Carlos Tenreiro

Notas de Estatística

Coimbra, 2018

Abril de 2022

Versões anteriores: Jan. 2018, Jan. 2019, Mar. 2020

Nota prévia

Estas notas, que têm servido de texto de apoio às aulas da disciplina de Bioestatística lecionada a alunos do primeiro ano, segundo semestre, da Licenciatura em Farmácia Biomédica da Faculdade de Farmácia da Universidade de Coimbra, têm como ponto de partida o texto *Estatística: notas de apoio às aulas*, Coimbra, 2009, que continua disponível no endereço http://www.mat.uc.pt/∼tenreiro/. Apesar dos assuntos aqui tratados corresponderem, no essencial, ao lecionado, as matérias completas, que incluem os exercícios de aplicação que constam das folhas práticas da cadeira, foram expostas nas aulas. Informação complementar sobre os tópicos aqui abordados podem ser obtidos através da monografia de D.S. Moore e G.P. McCabe, *Introduction to the Practice of Statistics*, editada pela W.H. Freeman and Company, bem como nos textos a que fazemos referência no final de cada um dos capítulos destas notas.

Carlos Tenreiro

Índice

Introdução

O que é a Estatística? O que vamos aprender? Para que nos serve?

0.1 O que é a Estatística?

A palavra "estatística" deriva do latim "status" que significa "estado", "situação". Vejamos o que o Webster's Dictionary diz sobre a palavra "statistics" nas suas edições de 1828 e 1996:

> 1828 ([1]): *uma coleção de factos relativos ao estado da sociedade, à condição das pessoas no país, à sua saúde, longevidade, economia doméstica, propriedade, orientação política, ao estado do país, etc.*

> 1996 ([2]): *a ciência que trata da recolha, classificação, análise e interpretação de factos ou dados numéricos, e que, pela utilização da teoria matemática da probabilidade, procura e estabelece regularidades em conjuntos mais ou menos dispersos de elementos.*

Reparemos no significado atribuído à palavra "estatística" na edição de 1828 deste dicionário, em que ela serve para designar, não uma disciplina científica ou um conjunto de técnicas utilizadas para interpretar um conjunto de dados, mas tão só um conjunto de factos ou dados relevantes para a organização dos estados. Atualmente, utilizamos o plural **estatísticas** com um significado próximo do anterior. Mais precisamente, usamo-lo para designar um conjunto de dados numéricos, agrupados e classificados, referentes aos factos em estudo, ou ainda, descrições quantitativas duma realidade ou domínio. Reparemos na evolução do significado da palavra "estatística" patente na edição de 1996, onde se faz referência não só ao papel **descritivo** da disciplina, quando

[1] http://webstersdictionary1828.com/.

[2] Webster's Dictionary, Random House, New York, 1996.

se refere a classificação, análise e interpretação de dados numéricos, mas também ao seu papel **inferencial**, quando se menciona a **teoria da probabilidade** como instrumento matemático que permite a procura de regularidades ou padrões.

Estes dois aspetos são também referidos nos dicionários seguintes:

> Petit Robert (1993) ([3]): *estudo metódico de factos sociais, através de procedimentos numéricos (classificação, descrição, inventariação, recenseamento), destinado a informar e ajudar os governos (1832); campo da matemática aplicada que utiliza o cálculo das probabilidades para formular hipóteses a partir de acontecimentos reais e fazer previsões.*

> Dicionário da Academia das Ciências de Lisboa (2001) ([4]): *Estudo metódico que tem por objeto a observação de certo número de factos sociais, de uma realidade e a respetiva ordenação, análise e interpretação dos dados numéricos obtidos. – Estatística Descritiva: a que pesquisa e reúne dados numéricos, calcula médias, índices. – Estatística Matemática: a que utiliza o cálculo das probabilidades.*

Para clarificar o papel inferencial da estatística, pensemos no que se passa nas vésperas duma eleição para a Assembleia da República em que várias **sondagens** são realizadas para prever as votações nos vários partidos no ato eleitoral que se avizinha. Contrariamente aos **censos** ou **recenseamentos** em que todos os indivíduos da população são inquiridos, na realização duma sondagem apenas uma pequena parcela da população, a que chamamos **amostra**, é inquirida. No caso das sondagens eleitorais a **população** ou **universo da sondagem** é idealmente constituída por todos os cidadãos eleitores. Os resultados obtidos na amostra são depois usados para estimar a verdadeira percentagem de votantes em cada um dos partidos. Se a recolha da amostra, isto é, se a **amostragem** for feita de forma adequada, é ainda possível quantificar a confiança que podemos ter na previsão efetuada.

O esquema seguinte resume o que acabámos de dizer. Além dos objetivos de cada uma das áreas da Estatística acima referidas, incluímos também os instrumentos utilizados em cada uma delas.

Apesar do processo de exploração dos dados, na busca de padrões e de observações que fogem a esses padrões, não coincidir necessariamente com o da sua descrição, a

[3] Le nouveau Petit Robert, Dictionnaires Le Robert, Paris, 1993.
[4] Dicionário da Academia das Ciências de Lisboa, Verbo, Lisboa, 2001.

Estatística		
↙		↘
Estatística Descritiva	modelação	Estatística Matemática
↕	⟶⟶	↕
Organização e redução dos dados de forma a realçar a informação importante neles contida		Tira conclusões sobre uma população a partir duma amostra, dando uma medida do erro que podemos cometer
↑		↑
Tabelas Gráficos Indicadores numéricos		Amostragem Planeamento de Experiências Estimação e Testes

Tabela 0.1.1: Objetivos e métodos da Estatística

Estatística Descritiva é também referida na literatura como Análise Exploratória de Dados. Estatística Indutiva e Estatística Inferencial são designações correntemente usadas em alternativa a Estatística Matemática.

De forma sucinta podemos dizer que a **Estatística** é uma área da matemática aplicada que se ocupa da recolha, tratamento e interpretação de **dados** numéricos, e que usa a teoria da probabilidade para inferir sobre a população de onde esses dados foram recolhidos.

0.2 O que vamos aprender?

Pretendendo concretizar um pouco mais as diversas etapas descritas no esquema anterior, e, simultaneamente, dar uma ideia sobre os assuntos que abordaremos, consideremos o exemplo seguinte de aplicação da Estatística à medicina.

Exemplo 0.2.1 Num estudo sobre os efeitos psico-somáticos na recuperação de jovens anoréxicas, pretende-se saber se o tratamento habitualmente usado dá melhores resultados em regime ambulatório do que em regime de internamento hospitalar (para mais detalhes sobre este exemplo ver Pestana e Velosa, 2002, pág. 34–36). Do quadro seguinte consta o peso, em quilogramas, de jovens anoréxicas, no início do tratamento

1	H	36.5	37.2	17	H	37.7	38.7	33	F	39.3	45.4
2	H	38.5	38.8	18	H	37.6	37.0	34	F	36.1	34.7
3	H	36.9	36.9	19	H	39.7	40.4	35	F	37.4	41.6
4	H	37.4	37.1	20	H	38.1	38.0	36	F	34.8	34.8
5	H	36.2	34.6	21	H	39.6	37.5	37	F	42.7	46.0
6	H	40.2	46.9	22	H	34.7	34.3	38	F	33.3	43.0
7	H	43.0	44.6	23	H	36.3	37.4	39	F	36.5	34.1
8	H	34.6	42.3	24	H	39.8	45.5	40	F	37.0	35.2
9	H	36.7	33.3	25	H	37.7	38.6	41	F	37.2	43.3
10	H	36.5	37.2	26	H	36.1	37.9	42	F	35.2	41.1
11	H	38.5	43.8	27	H	38.3	38.3	43	F	37.8	41.9
12	H	40.4	43.2	28	H	36.6	39.6	44	F	40.7	42.5
13	H	36.8	37.3	29	H	39.6	39.3	45	F	39.0	41.5
14	H	34.7	32.8	30	F	38.0	43.1	46	F	39.5	44.4
15	H	31.7	41.2	31	F	37.7	42.7				
16	H	36.4	32.3	32	F	39.0	41.4				

Tabela 0.2.2: Peso em Kg de jovens anoréxicas

e passado quatro semanas. Um grupo recebe o tratamento em internamento hospitalar (H) na companhia de um familiar e o outro recebe o tratamento residindo com a família (F). Apesar deste conjunto de dados não ser muito extenso, os dados são difíceis de ler e de interpretar mesmo para um conhecedor da anorexia. É assim importante estudarmos técnicas estatísticas para organizar, apresentar de forma clara e resumir os dados anteriores, de modo que deles sobressaia a informação mais relevante. Estamos naturalmente a falar da utilização de **tabelas**, **gráficos** e **indicadores numéricos**.

Pretendendo saber qual a eficácia de qualquer um dos tratamentos, ou ainda, se o tratamento ambulatório é, ou não, mais eficaz que o hospitalar, de modo a que o possamos, ou não, indicar a outros doentes, necessitamos de técnicas que nos permitam avaliar cada um dos tratamentos e também decidir por uma ou outra forma de tratamento, e ao mesmo tempo quantificar o erro que poderemos estar a cometer quando tomamos essa decisão. Referimo-nos desta vez aos **intervalos de confiança** e aos **testes de hipóteses**.

Um ponto fundamental de todo este procedimento de inferência, é a forma como as jovens foram escolhidas para integrar o estudo (de modo a avaliarmos o universo de jovens anoréxicas para o qual são válidos os resultados e conclusões do estudo), ou ainda, a forma como as jovens foram divididas pelos dois grupos de tratamento. Estamos neste caso a levantar a questão da **amostragem** e do **planeamento da experiência**. Todas estas questões serão por nós estudadas em capítulos futuros.

0.3 Para que nos serve a Estatística?

A Estatística é hoje uma ferramenta essencial aos profissionais das mais diversas áreas de atividade. É-o para aqueles que a usam para fundamentar ou realizar estudos nas áreas da medicina, das ciências da terra, das engenharias, da psicologia, da pedagogia, etc, mas é-o também para aqueles que, no desempenho de cargos administrativos, precisam de interpretar, preferivelmente de forma crítica, informação estatística quer esta se apresente de forma gráfica ou não gráfica.

Exemplo 0.3.1 Para ilustrar a necessidade de conhecimentos na área da Estatística, por mais elementares que sejam, dum qualquer cidadão na interpretação duma simples **sondagem de opinião**, fica o exemplo da sondagem eleitoral realizada pela INTERCAMPUS para a TVI e PÚBLICO, com o objetivo de conhecer a opinião dos portugueses sobre diversos temas da política nacional, incluindo a intenção de voto para as eleições para o Presidente da República de 2016 ([5]). As intenções de votos obtidas para os candidatos colocados nas cinco primeiras posições da sondagem são apresentadas nos quadro e figura seguintes:

Candidatos	Intenção de voto
Marcelo Rebelo de Sousa	51,8%
Sampaio da Nóvoa	16,9%
Maria de Belém	10,1%
Marisa Matias	7,9%
Edgar Silva	4,6%

Ficha técnica:

Universo – População portuguesa, com 18 e mais anos de idade, eleitoralmente recenseada, residente em Portugal Continental.

Amostragem – Os lares foram selecionados aleatoriamente a partir de uma estratificação segundo a Região (7 regiões) e Habitat/Dimensão dos agregados populacionais (6 grupos); em cada lar os respondentes foram selecionados através do método de quotas, com base numa matriz que cruzou as variáveis Sexo e Idade (3 grupos).

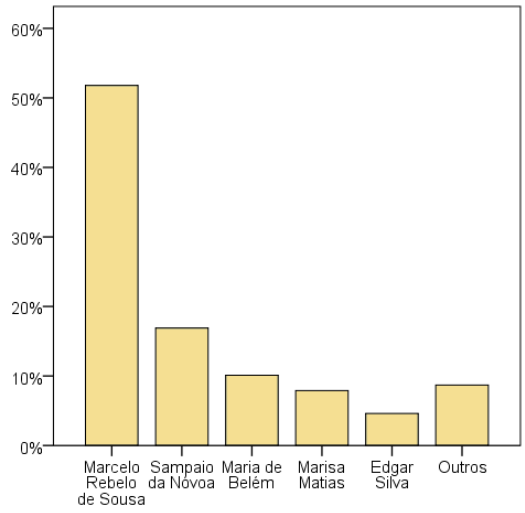
Amostra – Constituída por 1043 indivíduos.

Recolha da Informação – A informação foi recolhida através de entrevista directa e pessoal, com base em questionário estruturado e elaborado pela INTERCAMPUS, utilizando a técnica de simulação de voto em urna. Os trabalhos de campo decorreram entre 14 e 20 de Janeiro de 2016.

[5] https://www.publico.pt/politica/noticia/marcelo-proximo-da-vitoria-a-primeira-volta-1721017

Margem de Erro – O erro máximo de amostragem deste estudo, para um intervalo de confiança de 95%, é de $\pm 3,0\%$.

Taxa de Resposta – A taxa de resposta obtida neste estudo foi de: $60,9\%$.



Uma sondagem é, como veremos, um caso particular duma classe mais vasta de problemas a que em Estatística se dá o nome de estimação por **intervalos de confiança**.

O gráfico de barras e o quadro são de interpretação simples dando-nos, de forma gráfica e não gráfica, respetivamente, as intenções de voto previstas para cada um dos candidatos indicados. As questões principais prendem-se com a compreensão da ficha técnica. Nesta identifica-se o universo da sondagem ou população, ou seja, o conjunto total de indivíduos para os quais os resultados da sondagem são aplicáveis. Neste caso o universo da sondagem não é constituído por todos os cidadãos eleitores, uma vez que nem os eleitores dos Açores e da Madeira, nem os cidadãos portugueses emigrados, foram incluídos no estudo. Sobre a amostra é dito que foram inquiridos 1043 indivíduos e que o processo usado para a selecionar utiliza dois métodos, um aleatório e outro não aleatório, a que faremos referência mais à frente: o método de amostragem estratificada e o método de amostragem por quotas. Centrando a nossa atenção no candidato Marcelo Rebelo de Sousa, ficamos também a saber que o intervalo de confiança para a verdadeira intenção de voto neste candidato é $[51{,}8 - 3{,}0; 51{,}8 + 3{,}0] = [48{,}8; 54{,}8]$ sendo de 95% o seu nível de confiança. Isto quer dizer que se se recolhessem várias amostras, cada uma delas com 1034 indivíduos, pelo método de amostragem referido, poderíamos construir outros tantos intervalos do tipo anterior, diferentes de amostra para amostra, 95% dos quais conteriam a verdadeira intenção de voto no candidato Marcelo Rebelo de Sousa. Conclusões análogas poderiam ser tiradas para os outros candidatos. Reparemos que esta quantificação da confiança nas

previsões da sondagem, tem a ver, não com as previsões particulares apresentadas, pois estas podem estar, ou não, corretas, mas com o que se passaria se a sondagem fosse repetida um grande número de vezes. Por outras palavras, a quantificação da confiança nos resultados duma sondagem tem a ver com o método utilizado para produzir as previsões. Voltaremos mais tarde a todas estas questões.

Por curiosidade, apresentamos na tabela seguinte intervalos de confiança (IC) para as intenções de voto que decorrem da sondagem anterior realizada entre 14 e 20 de Janeiro de 2016 e os resultados nacionais obtidos na eleição para o Presidente da República realizada em 24 de Janeiro de 2016 ([6]):

Candidatos	IC para a intenção de voto	Votação
Marcelo Rebelo de Sousa	[48,8; 54,8]	52,00%
Sampaio da Nóvoa	[13,8; 19,9]	22,88%
Maria de Belém	[7,1; 13,1]	4,24%
Marisa Matias	[4,9; 10,9]	10,12%
Edgar Silva	[1,6; 7,6]	3,94%

0.4 O *software* estatístico

O programa estatístico IBM SPSS Statistics que utilizamos nestes apontamentos encontra-se licenciado para todos os alunos da UC. Em alternativa a este *software*, poderão os alunos utilizar o *software* estatístico livre R e a livraria R-Commander que podem ser obtidos a partir do endereço https://www.r-project.org/

0.5 Bibliografia

Martins, M.E.G., Cerveira, A.G. (2000). *Introdução às Probabilidades e à Estatística*, Universidade Aberta.

Pestana, D.D., Velosa, S.F. (2002). *Introdução à Probabilidade e à Estatística*, Vol. I, Fundação Calouste Gulbenkian.

Vicente, P., Reis, E., Ferrão, F. (2001). *Sondagens: a amostragem como factor decisivo de qualidade*, Edições Sílabo.

[6] Fonte: Comissão Nacional de Eleições (http://www.cne.pt/).

1

Análise exploratória de dados: distribuição duma variável

Indivíduos e variáveis. Variáveis qualitativas e variáveis quantitativas. Distribuição duma variável. Frequências absolutas, relativas e percentuais. Tabela de frequências. Gráfico de barras. Gráfico circular. Gráfico de caule-e-folhas. Histograma. Distribuições simétricas e assimétricas, unimodais e bimodais. Média e mediana. Variância e desvio-padrão. Mínimo, máximo e amplitude. Quartis e amplitude interquartil. Observações discordantes. Gráfico de extremos-e-quartis. Alteração da unidade de medida.

1.1 Indivíduos e variáveis

A informação contida na Tabela 1.1.1 diz respeito a 30 cidadãos nacionais que responderam a um questionário ([1]). Qualquer conjunto de dados como este, contém informação acerca dum grupo de **indivíduos**, informação essa que está organizada em **variáveis**.

No caso particular da Tabela 1.1.1, temos informação sobre 5 variáveis (residência, idade, estado civil, número de filhos, sexo), observadas em 30 indivíduos. Por **indivíduo** queremos designar qualquer objeto descrito por um conjunto de dados. Os indivíduos podem ser pessoas, animais, ou coisas. As **variáveis** são caraterísticas que observamos nos diversos indivíduos, variando os seus valores de indivíduo para indivíduo.

Sendo a informação contida na Tabela 1.1.1 relativa a uma parte dos cidadãos nacionais que responderam ao questionário entregue, dizemos que tal informação é relativa a uma **amostra** desse conjunto mais vasto de cidadãos. Ao número de indivíduos da amostra, chamamos **dimensão da amostra**. No caso presente, temos uma amostra de dimensão 30.

[1] Dados adaptados de Ferreira, I., Gonçalves, V.P., *Métodos Quantitativos*, Texto Editora, 2006.

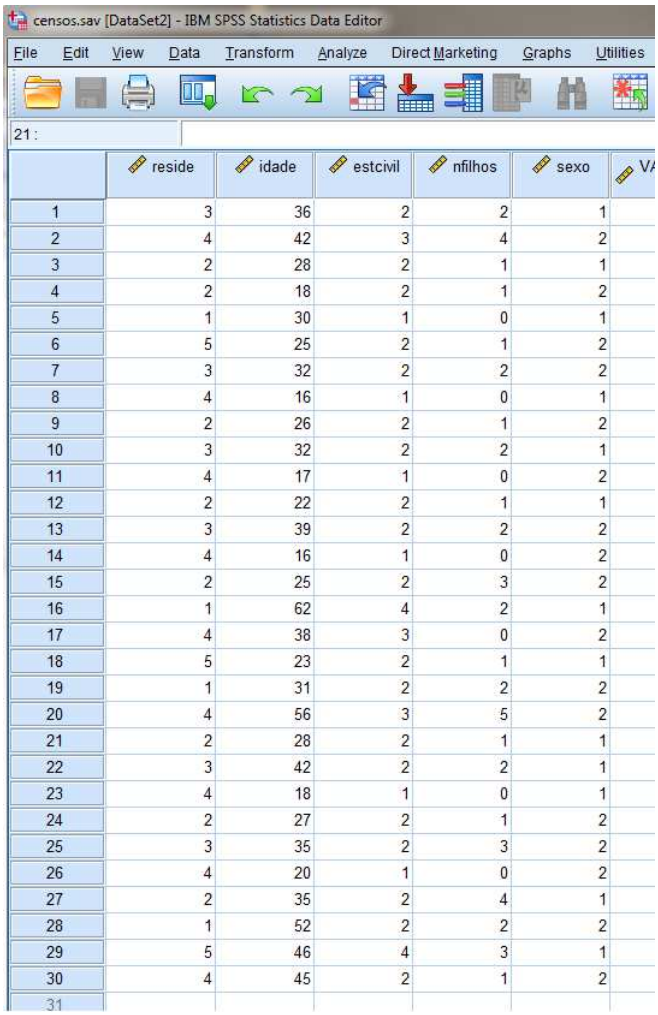
Residência	Idade	Estado civil	Nº de filhos	Sexo
Coimbra	36	casado	2	M
Lisboa	42	divorciado	4	F
Porto	28	casado	1	M
Porto	18	casado	1	F
Braga	30	solteiro	0	M
Faro	25	casado	1	F
Coimbra	32	casado	2	F
Lisboa	16	solteiro	0	M
Porto	26	casado	1	F
Coimbra	32	casado	2	M
Lisboa	17	solteiro	0	F
Porto	22	casado	1	M
Coimbra	39	casado	2	F
Lisboa	16	solteiro	0	F
Porto	25	casado	3	F
Braga	62	viúvo	2	M
Lisboa	38	divorciado	0	F
Faro	23	casado	1	M
Braga	31	casado	2	F
Lisboa	56	divorciado	5	F
Porto	28	casado	1	M
Coimbra	42	casado	2	M
Lisboa	18	solteiro	0	M
Porto	27	casado	1	F
Coimbra	35	casado	3	F
Lisboa	20	solteiro	0	F
Porto	35	casado	4	M
Braga	52	casado	2	F
Faro	46	viúvo	3	M
Lisboa	45	casado	1	F

Tabela 1.1.1: Dados relativos a 30 cidadãos nacionais

Algumas variáveis como "residência", "sexo" ou "estado civil", apenas distribuem os indivíduos em categorias de acordo com qualidades desses mesmos indivíduos. Tais variáveis dizem-se por isso **qualitativas** ou **categóricas**. Os valores ou modalidades assumidos por uma variável qualitativa são assim identificadores de qualidades, modalidades ou atributos do indivíduo observado. No caso da variável "sexo", em vez das letras M e F para designar masculino e feminino, poderíamos utilizar números como 1 e 2, desde que indiquemos as modalidades representadas por cada número. Estes números expressam apenas um atributo do indivíduo observado, não fazendo sentido realizar operações numéricas sobre tais números como, por exemplo, o cálculo duma média. Apesar de neste caso a utilização das letras M e F ser mais sugestiva, casos há em que é mais fácil utilizar números como identificadores dos valores assumidos por

uma variável qualitativa.

Outras variáveis como "idade" ou "número de filhos", tomam valores numéricos com os quais faz sentido realizar operações aritméticas. Fará, por exemplo, sentido calcular a idade média dos indivíduos observados. A estas variáveis chamamos **variáveis quantitativas**.



	reside	idade	estcivil	nfilhos	sexo	VA
1	3	36	2	2	1	
2	4	42	3	4	2	
3	2	28	2	1	1	
4	2	18	2	1	2	
5	1	30	1	0	1	
6	5	25	2	1	2	
7	3	32	2	2	2	
8	4	16	1	0	1	
9	2	26	2	1	2	
10	3	32	2	2	1	
11	4	17	1	0	2	
12	2	22	2	1	1	
13	3	39	2	2	2	
14	4	16	1	0	2	
15	2	25	2	3	2	
16	1	62	4	2	1	
17	4	38	3	0	2	
18	5	23	2	1	1	
19	1	31	2	2	2	
20	4	56	3	5	2	
21	2	28	2	1	1	
22	3	42	2	2	1	
23	4	18	1	0	1	
24	2	27	2	1	2	
25	3	35	2	3	2	
26	4	20	1	0	2	
27	2	35	2	4	1	
28	1	52	2	2	2	
29	5	46	4	3	1	
30	4	45	2	1	2	
31						

A figura anterior mostra o aspeto do ficheiro SPSS *censos.sav* que comporta a informação incluída no quadro da Tabela 1.1.1. Reparemos que não só na variável "sexo" foram usadas etiquetas para representar as suas modalidades. Tal acontece também com as variáveis "residência" e "estado civil". No caso da variável "residência", usámos as etiquetas 1, 2, 3, 4, e 5, para representar as cidades "Braga", "Porto", "Coimbra", "Lisboa" e "Faro", respetivamente.

1.2 Representação gráfica duma distribuição

O **padrão de variação** duma variável, a que chamaremos **distribuição da variável**, é uma informação importante sobre essa variável. A **distribuição duma variável** dá-nos conta dos valores que a variável toma, bem como a frequência com que os toma. Os métodos de representação de dados que vamos estudar nos parágrafos seguintes, permitir-nos-ão descrever a distribuição da variável em estudo, pondo em evidência as suas principais caraterísticas.

1.2.1 Gráficos para variáveis qualitativas

Os valores que uma variável qualitativa toma são etiquetas ou rótulos para as modalidades ou categorias respeitantes a essa variável. Um modo de resumir os dados observados para uma variável qualitativa é contar o número de vezes que ocorre cada um dos valores assumidos pela variável. Esse número é dito **efetivo**, **frequência absoluta** ou, simplesmente, **frequência** desse valor.

