2.}O_{.1}O_{.2}} \sim \chi^2_1$
	$\rho_S = 0$	Correlação ordinal de Spearman	$n < 10$	$R = R_s$ (ver pontos críticos em tabela própria)
			$n \geq 10$	$T = \frac{R_s}{\sqrt{\frac{1 - R_s^2}{n-2}}} \sim t_{n-2}$

Tabela 10.2: *Quadro resumo dos testes de hipótese não paramétricos. (continuação)*

Aplicação (n.º de amostras)	H_0	Teste	População	Dimensão da(s) amostra	Estatística de Teste
Localização (1 amostra)	$\tilde{\mu} = \tilde{\mu}_0$	Sinais	Qualquer	Qualquer	$S^+ =$ Número total de sinais + que ocorrem $\sim B(n'; 0{,}5)$ (utilizar tabela da Binomial)
		Wilcoxon	Simétrica	$n \leq 30$	$W =$ Soma das ordens dos $\lvert D_i \rvert$ com sinal + (ver pontos críticos em tabela própria)
				$n > 30$	$Z = \dfrac{W - \frac{n'(n'+1)}{4}}{\sqrt{\frac{n'(n'+1)(2n'+1)}{24}}} \overset{\circ}{\sim} N(0;1)$
Localização (2 emparelhadas)	$\tilde{\mu}_1 = \tilde{\mu}_2$	Sinais	Qualquer	Qualquer	$S^+ =$ Número total de sinais + que ocorrem $\sim B(n'; 0{,}5)$ (utilizar tabela da Binomial)
	$\tilde{\mu}_1 = \tilde{\mu}_2$	Wilcoxon	Simétrica	$n \leq 30$	$W =$ Soma das ordens dos $\lvert D_i \rvert$ com sinal + (ver pontos críticos em tabela própria)
				$n > 30$	$Z = \dfrac{W - \frac{n'(n'+1)}{4}}{\sqrt{\frac{n'(n'+1)(2n'+1)}{24}}} \overset{\circ}{\sim} N(0;1)$

Tabela 10.2: *Quadro resumo dos testes de hipótese não paramétricos. (continuação)*

Aplicação (n.º de amostras)	H_0	Teste	Dimensão da(s) amostra	Estatística de Teste
Localização (2 independentes)	$\tilde{\mu}_1 = \tilde{\mu}_2$	Mann-Whitney U	$n_1 \leq 30$ e $n_2 \leq 30$	$U = min\left\{R_1 - \frac{n_1(n_1+1)}{2}; R_2 - \frac{n_2(n_2+1)}{2}\right\}$ (ver pontos críticos em tabela própria)
			$n_1 > 30$ e $n_2 > 30$	$Z = \frac{R_1 - \frac{n_2(n+1)}{2}}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}} \overset{\circ}{\sim} N(0;1)$
Localização (3 ou mais independentes)	$\mu_1 = \mu_2 = \cdots = \mu_K$	Kruskall-Wallis	$n_i \geq 5,$ $i = 1, 2, \ldots, K$	$\chi^2 = \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1) \overset{\circ}{\sim} \chi^2_{K-1}$
Simetria (1)	$\gamma_1 = 0$	Simetria		$Z = \frac{g_a}{EP_{g_a}} \overset{\circ}{\sim} N(0;1)$
Achamento (1)	$\gamma_2 = 0$	Achatamento		$Z = \frac{k_a}{EP_{k_a}} \overset{\circ}{\sim} N(0;1)$

10.7 Exercícios resolvidos

10.7.1 Teste de ajustamento do Qui-quadrado

1. Um determinado modelo de automóveis é vendido em 4 versões diferentes:
   - Descapotável (V1)
   - Clássico de 3 portas (V2)
   - Clássico de 5 portas (V3)
   - Comercial de 2 lugares (V4)

Para apreciar a popularidade das várias versões, analisou-se uma amostra relativa às preferências de 200 consumidores:

Versão preferida	V1	V2	V3	V4
N.º de consumidores	40	47	59	54

Para α = 1%, ensaie a hipótese de todas as versões serem igualmente populares.

Resolução:

Sejam:

- X_1 a v. a. que representa o número de consumidores, em 200, que preferem a versão V1;
- X_2 a v. a. que representa o número de consumidores, em 200, que preferem a versão V2;
- X_3 a v. a. que representa o número de consumidores, em 200, que preferem a versão V3;
- X_4 a v. a. que representa o número de consumidores, em 200, que preferem a versão V4.

$n = 200$ consumidores.

H_0: As versões V1, V2, V3 e V4 são igualmente populares *vs*
H_1: As versões V1, V2, V3 e V4 não são igualmente populares

ou equivalentemente:

H_0: $p_1 = p_2 = p_3 = p_4 = 0{,}25$ vs H_1: Nem todos os p_i são iguais, $i = 1, \ldots, 4$.

Para ensaiar a hipótese pretendida recorre-se ao teste de ajustamento do Qui-quadrado, visto que se pretende avaliar se os dados se ajustam a uma distribuição em que a probabilidade de um consumidor preferir cada uma das marcas é a mesma.

Estatística de teste:

$$\chi^2 = \sum_{i=1}^{K=4} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-p-1}.$$

Versão preferida	o_i (n.º de consum.)	p_i^*	$e_i = np_i^*$	$o_i - e_i$	$\frac{(o_i - e_i)^2}{e_i}$
V1	40	0,25	50	-10	2
V2	47	0,25	50	-3	0,18
V3	59	0,25	50	9	1,62
V4	54	0,25	50	4	0,32
Total	n = 200	1	n = 200	0	$\chi^2_{obs} = 4{,}12$

$K = 4; p = 0;\ \alpha = 0{,}01.$

Todos os $e_i \geq 5$, logo são verificados os pressupostos para a aplicação do teste de ajustamento do Qui-quadrado.

$\chi^2_{K-p-1;\,1-\alpha} = \chi^2_{3;\,0,99} = 11{,}34$. Logo, $R.A.: \,]-\infty;\ 11{,}34[$ e $R.R.: [11{,}34;\ +\infty[$.

Como $\chi^2_{obs} = 4{,}12 \in R.A.$, não rejeitar H_0 para $\alpha = 1\%$. Portanto, ao nível de significância de 1%, as 4 versões são igualmente preferidas pelos consumidores.

(SPSS) Analyse → Nonparametric Tests → Legacy Dialogs → Chi-square…
(Test Variable List: Versão; Expected Range: ⊙ Get from data; Expected Values: ⊙ All categories equal)

Chi-Square Test

Versão

	Observed N	Expected N	Residual
V1	40	50,0	-10,0
V2	47	50,0	-3,0
V3	59	50,0	9,0
V4	54	50,0	4,0
Total	200		

Test Statistics

	Versão
Chi-Square	4,120[a]
df	3
Asymp. Sig.	,249

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 50,0.

No primeiro quadro são apresentados os valores de o_i, e_i e $(o_i - e_i)$. Na nota de rodapé são verificados os pressupostos de aplicação do teste (0% de células com $e_i < 5$ e o menor e_i é 50), χ^2_{obs} é dado na linha *Chi-Square*, os graus de liberdade na linha *df*, e o valor p na linha *Aymp. Sig.* Como o valor $p = 0{,}249$ (2º quadro), H_0 só é rejeitada se $\alpha \geq 24{,}9\%$.

2. A loja "Grandes vendas" tem verificado nos últimos anos que 15% dos seus clientes pagam as suas compras com cheque, 38% com cartão de crédito, 32% com cartão de débito e 15% em dinheiro.
Uma amostra de 160 vendas realizadas na semana anterior ao Natal revelou os seguintes resultados:

Tipo de pagamento	Cheque	Cartão de crédito	Cartão de débito	Dinheiro
N.º de vendas	27	65	48	20

Será que o tipo de pagamento que os clientes da loja "Grandes vendas" utilizam na época natalícia é concordante com a informação que a loja tem ($\alpha = 10\%$)?

Resolução:

Sejam:

- X_1 a v. a. que representa o número de clientes, em 160, que pagaram com cheque;
- X_2 a v. a. que representa o número de clientes, em 160, que pagaram com cartão de crédito;
- X_3 a v. a. que representa o número de clientes, em 160, que pagaram com cartão de débito;
- X_4 a v. a. que representa o número de clientes, em 160, que pagaram com dinheiro.

$n = 160$ clientes.

$H_0: p_1 = 0{,}15$ e $p_2 = 0{,}38$ e $p_3 = 0{,}32$ e $p_4 = 0{,}15$ *vs*
$H_1: p_1 \neq 0{,}15$ ou $p_2 \neq 0{,}38$ ou $p_3 \neq 0{,}32$ ou $p_4 \neq 0{,}15$.

Para ensaiar a hipótese pretendida recorre-se ao teste de ajustamento do Qui-quadrado, visto que se pretende avaliar se os dados se ajustam a uma distribuição.

Estatística de teste:

$$\chi^2 = \sum_{i=1}^{K=4} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-p-1}.$$

Tipo de pagamento	o_i (n.º de vendas)	p_i^*	$e_i = np_i^*$	$o_i - e_i$	$\frac{(o_i - e_i)^2}{e_i}$
Cheque	27	0,15	24	3	0,375
Cartão de crédito	65	0,38	60,8	4,2	0,290
Cartão de débito	48	0,32	51,2	-3,2	0,200
Dinheiro	20	0,15	24	-4	0,667
Total	n = 160	1	n = 160	0	$\chi^2_{obs} = 1{,}532$

$K = 4; p = 0;\ \alpha = 0{,}1.$

Todos os $e_i \geq 5$, logo são verificados os pressupostos para a aplicação do teste de ajustamento do Qui-quadrado.

$\chi^2_{K-p-1;\,1-\alpha} = \chi^2_{3;\,0{,}90} = 6{,}251$. Logo, $R.A.:]-\infty;\ 6{,}251[$ e $R.R.: [6{,}251; +\infty[$.

Como $\chi^2_{obs} = 1{,}532 \in R.A.$, não rejeitar H_0 para $\alpha = 10\%$. Portanto, ao nível de significância de 1%, o tipo de pagamento utilizado pelos clientes na época natalícia é concordante com a informação que a loja tem.

(SPSS) Analyse → Nonparametric Tests → Legacy Dialogs → Chi-square…
(Test Variable List: Tipo_pagamento; Expected Range: ⊙ Get from data; Expected Values: ⊙ Values: 0,15; 0,38; 0,32; 0,15)

Chi-Square Test
Tipo de pagamento

	Observed N	Expected N	Residual
Cheque	27	24,0	3,0
Cartão de crédito	65	60,8	4,2
Cartão de débito	48	51,2	-3,2
Dinheiro	20	24,0	-4,0
Total	160		

Test Statistics

	Tipo de pagamento
Chi-Square	1,532[a]
df	3
Asymp. Sig.	,675

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 24,0.

3. Determinada empresa seguradora baseia o seu sistema de prémios para determinado risco na premissa de que o número de sinistros por apólice tem distribuição Poisson de parâmetro $\lambda = 0{,}2$. Numa amostra de 1000 apólices referentes ao ano anterior observou-se:

N.º sinistros por apólice	0	1	2	3
N.º apólices	800	175	21	4

Teste, ao nível de significância de 1%, a hipótese da suposição da empresa seguradora.

Resolução:

Seja X a v. a. que representa o número de sinistros por apólice.

H_0: $X \sim P(\lambda = 0{,}2)$ vs H_1: X não tem distribuição $P(\lambda = 0{,}2)$.

Estatística de teste:

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-p-1}.$$

Cálculo de p_i^*: Se H_0 verdadeiro, então $X \sim P(\lambda = 0{,}2)$ e, portanto,

$$p_i^* = P(X = i) = e^{-0{,}2}\frac{0{,}2^i}{i!}.$$

$p_0^* = P(X = 0) = 0{,}8187;$
$p_1^* = P(X = 1) = 0{,}1637;$
$p_2^* = P(X = 2) = 0{,}0164;$
$p_{3\ ou\ mais}^* = P(X \geq 3) = 1 - P(X \leq 2) = 0{,}0012.$

N.º sinistros por apólice	o_i	p_i^*	$e_i = np_i^*$
0	800	0,8187	818,7
1	175	0,1637	163,7
2	21	0,0164	16,4
≥ 3	4	0,0012	1,2
Total	n = 1000	1	1000

Uma vez que para se poder aplicar o teste apenas se pode ter 20% das classes com $e_i < 5$ (que neste caso seria 0,8 classes), agrupa-se a última classe com a 3ª classe.
Agora $p_{2\ ou\ mais}^* = P(X \geq 2) = 1 - P(X \leq 1) = 0{,}0176$.

N.º sinistros por apólice	o_i	p_i^*	$e_i = np_i^*$	$o_i - e_i$	$\frac{(o_i - e_i)^2}{e_i}$
0	800	0,8187	818,7	-18,7	0,4271
1	175	0,1637	163,7	11,3	0,7800
≥ 2	25	0,0176	17,6	7,4	3,1114
Total	$n = 1000$	1,00	n = 1000		$\chi^2_{obs} = 4{,}3185$

$K = 3; p = 0;\ \alpha = 0{,}01.$

Todos os $e_i \geq 5$, logo são verificados os pressupostos para a aplicação do teste de ajustamento do Qui-quadrado.

$\chi^2_{K-p-1;\,1-\alpha} = \chi^2_{2;\,0{,}99} = 9{,}21$. Logo, $R.A.$: $]-\infty;\ 9{,}21[$ e $R.R.$: $[9{,}21; +\infty[$.

Como $\chi^2_{obs} = 4{,}3185 \in R.A.$, não rejeitar H_0 para $\alpha = 1\%$. Portanto, ao nível de significância de 1%, não existem evidências de que a suposição da seguradora esteja incorreta.

(SPSS) Analyse → Nonparametric Tests → Legacy Dialogs → Chi-square…
(Test Variable List: N_sinistros; Expected Range: ⊙ Get from data; Expected Values: ⊙ Values: 0,8187; 0,1637; 0,0164; 0,0012)

Chi-Square Test

N.º sinistros por apólice

	Observed N	Expected N	Residual
0	800	818,7	-18,7
1	175	163,7	11,3
2	21	16,4	4,6
3 ou mais	4	1,2	2,8
Total	1000		

Test Statistics

	N.º sinistros por apólice
Chi-Square	9,031[a]
df	3
Asymp. Sig.	,029

a. 1 cells (25,0%) have expected frequencies less than 5. The minimum expected cell frequency is 1,2.

Devido ao conteúdo da mensagem sobre as condições de aplicabilidade, que surge em rodapé no segundo quadro, é necessário proceder ao agrupamento de classes e voltar a realizar o teste.

Chi-Square Test

N.º sinistros por apólice

	Observed N	Expected N	Residual
0	800	818,7	-18,7
1	175	163,7	11,3
2 ou mais	25	17,6	7,4
Total	1000		

Test Statistics

	N.º sinistros por apólice
Chi-Square	4,319[a]
df	2
Asymp. Sig.	,115

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 17,6.

4. Analisaram-se 250 medições de partículas totais em suspensão ($\mu g/m^3$) num jardim público, tendo-se obtido os seguintes resultados:

Partículas ($\mu g/m3$)	[0; 12[	[12; 14[	[14; 16[	[16; 18[	[18; 20[	[20; 22[	[22; +∞[
Frequência observada	7	22	55	81	53	24	8

Ao nível de significância de 10%, teste a hipótese dos dados amostrais provirem de uma população com distribuição:

a) Normal com média 18 e desvio padrão 3,5?

b) Normal?

Resolução:

Seja X a v. a. que representa a concentração de partículas totais em suspensão ($\mu g/m^3$) num jardim público.

a) $H_0: X \sim N(\mu = 18;\ \sigma = 3{,}5)$ vs $H_1: X$ não tem distribuição $N(\mu = 18;\ \sigma = 3{,}5)$.

Estatística de teste:

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-p-1}.$$

Cálculo de p_i^*: Se H_0 verdadeiro, então $X \sim N(\mu = 18;\ \sigma = 3{,}5)$ e, portanto,

$$p_x^* = P(X \leq x) = \Phi\left(\frac{x - 18}{3{,}5}\right).$$

$$p^*_{<12} = P(X < 12) = \Phi\left(\frac{12 - 18}{3{,}5}\right) = \Phi(-1{,}714) = 1 - \Phi(1{,}714) = 0{,}0432;$$

$$p^*_{[12;\ 14[} = P(12 \leq X < 14) = \Phi\left(\frac{14 - 18}{3{,}5}\right) - \Phi\left(\frac{12 - 18}{3{,}5}\right) = \Phi(-1{,}143) - \Phi(-1{,}714) = 0{,}0833;$$

$$p^*_{[14;\ 16[} = P(14 \leq X < 16) = \Phi\left(\frac{16 - 18}{3{,}5}\right) - \Phi\left(\frac{14 - 18}{3{,}5}\right) = \Phi(-0{,}571) - \Phi(-1{,}143) = 0{,}1573;$$

$$p^*_{[16;\ 18[} = P(16 \leq X < 18) = \Phi\left(\frac{18 - 18}{3{,}5}\right) - \Phi\left(\frac{16 - 18}{3{,}5}\right) = \Phi(0) - \Phi(-0{,}571) = 0{,}2161;$$

$$p^*_{[18;\ 20[} = P(18 \leq X < 20) = \Phi\left(\frac{20 - 18}{3{,}5}\right) - \Phi\left(\frac{18 - 18}{3{,}5}\right) = \Phi(0{,}571) - \Phi(0) = 0{,}2161;$$

$$p^*_{[20;\ 22[} = P(20 \leq X < 22) = \Phi\left(\frac{22 - 18}{3{,}5}\right) - \Phi\left(\frac{20 - 18}{3{,}5}\right) = \Phi(1{,}143) - \Phi(0{,}571) = 0{,}1573;$$

$$p^*_{[22;\ +\infty[} = P(X \geq 22) = P(X < 22) = 1 - \Phi(1{,}143) = 0{,}1265.$$

Partículas (μg/m3)	o_i (n.º de medições)	p_i^*	$e_i = np_i^*$	$o_i - e_i$	$\frac{(o_i - e_i)^2}{e_i}$
[0; 12[	7	0,0432	10,8095	-3,8095	1,3370
[12; 14[	22	0,0833	20,8277	1,1723	0,0663
[14; 16[	55	0,1573	39,3264	15,6736	6,2481
[16; 18[	81	0,2161	54,0364	26,9636	13,4688
[18; 20[	53	0,2161	54,0364	-1,0364	0,0194
[20; 22[	24	0,1573	39,3264	-15,3264	5,9722
[22; +∞[	8	0,1265	31,6372	-23,6372	17,6487
Total	$n = 250$	1,00	n =250		$\chi^2_{obs} = 44{,}7605$

$K = 7; p = 0;\ \alpha = 0{,}1.$

Todos os $e_i \geq 5$, logo são verificados os pressupostos para a aplicação do teste de ajustamento do Qui-quadrado.

$\chi^2_{K-p-1;\ 1-\alpha} = \chi^2_{6;\ 0{,}9} = 10{,}64$, logo $R.A.:]-\infty; 10{,}64[$ e $R.R.: [10{,}64; +\infty[$.

Como $\chi^2_{obs} = 44{,}7605 \in R.R.$, rejeitar H_0 para $\alpha = 10\%$. Ao nível de significância de 10%, as concentrações de partículas totais em suspensão não provêm de uma população com distribuição $N(\mu = 18;\ \sigma = 3{,}5)$.

(SPSS) Analyse → Nonparametric Tests → Legacy Dialogs → Chi-square…
(Test Variable List: Particulas; Expected Range: ⊙ Get from data; Expected Values: ⊙ Values: 0,0432; 0,0833; ...)

Chi-Square Test

Partículas

	Observed N	Expected N	Residual
[0; 12[	7	10,8	-3,8
[12; 14[	22	20,8	1,2
[14; 16[	55	39,3	15,7
[16; 18[	81	54,0	27,0
[18; 20[	53	54,0	-1,0
[20; 22[	24	39,3	-15,3
[22; +inf[	8	31,6	-23,6
Total	250		

Test Statistics

	Partículas
Chi-Square	44,752[a]
df	6
Asymp. Sig.	,000

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 10,8.

b) H_0: X não tem distribuição Normal *vs* H_1: X não tem distribuição Normal.

Estatística de teste:

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-p-1}.$$

Cálculo de p_i^*: Se H_0 verdadeiro, então $X \sim N(\mu;\ \sigma)$ e, portanto,

$$p_x^* = P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Como se desconhecem μ e σ =σ, estes têm que ser estimados:

$$\hat{\mu} = \bar{x} = \sum_{i=1}^{K} \frac{O_i x_i'}{n} = 16{,}9 \quad \text{e} \quad \hat{\sigma} = s = \sqrt{\sum_{i=1}^{K} \frac{O_i\left(x_i' - \bar{x}\right)^2}{n-1}} = 3{,}04.$$

E assim,

$$p_x^* = P(X \leq x) = \Phi\left(\frac{x-16{,}9}{3{,}04}\right).$$

$$p^*_{<12} = P(X < 12) = \Phi\left(\frac{12-16{,}9}{3{,}04}\right) = \Phi(-1{,}612) = 1 - \Phi(1{,}612) = 0{,}0535;$$

$$p^*_{[12;\ 14[} = P(12 \leq X < 14) = \Phi\left(\frac{14-16{,}9}{3{,}04}\right) - \Phi\left(\frac{12-16{,}9}{3{,}04}\right) = \Phi(-0{,}954) - \Phi(-1{,}612) = 0{,}1166;$$

$$p^*_{[14;\ 16[} = P(14 \leq X < 16) = \Phi\left(\frac{16-16{,}9}{3{,}04}\right) - \Phi\left(\frac{14-16{,}9}{3{,}04}\right) = \Phi(-0{,}296) - \Phi(-0{,}954) = 0{,}2135;$$

$$p^*_{[16;\ 18[} = P(16 \leq X < 18) = \Phi\left(\frac{18-16{,}9}{3{,}04}\right) - \Phi\left(\frac{16-16{,}9}{3{,}04}\right) = \Phi(0{,}362) - \Phi(-0{,}296) = 0{,}2577;$$

$$p^*_{[18;\ 20[} = P(18 \leq X < 20) = \Phi\left(\frac{20-16{,}9}{3{,}04}\right) - \Phi\left(\frac{18-16{,}9}{3{,}04}\right) = \Phi(1{,}020) - \Phi(0{,}362) = 0{,}2048;$$

$$p^*_{[20;\ 22[} = P(20 \leq X < 22) = \Phi\left(\frac{22-16{,}9}{3{,}04}\right) - \Phi\left(\frac{20-16{,}9}{3{,}04}\right) = \Phi(1{,}678) - \Phi(1{,}020) = 0{,}1072$$

$$p^*_{[22;\ +\infty[} = P(X \geq 22) = P(X < 22) = 1 - \Phi(1{,}678) = 0{,}0467$$

Partículas (µg/m3)	o_i	p_i^*	$e_i = np_i^*$	$o_i - e_i$	$\frac{(o_i - e_i)^2}{e_i}$
[0; 12[	7	0,0535	13,3745	-6,3745	3,0382
[12; 14[	22	0,1166	29,1392	-7,1392	1,7491
[14; 16[	55	0,2135	53,3850	1,6150	0,0489
[16; 18[	81	0,2577	64,4175	16,5825	4,2687
[18; 20[	53	0,2048	51,2021	1,7979	0,0631
[20; 22[	24	0,1072	26,8043	-2,8043	0,2934
[22; +∞[	8	0,0467	11,6774	-3,6774	1,1581
Total	n = 250	1,0000	250		$\chi^2_{obs} = 10{,}6195$

$K = 7; p = 2;\ \alpha = 0{,}1.$

Todos os $e_i \geq 5$, logo são verificados os pressupostos para a aplicação do teste de ajustamento do Qui-quadrado.

$\chi^2_{K-p-1;\,1-\alpha} = \chi^2_{4;\,0,9} = 7{,}779$. Logo, $R.A.:]-\infty;\ 7{,}779[$ e $R.R.: [7{,}779; +\infty[$.

Como $\chi^2_{obs} = 10{,}6195 \in R.R.$, rejeitar H_0 para $\alpha = 10\%$. Ao nível de significância de 10%, as concentrações de partículas totais em suspensão não provêm de uma população com distribuição Normal.

(SPSS) Analyse → Descriptive Statistics → Frequencies…
(Variable: Particulas; Statistics → ☑ Mean; ☑ Std. deviation)

Statistics

Partículas

N	Valid	250
	Missing	0
Mean		16,90
Std. Deviation		3,040

(SPSS) Analyse → Nonparametric Tests → Legacy Dialogs → Chi-square…
(Test Variable List: Particulas; Expected Range: ⊙ Get from data; Expected Values: ⊙ Values: 0,0535; 0,1166; ...)

Chi-Square Test

Partículas

	Observed N	Expected N	Residual
[0; 12[	7	13,4	-6,4
[12; 14[	22	29,2	-7,2
[14; 16[	55	53,4	1,6
[16; 18[	81	64,4	16,6
[18; 20[	53	51,2	1,8
[20; 22[	24	26,8	-2,8
[22; +inf[	8	11,7	-3,7
Total	250		

Test Statistics

	Partículas
Chi-Square	10,619[a]
df	6
Asymp. Sig.	,101

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 11,7.

Observação: Como no SPSS não é possível indicar que foram estimados 2 parâmetros (média e desvio padrão), tanto os graus de liberdade como o valor p apresentados no segundo quadro estão incorretos.

10.7.2 Teste de Kolmogorov-Smirnov

Pesaram-se 10 alunos, escolhidos aleatoriamente, do curso de Ciências do Desporto, tendo-se obtido os seguintes valores, em kg:

55 65 79 78 65 90 65 58 60 80

Teste ao nível de significância de 5% se os pesos provêm de uma população com distribuição:

a) Normal com média 65 kg e desvio padrão 10 kg.
b) Normal.

Resolução:

Seja X a v. a. que representa o peso, em kg, dos alunos.

a) $H_0: X \sim N(\mu = 65;\ \sigma = 10)$ vs $H_1: X$ não tem distribuição $N(\mu = 65;\ \sigma = 10)$.

Estatística de teste:

$$D = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|.$$

$x_{i:n}$	$F_n(x_{i:n})$	$F_0(x_{i:n})$	$\lvert F_n(x_{i:n}) - F_0(x_{i:n})\rvert$
55	0,1	0,1587	0,0587
58	0,2	0,2420	0,0420
60	0,3	0,3085	0,0085
65	0,4	0,5000	0,1000
65	0,5	0,5000	0,0000
65	0,6	0,5000	0,1000
78	0,7	0,9032	0,2032
79	0,8	0,9192	0,1192
80	0,9	0,9332	0,0332
90	1	0,9938	0,0062

$n = 10$ e $d_{10;\,0,95} = 0{,}409$, logo $R.A.: [0; 0{,}409[$ e $R.R.: [0{,}409; +\infty[$.

Como $d_{obs} = 0{,}2032 \in R.A.$, não rejeitar H_0 para $\alpha = 5\%$. Portanto, ao nível de significância de 5%, não existe evidência de que os pesos amostrais não provenham de uma população com distribuição Normal com média 65 kg e desvio padrão 10 kg.

Notas:

- Esta amostra tinha inicialmente 50 valores: aqui apenas se utilizaram os primeiros dez por simplicidade de cálculo. Se a dimensão desta amostra fosse reduzida, dever-se-ia optar pelo teste de Shapiro-Wilks (conforme referido anteriormente).
- Não é possível realizar no SPSS o teste de Kolmogorov-Smirnov quando as hipóteses estão completamente especificadas, i. e., não é possível definir os parâmetros da distribuição.

b) $H_0: X$ não tem distribuição Normal *vs* $H_1: X$ não tem distribuição Normal.

Estatística de teste:

$$D = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|.$$

Para realizar o teste é preciso especificar completamente a distribuição, ou seja, é necessário indicar qual a média e o desvio padrão. Para tal os parâmetros vão estimados a partir dos dados amostrais:

$$\hat{\mu} = \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = 69{,}5 \quad \text{e} \quad \hat{\sigma} = s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} = 11{,}482.$$

Dado que neste caso é necessário proceder à correção de Lilliefors no teste de Kolmogorov--Smirnov, uma vez que foi necessário estimar os parâmetros da distribuição Normal, a obtenção dos resultados será efetuada apenas com recurso ao programa SPSS.

(SPSS) Analyse → Nonparametric Tests → Legacy Dialogs → 1-Sample K-S...
(Test Variable List: Peso; Test Distribution: ☑ Normal)

One-Sample Kolmogorov-Smirnov Test

		Peso (em kg)
N		10
Normal Parameters[a,b]	Mean	69,50
	Std. Deviation	11,482
Most Extreme Differences	Absolute	,252
	Positive	,252
	Negative	-,170
Test Statistic		,252
Asymp. Sig. (2-tailed)		,070[c]

a. Test distribution is Normal.
b. Calculated from data.
c. Lilliefors Significance Correction.

Como $d_{obs} = 0{,}252$ (test Statistic) e valor $p = 0{,}07$, não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência de que os pesos não provenham de uma população com distribuição Normal.

10.7.3 Teste de Shapiro-Wilk

Pesaram-se 10 alunos, escolhidos aleatoriamente, do curso de Ciências do Desporto, tendo-se obtido os seguintes valores, em kg:

55 65 79 78 65 90 65 58 60 80

Teste ao nível de significância de 5% se os pesos provêm de uma população com distribuição Normal.

Resolução:

Seja X a v. a. que representa o peso, em kg, dos alunos.

H_0: X não tem distribuição Normal *vs* H_1: X não tem distribuição Normal.

A obtenção dos resultados será efetuada apenas com recurso ao programa SPSS.

(SPSS) Analyse → Descriptive Statistics → Explore...
(Dependent list: Peso; ⊙ Plots;
Plots → ☑ Normality plots with tests)

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Pesos (em kg)	,252	10	,070	,917	10	,333

a. Lilliefors Significance Correction

Como $w_{obs} = 0{,}917$ (Statistic) e valor $p = 0{,}333$, não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe evidência de que os pesos não provenham de uma população com distribuição Normal.

10.7.4 Teste de independência do Qui-quadrado

1. Com o objetivo de participarem numa dada atividade social, os estudantes de uma escola, foram submetidos a dois testes: um psicotécnico e um sobre regras de conduta. Obtiveram-se os seguintes resultados:

	Psicotécnico	
Regras de conduta	Aprovado	Reprovado
Aprovado	54	73
Reprovado	47	167

Considera que existe relação entre os resultados obtidos nos dois testes (utilize α = 5%)?

Resolução:

Sejam:

- X a v. a. que representa o resultado do teste sobre regras de conduta,
- Y a v. a. que representa o resultado do teste psicotécnico.

H_0: Não existe relação entre os resultados obtidos nos dois testes *vs*
H_1: Existe relação entre os resultados obtidos nos dois testes

$\Leftrightarrow H_0$: X e Y são independentes *vs* H_1: X e Y não são independentes.

$L = 2; C = 2$ e $\alpha = 0{,}05$.

Estatística de teste (tabela 2×2):

$$\chi^2 = \frac{n(|O_{11}O_{22} - O_{12}O_{21}| - 0{,}5n)^2}{O_{1.}O_{2.}O_{.1}O_{.2}} \sim \chi^2_{(L-1)(C-1)=1}.$$

o_{ij}	Psicotécnico		
Regras de conduta	Aprovado	Reprovado	Totais
Aprovado	54	73	127
Reprovado	47	167	214
Totais	101	240	$n = 341$

Todos os $e_{ij} \geq 5$, logo são verificados os pressupostos para a aplicação do teste de independência do Qui-quadrado.

$$\chi^2_{obs} = \frac{341(|54 \times 167 - 73 \times 47| - 0{,}5 \times 341)^2}{127 \times 214 \times 101 \times 240} = 11{,}186.$$

$\chi^2_{(L-1)(C-1);\,1-\alpha} = \chi^2_{1;\,0{,}95} = 3{,}841$, logo $R.A.: [0;\ 3{,}841[$ e $R.R.: [3{,}841; +\infty[$.

Como $\chi^2_{obs} \in R.R.$, rejeitar H_0 para $\alpha = 5\%$. Portanto, ao nível de significância de 5%, existe relação entre os resultados obtidos nos dois testes.

(SPSS) Analyse → Descriptive Statistics → Crosstabs…
(Row(s): Conduta; Column(s): Psicotécnico; Statistics → ☑ Chi-square; Cells → Counts: ☑ Observed; ☑ Expected)

Regras de conduta * Psicotécnico Crosstabulation

			Psicotécnico		
			Aprovado	Reprovado	Total
Regras de conduta	Aprovado	Count	54	73	127
		Expected Count	37,6	89,4	127,0
	Reprovado	Count	47	167	214
		Expected Count	63,4	150,6	214,0
Total		Count	101	240	341
		Expected Count	101,0	240,0	341,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	16,157[a]	1	,000		
Continuity Correction[b]	15,186	1	,000		
Likelihood Ratio	15,863	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	16,110	1	,000		
N of Valid Cases	341				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 37,62.
b. Computed only for a 2x2 table

Nas tabelas 2×2, os resultados do teste de independência do Qui-quadrado lêem-se na segunda linha (*Continuity Correction*). Assim, $\chi^2_{obs} = 15{,}186$, ao qual está associado um valor $p < 0{,}001$. Portanto, aos níveis usuais de significância (1%, 5% e 10%) rejeita-se H_0, donde se conclui que existe relação entre os resultados obtidos nos dois testes.

2. Os dados apresentados na tabela de contingência, são o resultado de um inquérito acerca da opinião dos consumidores, residentes na capital e no interior, sobe um novo detergente. Teste, ao nível de 5%, a hipótese de que não há divergência de opinião, entre os consumidores do interior e da capital, no que se refere à qualidade do produto.

Região	Bom	Regular	Insatisfatório
Capital	20	16	9
Interior	30	17	8

Resolução:

Sejam:

- X a v. a. que representa a região,
- Y a v. a. que representa a opinião sobre a qualidade do produto.

H_0: X e Y são independentes *vs* H_1: X e Y não são independentes.

$L = 2; C = 3$ e $\alpha = 0{,}05$.

Estatística de teste:

$$\chi^2 = \sum_{i=1}^{L}\sum_{j=1}^{C} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(L-1)(C-1)}.$$

$$e_{11} = \frac{45 * 50}{100} = 22{,}5; \qquad e_{12} = \frac{45 * 33}{100} = 14{,}85; \qquad e_{13} = \frac{45 * 17}{100} = 7{,}65;$$

$$e_{21} = \frac{55 * 50}{100} = 100; \qquad e_{22} = \frac{55 * 33}{100} = 18{,}15; \qquad e_{23} = \frac{55 * 17}{100} = 9{,}35.$$

e_{ij}:

Região	Bom	Regular	Insatisfatório	Total
Capital	22,5	14,85	7,65	45
Interior	27,5	18,15	9,35	55
Total	50	33	17	100

Todos os $e_{ij} \geq 5$, logo são verificados os pressupostos para a aplicação do teste de independência do Qui-quadrado.

$$\chi^2_{obs} = \frac{(20-22{,}5)^2}{22{,}5} + \frac{(16-14{,}85)^2}{14{,}85} + \frac{(9-7{,}65)^2}{18{,}15} + \frac{(30-27{,}5)^2}{27{,}5} + \frac{(17-18{,}15)^2}{7{,}65} + \frac{(8-9{,}35)^2}{9{,}35} = 1{,}1.$$

$\chi^2_{(L-1)(C-1);\,1-\alpha} = \chi^2_{2;\,0{,}95} = 5{,}991$, logo $[0;\ 5{,}991[$ e $R.R.: [5{,}991; +\infty[$.

Como $\chi^2_{obs} \in R.A.$, não rejeitar H_0, para $\alpha = 5\%$. Portanto, ao nível de significância de 5%, a opinião sobre a qualidade do produto é independente da zona de residência.

(SPSS) Analyse → Descriptive Statistics → Crosstabs…
(Row(s): Região; Column(s): Opinião; Statistics → ☑ Chi-square; Cells → Counts: ☑ Observed; ☑ Expected)

Região de residência * Opinião Crosstabulation

			Psicotécnico			
			Bom	Regular	Insatisfatório	Total
Regras de conduta	Capital	Count	20	16	9	45
		Expected Count	22,5	14,9	7,7	45,0
	Interior	Count	30	17	8	55
		Expected Count	27,5	18,2	9,4	55,0
Total		Count	50	33	17	100
		Expected Count	50,0	33,0	17,0	100,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1,100[a]	2	,577
Likelihood Ratio	1,101	2	,577
Linear-by-Linear Association	1,057	1	,304
N of Valid Cases	100		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 7,65.

Os resultados do teste de ajustamento lêem-se na primeira linha (*Pearson Chi-Square*), sendo apresentado na coluna *Value* o valor de χ^2_{obs}, *df* os graus de liberdade e na última coluna o valor p. Como o valor $p = 0{,}577$ não se rejeita H_0 para $\alpha < 0{,}577$.

10.7.5 Teste de correlação ordinal de Spearman

1. Um grupo de 6 professores de estatística foi avaliado segundo a sua capacidade pedagógica pelo coordenador da disciplina e por um grupo de estudantes. Os resultados obtidos foram:

Professor	A	B	C	D	E	F
Nota do coordenador	9,4	9,7	9,6	9,9	9,1	9,0
Nota dos estudantes	9,5	9,7	9,9	9,8	9,2	9,3

Ao nível de significância de 5%, considera que existe correlação significativa entre as duas variáveis?

Resolução:

Sejam:

- X a v. a. que representa a nota do coordenador,
- Y a v. a. que representa a nota dos estudantes.

Com $\alpha = 5\%, \rho \neq 0$ (existe correlação)?

H_0: $\rho_S = 0$ vs H_1: $\rho_S \neq 0$.

Estatística de teste $(n = 6)$:

$$R = R_s$$

Professor	x_i	y_i	o_{x_i} (ordem x)	o_{y_i} (ordem y)	$d_i = o_{x_i} - o_{y_i}$	d_i^2
A	9,4	9,5	3	3	0	0
B	9,7	9,7	5	4	1	1
C	9,6	9,9	4	6	2	4
D	9,9	9,8	6	5	1	1
E	9,1	9,2	2	1	1	1
F	9,0	9,3	1	2	-1	1

$$\sum_{i=1}^{5} d_i^2 = 8$$

Coeficiente de correlação de Spearman:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 * 8}{6(6^2-1)} = 1 - 0{,}2286 = 0{,}7714.$$

Consultando a tabela, $r_{n;1-\frac{\alpha}{2}} = r_{6;0,975} = 0{,}886$. Logo $R.A.:]-0{,}886;\ 0{,}886[$ e $R.R.:]-\infty;\ -0{,}886] \cup [0{,}886;\ +\infty[$.

Como $r_{obs} = r_s = 0{,}7714 \in R.A.$, não rejeitar H_0 para $\alpha = 5\%$. Portanto, ao nível de significância de 5%, não existe relação significativa entre as duas variáveis.

(SPSS) Analyse → Correlate → Bivariate…
(Variables: Nota_coordenador, Nota_dos_estudantes; Correlation Coefficiens: ☑ Spearman; Test of Significance: ☑ Two-tailed)

Correlations

			Nota do coordenador	Nota dos estudantes
Spearman's rho	Nota do coordenador	Correlation Coefficient	1,000	,771
		Sig. (2-tailed)	.	,072
		N	6	6
	Nota dos estudantes	Correlation Coefficient	,771	1,000
		Sig. (2-tailed)	,072	.
		N	6	6

Como o valor $p = 0{,}072$ não se rejeita H_0 para $\alpha < 0{,}072$.

2. Pretende-se avaliar a relação entre o tempo de resolução de puzzles e a aptidão para o raciocínio matemático. Para tal, pediu-se a um grupo de 10 alunos do 1º ciclo para resolver um determinado puzzle. Na tabela seguinte apresentam-se os resultados obtidos para cada um dos alunos relativamente ao *tempo* de resolução do *puzzle* e a *nota* obtida em matemática:

Aluno	A	B	C	D	E	F	G	H	I	J
Tempo	10	15	40	30	20	35	13	25	9	30
Nota	Muito Bom	Bom	Muito mau	Mau	Satisfaz	Mau	Bom	Satisfaz	Muito bom	Mau

Ao nível de significância de 1%, existe correlação negativa significativa entre as duas variáveis?

Resolução:

Sejam:

- X a v. a. que representa o tempo de resolução do puzzle,
- Y a v. a. que representa a nota a matemática (1 – muito mau, 2 – mau, 3 – satisfaz, 4 – bom, 5 – muito bom).

Para $\alpha = 1\%, \rho_s < 0$ (existe correlação negativa)?

H_0: $\rho_S \geq 0$ vs H_1: $\rho_S < 0$ (Teste unilateral esquerdo).

Estatística de teste $(n = 10)$:

$$T = \frac{R_s}{\sqrt{\frac{1 - R_s^2}{n-2}}} \sim t_{n-2}.$$

Aluno	Tempo x_i	Nota y_i	o_{x_i} (ordem x)	o_{y_i} (ordem y)	$d_i = o_{x_i} - o_{y_i}$	d_i^2
1	10	5 (muito bom)	2	9,5	-7,5	56,3
2	15	4 (bom)	4	7,5	-3,5	12,3
3	40	1 (muito mau)	10	1	9	81
4	30	2 (mau)	7,5	3	4,5	20,3
5	20	3 (satisfaz)	5	5,5	-0,5	0,3
6	35	2 (mau)	9	3	6	36
7	13	4 (bom)	3	7,5	-4,5	20,3
8	25	3 (satisfaz)	6	5,5	0,5	0,3
9	9	5 (muito bom)	1	9,5	-8,5	72,3
10	30	2 (mau)	7,5	3	4,5	20,3

$$\sum_{i=1}^{10} d_i^2 = 319$$

Coeficiente de correlação de Spearman:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 319}{10(10^2-1)} = -0{,}933.$$

$\alpha = 10\%, t_{n-2;1-\alpha} = t_{8;0,9} = 1{,}397$, logo $R.A.:]-1{,}397; 1]$ e $R.R.: [-1;\ -1{,}397]$.

$$t_{obs} = \frac{-0{,}933}{\sqrt{\frac{1 - 0{,}933^2}{10-2}}} = -7{,}353$$

Como $t_{obs} = -7{,}353 \in R.R.$, rejeitar H_0 para $\alpha = 10\%$. Portanto, ao nível de significância de 10%, pode-se considerar que existe correlação negativa entre as duas variáveis, i. e., os alunos que demoram pouco tempo a resolver os puzzles são os que têm notas mais elevadas a matemática, e vice-versa.

(SPSS) Analyse → Correlate → Bivariate…
(Variables: Tempo, Nota; Correlation Coefficiens: ☑ Spearman; Test of Significance: ☑ One-tailed)

Correlations

			Tempo na resolução do puzzle	Nota de matemática
Spearman's rho	Nota do coordenador	Correlation Coefficient	1,000	-,982**
		Sig. (2-tailed)	.	,000
		N	10	10
	Nota dos estudantes	Correlation Coefficient	-,982**	1,000
		Sig. (2-tailed)	,000	.
		N	10	10

**. Correlation is significant at the 0.01 level (1-tailed).

Como o valor $p < 0{,}001$ rejeita-se H_0 aos níveis usuais de significância.