Exemplo 1.2.1 Centrando a nossa atenção na variável "residência" da Tabela 1.1.1, apresentamos na tabela seguinte o resultado de tais contagens. Além da frequência de cada uma das modalidades que a variável "residência" assume, a tabela apresenta também as chamadas **frequência relativa** e **frequência percentual** . É por isso dita **tabela de frequências**.

residência

	Frequency	Relative Frequency	Percentage Frequency
Braga	4	,133	13,3
Porto	8	,267	26,7
Coimbra	6	,200	20,0
Lisboa	9	,300	30,0
Faro	3	,100	10,0
Total	30	1,000	100,00

Vejamos como, em geral, efetuamos o cálculo das frequências relativa e percentual. Comecemos pela **frequência relativa** que se obtém dividindo a frequência (absoluta) pelo número de observações:

$$\text{frequência relativa} = \frac{\text{frequência}}{\text{número de observações}}.$$

A frequência relativa é por isso um número maior ou igual que 0 e menor ou igual que 1. A **frequência percentual**, exprime-se em percentagem, e não é mais do que a frequência relativa multiplicada por 100:

$$\text{frequência percentual} = \text{frequência relativa} \times 100\ \%$$

A informação contida numa tabela de frequência pode ser apresentada graficamente através dum **gráfico de barras**.

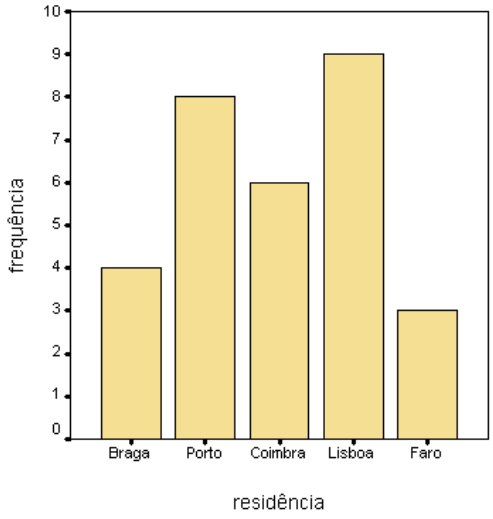
Construção dum gráfico de barras:

⊙ marcar no eixo dos xx dum sistema de eixos coordenados os valores ou modalidades assumidos pela variável em estudo;

⊙ colocar por cima desses valores barras verticais de altura igual à sua frequência, à sua frequência relativa ou à sua frequência percentual.

Notemos que num gráfico de barras a largura das barras não tem qualquer significado, apenas a altura o tem. Por isso, todas as barras devem ter a mesma largura.

Exemplo 1.2.1 (cont.) A informação contida na tabela de frequências da variável "residência"dá origem ao gráfico de barras de frequências absolutas seguinte:



Uma representação alternativa muito corrente é a representação em **gráfico circular**. Esta representação tem por base o círculo.

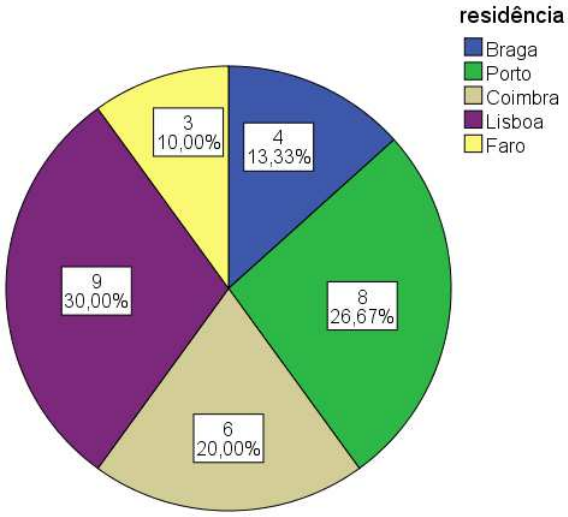
Construção dum gráfico circular:

⊙ dividir o círculo em tantos setores quantos os valores ou modalidades que a variável toma;

⊙ os ângulos desses setores são obtidos multiplicando a frequência relativa respetiva por 360 graus:

$$\text{ângulo dum setor} = \text{frequência relativa} \times 360^o$$

Exemplo 1.2.1 (cont.) Para a variável "residência" obtemos o gráfico circular de frequências absolutas e percentuais seguinte:



Os ângulos de cada um dos setores marcados no gráfico anterior são apresentados na tabela seguinte. Na primeira coluna o ângulo é calculado a partir do valor aproximado da frequência relativa que na tabela de frequência foram aproximados às milésimas. Na segunda coluna o mesmo cálculo é feito utilizando o valor exato da frequência relativa.

	ângulo (aproxi.)	ângulo (exato)
Braga	47,88	48,00
Porto	96,12	96,00
Coimbra	72,00	72,00
Lisboa	108,00	108,00
Faro	36,00	36,00
Total	360	360

A discrepância encontrada para os valores dos dois primeiros ângulos é devida aos erros de arredondamento presentes na frequência relativa. Sempre que efetuamos cálculos

utilizando uma calculadora, é preferível trabalhar com os valores exatos ou guardar na memória da máquina os resultados de cálculos anteriores. Em particular, se esses resultados são números com muitas casas decimais, estamos, ao proceder assim, a diminuir possíveis erros de arredondamento quando usamos tais resultados em novos cálculos.

1.2.2 Gráficos para variáveis quantitativas

Os gráficos anteriores permitem uma perceção rápida sobre a distribuição da variável em estudo. No entanto, eles não são essenciais para compreendermos a distribuição duma variável qualitativa uma vez que esta é normalmente fácil de apreender a partir exclusivamente da tabela de frequências. A importância da representação gráfica na descrição da distribuição duma variável será mais relevante no caso das variáveis quantitativas.

Uma representação gráfica muito utilizada para descrever a distribuição duma variável quantitativa, especialmente quando o número de observações é pequeno, é o diagrama ou **gráfico de caule-e-folhas**. Vejamos como construir um gráfico de caule-e-folhas.

Construção dum gráfico de caule-e-folhas:

⊙ separar cada observação num **caule**, formado pelos algarismos dominantes do número, e numa **folha**, formada pelos restantes algarismos;

⊙ colocar os caules numa coluna por ordem crescente de cima para baixo, e desenhar uma linha vertical à direita dessa coluna de números;

⊙ colocar à direita de cada caule as respetivas folhas, por ordem crescente da esquerda para a direita.

Exemplo 1.2.2 Consideremos o seguinte conjunto de dados relativo ao peso em gramas de 42 ratos diabéticos ([2]):

40, 46, 45, 46, 43, 47, 52, 39, 45, 42, 42, 44, 40, 41, 51, 42, 41, 38, 45, 48, 39
49, 38, 38, 42, 48, 49, 40, 38, 46, 42, 38, 51, 48, 44, 48, 40, 44, 38, 41, 45, 52

A variável em estudo é o "peso" e os indivíduos são os ratos observados. Seguindo o procedimento acima descrito, façamos a representação dos dados anteriores através

[2]Fonte: Pestana e Velosa, 2002, p. 115.

dum gráfico de caule-e-folhas. Neste caso a separação das observações em caule e folha é simples. O caule é o algarismo das dezenas, enquanto que a folha é o algarismo das unidades. Os passos atrás descritos dão origem às etapas seguintes:

1)

3	|
4	|
5	|

2)

3	|	88888899
4	|	00001112222234445555666788899
5	|	1122

Reparemos que cada caule tem aqui uma amplitude de 10 unidades. Isto quer dizer que o número representado pelo caule 3 é $3 \times 10 = 30$. Dizemos que a **amplitude do caule** é 10. Além disso, cada folha representa uma só observação.

O gráfico anterior dá uma pobre ideia da distribuição da variável na parte central do mesmo. Neste caso é habitual separar cada caule em semicaules. No caso do exemplo anterior, isto corresponderia a considerar os semicaules 3, 3, 4, 4, 5 e 5, e a associar ao primeiro semicaule as folhas 0, 1, 2, 3 e 4, e ao segundo semicaule as folhas 5, 6, 7, 8 e 9. Eis o gráfico de caule-e-folhas resultante:

3	|	88888899
4	|	0000111222223444
4	|	55556667888899
5	|	1122

Por vezes justifica-se ainda dividir cada caule em 5 subcaules. Ao primeiro subcaule associávamos as folhas 0 e 1, ao segundo as folhas 2 e 3, ao terceiro as folhas 4 e 5, ao quarto as folhas 6 e 7, e, finalmente, ao quinto subcaule associávamos as folhas 8 e 9.

```
peso Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     8,00        3 .  88888899
    16,00        4 .  0000111222223444
    14,00        4 .  55556667888899
     4,00        5 .  1122

Stem width:        10
Each leaf:        1 case(s)
```

Quando esta tarefa é executada por um *software* estatístico, a separação das observações em caule e folhas é feita de modo automático. No caso do SPSS o gráfico produzido é o segundo dos gráficos anteriores. O SPSS inclui no gráfico a amplitude de cada caule, a informação de que cada folha corresponde a uma observação, e também a frequência de cada caule.

Casos há em que não é óbvia a separação das observações em caule e folhas. O exemplo seguinte ilustra este facto.

Exemplo 1.2.3 Para testar uma nova farinha para pintos, de um grupo de 40 pintos com um dia selecionaram-se 20 aos quais foi administrada a nova farinha — grupo experimental —, tendo aos restantes sido dada a ração habitual — grupo de controlo. Passadas três semanas os pintos foram pesados tendo-se obtido os seguintes ganhos no peso (em gramas) ([3]):

Grupo de controlo				Grupo experimental			
383	325	360	351	362	443	404	376
285	343	405	468	438	407	392	424
352	414	326	392	409	313	464	406
356	386	313	279	421	423	475	398
348	452	363	432	434	336	417	322

Pretendendo-se representar a distribuição dos pesos dos pintos do grupo de controlo por um gráfico de caule-e-folhas, surgem duas possibilidades para separar as observações em caule e folhas. Tomando a observação 383 para exemplificar, podemos optar por considerar 3 o caule e 83 a folha, ou, em alternativa, considerar 38 o caule e 3 a folha. A segunda opção é desapropriada uma vez que levaria a um gráfico com demasiados caules e poucas folhas por caule. Tomando então a primeira opção, somos conduzidos ao gráfico seguinte em que cada caule tem uma amplitude de 100:

2 | 79 85
3 | 13 25 26 43 48 51 52 56 60 63 83 86 92
4 | 05 14 32 52 68

Para facilitar a leitura representamos cada folha por um só algarismo o que neste caso corresponde a desprezarmos o algarismo das unidades. Obtemos então o gráfico de caule-e-folhas simplificado:

2 | 78
3 | 1224455566889
4 | 01356

[3]fonte: Martins e Cerveira, 2000, p. 67.

Reparemos que, contrariamente aos gráficos anteriores, neste gráfico simplificado não são registadas as verdadeiras observações uma vez que estas aparecem truncadas. Tal como no Exemplo 1.2.2 podemos ainda dividir cada caule em semicaules:

2	|	78
3	|	12244
3	|	55566889
4	|	013
4	|	56

A amplitude do caule é neste caso de 100.

Para cada um dos grupos de controlo e experimental apresentamos a seguir os gráficos de caule-e-folhas produzidos pelo SPSS:

Stem-and-Leaf Plot for
aumento de peso grupo= controlo

Frequency	Stem &	Leaf
2,00	2 .	78
5,00	3 .	12244
8,00	3 .	55566889
3,00	4 .	013
2,00	4 .	56

Stem width: 100
Each leaf: 1 case(s)

Stem-and-Leaf Plot for
aumento de peso grupo= experimental

Frequency	Stem &	Leaf
1,00 Extremes		(=<313)
2,00	3 .	23
4,00	3 .	6799
11,00	4 .	00001222334
2,00	4 .	67

Stem width: 100
Each leaf: 1 case(s)

Reparemos que a observação 313 do grupo experimental é marcada de forma especial sendo rotulada de "extrema". Como teremos oportunidade de estudar um pouco mais à frente, isto quer dizer que este valor é suspeito de não seguir o padrão revelado pelas restantes observações. Poder-se-á, por exemplo, tratar dum erro de observação, dum valor incorretamente registado, ou dum valor incorretamente incluído no conjunto de dados. Diremos por isso que se trata duma **observação discordante**. Devido à influência que tais observações podem ter, por si só, no resultado de diversas metodologias estatísticas, este tipo de observações exige uma análise especial. Em particular, estes valores devem ser confirmados ou corrigidos antes de continuarmos o estudo. No caso de ser um valor incorretamente incluído no conjunto de dados, ele deve ser excluído.

Uma das aplicações mais interessantes dos gráficos de caule-e-folhas, é a possibilidade de comparar dois conjuntos de observações conjugando os gráficos de caule-e-folhas respetivos. O gráfico seguinte permite uma comparação simples dos grupos de controlo e experimental, revelando evidências de que para os pintos considerados a nova farinha é preferível à antiga. Para que esta comparação seja válida é importante que o número

de observações em cada um dos grupos seja aproximadamente o mesmo. O SPSS não executa este tipo gráfico.

Grupo de controlo		Grupo experimental
87	2	
44221	3	1
98866555	3	6799
310	4	00001222334
65	4	67

Gráficos de caule-e-folhas paralelos

Por razões que decorrem da construção dum gráfico de caule-e-folhas, em particular pelo facto de todas as observações estarem nele representadas, este tipo de gráfico revela-se desapropriado para grandes conjuntos de dados a não ser que se disponha de um computador para executar esta tarefa. Neste caso, quando o número de observações é elevado o gráfico é habitualmente construído associando a uma folha várias observações.

Exemplo 1.2.4 O gráfico de caule-e-folhas seguinte é relativo à distribuição dos pesos (em gramas) de 1130 pacotes de açúcar empacotados por uma máquina. Como podemos verificar cada folha corresponde a (aproximadamente) 3 observações.

```
Frequency    Stem &  Leaf

    4,00 Extremes    (=<973)
    6,00       97 .  67
   10,00       97 .  889
   17,00       98 .  000111
   17,00       98 .  222233
   27,00       98 .  444445555
   32,00       98 .  66666777777
   46,00       98 .  8888889999999999
   61,00       99 .  00000000111111111111
   64,00       99 .  222222222233333333333
   87,00       99 .  44444444444444455555555555555
   94,00       99 .  66666666666667777777777777777
   93,00       99 .  88888888888888899999999999999
   75,00      100 .  000000000011111111111111
   99,00      100 .  22222222222222222222333333333333
   81,00      100 .  44444444444444455555555555
   80,00      100 .  66666666666666667777777777
   62,00      100 .  888888888899999999
   51,00      101 .  00000000011111111
   42,00      101 .  22222233333333
   26,00      101 .  444445555
   17,00      101 .  666777
   18,00      101 .  888899
    9,00      102 .  011
   10,00      102 .  2233
    2,00 Extremes    (>=1026)

Stem width:     10,00
Each leaf:       3 case(s)
```

Quando o número de observações é elevado é habitual utilizar uma outra repre-

sentação gráfica a que chamamos **histograma de frequências** ou simplesmente **histograma**.

Construção dum histograma de frequências:

⊙ dividir as observações em **classes** justapostas de igual amplitude e calcular o efetivo de cada classe;

⊙ marcar as classes no eixo dos xx dum sistema de eixos coordenados;

⊙ por cima de cada classe colocar uma barra que cubra toda a classe e cuja altura é igual ou proporcional à frequência (à frequência relativa ou à frequência percentual) da classe.

Apesar das classes poderem, em geral, ter amplitudes ou tamanhos diferentes, vamos, por simplicidade, considerar sempre classes com iguais amplitudes. Um histograma é assim um gráfico idêntico ao gráfico de barras mas em que as barras surgem justapostas, sem qualquer espaço entre elas a não ser que uma das classes consideradas não tenha qualquer efetivo.

Exemplo 1.2.2 (cont.) Retomemos os dados relativos ao peso dos ratos diabéticos e façamos a sua representação através dum histograma. Tomando como referência o gráfico de caule-e-folhas executado pelo SPSS para este mesmo conjunto de dados em que foram usados 4 caules, comecemos por dividir os dados em 4 classes. Como as observações variam entre 38 e 52 gramas, vamos considerar as seguintes classes de amplitude 4 gramas:

$$]37, 41[,\ [41, 45[,\ [45, 49[,\ [49, 53[.$$

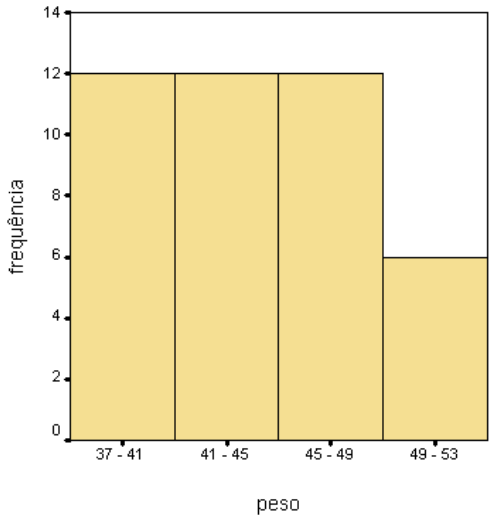
As frequências de cada uma das classes são apresentadas na tabelas de frequências seguinte:

classes	frequência	percentagem
$[37, 41[$	12	28,6
$[41, 45[$	12	28,6
$[45, 49[$	12	28,6
$[49, 53]$	6	14,3
Total	42	100,1

A soma das frequências percentuais (indicadas na tabela anterior por percentagens por simplicidade de linguagem) de todas as classes deveria ser igual a 100%.

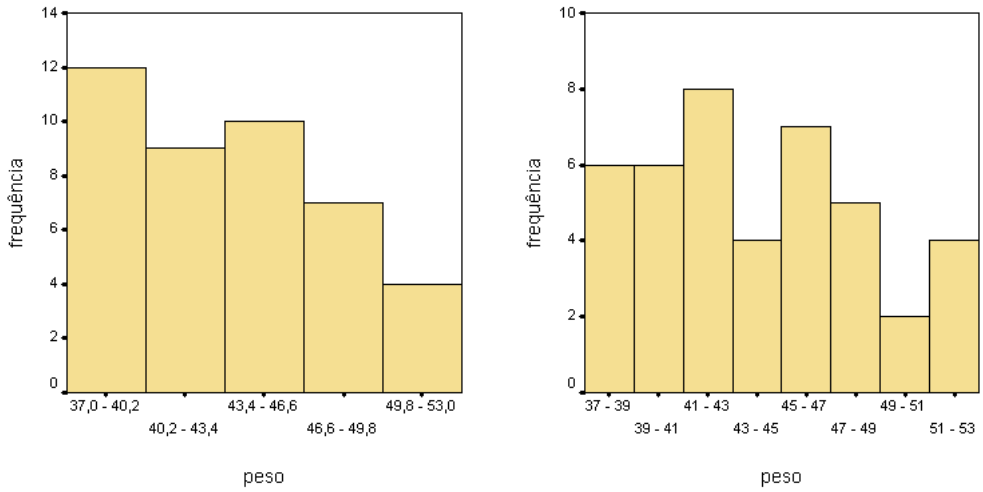
Tal não acontece devido a erros de arredondamento, uma vez que cada uma das percentagens associadas a cada classe, estando arredondada às décimas, introduz um erro na soma. Neste caso particular, 28,6 e 14,3 são aproximações por excesso de 12/46 e 6/42, respetivamente. Casos há, em que erros de arredondamento por defeito e por excesso se compensam permitindo obter uma soma de 100. Por exemplo, um arredondamento às centésimas das percentagens de cada classe dá origem a: 28,57 + 28,57 + 28,57 + 14,29 = 100. Reparemos que 28,57 é uma aproximação por defeito de 12/42, enquanto que 14,29 é uma aproximação por excesso de 6/42.

O histograma produzido pelo SPSS para as classes anteriores tem o aspeto seguinte:



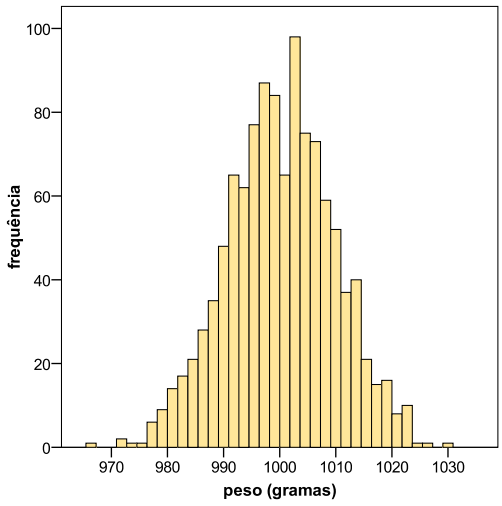
Tal como no gráfico de caule-e-folhas, em que não há uma regra ideal para calcular o número de caules ou semicaules a considerar, também para o histograma não há nenhuma regra universalmente aceite sobre o número de classes em que devemos dividir as observações. Refira-se no entanto que um número demasiado elevado de classes conduz a um histograma muito irregular com poucas observações em cada classe, enquanto que um número demasiado pequeno de classes conduz a um histograma demasiado suave com muitas observações em cada classe.

Os gráficos seguintes são histogramas obtidos por divisão das observações em 5 e em 8 classes, respetivamente. O gráfico com 8 classes é o que é feito de forma automática pelo SPSS. Apesar destes histogramas descreverem o mesmo conjunto de dados, fica claro que o aspeto do histograma é bastante influenciado pela escolha do número de classes a considerar. Tal influência é maior quando o número de observações é pequeno. Este é o caso do exemplo presente.



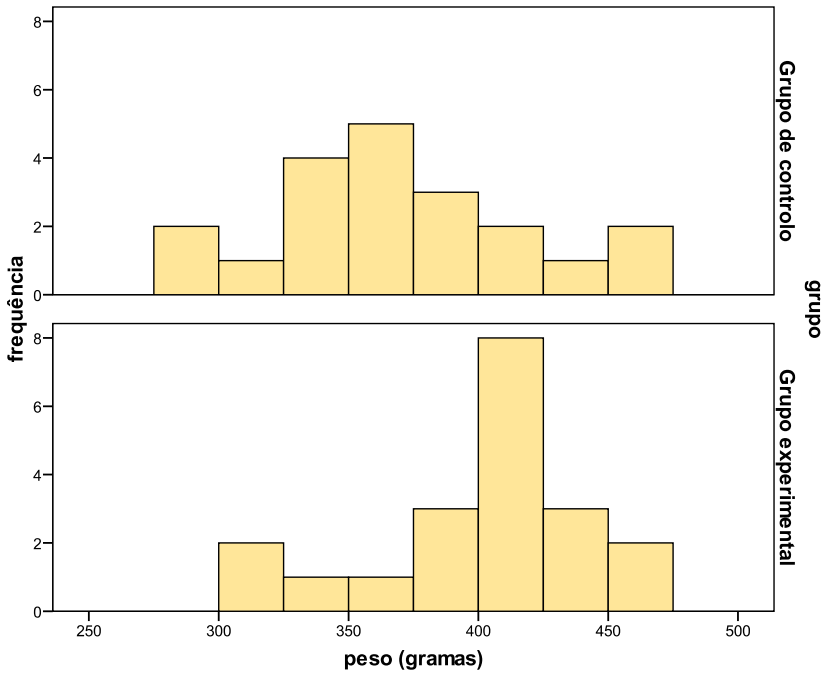
De uma forma geral, quando o número de observações é pequeno, a representação em gráfico de caule-e-folhas revela-se mais informativa do que a representação em histograma. Quanto mais não seja pelo facto de que num gráfico de caule-e-folhas o valor das observações é incluído no gráfico. Quando comparamos qualquer dos histogramas anteriores com o gráfico de caule-e-folhas construído no Exemplo 1.2.2, constatamos que a observação anterior é particularmente adequada a este exemplo.

Exemplo 1.2.4 (cont.) O histograma seguinte é construído a partir do mesmo conjunto de observações que o gráfico de caule-e-folhas atrás considerado. Atendendo ao grande número de observações envolvido a informação dada pelos dois gráficos é muito semelhantes. Neste caso é mais habitual optar pelo histograma para representar graficamente a distribuição dos dados.



Tal como nos gráficos de caule-e-folhas, podemos usar histogramas para comparar duas distribuições de dados. Para ser mais fácil e fiável a comparação dos gráficos respetivos, devemos considerar em ambos intervalos de variação com igual amplitude, quer no eixo dos xx, quer no eixo dos yy e ambos os grupos devem ter dimensões semelhantes.

Exemplo 1.2.3 (cont.) Os histogramas paralelos seguintes permitem uma análise comparativa das distribuições dos grupos de controlo e experimental em tudo semelhante à efetuada a partir dos gráficos de caule-e-folhas paralelos.



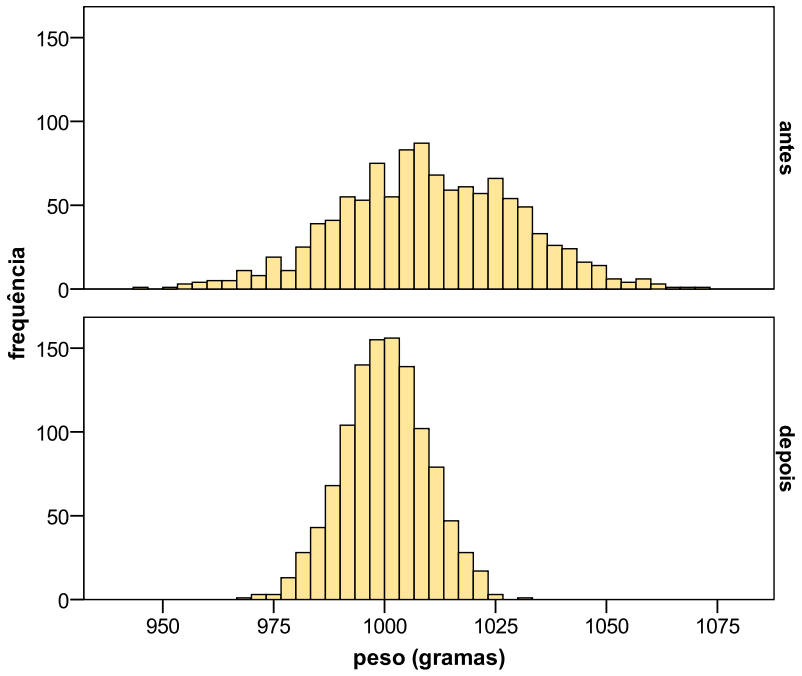
1.2.3 Caraterísticas gráficas mais relevantes

A representação gráfica duma variável não é um fim em si mesma. Ela tem como objetivo primeiro a compreensão da distribuição dessa variável, ou seja, a compreensão dos dados. Algumas caraterísticas importantes dessa distribuição são, por exemplo, a **forma**, o **centro**, a **dispersão** ou **variabilidade** e a presença de **observações discordantes**.

Relativamente ao **centro** e à **dispersão** da distribuição, veremos na próxima secção como caraterizá-los numericamente. Por agora, fiquemos com a ideia que o **centro da distribuição** pode ser descrito por um ponto abaixo do qual estão metade das observações e acima do qual está a outra metade. A **dispersão** ou **variabilidade**

da distribuição pode ser descrita pela distância entre a mais pequena e a maior das observações.

Exemplo 1.2.5 Para ilustrar graficamente estes dois conceitos, consideremos os histogramas paralelos seguintes relativos à distribuição dos pesos (em gramas) de pacotes de açúcar empacotados por uma máquina antes e depois de ter sido calibrada (em cada uma das situações foram recolhidas amostras de dimensão 1130).

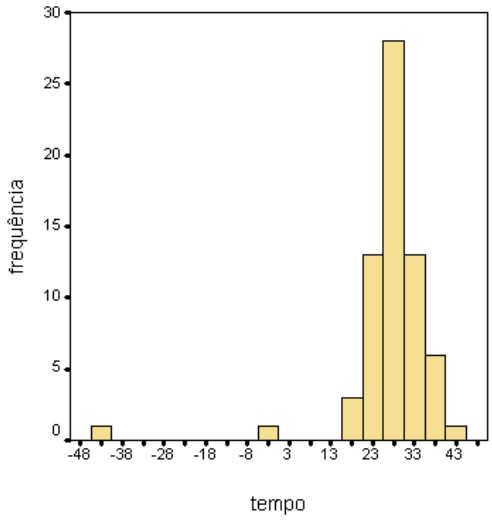


Estes gráficos revelam que o centro da distribuição do peso dos pacotes de açúcar antes da calibragem da máquina estava acima dos 1000 gramas (peso indicado no rótulo dos pacotes). Dizemos, por isso, que há um **enviesamento**, neste caso por excesso, relativamente ao peso de referência. Este enviesamento parece ter sido corrigido com a calibragem da máquina. Por outro lado, é claro também que o segundo gráfico revela uma menor dispersão dos pesos dos pacotes produzidos depois da calibragem, indicando uma maior precisão da máquina.

Como já referimos, além do centro e da dispersão duma distribuição, outra característica importante duma distribuição que pode ser analisada a partir dum gráfico de caule-e-folhas ou dum histograma, é a identificação de **observações discordantes**, isto é, observações que, por serem demasiado grandes ou pequenas, não seguem o padrão revelado pelas restantes observações. Na próxima secção daremos uma regra numérica que nos permite identificar observações discordantes. Por agora, no que respeita à sua

deteção gráfica, é relevante o facto destas observações serem caraterizadas por serem excecionalmente grandes ou pequenas relativamente às restantes observações.

Exemplo 1.2.6 O conjunto de dados que consideramos para ilustrar a presença de observações suspeitas de serem discordantes, é relativo a 66 medições feitas por Newcomb em 1882 para estimar a velocidade da luz ([4]). Mais precisamente, Newcomb mediu o tempo, expresso numa apropriada unidade de medida, que a luz levou a percorrer 7400 metros. No histograma seguinte, que resume as observações feitas, sobressaem as duas observações mais à esquerda que podemos considerar tratar-se de observações discordantes.



Pretendendo apresentar uma aproximação para a velocidade da luz, surge de forma natural a ideia de considerar a média das observações realizadas. A média das 66 observações é 26,21. Newcomb decidiu considerar a mais pequena das observações como discordante, não a tendo incluído no cálculo da média. A média das restantes 65 observações é 27,29. É clara a **influência** que, por si só, esta observação tem no cálculo da média. Este facto foi talvez a principal razão para que ela tenha sido excluída.

Finalmente, falemos da **forma da distribuição** que não é mais do que a forma ou padrão revelados pelo histograma ou pelo gráfico de caule-e-folhas respetivos. No caso deste último, estamos a admitir que o rodamos 90 graus no sentido contrário dos ponteiros do relógio. A distribuição pode ser aproximadamente **simétrica** quando os gráficos são aproximadamente simétricos relativamente ao centro da distribuição, ou **assimétrica** quando uma das "caudas" dos gráficos é muito maior do que a outra. No caso da cauda direita (valores grandes) ser muito maior do que a esquerda (valores

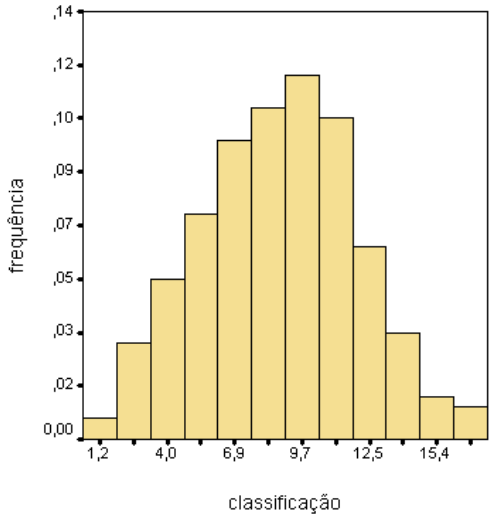
[4]Fonte: Moore e Mccabe, 2003, pp. 8-9.

pequenos), dizemos que temos uma **assimetria positiva**. Quando é a cauda esquerda que é mais longa que a direita, diremos que ocorre uma **assimetria negativa**.

Exemplos de distribuições simétricas são-nos dados nas figuras do Exemplo 1.2.5. Além de simétricas estas distribuições têm uma forma aproximada de "sino". A mesma forma tem a distribuição dos dados do Exemplo 1.2.2, cujo gráfico de caule-e-folhas apresentamos a seguir rodado de 90 graus em sentido contrário aos ponteiros do relógio:

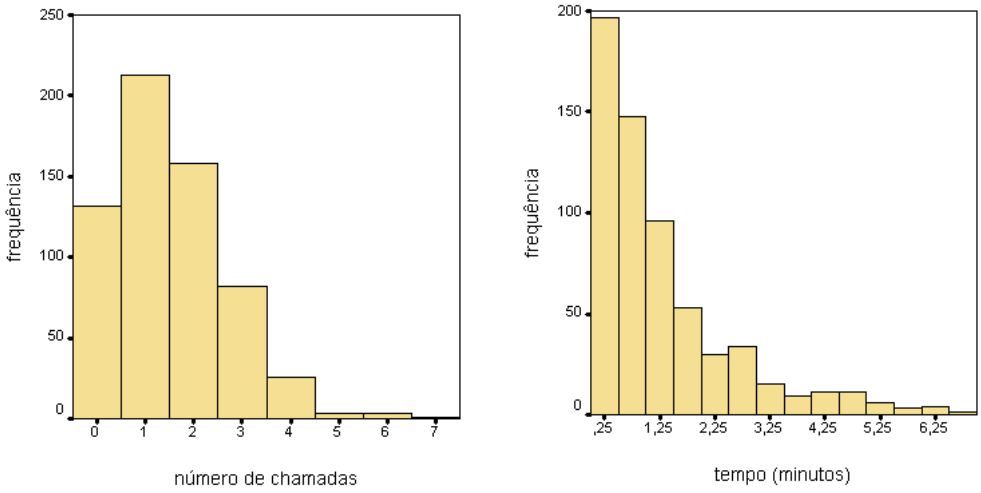
```
peso Stem-and-Leaf Plot

 Frequency   Stem &  Leaf

     8,00     3 .  88888899
    16,00     4 .  0000111222223444
    14,00     4 .  55556667888899
     4,00     5 .  1122

 Stem width:     10
 Each leaf:      1 case(s)
```

Exemplo 1.2.7 Aproximadamente simétrica é também a distribuição das classificações obtidas por 205 alunos numa frequência de Análise Matemática:



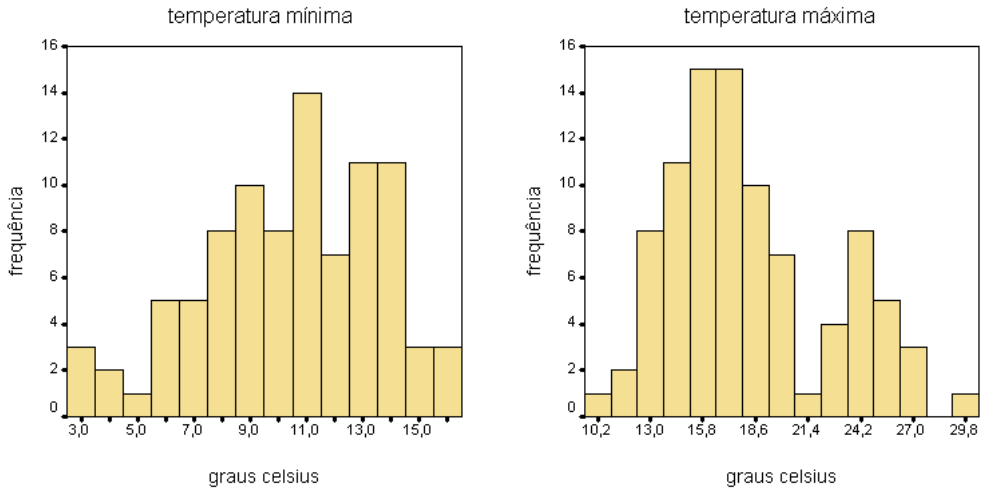
Exemplo 1.2.8 Assimetrias marcadas são reveladas pela distribuição do número de chamadas telefónicas por minuto que chegam a uma central telefónica dum determinado serviço público, ou pela distribuição do tempo (em minutos) que medeia a chegada

de dois clientes consecutivos a uma caixa dum hipermercado. Dos gráficos seguinte constatamos que tais distribuições são positivamente assimétricas.



Uma caraterística comum a todas as distribuições anteriores é a dos gráficos respetivos terem um único "pico" ou **moda** . Tais distribuições são por isso ditas **unimodais**. A moda corresponde à observação ou a uma zona de observações mais frequentes. No exemplo seguinte encontramos uma distribuição com duas modas, dita por isso **bimodal**.

Exemplo 1.2.9 Nos histogramas seguintes descrevem-se as distribuições das temperaturas mínima e máxima ocorridas em Coimbra no Outono de 2000 (dados do Instituto Geofísico da UC). A distribuição da temperatura mínima revela uma assimetria negativa, enquanto que a da temperatura máxima é claramente bimodal.



Uma análise das temperaturas máximas observadas revela que a segunda moda é devida a um conjunto de temperaturas mais elevadas que ocorreram no período popularmente conhecido por "verão de S. Martinho".

1.3 Caraterísticas numéricas duma distribuição

Na secção anterior estudámos formas de resumir graficamente a distribuição duma variável quantitativa. Nesse contexto falámos do **centro** e da **dispersão** duma distribuição. Nesta secção vamos estudar medidas do centro e da dispersão ou variabilidade duma distribuição. Tal como os gráficos, estes resumos numéricos são muito importantes na descrição e interpretação dum conjunto de dados.

1.3.1 Medidas do centro da distribuição

A **média** é a medida mais utilizada do centro duma distribuição. Se denotarmos por $x_1, x_2, \ldots, x_n$ os n valores observados, a média respetiva não é mais do que a soma de todos esses valores dividida pelo número total de observações. A média denota-se por $\bar{x}$ e, de acordo com a definição anterior, é calculada a partir da fórmula seguinte onde o símbolo $\sum x_i$ representa a soma de todos os valores $x_1, x_2, \ldots, x_n$:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}.$$

O cálculo da média só é simples de ser executado sem auxílio dum computador quando número de observações é pequeno, ou quando, sendo grande, o número de observações distintas é pequeno. Neste último caso, se denotarmos por $y_1, y_2, \ldots, y_k$ os valores distintos que ocorrem em $x_1, x_2, \ldots, x_n$, e por $n_1, n_2, \ldots, n_k$ o número de vezes que cada um desses valores ocorre, a fórmula anterior para o cálculo da média reduz-se a

Cálculo da média:

$$\bar{x} = \frac{n_1 y_1 + n_2 y_2 + \ldots + n_k y_k}{n} = \frac{\sum n_i y_i}{n}.$$

Exemplo 1.3.1 Retomemos os dados relativos ao peso dos ratos diabéticos apresentados no Exemplo 1.2.2 (pág. 15). Neste conjunto de 42 observações surgem várias

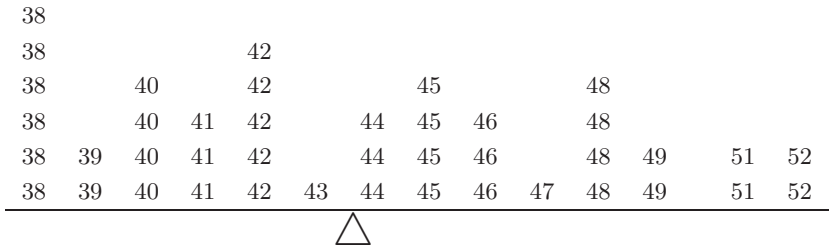
observações repetidas. Neste caso, o cálculo da média é simples de ser executado a partir da tabela de frequências da variável peso

y_i	38	39	40	41	42	43	44	45	46	47	48	49	51	52	Σ
n_i	6	2	4	3	5	1	3	4	3	1	4	2	2	2	42
$n_i y_i$	228	78	160	123	210	43	132	180	138	47	192	98	102	104	1835

Utilizando a segunda das fórmulas anteriores, obtemos

$$\bar{x} = \frac{6 \times 38 + 2 \times 39 + \ldots + 2 \times 52}{42} = \frac{1835}{42} \approx 43.690.$$

A média pode ser interpretada geometricamente de forma simples. Lançando mão das observações anteriores, imaginemos que as colocamos sobre uma barra graduada. A média $\bar{x}$ é o ponto da barra que a mantém em equilíbrio.



Se em vez das observações tivermos acesso ao respetivo histograma, podemos também dizer que a média é o ponto do eixo dos xx que mantém a "figura em equilibrio".

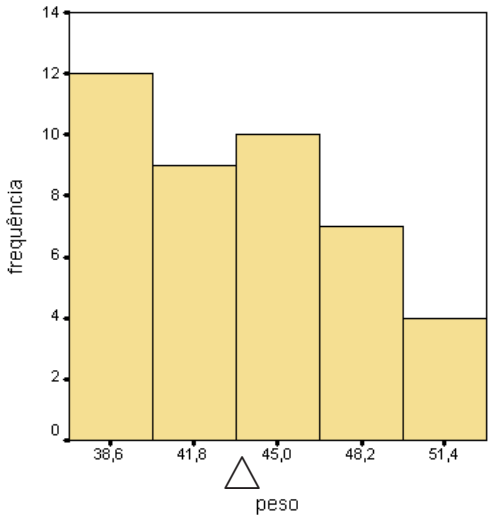


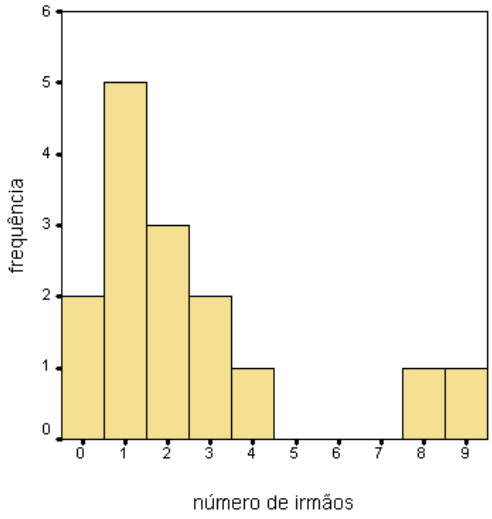
Figura 1.3.2: Localização gráfica da média

A média é uma boa medida do centro da distribuição quando esta é simétrica. No entanto, como vimos no Exemplo 1.2.6, a média é muito sensível à presença de valores muito grandes ou muito pequenos no conjunto das observações. Dizemos então que a média é uma medida pouco **resistente** ou **robusta** do centro da distribuição.

No exemplo seguinte, este facto é mais uma vez realçado.

Exemplo 1.3.3 Os valores seguintes dizem respeito ao número de irmãos de cada um dos 15 alunos duma turma ([5]):

1, 2, 0, 1, 0, 4, 1, 3, 1, 3, 1, 2, 8, 2, 9



Tendo em conta o que dissemos atrás, e sendo as observações 8 e 9 significativamente maiores que as restantes, antes de efetuarmos qualquer cálculo devemos certificar-nos se se tratam, ou não, de verdadeiras observações ou observações corretamente registadas. Devemos, por isso, confirmar estes valores.

Admitindo que os valores são verdadeiros, surge o problema de saber se na presença de tais observações num conjunto de dados tão pequeno, a média é ainda uma boa medida do centro da distribuição. A média das 15 observações é igual $38/15 \approx 2{,}53$. Dizer que os alunos da turma têm em média 2,53 irmãos, isto é, mais de dois irmãos, parece distorcer a realidade pois dos 15 alunos apenas 5 têm mais de 2 irmãos. A presença das observações 8 e 9 faz deslocar a média para a direita de forma muito significativa. Com efeito, se em vez das observações 8 e 9 tivessem sido observados os valores 3 e 4, por exemplo, a média seria igual a $28/15 \approx 1{,}87$. Nesse caso, para descrever o centro da distribuição talvez seja preferível usar uma medida do centro da distribuição que não seja tão sensível a valores muito grandes ou muito pequenos.

[5]Fonte: Martins e Cerveira, 2000, p. 85.

A **não robustez** da **média** como medida do centro da distribuição, é uma propriedade negativa da média. Para contornar esta dificuldade, uma outra medida do centro da distribuição é utilizada em alternativa à média. Trata-se da **mediana**. A **mediana** é um ponto em que aproximadamente metade das observações são menores ou iguais a ele e a outra metade são maiores ou iguais a ele. A mediana é habitualmente representada pela letra M.

Cálculo da mediana:

⊙ ordenar as observações da mais pequena para a maior;

⊙ se o número n de observações é ímpar, a mediana é a observação que está no centro da lista das observações ordenadas; a mediana está assim colocada na posição $(n+1)/2 = n/2 + 1/2$ dessa lista;

⊙ se o número n de observações é par, a mediana é a média das duas observações que estão no centro da lista das observações ordenadas; como estas observações estão colocadas nas posições $n/2$ e $n/2+1$ da lista, dizemos que a mediana está colocada na posição $n/2 + 1/2 = (n+1)/2$ dessa lista.

Reparemos que quando o número n de observações é par, o número $(n+1)/2$ é sempre um número fracionário. Dizer que a mediana está colocada na posição $(n+1)/2$ da lista das observações ordenadas é apenas uma simplificação de linguagem. O que queremos efetivamente dizer é que a mediana é a média das duas observações que estão colocadas nas posições $n/2$ e $n/2+1$ da lista. Como veremos, esta forma de dizer, além da simplificação evidente de linguagem, trará outras vantagens.

Exemplo 1.3.3 (cont.) Para calcular a mediana das observações

$$1,\ 2,\ 0,\ 1,\ 0,\ 4,\ 1,\ 3,\ 1,\ 3,\ 1,\ 2,\ 8,\ 2,\ 9$$

comecemos por ordená-las por ordem crescente:

$$0,\ 0,\ 1,\ 1,\ 1,\ 1,\ 1,\ 2,\ 2,\ 2,\ 3,\ 3,\ 4,\ 8,\ 9.$$

Sendo o número de observações ímpar, $n = 15$, a mediana é a observação central, isto é, é a observação colocada na posição $(15+1)/2 = 8$. Assim $M = 2$. Reparemos que, contrariamente à média, se em vez das observações 8 e 9 tivessem sido observados os valores 3 e 4, a mediana anterior não se alterava. O mesmo aconteceria se alguma, ou

ambas, das observações 8 ou 9 fosse substituída por uma observação grande, por muito grande que ela fosse. Com efeito, a mediana não é sensível às observações que são muito maiores ou muito menores que as restantes. Por isso, dizemos que **a mediana** é uma **medida robusta** do centro da distribuição.

No exemplo anterior constatámos que a média é superior à mediana. Vimos que tal acontece porque, contrariamente à mediana, a média é muito sensível à presença no conjunto das observações de valores grandes. Em geral, sempre que, tal como para a distribuição do número de irmãos, a distribuição é positivamente assimétrica, a média é maior que a mediana. Por razões análogas, se a distribuição é negativamente assimétrica a média é inferior à mediana. Finalmente, se a distribuição é aproximadamente simétrica, a média e a mediana são valores próximos um do outro.

Exemplo 1.3.4 O gráfico de caule-e-folhas e o histograma relativos à distribuição do peso dos ratos considerados no Exemplo 1.2.2 (ver pág. 15, 16 e 21), apesar de não revelarem uma simetria clara da distribuição do peso dos ratos, também não revelam uma assimetria marcada, quer negativa, quer positiva, dessa distribuição. Calculemos a mediana da distribuição dos pesos dos ratos, e verifiquemos que, tal como dissemos atrás, obtemos para mediana um valor próximo do peso médio dos ratos que vimos ser igual a $\bar{x} \approx 43{,}69$. Como o número de observações é par, $n = 42$, a mediana está colocada na posição $(42 + 1)/2 = 21{,}5$ da lista. Como referimos, isto quer dizer que a mediana é a média das observações colocadas nas posições 21 e 22 da lista ordenada das observações. Usando a Tabela 1.3.1 verificamos que tais posições são ocupadas pelas observações 43 e 44.

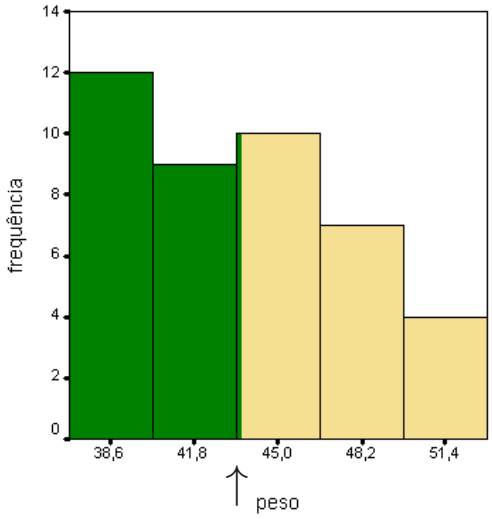


Figura 1.3.5: Localização gráfica da mediana

Assim

$$M = \frac{43 + 44}{2} = 43{,}5.$$

Tal como fizemos para a média, é possível localizar geometricamente a mediana a partir do histograma da distribuição em estudo. A mediana é (aproximadamente) o ponto do eixo dos xx em que a área da porção do histograma à sua esquerda é igual à área da porção do histograma à sua direita.

O exemplo seguinte é também interessante para compreendermos que a média e a mediana, como medidas distintas do centro da distribuição, nos dão informações distintas sobre a realidade que se propõem resumir. Como é natural, devemos escolher aquela que mais relevante seja na descrição dessa realidade.