10.7.6 Testes dos Sinais e de Wilcoxon

O departamento de qualidade de um determinado hospital, pretende fazer um estudo sobre o tempo que os utentes encaminhados para cirurgia pelos seus médicos de família demoram a ser efetivamente operados. Estes utentes têm que primeiro fazer uma consulta de referência no hospital (X dias após o encaminhamento) e só posteriormente são operados (Y dias após a consulta de referência). Escolheram-se aleatoriamente 15 utentes nestas condições, tendo sido reportados os seguintes tempos de espera:

x	69	76	51	34	62	13	40	17	64	41	54	36	50	34	44
y	28	64	7	26	38	18	40	20	44	32	31	32	36	25	73

a) Ao nível de significância de 5%, pode concluir-se que metade dos utentes teve um tempo de espera até à consulta de referência superior a 50 dias?
b) Admitindo que a distribuição dos tempos de espera é simétrica, qual a resposta à alínea anterior?
c) Ao nível de significância de 5%, pode afirmar-se que existe diferença entre os tempos de espera entre o encaminhamento e a consulta de referência, e entre a consulta de referência e a cirurgia?
d) Admitindo que a distribuição da diferença entre os tempos de espera é simétrica, qual a resposta à alínea anterior?

Resolução:

Sejam:
- X a v. a. que representa o tempo de espera entre o encaminhamento e a consulta de referência (em dias),
- Y a v. a. que representa o tempo de espera entre a consulta de referência e a cirurgia (em dias).

$n = 15$.

a) $\alpha = 5\%,\ \tilde{\mu} > 50$?

$H_0: \tilde{\mu} \leq 50$ vs $H_1: \tilde{\mu} > 50$ (Teste unilateral direito)
$\Leftrightarrow H_0: P(+) \leq P(-)$ vs $H_1: P(+) > P(-)$.

Recorre-se ao teste dos Sinais uma vez que nada se sabe quanto à distribuição das classificações.

Estatística de teste (amostra pequena):

$$S^+ = \text{Número total de sinais} + \text{que ocorrem} \sim B(n'; 0{,}5).$$

x_i	69	76	51	34	62	13	40	17	64	41	54	36	50	34	44
$d_i = x_i - 50$	19	26	1	-16	12	-37	-10	-33	14	-9	4	-14	0	-16	-6
Sinal	+	+	+	–	+	–	–	–	+	–	+	–		–	–

Há 1 empate, logo $n' = n -$ n.º de empates $= 15 - 1 = 14$.

$s_{obs}^{+} = 6$.

Determinação de $b_{1-\alpha} = b_{0,95}$ sabendo que $S^{+} \sim B(n' = 14; 0{,}5)$, i. e., o menor inteiro a que verifica

$$P(S^{+} \geq a) = \sum_{x=a}^{14} {}^{14}C_x 0,5^{14} \leq \alpha = 0{,}05.$$

$P(S^{+} \geq 14) = P(S^{+} = 14) = {}^{14}C_{14} 0,5^{14} = 0{,}0001 \leq \alpha = 0{,}05;$

$$P(S^{+} \geq 13) = \sum_{x=13}^{14} {}^{14}C_x 0,5^{14} = 0{,}0009 \leq \alpha = 0{,}05;$$

$$P(S^{+} \geq 12) = \sum_{x=12}^{14} {}^{14}C_x 0,5^{14} = 0{,}0065 \leq \alpha = 0{,}05;$$

$$P(S^{+} \geq 11) = \sum_{x=11}^{14} {}^{14}C_x 0,5^{14} = 0{,}0287 \leq \alpha = 0{,}05;$$

$$P(S^{+} \geq 10) = \sum_{x=10}^{14} {}^{14}C_x 0,5^{14} = 0{,}0898 > \alpha = 0{,}05 \Longrightarrow \text{STOP: } b_{0,95} = 11.$$

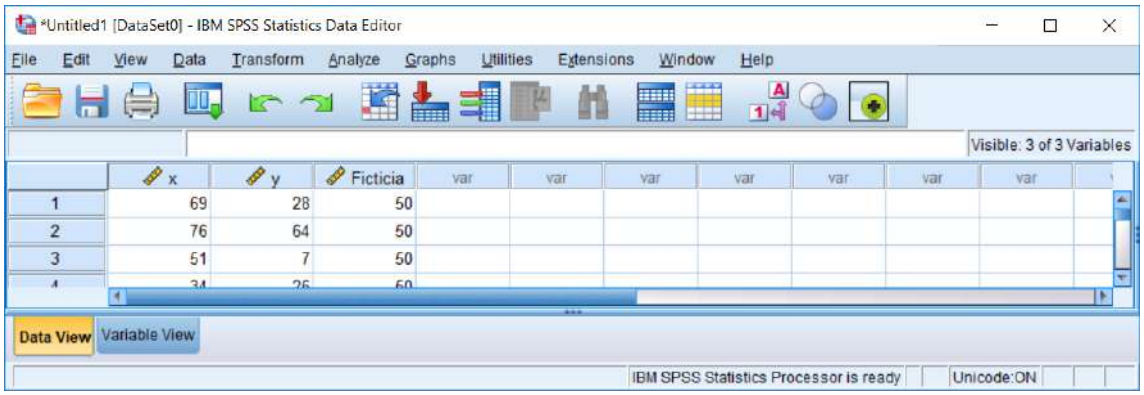
Logo, $R.A.: \{0, 1, 2, \ldots, 10\}$ e $R.R.: \{11, 12, 13, 14\}$.

Como $s_{obs}^{+} = 8 \in R.A.$ não rejeitar H_0.

Em alternativa:
valor $p = P(S^{+} \geq s_{obs}^{+}) = P(S^{+} \geq 6) = 1 - P(S^{+} \leq 5) = 1 - 0{,}3953 = 0{,}6047$.
Logo rejeitar H_0 quando $\alpha \geq 60{,}47\%$.

Portanto, ao nível de significância de 5%, metade dos utentes esperou no máximo 50 dias até à consulta de referência.

(SPSS) Criar uma variável fictícia com o valor da mediana a testar.



(SPSS) Analyze → Nonparametric Tests → Legacy Dialogs → 2 Related Samples…
(Variable 1: x; Variable 2: Ficticia; Test Type: ☑ Sign;
Exact → ⊙ Exact)

Sign Test
Frequencies

		N
x - Ficticia	Negative Differences[a]	6
	Positive Differences[b]	8
	Ties[c]	1
	Total	15

a. Ficticia < x
b. Ficticia > x
c. Ficticia = x

Test Statistics[a]

	Ficticia - x
Exact Sig. (2-tailed)	,791[b]
Exact Sig. (1-tailed)	,395
Point Probability	,183

a. Sign Test
b. Binomial distribution used.

b) $\alpha = 5\%,\ \tilde{\mu} > 50$?

$H_0: \tilde{\mu} \leq 50$ vs $H_1: \tilde{\mu} > 50$ (Teste unilateral direito)

Recorre-se ao teste de Wilcoxon uma vez que distribuição das classificações é simétrica.

Estatística de teste (amostra pequena):

$$W = \text{Soma das ordens dos } |D_i| \text{ com sinal } +$$

x_i	69	76	51	34	62	13	40	17	64	41	54	36	50	34	44
$d_i = x_i - 50$	19	26	1	-16	12	-37	-10	-33	14	-9	4	-14	0	-16	-6
$\lvert d_i \rvert$	19	26	1	16	12	37	10	33	14	9	4	14	0	16	6
sinal	+	+	+	–	+	–	–	–	+	–	+	–		–	–

$\lvert d_i \rvert$ ordenados	1	4	6	9	10	12	14	14	16	16	19	26	33	37
Ordem	1	2	3	4	5	6	7,5	7,5	9,5	9,5	11	12	13	14
Sinal	+	+	–	–	–	+	+	–	–	–	+	+	–	–

Há 1 empate, logo $n' = n - \text{n.º de empates} = 15 - 1 = 14$.

Pela tabela, $w_{n';\,1-\alpha} = w_{14;\,0,95} = 79$. Logo, $R.A.: \{0, 1, 2, \ldots, 78\}$ e $R.R.: \{79, \ldots, 105\}$, pois $\frac{n'(n'+1)}{2} = 105$.

Como $w_{obs} = 1 + 2 + 6 + 7{,}5 + 11 + 12 = 39{,}5 \in R.A.$, então não rejeitar H_0. Portanto, ao nível de significância de 5%, metade dos utentes esperou no máximo 50 dias até à cirurgia (tempo global).

(SPSS) Criar uma variável fictícia com o valor da mediana a testar (ver alínea anterior).
Analyze → Nonparametric Tests → Legacy Dialogs → 2 Related Samples…
(Variable 1: Ficticia; Variable 2: x; Test Type: ☑ Wilcoxon;
Exact → ⊙ Exact)

Wilcoxon Signed Ranks Test
Ranks

		N	Mean Rank	Sum of Ranks
x - Ficticia	Negative Ranks	8[a]	8,19	65,50
	Positive Ranks	6[b]	6,58	39,50
	Ties	1[c]		
	Total	15		

a. x < Ficticia
b. x > Ficticia
c. x = Ficticia

Test Statistics[a]

	x - Ficticia
Z	-,816[b]
Asymp. Sig. (2-tailed)	,414
Exact Sig. (2-tailed)	,435
Exact Sig. (1-tailed)	,217
Point Probability	,009

a. Wilcoxon Signed Ranks Test
b. Based on positive ranks.

Observação: Para amostras de grande dimensão z_{obs} é apresentado na linha Z e o valor p correspondente, para um teste bilateral, na linha *Asymp. Sig. (2-tailed)*, do segundo quadro.

c) $\alpha = 5\%,\ \tilde{\mu}_1 \neq \tilde{\mu}_2$?
$H_0: \tilde{\mu}_1 = \tilde{\mu}_2$ *vs* $H_1: \tilde{\mu}_1 \neq \tilde{\mu}_2$ (Teste bilateral)
$\Leftrightarrow H_0: P(+) = P(-)$ *vs* $H_1: P(+) \neq P(-)$
$\Leftrightarrow H_0: p = 0{,}5$ *vs* $H_1: p \neq 0{,}5$.

Recorre-se ao teste dos Sinais uma vez que as duas amostras são emparelhadas e nada se sabe quanto à distribuição da diferença das classificações.

Estatística de teste (amostra pequena):

$$S^+ = \text{Número total de sinais} + \text{que ocorrem} \sim B(n'; 0{,}5).$$

x_i	69	76	51	34	62	13	40	17	64	41	54	36	50	34	44
y_i	28	64	7	26	38	18	40	20	44	32	31	32	36	25	73
$d_i = x_i - y_i$	41	12	44	8	24	-5	0	-3	20	9	23	4	14	9	-29
Sinal	+	+	+	+	+	–		–	+	+	+	+	+	+	–

Há 1 empate, logo $n' = n - \text{n.º de empates} = 15 - 1 = 14$.

$s^+_{obs} = 11$.

Determinação de $b_{\frac{\alpha}{2}} = b_{0{,}025}$ sabendo que $S^+ \sim B(n' = 14; 0{,}5)$, i. e., o maior inteiro a que verifica

$$P(S^+ \leq a) = \sum_{x=0}^{a} {}^{14}C_x 0{,}5^{14} \leq \frac{\alpha}{2} = 0{,}025.$$

$P(S^+ \leq 0) = P(S^+ = 0) = {}^{14}C_0 0{,}5^{14} = 0{,}0001 \leq \frac{\alpha}{2} = 0{,}025;$

$P(S^+ \leq 1) = \sum_{x=0}^{1} {}^{14}C_x 0{,}5^{14} = 0{,}0009 \leq \frac{\alpha}{2} = 0{,}025;$

$P(S^+ \leq 2) = \sum_{x=0}^{2} {}^{14}C_x 0{,}5^{14} = 0{,}0065 \leq \frac{\alpha}{2} = 0{,}025;$

$P(S^+ \leq 3) = \sum_{x=0}^{3} {}^{14}C_x 0{,}5^{14} = 0{,}0287 > \frac{\alpha}{2} = 0{,}025 \Longrightarrow \text{STOP: } b_{0{,}025} = 2.$

Determinação de $b_{1-\frac{\alpha}{2}} = b_{0{,}975}$ sabendo que $S^- \sim B(n' = 14; 0{,}5)$, i. e., o menor inteiro b que verifica

$$P(S^+ \geq b) = \sum_{x=b}^{14} {}^{14}C_x 0{,}5^{14} \leq \frac{\alpha}{2} = 0{,}025 \Leftrightarrow b_{1-\frac{\alpha}{2}} = n' - b_{\frac{\alpha}{2}}.$$

Mas como $b_{1-\frac{\alpha}{2}} = n' - b_{\frac{\alpha}{2}} \Leftrightarrow\ b_{0{,}975} = 14 - 2 = 12.$

Logo, $R.A.: \{3, 4, \ldots, 10, 11\}$ e $R.R.: \{0, 1, 2, 12, 13, 14\}$.

Como $s_{obs}^{+} = 3 \in R.A.$ não rejeitar H_0.

Em alternativa:

valor $p = 2 \times min\{P(S^+ \leq s_{obs}^{+});\ P(S^+ \geq s_{obs}^{+})\} = 2 \times min\{P(S^+ \leq 11);\ P(S^+ \geq 11)\}$
$= 2 \times min\{0{,}9935;\ 0{,}0287\} = 0{,}057.$

Logo rejeitar H_0 quando $\alpha \geq 5{,}74\%$.

Portanto, ao nível de significância de 5%, não existe diferença entre as classificações obtidas pelos alunos na parte teórica e na parte prática.

(SPSS) Analyze → Nonparametric Tests → Legacy Dialogs → 2 related samples
(Variable 1: y; Variable 2: x; Test Type: ☑ Sign;
Exact → ⊙ Exact)

Sign Test
Frequencies

		N
x - y	Negative Differences[a]	3
	Positive Differences[b]	11
	Ties[c]	1
	Total	15

a. x < y
b. x > y
c. x = y

Test Statistics[a]

	x - y
Exact Sig. (2-tailed)	,057[b]
Exact Sig. (1-tailed)	,029
Point Probability	,022

a. Sign Test
b. Binomial distribution used.

d) $\alpha = 5\%,\ \tilde{\mu}_1 \neq \tilde{\mu}_2$?

$H_0: \tilde{\mu}_1 = \tilde{\mu}_2$ *vs* $H_1: \tilde{\mu}_1 \neq \tilde{\mu}_2$ (Teste bilateral).

Recorre-se ao teste de Wilcoxon uma vez que as duas amostras são emparelhadas e a distribuição da diferença das classificações é simétrica.

Estatística de teste (amostra pequena):

$W =$ Soma das ordens dos $|D_i|$ com sinal $+$

x_i	69	76	51	34	62	13	40	17	64	41	54	36	50	34	44
y_i	28	64	7	26	38	18	40	20	44	32	31	32	36	25	73
$d_i = x_i - y_i$	41	12	44	8	24	-5	0	-3	20	9	23	4	14	9	-29
$\lvert d_i \rvert$	41	12	44	8	24	5		3	20	9	23	4	14	9	29
Sinal	+	+	+	+	+	–		–	+	+	+	+	+	+	–

$\lvert d_i \rvert$ ordenados	3	4	5	8	9	9	12	14	20	23	24	29	41	44
Ordem	1	2	3	4	5,5	5,5	7	8	9	10	11	12	13	14
Sinal	-	+	-	+	+	+	+	+	+	+	+	-	+	+

Há 1 empate, logo $n' = n -$ n.º de empates $= 15 - 1 = 14$.

Consultando a tabela, $w_{n';\frac{\alpha}{2}} = w_{14;\,0{,}025} = 21$ e $w_{n';\,1-\frac{\alpha}{2}} = w_{14;\,0{,}025} = 84$.

Logo, $R.A.: \{22, \dots, 83\}$ e $R.R.: \{0, 1, \dots, 21, 84, 85, \dots, 105\}$, pois $\frac{n\prime(n'+1)}{2} = 105$.

Como $w_{obs} = 2 + 4 + 5{,}5 + 5{,}5 + 7 + 8 + 9 + 10 + 11 + 13 + 14 = 89 \in R.R.$, então rejeitar H_0. Portanto, ao nível de significância de 5%, existe diferença entre as classificações obtidas pelos alunos na parte teórica e na parte prática.

(SPSS) Analyze → Nonparametric Tests → Legacy Dialogs → 2 Related Samples…
(Variable 1: Ficticia; Variable 2: x; Test Type: ☑ Wilcoxon;
Exact → ⊙ Exact)

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
x - y	Negative Ranks	3[a]	5,33	16,00
	Positive Ranks	11[b]	8,09	89,00
	Ties	1[c]		
	Total	15		

a. x < y
b. x > y
c. x = y

Test Statistics[a]

	x - y
Z	-2,292[b]
Asymp. Sig. (2-tailed)	,022
Exact Sig. (2-tailed)	,019
Exact Sig. (1-tailed)	,010
Point Probability	,001

a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.

Pela informação da segunda tabela, $valor\text{-}p = 0{,}019$. Logo, rejeitar H_0 quando $\alpha \geq 1{,}9\%$.

10.7.7 Teste de Mann-Whitney U

A disciplina de Estatística é lecionada a dois cursos, A e B, tendo os alunos realizado o mesmo exame. Recolheu-se uma amostra aleatória de 15 alunos de cada curso tendo-se observado as seguintes classificações no exame:

Turma A	69	76	51	34	62	13	40	7	64	41	64	26	40	44	48
Turma B	28	64	7	26	38	18	40	20	44	32	31	32	36	25	73

Ao nível de significância de 10%, pode afirmar-se que existe diferença entre as classificações dos dois cursos?

Resolução:

Sejam:

- X a v. a. que representa a classificação obtida pelos alunos do curso A,
- Y a v. a. que representa a classificação obtida pelos alunos do curso B.

$\alpha = 10\%,\ \tilde{\mu}_1 \neq \tilde{\mu}_2$?

$H_0: \tilde{\mu}_1 = \tilde{\mu}_2$ *vs* $H_1: \tilde{\mu}_1 \neq \tilde{\mu}_2$ (Teste bilateral).

Aplica-se o teste de Mann-Whitney pois as 2 amostras são independentes.

Estatística de teste ($n_1 = 15$ e $n_2 = 15$):

$$U = min\{U_1;\ U_2\},$$

com

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}, \qquad U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

e R_i a soma das ordens da amostra i, $i = 1, 2$.

Turma	Ordens															Soma
A	28	30	23	13	24	3	17	1,5	26	19	26	7,5	17	20,5	22	277,5
B	9	26	1,5	7,5	15	4	17	5	20,5	11,5	10	11,5	14	6	29	187,5

$$r_1 = 277{,}5;\ u_1 = 277{,}5 - \frac{15(15+1)}{2} = 157{,}5;$$

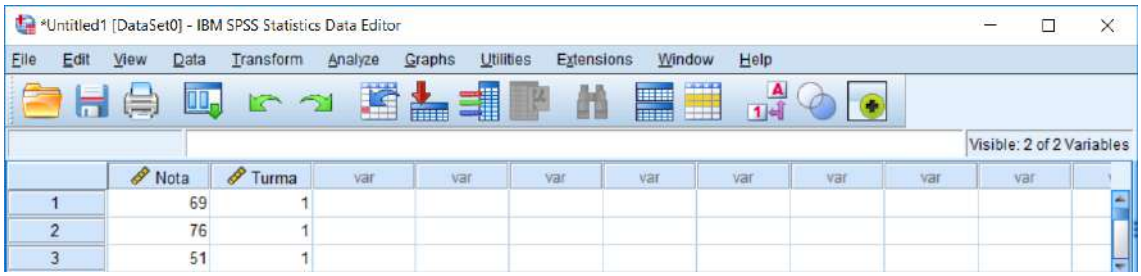
$$r_2 = 187{,}5;\ u_2 = 187{,}5 - \frac{15(15+1)}{2} = 67{,}5.$$

Logo $u_{obs} = min\{u_1;\ u_2\} = 67{,}5$.

Pela tabela $u_{n_1;\, n_2;\, \frac{\alpha}{2}} = u_{15;\, 15;\, 0{,}05} = 72$ e $u_{n_1;\, n_2;\, 1-\frac{\alpha}{2}} = n_1 n_2 - u_{n_1;\, n_2;\, \frac{\alpha}{2}} = 15 \times 15 - 72 = 153$. Logo, $R.A.: \{73, \ldots, 152\}$ e $R.R.: \{0, \ldots, 72, 153, \ldots, 225\}$.

Como $u_{obs} = 67{,}5 \in R.R.$, rejeitar H_0. Portanto, ao nível de significância de 10%, existe diferença entre as classificações das duas turmas.

(SPSS)



(SPSS) Analyze → Nonparametric Tests → Legacy Dialogs 2 → 2 Independent Samples
(Test Variable List: Nota; Grouping Variable: Turma; Define Groups: Group 1: 1; Group 2: 2; Test Type: ☑ Mann-Whitney U;
Exact → ⊙ Exact)

Mann-Whitney Test
Ranks

	Turma	N	Mean Rank	Sum of Ranks
Classificação no exame	A	15	18,50	277,50
	B	15	12,50	187,50
	Total	30		

Test Statistics[a]

	Nota
Mann-Whitney U	67,500
Wilcoxon W	187,500
Z	-1,869
Asymp. Sig. (2-tailed)	,062
Exact Sig. [2*(1-tailed Sig.)]	,061[b]
Exact Sig. (2-tailed)	,062
Exact Sig. (1-tailed)	,031
Point Probability	,001

a. Grouping Variable: Turma
b. Not corrected for ties.

Pelo *output* fornecido pelo SPSS, o valor $p = 0{,}062$ (linha *Exact Sig. (2-tailed)* da segunda tabela) que nos indica para rejeitar H_0 quando $\alpha \geq 0{,}062$. Logo, para $\alpha = 10\%$ rejeitar H_0.
Observação: Para amostras de grande dimensão o valor observado para a estatística de teste é apresentado na linha Z e o valor p correspondente, para um teste bilateral, na linha *Asymp. Sig. (2-tailed)*.

10.7.8 Teste de Kuskall-Wallis

Com vista a comparar as larguras máximas dos crânios de homens egípcios de diferentes épocas, recolheu-se uma amostra aleatória de dimensão 9, de crânios datados de 4000 a. C., 1850 a. C. e 150 d. C., tendo-se obtido os seguintes resultados (adaptado de Triola, 2017):

4000 a. C.	131	138	125	129	132	135	132	134	138
1850 a. C.	129	134	136	137	137	129	136	138	134
150 d. C.	128	138	136	139	141	142	137	145	137

Ao nível de significância de 5%, teste a afirmação de que as três amostras provêm de populações idênticas. (Utilize o SPSS)

Resolução:

Sejam:
- X_1 a v. a. que representa a largura dos crânios datados de 4000 a. C.,
- X_2 a v. a. que representa a largura dos crânios datados de 1850 a. C.,
- X_3 a v. a. que representa a largura dos crânios datados de 150 d. C.

$\alpha = 5\%,\ \tilde{\mu}_1 = \tilde{\mu}_2 = \tilde{\mu}_3$?

H_0: Os 3 grupos provêm da mesma população ou de populações idênticas *vs*
H_1: Nem todos os 3 grupos provêm da mesma população ou de populações idênticas
$\Leftrightarrow H_0$: $\tilde{\mu}_1 = \tilde{\mu}_2 = \tilde{\mu}_3$ *vs* Nem todos os $\tilde{\mu}_i$ são iguais, i = 1, 2, 3 (Teste bilateral).

Aplica-se o teste de Kruskal-Wallis pois as 3 amostras são independentes.

Estatística de teste (não corrigida para empates):

$$\chi^2 \frac{12}{n(n+1)} \sum_{i=1}^{K} \frac{R_i^2}{n_i} - 3(n+1) \overset{\circ}{\sim} \chi^2_{K-1}.$$

Data	Ordens									Soma
4000 a. C.	6	21,5	1	4	7,5	12	7,5	10	21,5	91
1850 a. C.	4	10	14	17,5	17,5	4	14	21,5	10	112,5
150 d. C.	2	21,5	14	24	25	26	17,5	27	17,5	174,5

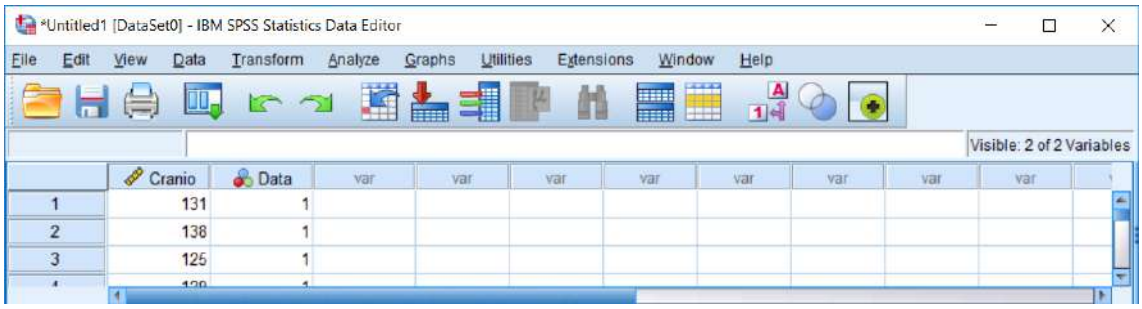
$$K = 3;\ n_1 = n_2 = n_3 = 9 \implies n = \sum_{i=1}^{3} n_i = 27.$$

$$\chi^2_{obs} = \frac{12}{27(27+1)}\left(\frac{91^2}{9} + \frac{112,5^2}{9} + \frac{174,5^2}{9}\right) - 3(27+1) = 6{,}631.$$

Pela tabela, $\chi^2_{K-1;\,1-\alpha} = \chi^2_{2;\,0,95} = 5{,}991$. Logo, $R.A.$: $[0;\ 5{,}991[$ e $R.R.$: $[5{,}991;\ +\infty[$.

Como $\chi^2_{obs} = 6{,}631 \in R.R.$, rejeitar H_0. Portanto, ao nível de significância de 5%, existe diferença entre as populações. As 4 amostras não são provenientes de populações idênticas.

(SPSS)



(SPSS) Analyze → Nonparametric Tests → Legacy dialogs → k Independent Samples…
(Test Variable List: Cranio; Grouping Variable: Data; Define Range: Minimum: 1; Maximum: 3; Test Type: ☑ Kruskal–Wallis H)

Kruskal-Wallis Test

Ranks

	Data	N	Mean Rank
Cranio	4000 a. C.	9	10,11
	1850 a. C.	9	12,50
	150 d. C.	9	19,39
	Total	27	

Test Statistics[a,b]

	Cranio
Kruskal-Wallis H	6,698
df	2
Asymp. Sig.	,035

a. Kruskal Wallis Test
b. Grouping Variable: Data

Observação: No SPSS a estatística de teste usada é a corrigida para empates, ao contrário da que foi calculada, daí que o resultado obtido seja diferente. Desta forma sempre que possível, deve-se usar os resultados devolvidos pelo SPSS. Neste caso, $\chi^2_{obs} = 6{,}698$ e valor $p = 0{,}035$, logo deve rejeitar-se H_0 quando $\alpha \geq 0{,}035$.

10.7.9 Teste à simetria e achatamento

Pesaram-se 10 alunos, escolhidos aleatoriamente, do curso de Ciências do Desporto, tendo-se obtido os seguintes valores, em kg:

55 65 79 78 65 90 65 58 60 80

Teste ao nível de significância de 5% se os pesos provêm de uma distribuição:

a) Simétrica
b) Mesocúrtica.

Resolução:

Seja X a v. a. que representa o peso, em kg, dos alunos.

a) $H_0: \gamma_1 = 0$ vs $H_1: \gamma_1 \neq 0$.

A obtenção dos resultados será efetuada apenas com recurso ao programa SPSS.

(SPSS) Analyse → Descriptive Statistics → Descriptives…
(Variables: Peso; ⊙ Plots;
Options → Characterize Posterior Distribution: ☑ Skewness)

Descriptive Statistics

	N	Skewness	
	Statistic	Statistic	Std. Error
Peso	10	,509	,687
Valid N (listwise)	10		

$g_a = 0{,}509;\ \ EP_{g_a} = 0{,}687;\ \ z_{obs} = \dfrac{0{,}509}{0{,}687} = 0{,}741.$

Pela tabela $z_{1-\frac{\alpha}{2}} = z_{0{,}975} = 1{,}96$. Logo, $R.A.:]-1{,}96;\ 1{,}96[$ e $R.R.:]-\infty; -1{,}96] \cup [1{,}96; +\infty[$.

Como $z_{obs} = 0{,}741 \in R.A.$, não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe não existe evidência de que os pesos amostrais não provenham de uma distribuição simétrica.

b) $H_0: \gamma_2 = 0$ *vs* $H_1: \gamma_2 \neq 0$.

A obtenção dos resultados será efetuada apenas com recurso ao programa SPSS.

(SPSS) Analyse → Descriptive Statistics → Descriptives…
(Variables: Peso; ⊙ Plots;
Options → Characterize Posterior Distribution: ☑ Kurtosis)

Descriptive Statistics

	N	Kurtosis	
	Statistic	Statistic	Std. Error
Peso	10	-,909	1,334
Valid N (listwise)	10		

$k_a = -0{,}909;\ \ EP_{k_a} = 1{,}334;\ \ z_{obs} = \dfrac{-0{,}909}{1{,}334} = -0{,}6814.$

Pela tabela $z_{1-\frac{\alpha}{2}} = z_{0{,}975} = 1{,}96$. Logo, $R.A.:]-1{,}96;\ 1{,}96[$ e $R.R.:]-\infty; -1{,}96] \cup [1{,}96; +\infty[$.

Como $z_{obs} = -0{,}6814 \in R.A.$, não rejeitar H_0. Portanto, ao nível de significância de 5%, não existe não existe evidência de que os pesos amostrais não provenham de uma distribuição mesocúrtica.

10.8 Exercícios propostos

1. Recolheu-se uma amostra aleatória de 3000 indivíduos de uma determinada população, tendo-se observado o seu grupo sanguíneo.

Grupo sanguíneo	A	B	AB	O
Frequência observada	1300	300	100	1300

Teste a hipótese de $p_A = 0{,}4$, $p_B = 0{,}15$, $p_{AB} = 0{,}05$ e $p_O = 0{,}4$, ao nível de significância de 1%.

2. No relatório mensal, publicado pela empresa de audimetria Contagem, relativo à análise das audiências durante o mês de abril afirma-se que a percentagem das audiências (share), por canal, nos sábados entre as 20:00 e as 21:00 foi:

Canal	TV1	TV2	TV3	TV4
Audiência em %	29%	7%	48%	16%

No início do mês de maio a TV1 reviu a sua programação dos sábados à noite. No final desse mês recolheu-se amostra aleatória de 300 lares, tendo-se obtido os seguintes resultados de audiência:

Canal	TV1	TV2	TV3	TV4
Audiência (n.º de pessoas)	95	10	150	45

Teste, ao nível se significância de 5%, se as proporções das audiências se alteraram após a revisão da programação efetuada pelo canal TV1.

3. Um jornalista desportivo, que acompanhou o percurso de preparação da seleção do país X para o Euro 2004, afirma que nos jogos 50% dos remates efetuados por esta equipa são à baliza, 20% para fora e os restantes são desviados.
Nos 2 jogos, entretanto já disputados por esta equipa, verificou-se que dos 34 remates que esta equipa fez 16 foram à baliza, 9 para fora e 9 foram desviados.
Com base nos resultados obtidos concorda com a afirmação do jornalista, ao nível de significância de 5%?

4. Numa determinada comunidade a autarquia pretende abrir ao público um parque didático. De forma a desenvolver um plano de afetação de pessoal, realizou-se um inquérito a 140 pessoas acerca da preferência sobre o dia mais provável para visitar o parque. Os resultados obtidos foram:

Dia mais provável de visita	Dia útil	Sábado	Domingo	Feriado
n.º de respostas	20	20	40	60

Deve o gestor do parque fazer a planificação do pessoal considerando que o número de visitantes é o mesmo todos os dias? Justifique a sua opinião com base num teste estatístico usando um nível de significância de 5%.

5. O registo do número de acidentes ocorridos numa fábrica durante 100 dias consta no quadro seguinte:

nº de acidentes	0	1	2	3	4	5	6	7
nº de dias	5	39	22	13	11	6	2	2

Teste a hipótese de o número de acidentes seguir uma lei de Poisson.

6. Um serviço de consultas externas de um hospital registou qual o número de utentes que, tendo consulta marcada acabaram por não fazer comparecer. Para 100 dias escolhidos aleatoriamente, os resultados foram os seguintes:

nº de ausências	0	1	2	3	4	5	6
nº de utentes	21	36	23	13	4	2	1

Teste, ao nível de significância de 5%, se a distribuição do número de ausências por utente segue uma distribuição de Poisson com média 1.

7. Numa escola realizou-se um teste de leitura a 82 crianças tendo-se obtido os seguintes resultados (valores entre 0 e 100):

resultado	35 – 45	45 – 55	55 – 65	65 – 75	75 – 85
n.º de crianças	12	26	31	11	2

Teste a hipótese dos dados amostrais provirem de uma população com distribuição Normal com $\mu = 55$ e $\sigma = 10$ (considere $\alpha = 5\%$).

8. Espera-se que numa determinada etapa do Mega Paris-Dakar se gaste cerca de 100 litros de combustível, em carros que concorrem na mesma categoria e com a mesma cilindrada. Crê-se que os desvios cometidos sigam distribuição Normal. Para testar esta hipótese efetuou-se a análise de 595 carros, tendo-se obtido os seguintes resultados:

Combustível (l)	[-∞; -3[	[-3; -1[	[-1; 0[	[0; 1[	[1; 3[	[3; +∞[
N.º de carros	10	95	200	190	90	10

Podem os dados ser considerados uma amostra aleatória de uma população com distribuição Normal (considere $\alpha = 10\%$)?

9. Os dados que se seguem referem-se ao comprimento (em cm) de um grupo de bebés prematuros (idade gestacional inferior a 36 semanas) nascidos durante um mês numa maternidade.

21,7	30	29,7	23,4	35,2	24,7	20,4	41,4	13,6	32,2
37,8	23,8	33,3	31,6	20,1	19,1	32,7	33,5	18,3	25,4
40,2	36,8	33,1	17,2	13,3	33,7	12,6	21,6	24,6	19,6
24,1	37,4	28,1	16,2	33,7	28,2	21	27,3	24,3	29,9

Ao nível de significância de 10%, teste a hipótese dos dados amostrais provirem de uma população com distribuição:

a) Normal com média 29 e desvio padrão 6.
b) Normal.

10. Num levantamento de opinião pública, foi realizado um inquérito a 1000 pessoas composto por duas questões:

- "É a favor da coincineração?"
- "Reside no distrito de Setúbal ou Coimbra?"

Na tabela seguinte apresentam-se os resultados obtidos:

Coincineração	Distrito de Setúbal	Distrito de Coimbra
A favor	20	300
Não é a favor	80	600

Teste a hipótese de não existir associação entre as respostas às duas questões ($\alpha = 10\%$).

11. Suspeita-se que o alumínio seja um fator que contribui para o desenvolvimento da doença de Alzheimer. Num estudo, investigadores compararam um grupo de pacientes com outro grupo, cuidadosamente selecionado, de pessoas com características comuns mas que não são portadoras daquela doença. O interesse do estudo reside na utilização de antiácidos que contêm alumínio. Cada indivíduo foi classificado de acordo com o uso desses antiácidos. Os resultados encontram-se na tabela abaixo.

	Antiácido contendo alumínio			
	Nenhum	Baixo	Médio	Alto
Com doença de Alzheimer	112	3	5	8
Grupo de Controle	114	9	3	2

Estará o uso de antiácidos contendo alumínio relacionado com a doença de Alzheimer?

12. Presume-se que determinado composto é eficaz no tratamento de constipações. Dois medicamentos que diferem pela quantidade daquele composto foram utilizados numa experiência em 164 indivíduos. Os resultados das reações observadas constam no quadro seguinte:

	Reação observada		
Medicamento	Melhorou	Piorou	Não teve efeito
A	50	10	22
B	44	12	26

Teste a hipótese de independência entre o resultado e o tipo de medicamento.

13. Realizou-se um teste de inflamabilidade para as roupas de dormir para crianças, através da queima de pedaços de tecido, sob determinadas condições. Seguidamente mediu-se e registou-se o comprimento da porção carbonizada. Este teste foi realizado com os mesmos tecidos em 2 laboratórios diferentes, tendo-se obtido os seguintes resultados:

Tecido	1	2	3	4	5	6	7	8	9	10
Lab. 1	2,9	3,1	3,1	3,7	3,1	4,2	3,7	3,9	3,2	3,1
Lab. 2	2,7	3,4	3,6	3,2	4	4,1	3,8	3,8	4,3	3,4

a) Como o tecido utilizado foi o mesmo, os resultados obtidos pelos 2 laboratórios deveriam ser os mesmos. Foi esse o caso (considere $\alpha = 10\%$).
b) Sabendo que o coeficiente de correlação amostral de Spearman é 0,514, considera que, ao nível de significância de 1%, existe relação entre as duas variáveis?

14. Relativamente às últimas eleições para o Parlamento Europeu realizadas no país Z, foi publicado a seguinte manchete num jornal: "O partido X foi mais votado do que o partido Y".
Uma equipa de estatísticos selecionou aleatoriamente 35 concelhos e registou as percentagens de votos obtidas por estes partidos.
Admita que a Normalidade dos dados não é verificada, mas que a distribuição das diferenças é simétrica. Pronuncie-se quanto à manchete do jornal, considerando $\alpha = 5\%$ e que o valor observado para a estatística de teste foi 2,28. Que teste realizou?

15. Recolheu-se uma amostra aleatória de 9 alunos do curso de Sociologia tendo-se observado as classificações obtidas nas duas frequências da disciplina Estatística para Sociólogos II:

1ª Frequência	16,3	12,2	14,4	8,9	12,1	11,6	14,3	11,0	15,9
2ª Frequência	15,1	12,4	11,7	8,0	8,4	10,1	6,7	8,2	14

Ao nível de significância de 1%, pode considerar-se que existe diferença entre as classificações das duas frequências?

16. Na tabela seguinte apresentam-se o número de idas ao Ginásio de um casal desportista, num mês.

Casal	1	2	3	4	5	6
N º idas (mulher)	12	20	7	9	4	5
N º idas (homem)	10	14	9	12	13	10

a) Teste a existência de ligação significativa entre as duas variáveis, ao nível de significância de 5%.

b) Pode-se concluir que as mulheres que frequentam mais o ginásio têm como parceiros os homens também mais assíduos?

17. Um centro de explicações disponibiliza um curso intensivo de matemática que afirmam ser bastante eficaz no combate às más notas obtidas na prova específica de matemática. Escolheram-se aleatoriamente 7 estudantes do 12º ano para frequentarem o referido curso. Para avaliar a eficácia do curso, os alunos realizaram dois testes equivalentes: um teste antes e outro depois de frequentarem o curso. O valor obtido para o coeficiente de correlação amostral de Spearman foi de 0,810.
Ao nível de significância de 10%, considera que existe relação linear positiva entre as duas variáveis?

18. Com o aproximar da época balnear, são várias as pessoas que iniciam programas de dieta para perder as gorduras acumuladas durante o inverno. Na TVCompras é anunciado um medicamento, à base de estratos naturais de plantas, cujo *slogan* é "Perca mais de 10 kg em 10 dias".
Selecionaram-se aleatoriamente 31 pessoas às quais foi administrado este medicamento, tendo-se observado os seus pesos antes de tomarem o medicamento e 10 dias depois, tendo-se observado para o coeficiente de correlação amostral de Spearman 0,927
Ao nível de significância de 10%, considera que existe relação linear positiva entre as duas variáveis?

19. Um determinado cientista suspeita que os homens têm maior propensão para o raciocínio abstrato. Para fundamentar a sua suspeita, recolheu uma amostra aleatória de 8 casais e submeteu cada um deles a uma prova, tendo-os classificado da seguinte forma:

Casais	1	2	3	4	5	6	7	8
Classificação	H > M	H = M	H > M	H > M	H > M	H = M	H > M	H < M

Com base nos resultados, concorda com a suspeita do cientista. Considere $\alpha = 5\%$.

20. Um hipermercado, no dia do seu 2º aniversário sorteou, entre os seus clientes que realizaram compras, a atribuição de 10 automóveis. Como é usual, na presença de um representante do Governo Civil sortearam-se os 10 números, correspondentes aos talões das compras. Os montantes, em euros, das compras desses clientes foram:

68,45 42,52 77,10 62,51 47,60 102,20 55,50 82,00 41,00 110,30

Ao nível de significância de 5%, pode concluir-se que metade dos seus clientes gastou menos de 60 euros?

21. De forma a controlar as quantidades calóricas das gorduras contidas nas refeições dos refeitórios de 3 escolas primárias, analisaram-se algumas refeições, escolhidas ao acaso, em cada uma das escolas, tendo-se observado os seguintes resultados:

Escolas	Quantidades calóricas (Kj)					
1	118	122	140	133	126	128
2	133	131	135	143	146	149
3	136	137	141	146	154	

Sabendo que a quantidade calórica das refeições não tem distribuição Normal:

a) Teste a hipótese de a distribuição da quantidade calórica ser idêntica nas 3 escolas (use $\alpha = 1\%$).
b) O diretor da escola 1 afirmou que a quantidade calórica mediana era inferior a 130 kj. Ao nível de significância de 5%, concorda com a afirmação?

11 Regressão linear simples

Existem dois objetivos possíveis e usuais para a utilização da regressão linear. O primeiro é descrever a relação linear entre duas variáveis, onde uma assume o papel de variável independente e outra de dependente; o segundo objetivo é prever valores futuros da variável dependente baseados no valor da variável independente.

Modelo de regressão linear simples: Descreve uma relação linear entre uma variável independente, x, e uma variável dependente, Y, i. e.:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

onde β_0 e β_1 são constantes, x_i são conhecidos e ε_i representa o erro aleatório associado ao valor observado para Y.

Alternativamente este modelo pode ser escrito da seguinte forma:

$$E(Y_i|x = x_i) = \beta_0 + \beta_1 x_i.$$

Observação: Existem autores que distinguem se estão perante um problema de estimação de um modelo de regressão (apresentado na definição anterior) ou um problema de correlação (Galvão de Mello, 1997). Assim:

- Num problema estatístico de regressão estudam-se as distribuições de frequências de uma variável para certos valores fixos de outra variável, dita controlada, procurando determinar a forma de relação entre as variáveis. Um objetivo importante num problema de regressão é predizer ou estimar o valor mais provável da variável não controlada correspondente a um valor dado da outra variável: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$;
- Num problema de correlação analisa-se a variação conjunta de duas variáveis, nenhuma das quais controladas pelo investigador, sendo neste caso $Y = \beta_0 + \beta_1 X + \varepsilon$.