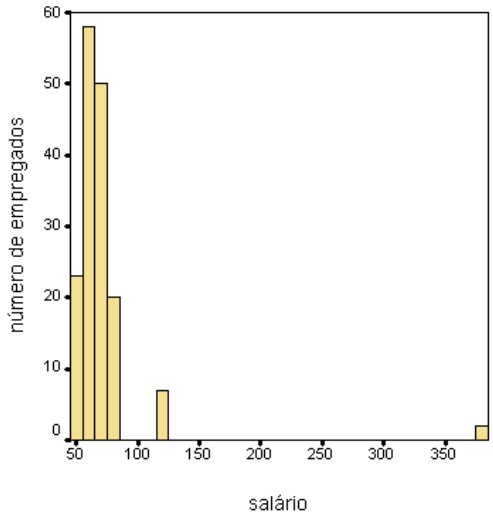
Exemplo 1.3.6 Os salários, em milhares de escudos (1000 escudos correspondem a aproximadamente 5 euros), dos 160 empregados duma empresa, distribuem-se de acordo com a seguinte tabela de frequências ([6]):

Salário	50	60	70	80	120	380
N^o de empregados	23	58	50	20	7	2

Concluímos facilmente que

$$\bar{x} \approx 70{,}81 \quad \text{e} \quad M = 60.$$

A discrepância evidente entre estas duas medidas do centro da distribuição pode ser facilmente compreendida a partir do histograma de frequências seguinte (porquê?):



[6]Fonte: Martins e Cerveira, 2000, p. 92.

Como aproximadamente metade das observações são inferiores ou iguais à mediana, a mediana, $M = 60$, exprime o facto de pelo menos metade dos trabalhadores receberem salários inferiores ou iguais a 60 mil escudos. Esta realidade não é traduzida pela média, $\bar{x} \approx 70{,}81$, uma vez que dos 160 trabalhadores, 81 deles têm salários significativamente inferiores a 70,81 mil escudos. Se o nosso objetivo é conhecer a massa salarial global desta empresa, a medida do centro da distribuição que nos interessa é a média, pois a massa salarial global é dada por

$$\text{massa salarial global} = 160 \times \bar{x} \approx 160 \times 70{,}81 = 11329{,}6 \text{ (milhares de escudos)}.$$

Exemplo 1.3.7 O cálculo da média e da mediana é simples de fazer, sem auxílio de computador, para um conjunto pequeno de observações. Torna-se no entanto impraticável efetuar tal cálculo quando o número de observações é elevado. Tal acontece, por exemplo, caso pretendamos calcular a média e a mediana das distribuições dos pesos dos pacotes de açúcar, antes e depois da calibragem da máquina de empacotamento, descritas no Exemplo 1.2.5 (pág. 24). Num e noutro casos temos 1130 observações. Recorrendo ao SPSS obtemos facilmente o quadro seguinte. Tal como referimos a propósito dos gráficos do Exemplo 1.2.5, ambas as medidas, média e mediana, apontam para que o procedimento de calibragem foi executado com sucesso, uma vez que ambas as medidas do centro da distribuição do peso dos pacotes de açúcar depois da calibragem se aproximam do valor de referência de 1000 gramas. De acordo com o que vimos atrás, as médias anteriores são próximas das medianas respetivas uma vez que ambas as distribuições são simétricas.

Descriptives

	maquina		Statistic
peso (gramas)	antes	Mean	1010,0639
		Median	1009,2444
	depois	Mean	1000,2357
		Median	1000,2773

1.3.2 Medidas de dispersão

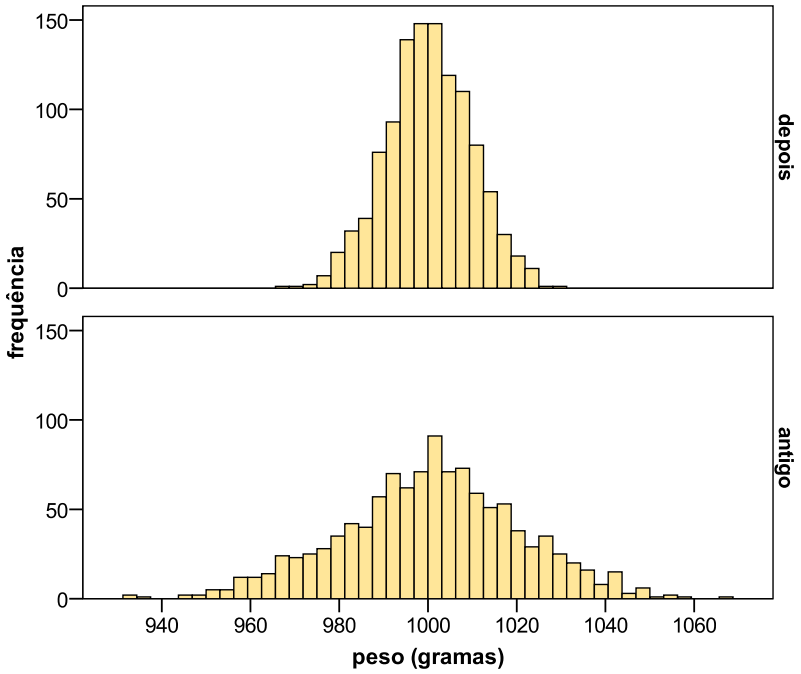
As duas medidas de localização do centro duma distribuição que estudámos na secção anterior, não nos dão qualquer informação sobre a variabilidade do conjunto das observações.

Exemplo 1.3.8 Retomando os dados do Exemplo 1.2.5 (pág. 24), as médias e medianas das distribuições do peso dos pacotes de açúcar empacotados pela máquina depois de calibrada e por uma máquina dum modelo antigo são dadas por:

Descriptives

	maquina		Statistic
peso (gramas)	depois	Mean	1000,2357
		Median	1000,2773
	antigo	Mean	1000,5836
		Median	1001,0458

Comparemos as respetivas distribuições através de histogramas paralelos:



Apesar das média e medianas anteriores serem próximas e das formas das distribuições serem semelhantes, é visível que os pesos dos pacotes de açúcar empacotados pela máquina de modelo mais antigo apresentam maior variabilidade do que os relativos à máquina mais moderna. Por outras palavras, a máquina de modelo mais recente é **mais precisa** do que a de modelo mais antigo.

As medidas de localização, apesar de fundamentais para a compreensão da distribuição dos dados, não nos dão, por si só, um resumo adequado do conjunto das observações. Esse resumo numérico pode ser enriquecido se à medida do centro da distribuição juntarmos uma medida da variabilidade dos dados. Neste parágrafo estudamos **medidas da variabilidade** dum conjunto de dados, ditas também **medidas de dispersão**.

Desvio-padrão e variância

O **desvio-padrão** é uma das medidas de dispersão ou variabilidade mais utilizadas. O **desvio-padrão** mede essa variabilidade relativamente à média $\bar{x}$ do conjunto das observações em causa. Por outras palavras, o desvio-padrão dá-nos informação de quão afastadas da média estão as observações. A sua utilização restringe-se, por isso, ao caso em que a média tenha sido escolhida como medida do centro da distribuição.

Se $x_1, x_2, \ldots, x_n$ são os n valores observados, o seu **desvio-padrão** denota-se por s e é definido por

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}.$$

Por vezes utiliza-se o quadrado do desvio-padrão, a que chamamos **variância**, como medida da variabilidade do conjunto das observações. Reparemos que contrariamente à variância, o desvio-padrão vem expresso nas mesmas unidades que os dados iniciais. Por exemplo, se as observações x_i são expressas em metros, o desvio-padrão vem expresso em metros, enquanto que a variância vem expressa em metros quadrados.

A **variância** denota-se por s^2 e, de acordo com a definição anterior, é dada por

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}.$$

Reparemos que as observações mais afastadas da média contribuem mais para o desvio-padrão (e também para a variância) do que observações próximas da média.

Exemplo 1.3.8 (cont.) Não será assim de estranhar que o desvio-padrão dos dados descritos pelos histogramas do Exemplo 1.2.5 (pág. 24) seja inferior ao dos dados descritos no Exemplo 1.3.8 anterior:

Descriptives

	maquina		Statistic
peso (gramas)	depois	Mean	1000,2357
		Variance	91,646
		Std. Deviation	9,57317
	antigo	Mean	1000,5836
		Variance	412,274
		Std. Deviation	20,30453

Quando pretendemos efetuar o cálculo do desvio-padrão sem auxílio dum computador, a fórmula anterior não é a mais adequada para o efeito. Em vez dela deve ser usada uma das fórmulas seguintes:

Cálculo do desvio-padrão:

$$s = \sqrt{\frac{\sum x_i^2 - n\,\bar{x}^2}{n-1}} = \sqrt{\frac{\sum x_i^2 - (\sum x_i)^2/n}{n-1}}.$$

Tal como para média, o cálculo do desvio-padrão só é simples de ser executado sem auxílio dum computador quando número de observações é pequeno, ou quando, sendo grande, o número de valores distintos é pequeno. Neste último caso, se denotarmos por $y_1, y_2, \ldots, y_k$ os valores distintos que ocorrem em $x_1, x_2, \ldots, x_n$, e por $n_1, n_2, \ldots, n_k$ a frequência absoluta de cada um desses valores, as fórmulas anteriores para o cálculo do desvio-padrão reduzem-se a

Cálculo do desvio-padrão:

$$s = \sqrt{\frac{\sum n_i\, y_i^2 - n\,\bar{x}^2}{n-1}} = \sqrt{\frac{\sum n_i\, y_i^2 - (\sum n_i\, y_i)^2/n}{n-1}}.$$

Exemplo 1.3.9 Retomemos os dados relativos ao peso dos ratos diabéticos do Exemplo 1.2.2 (pág. 15). Como neste conjunto de 42 observações surgem várias observações repetidas, o cálculo do desvio-padrão é simples de ser executado utilizando a penúltima das fórmulas anteriores, a partir da tabela de frequências da variável peso dada a seguir. Obtemos então:

$$s = \sqrt{\frac{80911 - 42 \times (1835/42)^2}{42-1}} \approx 4{,}245.$$

Reparemos que em vez do valor 43,69 anteriormente obtido para a média, estamos a utilizar na fórmula anterior o verdadeiro valor da média. Deste modo, evitamos que o erro de arredondamento que o valor 43,69 comporta se propague ao cálculo do desvio-padrão:

$$\sqrt{\frac{80911 - 42 \times (43{,}69)^2}{42-1}} \approx 4{,}250.$$

y_i	n_i	$n_i y_i$	y_i^2	$n_i y_i^2$
38	6	228	1444	8664
39	2	78	1521	3042
40	4	160	1600	6400
41	3	123	1681	5043
42	5	210	1764	8820
43	1	43	1849	1849
44	3	132	1936	5808
45	4	180	2025	8100
46	3	138	2116	6348
47	1	47	2209	2209
48	4	192	2304	9216
49	2	98	2401	4802
51	2	102	2601	5202
52	2	104	2704	5408
Σ	42	1835	28155	80911

Amplitude

Como já referimos no início do parágrafo 1.2.3, a dispersão de uma distribuição pode também ser medida pela diferença entre a maior e a menor observação. Ao valor obtido pela diferença entre os valores **máximo** e o **mínimo** do conjunto das observações chamamos **amplitude**, e vamos denotá-la por A:

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

Por razões análogas às avançadas a propósito da média, o desvio-padrão é uma medida de dispersão pouco robusta. Da definição de amplitude que acabámos de dar, é claro que também a amplitude é uma medida pouco robusta pois o máximo e o mínimo são muito sensíveis à presença de valores, respetivamente, muito grandes ou muito pequenos, no conjunto de dados. Em particular, a amplitude não deve ser usada para comparar a dispersão de dois conjuntos de dados a não ser que tenham a mesma dimensão, uma vez que a amplitude tende a aumentar à medida que a dimensão da amostra aumenta.

Quartis e amplitude interquartil

Uma medida de dispersão mais robusta do que as anteriores é a **amplitude interquartil**. Para o seu cálculo é necessário obter os **primeiro e terceiro quartis** do conjunto das observações.

Os **quartis**, que denotamos por Q_1, Q_2 e Q_3, são quantidades numéricas caraterizadas pelo facto de (aproximadamente) 25%, 50% e 75% das observações, respetivamente, serem menores ou iguais a elas. De acordo com a definição de **mediana**, concluímos que o **segundo quartil** é precisamente a mediana. Por vezes Q_1, $Q_2 = M$ e Q_3 são também referidos como sendo **percentis** de ordens 25, 50 e 75, respetivamente. Com efeito, sendo p um número inteiro maior que 0 e inferior a 100, o **percentil de ordem** p é caraterizado pelo facto de (aproximadamente) p% das observações serem menores ou iguais a ele. Mais geralmente, sendo p um número entre 0 e 1, o **quantil de ordem** p é caraterizado por uma proporção p de observações ser inferior ou igual a ele. Assim, Q_1, $Q_2 = M$ e Q_3 são os quantis de ordem 0,25, 0,5 e 0,75, respetivamente.

Como já referimos, para o cálculo da amplitude interquartil precisamos de calcular os quartis Q_1 e Q_3. Vejamos agora como proceder:

Cálculo dos quartis Q_1 e Q_3 (Método de Tukey):

⊙ ordenar as observações da mais pequena para a maior;

⊙ calcular a posição da mediana M na lista ordenada das observações;

⊙ o primeiro quartil, Q_1, é a mediana das **observações** cujas posições, na lista ordenada das observações, são inferiores ou iguais à posição de M;

⊙ o terceiro quartil, Q_3, é a mediana das **observações** cujas posições, na lista ordenada das observações, são superiores ou iguais à posição de M.

De forma análoga ao que fizemos para a mediana, podemos verificar que o primeiro e o terceiro quartis são pouco sensíveis à presença nos dados de observações muito grandes ou muito pequenas em comparação com as restantes observações.

Exemplo 1.3.10 Calculemos os quartis Q_1 e Q_3 do seguinte conjunto de dados:

$$10, 10, 11, 12, 12, 13, 13, 13, 14, 15, 16, 17, 17, 18.$$

Como temos 14 observações, a mediana está colocada na posição $(14+1)/2 = 7.5$. O primeiro quartil é então a mediana das observações colocadas nas posições $1, 2, \ldots, 7$ uma vez que são estas as posições inferiores ou iguais à posição da mediana: $10, 10, 11, 12, 12, 13, 13$. Assim $Q_1 = 12$. De forma análoga $Q_3 = 16$, pois 16 é a mediana das observações colocadas nas posições $8, 9, \ldots, 13, 14$ uma vez que são estas as posições superiores ou iguais à posição da mediana: $13, 14, \ldots, 17, 18$.

Para as observações

$$10, 10, 11, 12, 12, 13, 13, 13, 14, 15, 16, 17, 17,$$

a mediana está colocada na posição $(13 + 1)/2 = 7$. O primeiro quartil é então a mediana das observações colocadas nas posições $1, 2, \ldots, 7$: $10, 10, 11, 12, 12, 13, 13$. Assim $Q_1 = 12$. De forma análoga Q_3 é a mediana das observações colocadas nas posições $7, 8, \ldots, 12, 13$: $13, 13, \ldots, 17, 17$. Assim $Q_3 = 15$.

Notemos que tal como fizemos para a mediana, os quartis podem ser aproximadamente localizados a partir dum histograma (ver figura seguinte). Q_1 e Q_3 são (aproximadamente) os ponto do eixo dos xx em que a área da porção do histograma à sua esquerda é igual a 1/4 e 3/4, respetivamente, da área total.

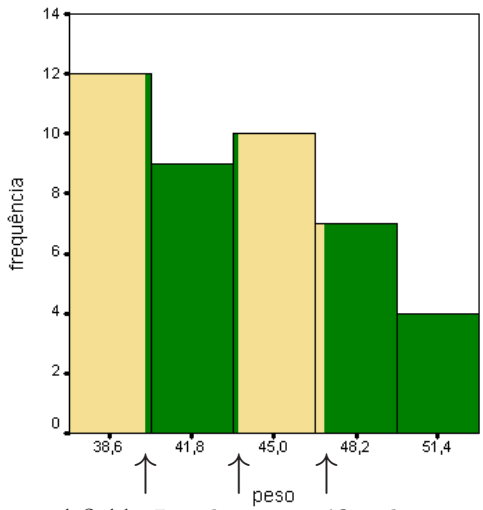


Figura 1.3.11: Localização gráfica dos quartis

Estamos agora em condições de definir a amplitude interquartil, que vamos denotar por AIQ. A **amplitude interquartil** é a diferença entre os terceiro e primeiro quartis:

$$AIQ = Q_3 - Q_1.$$

A robustez desta medida de dispersão é consequência da robustez dos primeiro e terceiro quartis.

1.3.3 Identificação de observações discordantes

A amplitude interquartil intervém na regra habitualmente utilizada para identificar observações discordantes.

Regra para a identificação de observações discordantes:

Uma observação é considerada discordante se estiver fora do intervalo

$$[Q_1 - 1{,}5 \times AIQ, Q_3 + 1{,}5 \times AIQ].$$

Como exemplificaremos mais à frente, as observações identificadas pela regra anterior não são necessariamente falsas observações ou observações mal registadas, casos em que o valores em causa devem ser excluídos ou corrigidos, respetivamente. Tal é em particular verdade quando a distribuição dos dados é bastante assimétrica. Neste caso, a regra anterior identifica com frequência observações na parte direita do conjunto de dados quando a distribuição é positivamente assimétrica, ou na parte esquerda do conjunto de dados quando a distribuição é negativamente assimétrica. No caso do valor discordante corresponder a uma verdadeira observação, a sua inclusão ou exclusão do conjunto dos dados depende da influência que tal observação tenha nas metodologias estatísticas que estejam a ser usadas. Tratando-se duma observação que, por si só, determina as conclusões do estudo em curso, será mais prudente retirá-la do conjunto dos dados (a este propósito ver o Exemplo 1.2.6).

Exemplo 1.3.12 No segundo dos gráficos de caule-e-folhas apresentados no Exemplo 1.2.3 relativo ao grupo experimental (pág. 18), a observação 313 é, como vimos, discordante. Confirmemos este facto a partir da regra anterior. Os primeiro e terceiro quartis da distribuição dos pesos são dados por $Q_1 = 384$ e $Q_3 = 429$. Como $AIQ = 429 - 384 = 45$, as observações inferiores a $384 - 1{,}5 \times 45 = 316{,}5$ e superiores a $429 + 1{,}5 \times 45 = 496{,}5$ são consideradas discordantes. Como podemos confirmar a partir dos dados do Exemplo 1.2.3 (pág. 17), apenas a observação 313 está nestas condições.

Para o cálculo da mediana, começámos por calcular a sua posição na lista ordenada das observações. Podemos proceder de igual modo no cálculo dos quartis. Das regras anteriores para o cálculo de Q_1 e Q_3 deduz-se que sendo k o número de observações usadas para calcular Q_1, a **posição de** Q_1 na lista ordenada das observações é $(k+1)/2$. De forma análoga, como k é também o número de observações usadas para calcular Q_3, começando agora a contar da maior para a menor observação a **posição de** Q_3 é também $(k+1)/2$.

Exemplo 1.3.13 Retomemos os dados relativos ao peso dos ratos diabéticos considerados no Exemplo 1.3.9 (pág. 37). Calculemos Q_1 e Q_3, começando pela determinação das suas posições na lista ordenada de todas as observações. Sendo 42 o número total

de observações, as 21 primeiras intervêm no cálculo de Q_1 e as últimas 21 intervêm no cálculo de Q_3. Assim, como a posição de Q_1 na lista ordenada das observações, é $(21+1)/2 = 11$, usando a tabela de frequências apresentada no Exemplo 1.3.9, concluímos que $Q_1 = 40$ e $Q_3 = 47$. A amplitude interquartil é igual a $AIQ = 47-40 = 7$. Neste caso $Q_1 - 1{,}5 \times AIQ = 40 - 1{,}5 \times 7 = 29{,}5$ e $Q_3 + 1{,}5 \times AIQ = 47 + 1{,}5 \times 7 = 57{,}5$, o que significa que nenhuma observação é considerada discordante.

Regra alternativa para a determinação dos quartis

O cálculo dos quartis pode ser feito utilizando métodos ligeiramente diferentes do que demos atrás. Um desses métodos passa pelo cálculo das posições do primeiro e terceiro quartis a partir das fórmulas

$$\text{Posição de } Q_1 = \frac{1}{4}(n+1) \quad \text{e} \quad \text{Posição de } Q_3 = \frac{3}{4}(n+1),$$

fórmulas estas que estendem de forma natural a fórmula que usámos para o cálculo da posição da mediana

$$\text{Posição de } M = \frac{1}{2}(n+1).$$

Vejamos através de exemplos simples como podemos usar os valores anteriores para calcular Q_1 e Q_3.

Exemplo 1.3.14 Retomemos o conjunto de dados

$$10, 10, 11, 12, 12, 13, 13, 13, 14, 15, 16, 17, 17, 18,$$

que considerámos no Exemplo 1.3.10. Como temos 14 observações, o primeiro quartil está colocado na posição $\frac{1}{4}(14+1) = 3{,}75$. Assim, Q_1 está situado entre as observações 11 e 12, que são as observações colocadas nas 3.ª e 4.ª posições da lista ordenada das observações. No entanto como a parte decimal da posição de Q_1 é 0,75, isto significa que Q_1 estará mais próximo da 4.ª do que da 3.ª observação, mais precisamente, Q_1 será dado por

$$Q_1 = 11 + 0{,}75 \times (12 - 11) = 11{,}75.$$

Relativamente ao terceiro quartil, este está colocado na posição $\frac{3}{4}(14+1) = 11{,}25$. Q_3 está assim situado entre as observações 16 e 17, que são as observações colocadas nas 11.ª e 12.ª posições da lista ordenada das observações. No entanto como a parte decimal da posição de Q_3 é 0,25, isto significa que Q_3 estará mais próximo de 16 do que de 17, ou seja,

$$Q_3 = 16 + 0{,}25 \times (17 - 16) = 16{,}25.$$

Recordemos que os valores de Q_1 e Q_3 obtidos pela regra de Tukey eram 12 e 16, respectivamente.

Exemplo 1.3.15 Consideremos agora as observações do grupo experimental do Exemplo 1.2.3 (pág. 18). Sendo 20 o número total de observações, o primeiro e terceiro quartis têm posições $\frac{1}{4}(20+1) = 5{,}25$ e $\frac{3}{4}(20+1) = 15{,}75$. Assim, de acordo com a regra alternativa que usámos atrás para o cálculo dos quartis, temos $Q_1 = 37+0{,}25\times(39-37) = 37{,}5$ e $Q_3 = 42 + 0{,}75 \times (43 - 42) = 42{,}75$.

Exemplo 1.3.16 Retomemos os dados relativos ao peso dos ratos diabéticos considerados no Exemplo 1.3.13 (pág. 41). Calculemos Q_1 e Q_3, começando pela determinação das suas posições na lista ordenada de todas as observações usando as fórmulas anteriores. Sendo 42 o número total de observações, o primeiro e terceiro quartis têm posições $\frac{1}{4}(42+1) = 10{,}75$ e $\frac{3}{4}(42+1) = 32{,}25$. Assim, $Q_1 = 40 + 0{,}75 \times (40 - 40) = 40$ e $Q_3 = 47 + 0{,}25 \times (48 - 47) = 47{,}25$.

Na tabela seguinte produzida pelo SPSS, os quartis são calculados pelos dois métodos expostos. Contrariamente ao método de Tukey, exclusivamente destinado ao cálculo dos quartis, o método alternativo usado atrás pode ser facilmente adaptado ao cálculo de outros quantis.

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	peso	38,00	38,00	40,00	43,50	47,25	50,40	51,85
Tukey's Hinges	peso			40,00	43,50	47,00		

Como podemos confirmar pela tabela seguinte, o SPSS usa os valores da primeira linha do quadro anterior para calcular a amplitude interquartil, o que conduz a um valor diferente do que calculámos no Exemplo 1.3.13 (pág. 41).

Descriptives

		Statistic
peso	Mean	43,69
	Median	43,50
	Variance	18,024
	Std. Deviation	4,245
	Minimum	38
	Maximum	52
	Range	14
	Interquartile Range	7,25

1.3.4 Gráfico de extremos-e-quartis

Decorre das definições anteriores, que o primeiro quartil, a mediana e o terceiro

quartil, dividem o conjunto das observações em quatro partes, cada uma das quais contendo, aproximadamente, 25% das observações. Esquematicamente:

	25%		25%		25%		25%	
mínimo		Q_1		M		Q_3		máximo

Estes números, ditos **cinco números de resumo duma distribuição**, dão-nos uma informação bastante completa sobre a distribuição subjacente aos dados: a mediana descreve o centro da distribuição; os quartis permitem descrever, através da amplitude interquartil, a variabilidade da metade central da distribuição; o mínimo e o máximo permitem descrever, através da amplitude, a variabilidade de todo o conjunto dos dados. Devido à sua robustez, a amplitude interquartil é também usada, como alternativa à amplitude, como medida da variabilidade de todo o conjunto dos dados.