Estas duas situações implicam algumas diferenças nos desenvolvimentos e nos formalismos destes modelos. Neste trabalho optou-se por apenas se apresentar o primeiro modelo referido.

11.1 Reta de regressão ajustada

Com base numa amostra de n pares de observações $(x_i;\ y_i), i = 1, \dots, n$, é possível ajustar diferentes retas à nuvem de pontos que dependem, obviamente, dos valores considerados para a ordenada na origem, $\hat{\beta}_0$, e para o declive, $\hat{\beta}_1$.

A equação da **reta de regressão ajustada** é:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

A diferença entre o valor observado para Y e o valor estimado pela reta de regressão designa-se por resíduo de estimação ou erro de previsão (Figura 11.1):

O **resíduo** ou **erro** é dado por**:**

$$\varepsilon_i = Y_i - \hat{Y}_i.$$

Os resíduos de estimação, ou erro de previsão, são representados por e_i, sendo $e_i = \hat{\varepsilon}_i = y_i - \hat{y}_i$.

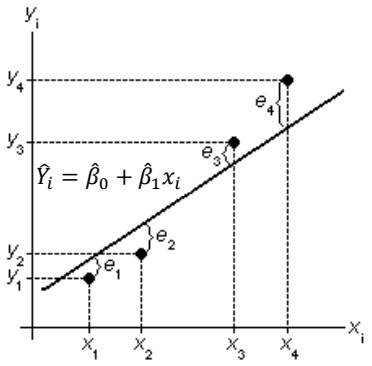


Figura 11.1: *Desvios entre os valores observados e os valores estimados.*

A questão que se coloca é então saber qual a reta que melhor se ajusta aos dados, ou seja, que valores considerar para β_0 e β_1? Existem vários métodos para estimar os parâmetros β_0 e β_1, mas mais adiante veremos apenas um desses métodos, o método dos mínimos quadrados.

11.2 Pressupostos do modelo

Os pressupostos subjacentes ao modelo de regressão linear simples são:

1. A relação entre X e Y tem que ser linear;
2. $E(\varepsilon_i) = 0$;
3. $Var(\varepsilon_i) = \sigma^2$;
4. ε_i são independentes, i. e., $Cov(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$;
5. ε_i seguem distribuição Normal;

para $i, j = 1, \dots, N$. Os pressupostos 2 a 5 podem ser escritos apenas como ε_i são independentes e identicamente distribuídos (*i.i.d*) com $\varepsilon_i \sim N(0;\ \sigma)$.

Com base nos pressupostos do modelo, verifica-se que:

- $E(Y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = E(\beta_0) + E(\beta_1 x_i) + E(\varepsilon_i) = \beta_0 + \beta_1 x_i \Rightarrow Y_i = E(Y_i) + \varepsilon_i$;
- $Var(Y_i) = E\left(\left(Y_i - E(Y_i)\right)^2\right) = E((Y_i - \beta_0 + \beta_1 x_i)^2) = E((\varepsilon_i)^2) = Var(\varepsilon_i) = \sigma^2$;
- $Cov(Y_i, Y_j) = Cov\left(\left(Y_i - E(Y_i)\right), \left(Y_j - E(Y_j)\right)\right) = E(\varepsilon_i, \varepsilon_j) = Cov(\varepsilon_i, \varepsilon_j) = 0$;
- $Y_i \sim N(\beta_0 + \beta_1 x_i;\ \sigma)$ uma vez que os Y_i são combinações lineares de v. a. Normais independentes.

Uma forma de validar os pressupostos do modelo é através da análise gráfica dos resíduos. Como se trata da análise de gráficos, por vezes a sua interpretação pode ser subjetiva sendo a experiência fundamental nesta análise. Existem alguns testes estatísticos que podem ser usados, mas que não serão abordados neste livro.

1. Linearidade entre as variáveis X e Y:
   - Representar os valores observados $(x_i;\ y_i)$ num diagrama de dispersão. Os pontos devem posicionar-se aproximadamente sobre uma reta; ou,
   - Calcular o coeficiente de correlação linear de Pearson. Deve obter-se um valor absoluto elevado para este coeficiente.
2. Resíduos têm média nula
   - Representar os valores $(\hat{y}_i;\ e_i)$ num diagrama de dispersão. Os pontos devem dispor-se em torno de uma reta de declive nulo e que passa na origem.
3. Homogeneidade da variância, i.e., homocedasticidade:

- Representar os valores $(\hat{y}_i;\ e_i)$ num diagrama de dispersão. A variância dos resíduos deve ser a mesma ao longo dos valores representados no eixo do horizontal $(\hat{y}_i)$, ou seja, os pontos devem formar uma banda horizontal em torno do valor zero e o gráfico não deve apresentar qualquer padrão.

4. Independência entre os erros:
   - Representar os valores $(\hat{y}_i;\ e_i)$ num diagrama de dispersão. Deve ser observada uma nuvem de pontos aleatórios; ou,
   - Quando as observações são feitas de acordo com uma ordem, então representar os valores $(e_i; i)$ num diagrama de dispersão. O gráfico não deve apresentar qualquer tendência, padrão ou alternância entre os sinais dos resíduos.
5. Normalidade dos resíduos

   O estudo da normalidade dos resíduos é feita sobre os resíduos estandardizados:

$$e_i^* = \frac{e_i}{\sqrt{MQE\left(1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{(n-1)s_x^2}\right)}}$$

   - Traçar os resíduos estandardizados (e_i^*) no gráfico quantil-quantil comparando-os com quantis da distribuição Normal. Os pontos devem posicionar-se em trono de uma linha reta; ou,
   - Representar os resíduos estandardizados (e_i^*) num histograma e este deve apresentar o comportamento de uma distribuição Normal; ou,
   - Testar a normalidade dos resíduos estandardizados recorrendo a um teste de hipótese (ver capítulo 10).

Por vezes existem pares de observações que se destacam dos restantes e que podem colocar em causa a qualidade do modelo de regressão ajustado. Designam-se por observações *influentes* os pontos cuja remoção provoca uma alteração significativa na reta de regressão. Designam-se por *outliers*, ou observações atípicas, os pontos que saem fora do padrão geral dos dados.

Algumas observações podem ser influentes e *outliers* em simultâneo. Esses casos merecem uma atenção especial.

11.3 Estimadores dos mínimos quadrados

Com o **método dos mínimos quadrados** interessa encontrar os valores de β_0 e β_1 que minimizam a soma dos quadrados dos erros (SQE), ou seja,

$$\min_{\hat{\beta}_0, \hat{\beta}_1} SQE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Os valores de β_0 e β_1 que minimizam a soma dos quadrados dos erros chamam-se **estimadores dos mínimos quadrados** de β_0 e β_1, do modelo de regressão linear simples, e representam-se por $\hat{\beta}_0$ e $\hat{\beta}_1$ (Figura 11.2).

Estimadores dos mínimos quadrados de $\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{x}\,\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{n\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xY}}{S_x^2} = R\frac{S_Y}{S_x}$$

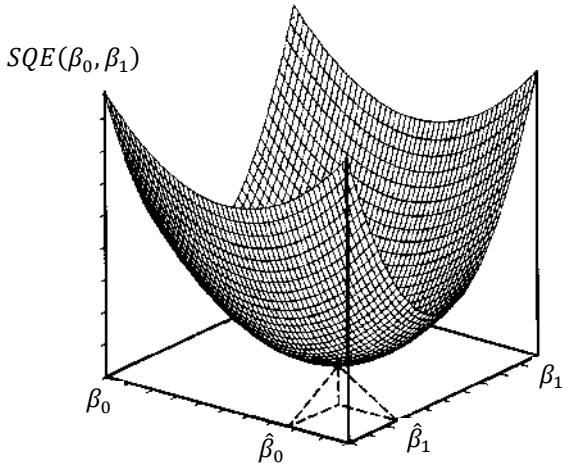


Figura 11.2: *Função quadrática e minimização dos valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ (Griffiths et al., 1993).*

Demonstração: Os estimadores dos mínimos quadrados são obtidos através da resolução das condições de 1ª ordem (designadas por equações normais):

$$\begin{cases} \dfrac{\partial SQE}{\partial \hat{\beta}_0} = 0 \\ \dfrac{\partial SQE}{\partial \hat{\beta}_1} = 0 \end{cases} \Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} 2(-1)\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \\ \displaystyle\sum_{i=1}^{n} 2(-x_i)\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \end{cases} \Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \\ \displaystyle\sum_{i=1}^{n} x_i\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \\ \displaystyle\sum_{i=1}^{n} x_i Y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{\beta}_0 = \dfrac{\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} \\ \displaystyle\sum_{i=1}^{n} x_i Y_i - \left(\overline{Y} - \hat{\beta}_1 \overline{x}\right) \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x} \\ \hat{\beta}_1 \left(\displaystyle\sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} x_i Y_i - \overline{Y} \sum_{i=1}^{n} x_i \end{cases} \Leftrightarrow \begin{cases} - \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i Y_i - \overline{Y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i} \end{cases}$$

$$\Leftrightarrow \begin{cases} \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i Y_i - n\overline{x}\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \cdot \end{cases}$$

Observação: As segundas derivadas são positivas.

A reta de regressão estimada resulta da substituição, no modelo de regressão linear simples, dos coeficientes β_0 e β_1, pelas suas estimativas.

Propriedades do modelo estimado:

1. A reta de regressão estimada passa pelo ponto $\left(\overline{x}, \overline{Y}\right)$:
$$E(Y|x = \overline{x})E = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} = \left(\overline{Y} - \hat{\beta}_1 \overline{x}\right) + \hat{\beta}_1 \overline{x} = \overline{Y}.$$
2. $\displaystyle\sum_{i=1}^{n} e_i = 0$, sendo $e_i = \hat{\varepsilon}_i = Y_i - \hat{Y}_i$.

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right) = \sum_{i=1}^{n}\left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right) = \sum_{i=1}^{n}\left(Y_i - \left(\left(\overline{Y} - \hat{\beta}_1\overline{x}\right) + \hat{\beta}_1 x_i\right)\right)$$
$$= \sum_{i=1}^{n}\left(\left(Y_i - \overline{Y}\right) - \hat{\beta}_1(x_i - \overline{x})\right) = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right) - \hat{\beta}_1\sum_{i=1}^{n}(x_i - \overline{x}) = 0.$$

11.3.1 Propriedades dos estimadores

Facilmente se verifica que os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ são combinações lineares de Y_i, pelo que são v. a.:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{x} = \sum_{i=1}^{n}\frac{Y_i}{n} - \hat{\beta}_1\sum_{i=1}^{n}\frac{x_i}{n},$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sum_{i=1}^{n}\frac{(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\left(Y_i - \overline{Y}\right) = \sum_{i=1}^{n} w_i Y_i.$$

Como Y_i seguem uma distribuição normal, então $\hat{\beta}_0$ e $\hat{\beta}_1$ também têm distribuição Normal, resta determinar os parâmetros.

No caso de a amostra ser grande, então mesmo que os Y_i não sigam uma distribuição Normal, pelo Teorema do Limite Central, $\hat{\beta}_0$ e $\hat{\beta}_1$ seguirão aproximadamente uma distribuição Normal.

$$\hat{\beta}_0 \sim N\left(\beta_0, \sqrt{\frac{\sigma^2\sum_{i=1}^{n}x_i^2}{n(n-1)s_x^2}}\right),$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \sqrt{\frac{\sigma^2}{(n-1)\,s_x^2}}\right),$$

$$Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\sigma^2\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

ou seja, $\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores centrados para os parâmetros β_0 e β_1, respetivamente.

11.4 Teorema de Gauss-Markov

Definição: Considere-se o modelo de regressão linear simples

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

sob os pressupostos enunciados anteriormente (1 a 5).

De todos os estimadores possíveis de β_0 e $\beta_1$que são centrados e lineares em Y_i, os estimadores dos mínimos quadrados são os que têm menor variância, e por consequência são os mais eficientes. Além disso, qualquer combinação linear de $\hat{\beta}_0$ e $\hat{\beta}_1$ tem variância mínima na classe de todos os estimadores que são centrados e lineares em Y_i.

Desta forma, os estimadores dos mínimos quadrados dizem-se **os melhores estimadores lineares centrados** (BLUE – *Best Linear Unbiased Estimators*).

11.5 Decomposição da variação total

A variabilidade total que ocorre no conjunto de dados pode ser decomposta em duas componentes: uma explicada pela reta de regressão e outra que não é explicada pela reta de regressão (Figura 11.3)

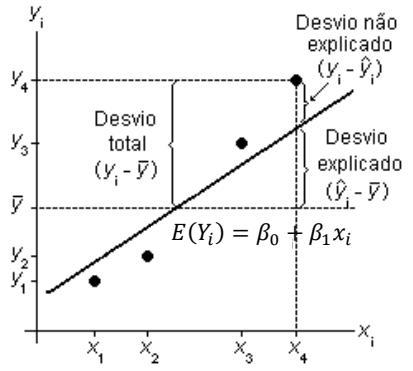


Figura 11.3: *Decomposição do desvio total.*

$$\begin{gathered}\text{Variação total} = \text{Variação explicada pela regressão} + \text{Variação não explicada pela regressão} \\ \Leftrightarrow \text{Soma dos Quadrados Totais } (SQT) = \text{Soma dos Quadrados da Regressão } (SQR) + \text{Soma dos Quadrados dos Erros } (SQE) \\ \Leftrightarrow \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.\end{gathered}$$

11.5.1 Coeficiente de determinação

O **coeficiente de determinação**, $\boldsymbol{r^2}$, associado à reta ajustada, representa a proporção da variabilidade amostral da variável dependente que é explicada pela equação de regressão:

$$r^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} = \left(\hat{\beta}_1 \frac{s_x}{s_Y}\right)^2, \text{ sendo } 0 \leq r^2 \leq 1.$$

Quanto maior o valor de r^2 maior é o poder de explicação da regressão.

Observação: r^2 é o quadrado do coeficiente de correlação r.

11.5.2 Tabela ANOVA

Na tabela ANOVA (Tabela 11.1) apresenta-se um resumo da informação relativa à variação que ocorre nos dados.

Tabela 11.1: *Tabela ANOVA.*

Fonte de Variação	Soma dos Quadrados	Graus de liberdade	Média dos Quadrados	F
Explicada	SQR	1	$MQR = SQR$	$f_{obs} = \frac{MQR}{MQE}$
Não explicada	SQE	$n-2$	$MQE = \frac{SQE}{n-2}$	
Total	SQT	$n-1$	———	———

11.6 Inferência estatística

11.6.1 Estimação da variância do erro, σ^2

Habitualmente σ^2, a variância do erro do modelo, é desconhecida pelo que é necessário estimá-la para se poder prosseguir o estudo com a inferência estatística.

Um estimador centrado da variância do erro do modelo, σ^2, é:

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{e_i^2}{n-2} = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-2} = \sum_{i=1}^{n} \frac{\left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2}{n-2} = MQE.$$

Consequentemente,

$$\hat{\sigma}^2_{\hat{\beta}_0} = \widehat{Var}(\hat{\beta}_0) = \frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{n(n-1)S_x^2}$$

e

$$\hat{\sigma}^2_{\hat{\beta}_1} = \widehat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{(n-1)\, S_x^2}.$$

Portanto,

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\widehat{Var}(\hat{\beta}_0)}} \sim t_{n-2},$$

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{n-2}.$$

11.6.2 Intervalos de confiança para $\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$

O **I. C. a $100(1-\alpha)\%$ para β_0** é dado por:

$$\left]\hat{\beta}_0 - t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\beta}_0)};\ \hat{\beta}_0 + t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\beta}_0)}\right[.$$

O **I. C. a $100(1-\alpha)\%$ para β_1** é dado por:

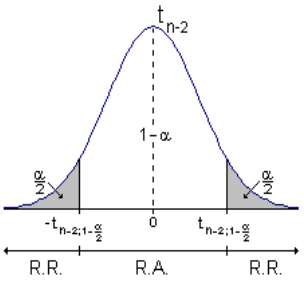
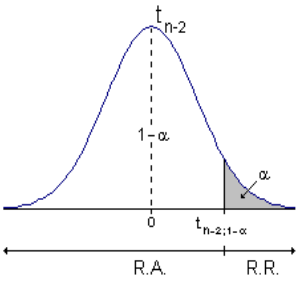
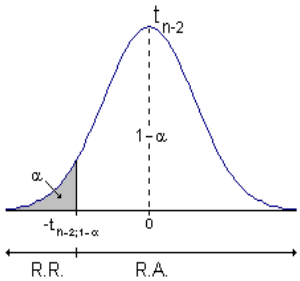
$$\left]\hat{\beta}_1 - t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\beta}_1)};\ \hat{\beta}_1 + t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\beta}_1)}\right[.$$

11.6.3 Testes de hipóteses

11.6.3.1 Testes de hipóteses para $\boldsymbol{\beta_0}$

A estatística de teste a utilizar, quando se realiza um teste de hipótese para o parâmetro β_0, é:

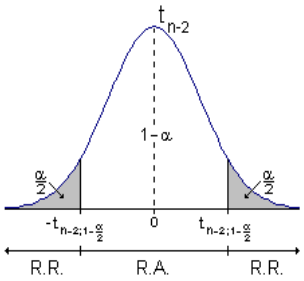
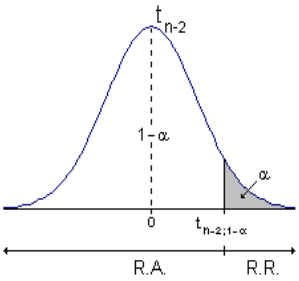
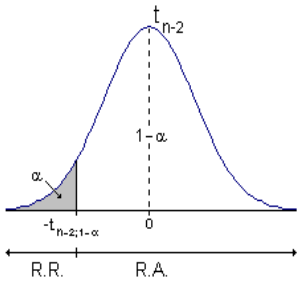
$$T = \frac{\hat{\beta}_0 - b_0}{\sqrt{\widehat{Var}(\hat{\beta}_0)}} \sim t_{n-2}.$$

T. bilateral	T. unilateral direito	T. unilateral esquerdo
Hipóteses a testar: $H_0: \beta_0 = b_0$ *vs* $H_1: \beta_0 \neq b_0$	**Hipóteses a testar:** $H_0: \beta_0 \leq b_0$ *vs* $H_1: \beta_0 > b_0$	**Hipóteses a testar:** $H_0: \beta_0 \geq b_0$ vs $H_1: \beta_0 < b_0$
Regiões críticas:  $R.A.: \left]-t_{n-2;1-\frac{\alpha}{2}};\ t_{n-2;1-\frac{\alpha}{2}}\right[$ $R.R.: \left]-\infty;\ -t_{n-2;1-\frac{\alpha}{2}}\right] \cup \left[t_{n-2;1-\frac{\alpha}{2}};+\infty\right[$	**Regiões críticas:**  $R.A.: \left]-\infty;\ t_{n-2;1-\alpha}\right[$ $R.R.: \left[t_{n-2;1-\alpha};+\infty\right[$	**Regiões críticas:**  $R.A.: \left]-t_{n-2;1-\alpha};\ +\infty\right[$ $R.R.: \left]-\infty;\ -t_{n-2;1-\alpha}\right]$
Regra de decisão: Rejeitar H_0 quando $\lvert t_{obs}\rvert \geq t_{n-2;1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $t_{obs} \geq t_{n-2;1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $t_{obs} \leq -t_{n-2;1-\alpha}$
Cálculo do valor p: valor $p = 2 \times P(T \geq \lvert t_{obs}\rvert)$	**Cálculo do valor p:** valor $p = P(T \geq t_{obs})$	**Cálculo do valor p:** valor $p = P(T \leq t_{obs})$

11.6.3.2 Testes de hipóteses para β_1

A estatística de teste a utilizar, quando se realiza um teste de hipótese para o parâmetro β_1, é:

$$T = \frac{\hat{\beta}_1 - b_0}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{n-2}.$$

T. bilateral	T. unilateral direito	T. unilateral esquerdo
Hipóteses a testar: $H_0: \beta_1 = b_0$ *vs* $H_1: \beta_1 \neq b_0$	**Hipóteses a testar:** $H_0: \beta_1 \leq b_0$ *vs* $H_1: \beta_1 > b_0$	**Hipóteses a testar:** $H_0: \beta_1 \geq b_0$ vs $H_1: \beta_1 < b_0$
Regiões críticas:  	**Regiões críticas:**  	**Regiões críticas:** 

$R.A.: \left]-t_{n-2;1-\frac{\alpha}{2}};\ t_{n-2;1-\frac{\alpha}{2}}\right[$ $R.R.: \left]-\infty;\ -t_{n-2;1-\frac{\alpha}{2}}\right] \cup \left[t_{n-2;1-\frac{\alpha}{2}};+\infty\right[$	$R.A.: \left]-\infty;\ t_{n-2;1-\alpha}\right[$ $R.R.: \left[t_{n-2;1-\alpha};+\infty\right[$	$R.A.: \left]-t_{n-2;1-\alpha};\ +\infty\right[$ $R.R.: \left]-\infty;\ -t_{n-2;1-\alpha}\right]$
Regra de decisão: Rejeitar H_0 quando $\lvert t_{obs}\rvert \geq t_{n-2;1-\frac{\alpha}{2}}$	**Regra de decisão:** Rejeitar H_0 quando $t_{obs} \geq t_{n-2;1-\alpha}$	**Regra de decisão:** Rejeitar H_0 quando $t_{obs} \leq -t_{n-2;1-\alpha}$
Cálculo do valor p: $\text{valor } p = 2 \times P(T \geq \lvert t_{obs}\rvert)$	**Cálculo do valor p:** $\text{valor } p = P(T \geq t_{obs})$	**Cálculo do valor p:** $\text{valor } p = P(T \leq t_{obs})$

O teste de hipótese *bilateral* para o parâmetro β_1

$$H_0: \beta_1 = 0 \text{ } vs \text{ } H_1: \beta_1 \neq 0$$

pode ser realizado, alternativamente, utilizando a estatística de teste:

$$F = \frac{MQR}{MQE} \sim F_{1;n-2}.$$

Neste caso a regra de decisão é:

$$\text{rejeitar } H_0 \text{ quando } f_{obs} \geq f_{1;\,n-2;1-\alpha}.$$

O valor p é dado por:

$$\text{valor } p = P(F \geq f_{obs}).$$

11.7 Previsão

O objetivo final da regressão final é o de prever o valor ou valores para a variável dependente, Y, quando a variável independente, X, toma um certo valor e assumindo como verdadeira a reta de regressão estimada. Portanto, o que se pretende é estimar Y. Tal como já foi referido no capítulo da estimação, também aqui existem dois tipos de estimação (ou previsão) baseados:

- *Estimação pontual*: produção de um só valor para Y;
- *Estimação intervalar*: construção de um intervalo que, com certo grau de certeza previamente estipulado, contenha o verdadeiro valor de Y.

Além disso, a previsão pode ser encarada de 2 formas:

- *Previsão individual*: quando a variável X não se repete, então o que se pretende é estimar o *valor* a variável dependente, Y, quando a variável independente, X, assume um determinado valor.
- *Previsão média*: quando existem várias observações de Y para o mesmo valor da variável X, então o que se pretende é estimar o *valor médio* (valor esperado) para a variável dependente, Y, quando a variável independente, X, assume um determinado valor.

O intervalo de confiança para a previsão em média tem menor amplitude do que o intervalo de confiança para a previsão individual (Figura 11.4).

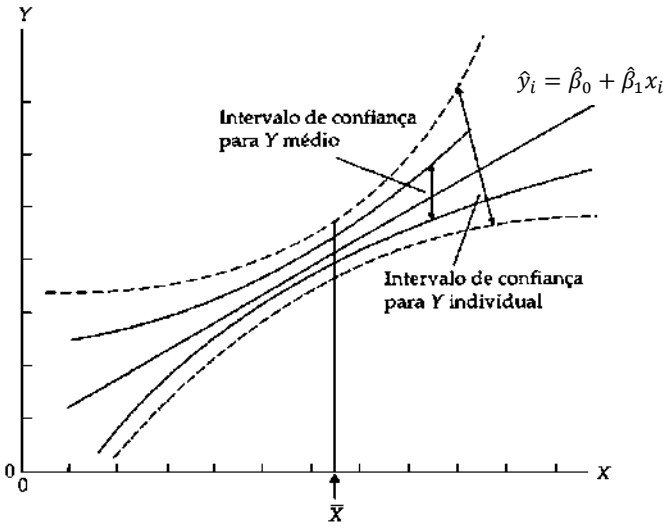


Figura 11.4: *Intervalos de confiança para valores Y individual e Y médio* (Gujarati, 2000).

Exemplo: Seja X o rendimento de uma família e Y o consumo dessa família de certos tipos de bens, em unidades momentárias (u. m.). Considere-se que $x = 24$ u. m.

- A *previsão individual* corresponde à previsão do *consumo de uma certa família* com rendimento igual a 24 u. m.
- A *previsão média* consiste em prever o *consumo médio das famílias* com rendimento igual a 24 u. m.

Para um certo valor x_h, a **melhor estimativa pontual** (individual ou média) **para $\boldsymbol{Y}$** é dada por:

$$\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h.$$

11.7.1 Intervalo de confiança para a previsão individual de $\boldsymbol{Y}$

O **I. C. a $\boldsymbol{100(1-\alpha)\%}$ para a previsão individual de $\boldsymbol{Y}$** é dado por:

$$\left]\hat{Y}_h - t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x_h-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)};\ \hat{Y}_h + t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x_h-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}\right[.$$

11.7.2 Intervalo de confiança para a previsão em média de $\boldsymbol{Y}$

O **I. C. a $\boldsymbol{100(1-\alpha)\%}$ para a previsão em média de $\boldsymbol{Y}$**, isto é, de $E(Y|X=x_h)$, é dado por:

$$\left]\hat{Y}_h - t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{(x_h-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)};\ \hat{Y}_h + t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{(x_h-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}\right[.$$

11.8 Exercícios resolvidos

1. Pensa-se que o número de embalagens vendidas de um determinado medicamento genérico (Y) depende do seu preço (X, em euros). Para o efeito observou-se durante 12 semanas os valores destas variáveis, tendo-se obtido os seguintes resultados:

y	892	1012	1060	987	680	739	809	1275	946	874	720	1096
x	1,23	1,15	1,10	1,20	1,35	1,25	1,28	0,99	1,22	1,25	1,30	1,05

a) Estime a reta de regressão linear, pelo método dos mínimos quadrados. Interprete os coeficientes de regressão.
b) Represente graficamente a nuvem de pontos e a reta ajustada.
c) Obtenha os resíduos de estimação.
d) Valide os pressupostos subjacentes ao modelo de regressão linear.
e) Determine o coeficiente de correlação e interprete o valor obtido.
f) Estime a variância do erro do modelo.
g) Construa a tabela ANOVA.
h) Construa um intervalo de confiança a 99% para β_0.
i) Complete: "Com 95% de confiança o verdadeiro valor de β_1 situa-se entre ... e ...".
j) A partir de que nível de significância é rejeitada a hipótese do coeficiente β_0 ser nulo?
k) Ensaie a hipótese de que o preço não influência linearmente o número de embalagens vendidas (considere $\alpha = 1\%$).
l) Determine e interprete o coeficiente de determinação.
m) Considere que o preço de cada embalagem é atualmente 1,23 euros.
n) Quantas embalagens espera vender?
o) Determine o intervalo de confiança a 95% para o valor médio de embalagens que se preveem vender.

Resolução:

Sejam:

- X a v. a. que representa o preço de cada embalagem,
- Y a v. a. que representa o número de embalagens vendidas.

$n = 12;\ \overline{x} = 1{,}198;\ \overline{y} = 924{,}167;\ s_x = 0{,}106;\ s_y = 174{,}669;\ s_{xy} = -17{,}816.$

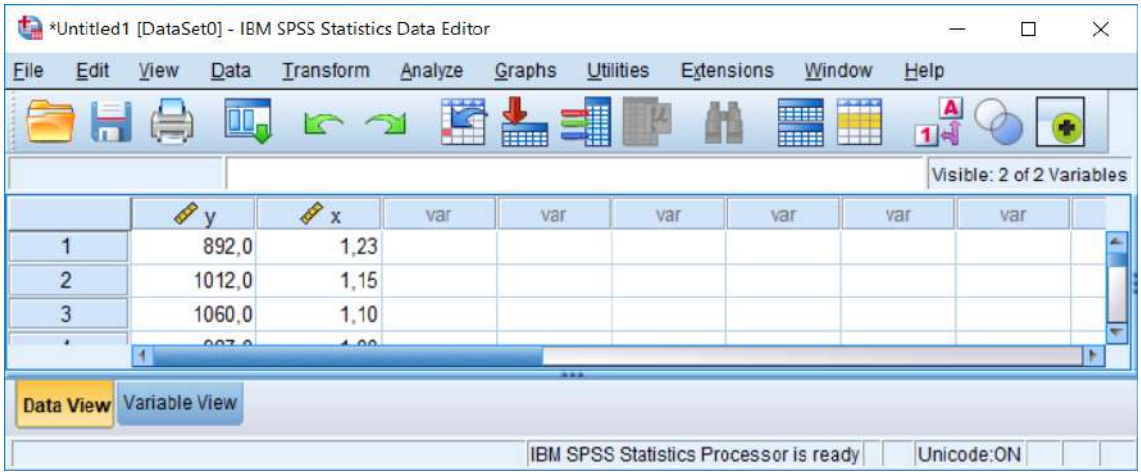
a) $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \Leftrightarrow \hat{y}_i = 2813{,}32 - 1577{,}58 x_i$, porque

$$\hat{\beta}_1 = \frac{s_{xY}}{s_x^2} = \frac{-16{,}331}{0{,}102^2} = -1577{,}58,$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 924{,}167 - (-1577{,}58) \times 1{,}1975 = 2813{,}32.$$

Interpretação: Quando o preço das embalagens for 0 euros espera-se vender em média 2813,32 embalagens (neste caso não faz sentido a interpretação de $\hat{\beta}_1$). Por cada aumento de 1 euro (uma unidade) no preço das embalagens espera-se vender, em média, aproximadamente menos 1578 embalagens do referido medicamento genérico.

(SPSS)



(SPSS) Analyze → Regression → Linear…
(Dependent: y; Block 1 of 1: Independent(s): x;
Statistics → ☑ Estimates)

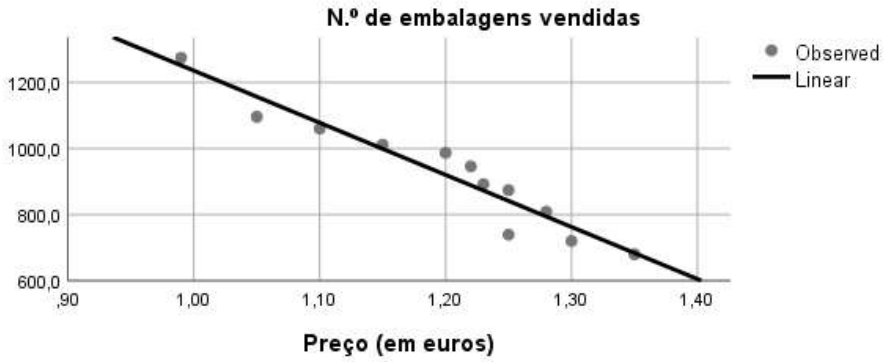
Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2813,320	175,324		16,046	,000
	Preço (em euros)	-1577,581	145,883	-,960	-10,814	,000

a. Dependent Variable: N.º de embalagens vendidas

O valor dos parâmetros estimados é apresentado na coluna B do Unstandardized Coefficients: o valor de $\hat{\beta}_0$ é dado na linha Constant e o valor de $\hat{\beta}_1$ na linha Preço (em euros).

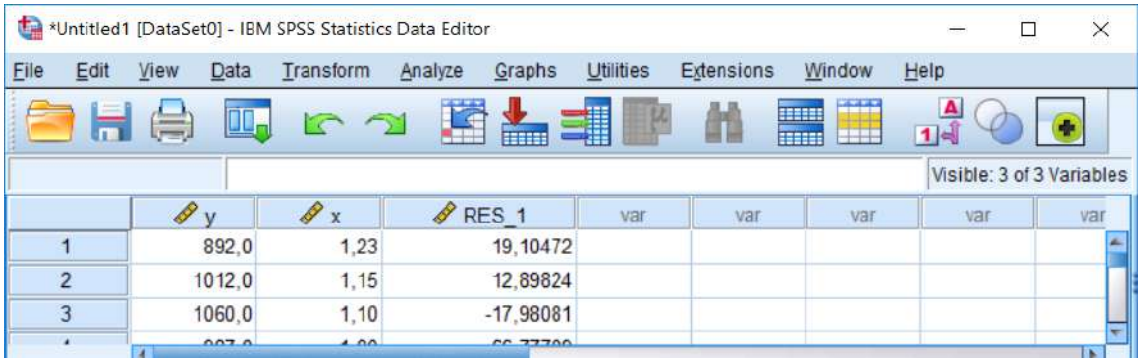
b) (SPSS) Analyze → Regression → Curve Estimation…
(Dependent: y; Independent: ⊙ Variable: x; ☑ Plot models; Models: ☑ Linear)



c) Resíduos de estimação: $e_i = y_i - \hat{y}_i$.

y_i	x_i	$\hat{y}_i$	e_i	y_i	x_i	$\hat{y}_i$	e_i
892	1,23	872,8953	19,1047	809	1,28	794,0162	14,9838
1012	1,15	999,1018	12,8982	1275	0,99	1251,5147	23,4853
1060	1,10	1077,9808	-17,9808	946	1,22	888,6711	57,3289
987	1,20	920,2227	66,7773	874	1,25	841,3437	32,6563
680	1,35	683,5856	-3,5856	720	1,30	762,4646	-42,4646
739	1,25	841,3437	-102,3437	1096	1,05	1156,8599	-60,8599

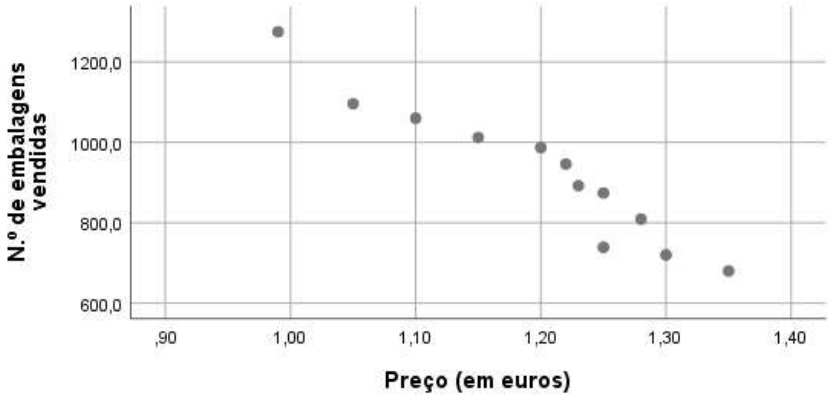
(SPSS) Analyze → Regression → Linear…
(Dependent: y; Block 1 of 1: Independent(s): x;
Save → Residuals: ☑ Unstandartized)



Os resíduos são apresentados numa nova coluna na janela de Data Editor.

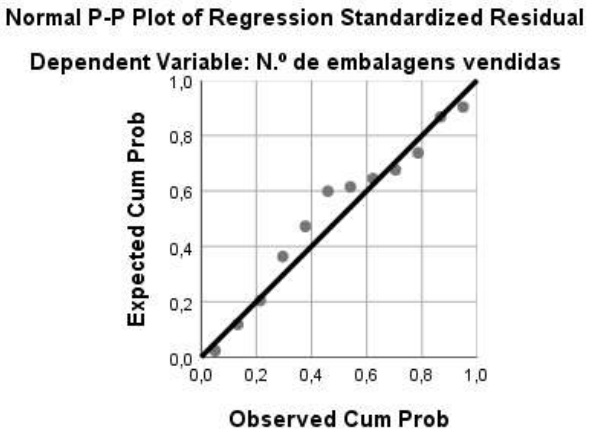
d) Validação dos pressupostos

(SPSS) Graphs → Legacy Dialogs → Scatter/Dot…
(Y Axis: y; X Axis: x)



Linearidade: os pontos dispõem-se aproximadamente sobre uma reta com declive negativo. Logo, podemos considerar que há uma relação linear negativa entre as duas variáveis.

(SPSS) Analyze → Regression → Linear…
(Dependent: y; Block 1 of 1: Independent(s): x;
Plots → Standardized Residual Plots: ☑ Normal probability plot)



Normalidade: pela análise do gráfico, parece haver alguma diferença na parte central entre os quantis observados e os esperados caso os dados fossem provenientes de uma população com distribuição Normal. Contudo, como a amostra é muito pequena podemos considerar que estes desvios são pequenos e que não colocam em causa o pressuposto de normalidade dos resíduos de estimação.

Em alternativa poderia ser usado um teste de hipóteses.

(SPSS) Analyze → Descriptive Statistics → Explore…
(Dependent List: RES_1; Display: ⊙ Plots;
Plots → ☑ Normality plots with tests)

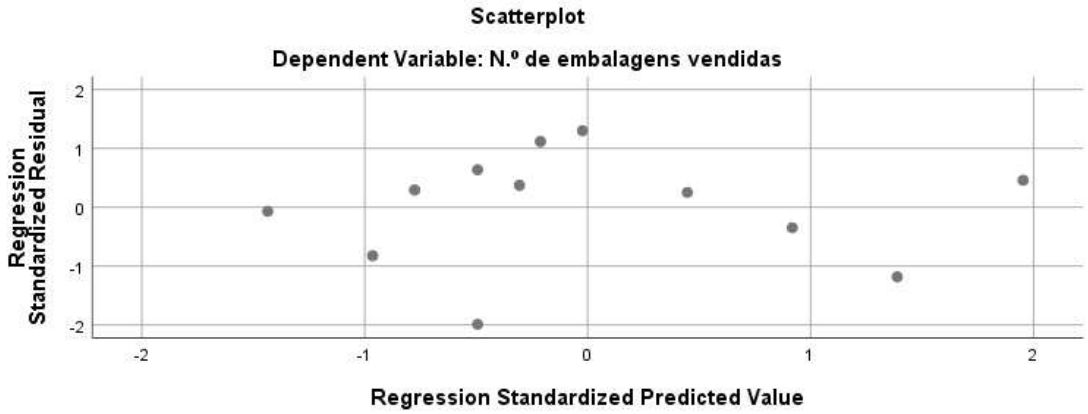
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	,187	12	,200*	,947	12	,598

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

Como a amostra é pequena, deve-se optar pelo teste de Shapiro-Wilk. Para este teste, o valor $p = 0{,}598$. Logo aos níveis usuais de significância, não há evidência estatística para afirmar que os resíduos de estimação não seguem uma população com distribuição Normal.

(SPSS) Analyze → Regression → Linear…
(Dependent: y; Block 1 of 1: Independent(s): x;
Plots → Scatter 1 of 1: Y: *ZRESID; X: *ZPRED)



Homocedasticidade: os pontos parecem dispor-se numa banda horizontal em torno do eixo horizontal.
Independência: O gráfico não parece apresentar um padrão nem tendência.
Portanto, não há violação dos pressupostos.

e) Coeficiente de correlação:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-17{,}816}{0{,}106 \times 174{,}669} = -0{,}96.$$

Existe relação linear negativa forte entre o preço das embalagens e número de embalagens vendidas. Portanto, para um preço elevado (baixo) da embalagem espera-se um número baixo (elevado) de embalagens vendidas.

(SPSS) Analyze → Correlate → Bivariate…
(Variables: y e x; Correlation Coefficients: ☑ Pearson)

Correlations

		N.º de embalagens vendidas	Preço (em euros)
N.º de embalagens vendidas	Pearson Correlation	1	-,960**
	Sig. (2-tailed)		,000
	N	12	12
Preço (em euros)	Pearson Correlation	-,960**	1
	Sig. (2-tailed)	,000	
	N	12	12

**. Correlation is significant at the 0.01 level (2-tailed).

f) Variância do erro do modelo:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{26437{,}23}{12-2} = 2643{,}723.$$

(SPSS) Analyze → Regression → Linear…
(Dependent: y; Block 1 of 1: Independent(s): x;
Statistics → ☑ Model fit)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,960[a]	,921	,913	51,4171

a. Predictors: (Constant), Preço (em euros)
b. Dependent Variable: N.º de embalagens vendidas

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	309166,437	1	309166,437	116,944	,000[b]
	Residual	26437,230	10	2643,723		
	Total	335603,667	11			

a. Dependent Variable: N.º de embalagens vendidas
b. Predictors: (Constant), Preço (em euros)

O valor de $\hat{\sigma}$ é apresentado na coluna Std. Error of the Estimate da primeira tabela. O valor de $\hat{\sigma}^2$ é apresentado na coluna Mean Square na linha Residual da segunda tabela.

g) $SQE = \sum_{i=1}^{n} e_i^2 = 26437{,}230;$

$$MQE = \frac{SQE}{n-2} = \hat{\sigma}^2 = 2643{,}723;$$

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 335603{,}667;$$

$$SQT = SQR + SQE \Leftrightarrow SQR = SQT - SQE$$

$$\Leftrightarrow 335603{,}667 - 26437{,}230 = 309166{,}437;$$

$$MQE = SQR = 309166{,}437;$$

$$f_{obs} = \frac{MQR}{MQE} = \frac{309166{,}437}{2643{,}723} = 116{,}944.$$

(SPSS) Ver alínea anterior (segunda tabela)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	309166,437	1	309166,437	116,944	,000[b]
	Residual	26437,230	10	2643,723		
	Total	335603,667	11			

a. Dependent Variable: N.º de embalagens vendidas
b. Predictors: (Constant), Preço (em euros)

h) I. C. a $100(1-\alpha)\%$ para β_0:

$$\left]\hat{\beta}_0 - t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\beta}_0)};\ \hat{\beta}_0 + t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\beta}_0)}\right[$$

Como,

$$\widehat{Var}(\hat{\beta}_0) = \frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{n(n-1)s_x^2} = \frac{2643{,}723 \times 17{,}332}{12 \times 11 \times 0{,}106^2} = 30738{,}445,$$

o I. C. a 99% para β_0 é

$$\begin{aligned}&\left]2813{,}320 - t_{10;\,0{,}995}\sqrt{30738{,}445};\ 2813{,}320 + t_{10;\,0{,}995}\sqrt{30738{,}445}\right[\\ &=]2813{,}320 - 3{,}169 \times 175{,}324;\ 2813{,}320 + 3{,}169 \times 175{,}324[\\ &=]2257{,}671;\ 3368{,}969[.\end{aligned}$$

Com 99% de confiança o verdadeiro valor de β_0 situa-se entre 2257,671 e 3368,969.