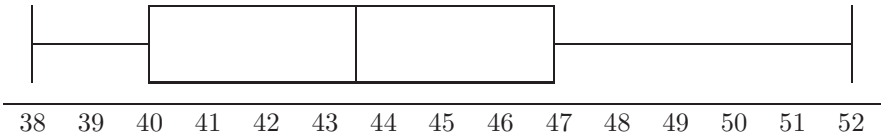
Estes cinco números de resumo dão origem a uma representação gráfica bastante interessante. Trata-se do **gráfico de extremos-e-quartis** que tem o aspeto seguinte:

mínimo Q_1 M Q_3 máximo

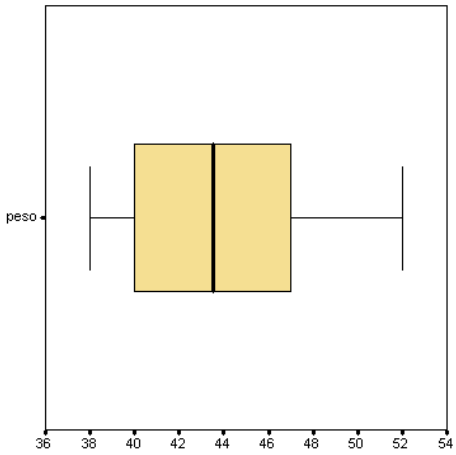
Exemplo 1.3.17 Vimos no Exemplo 1.3.9 (pág. 41) que os cinco números de resumo da distribuição do peso dos ratos são dados por:

$$\text{mínimo} = 38, \quad Q_1 = 40, \quad M = 43{,}5, \quad Q_3 = 47, \quad \text{máximo} = 52.$$

O gráfico de extremos-e-quartis correspondente é dado por



Para esta distribuição, o SPSS produz o seguinte gráfico de extremos-e-quartis:

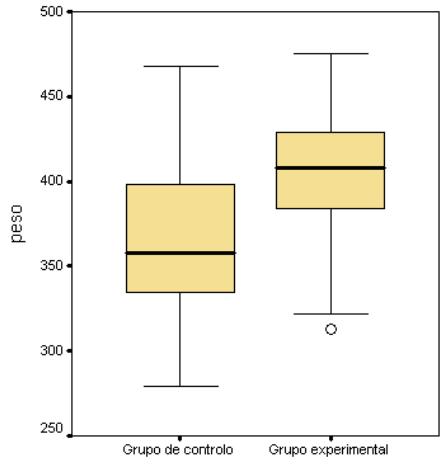


Os números de resumo, máximo e mínimo, incluídos na representação em gráfico de extremos-e-quartis, são muito sensíveis à presença nos dados de observações discordantes. Para que o aspeto do gráfico não dependa em demasia destas observações, é habitual que as barras exteriores do gráfico sejam marcadas, não no máximo ou no mínimo, mas sim, na menor e na maior observação não discordante. Neste novo **gráfico de extremos-e-quartis**, as observações discordantes são representadas individualmente (através de asteriscos ou pequenos círculos).

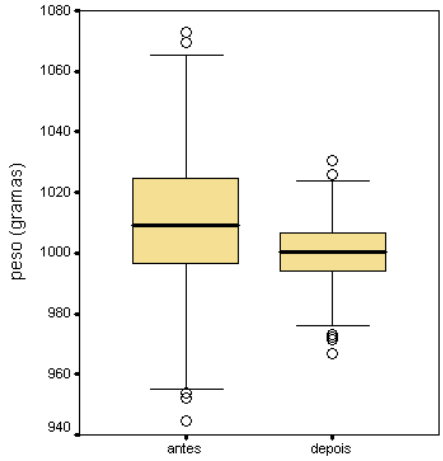
mais pequena observação não discordante — Q_1 — M — Q_3 — maior observação não discordante

Os gráficos de extremos-e-quartis são também de extrema importância na comparação de várias distribuições.

Exemplo 1.3.18 Ilustremos estes factos apresentando na figura seguinte os **gráficos de extremos-e-quartis paralelos** correspondentes ao grupo de controlo e ao grupo experimental do Exemplo 1.2.3 (pág. 17). Reparemos no gráfico respeitante ao grupo experimental em que a observação discordante é marcada individualmente. A conclusão tirada a partir dos gráficos de caule-e-folhas paralelos (ver pág. 19), de que há boas razões para concluir que a nova farinha é preferível à antiga, é agora reforçada. Reparemos que não só a mediana do grupo experimental é superior à mediana do grupo de controlo, como a dispersão do grupo experimental é inferior à do grupo de controlo (porquê?).

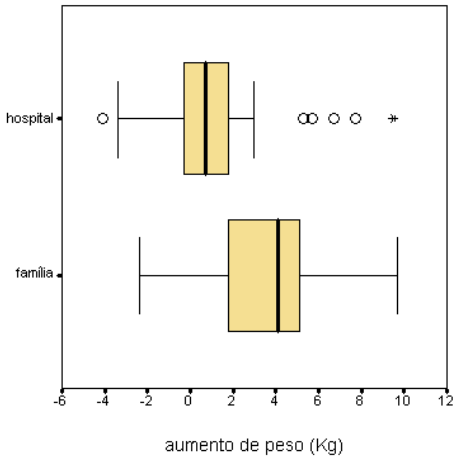


Exemplo 1.3.19 Apresentamos de seguida os gráfico de extremos-e-quartis paralelos relativos à distribuição dos pesos (em gramas) de pacotes de açúcar empacotados por uma máquina antes e depois de ter sido calibrada, cujos histogramas apresentámos no Exemplo 1.2.5 (pág. 24). As conclusões retiradas a partir dos gráficos aí apresentados, são análogas às que podemos tirar dos gráficos seguintes.

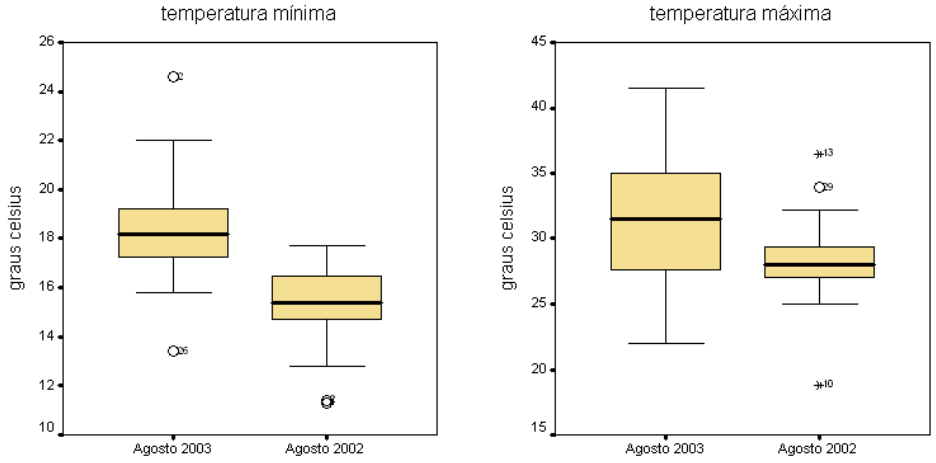


Exemplo 1.3.20 Na Tabela 0.2.2 (pág. 4) apresentámos os pesos, em quilogramas, de dois grupos de jovens anoréxicas, no início do tratamento e passado quatro semanas. Um grupo recebe o tratamento em internamento hospitalar na companhia dum familiar e o outro recebe o tratamento residindo com a família. Os gráficos de extremos-e-quartis paralelos, por grupo de tratamento, para a distribuição das diferenças de peso verificadas (final-inicial), indiciam que, para os grupos de estudo considerados, o tratamento produziu mais efeito quando a doente continuou a residir com a família. Como podemos observar, o SPSS distingue as observações discordantes, assinalando de forma diferente

as mais extremas.

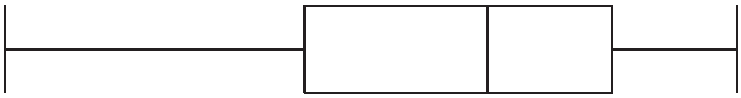


Exemplo 1.3.21 Os gráficos de extremos-e-quartis seguintes relativos às temperaturas máximas e mínimas ocorridas em Coimbra nos meses de Agosto de 2002 e 2003, revela que o mês de Agosto de 2003 foi muito mais quente que o de 2002 (porquê?).

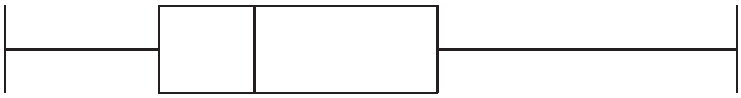


Os gráficos de extremos-e-quartis são também úteis na descrição da forma da distribuição. Para distribuições simétricas, assimétricas negativas e assimétricas positivas, é o seguinte o aspeto dos gráficos de extremos-e-quartis correspondentes:

Distribuição simétrica

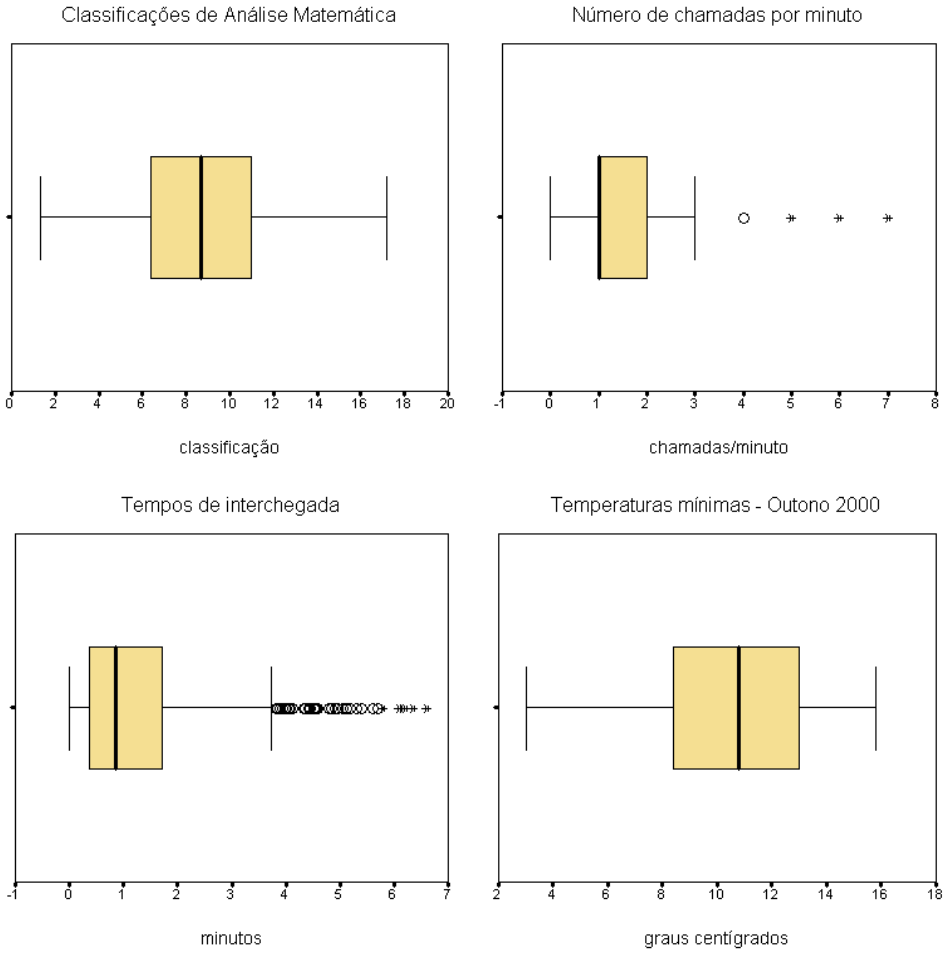


Distribuição assimétrica negativa



Distribuição assimétrica positiva

Os gráficos de extremos-e-quartis seguintes, são relativos às distribuições descritas nos Exemplos 1.2.7, 1.2.8 e 1.2.9. No primeiro caso a distribuição é simétrica, enquanto que nos dois casos seguintes os gráficos revelam distribuições fortemente assimétricas positivas. No último caso, a distribuição é negativamente assimétrica.



1.4 Alteração da unidade de medida

Quando na observação de determinada caraterística dos indivíduos em estudo efetuamos uma medição, essa medição pode habitualmente ser registada em diferentes unidades.

No Exemplo 1.2.5, o peso dos pacotes de açúcar foi registado em gramas mas poderia ter sido registado em quilogramas ou em libras. No registo de temperaturas, estas podem ser registadas em graus Fahrenheit ou, mais habitualmente, em graus Celsius ou centígrados. Na medição de distâncias, os europeus continentais utilizam o quilómetro enquanto que os britânicos e americanos utilizam a milha. Os americanos utilizam o galão como medida para líquidos enquanto que nós utilizamos o litro.

Em todas estas situações, para um mesmo indivíduo i, os dois valores x_i e y_i registados, correspondentes a unidades diferentes de medida, estão relacionados através duma equação do tipo

$$y_i = a\, x_i + b,$$

para determinados valores reais $a > 0$ e b. Esta relação exprime a **alteração da unidade de medida** verificada. Dizemos que os valores originais x_i foram transformados nos novos valores y_i através duma **transformação linear**.

Exemplo 1.4.1 Se x_i representar o peso em gramas e y_i o peso em quilogramas sabemos que

$$y_i = \frac{x_i}{1000} = \frac{1}{1000}\, x_i.$$

Se x_i representar o peso em quilogramas e y_i o peso em libras, então

$$y_i = 2{,}2046\; x_i.$$

No primeiro caso $a = 1/1000$, enquanto que no segundo $a = 2{,}2046$. Em ambos os casos $b = 0$.

Exemplo 1.4.2 Se x_i for a temperatura em graus Fahrenheit, a temperatura em graus Celsius é dada por

$$y_i = \frac{5}{9}\,(x_i - 32).$$

Neste caso $a = 5/9$ e $b = -160/9$ (porquê?).

A questão que colocamos neste parágrafo é a de saber como variam a forma da distribuição e os seus resumos numéricos do centro e de dispersão, quando os dados são transformados através duma transformação linear.

Comecemos por analisar o efeito produzido por uma transformação do tipo

$$y_i = x_i + b$$

isto é, a cada uma das observações originais x_i foi adicionado um mesmo valor b. Como sabemos, a operação de adicionar a constante b a todos os dados x_i produz uma translação deste conjunto de dados. Os novos valores y_i estão assim distanciados dos correspondentes valores x_i de b unidades, e estão à direita daqueles se b é positivo, e à sua esquerda se b é negativo. As medidas do centro da distribuição, média e mediana, da nova distribuição de dados y_i devem assim ser obtidas das anteriores adicionando-lhes b unidades, isto é, $\bar{y} = \bar{x} + b$ e $M_y = M_x + b$. Por outro lado, como a posição relativa dos dados x_i é precisamente a mesma que a dos dados y_i, tendo-se mantido inalteradas as distâncias correspondentes, as medidas de dispersão, desvio-padrão e amplitude interquartil, mantêm-se inalteradas, ou seja, $s_y = s_x$ e $AIQ_y = AIQ_x$. Finalmente, o histograma relativo aos novos dados surge deslocado de a unidades relativamente ao histograma original. A forma da distribuição não sofre assim qualquer alteração.

Vejamos agora o efeito, sobre as caraterísticas distribucionais anteriores, da transformação

$$y_i = a\, x_i,$$

isto é, a cada uma das observações originais x_i foi multiplicada por um mesmo valor $a > 0$. A operação de multiplicar todos os pontos x_i por um número a, corresponde a uma homotetia, de razão a e centro na origem, deste conjunto de pontos (contração do conjunto de pontos se $a < 1$ e dilatação se $a > 1$). Será assim de esperar que as medidas de localização central das observações transformadas sejam afetadas pela mesma homotetia de razão a e centro na origem, isto é, $\bar{y} = a\,\bar{x}$ e $M_y = a\, M_x$. A posição relativa dos pontos y_i é precisamente a mesma que a dos pontos x_i, mas a distância entre duas quaisquer das novas observações é igual à distância entre as observações originais correspondentes multiplicada por a. Assim, as novas medidas de localização e dispersão, obtêm-se das originais multiplicando-as por a, ou seja, $s_y = a\, s_x$ e $AIQ_y = a\, AIQ_x$. A forma da distribuição não sofrerá também qualquer alteração.

Tendo agora em conta que a transformação $y_i = a\, x_i + b$ se obtém efetuando em primeiro lugar a transformação $z_i = a\, x_i$, e depois a transformação $y_i = z_i + b$, das conclusões anteriores podemos obter o quadro seguinte:

Efeito da transformação linear $y_i = a\,x_i + b$ $(a > 0)$:

- medidas de localização central:

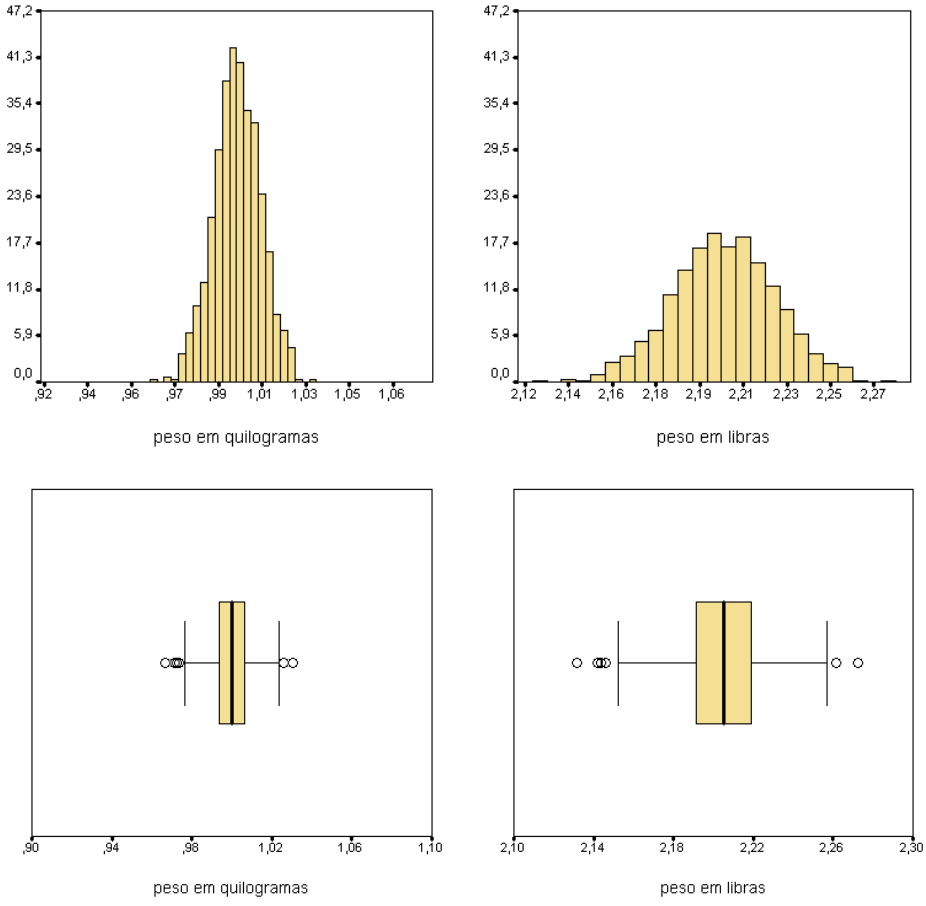
$$\bar{y} = a\,\bar{x} + b, \quad M_y = a\,M_x + b;$$

- medidas de dispersão:

$$s_y = a\,s_x, \quad AIQ_y = a\,AIQ_x;$$

- a forma da distribuição não sofre alteração.

Exemplo 1.4.1 (cont.) Os histogramas e os gráficos de extremos-e-quartis seguintes, dizem respeito ao peso, em quilogramas e em libras, dos pacotes de açúcar considerados no Exemplo 1.2.5 depois da calibragem da máquina. Para facilitar a sua comparação, os intervalos correspondentes aí considerados têm igual amplitude.



Da comparação dos gráficos é claro o aumento da média, da mediana, do desvio-padrão e da amplitude interquartil da distribuição do peso em libras relativamente à distribuição do peso em quilogramas. Como esperado, a forma mantém-se inalterada.

Do quadro seguinte podemos ainda confirmar que a média, a mediana, o desvio-padrão e a amplitude interquartil da distribuição do peso em libras, se obtém dos correspondentes valores da distribuição do peso em quilogramas multiplicando-os por 2,2046.

Descriptives

	maquina		Statistic
peso (quilos)	depois	Mean	1,0002357
		Median	1,0002773
		Std. Deviation	,00957317
		Interquartile Range	,01260
peso (libras)	depois	Mean	2,2051197
		Median	2,2052113
		Std. Deviation	,02110500
		Interquartile Range	,02778

1.5 Bibliografia

Daniel, W.W. (2009). *Biostatistics: a foundation for analysis in the health sciences*, Wiley.

Dawson, B., Trapp, R.G. (2004). *Basic & Clinical Biostatistics*, McGraw-Hill.

Martins, M.E.G., Cerveira, A.G. (2000). *Introdução às Probabilidades e à Estatística*, Universidade Aberta.

Moore, D.S., McCabe, G.P. (2003). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

Moore, D.S., McCabe, G.P. (2006). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

Murteira, B.J.F. (1993). *Análise Exploratória de Dados. Estatística Descritiva*, McGraw-Hill.

Pestana, D.D., Velosa, S.F. (2002). *Introdução à Probabilidade e à Estatística*, Vol. I, Fundação Calouste Gulbenkian.

2

A recolha dos dados

Estudos observacionais e experiências. Planeamento de experiências. Fator, nível, tratamento. Experiências controladas. Números aleatórios e aleatorização na constituição dos grupos. Constituição de blocos e emparelhamento. Planeamento de estudos por amostragem. População, amostra, parâmetro, estatística. Amostragem aleatória simples, estratificada e em várias etapas. Métodos de amostragem não-aleatórios. Viés e variabilidade. Distribuição amostral.

2.1 A importância duma adequada recolha de dados

Até ao momento estivemos interessados em descrever gráfica e numericamente um conjunto de dados provenientes da observação de determinadas variáveis num conjunto de indivíduos. A descrição da distribuição duma variável, a identificação de observações discordantes ou influentes, e a comparação das distribuições duma variável observada em duas populações, foram, de um modo geral, os objetivos principais do estudo até agora realizado.

A natureza exploratória e descritiva desse estudo não permite que as suas conclusões permaneçam válidas para além dos dados considerados. Em alguns dos exemplos focados, é claro que a análise até agora feita só parcialmente responde às principais questões colocadas quando recolhemos os dados. Exemplos do que acabámos de referir são o Exemplo 1.2.3, onde se pretende avaliar os efeito duma nova farinha para a alimentação de pintos, e o exemplo da sondagem de opinião a que fizemos referência no capítulo introdutório. Quando utilizámos gráficos de caule-e-folhas e de extremos-e-quartis para comparar os dois grupos de pintos do Exemplo 1.2.3 (ver pág. 45), o objetivo principal era saber se a nova farinha deve ser utilizada na alimentação de todos os pintos do aviário, ou mesmo na de pintos de outros aviários que estejam em condições semelhantes às dos pintos observados. O mesmo se passa na realização duma sondagem eleitoral, dum estudo de saúde pública ou dum estudo sobre o consumo. O

seu interesse principal está na possibilidade de extrapolar para toda a população, os resultados obtidos para os indivíduos inquiridos.

Para que esse procedimento inferencial não conduza a resultados inválidos, é necessário que os dados sejam recolhidos de forma apropriada, que permita, em particular, a utilização de metodologias estatísticas adequadas para os analisar. Sobre algumas dessas metodologias falaremos em capítulos futuros.

Exemplo 2.1.1 Para realçar a importância duma adequada recolha de dados, retomemos o Exemplo 1.2.3 (pág. 17) e imaginemos que os 40 pintos selecionados para testar a nova farinha se encontravam numa caixa grande, da qual retirávamos, um a um, 20 pintos para formar o grupo ao qual era ministrada a farinha habitual. Para evitar fatores pessoais na escolha dos pintos, suponhamos que, sem olhar, introduzíamos a mão na caixa e retirávamos o primeiro pinto que apanhássemos. Este procedimento de seleção dos grupos experimentais não é o mais apropriado, encerrando vícios que podem deturpar o resultado do estudo. Por exemplo, será de esperar que os pintos mais fracos se deixem apanhar mais facilmente, ficando no grupo ao qual era ministrada a nova farinha, os pintos mais fortes e saudáveis. Não seria de estranhar que desse grupo proviessem os pintos mais gordos quando, passado alguns dias, todos eles fossem pesados.

Nos próximos parágrafos abordaremos sucintamente questões relacionadas com a recolha ou produção de dados em **estudos observacionais por amostragem**, cujo principal objetivo é o de recolher informação sobre um conjunto de indivíduos a partir da observação de uma pequena parte destes, e nos quais – por isso serem chamados de **estudos observacionais** –, os indivíduos são observados e as suas caraterísticas registadas sem lhes impor qualquer regime específico, e num **estudo experimental** em que o observador impõe deliberadamente um tratamento ou regime específico aos indivíduos intervenientes no estudo com o objetivo de observar a sua resposta.

2.2 Planeamento de experiências

Como referimos atrás, um estudo diz-se uma **experiência** quando uma ou várias condições experimentais específicas são deliberadamente impostas aos indivíduos, também chamados de **unidades experimentais**, de modo a observar a sua resposta. A cada uma dessas condições experimentais chamamos **tratamento**. Assim cada um destes tratamentos resulta da alteração de uma ou de várias variáveis, ditas também **independentes** ou **explicativas**, e que no contexto das experiências são também chamadas de **fatores**. Os diferentes valores que os fatores tomam dizem-se também **níveis**

desse fator. Os níveis não são assim mais do que os diferentes valores que as variáveis explicativas, quantitativas ou qualitativas, tomam. O objetivo duma experiência é estudar o efeito dessas alterações na variável **resposta** a que chamamos também variável **dependente**.

Exemplo 2.2.1 A comparação dos dois tratamentos para a recuperação de jovens anoréxicas apresentados no Exemplo 0.2.1 (pág. 3), é um exemplo típico duma experiência. Os dois tratamentos em confronto dizem respeito à forma como é aplicada a terapia habitual. A variável resposta é aqui o peso e há apenas um fator, a terapia, com dois níveis respeitantes ao regime, ambulatório ou de internamento, em que a terapia habitual é aplicada.

Exemplo 2.2.2 A comparação das duas dietas para os pintos do Exemplo 1.2.3 (pág. 17), é outro exemplo duma experiência. Os tratamentos são constituídos aqui pelas duas dietas impostas aos pintos. A variável resposta é o peso e há apenas um fator com dois níveis respeitantes ao tipo de dieta aplicado. Além do efeito da farinha, poderíamos estar também interessados no efeito produzido por um complexo proteico que era, ou não, adicionado à farinha. Teríamos assim mais um fator com dois níveis possíveis. A conjugação destes dois **fatores**, cada um deles com dois **níveis**, daria origem a quatro tratamentos diferentes.

Algumas questões importantes relativas ao **planeamento das experiências** anteriores ou de quaisquer outras experiências, podem ser levantadas. Tais questões têm, no essencial, a ver com o objetivo de controlar a variação de variáveis distintas da variável que está a ser medida (variável resposta) que podem ter influência nessa variável. É importante para a validade do estudo que, quer no início, quer durante a aplicação dos diferentes tratamentos, o efeito dessas variáveis, a ocorrer, se manifeste de igual forma nos indivíduos dos vários grupos de tratamento. Dizemos neste caso que a **experiência** está **controlada**.

Exemplo 2.2.2 (cont.) Retomemos o exemplo dos pintos, e suponhamos que a nova farinha era dada a 20 pintos que manteríamos afastados dos restantes pintos do aviário para garantir que estes só se alimentavam com a nova farinha, e que passado alguns dias comparávamos o seu peso com o de 20 outros pintos selecionados no aviário. Ao planearmos a experiência desta forma, os pintos que comem a nova farinha vivem sob condições diferentes das dos restantes pintos do aviário. Podem ter mais ou menos espaço, mais ou menos calor, mais ou menos quantidade de alimento, etc. Como todas estas variáveis podem influenciar o seu crescimento, no final da experiência ficaríamos sem saber se as possíveis diferenças observadas na variável resposta se deviam às diferentes farinhas utilizadas, ou ao efeito de algumas das variáveis que não foram controladas.

Neste exemplo concreto, o controlo dessas variáveis pode passar por garantir que os pintos de ambos os grupos vivam sob condições semelhantes durante a realização do estudo.

Outra questão importante é relativa à forma como os indivíduos são divididos pelos vários grupos de tratamento. Como já fizemos notar na secção anterior, esta é uma questão importante que quando não é tida em conta, pode conduzir ao favorecimento sistemático de determinado resultado. Nesse caso dizemos que há um enviesamento dos resultados do estudo. Em populações humanas a constituição dos grupos é por vezes feita de forma a que esses grupos sejam semelhantes relativamente a algumas variáveis tidas como possivelmente influentes na resposta ao tratamento. Por razões já avançadas, este pode não ser o método mais adequado para constituir os grupos pois pode haver **variáveis omissas** que influenciem fortemente a resposta aos diferentes tratamentos.

A **aleatorização** na constituição dos grupos experimentais é uma forma simples de evitar o problema anterior. Isto quer dizer que os indivíduos a incluir em cada um dos grupos devem ser escolhidos ao acaso, isto é, de forma aleatória, evitando-se assim escolhas pessoais ou subjectivas na sua seleção. Voltando ao exemplo dos pintos, a aleatorização produz grupos de pintos que devem ser semelhantes em todos os aspetos antes de começar o estudo. Desta forma estamos a esbater diferenças que surgem sempre entre os indivíduos. No caso dos pintos haverá, por exemplo, pintos com mais tendência a engordar do que outros. No entanto a probabilidade de tais pintos serem selecionados para um dos grupos é igual à de serem selecionados para o outro grupo. Ao procedermos da forma anterior, as diferenças observadas no final do estudo serão devidas aos diferentes tratamentos ou ao papel desempenhado pelo acaso na constituição dos grupos. Quando uma tal diferença é tão grande que raramente poderia ocorrer por acaso, dizemos que se trata duma diferença estatisticamente significativa.

Para proceder à constituição dos grupos experimentais por métodos aleatórios, devemos começar por numerar, da forma mais simples possível, todos os indivíduos intervenientes no estudo. Destes devemos escolher ao acaso alguns que integrarão um dos grupos experimentais. Para os outros grupos procede-se da mesma maneira. Ao dizermos que escolhemos ao acaso alguns indivíduos, digamos m, queremos dizer que todas as possíveis amostras com m indivíduos deverão ter todas a mesma possibilidade de serem selecionadas. A maior partes das aplicações informáticas com rotinas de estatística, ou mesmo uma calculadora mais evoluída, têm uma função (chamada random ou aleatório) para executar a tarefa anterior.

Exemplo 2.2.2 (cont.) No exemplo dos pintos, os 40 indivíduos podem ser numerados de 1 a 40. Destes 40 números, devem ser escolhidos 20 ao acaso que integrarão um

dos grupos do estudo. Usando o SPSS obtemos os seguintes números (excluídas as repetições): 36, 28, 33, 06, 32, 01, 30, 18, 12, 29, 02, 17, 16, 27, 15, 20, 35, 13, 08, 19.

Sem auxílio dum computador podemos também efetuar a aleatorização dos grupos usando uma **tabela de números aleatórios**:

> Uma **tabela de números aleatórios** é uma lista dos algarismos $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ com as propriedades seguintes:
>
> • qualquer posição da lista é ocupada com igual possibilidade por qualquer um dos algarismos $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$;
>
> • algarismos colocados em diferentes posições na lista são independentes, no sentido em que o valor de um não influencia o valor de outro.

Das propriedades anteriores deduz-se ainda que: i) qualquer par de algarismos selecionado tem igual possibilidade de ser um dos pares $00, 01, 02, \ldots, 98, 99$; ii) qualquer terno de algarismos tem igual possibilidade de ser um dos ternos $000, 001, 002, 998, 999$; iii) valem propriedades análogas para grupos de quatro ou mais algarismos.

A Tabela A (pág. 305) é um exemplo duma tabela de números aleatórios. Apesar desta ter sido gerada por computador, uma tabela deste tipo poderia ter sido construída com o auxílio duma esfera de extração de bolas da lotaria (ou outro sistema análogo), na qual introduzíamos 10 bolas com os algarismos $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ e da qual extraíamos uma bola registando o seu algarismo. Depois de repor na esfera a bola extraída, o processo seria repetido tanta vezes quantos os algarismos que desejássemos incluir na tabela.

Uma tabela de números aleatórios pode ser lida por qualquer ordem, ao longo duma linha, ao longo duma coluna, da esquerda para a direita, da direita para a esquerda, etc. Para fixar um modo de leitura que todos utilizemos, vamos ler a tabela por colunas, de cima para baixo e da esquerda para a direita, só passando às colunas seguintes da tabela depois de esgotar todas as linhas da tabela.

Exemplo 2.2.2 (cont.) Relativamente ao exemplo dos pintos, iniciando a leitura da tabela na linha 01 da coluna 11, e agrupando os algarismos em grupos de dois, obtemos os 20 números seguintes (excluídas as repetições e os agrupamentos de dois algarismos $00, 41, 42, \ldots, 98, 99$): 03, 14, 15, 32, 04, 10, 11, 36, 40, 23, 12, 33, 22, 02, 39, 16, 18, 17, 24, 31.

O processo de aleatorização na constituição dos grupos que acabámos de descrever, é, como referimos, um método simples de constituir grupos que sejam homogéneos relativamente a variáveis, identificadas, ou não, à partida, que possam influenciar a resposta dos indivíduos aos diferentes tratamentos.

No entanto, a aleatorização na constituição dos grupos não nos guarda da possibilidade de obtermos grupos que sejam pouco homogéneos relativamente a variáveis influentes, omissas ou não, na variável que medimos. No caso particular de algumas dessas variáveis influentes estarem identificadas à partida, processos diferentes para a constituição dos grupos podem ser utilizados.

Se, por exemplo, pretendemos comparar duas dietas numa população humana de indivíduos entre os 25 e os 30 anos, e tivermos razão para acreditar que a variável sexo possa influenciar fortemente os resultados, em vez de se proceder à aleatorização na constituição dos grupos a partir de todo o conjunto de indivíduos independentemente do seu sexo, é preferível conduzir experiências separadas para homens e para mulheres, e proceder à aleatorização na constituição dos grupos dentro de cada um destes grupos, a que chamamos habitualmente **blocos**. Ao constituirmos **blocos** de indivíduos que são semelhantes relativamente a uma variável que afeta a resposta ao tratamento, podemos ainda tirar conclusões separadas acerca de cada um desses blocos.

Na comparação de dois tratamentos, é por vezes possível aplicar ambos os tratamentos num mesmo indivíduo ou em dois indivíduos que sejam semelhantes relativamente às variáveis influentes na variável resposta. Dizemos então que os indivíduos, as observações, ou as amostras, estão **emparelhados**. No primeiro caso, em algumas experiências os dois tratamentos são aplicados ao indivíduo por ordem aleatória, enquanto que no segundo caso os indivíduos emparelhados são afectos a um ou a outro dos grupos de forma aleatória.

2.3 Planeamento de estudos por amostragem

Neste parágrafo centramos a nossa atenção nos chamados **estudos por amostragem** que têm como objetivo tirar conclusões gerais acerca dum conjunto finito de indivíduos a partir da observação dum números restrito desses indivíduos. Contrariamente a uma experiência em que certas condições são impostas aos indivíduos de modo a observar a sua resposta, nos estudos por amostragem os indivíduos são observados nas condições habituais com o objetivo de determinar alguma ou algumas caraterísticas particulares da população.

O conjunto total dos indivíduos, ou **unidades individuais**, sobre o qual queremos obter informação é denominado **população**. À parte da população que é sujeita

a observação chamamos **amostra**. Ao número de indivíduos da amostra chamamos **dimensão** ou **tamanho** da amostra.

Num estudo deste género, pretendemos normalmente obter informação sobre características numéricas dessa população, a que chamamos **parâmetros**. Para o efeito utilizamos as caraterísticas amostrais correspondentes a que chamamos **estatísticas**. As estatísticas são assim funções da amostra que não dependem dos parâmetros populacionais.

Exemplo 2.3.1 A título de exemplo, suponhamos que algum tempo antes das eleições para a AAC pretendemos conhecer a percentagem p de estudantes que vão votar. A população é aqui constituída por todos os alunos da UC. A percentagem de alunos que vão votar é aqui o parâmetro em que estamos interessados. Para o conhecermos teríamos de inquirir todos os alunos da UC, isto é, teríamos de realizar um **censo**. Sendo esta tarefa difícil, ou mesmo impossível, de ser realizada num período curto de tempo, seríamos conduzidos a inquirir alguns, não muitos, alunos da UC aos quais perguntávamos se iriam votar no dia das eleições. Para uma tal amostra é fácil calcular a percentagem de alunos que vão votar. Essa percentagem é uma estatística.

Um ponto essencial do **planeamento dum estudo por amostragem** é o da escolha do método a utilizar para recolher a amostra. A fase da recolha da amostra é de grande importância pois esta deve ser, na medida do possível, representativa da população que se pretende estudar, isto é, a estatística calculada deve ser uma aproximação razoável da caraterística populacional de interesse. Amostras representativas da população dizem-se **sem viés** ou **não enviesadas**. Caso contrários dizemos que as amostras são **enviesadas**. O caso das **amostras de resposta voluntária**, que ocorrem quando em programas televisivos é lançada uma questão para ser respondida pelo espectadores, são exemplos de amostras que apresentam enviesamentos claros favorecendo de forma sistemática um dos resultados.

Tal como para o caso das experiências que abordámos no parágrafo anterior, uma forma simples de evitar o enviesamento da amostra, evitando preferências pessoais na sua escolha ou o problema da resposta voluntária, é proceder à sua seleção por **métodos aleatórios** ou **probabilísticos**.

Descrevemos a seguir três destes métodos, a **amostragem aleatória simples**, a **amostragem estratificada** e a **amostragem em várias etapas**, para os quais indicamos algumas vantagens e desvantagens.

A **amostragem aleatória simples** é um dos métodos mais simples de seleção de amostras de tamanho fixo n duma população. Uma amostra aleatória simples obtém-se selecionando ao acaso, e sem reposição, os elementos da amostra tendo por base a população. Em particular, todas as possíveis amostras com n elementos têm a mesma

possibilidade de ser selecionadas. Este foi precisamente o método utilizado na aleatorização dos grupos numa experiência. A aleatorização na constituição dos grupos numa experiência não é mais do que uma amostragem aleatória simples que tem por base o conjunto dos indivíduos intervenientes no estudo. Para obter uma amostra aleatória simples, é necessário listar todos os indivíduos da população atribuindo um número a cada um deles. A seguir utilizamos uma tabela de números aleatórios e selecionamos a amostra com o tamanho desejado.

A **amostragem estratificada** realiza-se quando possuímos informação suplementar sobre a população que permita fazer a sua divisão em subpopulações ou **estratos**. A ideia da amostragem estratificada é a de selecionar em cada um desses estratos uma amostra aleatória simples, combinando depois essas diferentes amostras para obter informação sobre a população. Como vantagens da estratificação da população podemos referir o facto dela permitir obter informação sobre cada um dos estratos, tornar o processo de amostragem mais simples, e oferecer mais garantia de representatividade à amostra uma vez que uma amostra aleatória simples com base na população poderia não conter qualquer elemento de um dos estratos. Pode provar-se matematicamente que a amostragem estratificada permite obter resultados mais exatos do que a amostragem aleatória simples quando a população é muito heterogénea mas as subpopulações que integram os estratos são razoavelmente homogéneas relativamente à variável de interesse. Ao pretendermos obter uma amostra estratificada de dimensão n, é preciso saber a dimensão das amostras a recolher em cada estrato. Uma forma de o fazer, conhecida como **afetação proporcional**, consiste em recolher em cada estrato uma amostra de dimensão proporcional à dimensão do estrato. Notemos, no entanto, que a afetação proporcional nem sempre é a mais indicada. É razoável pensar que em estratos homogéneos relativamente à caraterística em estudo, a dimensão da amostra a recolher poderá ser mais pequena do que em estratos mais heterogéneos.

Os dois métodos de amostragem anteriores, exigem que a população, ou melhor, que as suas unidades individuais estejam listadas. Casos há, em que apesar de não ser possível listar toda a população é possível identificar grupos de indivíduos e listar tais grupos. A **amostragem aleatória em várias etapas**, também designada por **amostragem por grupos**, é um método de amostragem aleatória em que a escolha aleatória da amostra é feita em várias fases. Para a sua utilização a população é dividida em grupos ditos **unidades amostrais**. Esta começa por ser dividida em **unidades primárias**, cada um destes grupos pode ser dividido em subgrupos ditos **unidades secundárias**, e assim sucessivamente. Cada unidade corresponde a uma etapa do processo de amostragem, etapas essas que vão sendo percorridas até se chegar às **unidades finais** que são as únicas a serem inquiridas. Em cada etapa a seleção das

unidades a considerar pode ser feita por amostragem aleatória simples ou por outros métodos de amostragem aleatórios. Como as unidades finais são as únicas a serem inquiridas, apenas estas necessitam de ser listadas.

Exemplo 2.3.1 (cont.) No quadro seguinte indica-se o número de alunos por cada uma das Faculdades da UC ([1]):

Faculdade	nº de alunos	nº de licenciaturas
FL	4606	17
FD	3145	2
FM	1512	2
FCT	7669	23
FF	934	1
FE	2460	4
FPCE	1271	2
FCDEF	475	1
Total	22072	52

Havendo listas de todos os alunos da UC, qualquer um dos métodos anteriores de amostragem pode ser aplicado. Apenas no sentido de ilustrar a sua aplicação, admitamos que pretendíamos recolher uma amostra de tamanho 100 para estimar a percentagem p de estudantes que neste momento pensam ir votar nas próximas eleições da AAC.

Usando a **amostragem aleatória simples**, teríamos que numerar todos os alunos, por exemplo de 00001 a 22072 (FL: 00001 a 04606, FD: 04606 a 07751, FM: 07752 a 09263, FCT: 09264 a 16932, etc), e usar um computador ou uma tabela de números aleatórios para selecionar a amostra. Usando a Tabela A e iniciando a leitura na primeira linha da primeira coluna, os alunos selecionados são os numerados por: 15685, 14768, 05374, 15252, 07908,....

Usando agora a **amostragem estratificada** com afetação proporcional, em que os estratos são as diversas Faculdades, é preciso começar por determinar o número de alunos de cada uma das Faculdades que devemos incluir na amostra. Sendo a afetação proporcional, obtemos as afetações FL: 21, FD: 14, FM: 9, FCT: 35, FF: 4, FE: 11, FPCE: 6, FCDEF: 2. Para extrair uma amostra aleatória simples de dimensão 21 da Faculdade de Letras, listamos os seus alunos da 0001 a 4606. Iniciando a leitura da Tabela A na primeira linha da primeira coluna (por exemplo), os alunos a incluir na amostra são o 4156, 4596, 1568, 2581, 1476,.... De igual modo procederíamos para as restantes Faculdades.

[1] Fonte: Prospeto da UC de 2003/04.

Usando agora a **amostragem aleatória a várias etapas**, poderíamos considerar as Faculdades as unidades primárias, e os alunos dessas Faculdades as **unidades secundárias** e **finais**. Neste caso, este tipo de amostragem é também conhecido por **amostragem aleatória bietápica**. Na primeira etapa escolheríamos algumas Faculdades, e na segunda escolheríamos alguns alunos das Faculdades selecionadas. Em vez duma amostragem em duas etapas, poderíamos ter também considerado uma amostragem a três etapas em que as **unidades terciárias e finais** seriam os alunos de cada uma das licenciaturas das Faculdades. Assim, enquanto que na primeira etapa escolhíamos algumas das Faculdades, na segunda etapa, para cada uma das Faculdades escolhidas na etapa anterior, escolhíamos alguma ou algumas das suas licenciaturas, e na etapa final seriam selecionados aleatoriamente alguns alunos das licenciaturas escolhidas.

Uma segunda classe de métodos para seleção de amostras é bastante utilizada na prática pela sua maior facilidade de implementação e economia, quando comparados com os métodos aleatórios. Nestes, a amostra é escolhida de modo que, segundos determinados critérios, mais ou menos subjetivos, se assemelhe à população. Contrariamente aos métodos aleatórios, estes métodos de amostragem, ditos **não aleatórios** ou **não probabilísticos**, não permitem medir o grau de confiança que podemos ter nos resultados que com base neles obtemos. Entre os métodos não-aleatórios mais utilizados encontramos a **amostragem de resposta voluntária**, que surge na forma de questionários de rua, questionários incluídos em revistas, questionários televisivos, etc, a **amostragem de conveniência**, em que a amostra é formado por sujeitos facilmente acessíveis, e a **amostragem por quotas**, que é usada com frequência nos estudos de mercado, em que para o entrevistador são definidas quotas de indivíduos a entrevistar para os diferentes grupos de indivíduos em que a população foi dividida. Apesar de semelhante à amostragem estratificada, uma vez que a população em estudo também é dividida em estratos, difere desta pelo facto dos indivíduos serem selecionados de forma não aleatória dentro de cada estrato.

Em situações práticas os métodos de seleção de amostras podem na prática ser combinados. A sondagem de opinião descrita no Exemplo 0.3.1 (pág. 5) é ilustrativa de uma tal situação, em que os lares foram escolhidos por métodos de amostragem aleatórios enquanto que em cada um dos lares selecionados o sujeito inquirido foi escolhido pelo método das quotas.

Um conveniente planeamento dum estudo de amostragem não se resume apenas à escolha dum método apropriado de amostragem, que como vimos deve ser aleatório para evitar o enviesamento da amostra. Outras fontes de enviesamento da amostra devem ser acauteladas, como são o problema da **não cobertura**, que ocorre quando

a população que realmente foi alvo do estudo não coincide com a população que se pretende estudar, e o problema da **não resposta**, que ocorre em populações humanas, quando um indivíduo selecionado para integrar a amostra se recusa a participar no estudo. Em estudos de amostragem que envolvam a resposta a um questionário, o comportamento do entrevistador e do entrevistado, bem como a clareza das questões que são formuladas, podem influenciar fortemente a qualidade do estudo.

2.4 Viés, variabilidade e distribuição amostral

Os métodos aleatórios utilizados nos dois parágrafos anteriores a propósito da constituição de grupos de tratamento numa experiência ou da seleção duma amostra num estudo por amostragem, foram motivadas pelo objetivo comum de evitar o enviesamento dos resultados obtidos nesses estudos.

Com o duplo objetivo de precisar um pouco mais a noção de **enviesamento** e de motivar a noção de **variabilidade**, vamos centrar-nos num estudo observacional por amostragem em que, para uma determinada população, pretendemos conhecer a proporção p de indivíduos que possuem determinada caraterística. Essa proporção é o **parâmetro de interesse**. Admitamos que utilizamos o método de amostragem aleatória simples para recolher uma amostra. A partir da amostra recolhida podemos calcular a estatística $\widehat{p}$ associada ao parâmetro de interesse que, neste caso, não é mais do que a proporção de indivíduos nessa amostra que possuem essa caraterística. A $\widehat{p}$ chamamos **proporção amostral**. Não havendo enviesamento no que respeita à amostragem, esperamos que esta estatística nos dê uma boa informação sobre o parâmetro desconhecido p.

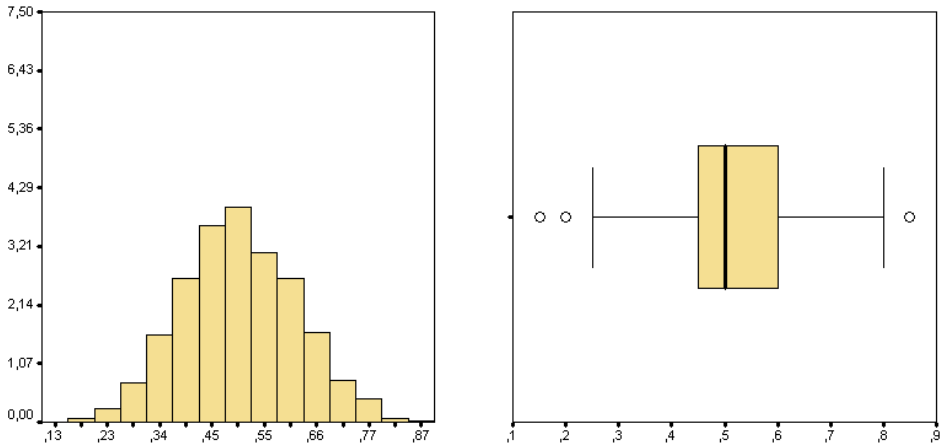
Para ir um pouco mais longe nesta interpretação, admitamos que várias amostras aleatórias simples, todas com a mesma dimensão, são recolhidas, e que para cada uma delas calculamos $\widehat{p}$. Será de esperar que diferentes valores de $\widehat{p}$ sejam obtidos para as diferentes amostras. Este facto é conhecido por **variabilidade amostral**. Estes vários valores obtidos para a estatística $\widehat{p}$ podem ser interpretados como diferentes observações da estatística $\widehat{p}$. Utilizando os métodos do Capítulo 1 será possível saber mais sobre esta estatística? Qual o centro da sua distribuição? E a sua variabilidade? Qual é a sua forma? Uma vez que diferentes valores de $\widehat{p}$ são obtidos a partir de diferentes amostras, à distribuição de $\widehat{p}$ chamamos **distribuição amostral** de $\widehat{p}$. A distribuição amostral duma estatística descreve assim o comportamento da estatística em sucessivas repetições do processo de amostragem.