(SPSS) Analyze → Regression → Linear…
(Dependent: y; Block 1 of 1: Independent(s): x;
Statistics → Regression Coefficient: ☑ Estimates; ☑ Confidence Intervals; Level (%): 99)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	99,0% Confidence Interval for B Lower Bound	99,0% Confidence Interval for B Upper Bound
1	(Constant)	2813,320	175,324		16,046	,000	2257,671	3368,969
	Preço (em euros)	-1577,581	145,883	-,960	-10,814	,000	-2039,923	-1115,239

a. Dependent Variable: N.º de embalagens vendidas

i) I. C. a $100(1-\alpha)\%$ para β_1:

$$\left]\hat{\beta}_1 - t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\beta}_1)};\ \hat{\beta}_1 + t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\beta}_1)}\right[.$$

Como,

$$\widehat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{(n-1)\,s_x^2} = \frac{2643{,}723}{11 \times 0{,}106^2} = 21281{,}731,$$

o I. C. a 95% para β_1 é

$$\begin{aligned}&\left]-1577{,}58 - t_{10;\,0{,}995}\sqrt{21281{,}731};\ -1577{,}58 + t_{10;\,0{,}995}\sqrt{21281{,}731}\right[\\ &=]-1577{,}58 - 3{,}169 \times 145{,}883;\ -1577{,}58 + 3{,}169 \times 145{,}883[\\ &=]-1902{,}628;\ -1252{,}534[.\end{aligned}$$

Com 95% de confiança o verdadeiro valor de β_1 situa-se entre -1902,628 e -1252,534.

(SPSS) Analyze → Regression → Linear…
(Dependent: y; Block 1 of 1: Independent(s): x;
Statistics → Regression Coefficient: ☑ Estimates; ☑ Confidence Intervals; Level (%): 95)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95,0% Confidence Interval for B Lower Bound	95,0% Confidence Interval for B Upper Bound
1	(Constant)	2813,320	175,324		16,046	,000	2422,674	3203,966
	Preço (em euros)	-1577,581	145,883	-,960	-10,814	,000	-1902,628	-1252,534

a. Dependent Variable: N.º de embalagens vendidas

j) valor p = ?

H_0: $\beta_0 = 0$ vs H_1: $\beta_0 \neq 0$ (teste bilateral)

Estatística de teste:

$$T = \frac{\hat{\beta}_0 - b_0}{\sqrt{\widehat{Var}(\hat{\beta}_0)}} \sim t_{n-2=10}.$$

$$t_{obs} = \frac{2813{,}320 - 0}{\sqrt{30738{,}445}} = \frac{2813{,}320 - 0}{175{,}324} = 16{,}046.$$

valor $p = 2 \times P(T \geq |t_{obs}|) = 2 \times P(T \geq 16{,}046) = 2 \times \big(1 - P(T < 16{,}046)\big) < 0{,}001.$

Rejeitar H_0 aos níveis usuais de significância, i. e., existe evidência estatística para afirmar que o coeficiente β_0 não é nulo.

(SPSS) Ver alíneas f ou g (valor na linha Constant coluna Sig.)

k) $\alpha = 1\%$, $\beta_1 = 0$?

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ (teste bilateral)

Estatística de teste:

$$T = \frac{\hat{\beta}_1 - b_0}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{n-2=10} \quad \text{ou} \quad F = \frac{MQR}{MQE} \sim F_{1;n-2=1;10}$$

$$t_{obs} = \frac{-1577{,}58 - 0}{\sqrt{21281{,}731}} = \frac{-1577{,}58 - 0}{145{,}883} = -10{,}814 \qquad f_{obs} = \frac{309166{,}437}{2643{,}723} = 116{,}944$$

Rejeitar H_0 pois

$t_{n-2;1-\frac{\alpha}{2}} = t_{10;0{,}995} = 3{,}169$

e

$|t_{obs}| = 10{,}814 \geq 3{,}169$

ou

$f_{1;\, n-2;1-\alpha} = f_{1;\, 10;0{,}99} = 10$

e

$f_{obs} = 116{,}944 \geq 10.$

Ao nível de significância de 1%, existe evidência estatística para afirmar que preço influência de forma linear o número de embalagens vendidas.

(SPSS) Ver alíneas f ou g (valor na linha Preço (em euros) coluna Sig.)

l) Coeficiente de determinação:

$r^2 = (-0{,}9598)^2 = 0{,}921.$

Pode-se considerar que se fez um bom ajustamento, pois 92,1% da variabilidade que ocorre no n.º de embalagens vendidas é explicada pela relação linear com o preço das embalagens.

(SPSS) Ver alínea e (valor na coluna R square da primeira tabela)

m) $x_h = 1{,}23 \Rightarrow \hat{y}_h = ?$

i) $\hat{y}_h = 2813{,}320 - 1577{,}58 \times 1{,}23 = 872{,}895.$
Quando o preço é 1,23 euros espera-se vender, em média, 872,895 embalagens.

ii) $x_h = 1{,}23 \Rightarrow \hat{y}_h = 872{,}895.$

I. C. a $100(1-\alpha)\%$ para a previsão média de Y é dado por:

$$\left]\hat{Y}_h - t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\left(\frac{1}{n} + \frac{(x_h - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)\hat{\sigma}^2};\ \hat{Y}_h + t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\left(\frac{1}{n} + \frac{(x_h - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)\hat{\sigma}^2}\right[.$$

Como

$$s_x^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1} \Leftrightarrow \sum_{i=1}^{n} (x_i - \overline{x})^2 = (n-1)s_x^2,$$

então

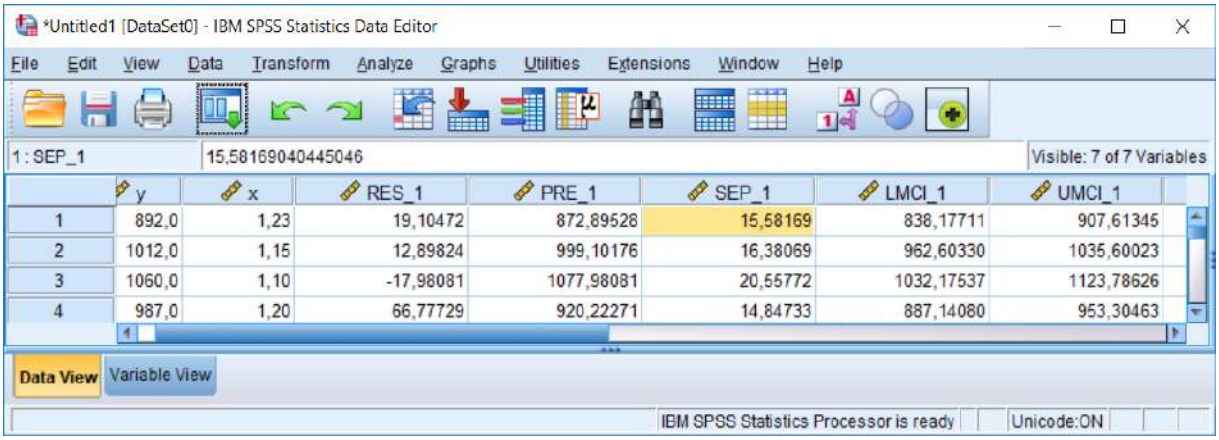
$$\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_s - \overline{x})^2}{(n-1)s_x^2} \right) = 2643{,}230 \left(\frac{1}{12} + \frac{(1{,}23 - 1{,}198)^2}{(12-1) \times 0{,}106^2} \right) = 242{,}789,$$

e o I. C. a 95% para a previsão média de Y é:

$$]872{,}895 - t_{10;\,0{,}975}\sqrt{242{,}789};\ 872{,}895 + t_{10;\,0{,}975}\sqrt{242{,}789}[$$
$$=]872{,}895 - 2{,}228 \times 15{,}517;\ 872{,}895 + 2{,}228 \times 15{,}517[=]838{,}177;\ 907{,}613[.$$

(SPSS) Verificar se na coluna dos valores x existe um valor $x = x_h$. Se não existir, então adicionar uma linha ao conjunto de colocando o valor de x_h na coluna x. Neste exercício, já existe na coluna x o valor $1{,}23 = x_h$, não sendo por isso necessário adicionar qualquer valor na coluna x.

(SPSS) Analyze → Regression → Linear…
(Dependent: y; Block 1 of 1: Independent(s): x;
Save → Predicted values: ☑ Unstandartized; ☑ S.E. of mean predictions; Prediction Intervals: ☑ Mean: Confidence Interval: 95%)



A informação sobre a $\hat{y}_h$ quando $x_h = 1{,}23$ é dada na linha em que $x = 1{,}23$ (neste caso na primeira linha). O valor de $\hat{y}_h$ é apresentado na coluna PRE_1, o valor de $\sqrt{\left(\frac{1}{n} + \frac{(x_h - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)\hat{\sigma}^2}$ na coluna SEP_1 e os limites do I. C. para a previsão média de Y nas colunas LMCI_1 (limite inferior) e UMCI_1 (limite superior).

2. Um vendedor de bebidas está interessado em estudar o efeito que o preço (em euros) dum whisky de primeira qualidade tem na quantidade vendida. Para o efeito registou os valores destas variáveis durante 8 semanas, e aplicou uma regressão linear, com o auxílio do software Excel. Infelizmente o computador foi atacado por um vírus, pelo que apenas obteve os seguintes resultados:

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression					,00[b]
	Residual	28,613				
	Total	529,689				

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	135,691			12,024	,000		
	x	-5,903	0,576			,000		

a) Apresente a equação da reta de regressão ajustada.
b) Interprete $\hat{\beta}_2$no modelo estimado.
c) Complete a tabela ANOVA.
d) Estime a variância do erro do modelo.
e) Determine e interprete o coeficiente de determinação.
f) Determine o coeficiente de correlação e interprete o valor obtido.
g) Ensaie a hipótese de que a reta de regressão estimada passa pela origem. (Utilize um nível de significância de 1%)
h) Teste a hipótese de o preço não explicar de forma linear a quantidade vendida.
i) Construa um intervalo de 95% de confiança para β_0. Interprete-o.
j) Construa um intervalo de 95% de confiança para β_1. Interprete-o.
k) Complete a segunda tabela.
l) Que quantidade de *whisky* espera vender num dia em que o preço é de 10 €?

Resolução:

Sejam:

- X a v. a. que representa o preço (em euros) do *whisky*,
- Y a v. a. que representa a quantidade vendida.

$n = 8$.

a) $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \Leftrightarrow \hat{y}_i = 135{,}691 - 5{,}903 x_i$.

b) $\hat{\beta}_1 = -5{,}903$. Logo, por cada aumento de 1 euro (uma unidade) no preço do *whisky* espera-se um decréscimo *médio* de 5,903 na quantidade vendida.

c) $gl_R = df_R = 1$;

$gl_E = df_E = n - 2 = 8 - 2 = 6$;

$gl_T = df_T = n - 1 = 8 - 1 = 7$;

$SQR = SQT - SQE = 529{,}689 - 28{,}613 = 501{,}076$;

$$MQR = \frac{SQR}{gl_R} = SQR = 501{,}076;$$

$$MQE = \frac{SQE}{gl_E} = SQR = \frac{28{,}613}{6} = 4{,}769;$$

$$f_{obs} = \frac{MQR}{MQE} = \frac{501{,}076}{4{,}769} = 105{,}073.$$

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	501,076	1	501,076	105,073	,00[b]
	Residual	28,613	6	4,769		
	Total	529,689	7			

d) $\hat{\sigma}^2 = MQE = 4{,}769$.

e) Coeficiente de determinação:

$$r^2 = \frac{SQR}{SQT} = \frac{501{,}706}{529{,}689} = 0{,}946.$$

Assim, 94,6% da variabilidade que se verifica na quantidade vendida é explicada pela reta de regressão ajustada, ou seja, pela relação linear com o preço do *whisky*. Portanto, foi efetuado um bom ajustamento.

f) $r = \pm\sqrt{r^2}$.

Como r tem sempre o mesmo sinal que $\hat{\beta}_1$ então, neste caso $r = -\sqrt{r^2} = -\sqrt{0{,}946} = -0{,}973$.

Existe relação linear negativa muito forte entre a quantidade vendida e o preço, em euros, do whisky, ou seja, para um preço elevado do whisky espera-se que a quantidade vendida seja baixa, enquanto que para um preço baixo espera-se uma quantidade vendida elevada.

g) $\alpha = 1\%$, $\beta_0 = 0$?

H_0: $\beta_0 = 0$ vs H_1: $\beta_0 \neq 0$ (teste bilateral)

Estatística de teste:

$$T = \frac{\hat{\beta}_0 - b_0}{\sqrt{\widehat{Var}(\hat{\beta}_0)}} \sim t_{n-2=6}.$$

$t_{obs} = 12{,}024$.

valor $p < 0{,}001$.

Como $\alpha = 0{,}01 \geq$ valor p rejeita-se H_0. Portanto, ao nível de significância de 1%, existe evidência estatística para afirmar que a reta de regressão não passa pela origem.

h) $\alpha = 1\%$, $\beta_1 \neq 0$?

H_0: $\beta_1 = 0$ *vs* H_1: $\beta_1 \neq 0$ (teste bilateral)

Estatística de teste:

$$T = \frac{\hat{\beta}_1 - b_0}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{n-2=6}$$

$t_{obs} = \dfrac{-5{,}903 - 0}{0{,}576} = -10{,}248.$

$valor\text{-}p < 0{,}001$.

Como $\alpha = 0{,}01 \geq$ valor p rejeita-se H_0. Portanto, ao nível de significância de 1%, existe evidência estatística para afirmar que o preço influencia de forma linear a quantidade vendida.

i) I. C. a $100(1-\alpha)\%$ para β_0:

$$\left]\hat{\beta}_0 - t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\beta}_0)};\ \hat{\beta}_0 + t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\beta}_0)}\right[$$

Como,

$$\sqrt{\widehat{Var}(\hat{\beta}_0)} = \frac{\hat{\beta}_0}{t_{obs_{\beta_0}}} = \frac{2135{,}691}{12{,}024} = 11{,}285,$$

o I. C. a 95% para β_0 é:

$$\left]135{,}691 - t_{6;\,0{,}975} \times 11{,}285;\ 135{,}691 + t_{6;\,0{,}975} \times 11{,}285\right[$$
$$= \left]135{,}691 - 2{,}447 \times 11{,}285;\ 135{,}691 + 2{,}447 \times 11{,}285\right[= \left]108{,}077;\ 163{,}305\right[.$$

Com 95% de confiança o verdadeiro valor de β_0 situa-se entre 108,077 e 163,305.

j) I. C. a $100(1-\alpha)\%$ para β_1:

$$\left]\hat{\beta}_1 - t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\beta}_1)};\ \hat{\beta}_1 + t_{n-2\,;\,1-\frac{\alpha}{2}}\sqrt{\widehat{Var}(\hat{\beta}_1)}\right[.$$

Logo, o I. C. a 95% para β_1 é:

$$]-5{,}903 - t_{6;\,0{,}975} \times 0{,}576;\ -5{,}903 + t_{6;\,0{,}975} \times 0{,}576[$$
$$=]-5{,}903 - 2{,}447 \times 0{,}576;\ -5{,}903 + 2{,}447 \times 0{,}576[=]-7{,}312;\ -4{,}494[.$$

Com 95% de confiança o verdadeiro valor de β_1 situa-se entre -7,3122 e -4,4938.

k) Usando os valores obtidos nas alíneas anteriores:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	135,691	11,285	12,024	,000	108,077	163,305	135,691
	x	-5,903	0,576	-10,250	,000	-7,312	-4,494	-5,903

l) $x_h = 1{,}23 \Rightarrow \hat{y}_h = ?$

$\hat{y}_h = 135{,}691 - 5{,}903 \times 10 = 76{,}661.$

Quando o preço do *whisky* é 10 euros espera-se que a quantidade vendida seja, *em média*, 76,661.

11.9 Exercícios propostos

1. Durante o desenvolvimento de um novo medicamento para alergias, foi realizada uma experiência para estudar o efeito de diferentes dosagens, no período de tempo que os doentes se libertam dos sintomas alérgicos. Foram incluídos na experiência 10 pacientes. A cada um foi dada uma dosagem especifica do medicamento, e foi-lhes pedido que comunicassem, de imediato, assim que o efeito desaparecesse.
As dosagens (x) foram medidas em miligramas e o tempo de duração do medicamento (y) em dias. Os resultados obtidos foram os seguintes:

$$\sum_{i=1}^{10} x_i = 59; \qquad \sum_{i=1}^{10} y_i = 151; \qquad \sum_{i=1}^{10} x_i^2 = 389; \qquad \sum_{i=1}^{10} y_i^2 = 2651; \qquad \sum_{i=1}^{10} x_i\, y_i = 1003.$$

a) Construa o modelo de regressão linear simples que explique a duração do efeito do medicamento em função da sua dosagem.
b) Determine os coeficientes de correlação e de determinação. Interprete.
c) Teste a significância da regressão.
d) Qual a previsão para o número de dias que um paciente se liberta dos sintomas alérgicos, se lhe forem administrados 6,5 mg do medicamento.

2. Suponha que se utiliza um novo método para determinar o montante de magnésio na água do mar. Se o método for bom, haverá uma forte relação entre o montante real de magnésio na água do mar e o montante indicado por este novo método. Foram preparadas 10 amostras de água do mar, cada uma contendo um montante conhecido de magnésio a fim de serem testadas pelo novo método.
Os dados desta experiência são apresentados na forma de estatísticas, onde x representa o montante real de magnésio presente e y o montante determinado pelo novo método.

$$\sum_{i=1}^{10} x_i = 311; \quad \sum_{i=1}^{10} y_i = 310{,}1; \quad \sum_{i=1}^{10} (x_i - \overline{x})^2 = 427{,}9; \quad \sum_{i=1}^{10} (y_i - \overline{y})^2 = 438{,}89;$$

$$\sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y}) = 429{,}89.$$

a) Determine a equação de regressão, o desvio padrão estimado, $\hat{\sigma}$, e o coeficiente de determinação, r^2.
b) Construa intervalos de confiança a 95 % para β_0 e β_1 e teste as hipóteses de que a verdadeira equação tenha parâmetros $\beta_0 = 0$ e $\beta_1 = 1$.

3. Com base numa amostra de 306 estudantes num curso de enfermagem foi estimada a seguinte reta de regressão:

$$\hat{y}_i = 58{,}813 + 0{,}2875x_i,$$

Onde y representa a nota final do aluno no fim do curso e x a nota no teste escrito dado no início do curso. O coeficiente de determinação foi 0,1158, e o desvio padrão estimado para o estimador do declive da reta foi 0,04566.

a) Interprete o declive da reta de regressão estimada.
b) Interprete o coeficiente de determinação.
c) Ao nível de significância de 5%, teste a hipótese de ausência de capacidade explicativa da variável independente.

4. Os seguintes dados constituem os valores relativos ao volume de vendas de gelados (y) e da temperatura (x) observados em cinco dias, em Évora.

Vendas (euros)	324	224	324	249	299
Temperatura ao meio dia (°C)	40	24	36	28	32

a) Determine a reta dos mínimos quadrados e interprete os coeficientes de regressão.
b) Represente graficamente a nuvem de pontos e a reta ajustada.
c) Calcule o resíduo de estimação quando a temperatura for 36°C.
d) Calcule o coeficiente de correlação e comente.
e) Determine e interprete o coeficiente de determinação.
f) Estime a variância do erro do modelo.
g) Construa a tabela ANOVA.
h) Complete: "Com 95% de confiança o verdadeiro valor de β_0 situa-se entre ... e ...".
i) Construa um intervalo e confiança a 99% para β_1.
j) A partir de que nível de significância é rejeitada a hipótese do coeficiente β_0 ser nulo?
k) Ensaie a hipótese de que a temperatura não influencia linearmente o volume de vendas de gelados (considere $\alpha = 1\%$).
l) Estime o volume de vendas correspondente a uma temperatura de 35°C.

5. Um investigador desenvolveu um novo programa de ensino individualizado que acredita aumentar o QI dum indivíduo. Escolheram-se aleatoriamente 8 alunos do ensino do 2º ciclo para participar em tal programa. Na tabela seguinte apresentam-se os seus QI registados, de forma equivalente, antes (x) e depois (y) do referido programa:

Aluno	1	2	3	4	5	6	7	8
Antes	81	89	90	97	108	111	118	124
Depois	89	88	94	96	118	111	121	121

O mesmo investigador acredita que a nota obtida pelo aluno depois de participar no referido programa está relacionada de forma linear com a nota obtida antes da participação.

a) Determine a reta dos mínimos quadrados e interprete os coeficientes de regressão.
b) Represente graficamente a nuvem de pontos e a reta ajustada.
c) Calcule o resíduo de estimação quando o QI antes é 90.
d) Calcule o coeficiente de correlação e comente.
e) Determine e interprete o coeficiente de determinação. Considera que efetuou um bom ajustamento?
f) Estime a variância do erro do modelo.
g) Construa a tabela ANOVA.
h) Construa um intervalo e confiança a 99% para β_0.
i) Complete: "Com 95% de confiança o verdadeiro valor de β_1 situa-se entre ... e ...".
j) Teste, ao nível de significância de 1% a hipótese da reta de regressão passar pela origem ($\beta_1 = 0$).
k) A partir de que nível de significância é rejeitada a hipótese de que a nota antes do programa não influência linearmente a nota depois do programa.
l) A nota de um aluno antes de participar no programa é 93. Que nota espera que ele tenha depois de participar no programa?
m) Valide os pressupostos subjacentes ao modelo.

6. Pretende-se verificar se, nos jogos do campeonato do Mundo de 2018, o número de remates dependia do n.º de cruzamentos efetuados por jogo. Para tal selecionaram-se ao acaso 9 jogos, tendo-se obtido os seguintes resultados.

Número de remates	19	15	12	16	11	28	22	8	16
Número de cruzamentos	45	25	22	32	28	56	38	10	32

Admitindo que os pressupostos do modelo são verificados, responda às seguintes questões:

a) Através da análise gráfica, considera que existe relação linear entre o número de remates à e o número de cruzamentos? Justifique.
b) Ao nível de significância de 5%, considera que existe relação linear positiva entre as duas variáveis?
c) Qual a equação da reta de regressão ajustada?
d) Interprete os coeficientes de regressão estimados.
e) Considera que efetuou um bom ajustamento? Justifique.
f) Ao nível de significância de 1%, ensaie a hipótese de que a reta de regressão estimada passa pela origem.
g) Ao nível se significância de 5%, pode-se considerar que $\beta_1 = 0$?
h) Quantos remates espera que se tenham realizado num jogo em que se fizeram 56 cruzamentos?
i) Calcule e interprete o resíduo de estimação da alínea anterior.
j) Valide os pressupostos subjacentes ao modelo.

7. Sabe-se que a procura anual de café (Y) está relacionada, linearmente, com o respetivo preço (X) do seguinte modo: $Y = \beta_0 + \beta_1 x + \varepsilon$. Durante 10 anos e num determinado país, registou-se para a variável Y, o número médio de chávenas de café consumidas por pessoa e por dia e para a variável X o respetivo preço, expresso em unidades monetárias. Com base na informação de uma amostra e usando o método dos mínimos quadrados, obtiveram-se os seguintes resultados de estimação:

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2,691	0,122					
	x	-0,479	0,114					

a) Interprete $\hat{\beta}_2$no modelo estimado.
b) Construa um intervalo de 95% de confiança para β_0. Interprete-o.
c) Construa um intervalo de 95% de confiança para β_1. Interprete-o.
d) Para um nível de significância de 5% ensaie a hipótese de ausência de capacidade explicativa da variável X.
e) Ensaie a hipótese de que a reta de regressão estimada passa pela origem. (Utilize um nível de significância de 5%)
f) Admita que o $\text{valor } p$ para este exercício foi inferior a 0,001. Para um nível de significância de 5%, que decisão tomaria face a uma qualquer hipótese nula?
g) Complete a tabela anterior.

12 Um caso de estudo baseado no Inquérito Nacional de Saúde, utilizando o SPSS

Este capítulo pretende ser uma revisão de algumas das metodologias apresentadas neste livro, com uma forte componente prática, totalmente resolvido com base no SPSS (versão 25), utilizando um conjunto de dados reais e constituindo um apoio organizado na aplicação da Estatística. Para além do referido, apresenta-se no final deste capítulo uma secção de apoio à introdução de dados neste programa de estatística. Note-se que, neste capítulo, não é possível percorrer todas as matérias apresentadas ao longo desta obra, nem se pretende substituir a consulta de um manual de apoio a este programa.

12.1 Apresentação do caso de estudo

O Inquérito Nacional de Saúde (INS), promovido e financiado pelo Ministério da Saúde, é um instrumento de medida e de observação em saúde, que recolhe dados de base populacional, que permite gerar estimativas sobre alguns estados de saúde e de doença da população portuguesa, bem como as respetivas determinantes e estudar a sua evolução ao longo do tempo. O INS foi planeado e testado pela primeira vez entre 1980 e 1982. Após inquéritos de âmbito regional, conduzidos entre 1983 e 1985, realizou-se em 1987 o primeiro INS, cobrindo o Continente português (Dias e Graça, 2001).

Até à data foram já realizados cinco INS (1987, 1995/1996, 1998/1999, 2005/2006, 2014) utilizando amostras probabilísticas representativas da população de Portugal Continental onde o penúltimo (2005/2006) incluiu já as Regiões Autónomas dos Açores e Madeira, com mais de 41000 entrevistas efetuadas por inquérito.

Como forma de exemplificar a utilidade da estatística na área específica da Saúde e também como forma de transmitir alguns conceitos base da utilização do SPSS (versão 25), disponibiliza-se um ficheiro com 1000 casos (cerca de 5%) retirados aleatoriamente do INS de 1995/1996 (Ministério da Saúde, 1997), disponibilizado em http://evunix.uevora.pt/~aafonso/ com informação de 16 variáveis apresentadas na Tabela 12.1. A representatividade desta amostra não está garantida e estes dados apenas poderão ser utilizados por questões pedagógicas como caso de estudo, sempre referindo a sua origem.

Tabela 12.1: *Apresentação das variáveis em análise com indicação das principais propriedades no SPSS.*

Nome da Variável (Name)	Descrição (Label)	Tipo[†] (Measure)	Codificação (Values)	Valores omissos[‡] (Missing)
Regiao	Região de residência	Nominal	1 = Norte; 2 = Centro; 3 = LVT; 4 = Alentejo; 5 = Algarve.	---------
Sexo	Sexo	Nominal	1 = Masculino; 2 = Feminino.	---------
Idade	Idade	Scale	---------	---------

[†] O SPSS só classifica as variáveis quantitativas em" Scale", não descriminando entre variáveis discretas e contínuas.
[‡] Valores utilizados como códigos para valores omissos.

Tabela 12.1: *Apresentação das variáveis em análise com indicação das principais propriedades no SPSS. (continuação)*

Nome da Variável (Name)	Descrição (Label)	Tipo[†] (Measure)	Codificação (Values)	Valores omissos[‡] (Missing)
Civil	Estado civil	Nominal	1 = Casado; 2 = Solteiro; 3 = Casado; 4 = Viúvo.	--------
Escolar	Número de anos de escolaridade concluídos com aproveitamento	Scale	--------	--------
Autoapreciacao	Autoapreciação do estado de saúde	Ordinal	1 = Muito Bom; 2 = Bom; 3 = Razoável; 4 = Mau; 5 = Muito Mau.	9
Altura	Altura	Scale	--------	999
Peso	Peso	Scale	--------	999
Servicos	Serviços a que recorre mais vezes para benefícios de saúde	Nominal	1 = ADSE; 2 = Militares; 3 = SAMS; 4 = SNS; 5 = Outros.	96, 99
Sofre_Diabetes	Sofre de diabetes	Nominal	1 = Sim; 2 = Não.	--------
Idade_Diabetes	Desde que idade sofre diabetes	Scale	--------	--------
Quem_Diabetes	Quem lhe disse que sofre de diabetes	Nominal	1= Medico ou Enfermeiro 2 = Outros.	--------
Sofre_Tensãoalta	Sofre de tensão alta	Nominal	1 = Sim; 2 = Não.	--------
Idade_Tensãoalta	Desde que idade sofre de tensão alta	Scale	--------	--------
Servico_Avalia	Como considera o serviço dos médicos nos Centros de Saúde	Ordinal	1 = Muito Bom; 2 = Bom; 3 = Razoável; 4 = Mau; 5 = Muito Mau.	9
Ndiasbebe_Semana	Nº dias em que bebeu bebidas alcoólicas na última semana	Scale	--------	9

As questões que se pretendem ver resolvidas apresentam-se de seguida assim como as respetivas soluções.

[†] O SPSS só classifica as variáveis quantitativas em" Scale", não descriminando entre variáveis discretas e contínuas.
[‡] Valores utilizados como códigos para valores omissos.

12.2 Análise do caso de estudo

Nesta secção pretende-se caracterizar o perfil dos inquiridos relativamente às seguintes questões:

- Qual ou quais as regiões mais amostradas?
- Como os inquiridos definem o seu estado de saúde?
- Quais os hábitos de consumo de bebidas alcoólicas? Será que é semelhante nos dois
- sexos?
- Será que o estado civil e a idade dos inquiridos se distribuem de igual forma nos dois
- sexos?
- Nos inquiridos, a diabetes é diagnosticada com base em pareceres médicos?
- Qual a distribuição do índice de massa corporal (IMC) por sexo?
- Qual a idade média da população portuguesa? Será que é idêntica nos dois sexos?
- E nas diferentes regiões do país?
- É admissível considerar que em média os homens portugueses têm 1,70m? E a altura
- média dos portugueses será semelhante em todas as regiões?
- Será que a diabetes e a tensão alta surgem em média na mesma idade?
- Estarão o estado civil e a diabetes relacionados com o sexo?
- Fará sentido considerar que existe relação entre a idade e a altura?
- Será que a forma como definem o seu estado de saúde estará relacionada com o IMC?
- Existirá relação linear entre a altura e o peso? Caracterize-a.

12.3 Caracterização da variável Região

Construção da tabela de frequências e elaboração de um gráfico.

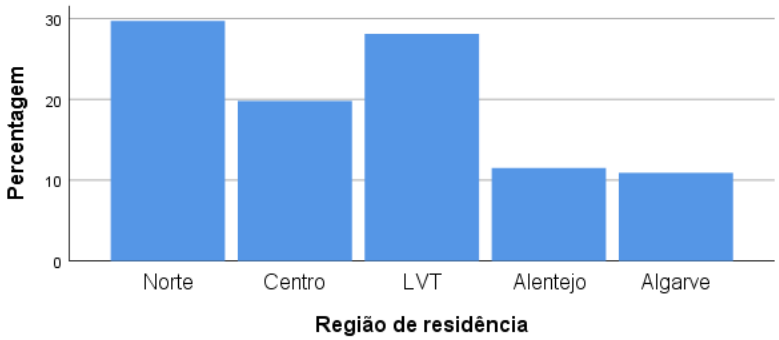
(SPSS) Analyze → Descriptive statistics → Frequencies
(Variable(s): Região de residência [regiao]; ☑ Display frequency tables; Charts → Chart Type: ⊙ Bar charts; Chart Values: ⊙ Percentages)

Statistics
região de residência

N	Valid	1000
	Missing	0

Região de residência

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Norte	297	29,7	29,7	29,7
	Centro	198	19,8	19,8	49,5
	LVT	281	28,1	28,1	77,6
	Alentejo	115	11,5	11,5	89,1
	Algarve	109	10,9	10,9	100,0
	Total	1000	100,0	100,0	



Pelos resultados anteriores verifica-se que não existem valores omissos (são os próprios entrevistadores que preenchem este campo) e a região mais amostrada foi o Norte (29,7%). A região LVT apresenta frequências próximas da região Norte (cerca de 28%), as regiões Alentejo e Algarve apresentam as frequências mais baixas (cerca de 11%) e a região Centro uma percentagem de 19,8%. A coluna "*Cumulative Percent*" não faz sentido ser analisada para variáveis do tipo Nominal (não têm ordem).

12.3.1 Caracterização da amostra em relação ao estado civil, autoapreciação do estado de saúde e número de dias em que bebeu bebidas alcoólicas na última semana

As tabelas de frequências fazem sentido para descrever a informação de variáveis que não tomem um n.º muito elevado de valores, categorias ou classes, distintos. Quando tal não acontece é necessário proceder a agrupamento/reagrupamento dos valores, categorias ou classes, de forma a tornar interpretável a leitura da informação. As variáveis *Civil* (nominal), *Autoapreciacao* (ordinal com 5 níveis) e *Ndiasbebe_Semana* (discreta com valores de 0 a 7), pelas suas naturezas respeitam este princípio.

(SPSS) Analyze → Descriptive statistics → Frequencies
(Variable(s): Estado civil [Civil], Autoapreciação do estado de saúde [Autoapreciacao], N.º de dias em que bebeu bebidas alcoólicas na última semana [Ndiasbebe_Semana])

Statistics

		Estado civil	Auto-apreciação do estado de saúde	N.º dias em que bebeu bebidas alcoolicas na última semana
N	Valid	1000	440	997
	Missing	0	560	3

Estado civil

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Casado	539	53,9	53,9	53,9
	Solteiro	396	39,6	39,6	93,5
	Separado	11	1,1	1,1	94,6
	Viúvo	54	5,4	5,4	100,0
	Total	1000	100,0	100,0	

Autoapreciação do estado de saúde

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Muito Bom	11	1,1	2,5	2,5
	Bom	117	11,7	26,6	29,1
	Razoável	196	19,6	44,5	73,6
	Mau	99	9,9	22,5	96,1
	Muito Mau	17	1,7	3,9	100,0
	Total	440	44,0	100,0	
Missing	System	560	56,0		
Total		1000	100,0		

N.º dias em que bebeu bebidas alcoolicas na última semana

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	587	58,7	58,9	58,9
	1	34	3,4	3,4	62,3
	2	21	2,1	2,1	64,4
	3	14	1,4	1,4	65,8
	4	11	1,1	1,1	66,9
	5	13	1,3	1,3	68,2
	6	3	,3	,3	68,5
	7	314	31,4	31,5	100,0
	Total	997	99,7	100,0	
Missing	9	3	,3		
Total		1000	100,0		

Observa-se que a variável estado civil não teve valores omissos, que a variável autoapreciação do estado de saúde teve 56% de valores omissos, uma percentagem claramente indiciando problemas na análise desta variável, e em relação ao consumo de álcool houve uma percentagem desprezável de 3% de valores omissos. No caso do estado civil interessa realçar que a maior parte das pessoas são casadas (cerca de 54%) ou solteiras (39,6%), perfazendo um total de 93,5%. As outras classes têm frequências muito baixas, sendo de estranhar a percentagens de separados (apenas 1,1%).

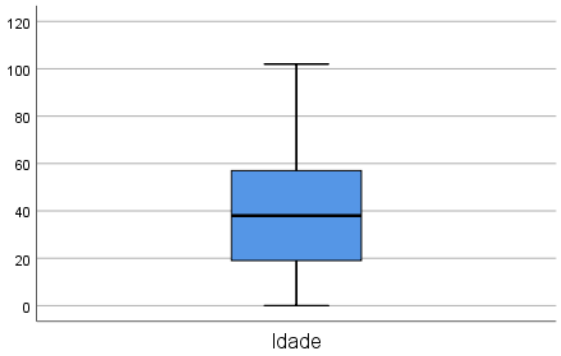
Na autoapreciação do estado de saúde observam-se pequenas percentagens nos extremos (muito bom ou muito mau) e um equilíbrio entre as classes bom e mau. A classe razoável é a mais representada (com 44,5% dos respondentes). Interessa relembrar a que percentagem muito elevada de valores omissos.

Em relação ao número de dias em que bebe álcool, verifica-se que nos extremos houve valores elevados: 58,5% dos respondentes amostrados nunca bebe álcool e 31,5% bebe todos os dias. É residual a quantidade de pessoas que têm um comportamento intermédio (cerca de 10% para os restantes dias).

12.3.2 Descrição da amostra recolhida em termos de idade

Como primeira abordagem deve utilizar-se a estatística descritiva, elaborando um gráfico e calculando as estatísticas que fazem sentido para o tipo variáveis em estudo, de forma descrever o conjunto de dados recolhido.

(SPSS) Graphs → Legacy dialogs → Boxplot
(Simple → ⊙ Summaries of separate variables; Boxes Represent: Idade)



(SPSS) Analyze → Descriptive statistics → Descriptives
(Variable(s): Idade;
Options… → ☑ Mean; ☑ Std. Deviation; ☑ Mininum; ☑ Maximum; ☑ Kurtosis; ☑ Skewness)

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Idade	1000	0	102	38,81	22,339	,165	,077	-1,056	,155
Valid N (listwise)	1000								

A idade varia entre 0 e 102 anos com uma média de 38,81 anos, um desvio *típico* em relação à média de 22,339 (desvio padrão) indicando que esta amostra abrange todas as classes etárias. Apresenta uma assimetria (*skewness*) à direita – tendência para uma estrutura etária mais nova (0,165>0) e um achatamento (*kurtosis*) não muito representativo dado que a variável não é simétrica.

12.3.3 Análise estatística das variáveis idade e estado civil pela variável sexo

A análise estatística compreende o cálculo das medidas descritivas bem como a representação gráfica e tabular.

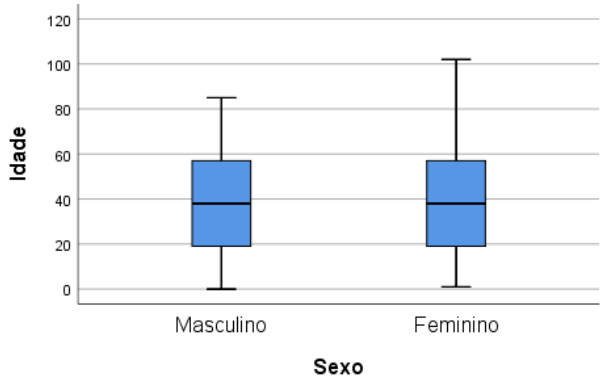
Idade por variável sexo:

(SPSS) Analyze → Descriptive statistics → Explore
(Dependent List: Idade [idade]; Factor List: Sexo [sexo]; Display: ⊙ Both;
Statistics → ☑ descriptive;
Plots → Boxplots: ⊙ Factor levels together)

Descriptives

	Sexo			Statistic	Std. Error
Idade	Masculino	Mean		38,56	,919
		95% Confidence Interval for Mean	Lower Bound	36,76	
			Upper Bound	40,37	
		5% Trimmed Mean		38,37	
		Median		38,00	
		Variance		490,008	
		Std. Deviation		22,136	
		Minimum		0	
		Maximum		85	
		Range		85	
		Interquartile Range		38	
		Skewness		,113	,101
		Kurtosis		-1,111	,203

	Feminino	Mean		39,14	1,105
		95% Confidence Interval for Mean	Lower Bound	36,97	
			Upper Bound	41,31	
		5% Trimmed Mean		38,71	
		Median		38,00	
		Variance		512,450	
		Std. Deviation		22,637	
		Minimum		1	
		Maximum		102	
		Range		101	
		Interquartile Range		38	
		Skewness		,232	,119
		Kurtosis		-,995	,238



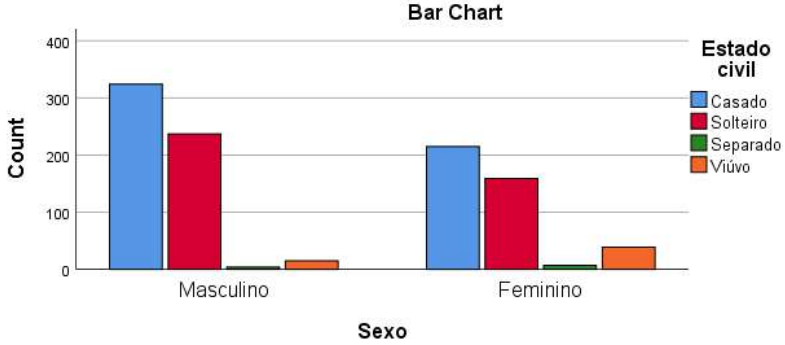
Em relação à idade por sexo verifica-se que as distribuições têm aparentemente características semelhantes com idades médias/medianas de cerca de 39 anos e desvio padrão de 22 anos. A diferença mais pertinente centra-se na idade máxima (85 anos para o sexo masculino e 102 anos para o feminino).

Estado civil pela variável sexo:

(SPSS) Analyze → Descriptive statistics → Crosstabs
(Row(s): Sexo [sexo]; Column(s): Estado Civil [Civil]; ☑ Display clustered bar charts;
Cells → Counts: ☑ Observed; Percentages: ☑ Row)

Sexo * Estado civil Crosstabulation

			Estado civil				
			Casado	Solteiro	Separado	Viúvo	Total
Sexo	Masculino	Count	324	237	4	15	580
		% within Sexo	55,9%	40,9%	0,7%	2,6%	100,0%
	Feminino	Count	215	159	7	39	420
		% within Sexo	51,2%	37,9%	1,7%	9,3%	100,0%
Total		Count	539	396	11	54	1000
		% within Sexo	53,9%	39,6%	1,1%	5,4%	100,0%



Em relação à variável estado civil por sexo verifica-se que as distribuições são diferentes em termos de frequências absolutas. No entanto, em termos percentuais as distribuições são similares, exceto no que diz respeito à categoria dos viúvos que é muito mais elevada no sexo feminino.

12.3.4 Os Portugueses diagnosticam a diabetes com base em pareceres médicos ou não?

Esta questão só tem significado para as pessoas que declararam sofrer de diabetes. Deste modo, primeiro é preciso selecionar este grupo de pessoas para depois analisar "Quem lhe disse que sofre de diabetes?" através, por exemplo de uma tabela de frequências.

(SPSS) Data → Select Cases
(⊙ If condition is satisfied; If → Sofre_diabetes = 1)
Analyze → Descriptive statistics → Frequencies
(Variable(s): Quem lhe disse que sofre de diabetes [Quem_diabetes]; ☑ Display frequency tables)

Quem lhe disse que sofre de diabetes

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Med ou Enf	48	100,0	100,0	100,0

Todas as pessoas que declararam sofrer desta doença, afirmaram que foram profissionais de saúde que a diagnosticaram.

Atenção: no final é necessário remover o filtro aplicado aos dados.
(SPSS) Data → Select Cases
(⊙ All cases)

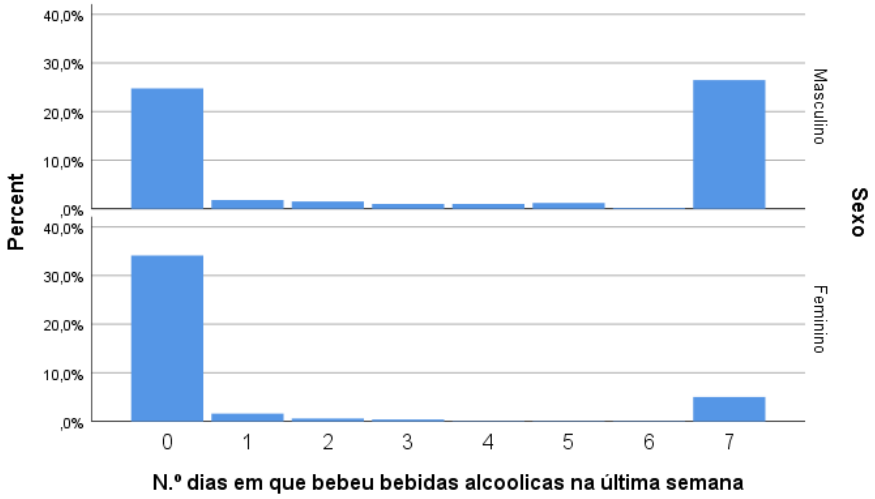
12.3.5 Será que a maioria das pessoas consume álcool todos os dias? Comportam-se de igual maneira nos dois sexos?