Exemplo 2.4.1 Para lançar algumas pistas de resposta a estas questões, vamos reduzir-nos ao exemplo concreto duma população de 10000 indivíduos, que numeramos de 1

a 10000, em que (estranhamente) os indivíduos numerados de 1 a 5000 possuem a caraterística em estudo, e os restantes, numerados de 5001 a 10000, não possuem essa caraterística. Neste caso $p = 1/2$. Para cada uma de 2000 amostras de dimensão 20 recolhidas desta população, calculámos $\widehat{p}$. Para as 85 primeiras obtivemos os valores:

0,45; 0,75; 0,55; 0,60; 0,40; 0,45; 0,50; 0,30; 0,65; 0,55; 0,50; 0,50; 0,50; 0,40; 0,65; 0,35;
0,50; 0,50; 0,35; 0,65; 0,35; 0,60; 0,35; 0,45; 0,55; 0,55; 0,65; 0,60; 0,60; 0,35; 0,50; 0,55;
0,40; 0,60; 0,60; 0,55; 0,65; 0,50; 0,60; 0,60; 0,60; 0,45; 0,45; 0,50; 0,70; 0,30; 0,70; 0,35;
0,60; 0,50; 0,40; 0,50; 0,55; 0,50; 0,50; 0,50; 0,60; 0,50; 0,35; 0,55; 0,50; 0,35; 0,50; 0,60;
0,50; 0,35; 0,40; 0,45; 0,45; 0,40; 0,45; 0,25; 0,50; 0,30; 0,65; 0,40; 0,50; 0,55; 0,55; 0,55

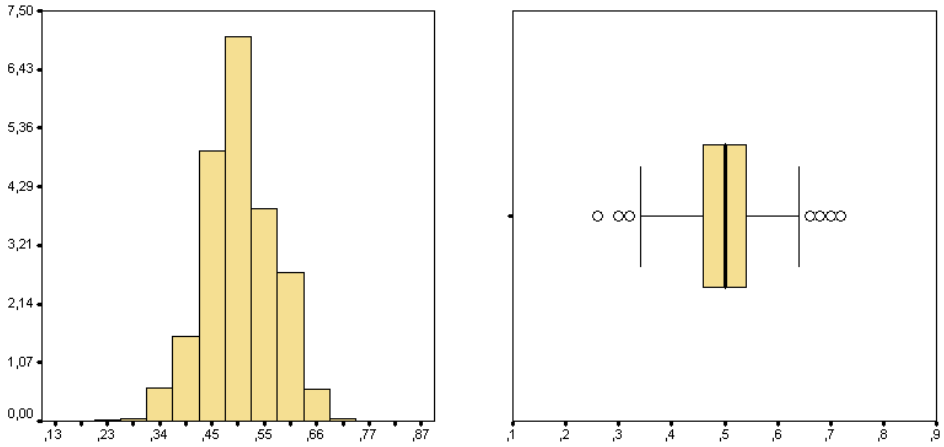
Usando todos os valores obtidos para $\widehat{p}$, apresentamos a seguir dois resumos gráficos da distribuição amostral de $\widehat{p}$:



Comecemos por notar que o centro da distribuição de $\widehat{p}$ é aproximadamente igual ao verdadeiro parâmetro $p = 1/2$. Dizemos assim que a estatística $\widehat{p}$ **não é enviesada** ou que **não tem viés**. Esta excelente propriedade é uma das consequências de termos usado um método de amostragem aleatório na seleção das amostras. Assim, dizer que os resultados do estudo por amostragem não são enviesados significa que a estatística de interesse, neste caso $\widehat{p}$, possui como centro da sua distribuição amostral o verdadeiro parâmetro p (desconhecido). O centro da distribuição duma estatística **enviesada** ou **com viés** não coincide com o verdadeiro valor do parâmetro de interesse.

A **variabilidade** da estatística $\widehat{p}$ é naturalmente descrita pela variabilidade da sua distribuição amostral. Uma menor variabilidade corresponde naturalmente a resultados mais precisos. Esta variabilidade depende da dimensão da amostra recolhida. Quanto maior a dimensão da amostra menor a variabilidade da estatística $\widehat{p}$. Este facto é ilustrado pelos gráficos seguinte relativos à distribuição amostral da estatística $\widehat{p}$ obtida

a partir de 2000 amostras de dimensão 50. Para facilitar a comparação, as escalas usadas nestes e nos gráficos anteriores são iguais.



As conclusões anteriores, válidas para a amostragem aleatória simples e para a estatística $\widehat{p}$, permanecem válidas para estatísticas apropriadamente escolhidas quando as amostras são selecionadas por métodos aleatórios, ou quando se procede à aleatorização na constituição de grupos de tratamento numa experiência. Assim sendo, tais estatísticas não têm viés e a sua variabilidade pode ser reduzida pelo aumento da dimensão da amostra ou do tamanho dos grupos experimentais.

Há no entanto outra enorme vantagem na utilização de procedimentos aleatórios que não é partilhada pelos métodos não-aleatórios de seleção de amostras a que fizemos breve referência no parágrafo anterior: a distribuição da estatística de interesse é em geral conhecida (pelo menos de forma aproximada).

No caso particular da estatística $\widehat{p}$, quando a seleção da amostra é feita por amostragem aleatória simples, a distribuição é simétrica, unimodal, com a forma dum sino como é ilustrado pelos histogramas anteriores. Veremos que uma tal distribuição é próxima de um tipo de distribuição a que chamaremos **distribuição normal**, também conhecida por distribuição gaussiana, de Gausse, ou de Gauss-Laplace, devido a Pierre-Simon Laplace (1744-1827) e Carl Friedrich Gauss (1777-1855). Utilizando a noção de probabilidade, veremos mais à frente que a distribuição de $\widehat{p}$ é aproximadamente normal independentemente do valor do parâmetro desconhecido p. No caso dos valores atrás obtidos para $\widehat{p}$, este facto é sugerido pelos histogramas e gráficos de extremos-e-quartis anteriores. Além disso, e também como é sugerido pelos gráficos anteriores, veremos que o **centro da distribuição** de $\widehat{p}$ é o parâmetro desconhecido p. Veremos ainda que a **variabilidade da distribuição** de $\widehat{p}$ pode ser aproximada a partir da amostra observada.

Na posse de toda esta informação poderemos afirmar com grande confiança (confiança esta que será medida usando a noção de probabilidade), que $\widehat{p}$ pertence a um intervalo do tipo $[p-\widehat{V}, p+\widehat{V}]$ (recorde que p é o centro da distribuição de $\widehat{p}$ e que esta é aproximadamente simétrica), ou seja,

$$p - \widehat{V} \leq \widehat{p} \leq p + \widehat{V},$$

onde a quantidade $\widehat{V}$ pode ser calculada a partir da amostra observada e está relacionada com a variabilidade da distribuição de $\widehat{p}$. Dito de outro modo, poderemos afirmar com grande confiança que

$$\widehat{p} - \widehat{V} \leq p \leq \widehat{p} + \widehat{V},$$

isto é, com grande confiança poderemos fazer uma afirmação sobre o valor desconhecido p: p pertence ao intervalo $[\widehat{p}-\widehat{V}, \widehat{p}+\widehat{V}]$. Para que esta afirmação tenha algum interesse prático $\widehat{V}$ não deverá ser grande. Veremos mais à frente que para que tal aconteça não poderemos exagerar no grau confiança que impomos às afirmações anteriores.

O conhecimento da distribuição da estatística de interesse é assim de primeira importância no procedimento inferencial de que temos vindo a falar, e que abordaremos em capítulos futuros. Esse conhecimento permitirá, em particular, medir o grau de confiança que podemos ter nos resultados que obtemos a partir dessa estatística.

Neste parágrafo, para podermos ter uma ideia sobre a forma da sua distribuição, admitimos que possuíamos várias observações dessa mesma estatística, o que só foi possível extraindo outras tantas amostras, todas com a mesma dimensão, da população que pretendemos estudar. Numa situação prática, apenas uma amostra é recolhida, isto é, apenas uma observação da estatística de interesse é conhecida. A partir dessa observação nada podemos dizer sobre a distribuição da estatística.

É por isso fundamental desenvolver métodos matemáticos que nos permitam, a partir de outra informação associada às observações que realizamos, ter acesso, mesmo que de forma aproximada, à distribuição da estatística de interesse. Tais métodos são baseados na noção de probabilidade que abordaremos no próximo capítulo.

2.5 Bibliografia

Daniel, W.W. (2009). *Biostatistics: a foundation for analysis in the health sciences*, Wiley.

Dawson, B., Trapp, R.G. (2004). *Basic & Clinical Biostatistics*, McGraw-Hill.

Levy, P. (1999). *Sampling of Populations: methods and applications*, Wiley.

Moore, D.S., McCabe, G.P. (2006). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

Pestana, D.D., Velosa, S.F. (2002). *Introdução à Probabilidade e à Estatística*, Vol. I, Fundação Calouste Gulbenkian.

Vicente, P., Reis, E., Ferrão, F. (2001). *Sondagens: a amostragem como factor decisivo de qualidade*, Edições Sílabo.

3

Introdução à probabilidade

Experiência aleatória e acontecimentos aleatórios. Acontecimentos e conjuntos. Princípios clássico e frequencista para atribuição de probabilidade a um acontecimento aleatório. Propriedades da probabilidade. Independência de acontecimentos. Testes de diagnóstico.

3.1 Experiência e acontecimentos aleatórios

O lançamento de um dado é um caso particular duma classe mais vasta de experiências, a que chamamos **experiências aleatórias**, que gozam das seguintes propriedades:

- podem repetir-se, mesmo que hipoteticamente, nas mesmas condições, ou em condições muito semelhantes;
- o resultado observado em cada uma dessas repetições é um de entre um conjunto de resultados possíveis conhecidos antes de realizar a experiência;
- esse resultado é consequência dum conjunto de fatores que não podemos, na totalidade, controlar, e que atribuímos ao **acaso**.

Os exemplos seguintes de experiências aleatórias, incluem exemplos já considerados nos capítulos anteriores:

1. lançamento duma moeda de um euro ao ar e observação da face que fica voltada para cima;
2. lançamento duma moeda de um euro ao ar 100 vezes consecutivas e registo do número de vezes que ocorreu a face europeia;

3. lançamento dum dado e observação do número de pontos obtidos;
4. extração duma carta dum baralho e observação das suas caraterísticas;
5. registo do número de lançamentos dum dado necessários à obtenção, pela primeira vez, da face 6;
6. registo do tempo de duração duma lâmpada;
7. tempo que medeia a chegada de dois clientes consecutivos a um caixa de supermercado (ver pág. 26);
8. registo do peso de pacotes de açúcar empacotados por uma máquina (ver pág. 24);
9. número de chamadas que por minuto chegam a uma central telefónica (ver pág. 26);
10. proporção de indivíduos numa amostra aleatória simples de tamanho 20 que possuem determinada caraterística (neste exemplo o acaso está presente no processo de amostragem; ver pág. 63).

A cada uma destas experiências aleatórias podemos associar **acontecimentos aleatórios**, isto é, acontecimentos que podem, ou não, ocorrer dependendo do resultado da experiência em causa. Os acontecimentos aleatórios são normalmente representados por letras maiúsculas: $A, B, C, \ldots$ Relativamente a cada uma das experiências anteriores, são exemplos de acontecimentos aleatórios:

1. A="ocorrência da face portuguesa";
2. A="mais de 45 e menos de 55 ocorrências"; B="95 ou mais ocorrências";
3. A="saída de 6"; B="saída de número par";
4. A="saída de naipe de paus"; B="saída de ás";
5. A="menos de 3 lançamentos"; B="mais de 5 lançamentos";
6. A="duração superior a 200 horas";
7. A="menos de 1 minuto"; B="mais de meio minuto";
8. A="peso superior a 1010 gramas"; B="peso superior a 980 gramas e inferior a 1020 gramas";

9. A="mais de 5 chamadas";

10. A="proporção superior a 3/8 e inferior a 5/8".

No caso da experiência aleatória 3., se sai 2 no lançamento do dado o acontecimento B realiza-se enquanto que o acontecimento A não se realiza. Na experiência 8., se um pacote tem 1015 gramas realizam-se ambos os acontecimentos A e B.

No estudo que vamos fazer, o nosso objetivo não é o de prever o resultado particular duma experiência aleatória. O que pretendemos é quantificar a maior ou menor possibilidade que cada um dos acontecimentos aleatórios associados à experiência tem de se realizar ou ocorrer. Por outras palavras, pretendemos associar a cada acontecimento um número, número esse que traduzirá essa maior ou menor possibilidade de realização. A esse número chamaremos **probabilidade** do acontecimento em causa.

3.2 Acontecimentos e conjuntos

A cada uma das experiências aleatória que descrevemos no parágrafo anterior, podemos associar um conjunto, que denotaremos por Ω, constituído por todos os resultados possíveis da experiência aleatória. Por outras palavras, cada resultado particular da experiência aleatória é representado por um e um só elemento de Ω. A este conjunto Ω chamamos **espaço dos resultados**.

Relativamente aos exemplos anteriores, podemos tomar:

1. $\Omega = \{E, P\}$, onde E representa a saída da face europeia, e P a saída da face portuguesa; ou então $\Omega = \{0, 1\}$, onde 0 representa a saída da face europeia, e 1 a saída da face portuguesa;

2. $\Omega = \{0, 1, 2, \ldots, 100\}$, onde, por exemplo, o número 34 significa que nos 100 lançamento da moeda, a face europeia ocorreu 34 vezes;

3. $\Omega = \{1, 2, 3, 4, 5, 6\}$, onde, por exemplo, o número 2 representa a saída da face com dois pontos;

4. $\Omega = \{A_p, R_p, V_p, D_p, 10_p, \ldots, 2_p, A_o, R_o, \ldots\}$;

5. $\Omega = \{1, 2, 3, 4, \ldots\}$;

6. $\Omega = [0, +\infty[$;

7. $\Omega = [0, +\infty[$;

8. $\Omega = [0, +\infty[$;

9. $\Omega = \{0, 1, 2, 3, 4, \ldots\}$;

10. $\Omega = \{0, 1/20, 2/20, \ldots, 19/20, 1\}$.

Notemos agora que cada um dos acontecimentos aleatórios considerados no parágrafo anterior, pode ser representado pelo subconjunto de Ω cujos elementos são **favoráveis à realização desse acontecimento**:

1. $A = \{E\}$; ou $A = \{0\}$;

2. $A = \{46, 47, \ldots, 54\}$;

3. $A = \{6\}$; $B = \{2, 4, 6\}$;

4. $A = \{A_p, R_p, V_p, D_p, 10_p, \ldots, 2_p\}$;

5. $A = \{1, 2\}$; $B = \{6, 7, \ldots\}$;

6. $A =]200, +\infty[$;

7. $A = [0, 1[$; $B =]0.5, +\infty[$;

8. $A =]1010, +\infty[$; $B =]980, 1020[$;

9. $A = \{5, 6, \ldots\}$;

10. $A = \{8/20, \ldots, 12/20\}$.

Há acontecimentos aos quais damos nomes especiais:

- Os acontecimentos constituídos por um só elemento dizem-se **acontecimentos elementares**. Este é o caso dos acontecimentos A das experiências 1. e 3..

- Um acontecimento que se realiza independentemente do resultado da experiência aleatória diz-se **acontecimento certo**. No lançamento da moeda, o acontecimento "saída de uma das duas faces", é um acontecimento certo. O subconjunto de Ω que devemos associar a um acontecimento certo deve conter todos os possíveis resultados da experiência. Assim, o acontecimento certo é representado pelo próprio Ω.

- Por oposição ao acontecimento certo, o **acontecimento impossível** é um acontecimento que, independentemente do resultado da experiência aleatória, não se realiza. No lançamento de um dado vulgar, o acontecimento "saída de face com 7 pontos", é um acontecimento impossível. Como nenhum resultado da experiência aleatória é favorável ao acontecimento impossível, o subconjunto de Ω que lhe

devemos associar não deve possuir nenhum elemento. Por outras palavras, ao acontecimento impossível associamos o conjunto vazio que representamos pelo símbolo $\emptyset$.

As operações usuais entre conjuntos que a seguir recordamos, **complementação**, **interseção** e **reunião**, permitem exprimir ou construir acontecimentos a partir de outros acontecimentos.

Exemplo 3.2.1 Para ilustrar o que acabámos de dizer, consideremos a experiência aleatória do lançamento dum dado vulgar e tomemos os acontecimentos aleatórios:

$$A = \text{“saída de número par”} = \{2, 4, 6\},$$
$$B = \text{“saída de número inferior a 3”} = \{1, 2\},$$
$$C = \text{“saída de número par superior a 3 ”} = \{4, 6\}.$$

O acontecimento aleatório

$$\text{“saída de número ímpar”} = \{1, 3, 5\},$$

realiza-se quando o acontecimento A não se realiza, isto é, o conjunto dos resultados da experiência que lhe são favoráveis, não são favoráveis ao acontecimento A. Este acontecimento diz-se **acontecimento contrário de** A. O subconjunto de Ω que lhe associamos é o conjunto dos elementos de Ω que não pertencem a A. Um tal conjunto é denotado por A^c e diz-se **complementar de** A:

$$\{1, 3, 5\} = \{2, 4, 6\}^c = A^c.$$

Reparemos que o acontecimento contrário do acontecimento certo é o acontecimento impossível, e que o acontecimento contrário do acontecimento impossível é o acontecimento certo.

O acontecimento aleatório

$$\text{“saída de número par inferior a 3”} = \{2\},$$

realiza-se quando ambos os acontecimentos A e B se realizam. Por outras palavras, os resultados da experiência aleatória favoráveis ao acontecimento anterior, são favoráveis a A e a B simultaneamente. O subconjunto de Ω que lhe associamos é o conjunto dos elementos que pertencem a A e a B simultaneamente. Um tal conjunto é denotado por $A \cap B$ e diz-se **interseção dos conjuntos** A **e** B:

$$\{2\} = \{2, 4, 6\} \cap \{1, 2\} = A \cap B.$$

De forma análoga, o conjunto dos resultados da experiência aleatória favoráveis à realização de B e C é

$$B \cap C = \{1, 2\} \cap \{4, 6\} = \emptyset.$$

Não havendo resultados da experiência aleatória favoráveis à realização simultânea de B e C, o acontecimento $B \cap C$ é impossível. Os acontecimentos B e C dizem-se por isso **acontecimentos incompatíveis**.

Reparemos que um acontecimento e o seu contrário são sempre acontecimentos incompatíveis.

O acontecimento aleatório

$$\text{“saída de número par ou de número inferior a 3”} = \{1, 2, 4, 6\},$$

realiza-se quando pelo menos um dos acontecimentos A ou B se realiza. Os resultados da experiência aleatória favoráveis ao acontecimento anterior, são favoráveis a pelo menos um dos acontecimentos A ou B. O subconjunto de Ω que lhe associamos é o conjunto dos elementos que pertencem a pelo menos um dos conjuntos A ou B. Um tal conjunto é denotado por $A \cup B$ e diz-se **reunião dos conjuntos** A **e** B:

$$\{1, 2, 4, 6\} = \{2, 4, 6\} \cup \{1, 2\} = A \cup B.$$

Atendendo à correspondência que podemos estabelecer entre acontecimentos aleatórios e subconjuntos do espaço dos resultados, daqui para a frente simplificaremos a linguagem usando a designação de acontecimento aleatório quer se trate do acontecimento aleatório em si mesmo, quer se trate do subconjunto do espaço dos resultados que lhe podemos associar. Neste sentido, falaremos da interseção de acontecimentos aleatórios, e não da interseção dos subconjuntos que podemos associar a esses acontecimentos aleatórios. Mais exemplos dessa simplificação de linguagem são dados a seguir:

acontecimento que se realiza quando A não se realiza	$\rightarrow$	complementar de A	$\rightarrow$	A^c
acontecimento que se realiza quando A e B se realizam simultaneamente	$\rightarrow$	interseção de A e B	$\rightarrow$	$A \cap B$
acontecimento que se realiza quando pelo menos um dos acontecimentos A e B se realiza	$\rightarrow$	reunião de A e B	$\rightarrow$	$A \cup B$

3.3 Atribuição de probabilidade

Como referimos no §3.1, o objetivo principal do estudo duma experiência aleatória é o da atribuição de probabilidade aos acontecimentos aleatórios que lhe estão associados. **A probabilidade dum acontecimento** A, que denotamos por $\mathrm{P}(A)$, não é mais do que um número real, que vamos supor pertencer ao intervalo $[0, 1]$, que **traduz a maior ou menor possibilidade do acontecimento** A **ocorrer**.

Neste parágrafo apresentaremos dois princípios fundamentais para atribuir probabilidade aos acontecimentos aleatórios duma experiência aleatória. Tais princípios são conhecidos por **definição clássica de probabilidade** e **definição frequencista de probabilidade**. No próximo capítulo falaremos também da atribuição de probabilidade utilizando curvas densidade.

O exemplo seguinte ilustra as principais características de cada um dos princípios anteriores.

Exemplo 3.3.1 Suponhamos que uma moeda portuguesa de um euro é lançada 50 vezes, tendo-se obtido 45 vezes a face europeia e 5 vezes a face portuguesa. Se lançarmos a moeda uma vez mais, qual é a probabilidade de sair a face europeia? Esta probabilidade pode ser obtida a partir de duas perspetivas distintas. Se estamos convencidos que a moeda é equilibrada, isto é, se julgamos haver igual possibilidade de ocorrer cada uma das faces, a resposta poderá ser 0.5. No entanto, é-nos dito também que nos 50 lançamentos efetuados ocorreu a face europeia em 45 deles. Utilizando esta informação podemos pensar em estimar a probabilidade de sair a face europeia por $45/50 = 0.9$.

Como veremos de seguida, a primeira das respostas anteriores utiliza o conceito clássico de probabilidade. Para a sua aplicação, usámos apenas o facto da experiência em causa ter dois resultados possíveis que avaliámos como sendo igualmente prováveis. Os resultados obtidos em anteriores realizações da experiência aleatória não tiveram qualquer influência na resposta dada. Na segunda resposta tivemos apenas em conta tais resultados, possivelmente por pensarmos que os resultados obtidos nas realizações anteriores da experiência revelam que a hipótese da moeda ser equilibrada é pouco credível. Usámos, por isso, o conceito frequencista de probabilidade.

3.3.1 Definição clássica de probabilidade

O primeiro princípio para atribuição de probabilidade de que vamos falar, é conhecido como **definição clássica de probabilidade** ou **definição de probabilidade de Laplace**. Apesar de ter ficado associada ao matemático e astrónomo francês Pierre-Simon Laplace (1749-1827), a definição clássica de probabilidade era usada mais de um século antes do nascimento de Laplace.

A utilização desta definição é limitada ao caso em que o conjunto dos resultados possíveis da experiência aleatória é finito sendo esses resultados **equiprováveis**, isto é, igualmente prováveis. Com estes pressupostos é natural quantificar a maior ou menor possibilidade de realização de um acontecimento A através do número de resultados da experiência aleatória que são favoráveis a A.

Definição clássica de probabilidade:

Numa experiência aleatória com um número finito de resultados possíveis e equiprováveis, a probabilidade de um acontecimento A é dada pelo quociente entre os resultados favoráveis a A e o número total de resultados possíveis:

$$\mathrm{P}(A) = \frac{\text{número dos resultados favoráveis a } A}{\text{número de resultados possíveis}}.$$

Vejamos dois exemplos de aplicação da definição clássica de probabilidade.