Pode abordar-se este problema pedindo primeiro uma tabela de frequências da variável "N.º dias em que bebeu bebidas alcoólicas na última semana" (já efetuada na secção 12.3.1) e um gráfico de barras (ver instruções na secção 12.3).

Como já tinha sido referido na secção 12.3.1, existem comportamentos extremos, ou consumem todos os dias (31,4%) ou nunca consomem (58,9%), correspondendo a 90,4% dos respondentes a esta questão. Os restantes comportamentos (entre 1 a 6 dias) são residuais.

Para analisar esta variável por sexos, propõe-se aqui uma análise gráfica.

(SPSS) Graphs → Legacy dialogs → Bar
(Simple; ⊙ Summaries for Groups of cases
Bars represent: ⊙ % of cases; Category Axis: Nº dias em que bebeu bebidas alcoólicas na última semana [Ndiasbebe_semana]; Panel By: Rows: Sexo [Sexo])



As distribuições não são semelhantes nos extremos, continuando o comportamento intermédio a ser residual. No sexo masculino existe um equilibro entre as proporções de quem bebeu todos os dias e de quem não bebeu, enquanto que no sexo feminino já existe uma tendência clara para não ter bebido em nenhum dos dias.

12.3.6 Construção duma nova variável Índice de Massa Corporal (IMC) e sua codificação

A variável Índice de Massa Corporal (IMC) é determinada pela expressão

$$IMC = \frac{\text{peso, em kg}}{(\text{altura, em metros})^2}.$$

A interpretação habitual dos valores obtidos para o IMC é a apresentada na Tabela 12.2.

Tabela 12.2: *Interpretação dos valores de IMC.*

Definição	Condição
Abaixo do peso	< 20
No peso normal	20-25
Marginalmente acima do peso	25-28
Obeso	>28

O primeiro passo consiste em construir a variável IMC para todos os respondentes.

(SPSS) Transform → Compute Variable
(Target Variable: IMC; Numeric Expression: Peso/((Altura/100)**2))

Seguidamente será criada uma nova variável de tipo ordinal à qual será associada a codificação dos diferentes valores de IMC.

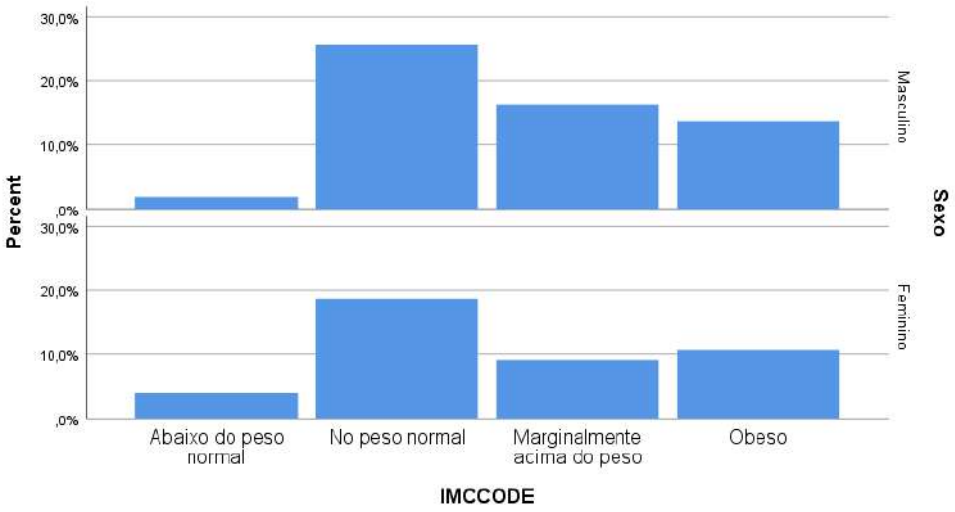
(SPSS) Transform → Recode into Different Variables
(Input Variable: IMC; Output Variable: Name: IMCCODE; Change;
Old and New Values → ⊙ Range, LOWEST through values: 19,999; New Value: Value: 1; Add;
⊙ Range 20 through 24,999; New Value: Value: 2; Add; ⊙ Range 25 through 27,999; New Value: Value: 3; Add; ⊙ Range, value through HIGHEST: 28; New Value: Value: 4; Add)

No separador *Variable View* da janela *SPSS Statistics Data Editor* classificar a variável corretamente: na coluna *Measure* escolher *Ordinal* e na coluna *Values* definir a denominação de cada valor:

1 = Abaixo do peso normal
2 = No peso normal
3 = Marginalmente acima do peso
4 = Obeso

12.3.7 Análise da distribuição do IMC codificado por sexo.

Esta análise pode ser efetuada através da construção de um gráfico de barras para cada sexo (ver instruções na secção 12.3.5)



Existem diferenças entre as distribuições nos dois sexos, com uma ligeira inversão de tendências nas classes superiores e existe uma maior percentagem de mulheres com um IMC inferior ao Normal. A classe mais observada em ambos os sexos é o peso normal, embora essa percentagem devesse ser superior para corresponder a uma situação favorável no âmbito da Saúde Pública.

12.3.8 Construção do intervalo de confiança a 90% para a média da idade da população portuguesa

(SPSS) Analyze → Compare Means → One-Sample T Test
(Test Variable(s): Idade [idade]; Test Value: 0;
Options → Confidence Interval: 90)

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Idade	1000	38,81	22,339	,706

One-Sample Test

	Test Value = 0					
					90% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Idade	54,933	999	,000	38,805	37,64	39,97

Com 90% de confiança, a idade média da população portuguesa situa-se entre 37,64 anos e 39,97.

12.3.9 Construção do intervalo de confiança a 95% para a idade média por sexo

Para fazer intervalos de confiança ou testes de hipótese por grupos, é necessário indicar qual a variável que contém os grupos.

(SPSS) Data → Split File
(⊙ Organize output by groups: Groups Based on: Sexo)

(SPSS) Analyze → Compare Means → One-Sample T Test
(Test Variable(s): Idade; Test Value: 0; Options → Confidence Interval: 95)

Sexo = Masculino

One-Sample Statistics[a]

	N	Mean	Std. Deviation	Std. Error Mean
Idade	580	38,56	22,136	,919

a. Sexo = Masculino

One-Sample Test[a]

	Test Value = 0					
			Sig.	Mean	95% Confidence Interval of the Difference	
	t	df	(2-tailed)	Difference	Lower	Upper
Idade	41,954	579	,000	38,562	36,76	40,37

a. Sexo = Masculino

Sexo = Feminino

One-Sample Statistics[a]

	N	Mean	Std. Deviation	Std. Error Mean
Idade	420	39,14	22,637	1,105

a. Sexo = Feminino

One-Sample Test[a]

	Test Value = 0					
			Sig.		95% Confidence Interval of the Difference	
	t	df	(2-tailed)	Mean Difference	Lower	Upper
Idade	35,434	419	,000	39,140	36,97	41,31

a. Sexo = Feminino

Com 95% de confiança a idade média da população masculina situa-se entre 36,76 anos e 40,37 anos, e a da população feminina encontra-se entre os 36,97 anos e os 41,31 anos.
Note-se que na secção 12.3.3, ao utilizar o comando Explore, o resultado também já continha a resposta a esta questão.

Atenção: no final é necessário remover a divisão da informação por grupos.

(SPSS) Data → Split File
(⊙ Analyze all cases, do not create groups)

12.3.10 Será que existe diferença entre as médias de idades por sexo, com 99% de confiança?

Para averiguar se, com 99% de confiança, existe evidência de diferença entre as médias de idades por sexo, pode recorrer-se a um intervalo de confiança, ou a um teste de hipótese, para a comparação de duas médias de amostras independentes.

(SPSS) Analyze → Compare Means → Independent-Samples T Test
(Test Variable(s): Idade; Grouping variable: Sexo; Define Groups → Group 1: 1; Group 2: 2; Options → Confidence Interval: 99)

Group Statistics

	Sexo	N	Mean	Std. Deviation	Std. Error Mean
Idade	Masculino	580	38,56	22,136	,919
	Feminino	420	39,14	22,637	1,105

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									99% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Idade	Equal variances assumed	,435	,510	-,404	998	,686	-,578	1,432	-4,274	3,117
	Equal variances not assumed			-,403	891,0	,687	-,578	1,437	-4,288	3,131

O primeiro passo a efetuar na análise dos resultados é verificar se a igualdade das variâncias pode ou não ser considerada. Como para o Teste de Levene o valor $p = 0{,}510$, pode assumir-se que as variâncias são iguais. Desta forma, a informação a reter será a da primeira linha (*Equal variances assumed*). Pela análise do valor $p = 0{,}686$, não existe evidência da diferença entre as médias das idades dos 2 sexos, para qualquer nível de significância usual. Esta conclusão também é sustentada pelo I. C. a 99 %, que contém o valor 0, onde a margem da possível diferença entre a idade média da população masculina e feminina situa-se no intervalo]-4,274; 3,117[.

12.3.11 Será que a altura média dos homens é de 1,70m?

(SPSS) Data → Select Cases
(⊙ If condition is satisfied; If → sexo = 1)
Analyze → Compare Means → One-Sample T Test
(Test Variable(s): Idade; Test Value: 170)

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Idade	580	38,56	22,136	,919

One-Sample Test

	Test Value = 170					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Idade	-142,999	579	,000	-131,438	-133,24	-129,63

Com um $valor\ p < 0{,}001$ rejeita-se a hipótese de que a altura média dos homens possa ser de 1,70m, para qualquer nível de significância não nulo. Esta é apenas uma das formas de resolver este problema (por exemplo: poderia manter-se o *Test Value* a 0 e calcular o I. C. a 95%, verificando posteriormente se o valor 170, estava ou não contido no I. C.).

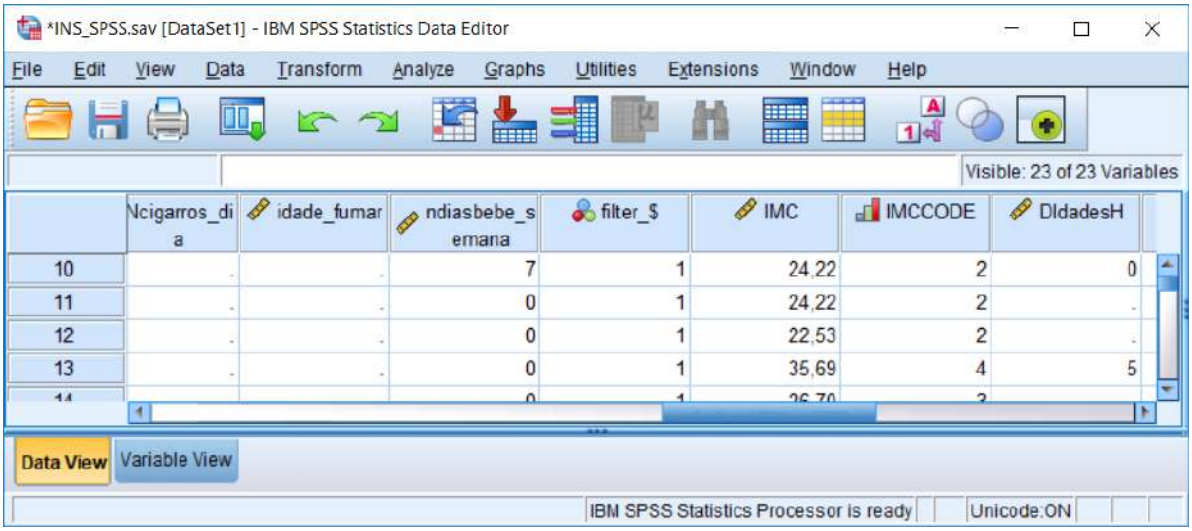
Atenção: no final é necessário remover o filtro aplicado aos dados.

(SPSS) Data → Select Cases
(⊙ All cases)

12.3.12 Será que a diabetes e a tensão alta surgem, em média, na mesma idade, com 95% de confiança, nos homens que sofrem das 2 doenças?

Uma vez que não se sabe quantos homens responderam a esta questão, podendo resultar num conjunto pequeno de respostas, para cada indivíduo vai ser criada uma nova variável que corresponde à diferença entre as idades a que surgiram as doenças.

(SPSS) Transform → Compute variable
(Target Variable: DIdadesH; Numeric Expression: idade_diabetes-idade_tensãoalta; If → ⊙ Include if case satisfies condition: sexo=1)



(SPSS) Analyze → Descritive Statistics → Explore…
(Dependent List: DIdadeH; Display: ⊙ Both;
Statistics → ☑ Descritives; Confidence Interval for Mean: 95%;
Plots → ☑ Normality plots with tests)

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
DIdadesH	16	1,6%	984	98,4%	1000	100,0%

Descriptives

			Statistic	Std. Error
DIdadesH	Mean		1,19	4,466
	95% Confidence Interval for Mean	Lower Bound	-8,33	
		Upper Bound	10,71	
	5% Trimmed Mean		1,99	
	Median		3,50	
	Variance		319,096	
	Std. Deviation		17,863	
	Minimum		-47	
	Maximum		35	
	Range		82	
	Interquartile Range		11	
	Skewness		-,965	,564
	Kurtosis		3,298	1,091

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
DIdadesH	,223	16	,032	,892	16	,061

a. Lilliefors Significance Correction

Pela informação da primeira tabela, apenas 16 homens responderam a ambas as questões (desde que idade sofrem de diabetes e tensão alta). Visto ser um conjunto de dados pequeno, é preciso avaliar se este conjunto de dados é proveniente de uma população Normal. Com base nos resultados do teste de Shapiro-Wilk (última tabela), $\text{valor } p = 0{,}061 > \alpha = 0{,}05$, ao nível de significância de 5% não há evidência estatística para afirmar que os dados não provêm de uma população com distribuição Normal. Desta forma, é possível obter um I.C. baseado na distribuição *t*-Student. Pela segunda tabela, o I.C. a 95% para a variável diferença entre as idades é $]-8{,}33; 10{,}71[$. O 0 está contido no I. C., logo não existe evidência da diferença entre as idades, i. e., no grupo dos homens assume-se que as doenças podem aparecer, em média, em idade semelhantes, com 95% de confiança.

O I.C. também poderia ser obtido sem ser necessário criar a variável diferença entre as idades, realizando os seguintes passos:

(SPSS) Data → Select Cases
(⊙ If condition is satisfied; If → sexo = 1)
Analyze → Compare Means → Paired-Sample T Test
(Paired Variables: Variable 1: Desde que idade começou a sofrer de diabetes [Idade_diabetes]; Variable 2: Desde que idade começou a sofrer de tensão alta [Idade_Tensãoalta]; Options → Confidence Interval: 95)

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Desde que idade sofre diabetes	50,88	16	16,657	4,164
	Desde que idade sofre de tensão alta	49,69	16	14,402	3,601

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Desde que idade sofre diabetes & Desde que idade sofre de tensão alta	16	,346	,190

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Desde que idade sofre diabetes - Desde que idade sofre de tensão alta	1,188	17,863	4,466	-8,331	10,706	,266	15	,794

Neste *output*, para além do I.C, também é apresentado o resultado do teste de hipótese para a igualdade entre as duas médias. Pela análise do valor $p = 0{,}794$, mantém-se a conclusão da não existência de diferença entre as médias para outros níveis de significância usualmente considerados (entre 1% e 10%).

Atenção: no final é necessário remover o filtro aplicado aos dados.

(SPSS) Data → Select Cases
(⊙ All cases)

12.3.13 Comparação da média de idades por região do país

Para comparar três ou mais médias em simultâneo pode utilizar-se a ANOVA (desde que se verifique a Normalidade e a homogeneidade das variâncias) ou em alternativa o teste de Kruskal-Wallis (caso os pressupostos anteriores não se verifiquem). Desta forma, o primeiro passo é verificar se os subgrupos seguem uma distribuição Normal ou, se pelo menos seguem distribuições simétricas e mesocúrticas.

(SPSS) Analyze → Descriptive Statistics → Explore
(Dependent list: Idade; Factor List: Região de residência [Regiao]; Display: ⊙ Plots; Plots → ☑ Normality plots with tests)

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Região de residência	Statistic	df	Sig.	Statistic	df	Sig.
Idade	Norte	,095	297	,000	,954	297	,000
	Centro	,095	198	,000	,956	198	,000
	LVT	,103	281	,000	,967	281	,000
	Alentejo	,093	115	,016	,965	115	,004
	Algarve	,092	109	,024	,959	109	,002

a. Lilliefors Significance Correction

Em virtude de todos os valores p serem ≤ 0,024 (através do teste de Kolmogorov-Smirnov com a correção de Lilliefors), não se pode assumir a Normalidade da idade em todas as regiões. No entanto, a ANOVA é robusta à Normalidade dos dados, desde que as subpopulações sejam simétricas e mesocúrticas.

(SPSS) Analyze → Compare Means → Means
(Dependent list: Idade; Independent List: Região de residência [Regiao]; Options → Cell Statistics: Mean, Number of Cases, Standard Deviation, Kurtosis, Std. Error of Kurtosis, Skweness, Std. Error of Skweness)

Report
Idade

Região de residência	Mean	N	Std. Deviation	Kurtosis	Std. Error of Kurtosis	Variance	Std. Error of Skewness
Norte	35,24	297	22,316	-,774	,282	497,998	,141
Centro	40,13	198	23,375	-1,151	,344	546,399	,173
LVT	38,52	281	21,458	-1,020	,290	460,457	,145
Alentejo	42,98	115	22,410	-1,086	,447	502,228	,226
Algarve	42,44	109	21,448	-1,087	,459	460,008	,231
Total	38,81	1000	22,339	-1,056	,155	499,012	,077

Também não se pode assumir que sejam simétricas e mesocúrticas, logo será mais adequado utilizar um teste não paramétrico Kruskall-Wallis.

(SPSS) Analyze → Nonparametric tests → Legacy Dialogs → K Independent Samples
(Test variable list: Idade; Grouping variable: Região de residência [Regiao]; Define Range: Minimum: 1; Maximum: 5; Test type: ☑ Kruskal-Wallis H)

Kruskal-Wallis Test
Ranks

	Região de residência	N	Mean Rank
Idade	Norte	297	452,83
	Centro	198	515,37
	LVT	281	498,58
	Alentejo	115	555,76
	Algarve	109	550,03
	Total	1000	

Test Statistics[a,b]

	Idade
Kruskal-Wallis H	16,047
df	4
Asymp. Sig.	,003

a. Kruskal Wallis Test
b. Grouping Variable: Região de residência

Com um valor $p = 0{,}003$ rejeita-se a hipótese de igualdade das distribuições e, consequentemente, das medianas. Para averiguar quais as medianas que diferem entre si, é necessário recorrer aos testes de comparações múltiplas apropriados (não abordados neste livro).

Apenas com um carácter pedagógico e exemplificativo irá analisar-se a igualdade das variâncias (teste de Levene) e comparar depois os resultados anteriores com a aplicação de um teste ANOVA.

(SPSS) Analyze → Compare Means → One-Way ANOVA
(Dependent list: Idade; Factor: Região de residência [Regiao];
Options → ☑ Homogeneity of variance test)

Oneway
Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Idade	Based on Mean	,713	4	995	,583
	Based on Median	,734	4	995	,569
	Based on Median and with adjusted df	,734	4	986,505	,569
	Based on trimmed mean	,712	4	995	,584

ANOVA
Idade

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7601,960	4	1900,490	3,852	,004
Within Groups	490911,015	995	493,378		
Total	498512,975	999			

Pela primeira tabela, como o valor $p = 0{,}583$ (linha *Based on Mean*) do teste de Levene não se rejeita a igualdade de variâncias. Logo pode-se interpretar a ANOVA baseada no teste F.

Pela segunda tabela, como o valor $p = 0{,}004$ rejeitamos a igualdade das médias de idades por regiões (*Observação:* valor p obtido está muito próximo do resultado do teste não paramétrico). Neste caso no SPSS é disponibilizado um conjunto de testes de comparação múltipla para explorar estas diferenças, aplicando-se como exemplo o teste de Scheffe, uma vez que as amostras têm dimensões diferentes.

(SPSS) Analyze → Compare Means → One-Way ANOVA
(Dependent list: Idade; Factor: Região de residência [Regiao];
Post Hoc → Equal Variances Assumed: ☑ Scheffe)

Post Hoc Tests
Multiple Comparisons
Dependent Variable: Idade
Scheffe

(I) Região de residência	(J) Região de residência	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Norte	Centro	-4,896	2,038	,218	-11,18	1,39
	LVT	-3,287	1,849	,531	-8,99	2,42
	Alentejo	-7,747*	2,440	,040	-15,28	-,22
	Algarve	-7,205	2,487	,079	-14,88	,47
Centro	Norte	4,896	2,038	,218	-1,39	11,18
	LVT	1,608	2,061	,962	-4,75	7,97
	Alentejo	-2,851	2,604	,878	-10,89	5,19
	Algarve	-2,309	2,649	,944	-10,48	5,87
LVT	Norte	3,287	1,849	,531	-2,42	8,99
	Centro	-1,608	2,061	,962	-7,97	4,75
	Alentejo	-4,459	2,459	,511	-12,05	3,13
	Algarve	-3,917	2,506	,655	-11,65	3,82
Alentejo	Norte	7,747*	2,440	,040	,22	15,28
	Centro	2,851	2,604	,878	-5,19	10,89
	LVT	4,459	2,459	,511	-3,13	12,05
	Algarve	,542	2,969	1,000	-8,62	9,71
Algarve	Norte	7,205	2,487	,079	-,47	14,88
	Centro	2,309	2,649	,944	-5,87	10,48
	LVT	3,917	2,506	,655	-3,82	11,65
	Alentejo	-,542	2,969	1,000	-9,71	8,62

*. The mean difference is significant at the 0.05 level.

Homogeneous Subsets
Idade
Scheffe[a,b]

Região de residência	N	Subset for alpha = 0.05: 1	2
Norte	297	35,24	
LVT	281	38,52	38,52
Centro	198	40,13	40,13
Algarve	109	42,44	42,44
Alentejo	115		42,98
Sig.		,067	,497

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 167,526.
b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Com base no teste de Scheffe, ao nível de significância de 5%, apenas é detetada diferença entre as médias das idades da população do Norte e do Alentejo. Podem assim ser considerados dois grupos distintos de regiões de igualdade das médias:

- Grupo 1 (menores idades): Norte, Centro, LVT e Algarve;
- Grupo 2 (maiores idades): Alentejo, Centro, LVT e Algarve.

12.3.14 Comparação da altura média entre as regiões do país

Esta questão é do mesmo tipo que a anterior onde se descreveram os processos e instruções a realizar, considerando agora que a variável de estudo é Altura. Neste caso, com vista a exemplificar como proceder quando o pressuposto de homogeneidade das variâncias não se verifica, vai ser assumida a Normalidade dos subgrupos.

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
Altura	Based on Mean	4,014	4	747	,003
	Based on Median	3,949	4	747	,004
	Based on Median and with adjusted df	3,949	4	715,072	,004
	Based on trimmed mean	3,999	4	747	,003

Como a igualdade de variâncias é rejeitada $(\text{valor } p = 0{,}003)$, aplica-se um teste com esta correcção (em vez de utilizar a ANOVA baseada no teste F). Optou-se aqui por pedir os dois testes disponíveis no SPSS.

(SPSS) Analyze → Compare Means → One-Way ANOVA
(Dependent list: Idade; Factor: Região de residência [Regiao];
Options → ☑ Welch; ☑ Brown-Forythe)

Robust Tests of Equality of Means

Altura

	Statistic[a]	df1	df2	Sig.
Welch	1,804	4	284,884	,128
Brown-Forsythe	1,917	4	556,015	,106

a. Asymptotically F distributed.

Com estes $valor\text{-}p$, não se rejeita a igualdade das médias das alturas por regiões, para qualquer nível de significância < 0,106, i. e., aos níveis usuais de significância. Atenção que caso a hipótese de igualdade das médias fosse rejeitada, para explorar estas diferenças teriam que ser utilizados os testes que comparação múltipla que assumem variâncias diferentes (Post Hoc → Equal Variances Not Assumed).

12.3.15 O estado civil está relacionado com o sexo?

Esta questão pode ser respondida através de um teste de independência do Qui-quadrado, dada a natureza qualitativa nominal das variáveis.

(SPSS) Analyze → Descriptive Statistics → Crosstabs
(Row(s): Sexo; Column(s): Estado Civil [Civil];
Statistics → ☑ Chi-square;
Cells → Counts: ☑ Observed; ☑ Expected)

Sexo * Estado civil Crosstabulation

			Estado civil				
			Casado	Solteiro	Separado	Viúvo	Total
Sexo	Masculino	Count	324	237	4	15	580
		Expected Count	312,6	229,7	6,4	31,3	580,0
	Feminino	Count	215	159	7	39	420
		Expected Count	226,4	166,3	4,6	22,7	420,0
Total		Count	539	396	11	54	1000
		Expected Count	539,0	396,0	11,0	54,0	1000,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	23,903[a]	3	,000
Likelihood Ratio	23,827	3	,000
Linear-by-Linear Association	14,975	1	,000
N of Valid Cases	1000		

a. 1 cells (12,5%) have expected count less than 5. The minimum expected count is 4,62.

As condições de aplicabilidade do teste de independência do Qui-quadrado estão satisfeitas, pois apenas 12,5% das frequências esperadas são inferiores a 5 e são todas superiores a 1 (a menor frequência esperada é 4,62). Rejeita-se a independência entre a variável sexo e o estado civil ($\text{valor } p < 0{,}001$), ou seja, existe uma relação entre estas duas variáveis.

12.3.16 Existe relação entre se sofre de diabetes e o sexo?

Esta questão também pode ser respondida através de um teste de independência do Qui-quadrado. Veja na pergunta anterior as instruções a realizar, representando-se agora em coluna a variável Sofre de diabetes [sofre_diabetes].

Sexo * Sofre de diabetes Crosstabulation

			Sofre de diabetes		
			Sim	Não	Total
Sexo	Masculino	Count	31	549	580
		Expected Count	27,9	552,1	580,0
	Feminino	Count	17	402	419
		Expected Count	20,1	398,9	419,0
Total		Count	48	951	999
		Expected Count	48,0	951,0	999,0

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,882[a]	1	,348		
Continuity Correction[b]	,623	1	,430		
Likelihood Ratio	,897	1	,344		
Fisher's Exact Test				,372	,216
Linear-by-Linear Association	,881	1	,348		
N of Valid Cases	999				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 20,13.
b. Computed only for a 2x2 table

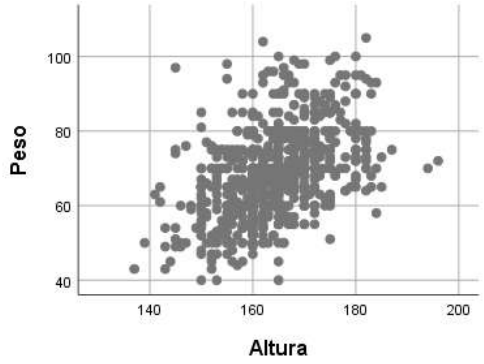
As condições de aplicabilidade do teste de independência do Qui-quadrado estão satisfeitas. Como é uma tabela 2×2 deve-se interpretar o $\text{valor } p$com correcção de continuidade (linha *Continuity Correction*, coluna *Asymptotic Significance*: $\text{valor } p = 0{,}430$), donde se conclui que não se rejeita a independência entre a variável sexo e a ocorrência da diabetes. Caso as condições de aplicabilidade não fossem verificadas, e *apenas*

em tabelas 2×2, deveria ser interpretado o teste exacto de Fisher (linha *Fisher's Exact Test*). Noutro tipo de tabelas de maior dimensão teriam que ser efetuados agrupamentos de classes.

12.3.17 Será que existe relação linear entre a altura e o peso? E a idade com a altura?

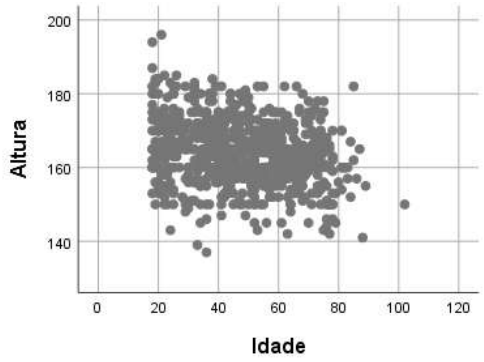
Visto tratarem-se de duas variáveis quantitativas, o estudo da relação linear entre as variáveis é efetuado através do coeficiente de correlação linear de Pearson. O passo inicial deve ser a representação gráfica dos dados e só posteriormente o cálculo do coeficiente de correlação.

(SPSS) Graphs → Legacy Dialogs → Scatter/Dot→ Simple Scatter
(Y Axis: Peso; X Axis: Altura)



Os pontos parecem posicionar-se, com algum afastamento, sobre uma linha reta com declive positivo. Portanto, neste caso faz sentido quantificar o grau de relação linear entre as duas variáveis.

(SPSS) Graphs → Legacy Dialogs → Scatter/Dot→ Simple Scatter
(Y Axis: Altura; X Axis: Idade)



Os pontos não se dispõem em torno de uma linha reta. Portanto, neste caso não faz sentido quantificar o grau de relação linear entre as duas variáveis uma vez que não se observa na nuvem de pontos um comportamento linear.

(SPSS) Analyze → Correlate → Bivariate
(Variables: Altura, Peso, Idade; Correlation Coefficients: ☑ Pearson; Test of Significance: ⊙ Two-tailed)

Correlations

		Altura	Peso	Idade
Altura	Pearson Correlation	1	,469**	-,248**
	Sig. (2-tailed)		,000	,000
	N	752	738	752
Peso	Pearson Correlation	,469**	1	,093*
	Sig. (2-tailed)	,000		,011
	N	738	750	750
Idade	Pearson Correlation	-,248**	,093*	1
	Sig. (2-tailed)	,000	,011	
	N	752	750	1000

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

A relação linear é positiva moderada entre o peso e a altura $(r = 0{,}469)$, havendo uma tendência para quanto mais altos forem os indivíduos mais pesados são, o que faz sentido.

Tal como se havia comprovado pela análise gráfica, a relação linear entre a idade e a altura é desprezável $(r = -0{,}093)$.

A significância das correlações não será analisada uma vez que os testes de hipótese para o coeficiente de correlação linear de Pearson se baseiam no pressuposto de que as variáveis são Normais, e este pressuposto é violado, conforme se pode observar na informação que se apresenta de seguida.

(SPSS) Analyze → Nonparametric Tests → Legacy Dialogs → 1-Sample K-S
(Test Variable List: Altura, Peso, Idade; Test Distribution: ☑ Normal)

One-Sample Kolmogorov-Smirnov Test

		Altura	Peso	Idade
N		752	750	1000
Normal Parameters[a,b]	Mean	164,39	68,94	38,81
	Std. Deviation	8,808	11,800	22,339
Most Extreme Differences	Absolute	,062	,076	,090
	Positive	,062	,076	,090
	Negative	-,054	-,031	-,058
Test Statistic		,062	,076	,090
Asymp. Sig. (2-tailed)		,000[c]	,000[c]	,000[c]

a. Test distribution is Normal.
b. Calculated from data.
c. Lilliefors Significance Correction.

A Normalidade é rejeitada para todas as variáveis analisadas (todos os valores $p \leq 0{,}006$), logo não se deve interpretar a significância da correlação.

12.3.18 Existe relação entre o que o IMC das mulheres e a sua autoapreciação do estado de saúde?

Para avaliar a relação entre uma variável qualitativa ordinal e uma variável quantitativa utiliza-se o coeficiente de correlação Spearman.

Primeiro tem que se efetuar uma seleção dos valores observados para o sexo feminino e só depois se pode calcular o coeficiente de Spearman entre as 2 variáveis ordinais.

(SPSS) Data → Select Cases
(Select: ⊙ If condition is satisfied; If → Sexo = 2; Output: ⊙ Filter out unselected cases)
Analyze → Correlate → Bivariate
(Variables: Autoapreciação do estado de saúde [autoapreciacao], IMC; Correlation Coefficients: ☑ Spearman; Test of Significance: ⊙ Two-tailed)

Correlations

			Autoapreciação do estado de saúde	IMC
Spearman's rho	Autoapreciação do estado de saúde	Correlation Coefficient	1,000	,182**
		Sig. (2-tailed)	.	,008
		N	218	208
	IMC	Correlation Coefficient	,182**	1,000
		Sig. (2-tailed)	,008	.
		N	208	313

**. Correlation is significant at the 0.01 level (2-tailed).

Para níveis de significância superiores ou iguais a 0,8%, considera-se que existe uma associação positiva muito fraca (sem interesse) entre as duas variáveis ($r_s = 0{,}182$).

Atenção: no final é necessário remover o filtro aplicado aos dados.
(SPSS) Data → Select Cases
(⊙ All cases)

12.3.19 De que forma poderá a altura explicar linearmente o peso?

Para descrever a relação linear entre duas variáveis, utiliza-se a técnica de regressão linear simples.

(SPSS) Analyze → Regression → Linear
(Dependent: Peso; Independent(s): Altura;
Statistics → Regression Coefficients: ☑ Estimates; ☑ Confidence Intervals; Level(%) 95; ☑ Model fit;
Plots → Scatter 1 of 1: Y: *ZRESID; X: *ZPRED; Standardized Residual Plots: ☑ Normal probability plot)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,469[a]	,220	,219	10,406

a. Predictors: (Constant), Altura
b. Dependent Variable: Peso

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22514,990	1	22514,990	207,924	,000[b]
	Residual	79697,607	736	108,285		
	Total	102212,598	737			

a. Dependent Variable: Peso
b. Predictors: (Constant), Altura

Coefficients[a]

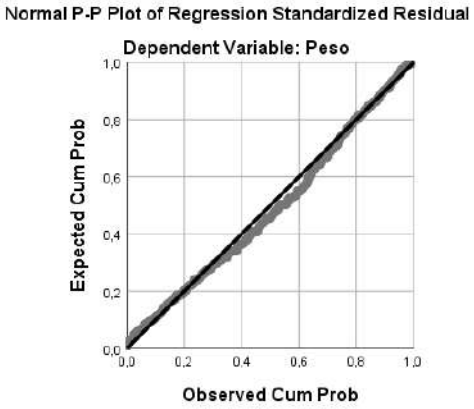
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-33,953	7,145		-4,752	,000	-47,980	-19,927
	Altura	,626	,043	,469	14,420	,000	,540	,711

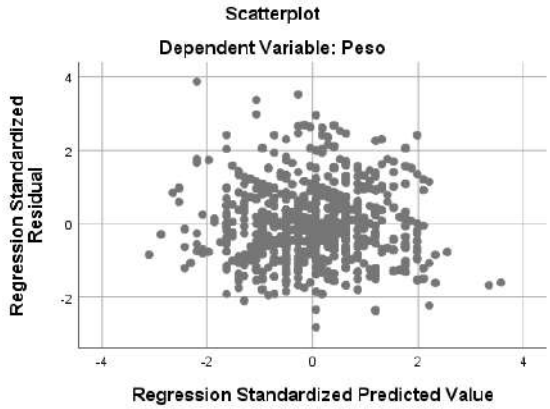
a. Dependent Variable: Peso

Residuals Statistics[a]

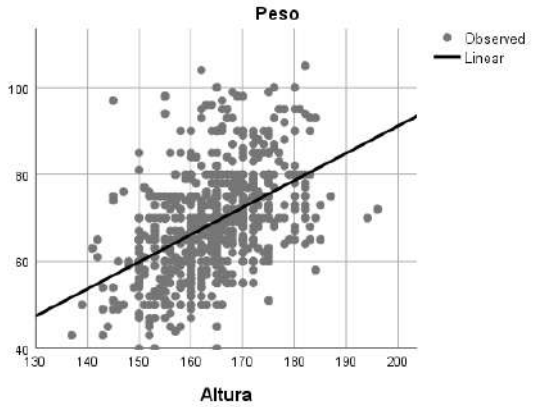
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	51,76	88,68	68,92	5,527	738
Residual	-29,283	40,230	,000	10,399	738
Std. Predicted Value	-3,104	3,574	,000	1,000	738
Std. Residual	-2,814	3,866	,000	,999	738

a. Dependent Variable: Peso





(SPSS) Analyze → Regression → Curve Estimation
(Dependent: Peso; Independent: ⊙ Variable: Altura; ☑ Include constant in equation; ☑ Plot models; Models: ☑ Linear)



Pela análise do gráfico de dispersão verifica-se os pontos parecem de certa forma dispor-se em torno de uma reta de declive positivo, verificando-se que à medida que aumenta a altura o peso também aumenta e vice-versa, indiciando que existe uma relação linear positiva fraca a moderada entre as variáveis.

Este modelo, estatisticamente significativo (valor $p < 0{,}001$, na tabela ANOVA), apresenta um poder explicativo fraco de apenas 22% (*R square* $= 0{,}220$). A linha sólida representa o modelo linear ajustado aos dados, cuja equação é $\widehat{Peso} = -33{,}953 + 0{,}626 \times Altura$. Estes coeficientes são significativos (ambos os valores $p < 0{,}001$). Os resíduos apresentam média 0 e apresentam um ligeiro desvio da distribuição Normal (análise do gráfico quantil-quantil).

12.4 Introdução de dados no SPSS

O SPSS é constituído por duas janelas principais:

- *SPSS Statistics Data Editor:* tem duas funcionalidades associadas aos separadores que a constituem:
  - Em *Data View* é possível visualizar e editar os dados em estudo;
  - Em *Variable View* é permitido definir ou alterar as características das variáveis.
- *SPSS Statistics Viewer:* onde se apresentam os resultados dos estudos realizados.

12.4.1 Introdução de um novo conjunto de dados

Para criar um novo conjunto de observações efetuar os seguintes passos:
menu *File* → submenu *New* → opção *Data*.
Na janela *SPSS Statistics Data Editor* é apresentada uma folha em branco disponível para editar dados.

12.4.1.1 Dados não agrupados

Para introduzir dados basta digitar valores, dentro das células apresentadas na janela *SPSS Statistics Data Editor*, e pressionar a tecla *Enter*.
Para introduzir uma *nova linha* (*coluna*) de observações (variáveis) entre linhas (colunas) já existentes:

1. Colocar o cursor na observação posterior à que se pretende criar;
2. Selecionar no menu *Data* a opção *Insert cases* (*Insert variable*).

Alternativamente pode selecionar a linha (coluna) posterior à que pretende criar e pressionar o botão direito do rato, para ativar o menu rápido, e escolher a opção *Insert cases* (*Insert Variable*). Após os passos anteriores é criada uma linha (coluna) sem observações na qual poderá introduzir os dados pretendidos.

12.4.1.2 Dados agrupados

Para introduzir dados agrupados em classes, efetuar os seguintes passos na janela *SPSS Statistics Data Editor:*

1. Introduzir numa das colunas os pontos médios de cada uma das classes;
2. Numa outra coluna digitar as frequências simples de cada uma das classes;
3. Ativar menu *Data* → opção *Weight Cases...* → marcar ⊙ *Weight cases by* → em *Frequency Variable* selecionar a variável associada às frequências.
4. Pressionar o botão *OK*.

No caso de os dados se apresentarem numa tabela de contingência, digitar o valor das frequências simples numa coluna. Nas duas colunas seguintes indicar a que categoria da variável 1 e 2 corresponde essa frequência. Seguidamente efetuar os passos 3, 4 e 5.

12.4.2 Definir as propriedades das variáveis

Para definir as características da variável em estudo na janela *SPSS Data Editor*, pressionar o botão esquerdo do rato duas vezes consecutivas em cima do nome da variável ou ativar o separador *Variable View*. Nesta janela pode definir em:

- *Name* – o nome (abreviado) pretendido para a variável;
- *Type* – o tipo da variável: numérico, data, texto, ...;
- *Width* – o tamanho (i. e., o número de caracteres) da variável;
- *Decimals* – o número de casas decimais da variável;

- *Label* – o nome completo ou descrição da variável;
- *Values* – associar descrições aos valores que a variável assume;
- *Missing* – valores que serão tratados os valores omissos;
- *Columns* – o número de caracteres a visualizar;
- *Align* – o alinhamento dos dados nas células: à esquerda, à direita ou centrado;
- *Measure* – o tipo de medida da variável: escala, ordinal ou nominal.

12.4.2.1 Legendar valores de uma variável

Quando se introduzem os dados no SPSS habitualmente as variáveis são todas codificadas para um valor numérico correspondendo cada um desses valores a uma categoria ou classe de valores, por exemplo:

1 = Aprovado	ou	450 = [300; 600[
2 = Reprovado		750 = [600; 900[

Existe, portanto, a necessidade de legendar os valores que a variável pode assumir. Para o fazer deve efetuar os seguintes passos na janela *SPSS Statistics Data Editor*:

1. Ativar o separador *Variable View*;
2. Pressionar o botão esquerdo direito do rato na célula *Values* da linha associada à variável para a qual pretende legendar os valores;
3. Na caixa de diálogo apresentada:
   a. Em *Value* escrever o valor ao qual quer associar uma legenda;
   b. Em *Value Label* escrever a legenda pretendida;
   c. Pressionar o botão *Add* para adicionar essa legenda;
4. Pressionar o botão *OK*.

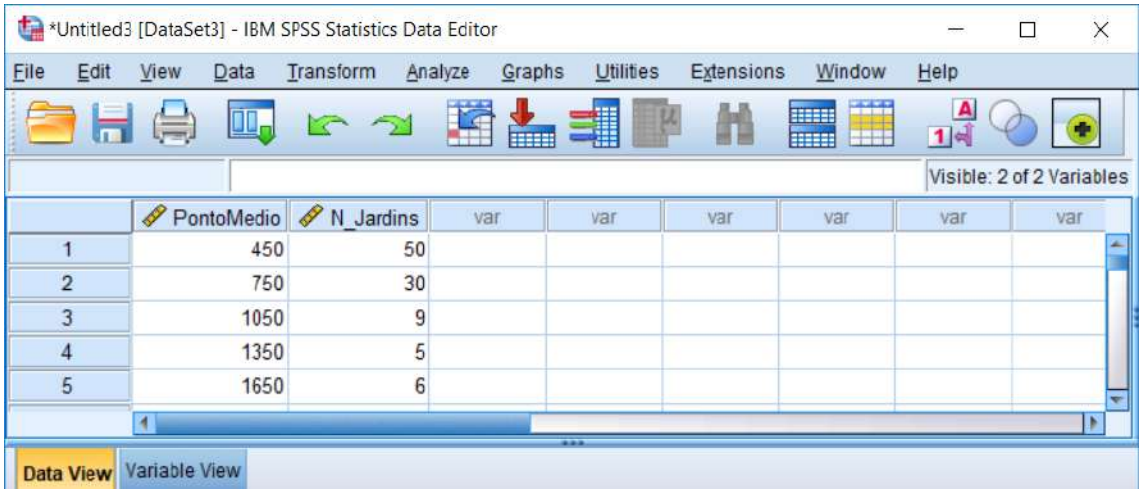
Caso pretenda:

- Alterar uma legenda: na caixa de diálogo selecione a legenda a alterar, proceda às alterações em *Value* e *Value Label* e pressione o botão *Change*.
- Eliminar uma legenda: selecione a legenda a eliminar e pressione o botão *Remove*.

12.4.3 Exemplos

12.4.3.1 Dados em tabela de frequências

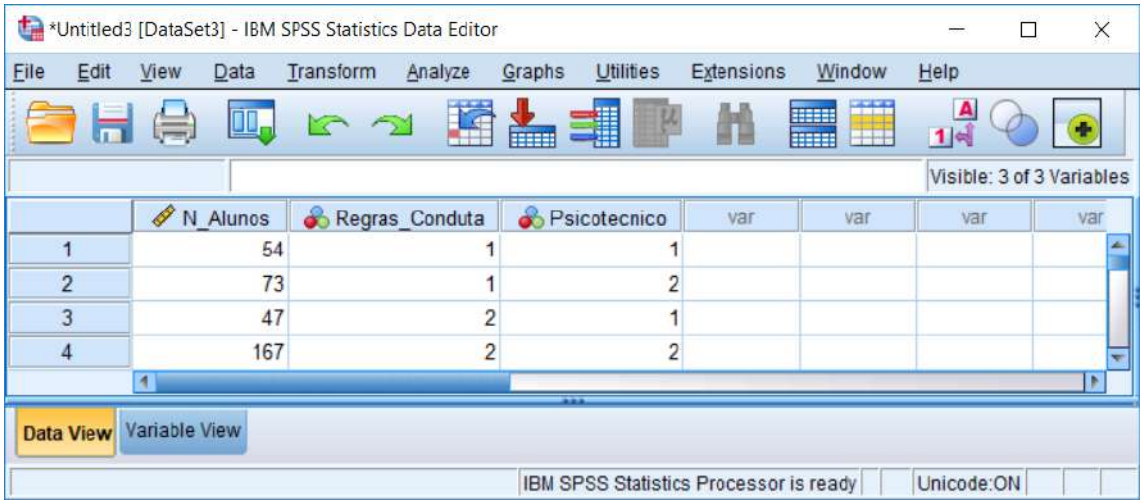
Considere o exercício da secção 2.1.5.4 sobre as áreas dos jardins aprovados:



1. Definição das propriedades:
   Janela *SPSS Statistics Data Editor* → separador *Variable View*:
   - 1ª Linha → *Name*: PontoMedio; *Type*: numeric; *Width*: 4; *Decimals*: 0; *Label*: Área (m2); *Values*: … *Value*: 450; *Label*: [300; 600[; *Add*; *Value*: 750; *Label*: [600; 900[; *Add*; *Value*: 1050; *Label*: [900; 1200[; *Add*; *Value*: 1350; *Label*: [1200; 1500[; *Add*; *Value*: 1650; *Label*: [1500; 1800[; *Add*; *OK*; *Measure*: Scale.
   - 2ª Linha → *Name*: N_Jardins; *Type*: numeric; *Width*: 2; *Decimals*: 0; *Label*: N.º de jardins; *Measure*: Scale.
2. Associar frequências aos valores da variável de estudo:
   menu *Data* → opção *Weight Cases...* → ⊙ *Weight cases by* → *Frequency Variable*: N_Jardins.

12.4.3.2 Dados em tabela de contingência

Considere o exercício da secção 10.7.4, sobre os resultados dos testes sobre regras de conduta e psicotécnico:



1. Definição das propriedades:
   Janela *SPSS Statistics Data Editor* → separador *Variable View*:
   - 1ª Linha → *Name*: N_Alunos; *Type*: numeric; *Width*: 4; *Decimals*: 0; *Measure*: Scale.
   - 2ª Linha → *Name*: Regras_Conduta; *Type*: numeric; *Width*: 1; *Decimals*: 0; *Label*: Regras de Conduta; *Values*: … *Value*: 1; *Label*: Aprovado; *Add*; *Value*: 2; *Label*: Reprovado; *Add*; *OK*; *Measure*: Nominal.
   - 3ª Linha → *Name*: Psicotecnico; *Type*: numeric; *Width*: 1; *Decimals*: 0; *Label*: Teste Psicoténico; *Values*: … *Value*: 1; *Label*: Aprovado; *Add*; *Value*: 2; *Label*: Reprovado; *Add*; *OK*; *Measure*: Nominal.
2. Associar frequências aos valores da variável de estudo:
   menu *Data* → opção *Weight Cases...* → ⊙ *Weight cases by* → *Frequency Variable*: N_Alunos.

Soluções

Capítulo 1

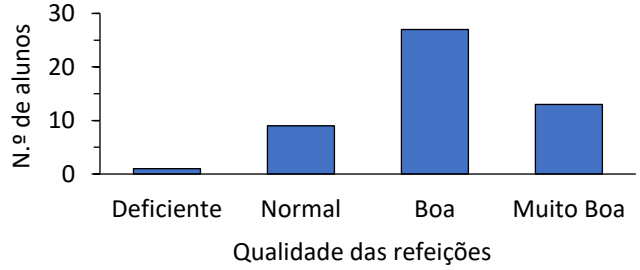
1. a) i) Todos os jovens portugueses inscritos no ensino superior no ano lectivo em que se realiza o estudo.
   ii) Por exemplo: os alunos de uma turma.
   iii) Indivíduos: cada um dos jovens portugueses inscrito no ensino superior.
   iv) Dados estatísticos: valor/resultado do inquérito (valor obtido da observação).
   b) i) Por ex.: Fuma? □ Sim; □ Não.
   ii) Por ex.: Qual a frequência com que fuma? □ Todos os dias; □ Às vezes; □ Raramente; □ Nunca.
   iii) Por ex.: Quantos cigarros fuma por dia? ____.
   c) i) Moda.
   ii) Moda e medidas de ordem (i.e., quantis).
   iii) Medidas de localização, dispersão, assimetria e achatamento.
2. Nome e Número são variáveis meramente informativas sendo todas as outras variáveis de interesse. Variáveis qualitativas nominais: Sexo, Curso, Almoço na cantina e Local de almoço; variáveis quantitativas discretas: Idade e Ano do curso; variável qualitativa ordinal: Qualidade da comida.
3. a) c) d) continuo; b) e) f) discreto.

Capítulo 2

1. a) Profissão do português radical: qualitativa nominal.
   b)

Profissão	n_i	f_i
Cargos técnico/científicos	640	0,64
Comércio e serviços	270	0,27
Empresários/Administradores/Gestores	70	0,07
Operários qualificados	20	0,02
Total	1000	1

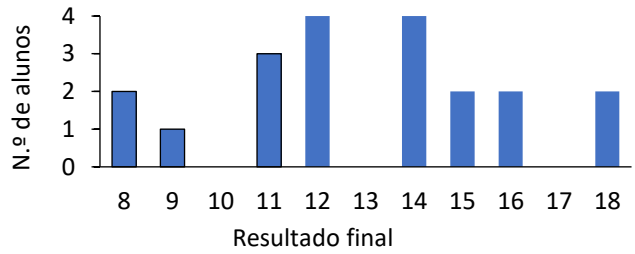
   c) 640.
   d) A maior parte dos portugueses radicais ocupa cargos técnico/científicos, cerca de ¼ ocupa cargos em comércio e serviços.
2. a) Opinião sobre a qualidade das refeições: qualitativa ordinal.
   b) Número total de alunos inquiridos. c) 54%.
   d)



   e) $\hat{x} =$ Boa e $\tilde{x} =$ Boa.
   f) O mais usual foi os alunos referirem que a comida era boa. Mais de 50% dos alunos disseram que a comida era no mínimo boa.
3. a) Quantitativos.
   b)

Resultados finais	n_i	N_i	f_i	F_i
8	2	2	0,10	0,10
9	1	3	0,05	0,15
11	3	6	0,15	0,30
12	4	10	0,20	0,50
14	4	14	0,20	0,70
15	2	16	0,10	0,80
16	2	18	0,10	0,90
18	2	20	0,10	1,00
Total	20		1	

c)

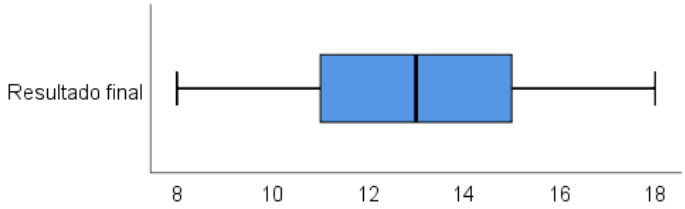


d) $\overline{x} = 13$; $\hat{x} = 12$ e 14; $\tilde{x} = 13$. e) $s^2 = 8{,}5263$ e $s = 2{,}92$.

f) 22,5%, que indicia que a média é representativa deste conjunto de dados.

g) $a = 10$ e $AIQ = 4$. h) $P_{48} = 12$ e $D_8 = P_{80} = 15{,}5$.

i)



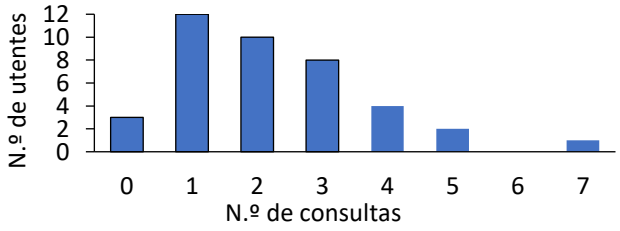
j) $g_B = 0$ (dist. simétrica), $k_P = 0{,}24$ (dist. quase mesocúrtica).

k) As notas dos alunos de estatística variam entre 8 e 18, com valores de tendência central (média, moda, mediana) similares em torno de 13 valores. Apresentam uma variação típica em relação à média de 3 valores, o que já representa alguma diversidade de notas. A distribuição é simétrica (comporta-se de forma igual em torno da média, sem tendências de concentração de valores elevados ou baixos) e sem grandes afastamentos de uma distribuição mesocúrtica. Não existem notas nem muito inferiores nem muito superiores (valores atípicos) às restantes.

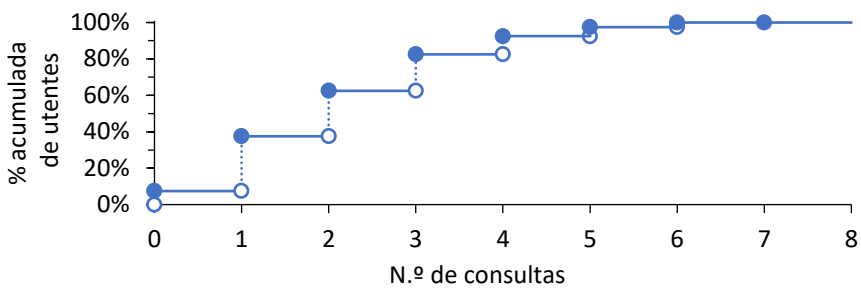
4. a)

N.º de consultas	n_i	N_i	f_i	F_i
0	3	3	0,075	0,075
1	12	15	0,300	0,375
2	10	25	0,250	0,625
3	8	33	0,200	0,825
4	4	37	0,100	0,925
5	2	39	0,050	0,975
7	1	40	0,025	1,000
Total	40		1	

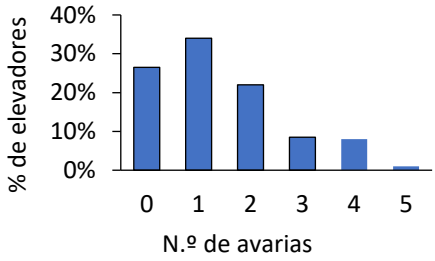
b)

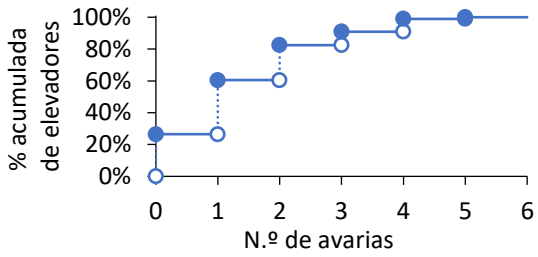


c)



5. a)



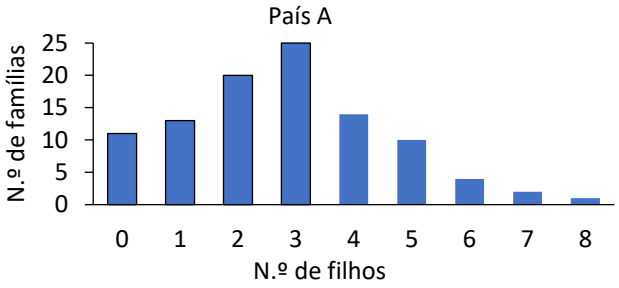


b) $\overline{x} = 1{,}405$; $s = 1{,}244$; $\hat{x} = 1$.

6. a) Número de filhos por família nos países A e B – variável quantitativa discreta. b) 2,8.

c) i) 5. ii) 0. d) 63,60%, a média é pouco representativa dos dados.

e)



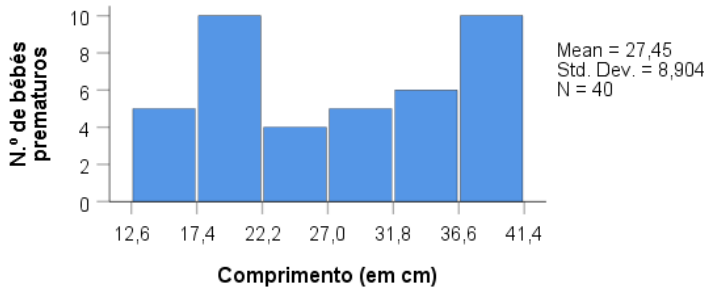
f) País A: $\overline{x} =$ 2,8; $\hat{x} = 3$; $\tilde{x} = 3$; País B: $\overline{x} = 2{,}56$; $\hat{x} = 1$; $\tilde{x} = 2$; as famílias do país A têm mais filhos do que as do país B, visto que todas as medidas de tendência central do país A são superiores às do país B.

g) $g_B = 0$ distribuição é simétrica.

7. a) $\overline{x} = 27{,}45$; $\tilde{x} = Q_2 = 28{,}5$; s = 8,90; $Q_1 = 19{,}25$; $Q_3 = 35{,}75$; $P_{66} = 33{,}3$; $P_{30} = D_3 = 19{,}55$.

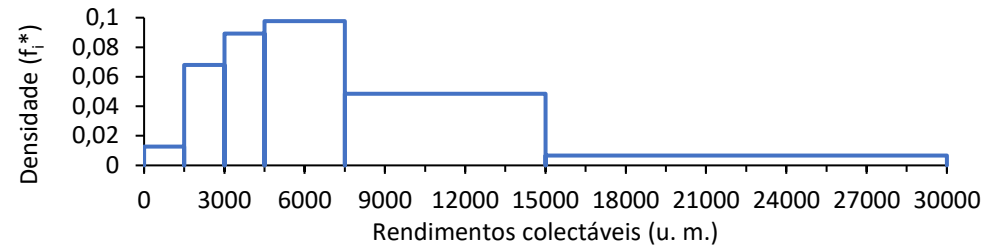
b) Regra de Sturges: $K = 6$; $a = 28{,}8$; $ac = 4{,}8$.

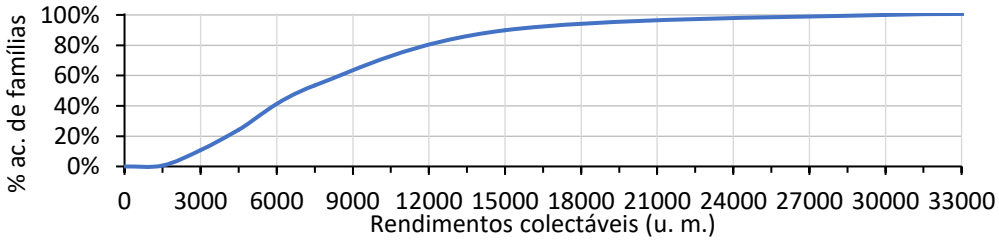
Comprimento em cm	Ponto médio	N.º bebés	dePropor. de bebés	N.º acum. bebés	dePropor. acum. de bebés
[12,6; 17,4[	15	5	0,125	5	0,125
[17,4; 22,2[	19,8	10	0,250	15	0,375
[22,2; 27[	24,6	4	0,100	19	0,475
[27; 31,8[	29,4	5	0,125	24	0,600
[31,8; 36,6[	34,2	6	0,150	30	0,750
[36,6; 41,4[	39	10	0,250	40	1,000
Total		40	1,000		



c) $\overline{x} = 27{,}84$; $\tilde{x} = 27{,}96$; $s^2 = 77{,}26$.

8. a)





b) $\overline{x} = 8862$; $\tilde{x} = 7141{,}64$; classe modal: [4500; 7500[, $\hat{x} = 4935{,}04$; $Q_3 = 11929{,}95$.

c) $s = 5625{,}72$; $CV = 63{,}5\%$. d) Falso, pois a mediana é inferior à média. e) i) 15148,5. ii) 44%.

9. a) $\overline{x} = 20 > \tilde{x} = 19{,}78 > \hat{x} = 19{,}32$, ligeira assimetria positiva.

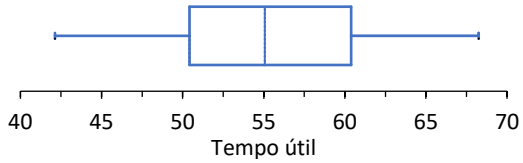
b) $Q_2 = \tilde{x} = 19{,}78$. Metade dos utentes demoraram no máximo 19,78 minutos a ser atendidos.

c) $s = 5{,}03$; $s^2 = 25{,}25$. d) $g_B = 0{,}013$; $g_p = 0{,}135$; a distribuição é quase simétrica.

10. a)

Tempo útil de jogo	x_i'	n_i	N_i	f_i	F_i
[41; 45[	43	2	2	0,0286	0,0286
[45; 49[	47	7	9	0,1000	0,1286
[49; 53[	51	18	27	0,2571	0,3857
[53; 57[	55	12	39	0,1714	0,5571
[57; 61[	59	15	54	0,2143	0,7714
[61; 65[	63	12	66	0,1714	0,9429
[65; 69]	67	4	70	0,0571	1,0000

b)



c) 50,4225. d) $\overline{x} = 55{,}551$ e classe modal: [49; 53[; e) $CD = 0{,}1084 < 0{,}5$. É representativa.

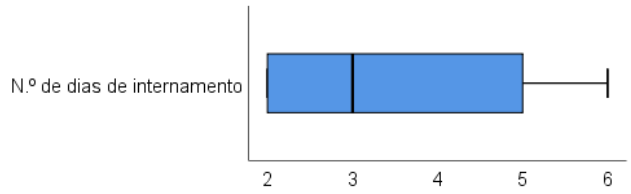
f) Não. $g_{SPSS} = 0{,}28$, não rejeitar a simetria do tempo útil de jogo.

g) $k_{SPSS} = -1{,}29$, não rejeitar a hipótese de a distribuição ser mesocúrtica.

11. a)

N.º de dias de internamento	N.º de parturientes	% de parturientes	N.º acum. de parturientes	% acum. de parturientes
2	18	45,0	18	45
3	4	10,0	22	55
4	5	12,5	27	67,5
5	7	17,5	34	85
6	6	15,0	40	100
Total	40	100,0		

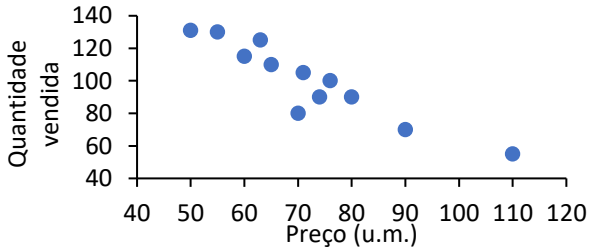
b)

c) 

$\hat{x} = 2$; $\overline{x} = 3{,}5$; $\tilde{x} = 3$; $s^2 = 2{,}46$.

d) i) 5 dias. ii) 2 a 3 dias ou de 3 a 6 dias (existem mais resultados possíveis, estes são apenas 2 exemplos).

12. a) $\overline{x} = 89{,}1$; classe modal: [85; 90[, $\hat{x} = 87{,}79$; $\tilde{x} = 88{,}31$. b) $s = 6{,}685$. c) 7%.
13. a) Humidade e temperatura – variáveis quantitativas contínuas. b) 15,895. c) 53,224.
    d) i) 21,997. ii) 24,613. iii) 33,137. iv) 19,44.
    e) $CD_{Temp} = 0{,}245$; $CD_{Hum} = 0{,}244$. As médias são representativas dos dados.
    f) Não, pois $g_{SPSS} = 2{,}25$, i. e., a distribuição é assimétrica positiva.
    g) $r_s =$ -0,84, correlação linear negativa forte, i. e., quanto maior a temperatura menor a humidade e vice-versa.
14. a)



    b) Sim.
    c) -0,926. A quantidade vendida apresenta uma correlação linear muito elevada, mas inversamente proporcional, com o preço do bem alimentar de 1ª necessidade, i. e., quanto mais caro se torna o bem menor é a quantidade que é vendida.
15. 0,443. A relação entre os 2 tempos de espera é moderada e positiva, i. e., existe uma tendência para quanto maior for o tempo de espera entre o encaminhamento e a consulta de referência, maior será também o tempo de espera até à cirurgia, o que poderá indiciar alguma consistência no processo tendo em conta, por exemplo, graus de urgência diferentes (terá que ser analisado este factor).
16. -0,982. Existe uma associação muito forte de tipo negativo entre o tempo de resolução do *puzzle* e a nota em matemática, i. e., as notas baixas de matemática estão associadas aos alunos que demoram mais tempo a resolver o *puzzle*, e vice-versa.

Capítulo 3

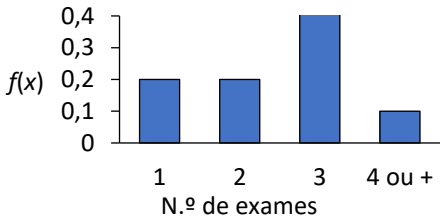
1. a) $\Omega = \{E, M, Q\}$ onde $E =$ estagnação, $M =$ melhoria e $Q =$ quebra. b) $P(E) = 2/5$; $P(M) = 2{,}5$; $P(Q) = 1/5$.
   c) Subjectivo.
2. a) 0,625. b) 0,5. c) 0,75.
3. 5/8.
4. a) $p = 0{,}4$. b) $p = 0{,}5$.
5. Amostragem: com reposição - $P(A) = 0{,}09$; sem reposição - $P(A) = 0{,}0909$ = 9/99.
6. a) 0,38. b) 0,33. c) 0,61. d) 0,0758. e) 0,0294. f) 0,8684. g) 0,9839.
   h) O resultado não é independente da doença.
7. a) 0,875. b) 0,375.
8. a) 0,48. b) 0,75.
9. a) 0,57. b) 0,4474.
10. a) 0,082. b) 0,0122.

Capítulo 4

1. -59.
2. a) 0,66. b) 0,05. c) 1,12. d) 1,20.
3. a) 0,6. b) 0,6. c) 0,7. d) 0,2222. e) $E(X) = 1{,}7$. f) $E(2X + 4) = 7{,}4$.

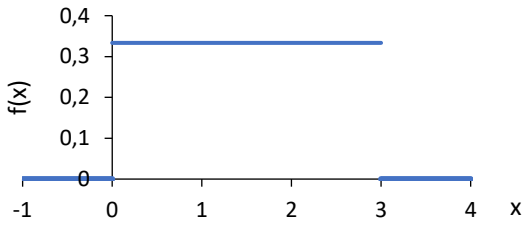
g) $Var(X) = 0{,}61$. $Var(2X + 4) = 2{,}44$.

4. a) $f(1) = 0{,}2$; $f(2) = 0{,}2$; $f(3) = 0{,}5$; $f(4) = 0{,}1$.



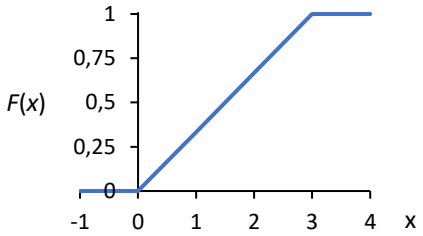
b) 0,1. c) 0,4. d) 0,2. e) 0,4444. f) 3. g) 2,5. h) $Var(X) = 0{,}85$; $E(2X) = 5$; $Var(3X) = 7{,}65$.

5. a) $f(x) = 1/6$, $x = 1, 2, 3, 4, 5, 6$. b) $E(X) = 3{,}5$ e $E(X^2) = 15{,}1667$;
   c) $E((X + 3)^2) = 45{,}1667$; $Var(3X - 2) = 26{,}25$. d) $E(Y) = 4{,}75$; $Var(Y) = 0{,}7292$.
6. a) $f(10) = f(11) = 0{,}3$, $f(12) = f(13) = 0{,}3$; $f(14) = 0{,}1$. b) 35800. c) 0,6.
7. a)



b)

$$F(x) = \begin{cases} 0, & x \leq 0 \\ \frac{1}{3}x, & 0 < x < 3 \\ 1, & x \geq 3 \end{cases}$$



c) 2/3. d) $E(X) = 3/2$; $Var(X) = 3/4$.

8. a) 1/9. b) 7/27. c) $E(X) = 2{,}25$; $Var(3X) = 3{,}0375$.
9. a)

x	1	2	3
$f_X(x) = P(X = x)$	1/4	1/2	1/4
$F_X(x) = P(X \leq x)$	1/4	3/4	1

y	2	3	4
$f_Y(y) = P(Y = y)$	1/3	1/3	1/3
$F_Y(y) = P(Y \leq y)$	1/3	2/3	1

b) $P(XY \text{ ser par}) = 2/3$. c) $F_{XY}(2;\ 3) = 5/12$.

d)

y	2	4
$f_{Y\|X=2}(y) = P(Y = y\|X = 2)$	1/3	2/3

e) $P(X = 3|Y \geq 3) = 1/6$. f) $P(X = 3) = 1/4$. g) $P(X = 2|X + Y \leq 5) = 1/3$.

h) Não são independentes. Por ex., $f_{XY}(2; 3) \neq f_X(2) f_Y(3)$.

10. a) 2/3. b) Sim. c) 76,85. d) 0,0058. e) 0,0476.
11. a) 0,05. b) 1050. c) 1500. d) 0,10. e) 0,30.

f) $f_{X|Y=1250}(1000) = 0{,}5$; $f_{X|Y=1250}(1050) = 0{,}3333$; $f_{X|Y=1250}(1100) = 0{,}1667$. g) $E(W) = 14300$.

h) Sim. Por ex., $f_X(1000) f_Y(750) \neq f_{XY}(1000; 750)$. i) $\sigma_{XY} = -2500$; $\rho_{XY} = -0{,}3333$.

12. a) 0,1. b) Sim. Por ex., $f_X(0) f_Y(0) \neq f_{XY}(0; 0)$. c) $E(Y) = 1{,}1$; $Var(X) = 0{,}49$.
13. a) $f_{XY}(1;\ 0) = 0{,}04$; $f_{XY}(1;\ 1) = 0{,}08$; $f_{XY}(1;\ 2) = 0{,}08$;

$f_{XY}(2;\ 0) = 0{,}04$; $f_{XY}(2;\ 1) = 0{,}08$; $f_{XY}(2;\ 2) = 0{,}08$;
$f_{XY}(3;\ 0) = 0{,}12$; $f_{XY}(3;\ 1) = 0{,}24$; $f_{XY}(3;\ 2) = 0{,}24$.

b) i)

z	3	4	5	6	7	8	9	10	11
$f(z)$	0,04	0,08	0,08	0,04	0,08	0,08	0,12	0,24	0,24

ii) $E(Z) = 8{,}4$; $Var(Z) = 6{,}32$.

14. a) 13/12. b) 1,84. c) Não são independentes.
15. a) 1/96. c) 15/384. d) $E(X) = 8/3$; $E(Y) = 31/9$. e) Sim.

Capítulo 5

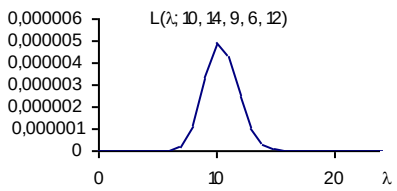
1. a) $f(x) = P(X = x) = {}^{n}C_{x} \times 0{,}8^{x} \times 0{,}2^{n-x},\ x = 0, 1, \ldots, n.$ b) i) 0. ii) 1. iii) 0,7718.
   c) $E(X) = 8$; $Var(X) = 1{,}6$.
2. a) 0,8244. b) 6. c) i) 0,1065. ii) 0,4287.
3. a) 0,9537. b) $n = 60$. c) $n = 6$. d) 2,5. e) 0,1444. f) 25. g) 0,0006.
4. a) 0,0005. b) 0,1461.
5. Sim, porque $P(X_N = 2;\ X_R = 4; X_O = 6) \approx 0$.
6. 0,18 (aprox. a Binomial).
7. a) 0,3394. b) 0,3151.
8. a) 0,1808. b) $E(X) = 15$; $Var(X) = 1{,}16$.
9. a) É o parâmetro da distribuição. Neste caso significa que em média ocorrem 7 actos de agressão por hora na tribo de chimpazés-pigmeus. b) 168. c) 0,0296.
10. a) 0,9291. b) 0,4455.
11. a) 0,2231. b) 0,4422.
12. a) $f(x) = \begin{cases} 4, & \text{se } 0{,}8 < x < 1{,}05 \\ 0, & \text{outros valores} \end{cases}$. b) $P(X < 1) = 0{,}8$. c) $E(X) = 0{,}925$.
13. a) $F(x) = \begin{cases} 0, & x \leq 21 \\ (x - 21)/4, & 21 < x < 25 \\ 1, & x \geq 25 \end{cases}$ b) $E(X) = 22$; $Var(X) = 4/3$. c) 0,5.
14. a) $\mu = 160$, logo $P(X \leq \mu) = P(X \leq 160) = 0{,}5 \Rightarrow P(X < 150) < 0{,}5$.
15. 0,9987.
16. a) 0,2514. b) 37. c) 0,3707. d) 0,2486. e) 118,25. f) 120,12.
17. a) Opção ii). b) 18,29.
18. a) i) 0,0228. ii) 0,0003. iii) 0,7865. b) 173,59. c) 186,41.
19. a) i) 0,0401. ii) 0,281. iii) 0,4649. b) i) 79,98. ii) 41,4.
20. a) 0,3085. b) $k = 20$.
21. a) 0,6827. b) 0,8758.
22. a) 580000. b) 0,8185. c) i) 0,5. ii) 0,1712. iii) 0,0705. iv) 9,7.
23. 0,0294.
24. $\mu_Y = 40$; $\sigma_Y = \sqrt{241}$.
25. a) 0,3085. b) 1,22 kg. c) $f_G(3) = 0{,}1587;\ f_G(4) = 0{,}5328;\ f_G(5) = 0{,}3085$. d) 4149,8 Euros. e) 0,2875. f) 0,0127. g) Aumenta.
26. 0,4865 (aprox. à distribuição Poisson).
27. a) 5. b) 0,1247. c) 0,0004.
28. a) $F(t) = 1 - e^{-\frac{1}{10}t}, t > 0$. b) 0,7769.
29. a) $f(t) = \frac{1}{15} e^{-\frac{1}{15}t}, t > 0$. b) 0,3679.
30. a) 0,1484. b) 0,4512.
31. a) 0,8. b) $a = 11{,}69$; $b = 38{,}08$. c) $E(X) = 23$; $Var(X) = 46$. d) $\chi^2_{23;\,0{,}05} = 13{,}09$; $\chi^2_{23;\,0{,}95} = 35{,}17$.
33. a) 0,025. b) –1,337.
34. a) 0,99. b) 0,4065.

Capítulo 6

1. 0,2512.
2. 0,0320.
3. a) 0,0793. b) 0,1271.
4. 0,0143.
5. 0,0427.
6. a) i) 0,0174. ii) 0,0174. b) Diminui em i) e ii). c) $n \geq 153$. d) Aumenta. e) i) 0,0321. ii) 0,0321. f) i) 0,3797. ii) 107,1414.
7. a) i) 0,9608. ii) 0,9606. iii) 29,27. b) Aumentam. c) É superior. d) i) 0,9555. ii) 0,9547. e) $n \geq 139$. f) Aumenta.
8. a) i) 0,0217. ii) 1. iii) 0,0001. b) $n = 71$.
9. a) 0,1271. b) 0,0055. c) Quanto maior a dimensão da amostra, menor é a diferença entre o valor amostral e o valor populacional.
10. a) i) 0,9292. ii) 0,9401. b) Com base nos desvios padrão populacionais existe menor incerteza. c) 0,9957.
11. a) i) 0,0073. ii) 0,0095. b) Quando se conhecem as variâncias populacionais existe menos incerteza e as diferenças são reflectidas pelo uso da distribuição Normal.
12. 0,695.
13. 0,3015.

Capítulo 7

1. a) $p_{\text{Abstenção}} = 0{,}64$. b) I.C. $p_{\text{Brancos/Nulos}} =]0{,}015;\ 0{,}023[$.
2. Opção b).
3. Opção b).
4. Um estimador dum parâmetro da população é uma variável aleatória (v. a.) que depende da informação amostral e cujas realizações fornecem aproximações para o parâmetro desconhecido. A um valor específico assumido por este estimador para uma amostra em concreto chama-se estimativa.
5. a) $T \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$. b) Sim, pois $E(T) = \mu$.
6. a) $E(\hat{\mu}_1) = E(\hat{\mu}_2) = E(\hat{\mu}_3) = \mu; Var(\hat{\mu}_2) = \sigma^2/2 > Var(\hat{\mu}_1) = 7\sigma^2/25; Var(\hat{\mu}_3) = \sigma^2/4$.
   b) $\hat{\mu}_1 = 8{,}6; \hat{\mu}_2 = 8{,}5; \hat{\mu}_3 = 9$.
7. a) $E(S^{*2}) = \frac{n-1}{n}\sigma^2; E(S^2) = \sigma^2$. b) $s^{*2} = 13{,}8765; s^2 = 15{,}6111$.
8. $\hat{p} = \sum_{i=1}^{n} \frac{X_i}{n}$.
9. a) $\hat{\lambda} = \bar{X}$. b) $\hat{\lambda} = \bar{X}$.
   c) i)



   ii) 10,2. iii) 0,0000042386.
10. a)]4,8052; 5,5948[;]4,7296; 5,6704[;]4,5818; 5,8182[.
    b)]4,7894; 5,6104[;]4,7046; 5,6954[;]4,5287; 5,8713[.
    c) Com base nos desvios padrão populacionais a incerteza é menor.
11. a) 4. b)]3,6322. 4,3678[. c) 68. d) 44. e) 40.
12. $n \geq 223$.
13. a) 8. b)]3,906. 24,638[.
14. a)]3,9235; 5,2965[. b)] 0,3514; 4,7775[.

15. a) 0,09. b) Não, pois $IC_{95\%}\ p =]0{,}0723;\ 0{,}1077[$.
16.]47,2; 57,8[. 0,62 $\notin$ I.C. Ambos os limites < 62%, logo com 90% de confiança, existe um declínio.
17. a) $IC_{90\%}\ p =]0{,}5207;\ 0{,}5793[$; $IC_{95\%}\ p =]0{,}5151;\ 0{,}5849[$; $IC_{98\%}\ p =]0{,}5086.\ 0{,}5914[$.
    b) A amplitude do I. C. aumenta com o grau de confiança.
    c)]0,5195. 0,5805[. A amplitude diminui com o aumento de n. d) $n \geq 1072$.
18.]-0,0358; 0,2025[. 0 $\in$ I. C. logo com 95 % de confiança não se pode concluir que sejam diferentes.
19. a)]-0,9995; 0,3614[. b)]2,8317; 4,5396[. c)]0,1057; 8,8846[. Não, pois 1 $\in$ ao I. C.
20. a)]-3,9371; 0,0799[. b)]-3,7599; 1,3885[. Não, pois 0 $\in$ ao I. C.
21. a)]0,0499; 4,0887[. b)]0,1869; 0,4531[. Sim, pois 0 $\notin$ I. C. c) 105.
22. a)]2,9556; 11,9459[. Não, pois 0 $\notin$ I. C.
    b)]0,3767. 5,0061[. 1 $\in$ I. C., logo não podemos concluir que as variâncias sejam diferentes.
    c)]3,1394. 11,7622[.
    d) Existe menor incerteza nos cálculos efectuados na alínea c), pois assume-se que se conhece mais informação sobre a população (desvios padrão populacionais), logo a amplitude do intervalo é menor.
23. a) 11,25. b) 2,1. c) 0,3. d)]10,6868; 11,8132[. e) 40. f)]0,1387; 0,4613[. g) 81.
    h)]-1,4031; 0,4031[. i)]-0,1361; 0,2161[. j)]0,3328; 1,0401[. Normalidade. k) Aumentando α.
24. a)]0,4045; 0,4855[. b)]-0,0146; 0,0606[. Sim, porque 0 $\in$ I. C.
    c) A 99% mantém-se a opinião. A 90% só se pode responder efectuando os cálculos, pois o I. C. 90% tem menor amplitude que o I. C. a 95%.
25.]0,1971; 0,6170[.
26.]0,5932; 0,8369[. Com 99% de confiança, existe correlação linear positiva.
27. a)]0,9187; 0,9748[. b) Sim, pois 0 $\notin$ I. C.

Capítulo 8

1. a) $H_1: \mu \neq 3{,}5;\ |z_{obs}| = 3{,}0246;\ z_{0{,}975} = 1{,}96$; rejeitar H_0 para $\alpha = 5\%$, i. e., contra.
   b) $H_1: \mu \neq 3{,}5;\ |t_{obs}| = 0{,}7740;\ t_{50;\,0{,}975} = 2{,}009$; não rejeitar H_0 para $\alpha = 5\%$, i. e., a favor.
2. a)]4,762; 5,971[. b) $H_1: \mu > 4{,}8;\ z_{obs} = 1{,}542;\ z_{0{,}95} = 1{,}645$; não rejeitar H_0. c) 0,6027.
3. a) $H_1: \mu \neq 150;\ |t_{obs}| = 2{,}4944;\ t_{13;\,0{,}995} = 3{,}012$; não rejeitar H_0 para $\alpha = 1\%$, i. e. afirmação verdadeira; $t_{13;\,0{,}95} = 1{,}771$; rejeitar H_0 para $\alpha = 10\%$, i. e. afirmação falsa.
   b) $\alpha = 1\%$: erro tipo II; $\alpha = 10\%$: erro tipo I. c) Construindo I. C. a 90% e 99% para μ. d) Opção ii. e) 0,0269.
   f) i) $H_1: \mu < 150;\ t_{obs} =$ -2,4944;
      $-t_{13;\,0{,}9} =$ -1,35; rejeitar H_0 para $\alpha = 10\%$, i. e. razão à defesa do consumidor;
      $-t_{13;\,0{,}95} =$ -1,771; rejeitar H_0 para $\alpha = 5\%$, i. e., razão aos gestores;
      $-t_{13;\,0{,}99} =$ -2,65; não rejeitar H_0 para $\alpha = 1\%$, i. e., razão aos gestores.
      ii) $\alpha = 1\%$: erro tipo II; $\alpha = 5\%$, 10%: erro tipo I. iii) opção i3.
4. Verdadeira, pois quanto mais próximo estiver o valor verdadeiro do valor testado na hipótese nula, menor é a capacidade do teste em detectar essa diferença.
5. a) i) $H_1: \mu \neq 1{,}3;\ |t_{obs}| = 1{,}479;\ t_{34;\,0{,}96} = 1{,}805$; não rejeitar H_0, i. e., não.
      ii) $H_1: \mu < 1{,}3;\ t_{obs} =$ -1,479; $-t_{34;\,0{,}92} =$ -1,436; rejeitar H_0, i. e., sim. b) i) 0,1483. ii) 0,0742.
6. a) $H_1: \mu_1 - \mu_2 \neq 0;\ |z_{obs}| = 0{,}943;\ z_{0{,}995} = 2{,}576$; não rejeitar H_0. b)]1698,02; 1821,98[.
7. a) Sim. b) $H_1: \mu \neq 15;\ |t_{obs}| = 2{,}451;\ t_{20;\,0{,}975} = 2{,}086$; rejeitar H_0.
   c) $H_1: \mu > 14;\ t_{obs} = 0{,}915;\ t_{15;\,0{,}99} = 2{,}602$; não rejeitar H_0.
   d) $H_1: \sigma_1^2 \neq \sigma_2^2;\ f_{obs} = 0{,}562;\ f_{20;\,15;\,0{,}025} = 0{,}39;\ f_{20;\,15;\,0{,}975} = 2{,}76$; não rejeitar H_0.
   e) $H_1: \mu_1 - \mu_2 \neq 0;\ |t_{obs}| = 1{,}670;\ t_{35;\,0{,}95} = 2{,}03$; não rejeitar H_0.
   f) i) $H_1: \mu_1 - \mu_2 < 0;\ t_{obs} = 1{,}670;\ t_{35;\,0{,}90} = 1{,}306$; rejeitar H_0, i. e., sim. ii) Não. iii) 0,052.
8. a) $H_1: \mu \neq 8;\ |t_{obs}| = 1{,}785;\ t_{50;\,0{,}95} = 1{,}676$; não rejeitar H_0. b) 0,080.
   c) i) $H_1: \sigma_1^2 \neq \sigma_2^2;\ f_{obs} = 1{,}778;\ f_{50;\,40;\,0{,}005} = 0{,}46;\ f_{50;\,40;0{,}995} = 2{,}23$; não rejeitar H_0.
      ii) $H_1: \mu_1 - \mu_2 > 0;\ t_{obs} = 1{,}3279;\ t_{90;\,0{,}9} = 1{,}291$; rejeitar H_0, i. e., não.
      iii) $H_1: \mu_1 - \mu_2 > 0;\ z_{obs} = 1{,}3284;\ z_{0{,}9} = 1{,}282$; rejeitar H_0, i. e., o medicamento está a ser eficaz.
9. a) $H_1: \mu >$ 10; $t_{obs} = 0{,}236;\ t_{7;\,0{,}95} = 1{,}895$; não rejeitar H_0.b) 0,9503.

c) $H_1: \sigma_1^2 \neq \sigma_2^2$; $f_{obs} =$ 0,801; $f_{7;\,8;\,0,025} = 0{,}20$; $f_{7;\,8;\,0,975} = 4{,}53$; não rejeitar H_0.

d) $H_1: \mu_1 - \mu_2 \neq 0$; $|t_{obs}| = 0{,}969$; $t_{15;\,0,995} = 2{,}947$; não rejeitar H_0.

10. a)]154,9252; 193,0748[. b) $H_1: \sigma_1^2 \neq \sigma_2^2$; $f_{obs} = 1{,}090$; $f_{4;\,4;\,0,005} = 0{,}04$; $f_{4;\,4;\,0,995} = 23{,}15$; não rejeitar H_0.

c)]-17,937; 25,937[. d) $H_1: \mu_1 - \mu_2 < 0$; $t_{obs} = 0{,}420$; $valor\ p = 0{,}6572$; não rejeitar H_0.

11. a) $H_0: p \geq 0{,}6\ \ vs.\ \ H_1: p < 0{,}6$.

b) $z_{obs} =$ -0,843; $z_{0,95} =$ -1,645; não rejeitar H_0. Erro tipo II. É dizer que são mais de 60% e estar enganado (realmente serem menos de 60 %).

12. a) $H_0: p \leq 0{,}5\ \ vs.\ \ H_1: p > 0{,}5$. b) $z_{obs} = 1{,}131$; $valor\ p = 0{,}129$; não rejeitar H_0.

13. $H_1: p < 0{,}5$; $z_{obs} =$ -16,975; $-z_{0,95} =$ -1,645; rejeitar H_0, i. e., fez baixar.

14. a) 1056. b) 0,59. c) $H_1: p \neq 0{,}64$; $|z_{obs}| = 3{,}385$; $z_{0,995} = 2{,}576$; rejeitar H_0. d) Diminuindo n.

e) i) $H_1: p_1 - p_2 > 0$; $z_{obs} = 2{,}229$; $z_{0,95} = 1{,}645$; rejeitar H_0, i. e., não. ii) Sim. iii) 0,013.

15. a) $H_1: p_2 \neq 0{,}5$; $|z_{obs}| = 1{,}233$; $z_{0,995} = 2{,}576$; não rejeitar H_0, i. e., sim.

b) $H_1: p_1 - p_2 < 0$; $z_{obs} =$ -1,278; $-z_{0,95} =$ -1,645; não rejeitar H_0, i. e., não.

16. a) $H_1: p < 0{,}8$; $z_{obs} =$ -2,5; $-z_{0,99} =$ -2,326; rejeitar H_0. b) 0,006.

c) $H_1: p_1 - p_2 \neq 0$; $|z_{obs}| = 3{,}275$; $z_{0,975} = 1{,}96$; rejeitar H_0.

17. a) $H_1: \mu_D < 5$; $t_{obs} = 0{,}444$; $-t_{23;\,0,95} =$ -1,714; não rejeitar H_0, i. e., falso.

b) $H_1: \rho \neq 0$; $|t_{obs}| = 19{,}402$; $t_{22;\,0,95} = 1{,}717$; rejeitar H_0, i. e., existe.

c) $H_1: \rho \neq 0{,}80$; $|z_{obs}| = 4{,}714$; $z_{0,95} = 1{,}645$; rejeitar H_0.

18. a) $H_1: \mu_D < 0$; $t_{obs} =$ -1,528; $valor\ p =$ 0,0852; não rejeitar H_0, i. e., não.

b) $H_1: \rho \neq 0$; $|t_{obs}| = 7{,}705$; $valor\ p < 0{,}001$; rejeitar H_0, i. e., existe.

c) $H_1: \rho > 0{,}90$; $z_{obs} = 0{,}875$; $z_{0,95} =$ 1,645; não rejeitar H_0.

Capítulo 9

1. a) $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$; $f_{obs} = 0{,}557$; $valor\ p = 0{,}649$; não rejeitar H_0.

b)

ANOVA

Quantidade calórica (Kj)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1486,833	3	495,611	9,186	,001
Within Groups	1079,000	20	53,950		
Total	2565,833	23			

c) $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$; $f_{obs} = 9{,}186$; $valor\ p = 0{,}001$; rejeitar H_0, i. e., existe. d) Sim.

e) $H_0: \mu_i = \mu_j, i \neq j$, e $i, j = 1, \ldots, 4$; Grupo 1: Escolas 3 e 2; Grupo 2: Escolas 2 e 1; Grupo 3: Escolas 1 e 4 (teste Tukey). f) 3 e 2; 1 e 4. g) A mesma.

2. a) $H_0: \mu_A = \mu_B = \mu_C = \mu_D$; $f_{obs} = 10{,}064$; $f_{3;\,24;\,0,95} = 3{,}01$; rejeitar H_0, i. e., idades médias diferentes.

b)

Fonte de Variação	Soma dos quadrados	Graus de liberdade	Média dos quadrados	F
Factor	72,964	3	24,321	10,064
Erro	58,000	24	2,417	
Total	130,964	27		

c) As amostras têm de ser independentes e $X_i \sim N(\mu_i; \sigma), i = 1, \ldots, 4$.

3. a) $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$; $f_{obs} = 0{,}665$; $valor\ p = 0{,}583$; não rejeitar H_0.

b)

ANOVA

Resistência do papel

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	382,792	3	127,597	19,605	,000
Within Groups	130,167	20	6,508		
Total	512,958	23			

c) $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$; $f_{obs} = 19{,}605$; $valor\ p < 0{,}001$; rejeitar H_0, i. e., há diferença. d) Sim.

e) $H_0: \mu_i = \mu_j, i \neq j$, e $i, j = 1, \ldots, 4$; Grupo 1: Concentração 5%, Grupo 2: Concentração 10% e 15%; Grupo 3: Concentração 20% (teste Tukey). f) 20%; 5%. g) A mesma.

4. a)

Fonte de Variação	Soma dos quadrados	Graus de liberdade	Média dos quadrados	F
Factor	64,615	2	32,308	10,422
Erro	62,000	20	3,100	
Total	126,615	22		

b) 9.c) H_0: $\mu_1 = \mu_2 = \mu_3$; $f_{obs} = 10{,}422$; $f_{2;\,20;\,0{,}95} = 3{,}49$; rejeitar H_0.

5. a)

Fonte de Variação	Soma dos quadrados	Graus de liberdade	Média dos quadrados	F
Factor	489,288	3	163,096	106,111
Erro	83,000	54	1,537	
Total	572,288	57		

b) 54. c) H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ *vs.* H_1: Nem todas as médias μ_i são iguais, $i = 1, \ldots, 4$.

d) $f_{obs} = 106{,}111$; $f_{3;\,54;\,0{,}95} = 2{,}78$; rejeitar H_0.

e) As amostras têm de ser independentes e $X_i \sim N(\mu_i; \sigma), i = 1, \ldots, 4$.

6. a) As amostras têm de ser independentes e $X_i \sim N(\mu_i; \sigma), i = M, T, N$.

b) i)

ANOVA

Duração

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3019,444	2	1509,722	4,567	,028
Within Groups	4958,333	15	330,556		
Total	7977,778	17			

ii) H_0: $\mu_M = \mu_T = \mu_N$; $f_{obs} = 4{,}567$; $valor\ p = 0{,}028$; rejeitar H_0.

iii) H_0: $\mu_i = \mu_j, i \neq j$ e $i,\ j = M,\ T,\ N$; Noite e tarde (pelo teste Tukey: Grupo 1 – noite e tarde, Grupo 2 – tarde e manhã).

7. a) H_0^A: A produção média é idêntica com os 3 fertilizantes; $f_{A_{obs}} = 6{,}237$; $valor\ p = 0{,}034$; não rejeitar H_0^A.

b) H_0^B: A produção média é idêntica com as 4 variedades de milho; $f_{B_{obs}} = 0{,}857$; $valor\ p = 0{,}512$; não rejeitar H_0^B.

Tests of Between-Subjects Effects

Dependent Variable: Produção

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	16,500[a]	5	3,300	3,009	,106
Intercept	554,880	1	554,880	505,970	,000
Fertilizante	13,680	2	6,840	6,237	,034
Centeio	2,820	3	,940	,857	,512
Error	6,580	6	1,097		
Total	577,960	12			
Corrected Total	23,080	11			

a. R Squared = ,715 (Adjusted R Squared = ,477)

8. a)

Tests of Between-Subjects Effects

Dependent Variable: N.º de jogos perdidos

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	11,667[a]	5	2,333	7,000	,017
Intercept	96,333	1	96,333	289,000	,000
Grupo	7,000	3	2,333	7,000	,022
Tipo	4,667	2	2,333	7,000	,027
Error	2,000	6	,333		
Total	110,000	12			
Corrected Total	13,667	11			

a. R Squared = ,854 (Adjusted R Squared = ,732)

b) H_0^A: O n.º médio de jogos perdidos é idêntico para os 3 tipos de jogo; $f_{A_{obs}} = 7$; $valor\ p = 0{,}027$; rejeitar H_0^A; $H_0: \mu_i = \mu_j,\ i \neq j$ e $i,\ j =$ Ofensivo, Misto, Defensivo (teste Tukey); Grupo 1 (perderam menos jogos): defensivo, misto; Grupo 2 (perderam mais jogos): Misto, ofensivo.

c) H_0^B: O n.º médio de jogos perdidos é idêntico nos 4 grupos; $f_{B_{obs}} = 7$; $valor\ p = 0{,}022$; rejeitar H_0^B; $H_0: \mu_i = \mu_j,\ i \neq j$ e $i,\ j = A, B, C, D$ (teste Tukey); Grupo 1 (perderam menos jogos): A, D, B; Grupo 2 (perderam mais jogos): B, C.

9. a) H_0^{AB}: Não existe interacção; H_0^A: O consumo médio é idêntico para os 3 jornalistas; H_0^B: O consumo médio é idêntico nas 2 zonas.

b) Interacção: $f_{AB_{obs}} = 3{,}709$; $valor\ p = 0{,}045$; rejeitar H_0^{AB}, i. e., existe interacção, logo não se testam as outras hipóteses.

Tests of Between-Subjects Effects

Dependent Variable: Consumo

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	58,933[a]	5	11,787	10,359	,000
Intercept	2181,227	1	2181,227	1917,094	,000
Jornalista	26,493	2	13,247	11,643	,001
Percurso	24,000	1	24,000	21,094	,000
Jornalista * Percurso	8,440	2	4,220	3,709	,045
Error	20,480	18	1,138		
Total	2260,640	24			
Corrected Total	79,413	23			

a. R Squared = ,742 (Adjusted R Squared = ,670)

10. a) 18;

b)

Fonte de Variação	Soma dos quadrados	Graus de liberdade	Média dos quadrados	F
Semestre	56,880	1	56,880	4,67
Método	98,912	2	49,456	4,07
Interacção	6,790	2	3,395	0,28
Erro	146,158	12	12,18	
Total	308,74	17		

c) Independência, Normalidade e homogeneidade das variâncias;

d) H_0^{AB} Não existe interacção; $f_{AB_{obs}} = 0{,}28$; $f_{2;\,12;\,0{,}99} = 6{,}93$; não rejeitar H_0^{AB};

H_0^A: Os resultados são idênticos nos 2 semestres; $f_{A_{obs}} = 4{,}67$; $f_{1;\,12;\,0{,}99} = 9{,}33$; não rejeitar H_0^B;

H_0^B: Os resultados são idênticos com os 3 métodos; $f_{B_{obs}} = 4{,}07$; $f_{2;\,12;\,0{,}99} = 6{,}93$; não rejeitar H_0^B.

Capítulo 10

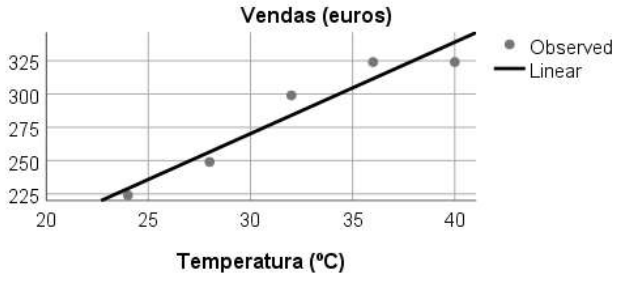
1. $H_0: p_A = 0{,}4$ e $p_B = 0{,}15$ e $p_{AB} = 0{,}05$ e $p_O = 0{,}4$; $\chi^2_{obs} = 83{,}333$; $valor\ p < 0{,}001$; rejeitar H_0.
2. $H_0: p_1 = 0{,}29$ e $p_2 = 0{,}07$ e $p_3 = 0{,}48$ e $p_4 = 0{,}16$; $\chi^2_{obs} = 6{,}935$; $valor\ p = 0{,}074$; não rejeitar H_0.
3. $H_0: p_B = 0{,}5$ e $p_F = 0{,}2$ e $p_D = 0{,}3$; $\chi^2_{obs} = 0{,}912$; $valor\ p = 0{,}634$; não rejeitar H_0.
4. $H_0: p_1 = p_2 = p_3 = p_4 = 0{,}25$; $\chi^2_{obs} = 31{,}429$; $valor\ p < 0{,}001$; rejeitar H_0, i. e., não.
5. $H_0: X \sim P(\lambda)$; $\hat{\lambda} = 2{,}22$; $\chi^2_{obs} = 16{,}401$; $valor\ p = 0{,}003$; rejeitar H_0.
6. $H_0: X \sim P(\lambda = 1)$; $\chi^2_{obs} = 29{,}345$; $valor\ p < 0{,}001$; rejeitar H_0.
7. $H_0: X \sim N(\mu = 55;\ \sigma = 10)$; $\chi^2_{obs} = 0{,}555$; $valor\ p = 0{,}968$; não rejeitar H_0.
8. $H_0: X \sim N(\mu;\ \sigma)$; $\hat{\mu} = -0{,}025$; $\hat{\sigma} = 1{,}5$; $\chi^2_{obs} = 61{,}731$; $valor\ p < 0{,}001$; rejeitar H_0.
9. a) $H_0: X \sim N(\mu = 29;\ \sigma = 6)$; Teste de Kolomogorov-Smirnov: $d_{obs} = 0{,}2382$; $d_{10;\,0{,}9} = 0{,}189$; rejeitar H_0.

   b) $H_0: X \sim N(\mu;\ \sigma)$; $\hat{\mu} = 147{,}9$; $\hat{\sigma} = 4{,}025$; Teste de Kolmogorov-Smirnov com correção de Lilliefors: $d_{obs} = 0{,}085$; $valor\ p > 0{,}200$; não rejeitar H_0; Teste de Shapiro-Wilk: $w_{obs} = 0{,}975$; $valor\ p = 0{,}494$; não rejeitar H_0.
10. H_0: A opinião é independente do distrito; $\chi^2_{obs} = 6{,}753$; $valor\ p = 0{,}009$; rejeitar H_0.
11. H_0: O uso de anti-ácidos contendo alumínio é independente da doença de Alzheimer; $\chi^2_{obs} = 6{,}573$; $valor\ p = 0{,}037$; rejeitar H_0 se $\alpha \geq 3{,}7\%$, i. e., para $\alpha = 1\%$ não, para $\alpha = 5\%$ e 10% sim.

12. H_0: O resultado é independente do tipo de medicamento; $\chi^2_{obs} = 0{,}898$; $valor\ p = 0{,}638$; não rejeitar H_0.
13. a) $H_1: P(+) \neq P(-) \Leftrightarrow H_1: \tilde{\mu}_1 - \tilde{\mu}_2 \neq$; $s^+_{obs} = 4$; $valor\ p = 0{,}754$; não rejeitar H_0, i. e., sim.
    b) $H_1: \rho_S \neq 0$; $t_{obs} = 1{,}695$, $valor\ p = 0{,}129$; rejeitar H_0, i .e. não.
14. $H_1: P(+) < P(-) \Leftrightarrow H_1: \tilde{\mu}_1 - \tilde{\mu}_2 < 0$; $z_{obs} = 2{,}28$; $z_{0{,}95} = 1{,}96$; rejeitar H_0, i. e., verdade; Teste de Wilcoxon.
15. $H_1: \tilde{\mu}_1 - \tilde{\mu}_2 \neq 0$; $s^+_{obs} = 8$; $valor\ p = 0{,}039$; não rejeitar H_0, i. e., não.
16. a) $H_1: \rho_S \neq 0$; $r_s = 0{,}232$; $r_{6;\,0{,}975} = 0{,}886$; não rejeitar H_0. b) Não.
17. $H_1: \rho_S > 0$; $r_s = 0{,}81$; $r_{7;\,0{,}95} = 0{,}714$; rejeitar H_0, i. e., sim.
18. $H_1: \rho_S > 0$; $t_{obs} = 13{,}31$; $t_{29;\,0{,}9} = 1{,}311$; rejeitar H_0, i. e., sim.
19. $H_1: P(+) > P(-)$; $s^+_{obs} = 5$; $valor\ p = 0{,}109$; não rejeitar H_0, i. e., não.
20. $H_1: \tilde{\mu} < 60$; $s^+_{obs} = 6$; $valor\ p = 0{,}828$; não rejeitar H_0, i. e., não.
21. a) $H_0: \tilde{\mu}_1 = \tilde{\mu}_2 = \tilde{\mu}_3$; $\chi^2_{obs} = 7{,}54$; $valor\ p = 0{,}023$; rejeitar H_0.
    b) $H_1: \tilde{\mu} < 130$; $s^-_{obs} = 2$; $valor\ p = 0{,}344$; não rejeitar H_0, i. e., não.

Capítulo 11

1. a) $\hat{y}_i =$ -1,071 + 2,741 x_i. Quando não é administrada dosagem de medicamento espera-se, em média, que o tempo de duração de efeito do medicamento seja de –1,071 dias, logo não faz sentido interpretar β_0. Por cada aumento de 1 miligrama de medicamento administrada espera-se em média que o tempo de duração de efeito do medicamento aumente 2,741 dias;
   b) $r = 0{,}910$, i. e., relação linear positiva forte. Quanto maior a dosagem de medicamento administrada mais tempo dura o efeito do mesmo. $r^2 = 0{,}828$. Bom ajustamento, pois 82,8% da variabilidade que ocorre na quantidade de dosagem de medicamento administrada é explicada pela relação linear com o tempo de duração do efeito do medicamento.
   c) $H_1: \beta_1 \neq 0$; $|t_{obs}| = 6{,}214$; $valor\ p < 0{,}001$, rejeitar H_0. d) 16,744.
2. a) $\hat{y} =$ -0,235 + 1,005 x; $\hat{\sigma}^2 = 0{,}875$; $r^2 \approx 0{,}984$.
   b) $IC_{95\%}\ \beta_0 =$]-3,549; 3,080[; $IC_{95\%}\ \beta_1 =$]0,900; 1,109[;
   $H_1: \beta_0 \neq 0$; $|t_{obs}| = 0{,}163$; $t_{8;\,0{,}975} = 2{,}306$; não rejeitar H_0;
   $H_1: \beta_1 \neq 1$; $|t_{obs}| = 0{,}103$; $t_{8;\,0{,}975} = 2{,}306$; não rejeitar H_0.
3. a) Por cada ponto a mais na nota do teste escrito no início do curso verifica-se um aumento médio de 0,2875 pontos na nota final do aluno no fim do curso.
   b) 0,1158. Mau ajustamento. c) $H_1: \beta_1 \neq 0$; $|t_{obs}| = 6{,}2965$; $t_{304;0{,}975} = 1{,}96$; rejeitar H_0.
4. a) $\hat{y} = 64 + 6{,}875\ x$.
   b)



   c) 12,5.
   d) 0,957, i. e., relação linear positiva muito forte.
   e) 0,917, i. e., bom ajustamento.
   f) 229,167.
   g)

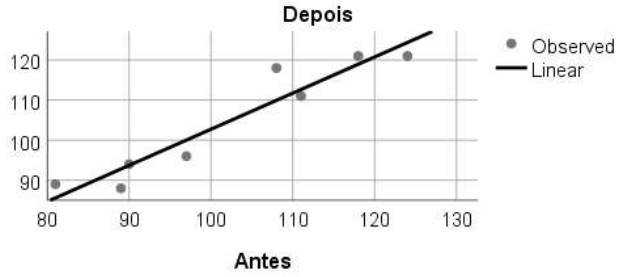
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7562,500	1	7562,500	33,000	,010[b]
	Residual	687,500	3	229,167		
	Total	8250,000	4			

a. Dependent Variable: Vendas (euros)
b. Predictors: (Constant), Temperatura (ºC)

h) -59,768 e 187,768. i)]-0,115. 13,865[. j) 0,198. k) $H_1: \beta_1 \neq 0$; $|t_{obs}| = 5{,}745$; $valor\ p = 0{,}01$; rejeitar H_0.
l) 304,625.

5. a) $\hat{y} = 12{,}566 + 0{,}902\ x$.
b)



c) $x = 90 \Rightarrow e = 0{,}2941$. d) 0,953, i. e., relação linear positiva muito forte. e) 0,908; sim. f) 22,359.
g)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1329,349	1	1329,349	59,456	,000[b]
	Residual	134,151	6	22,359		
	Total	1463,500	7			

a. Dependent Variable: Depois
b. Predictors: (Constant), Antes

h)]-32,189; 57,320[. i) 0,615 e 1,188. j) $H_1: \beta_0 \neq 0$; $|t_{obs}| = 1{,}041$; 0,338; não rejeitar H_0. k) <0,001.
l) 96,452.

6. a) Sim. b) $H_1: \rho > 0$; $t_{obs} = 6{,}9753$; $valor\ p < 0{,}001$; rejeitar H_0, i. e., sim. c) $\hat{y} = 2{,}772 + 0{,}424\ x$.
d) Num jogo em que o n.º de cruzamentos foi 0 efectuam-se, em média, 2,772 remates. Por cada cruzamento que se faça a mais num jogo verifica-se um aumento médio de 0,424 no n.º de remates realizados.
e) 0,873; sim. f) $H_1: \beta_0 \neq 0$; $t_{obs} = 1{,}322$; $valor\ p = 0{,}228$; não rejeitar H_0.
g) $H_1: \beta_1 \neq 0$; $t_{obs} = 6{,}952$; Não. h) 26,504. i) 1,496.

7. a) O aumento de 1 u. m. no preço do café origina um decréscimo médio de 0,479 no n.º médio de chávenas de café consumidas por dia.
b)]2,410; 2,972[. Com 95% de confiança, quando o preço do café é de 0 u. m., então o n.º médio de chávenas de café consumidas por dia encontra-se entre 2; 410 e 2; 872;
c)]-0,742; -0,216[. com 95% de confiança, estima-se o aumento de 1 u. m. no preço do café, o que origina um decréscimo médio entre 0; 216 e 0; 742 no n.º médio de chávenas de café consumidas por dia;
d) $H_1: \beta_1 \neq 0$; $|t_{obs}| = 4{,}202$; $valor\ p = 0{,}001$; rejeitar H_0.
e) $H_1: \beta_0 \neq 0$; $|t_{obs}| = 22{,}057$; $valor\ p < 0{,}001$; rejeitar H_0.
f) Rejeitar H_0.
g)

Coefficients

Model		Unstandardized Coefficients		t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error			Lower Bound	Upper Bound
1	(Constant)	2,691	0,122	22,057	0,000	2,410	2,972
	Preço (em euros)	-0,479	0,114	-4,202	0,001	-0,742	-0,216

Bibliografia

Brown, M. B., & Forsythe, A. B. (1974a). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, 16, 385-389.

Brown, M. B., and A. B. Forsythe. 1974b. Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69, 364-367.

Galvão de Mello, F. (2000). *Probabilidades e estatística: conceitos e métodos fundamentais.* Vol. I. Escolar Editora.

Galvão de Mello, F. (1997). *Probabilidades e estatística: conceitos e métodos* fundamentais. Vol. II. Escolar Editora.

Griffiths, W. E., Hill, R. C. e Judge, G. G. (1993). *Learning and practicing econometrics*. John Wiley & Sons, Inc.

Guimarães, R. C. e Cabral, J. A. S. (2010). *Estatística*. 2ª Edição. Verlag Dashöfer.

Instituto Nacional de Estatística, Direção Geral dos Estabelecimentos Escolares e Escola Secundária Tomaz Pelayo. *Acção local de estatística aplicada*. http://www.alea.pt. Consultado a 1 de Agosto de 2019.

Levene, H. (1960). Essays in Honor of Harold Hotelling. In *Contributions to probability and statistics* (I. Olkin *et al.* eds.). Stanford University Press, pp. 278-292.

Murteira, B. e Black, G. (1983). *Estatística descritiva*. McGraw-Hill.

Murteira, B., Ribeiro, C. S., Silva, J. A. e Pimenta, C. (2007). *Introdução à estatística*. McGraw-Hill.

Newbold, P., Carlson, W. e Thorne, B. (2013). *Statistics for business and economics*. 8ª Edição. Pearson.

Pestana, D. D. e Velosa, S. (2002). *Introdução à probabilidade e à estatística*. Volume 1, Fundação Calouste Gulbenkien.

Pestana, M. H. e Gageiro, J. N. (2014). *Análise de dados para ciências sociais. A complementaridade do SPSS*. 6ª Edição. Edições Sílabo.

Pires, A. M., & Amado, C. (2008). Interval estimators for a binomial proportion: Comparison of twenty methods. *REVSTAT–Statistical Journal*, 6(2), 165-197.

Reis, E. (2008). *Estatística descritiva*. 7ª Edição, Edições Sílabo.

Reis, E., Melo, P., Andrade, R. e Calapez, T. (1999). *Exercícios de estatística aplicada*. Volume I e II. 3ª Edição. Edições Sílabo.

Sheskin, D. J. (2011). *Handbook of parametric and nonparametric statistical procedures*. 5ª Edição. Chapman and Hall/CRC.

Silva, C. M. (1994). *Estatística aplicada à psicologia e ciências sociais*. McGraw-Hill.

Sociedade Portuguesa de Estatística e Associação Brasileira de Estatística. *Glossário Inglês-Português de Estatística*. http://glossario.spestatistica.pt/. Consultado a 1 de Agosto de 2019.

Triola, M. F. (2017). Introdução à Estatística. 12ª Edição. LTC.

Welch, B. L. 1947. The generalization of Student's problem when several different population *variances* are involved. *Biometrika*, 34, 28-35.

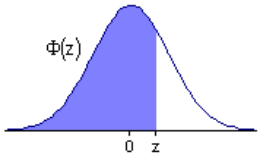
Welch, B. L. 1951. On the comparison of several mean values: an alternative approach. *Biometrika*, 38, 330-336.

Anexos

A Distribuição Normal Padrão

Valores da função de distribuição:

$$Z \sim N(0;1): \Phi(z) = F(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$



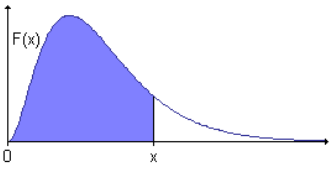
z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

z	1,282	1,645	1,96	2,326	2,576	3,09	3,291	3,891	4,417
$\Phi(z)$	0,90	0,95	0,975	0,99	0,995	0,999	0,9995	0,99995	0,999995
$2(1-\Phi(z))$	0,20	0,10	0,05	0,02	0,01	0,002	0,001	0,0001	0,00001

B Distribuição Qui-Quadrado

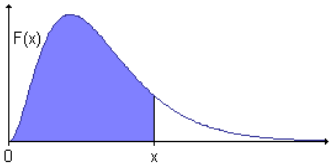
Valores da função de distribuição:

$$X \sim \chi^2_n : F(x) = P(X \leq x) = p$$



n \ p	0,0005	0,001	0,005	0,01	0,025	0,05	0,075	0,1	0,15	0,2	0,3	0,4
1	3,9E-07	1,6E-06	3,9E-05	1,6E-04	9,8E-04	0,004	0,009	0,016	0,036	0,064	0,148	0,275
2	0,0010	0,0020	0,0100	0,0201	0,0506	0,103	0,156	0,211	0,325	0,446	0,713	1,022
3	0,0153	0,0243	0,0717	0,115	0,216	0,352	0,472	0,584	0,798	1,005	1,424	1,869
4	0,0639	0,0908	0,207	0,297	0,484	0,711	0,897	1,064	1,366	1,649	2,195	2,753
5	0,158	0,210	0,412	0,554	0,831	1,145	1,394	1,610	1,994	2,343	3,000	3,656
6	0,299	0,381	0,676	0,872	1,237	1,635	1,941	2,204	2,661	3,070	3,828	4,570
7	0,485	0,599	0,989	1,239	1,690	2,167	2,528	2,833	3,358	3,822	4,671	5,493
8	0,710	0,857	1,344	1,647	2,180	2,733	3,144	3,490	4,078	4,594	5,527	6,423
9	0,972	1,152	1,735	2,088	2,700	3,325	3,785	4,168	4,817	5,380	6,393	7,357
10	1,265	1,479	2,156	2,558	3,247	3,940	4,446	4,865	5,570	6,179	7,267	8,295
11	1,587	1,834	2,603	3,053	3,816	4,575	5,124	5,578	6,336	6,989	8,148	9,237
12	1,935	2,214	3,074	3,571	4,404	5,226	5,818	6,304	7,114	7,807	9,034	10,18
13	2,305	2,617	3,565	4,107	5,009	5,892	6,524	7,041	7,901	8,634	9,926	11,13
14	2,697	3,041	4,075	4,660	5,629	6,571	7,242	7,790	8,696	9,467	10,82	12,08
15	3,107	3,483	4,601	5,229	6,262	7,261	7,969	8,547	9,499	10,31	11,72	13,03
16	3,536	3,942	5,142	5,812	6,908	7,962	8,707	9,312	10,31	11,15	12,62	13,98
17	3,980	4,416	5,697	6,408	7,564	8,672	9,452	10,09	11,12	12,00	13,53	14,94
18	4,439	4,905	6,265	7,015	8,231	9,390	10,21	10,86	11,95	12,86	14,44	15,89
19	4,913	5,407	6,844	7,633	8,907	10,12	10,97	11,65	12,77	13,72	15,35	16,85
20	5,398	5,921	7,434	8,260	9,591	10,85	11,73	12,44	13,60	14,58	16,27	17,81
21	5,895	6,447	8,034	8,897	10,28	11,59	12,50	13,24	14,44	15,44	17,18	18,77
22	6,404	6,983	8,643	9,542	10,98	12,34	13,28	14,04	15,28	16,31	18,10	19,73
23	6,924	7,529	9,260	10,20	11,69	13,09	14,06	14,85	16,12	17,19	19,02	20,69
24	7,453	8,085	9,886	10,86	12,40	13,85	14,85	15,66	16,97	18,06	19,94	21,65
25	7,991	8,649	10,52	11,52	13,12	14,61	15,64	16,47	17,82	18,94	20,87	22,62
26	8,537	9,222	11,16	12,20	13,84	15,38	16,44	17,29	18,67	19,82	21,79	23,58
27	9,093	9,803	11,81	12,88	14,57	16,15	17,24	18,11	19,53	20,70	22,72	24,54
28	9,656	10,39	12,46	13,56	15,31	16,93	18,05	18,94	20,39	21,59	23,65	25,51
29	10,23	10,99	13,12	14,26	16,05	17,71	18,85	19,77	21,25	22,48	24,58	26,48
30	10,80	11,59	13,79	14,95	16,79	18,49	19,66	20,60	22,11	23,36	25,51	27,44
31	11,39	12,20	14,46	15,66	17,54	19,28	20,48	21,43	22,98	24,26	26,44	28,41
32	11,98	12,81	15,13	16,36	18,29	20,07	21,30	22,27	23,84	25,15	27,37	29,38
33	12,58	13,43	15,82	17,07	19,05	20,87	22,12	23,11	24,71	26,04	28,31	30,34
34	13,18	14,06	16,50	17,79	19,81	21,66	22,94	23,95	25,59	26,94	29,24	31,31
35	13,79	14,69	17,19	18,51	20,57	22,47	23,76	24,80	26,46	27,84	30,18	32,28
36	14,40	15,32	17,89	19,23	21,34	23,27	24,59	25,64	27,34	28,73	31,12	33,25
37	15,02	15,97	18,59	19,96	22,11	24,07	25,42	26,49	28,21	29,64	32,05	34,22
38	15,64	16,61	19,29	20,69	22,88	24,88	26,25	27,34	29,09	30,54	32,99	35,19
39	16,27	17,26	20,00	21,43	23,65	25,70	27,09	28,20	29,97	31,44	33,93	36,16
40	16,91	17,92	20,71	22,16	24,43	26,51	27,93	29,05	30,86	32,34	34,87	37,13
50	23,46	24,67	27,99	29,71	32,36	34,76	36,40	37,69	39,75	41,45	44,31	46,86
60	30,34	31,74	35,53	37,48	40,48	43,19	45,02	46,46	48,76	50,64	53,81	56,62
80	44,79	46,52	51,17	53,54	57,15	60,39	62,57	64,28	66,99	69,21	72,92	76,19
100	59,89	61,92	67,33	70,06	74,22	77,93	80,41	82,36	85,44	87,95	92,13	95,81
150	99,46	102,1	109,1	112,7	118,0	122,7	125,8	128,3	132,1	135,3	140,5	145,0
200	140,7	143,8	152,2	156,4	162,7	168,3	172,0	174,8	179,4	183,0	189,0	194,3

B Distribuição Qui-Quadrado (continuação)



Valores da função de distribuição:

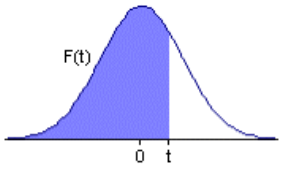
$$X \sim \chi^2_n : F(x) = P(X \leq x) = p$$

n \ p	0,500	0,600	0,700	0,800	0,850	0,900	0,925	0,95	0,975	0,99	0,995	0,999	0,9995
1	0,455	0,708	1,074	1,642	2,072	2,706	3,170	3,841	5,024	6,635	7,879	10,83	12,12
2	1,386	1,833	2,408	3,219	3,794	4,605	5,181	5,991	7,378	9,210	10,60	13,82	15,20
3	2,366	2,946	3,665	4,642	5,317	6,251	6,905	7,815	9,348	11,34	12,84	16,27	17,73
4	3,357	4,045	4,878	5,989	6,745	7,779	8,496	9,488	11,14	13,28	14,86	18,47	20,00
5	4,351	5,132	6,064	7,289	8,115	9,236	10,01	11,07	12,83	15,09	16,75	20,51	22,11
6	5,348	6,211	7,231	8,558	9,446	10,64	11,47	12,59	14,45	16,81	18,55	22,46	24,10
7	6,346	7,283	8,383	9,803	10,75	12,02	12,88	14,07	16,01	18,48	20,28	24,32	26,02
8	7,344	8,351	9,524	11,03	12,03	13,36	14,27	15,51	17,53	20,09	21,95	26,12	27,87
9	8,343	9,414	10,66	12,24	13,29	14,68	15,63	16,92	19,02	21,67	23,59	27,88	29,67
10	9,342	10,47	11,78	13,44	14,53	15,99	16,97	18,31	20,48	23,21	25,19	29,59	31,42
11	10,34	11,53	12,90	14,63	15,77	17,28	18,29	19,68	21,92	24,73	26,76	31,26	33,14
12	11,34	12,58	14,01	15,81	16,99	18,55	19,60	21,03	23,34	26,22	28,30	32,91	34,82
13	12,34	13,64	15,12	16,98	18,20	19,81	20,90	22,36	24,74	27,69	29,82	34,53	36,48
14	13,34	14,69	16,22	18,15	19,41	21,06	22,18	23,68	26,12	29,14	31,32	36,12	38,11
15	14,34	15,73	17,32	19,31	20,60	22,31	23,45	25,00	27,49	30,58	32,80	37,70	39,72
16	15,34	16,78	18,42	20,47	21,79	23,54	24,72	26,30	28,85	32,00	34,27	39,25	41,31
17	16,34	17,82	19,51	21,61	22,98	24,77	25,97	27,59	30,19	33,41	35,72	40,79	42,88
18	17,34	18,87	20,60	22,76	24,16	25,99	27,22	28,87	31,53	34,81	37,16	42,31	44,43
19	18,34	19,91	21,69	23,90	25,33	27,20	28,46	30,14	32,85	36,19	38,58	43,82	45,97
20	19,34	20,95	22,77	25,04	26,50	28,41	29,69	31,41	34,17	37,57	40,00	45,31	47,50
21	20,34	21,99	23,86	26,17	27,66	29,62	30,92	32,67	35,48	38,93	41,40	46,80	49,01
22	21,34	23,03	24,94	27,30	28,82	30,81	32,14	33,92	36,78	40,29	42,80	48,27	50,51
23	22,34	24,07	26,02	28,43	29,98	32,01	33,36	35,17	38,08	41,64	44,18	49,73	52,00
24	23,34	25,11	27,10	29,55	31,13	33,20	34,57	36,42	39,36	42,98	45,56	51,18	53,48
25	24,34	26,14	28,17	30,68	32,28	34,38	35,78	37,65	40,65	44,31	46,93	52,62	54,95
26	25,34	27,18	29,25	31,79	33,43	35,56	36,98	38,89	41,92	45,64	48,29	54,05	56,41
27	26,34	28,21	30,32	32,91	34,57	36,74	38,18	40,11	43,19	46,96	49,65	55,48	57,86
28	27,34	29,25	31,39	34,03	35,71	37,92	39,38	41,34	44,46	48,28	50,99	56,89	59,30
29	28,34	30,28	32,46	35,14	36,85	39,09	40,57	42,56	45,72	49,59	52,34	58,30	60,73
30	29,34	31,32	33,53	36,25	37,99	40,26	41,76	43,77	46,98	50,89	53,67	59,70	62,16
31	30,34	32,35	34,60	37,36	39,12	41,42	42,95	44,99	48,23	52,19	55,00	61,10	63,58
32	31,34	33,38	35,66	38,47	40,26	42,58	44,13	46,19	49,48	53,49	56,33	62,49	64,99
33	32,34	34,41	36,73	39,57	41,39	43,75	45,31	47,40	50,73	54,78	57,65	63,87	66,40
34	33,34	35,44	37,80	40,68	42,51	44,90	46,49	48,60	51,97	56,06	58,96	65,25	67,80
35	34,34	36,47	38,86	41,78	43,64	46,06	47,66	49,80	53,20	57,34	60,27	66,62	69,20
36	35,34	37,50	39,92	42,88	44,76	47,21	48,84	51,00	54,44	58,62	61,58	67,98	70,59
37	36,34	38,53	40,98	43,98	45,89	48,36	50,01	52,19	55,67	59,89	62,88	69,35	71,97
38	37,34	39,56	42,05	45,08	47,01	49,51	51,17	53,38	56,90	61,16	64,18	70,70	73,35
39	38,34	40,59	43,11	46,17	48,13	50,66	52,34	54,57	58,12	62,43	65,48	72,06	74,72
40	39,34	41,62	44,16	47,27	49,24	51,81	53,50	55,76	59,34	63,69	66,77	73,40	76,10
50	49,33	51,89	54,72	58,16	60,35	63,17	65,03	67,50	71,42	76,15	79,49	86,66	89,56
60	59,33	62,13	65,23	68,97	71,34	74,40	76,41	79,08	83,30	88,38	91,95	99,61	102,7
80	79,33	82,57	86,12	90,41	93,11	96,58	98,86	101,9	106,6	112,3	116,3	124,8	128,3
100	99,33	102,9	106,9	111,7	114,7	118,5	121,0	124,3	129,6	135,8	140,2	149,4	153,2
150	149,3	153,8	158,6	164,3	168,0	172,6	175,6	179,6	185,8	193,2	198,4	209,3	213,6
200	199,3	204,4	210,0	216,6	220,7	226,0	229,5	234,0	241,1	249,4	255,3	267,5	272,4

C Distribuição t-Student

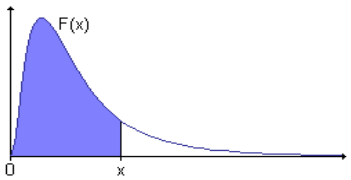
Valores da função de distribuição:

$T \sim t_n: F(t) = P(T \leq t) = p$



n \ p	0,6	0,7	0,8	0,9	0,925	0,95	0,975	0,99	0,995	0,999	0,9995
1	0,325	0,727	1,376	3,078	4,165	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,282	2,920	4,303	6,965	9,925	22,3	31,6
3	0,277	0,584	0,978	1,638	1,924	2,353	3,182	4,541	5,841	10,21	12,92
4	0,271	0,569	0,941	1,533	1,778	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	1,699	2,015	2,571	3,365	4,032	5,894	6,869
6	0,265	0,553	0,906	1,440	1,650	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,617	1,895	2,365	2,998	3,499	4,785	5,408
8	0,262	0,546	0,889	1,397	1,592	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,574	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,559	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,548	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,538	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,530	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,523	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,517	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,512	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,508	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,504	1,734	2,101	2,552	2,878	3,610	3,922
19	0,257	0,533	0,861	1,328	1,500	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,497	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,494	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,492	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,489	1,714	2,069	2,500	2,807	3,485	3,768
24	0,256	0,531	0,857	1,318	1,487	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,485	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,483	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,482	1,703	2,052	2,473	2,771	3,421	3,689
28	0,256	0,530	0,855	1,313	1,480	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,479	1,699	2,045	2,462	2,756	3,396	3,660
30	0,256	0,530	0,854	1,310	1,477	1,697	2,042	2,457	2,750	3,385	3,646
31	0,256	0,530	0,853	1,309	1,476	1,696	2,040	2,453	2,744	3,375	3,633
32	0,255	0,530	0,853	1,309	1,475	1,694	2,037	2,449	2,738	3,365	3,622
33	0,255	0,530	0,853	1,308	1,474	1,692	2,035	2,445	2,733	3,356	3,611
34	0,255	0,529	0,852	1,307	1,473	1,691	2,032	2,441	2,728	3,348	3,601
35	0,255	0,529	0,852	1,306	1,472	1,690	2,030	2,438	2,724	3,340	3,591
36	0,255	0,529	0,852	1,306	1,471	1,688	2,028	2,434	2,719	3,333	3,582
37	0,255	0,529	0,851	1,305	1,470	1,687	2,026	2,431	2,715	3,326	3,574
38	0,255	0,529	0,851	1,304	1,469	1,686	2,024	2,429	2,712	3,319	3,566
39	0,255	0,529	0,851	1,304	1,468	1,685	2,023	2,426	2,708	3,313	3,558
40	0,255	0,529	0,851	1,303	1,468	1,684	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,299	1,462	1,676	2,009	2,403	2,678	3,261	3,496
60	0,254	0,527	0,848	1,296	1,458	1,671	2,000	2,390	2,660	3,232	3,460
80	0,254	0,526	0,846	1,292	1,453	1,664	1,990	2,374	2,639	3,195	3,416
100	0,254	0,526	0,845	1,290	1,451	1,660	1,984	2,364	2,626	3,174	3,390
150	0,254	0,526	0,844	1,287	1,447	1,655	1,976	2,351	2,609	3,145	3,357
∞	0,253	0,524	0,842	1,282	1,440	1,645	1,960	2,326	2,576	3,090	3,291

D Distribuição F-Snedcor

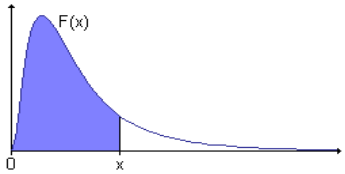


Valores da função de distribuição:

$F \sim F_{m;\,n}: F(x) = P(F \leq x) = p$

p	n \ m	1	2	3	4	5	6	7	8	9	10
0,9	**1**	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19
0,95		161	199	216	225	230	234	237	239	241	242
0,975		648	799	864	900	922	937	948	957	963	969
0,99		4052	5000	5403	5625	5764	5859	5928	5981	6022	6056
0,995		16212	19997	21614	22501	23056	23440	23715	23924	24091	24222
0,9	**2**	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39
0,95		18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
0,975		38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40
0,99		98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40
0,995		199	199	199	199	199,3	199	199	199	199	199
0,9	**3**	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23
0,95		10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
0,975		17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42
0,99		34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23
0,995		55,55	49,80	47,47	46,20	45,39	44,84	44,43	44,13	43,88	43,68
0,9	**4**	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92
0,95		7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
0,975		12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84
0,99		21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
0,995		31,33	26,28	24,26	23,15	22,46	21,98	21,62	21,35	21,14	20,97
0,9	**5**	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30
0,95		6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
0,975		10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62
0,99		16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
0,995		22,78	18,31	16,53	15,56	14,94	14,51	14,20	13,96	13,77	13,62
0,9	**6**	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94
0,95		5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
0,975		8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46
0,99		13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
0,995		18,63	14,54	12,92	12,03	11,46	11,07	10,79	10,57	10,39	10,25
0,9	**7**	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70
0,95		5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
0,975		8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76
0,99		12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
0,995		16,24	12,40	10,88	10,05	9,52	9,16	8,89	8,68	8,51	8,38
0,9	**8**	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54
0,95		5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
0,975		7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30
0,99		11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
0,995		14,69	11,04	9,60	8,81	8,30	7,95	7,69	7,50	7,34	7,21
0,9	**9**	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42
0,95		5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
0,975		7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96
0,99		10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
0,995		13,61	10,11	8,72	7,96	7,47	7,13	6,88	6,69	6,54	6,42
0,9	**10**	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32
0,95		4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
0,975		6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72
0,99		10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
0,995		12,83	9,43	8,08	7,34	6,87	6,54	6,30	6,12	5,97	5,85

D Distribuição F-Snedcor (continuação)



Valores da função de distribuição:

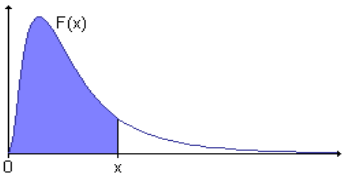
$F \sim F_{m;\,n}: F(x) = P(F \leq x) = p$

p	n \ m	1	2	3	4	5	6	7	8	9	10
0,9	**11**	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25
0,95		4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
0,975		6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53
0,99		9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
0,995		12,23	8,91	7,60	6,88	6,42	6,10	5,86	5,68	5,54	5,42
0,9	**12**	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19
0,95		4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
0,975		6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37
0,99		9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
0,995		11,75	8,51	7,23	6,52	6,07	5,76	5,52	5,35	5,20	5,09
0,9	**15**	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06
0,95		4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
0,975		6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06
0,99		8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
0,995		10,80	7,70	6,48	5,80	5,37	5,07	4,85	4,67	4,54	4,42
0,9	**20**	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94
0,95		4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
0,975		5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77
0,99		8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
0,995		9,94	6,99	5,82	5,17	4,76	4,47	4,26	4,09	3,96	3,85
0,9	**24**	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88
0,95		4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
0,975		5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64
0,99		7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
0,995		9,55	6,66	5,52	4,89	4,49	4,20	3,99	3,83	3,69	3,59
0,9	**30**	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82
0,95		4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
0,975		5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51
0,99		7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
0,995		9,18	6,35	5,24	4,62	4,23	3,95	3,74	3,58	3,45	3,34
0,9	**40**	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76
0,95		4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
0,975		5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39
0,99		7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
0,995		8,83	6,07	4,98	4,37	3,99	3,71	3,51	3,35	3,22	3,12
0,9	**60**	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71
0,95		4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
0,975		5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27
0,99		7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
0,995		8,49	5,79	4,73	4,14	3,76	3,49	3,29	3,13	3,01	2,90
0,9	**120**	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65
0,95		3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91
0,975		5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16
0,99		6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47
0,995		8,18	5,54	4,50	3,92	3,55	3,28	3,09	2,93	2,81	2,71
0,9	**∞**	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60
0,95		3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83
0,975		5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05
0,99		6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32
0,995		7,88	5,30	4,28	3,72	3,35	3,09	2,90	2,74	2,62	2,52

D Distribuição F-Snedcor (continuação)

Valores da função de distribuição:

$F \sim F_{m;\,n}: F(x) = P(F \leq x) = p$

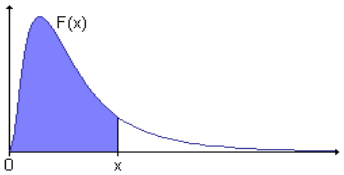


p	n \ m	11	12	15	20	24	30	40	60	120	∞
0,9	**1**	60,47	60,71	61,22	61,74	62,00	62,26	62,53	62,79	63,06	63,33
0,95		243	244	246	248	249	250	251	252	253	254
0,975		973	977	985	993	997	1001	1006	1010	1014	1018
0,99		6083	6106	6157	6209	6235	6261	6287	6313	6339	6366
0,995		24334	24427	24632	24837	24937	25041	25146	25254	25358	25466
0,9	**2**	9,40	9,41	9,42	9,44	9,45	9,46	9,47	9,47	9,48	9,49
0,95		19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
0,975		39,41	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50
0,99		99,41	99,42	99,43	99,45	99,46	99,47	99,48	99,48	99,49	99,50
0,995		199,41	199	199	199	199	199	199	199	199	200
0,9	**3**	5,22	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14	5,13
0,95		8,76	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
0,975		14,37	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95	13,90
0,99		27,13	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
0,995		43,52	43,39	43,08	42,78	42,62	42,47	42,31	42,15	41,99	41,83
0,9	**4**	3,91	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
0,95		5,94	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
0,975		8,79	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
0,99		14,45	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
0,995		20,82	20,70	20,44	20,17	20,03	19,89	19,75	19,61	19,47	19,32
0,9	**5**	3,28	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12	3,11
0,95		4,70	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
0,975		6,57	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
0,99		9,96	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
0,995		13,49	13,38	13,15	12,90	12,78	12,66	12,53	12,40	12,27	12,14
0,9	**6**	2,92	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
0,95		4,03	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
0,975		5,41	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
0,99		7,79	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
0,995		10,13	10,03	9,81	9,59	9,47	9,36	9,24	9,12	9,00	8,88
0,9	**7**	2,68	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49	2,47
0,95		3,60	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
0,975		4,71	4,67	4,57	4,47	4,41	4,36	4,31	4,25	4,20	4,14
0,99		6,54	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
0,995		8,27	8,18	7,97	7,75	7,64	7,53	7,42	7,31	7,19	7,08
0,9	**8**	2,52	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
0,95		3,31	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
0,975		4,24	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
0,99		5,73	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
0,995		7,10	7,01	6,81	6,61	6,50	6,40	6,29	6,18	6,06	5,95
0,9	**9**	2,40	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
0,95		3,10	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
0,975		3,91	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
0,99		5,18	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
0,995		6,31	6,23	6,03	5,83	5,73	5,62	5,52	5,41	5,30	5,19
0,9	**10**	2,30	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
0,95		2,94	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
0,975		3,66	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
0,99		4,77	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
0,995		5,75	5,66	5,47	5,27	5,17	5,07	4,97	4,86	4,75	4,64

D Distribuição F-Snedcor (continuação)

Valores da função de distribuição:

$F \sim F_{m;\,n}: F(x) = P(F \leq x) = p$



p	n \ m	11	12	15	20	24	30	40	60	120	∞
0,9	**11**	2,23	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
0,95		2,82	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
0,975		3,47	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
0,99		4,46	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
0,995		5,32	5,24	5,05	4,86	4,76	4,65	4,55	4,45	4,34	4,23
0,9	**12**	2,17	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
0,95		2,72	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
0,975		3,32	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
0,99		4,22	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
0,995		4,99	4,91	4,72	4,53	4,43	4,33	4,23	4,12	4,01	3,90
0,9	**15**	2,04	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76
0,95		2,51	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
0,975		3,01	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
0,99		3,73	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
0,995		4,33	4,25	4,07	3,88	3,79	3,69	3,59	3,48	3,37	3,26
0,9	**20**	1,91	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61
0,95		2,31	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
0,975		2,72	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
0,99		3,29	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
0,995		3,76	3,68	3,50	3,32	3,22	3,12	3,02	2,92	2,81	2,69
0,9	**24**	1,85	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53
0,95		2,22	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
0,975		2,59	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
0,99		3,09	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
0,995		3,50	3,42	3,25	3,06	2,97	2,87	2,77	2,66	2,55	2,43
0,9	**30**	1,79	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
0,95		2,13	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
0,975		2,46	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
0,99		2,91	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
0,995		3,25	3,18	3,01	2,82	2,73	2,63	2,52	2,42	2,30	2,18
0,9	**40**	1,74	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
0,95		2,04	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
0,975		2,33	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
0,99		2,73	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
0,995		3,03	2,95	2,78	2,60	2,50	2,40	2,30	2,18	2,06	1,93
0,9	**60**	1,68	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
0,95		1,95	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
0,975		2,22	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
0,99		2,56	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
0,995		2,82	2,74	2,57	2,39	2,29	2,19	2,08	1,96	1,83	1,69
0,9	**120**	1,57	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26	1,19
0,95		1,79	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
0,975		1,99	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
0,99		2,25	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
0,995		2,43	2,54	2,37	2,19	2,09	1,98	1,87	1,75	1,61	1,43
0,9	**∞**	2,17	1,55	1,49	1,42	1,38	1,34	1,30	1,24	1,17	1,00
0,95		2,72	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00
0,975		3,32	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00
0,99		4,22	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00
0,995		4,99	2,36	2,19	2,00	1,90	1,79	1,67	1,53	1,36	1,01

E Studentized range

Quantil 0,9:

$$X \sim q_{k;\,n}\colon F(x) = P\left(X \leq q_{k;\,n;\,0,9}\right) = 0{,}9$$

n \ k	2	3	4	5	6	7	8	9	10
1	8,929	13,437	16,358	18,488	20,150	21,504	22,642	23,621	24,477
2	4,129	5,733	6,772	7,538	8,139	8,633	9,049	9,409	9,725
3	3,328	4,467	5,199	5,738	6,162	6,511	6,806	7,062	7,287
4	3,015	3,976	4,586	5,035	5,388	5,679	5,926	6,139	6,327
5	2,850	3,717	4,264	4,664	4,979	5,238	5,458	5,648	5,816
6	2,748	3,558	4,065	4,435	4,726	4,966	5,168	5,344	5,499
7	2,679	3,451	3,931	4,280	4,555	4,780	4,971	5,137	5,283
8	2,630	3,374	3,834	4,169	4,431	4,646	4,829	4,987	5,126
9	2,592	3,316	3,761	4,084	4,337	4,545	4,721	4,873	5,007
10	2,563	3,270	3,704	4,018	4,264	4,465	4,636	4,783	4,913
11	2,540	3,234	3,658	3,965	4,205	4,401	4,567	4,711	4,838
12	2,521	3,204	3,621	3,921	4,156	4,349	4,511	4,652	4,776
13	2,504	3,179	3,589	3,885	4,116	4,304	4,464	4,602	4,724
14	2,491	3,158	3,563	3,854	4,081	4,267	4,424	4,560	4,679
15	2,479	3,140	3,540	3,828	4,052	4,235	4,390	4,524	4,641
16	2,469	3,124	3,520	3,804	4,026	4,207	4,360	4,492	4,608
17	2,460	3,110	3,503	3,784	4,003	4,182	4,334	4,464	4,579
18	2,452	3,098	3,487	3,766	3,984	4,161	4,310	4,440	4,553
19	2,445	3,087	3,474	3,751	3,966	4,142	4,290	4,418	4,530
20	2,439	3,077	3,462	3,736	3,950	4,124	4,271	4,398	4,510
21	2,433	3,069	3,451	3,724	3,936	4,109	4,255	4,380	4,491
22	2,428	3,061	3,441	3,712	3,923	4,095	4,239	4,364	4,474
23	2,424	3,054	3,432	3,701	3,911	4,082	4,226	4,350	4,459
24	2,420	3,047	3,423	3,692	3,900	4,070	4,213	4,336	4,445
25	2,416	3,041	3,416	3,683	3,890	4,059	4,201	4,324	4,432
26	2,412	3,036	3,409	3,675	3,881	4,049	4,191	4,313	4,420
27	2,409	3,030	3,402	3,667	3,873	4,040	4,181	4,302	4,409
28	2,406	3,026	3,396	3,660	3,865	4,032	4,172	4,293	4,399
29	2,403	3,021	3,391	3,654	3,858	4,024	4,163	4,284	4,389
30	2,400	3,017	3,386	3,648	3,851	4,016	4,155	4,275	4,381
31	2,398	3,013	3,381	3,642	3,845	4,009	4,148	4,268	4,372
32	2,396	3,010	3,376	3,637	3,839	4,003	4,141	4,260	4,365
33	2,393	3,006	3,372	3,632	3,833	3,997	4,135	4,253	4,357
34	2,391	3,003	3,368	3,627	3,828	3,991	4,129	4,247	4,351
35	2,389	3,000	3,364	3,623	3,823	3,986	4,123	4,241	4,344
36	2,388	2,998	3,361	3,619	3,819	3,981	4,117	4,235	4,338
37	2,386	2,995	3,357	3,615	3,814	3,976	4,112	4,230	4,332
38	2,384	2,992	3,354	3,611	3,810	3,972	4,107	4,224	4,327
39	2,383	2,990	3,351	3,608	3,806	3,967	4,103	4,220	4,322
40	2,381	2,988	3,348	3,605	3,802	3,963	4,099	4,215	4,317
48	2,372	2,973	3,330	3,583	3,778	3,937	4,070	4,185	4,285
60	2,363	2,959	3,312	3,562	3,755	3,911	4,042	4,155	4,254
80	2,353	2,945	3,294	3,541	3,731	3,885	4,014	4,125	4,223
120	2,344	2,930	3,276	3,520	3,707	3,859	3,986	4,096	4,191
240	2,335	2,916	3,258	3,499	3,684	3,834	3,959	4,066	4,160
∞	2,326	2,902	3,240	3,478	3,661	3,808	3,931	4,037	4,129

E Studentized range (continuação)

Quantil 0,9:

$$X \sim q_{k;\,n}: F(x) = P\left(X \leq q_{k;\,n;\,0,9}\right) = 0{,}9$$

n \ k	11	12	13	14	15	16	17	18	19	20
1	25,237	25,918	26,536	27,100	27,618	28,097	28,542	28,958	29,347	29,713
2	10,006	10,259	10,488	10,698	10,891	11,070	11,237	11,392	11,538	11,676
3	7,487	7,667	7,831	7,982	8,120	8,248	8,368	8,479	8,584	8,683
4	6,494	6,645	6,783	6,909	7,025	7,132	7,233	7,326	7,414	7,497
5	5,965	6,100	6,223	6,336	6,439	6,536	6,626	6,710	6,788	6,863
6	5,637	5,762	5,875	5,979	6,075	6,164	6,247	6,325	6,398	6,466
7	5,413	5,530	5,637	5,735	5,826	5,910	5,988	6,061	6,130	6,195
8	5,250	5,362	5,464	5,558	5,644	5,724	5,799	5,869	5,935	5,997
9	5,126	5,234	5,333	5,423	5,506	5,583	5,655	5,722	5,786	5,845
10	5,029	5,134	5,229	5,316	5,397	5,472	5,542	5,607	5,668	5,726
11	4,951	5,053	5,145	5,231	5,309	5,382	5,450	5,514	5,573	5,630
12	4,886	4,986	5,076	5,160	5,236	5,308	5,374	5,436	5,495	5,550
13	4,832	4,930	5,019	5,100	5,175	5,245	5,310	5,371	5,429	5,483
14	4,786	4,882	4,969	5,050	5,124	5,192	5,256	5,316	5,372	5,426
15	4,746	4,841	4,927	5,006	5,079	5,146	5,209	5,268	5,324	5,376
16	4,712	4,805	4,890	4,968	5,040	5,106	5,169	5,227	5,282	5,333
17	4,681	4,774	4,857	4,934	5,005	5,071	5,133	5,190	5,244	5,295
18	4,654	4,746	4,829	4,905	4,975	5,040	5,101	5,158	5,211	5,262
19	4,630	4,721	4,803	4,878	4,948	5,012	5,072	5,129	5,182	5,232
20	4,609	4,699	4,780	4,855	4,923	4,987	5,047	5,103	5,155	5,205
21	4,590	4,678	4,759	4,833	4,901	4,965	5,024	5,079	5,131	5,180
22	4,572	4,660	4,740	4,814	4,882	4,944	5,003	5,058	5,109	5,158
23	4,556	4,644	4,723	4,796	4,863	4,926	4,984	5,038	5,089	5,138
24	4,541	4,628	4,707	4,780	4,847	4,909	4,966	5,020	5,071	5,119
25	4,528	4,614	4,693	4,765	4,831	4,893	4,950	5,004	5,055	5,102
26	4,515	4,601	4,680	4,751	4,817	4,878	4,936	4,989	5,039	5,086
27	4,504	4,590	4,667	4,739	4,804	4,865	4,922	4,975	5,025	5,072
28	4,493	4,579	4,656	4,727	4,792	4,853	4,909	4,962	5,012	5,058
29	4,484	4,568	4,645	4,716	4,781	4,841	4,897	4,950	4,999	5,046
30	4,474	4,559	4,635	4,706	4,770	4,830	4,886	4,939	4,988	5,034
31	4,466	4,550	4,626	4,696	4,760	4,820	4,876	4,928	4,977	5,023
32	4,458	4,541	4,617	4,687	4,751	4,811	4,866	4,918	4,967	5,013
33	4,450	4,533	4,609	4,679	4,743	4,802	4,857	4,909	4,957	5,003
34	4,443	4,526	4,602	4,671	4,734	4,794	4,849	4,900	4,949	4,994
35	4,436	4,519	4,594	4,663	4,727	4,786	4,841	4,892	4,940	4,986
36	4,430	4,512	4,588	4,656	4,720	4,778	4,833	4,884	4,932	4,978
37	4,424	4,506	4,581	4,650	4,713	4,771	4,826	4,877	4,925	4,970
38	4,418	4,500	4,575	4,643	4,706	4,765	4,819	4,870	4,918	4,963
39	4,413	4,495	4,569	4,637	4,700	4,758	4,812	4,863	4,911	4,956
40	4,408	4,490	4,564	4,632	4,694	4,752	4,806	4,857	4,904	4,949
48	4,375	4,455	4,528	4,595	4,656	4,713	4,766	4,816	4,863	4,907
60	4,342	4,421	4,493	4,558	4,619	4,675	4,727	4,775	4,821	4,864
80	4,309	4,387	4,457	4,521	4,581	4,636	4,687	4,735	4,780	4,822
120	4,276	4,353	4,422	4,485	4,543	4,597	4,647	4,694	4,738	4,779
240	4,244	4,319	4,386	4,448	4,505	4,558	4,607	4,653	4,696	4,737
∞	4,211	4,285	4,351	4,412	4,468	4,519	4,568	4,612	4,654	4,694

E Studentized range (continuação)

Quantil 0,95:

$$X \sim q_{k;\,n}: F(x) = P\left(X \leq q_{k;\,n;\,0,95}\right) = 0{,}95$$

n \ k	2	3	4	5	6	7	8	9	10
1	17,969	26,976	32,819	37,082	40,408	43,119	45,397	47,357	49,071
2	6,085	8,331	9,798	10,881	11,734	12,435	13,027	13,539	13,988
3	4,501	5,910	6,825	7,502	8,037	8,478	8,852	9,177	9,462
4	3,926	5,040	5,757	6,287	6,706	7,053	7,347	7,602	7,826
5	3,635	4,602	5,218	5,673	6,033	6,330	6,582	6,801	6,995
6	3,460	4,339	4,896	5,305	5,628	5,895	6,122	6,319	6,493
7	3,344	4,165	4,681	5,060	5,359	5,606	5,815	5,997	6,158
8	3,261	4,041	4,529	4,886	5,167	5,399	5,596	5,767	5,918
9	3,199	3,948	4,415	4,755	5,024	5,244	5,432	5,595	5,738
10	3,151	3,877	4,327	4,654	4,912	5,124	5,304	5,460	5,598
11	3,113	3,820	4,256	4,574	4,823	5,028	5,202	5,353	5,486
12	3,081	3,773	4,199	4,508	4,750	4,950	5,119	5,265	5,395
13	3,055	3,734	4,151	4,453	4,690	4,884	5,049	5,192	5,318
14	3,033	3,701	4,111	4,407	4,639	4,829	4,990	5,130	5,253
15	3,014	3,673	4,076	4,367	4,595	4,782	4,940	5,077	5,198
16	2,998	3,649	4,046	4,333	4,557	4,741	4,896	5,031	5,150
17	2,984	3,628	4,020	4,303	4,524	4,705	4,858	4,991	5,108
18	2,971	3,609	3,997	4,276	4,494	4,673	4,824	4,955	5,071
19	2,960	3,593	3,977	4,253	4,468	4,645	4,794	4,924	5,037
20	2,950	3,578	3,958	4,232	4,445	4,620	4,768	4,895	5,008
21	2,941	3,565	3,942	4,213	4,424	4,597	4,743	4,870	4,981
22	2,933	3,553	3,927	4,196	4,405	4,577	4,722	4,847	4,957
23	2,926	3,542	3,914	4,180	4,388	4,558	4,702	4,826	4,935
24	2,919	3,532	3,901	4,166	4,373	4,541	4,684	4,807	4,915
25	2,913	3,523	3,890	4,153	4,358	4,526	4,667	4,789	4,897
26	2,907	3,514	3,880	4,141	4,345	4,511	4,652	4,773	4,880
27	2,902	3,506	3,870	4,130	4,333	4,498	4,638	4,758	4,864
28	2,897	3,499	3,861	4,120	4,322	4,486	4,625	4,745	4,850
29	2,892	3,493	3,853	4,111	4,311	4,475	4,613	4,732	4,837
30	2,888	3,486	3,845	4,102	4,301	4,464	4,601	4,720	4,824
31	2,884	3,481	3,838	4,094	4,292	4,454	4,591	4,709	4,812
32	2,881	3,475	3,832	4,086	4,284	4,445	4,581	4,698	4,802
33	2,877	3,470	3,825	4,079	4,276	4,436	4,572	4,689	4,791
34	2,874	3,465	3,820	4,072	4,268	4,428	4,563	4,680	4,782
35	2,871	3,461	3,814	4,066	4,261	4,421	4,555	4,671	4,773
36	2,868	3,457	3,809	4,060	4,255	4,414	4,547	4,663	4,764
37	2,865	3,453	3,804	4,054	4,249	4,407	4,540	4,655	4,756
38	2,863	3,449	3,799	4,049	4,243	4,400	4,533	4,648	4,749
39	2,861	3,445	3,795	4,044	4,237	4,394	4,527	4,641	4,741
40	2,858	3,442	3,791	4,039	4,232	4,388	4,521	4,634	4,735
48	2,843	3,420	3,764	4,008	4,197	4,351	4,481	4,592	4,690
60	2,829	3,399	3,737	3,977	4,163	4,314	4,441	4,550	4,646
80	2,814	3,377	3,711	3,947	4,129	4,277	4,402	4,509	4,603
120	2,800	3,356	3,685	3,917	4,096	4,241	4,363	4,468	4,560
240	2,786	3,335	3,659	3,887	4,063	4,205	4,324	4,427	4,517
∞	2,772	3,314	3,633	3,858	4,030	4,170	4,286	4,387	4,474

E Studentized range (continuação)

Quantil 0,95:

$$X \sim q_{k;\,n}: F(x) = P\left(X \leq q_{k;\,n;\,0,95}\right) = 0,95$$

n \ k	11	12	13	14	15	16	17	18	19	20
1	50,592	51,957	53,194	54,323	55,361	56,320	57,212	58,044	58,824	59,558
2	14,389	14,749	15,076	15,375	15,650	15,905	16,143	16,365	16,573	16,769
3	9,717	9,946	10,155	10,346	10,522	10,686	10,838	10,980	11,114	11,240
4	8,027	8,208	8,373	8,524	8,664	8,793	8,914	9,027	9,133	9,233
5	7,167	7,323	7,466	7,596	7,716	7,828	7,932	8,030	8,122	8,208
6	6,649	6,789	6,917	7,034	7,143	7,244	7,338	7,426	7,508	7,586
7	6,302	6,431	6,550	6,658	6,759	6,852	6,939	7,020	7,097	7,169
8	6,053	6,175	6,287	6,389	6,483	6,571	6,653	6,729	6,801	6,869
9	5,867	5,983	6,089	6,186	6,276	6,359	6,437	6,510	6,579	6,643
10	5,722	5,833	5,935	6,028	6,114	6,194	6,269	6,339	6,405	6,467
11	5,605	5,713	5,811	5,901	5,984	6,062	6,134	6,202	6,265	6,325
12	5,510	5,615	5,710	5,797	5,878	5,953	6,023	6,089	6,151	6,209
13	5,431	5,533	5,625	5,711	5,789	5,862	5,931	5,995	6,055	6,112
14	5,364	5,463	5,554	5,637	5,714	5,785	5,852	5,915	5,973	6,029
15	5,306	5,403	5,492	5,574	5,649	5,719	5,785	5,846	5,904	5,958
16	5,256	5,352	5,439	5,519	5,593	5,662	5,726	5,786	5,843	5,896
17	5,212	5,306	5,392	5,471	5,544	5,612	5,675	5,734	5,790	5,842
18	5,173	5,266	5,351	5,429	5,501	5,567	5,629	5,688	5,743	5,794
19	5,139	5,231	5,314	5,391	5,462	5,528	5,589	5,647	5,701	5,752
20	5,108	5,199	5,282	5,357	5,427	5,492	5,553	5,610	5,663	5,714
21	5,081	5,170	5,252	5,327	5,396	5,460	5,520	5,576	5,629	5,679
22	5,056	5,144	5,225	5,299	5,368	5,431	5,491	5,546	5,599	5,648
23	5,033	5,121	5,201	5,274	5,342	5,405	5,464	5,519	5,571	5,620
24	5,012	5,099	5,179	5,251	5,319	5,381	5,439	5,494	5,545	5,594
25	4,993	5,079	5,158	5,230	5,297	5,359	5,417	5,471	5,522	5,570
26	4,975	5,061	5,139	5,211	5,277	5,339	5,396	5,450	5,500	5,548
27	4,959	5,044	5,122	5,193	5,259	5,320	5,377	5,430	5,480	5,528
28	4,944	5,029	5,106	5,177	5,242	5,302	5,359	5,412	5,462	5,509
29	4,930	5,014	5,091	5,161	5,226	5,286	5,342	5,395	5,445	5,491
30	4,917	5,001	5,077	5,147	5,211	5,271	5,327	5,379	5,429	5,475
31	4,905	4,988	5,064	5,134	5,198	5,257	5,313	5,365	5,414	5,460
32	4,894	4,976	5,052	5,121	5,185	5,244	5,299	5,351	5,400	5,445
33	4,883	4,965	5,040	5,109	5,173	5,232	5,287	5,338	5,386	5,432
34	4,873	4,955	5,030	5,098	5,161	5,220	5,275	5,326	5,374	5,420
35	4,863	4,945	5,020	5,088	5,151	5,209	5,264	5,315	5,362	5,408
36	4,855	4,936	5,010	5,078	5,141	5,199	5,253	5,304	5,352	5,397
37	4,846	4,927	5,001	5,069	5,131	5,189	5,243	5,294	5,341	5,386
38	4,838	4,919	4,993	5,060	5,122	5,180	5,234	5,284	5,331	5,376
39	4,831	4,911	4,985	5,052	5,114	5,171	5,225	5,275	5,322	5,367
40	4,824	4,904	4,977	5,044	5,106	5,163	5,216	5,266	5,313	5,358
48	4,777	4,856	4,927	4,993	5,053	5,109	5,161	5,210	5,256	5,299
60	4,732	4,808	4,878	4,942	5,001	5,056	5,107	5,154	5,199	5,241
80	4,686	4,761	4,829	4,892	4,949	5,003	5,052	5,099	5,142	5,183
120	4,641	4,714	4,781	4,842	4,898	4,950	4,998	5,043	5,086	5,126
240	4,596	4,668	4,733	4,792	4,847	4,897	4,944	4,988	5,030	5,069
∞	4,552	4,622	4,685	4,743	4,796	4,845	4,891	4,934	4,974	5,012

E Studentized range (continuação)

Quantil 0,99:

$$X \sim q_{k;\,n}: F(x) = P\left(X \leq q_{k;\,n;\,0,99}\right) = 0,99$$

k / n	2	3	4	5	6	7	8	9	10
1	90,024	135,041	164,258	185,575	202,210	215,769	227,166	236,966	245,542
2	14,036	19,019	22,294	24,717	26,629	28,201	29,530	30,679	31,689
3	8,260	10,619	12,170	13,324	14,241	14,998	15,641	16,199	16,691
4	6,511	8,120	9,173	9,958	10,583	11,101	11,542	11,925	12,264
5	5,702	6,976	7,804	8,421	8,913	9,321	9,669	9,971	10,239
6	5,243	6,331	7,033	7,556	7,972	8,318	8,612	8,869	9,097
7	4,949	5,919	6,542	7,005	7,373	7,678	7,939	8,166	8,367
8	4,745	5,635	6,204	6,625	6,959	7,237	7,474	7,680	7,863
9	4,596	5,428	5,957	6,347	6,657	6,915	7,134	7,325	7,494
10	4,482	5,270	5,769	6,136	6,428	6,669	6,875	7,054	7,213
11	4,392	5,146	5,621	5,970	6,247	6,476	6,671	6,841	6,992
12	4,320	5,046	5,502	5,836	6,101	6,320	6,507	6,670	6,814
13	4,260	4,964	5,404	5,726	5,981	6,192	6,372	6,528	6,666
14	4,210	4,895	5,322	5,634	5,881	6,085	6,258	6,409	6,543
15	4,167	4,836	5,252	5,556	5,796	5,994	6,162	6,309	6,438
16	4,131	4,786	5,192	5,489	5,722	5,915	6,079	6,222	6,348
17	4,099	4,742	5,140	5,430	5,659	5,847	6,007	6,147	6,270
18	4,071	4,703	5,094	5,379	5,603	5,787	5,944	6,081	6,201
19	4,046	4,669	5,054	5,334	5,553	5,735	5,889	6,022	6,141
20	4,024	4,639	5,018	5,293	5,510	5,688	5,839	5,970	6,086
21	4,004	4,612	4,986	5,257	5,470	5,646	5,794	5,924	6,038
22	3,986	4,588	4,957	5,225	5,435	5,608	5,754	5,882	5,994
23	3,970	4,566	4,931	5,195	5,403	5,573	5,718	5,844	5,955
24	3,955	4,546	4,907	5,168	5,373	5,542	5,685	5,809	5,919
25	3,942	4,527	4,885	5,144	5,347	5,513	5,655	5,778	5,886
26	3,930	4,510	4,865	5,121	5,322	5,487	5,627	5,749	5,856
27	3,918	4,495	4,847	5,101	5,300	5,463	5,602	5,722	5,828
28	3,908	4,481	4,830	5,082	5,279	5,441	5,578	5,697	5,802
29	3,898	4,467	4,814	5,064	5,260	5,420	5,556	5,674	5,778
30	3,889	4,455	4,799	5,048	5,242	5,401	5,536	5,653	5,756
31	3,881	4,443	4,786	5,032	5,225	5,383	5,517	5,633	5,736
32	3,873	4,433	4,773	5,018	5,210	5,367	5,500	5,615	5,716
33	3,865	4,423	4,761	5,005	5,195	5,351	5,483	5,598	5,698
34	3,859	4,413	4,750	4,992	5,181	5,336	5,468	5,581	5,682
35	3,852	4,404	4,739	4,980	5,169	5,323	5,453	5,566	5,666
36	3,846	4,396	4,729	4,969	5,156	5,310	5,439	5,552	5,651
37	3,840	4,388	4,720	4,959	5,145	5,298	5,427	5,538	5,637
38	3,835	4,381	4,711	4,949	5,134	5,286	5,414	5,526	5,623
39	3,830	4,374	4,703	4,940	5,124	5,275	5,403	5,513	5,611
40	3,825	4,367	4,695	4,931	5,114	5,265	5,392	5,502	5,599
48	3,793	4,324	4,644	4,874	5,052	5,198	5,322	5,428	5,522
60	3,762	4,282	4,594	4,818	4,991	5,133	5,253	5,356	5,447
80	3,732	4,241	4,545	4,763	4,931	5,069	5,185	5,284	5,372
120	3,702	4,200	4,497	4,709	4,872	5,005	5,118	5,214	5,299
240	3,672	4,160	4,450	4,655	4,814	4,943	5,052	5,145	5,227
∞	3,643	4,120	4,403	4,603	4,757	4,882	4,987	5,078	5,157

E Studentized range (continuação)

Quantil 0,99:

$$X \sim q_{k;\,n}: F(x) = P\left(X \leq q_{k;\,n;\,0,99}\right) = 0,99$$

n \ k	11	12	13	14	15	16	17	18	19	20
1	253,15	259,98	266,17	271,81	277,00	281,80	286,26	290,43	294,33	298,00
2	32,589	33,398	34,134	34,806	35,426	36,000	36,534	37,034	37,502	37,943
3	17,130	17,526	17,887	18,217	18,522	18,805	19,068	19,315	19,546	19,765
4	12,567	12,840	13,090	13,318	13,530	13,726	13,909	14,081	14,242	14,394
5	10,479	10,696	10,894	11,076	11,244	11,400	11,545	11,682	11,811	11,932
6	9,300	9,485	9,653	9,808	9,951	10,084	10,208	10,325	10,434	10,538
7	8,548	8,711	8,860	8,997	9,124	9,242	9,353	9,456	9,553	9,645
8	8,027	8,176	8,311	8,436	8,552	8,659	8,760	8,854	8,943	9,027
9	7,646	7,784	7,910	8,025	8,132	8,232	8,325	8,412	8,495	8,573
10	7,356	7,485	7,603	7,712	7,812	7,906	7,993	8,075	8,153	8,226
11	7,127	7,250	7,362	7,464	7,560	7,648	7,731	7,809	7,883	7,952
12	6,943	7,060	7,166	7,265	7,356	7,441	7,520	7,594	7,664	7,730
13	6,791	6,903	7,006	7,100	7,188	7,269	7,345	7,417	7,484	7,548
14	6,663	6,772	6,871	6,962	7,047	7,125	7,199	7,268	7,333	7,394
15	6,555	6,660	6,756	6,845	6,927	7,003	7,074	7,141	7,204	7,264
16	6,461	6,564	6,658	6,744	6,823	6,897	6,967	7,032	7,093	7,151
17	6,380	6,480	6,572	6,656	6,733	6,806	6,873	6,937	6,997	7,053
18	6,309	6,407	6,496	6,579	6,655	6,725	6,791	6,854	6,912	6,967
19	6,246	6,342	6,430	6,510	6,585	6,654	6,719	6,780	6,837	6,891
20	6,190	6,285	6,370	6,449	6,523	6,591	6,654	6,714	6,770	6,823
21	6,140	6,233	6,317	6,395	6,467	6,534	6,596	6,655	6,710	6,762
22	6,095	6,186	6,269	6,346	6,417	6,482	6,544	6,602	6,656	6,707
23	6,054	6,144	6,226	6,301	6,371	6,436	6,497	6,553	6,607	6,658
24	6,017	6,105	6,186	6,261	6,330	6,394	6,453	6,510	6,562	6,612
25	5,983	6,070	6,150	6,224	6,292	6,355	6,414	6,469	6,522	6,571
26	5,951	6,038	6,117	6,190	6,257	6,319	6,378	6,432	6,484	6,533
27	5,923	6,008	6,087	6,158	6,225	6,287	6,344	6,399	6,450	6,498
28	5,896	5,981	6,058	6,129	6,195	6,256	6,314	6,367	6,418	6,465
29	5,871	5,955	6,032	6,103	6,168	6,228	6,285	6,338	6,388	6,435
30	5,848	5,932	6,008	6,078	6,142	6,202	6,258	6,311	6,361	6,407
31	5,827	5,910	5,985	6,055	6,119	6,178	6,234	6,286	6,335	6,381
32	5,807	5,889	5,964	6,033	6,096	6,155	6,211	6,262	6,311	6,357
33	5,789	5,870	5,944	6,013	6,076	6,134	6,189	6,240	6,289	6,334
34	5,771	5,852	5,926	5,994	6,056	6,114	6,169	6,220	6,268	6,313
35	5,755	5,835	5,908	5,976	6,038	6,096	6,150	6,200	6,248	6,293
36	5,739	5,819	5,892	5,959	6,021	6,078	6,132	6,182	6,229	6,274
37	5,725	5,804	5,876	5,943	6,004	6,061	6,115	6,165	6,212	6,256
38	5,711	5,790	5,862	5,928	5,989	6,046	6,099	6,148	6,195	6,239
39	5,698	5,776	5,848	5,914	5,974	6,031	6,084	6,133	6,179	6,223
40	5,685	5,764	5,835	5,900	5,961	6,017	6,069	6,118	6,165	6,208
48	5,606	5,681	5,750	5,814	5,872	5,926	5,977	6,024	6,069	6,111
60	5,528	5,601	5,667	5,728	5,784	5,837	5,886	5,931	5,974	6,015
80	5,451	5,521	5,585	5,644	5,698	5,749	5,796	5,840	5,881	5,920
120	5,375	5,443	5,505	5,561	5,614	5,662	5,708	5,750	5,790	5,827
240	5,300	5,366	5,426	5,480	5,530	5,577	5,621	5,661	5,699	5,735
∞	5,227	5,290	5,348	5,400	5,448	5,493	5,535	5,574	5,611	5,645

F Distribuição da estatística de Kolmogorov-Smirnov

Valores da função de distribuição:

$D \sim d_n: F(d) = P(D \leq d) = p$

n \ p	0,900	0,950	0,975	0,990	0,995
1	0,900	0,950	0,975	0,990	0,995
2	0,684	0,776	0,842	0,900	0,929
3	0,565	0,636	0,708	0,785	0,829
4	0,493	0,565	0,624	0,689	0,734
5	0,447	0,509	0,563	0,627	0,669
6	0,410	0,468	0,519	0,577	0,617
7	0,381	0,436	0,483	0,538	0,576
8	0,358	0,410	0,454	0,507	0,542
9	0,339	0,387	0,430	0,480	0,513
10	0,323	0,369	0,409	0,457	0,489
11	0,308	0,352	0,391	0,437	0,468
12	0,296	0,338	0,375	0,419	0,449
13	0,285	0,325	0,361	0,404	0,432
14	0,275	0,314	0,349	0,390	0,418
15	0,266	0,304	0,338	0,377	0,404
16	0,258	0,295	0,327	0,366	0,392
17	0,250	0,286	0,318	0,355	0,381
18	0,244	0,279	0,309	0,346	0,371
19	0,237	0,271	0,301	0,337	0,361
20	0,232	0,265	0,294	0,329	0,352
21	0,226	0,259	0,287	0,321	0,344
22	0,221	0,253	0,281	0,314	0,337
23	0,216	0,247	0,275	0,307	0,330
24	0,212	0,242	0,269	0,301	0,323
25	0,208	0,238	0,264	0,295	0,317
26	0,204	0,233	0,259	0,290	0,311
27	0,200	0,229	0,254	0,284	0,305
28	0,197	0,225	0,250	0,279	0,300
29	0,193	0,221	0,246	0,275	0,295
30	0,190	0,218	0,242	0,270	0,290
31	0,187	0,214	0,238	0,266	0,285
32	0,184	0,211	0,234	0,262	0,281
33	0,182	0,208	0,231	0,258	0,277
34	0,179	0,205	0,227	0,254	0,273
35	0,177	0,202	0,224	0,251	0,269
36	0,174	0,199	0,221	0,247	0,265
37	0,172	0,196	0,218	0,244	0,262
38	0,170	0,194	0,215	0,241	0,258
39	0,168	0,191	0,213	0,238	0,255
40	0,165	0,189	0,210	0,235	0,252
$n > 40$	$\frac{1,07}{\sqrt{n}}$	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,52}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

G Distribuição do coeficiente de correlação ordinal de Spearman

Valores da função de distribuição:

$R \sim R_n: F(r) = P(R \leq r) = p$

n \ p	**0,95**	**0,975**	**0,99**	**0,995**
5	0,900	-	-	-
6	0,829	0,886	0,943	-
7	0,714	0,786	0,893	-
8	0,643	0,738	0,833	0,881
9	0,600	0,683	0,783	0,883
10	0,564	0,648	0,745	0,794

H Distribuição da estatística W de Wilcoxon

Valores da função de distribuição:

$w \sim W_n: F(w) = P(W \leq w) = p$

n \ p	0,005	0,01	0,025	0,05	0,95	0,975	0,99	0,995
5	-	-	-	1	14	-	-	-
6	-	-	0	2	19	21	-	-
7	-	0	2	4	24	26	28	-
8	0	2	3	6	30	33	34	36
9	1	3	5	8	37	40	42	44
10	3	5	8	11	44	47	50	52
11	5	7	10	14	52	56	59	61
12	7	10	13	17	61	65	68	71
13	9	13	17	21	70	74	78	82
14	12	16	21	26	79	84	89	93
15	15	20	25	30	90	95	100	105
16	19	24	29	36	100	107	112	117
17	23	28	34	41	112	119	125	130
18	27	33	40	47	124	131	138	144
19	32	38	46	54	136	144	152	158
20	37	43	52	60	150	158	167	173
21	42	49	58	68	163	173	182	189
22	48	56	66	75	178	187	197	205
23	54	62	73	83	193	203	214	222
24	61	69	81	92	208	219	231	239
25	68	77	89	101	224	236	248	257
26	76	85	98	110	241	253	266	275
27	84	93	107	120	258	271	285	294
28	92	102	117	130	276	289	304	314
29	100	111	127	141	294	308	324	335
30	109	120	137	152	313	328	345	356

I Distribuição da estatística U de Mann-Whitney-Wilcoxon

Quantil 0,005:

$$F(u) = P\left(U \leq u_{n_1;n_2;\,0{,}005}\right) = 0{,}005$$

n_1 \ n_2	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
3	-	-	-	-	-	-	-	0	0	0	1	1	1	2	2	2	2	3	3
4	-	-			0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	-	-		0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	-	-	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	-	-	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	-	-	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	-	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	-	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	-	0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	46
12	-	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	-	1	3	7	10	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14	-	1	4	7	11	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15	-	2	5	8	12	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16	-	2	5	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17	-	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18	-	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19	0	3	7	12	17	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20	0	3	8	13	18	24	30	36	42	46	54	60	67	73	79	86	92	99	105

Quantil 0,025:

$$F(u) = P\left(U \leq u_{n_1;n_2;\,0{,}005}\right) = 0{,}005$$

n_1 \ n_2	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
2	-	-	-	-	-	-	0	0	0	0	1	1	1	1	1	2	2	2	2
3	-	-	-	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4	-	-	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5	-	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6	-	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7	-	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8	0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9	0	2	4	7	10	12	15	17	21	23	26	28	31	34	37	39	42	45	48
10	0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11	0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12	1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13	1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14	1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15	1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16	1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17	2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18	2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19	2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20	2	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127