Exemplo 3.3.2 No caso da extração de uma carta de um baralho de 52 cartas que supomos bem baralhadas, é natural admitir que cada carta tem igual possibilidade de ser escolhida. Assim,

$$\mathrm{P}(\text{“saída de paus”}) = \mathrm{P}(\{A_p, R_p, V_p, D_p, 10_p, \ldots, 2_p\}) = \frac{13}{52} = \frac{1}{4} = 0{,}25$$

e

$$\mathrm{P}(\text{“saída de ás”}) = \mathrm{P}(\{A_p, A_o, A_c, A_e\}) = \frac{4}{52} = \frac{1}{13} \approx 0{,}0769.$$

Exemplo 3.3.3 Voltemos à experiência aleatória do lançamento de um dado (ver Exemplo 3.2.1). Se tivermos boas razões para acreditar que o dado em questão é equilibrado (ou melhor, se não tivermos motivos para duvidar que ele seja equilibrado), a definição clássica de probabilidade pode ser utilizada. Nesse caso,

$$\mathrm{P}(\text{“saída da face 1”}) = \mathrm{P}(\{1\}) = \frac{1}{6} \approx 0{,}1667,$$

$$\mathrm{P}(\text{“saída de número inferior a 3”}) = \mathrm{P}(\{1, 2\}) = \frac{2}{6} \approx 0{,}3333$$

e

$$\mathrm{P}(\text{“saída de número par”}) = \mathrm{P}(\{2, 4, 6\}) = \frac{3}{6} = 0{,}5.$$

3.3.2 Frequência relativa e probabilidade

Uma das caraterísticas de uma experiência aleatória é, como já referimos, a possibilidade de ser repetida, mesmo que hipoteticamente, sempre nas mesmas condições. Ao repetirmos um determinado número de vezes uma experiência aleatória, podemos calcular a **frequência relativa** dum determinado acontecimento A, isto é, é possível calcular a proporção de ocorrências de A nas várias repetições da experiência. Por outras palavras, podemos calcular o quociente entre o número de vezes em que A ocorreu, a que chamamos **frequência absoluta do acontecimento** A, e o número de repetições da experiência aleatória:

$$\text{frequência relativa de } A = \frac{\text{número de ocorrências de } A}{\text{número de repetições}}.$$

Exemplo 3.3.4 Simulámos 10000 lançamentos dum dado equilibrado, tendo obtido as pontuações seguintes nos primeiros 500 lançamentos:

5, 4, 6, 5, 4, 6, 6, 2, 1, 6, 4, 5, 1, 3, 4, 3, 2, 1, 3, 1, 2, 3, 2, 1, 3, 2, 1, 6, 6, 5, 3, 5, 2, 3,
3, 6, 3, 2, 1, 3, 1, 5, 2, 2, 1, 5, 5, 2, 6, 1, 3, 1, 4, 4, 2, 1, 5, 2, 6, 5, 1, 3, 3, 3, 5, 5, 2, 1,
1, 3, 4, 2, 2, 5, 4, 2, 4, 2, 4, 4, 3, 1, 5, 6, 5, 6, 6, 4, 2, 6, 3, 3, 2, 5, 5, 6, 4, 1, 1, 5, 3, 4,
6, 4, 3, 4, 6, 1, 5, 4, 4, 1, 5, 2, 6, 3, 1, 6, 1, 3, 5, 3, 5, 1, 2, 3, 1, 6, 3, 1, 4, 6, 4, 4, 3, 6,
1, 3, 6, 5, 1, 3, 6, 5, 5, 5, 2, 5, 5, 2, 1, 4, 4, 5, 4, 6, 2, 4, 5, 5, 5, 2, 4, 2, 6, 6, 2, 1, 3, 2,
5, 3, 5, 5, 1, 3, 3, 2, 2, 2, 4, 3, 5, 1, 2, 2, 1, 3, 6, 5, 1, 5, 1, 5, 1, 6, 4, 2, 6, 1, 4, 5, 3, 3,
3, 4, 6, 6, 6, 1, 2, 3, 3, 6, 4, 5, 2, 4, 1, 2, 2, 2, 6, 3, 6, 6, 3, 4, 2, 3, 5, 6, 1, 2, 2, 4, 5, 1,
4, 5, 2, 6, 1, 5, 5, 4, 3, 6, 2, 4, 2, 4, 5, 1, 6, 5, 1, 2, 3, 2, 4, 2, 1, 5, 3, 3, 3, 1, 4, 1, 5, 5,
6, 6, 3, 5, 4, 5, 5, 5, 2, 6, 3, 1, 1, 2, 6, 1, 4, 3, 2, 2, 4, 3, 6, 6, 6, 3, 1, 3, 4, 6, 1, 3, 5, 4,
2, 3, 2, 6, 1, 4, 5, 4, 4, 5, 5, 4, 6, 3, 6, 2, 4, 3, 5, 4, 2, 4, 6, 3, 1, 4, 2, 1, 1, 6, 4, 2, 3, 6,
1, 3, 3, 6, 6, 1, 5, 5, 4, 4, 1, 3, 5, 4, 6, 3, 2, 1, 6, 2, 3, 6, 5, 5, 1, 5, 5, 5, 6, 1, 4, 1, 4, 1,
6, 4, 1, 4, 2, 4, 1, 3, 1, 6, 1, 6, 2, 2, 1, 2, 3, 4, 1, 1, 2, 2, 6, 6, 6, 5, 6, 4, 5, 4, 5, 5, 1, 6,
2, 2, 4, 3, 5, 4, 2, 5, 4, 3, 1, 4, 4, 3, 2, 5, 4, 3, 1, 3, 3, 1, 2, 3, 4, 1, 6, 3, 5, 6, 2, 6, 2, 5,
3, 6, 3, 5, 4, 6, 2, 5, 4, 6, 1, 5, 4, 5, 1, 4, 2, 4, 5, 3, 6, 3, 3, 6, 5, 1, 5, 6, 1, 6, 5, 4, 6, 1,
2, 4, 1, 3, 2, 4, 2, 3, 2, 6, 1, 3, 4, 2, 6, 2, 5, 6, 4, 1, 5, 2, 2, 4

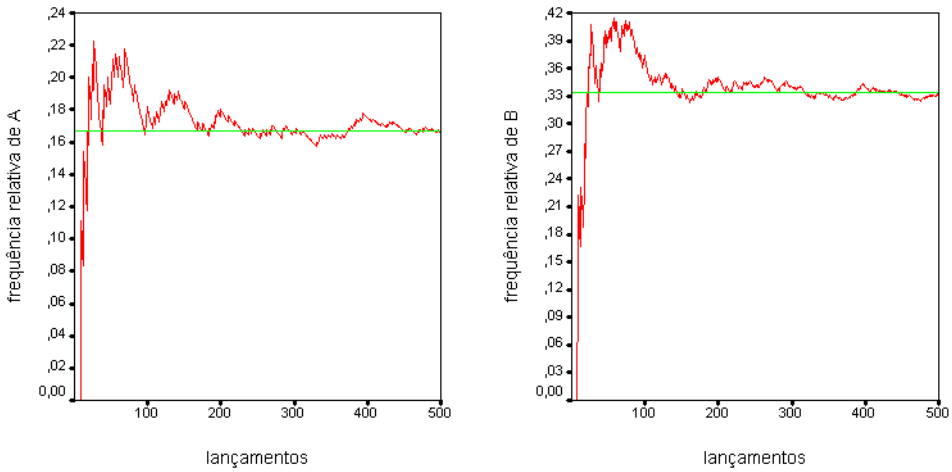
No quadro seguinte damos conta do número de ocorrências de cada uma das faces nos primeiros 100 e 1000 lançamentos do dado, bem como na totalidade dos 10000 lançamentos realizados:

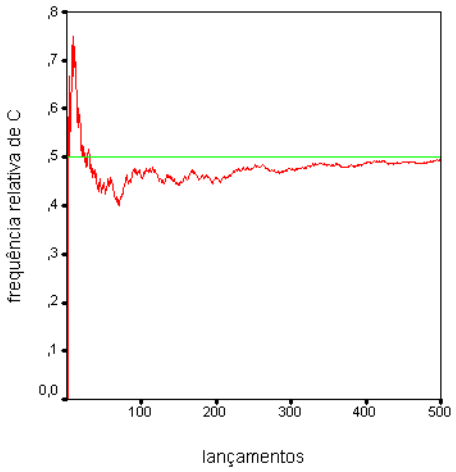
Faces \ Repetições	100	1000	10000
1	18	175	1722
2	19	164	1646
3	18	178	1661
4	13	157	1590
5	18	152	1769
6	14	174	1612

Para cada um dos acontecimentos $A = \{1\}$, $B = \{1, 2\}$ e $C = \{2, 4, 6\}$ considerados nos Exemplos 3.2.1 (pág. 73) e 3.3.3 (pág. 76), as suas frequências relativas são dadas no quadro seguinte:

Acontecimentos \ Repetições	100	1000	10000
A	0,18	0,175	0,1722
B	0,37	0,339	0,3368
C	0,46	0,495	0,4848

Para termos uma ideia mais clara sobre a evolução da frequência relativa de cada um dos acontecimentos anteriores à medida que o número de repetições da experiência aumenta, apresentamos os gráficos seguintes relativos às primeiras 500 observações realizadas:





Constatamos que à medida que repetimos a experiência mais e mais vezes, a frequência relativa de cada um dos acontecimentos anteriores aproxima-se da probabilidade desses acontecimentos que calculámos no Exemplo 3.3.3 (pág. 76).

O facto anterior, que verificámos ocorrer no exemplo anterior para os lançamentos que simulámos, ocorre também em outra qualquer sucessão de lançamentos. Mais geralmente, pode ser demonstrado matematicamente que para uma qualquer experiência aleatória, quando o número de repetições desta é muito elevado, a frequência relativa dum acontecimento aleatório aproxima-se, tanto quanto queiramos, da probabilidade desse acontecimento.

A este facto vamos nós chamar **definição frequencista de probabilidade**. Esta interpretação da noção de probabilidade é especialmente útil quando pouco conhecemos *a priori* sobre a experiência em causa, mas conhecemos os resultados obtidos na repetição da experiência aleatória, sempre nas mesmas condições, um grande número de vezes.

Definição frequencista de probabilidade:

A probabilidade de um acontecimento aleatório A, pode ser aproximada pela sua frequência relativa obtida pela repetição, um grande número de vezes, da experiência aleatória:

$$\mathrm{P}(A) \underset{n\approx\infty}{\approx} \text{frequência relativa de } A.$$

Exemplo 3.3.5 No lançamento de três dados equilibrados, 9 e 10 pontos podem ser obtidos de seis maneiras diferentes:

126	135	144	225	234	333
136	145	226	235	244	334

Por outro lado, as frequências absolutas desses acontecimentos indicam que a soma 9 ocorre menos vezes que a soma 10:

soma \ lançamentos	100	1000	10000	20000
9	12	109	1150	2296
10	10	147	1247	2529

À luz da interpretação frequencista de probabilidade como podem ser compatíveis os factos anteriores? Esta mesma questão foi colocada a Galileu Galilei (1564-1642) por volta de 1620, tendo este observado que a contagem dos casos favoráveis a cada uma das somas não está correta uma vez que os casos apresentados não têm todos a mesma possibilidade de ocorrerem. Por exemplo, a ocorrência de 333 tem seis vezes menos possibilidade de ocorrer que as configuração representadas por 126, uma vez que devemos ter em conta os dados em que esses números ocorrem. Assim, escondidos sob a designação 126 estão 6 casos igualmente prováveis (126, 162, 216, 261, 612, 621), o mesmo acontecendo em todas as situações anteriores com três números diferentes. Nos casos em que em dois dados ocorre o mesmo número e no dado restante ocorre um número diferente, temos 3 casos igualmente prováveis:

soma 9	casos igual.prov.	soma 10	casos igual.prov.
126	6	136	6
135	6	145	6
144	3	226	3
225	3	235	6
234	6	244	3
333	1	334	3
total	25	total	27

Usando a definição clássica podemos então calcular a probabilidade da ocorrência de "soma 9" e de "soma 10":

$$\mathrm{P}(\text{“soma 9”}) = \frac{25}{216} \approx 0{,}1157$$

e

$$\mathrm{P}(\text{“soma 10”}) = \frac{27}{216} = 0{,}125.$$

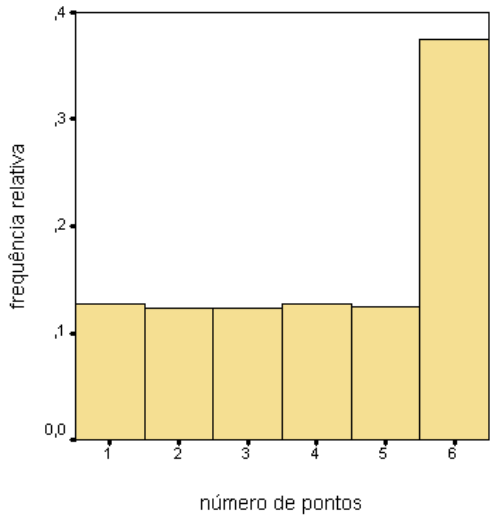
Apesar da frequência relativa ser tomada, para todos os efeitos, como probabilidade exata do acontecimento em causa, não nos devemos esquecer que ela não é mais do que uma aproximação para a verdadeira probabilidade (desconhecida) do acontecimento.

Exemplo 3.3.6 Suponhamos que em sucessivos lançamentos de um dado obtemos as seguintes frequências relativas para cada uma das faces:

Faces \ Repetições	100	1000	10000
1	0,10	0,118	0,1268
2	0,08	0,116	0,1228
3	0,08	0,125	0,1231
4	0,18	0,125	0,1278
5	0,11	0,132	0,1247
6	0,45	0,384	0,3748

Atendendo às frequências relativas anteriores, que representamos no gráfico seguinte para 10000 repetições da experiência, fica claro que a utilização da definição clássica de probabilidade terá aqui pouco sentido. É neste caso mais apropriado utilizar a interpretação frequencista de probabilidade e tomar para probabilidade de cada face a sua frequência relativa obtida em 10000 lançamentos do dado:

$$\begin{aligned} &\mathrm{P}(\{1\}) = 0{,}1268, \quad \mathrm{P}(\{2\}) = 0{,}1228, \quad \mathrm{P}(\{3\}) = 0{,}1231, \\ &\mathrm{P}(\{4\}) = 0{,}1278, \quad \mathrm{P}(\{5\}) = 0{,}1247, \quad \mathrm{P}(\{6\}) = 0{,}3748. \end{aligned}$$



Para cada um dos acontecimentos considerados nos parágrafos anteriores, as suas probabilidades são assim dadas por

$$\mathrm{P}(\{1\}) = 0{,}1268,$$

$$\mathrm{P}(\{1,2\}) = 0{,}1268 + 0{,}1228 = 0{,}2496$$

e

$$\mathrm{P}(\{2,4,6\}) = 0{,}1228 + 0{,}1278 + 0{,}3748 = 0{,}6254.$$

No cálculo destas probabilidades usámos novamente a interpretação frequencista de probabilidade e o facto da frequência relativa dos acontecimentos $\{1,2\}$ e $\{2,4,6\}$, ser a soma das frequências relativas dos acontecimentos $\{1\}$ e $\{2\}$, e $\{2\}$, $\{4\}$ e $\{6\}$, respetivamente.

Exemplo 3.3.7 No caso do lançamento dos três dados considerados no Exemplo 3.3.5 (pág. 80), havendo razões para admitir que algum dos dados era viciado, seria mais apropriado usar a definição frequencista para obter aproximações para as probabilidades da ocorrência da "soma 9" e da "soma 10". Usando os resultados obtidos em 20000 lançamentos dos três dados obteríamos

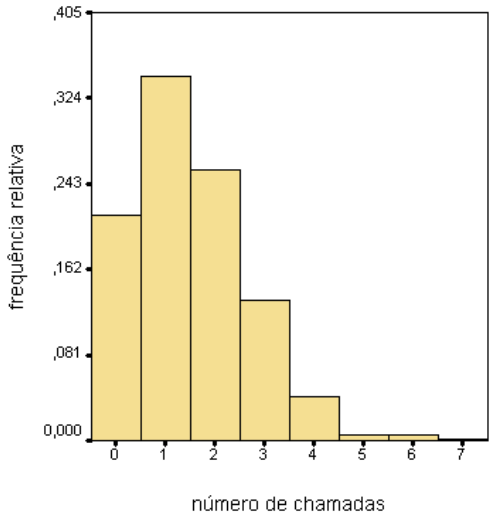
$$\mathrm{P}(\text{“soma 9”}) \approx \frac{2296}{20000} = 0{,}1148$$

e

$$\mathrm{P}(\text{“soma 10”}) \approx \frac{2529}{20000} = 0{,}12645.$$

Sendo estes valores muito próximos dos que calculámos pela definição clássica, é razoável pensar que os dados são efetivamente equilibrados.

Exemplo 3.3.8 Retomemos um exemplo anterior (Exemplo 1.2.8, pág. 26), relativo ao número de chamadas telefónicas por minuto que chegam a uma central telefónica dum determinado serviço público, cuja distribuição é descrita pelo histograma de frequências relativas seguinte:



As frequências relativas observadas são as seguintes:

número de chamadas/minuto

		Frequency	Relative Frequency
Valid	0	132	,2136
	1	213	,3447
	2	158	,2557
	3	82	,1327
	4	26	,0421
	5	3	,0049
	6	3	,0049
	7	1	,0016
	Total	618	1,0000

Tomando para probabilidade dum acontecimento a sua frequência relativa calculada a partir do número de chamadas verificadas nos 618 minutos observados, a probabilidade de, na central telefónica observada, ocorrerem mais que 5 chamadas num minuto é assim igual a

$$\mathrm{P}(\{6,7,8,\ldots\}) \approx 0{,}0049 + 0{,}0016 + 0{,}0000 + \ldots = 0{,}0065.$$

3.4 Propriedades da probabilidade

Nos parágrafos anteriores estudámos duas maneiras de atribuir probabilidade aos acontecimentos de uma experiência aleatória. Dito de outro modo, estudámos diferentes formas de modelar matematicamente uma experiência aleatória. Dizemos então que obtivemos um **modelo probabilístico** para a experiência aleatória em estudo. Como vimos, esse modelo é constituído pelo espaço dos resultados Ω, pela família de todos os acontecimentos aleatórios associados à experiência aleatória, e pela probabilidade P que a cada acontecimento A associa a sua probabilidade $\mathrm{P}(A)$.

Para qualquer uma das formas que estudámos de atribuir probabilidade aos acontecimentos duma experiência aleatória, a probabilidade P satisfaz algumas propriedades das quais realçamos as que constam do quadro seguinte.

As duas primeiras propriedades não levantam qualquer problema. São trivialmente verificadas por ambas as definições de probabilidade. Relativamente à terceira propriedade, pensemos, por exemplo, no caso da definição clássica. Se a experiência tem n resultados possíveis sendo m deles favoráveis a A, então os restantes $n - m$ resultados são contrários a A, isto é, são favoráveis a A^c. Assim

$$\mathrm{P}(A^c) = \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - \mathrm{P}(A).$$

P.1) *A probabilidade de qualquer acontecimento A é um número real maior ou igual a zero e menor ou igual a 1*

$$0 \leq \mathrm{P}(A) \leq 1.$$

P.2) *A probabilidade do acontecimento certo é igual a* 1:

$$\mathrm{P}(\Omega) = 1.$$

P.3) *A probabilidade do acontecimento contrário do acontecimento A é dada por:*

$$\mathrm{P}(A^c) = 1 - \mathrm{P}(A).$$

P.4) *A probabilidade do acontecimento impossível é igual a zero:*

$$\mathrm{P}(\emptyset) = 0.$$

P.5) *A probabilidade da reunião de dois acontecimentos A e B, é dada por:*

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

A propriedade P.4 é verificada por ambas as definições. Reparemos, no entanto, que se trata duma propriedade que não é independente das três primeiras já enunciadas. Qualquer forma de atribuir probabilidade aos acontecimentos duma experiência aleatória que satisfaça as três primeiras propriedades, satisfaz necessariamente esta quarta propriedade. Para justificar esta afirmação, basta ter em conta P.1 e P.2 e o facto do acontecimento impossível ser contrário ao acontecimento certo. Com efeito,

$$\mathrm{P}(\emptyset) = \mathrm{P}(\Omega^c) = 1 - \mathrm{P}(\Omega) = 1 - 1 = 0.$$

Relativamente à propriedade P.5, vejamos o que se passa com a interpretação frequencista de probabilidade. Suponhamos que a experiência aleatória é repetida n vezes, tendo ocorrido A em p repetições, B em q repetições e $A \cap B$ em r repetições. Significa isto que $p = p' + r$ e $q = q' + r$ onde p' representa o número de repetições da experiência em que ocorreu A mas não ocorreu B e q' representa o número de repetições da experiência em que ocorreu B mas não ocorreu A. Atendendo a que $p' + q' + r$ é o número de repetições da experiência em que $A \cup B$ ocorreu, então

$$\mathrm{P}(A \cup B) \quad = \quad \frac{p' + q' + r}{n} = \frac{p' + r + q' + r - r}{n}$$

$$\begin{aligned} &= \frac{p'+r}{n} + \frac{q'+r}{n} - \frac{r}{n} \\ &= \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B). \end{aligned}$$

Atendendo às propriedades P.4 e P.5 podemos ainda concluir que

P.5′) *A probabilidade da reunião de dois acontecimentos incompatíveis A e B, é igual à soma das suas probabilidade:*

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B), \text{ se } A \cap B = \emptyset.$$

Poderíamos sem grande esforço enunciar outras propriedades comuns às probabilidades definidas no §3.3 e que são verificadas por toda e qualquer forma de atribuir probabilidade a acontecimentos de uma experiência aleatória que satisfaça as propriedades anteriores. O facto de realçarmos estas põe em relevo a sua importância.

Até aqui estudámos duas formas de atribuir probabilidade aos acontecimentos duma experiência aleatória. Terminamos este parágrafo notando que as propriedades da probabilidade, conjuntamente com a observação da experiência aleatória, podem também ser usadas para atingirmos esse objetivo. Este facto é ilustrado no exemplo seguinte.

Exemplo 3.4.1 Atendendo às frequências relativas obtidas para cada uma das faces do dado Exemplo 3.3.6 (pág. 81), é perfeitamente razoável conjeturar que as faces 1, 2, 3, 4 e 5, têm igual probabilidade de ocorrer, e que a face 6 tem três vezes mais probabilidade de ocorrer que cada uma das outras:

$$\mathrm{P}(\{1\}) = \mathrm{P}(\{2\}) = \mathrm{P}(\{3\}) = \mathrm{P}(\{4\}) = \mathrm{P}(\{5\})$$

e

$$\mathrm{P}(\{6\}) = 3\mathrm{P}(\{1\}).$$

Tendo em conta P.2 e P.5′, sabemos também que

$$\mathrm{P}(\{1\}) + \mathrm{P}(\{2\}) + \mathrm{P}(\{3\}) + \mathrm{P}(\{4\}) + \mathrm{P}(\{5\}) + \mathrm{P}(\{6\}) = 1 \quad \text{(porquê?)}.$$

Concluímos então que

$$5\mathrm{P}(\{1\}) + 3\mathrm{P}(\{1\}) = 1,$$

ou seja,

$$\mathrm{P}(\{1\}) = \frac{1}{8}.$$

Assim

$$\mathrm{P}(\{1\}) = \mathrm{P}(\{2\}) = \mathrm{P}(\{3\}) = \mathrm{P}(\{4\}) = \mathrm{P}(\{5\}) = \frac{1}{8}$$

e

$$\mathrm{P}(\{6\}) = \frac{3}{8}.$$

Tal como fizemos atrás, calculemos agora a probabilidade dos acontecimentos $\{1\}$, $\{1,2\}$ e $\{2,4,6\}$. Usando a propriedade P.5′ relativa à probabilidade da reunião de acontecimentos incompatíveis obtemos:

$$\mathrm{P}(\{1\}) = \frac{1}{8} = 0{,}125,$$

$$\mathrm{P}(\{1,2\}) = \mathrm{P}(\{1\}) + \mathrm{P}(\{2\}) = \frac{1}{8} + \frac{1}{8} = \frac{2}{8} = \frac{1}{4} = 0{,}25$$

e

$$\mathrm{P}(\{2,4,6\}) = \mathrm{P}(\{2\}) + \mathrm{P}(\{4\}) + \mathrm{P}(\{6\}) = \frac{1}{8} + \frac{1}{8} + \frac{3}{8} = \frac{5}{8} = 0{,}625.$$

Apesar da modelização que agora fizemos da experiência aleatória ter sido distinta da do parágrafo anterior, o que se reflete nas diferentes probabilidades encontradas para os acontecimentos anteriores, notemos que tais probabilidades são próximas das então obtidas. Este facto reforça a conjetura em que baseámos a presente abordagem.

3.5 Probabilidade condicionada e independência de acontecimentos

A propriedade P.5′ anterior, dá-nos uma regra para calcular a probabilidade da reunião de dois acontecimentos exclusivamente a partir da probabilidade de cada um deles. Para aplicarmos essa regra é essencial que os acontecimentos em causa sejam incompatíveis.

Como vamos ver a seguir, há também uma regra que permite calcular a probabilidade da interseção de dois acontecimentos A e B, a partir exclusivamente da probabilidade de cada um deles. Para a podermos aplicar é necessário que a ocorrência, ou não ocorrência, de qualquer um dos acontecimentos não afete a probabilidade de realização do outro. Quando isto acontece, dizemos que os acontecimentos A e B são **independentes**.

Atendendo à propriedade P.3, se a ocorrência, ou não ocorrência, do acontecimento A não afeta a probabilidade de realização do acontecimento B, também não afeta a probabilidade de realização do acontecimento contrário B^c. Quer isto dizer, que se A e B são acontecimentos independentes, também A e B^c, A^c e B, e A^c e B^c, são pares de acontecimentos independentes.

Para formalizarmos esta noção de independência de acontecimentos vamos lançar mão da noção de **probabilidade condicionada** do acontecimento A por um acontecimento B que vamos denotar por $\mathrm{P}(A|B)$ para a distinguir da noção de probabilidade do acontecimento A. Uma forma simples de interpretar a probabilidade $\mathrm{P}(A|B)$ é pensar que esta representa a probabilidade do acontecimento A após termos conhecimento de que o acontecimento B se realizou, enquanto que $\mathrm{P}(A)$ representa a probabilidade de A ser termos informação sobre a realização, ou não, do acontecimento B.

Fixemos a nossa atenção no caso em que estamos a utilizar a definição clássica de probabilidade. Neste caso será natural tomar para probabilidade de A condicionada por B o quociente

$$\mathrm{P}(A|B) = \frac{\text{número de resultados favoráveis a } A \cap B}{\text{número de resultados favoráveis a } B},$$

uma vez que, como sabemos que B se realizou, o número de resultados possíveis da experiência reduz-se aos resultados que são favoráveis a B e o número de resultados favoráveis a A não é agora mais do que o número de resultados favoráveis a $A \cap B$.

Reescrevendo o quociente anterior na forma

$$\mathrm{P}(A|B) = \frac{\dfrac{\text{número de resultados favoráveis a } A \cap B}{\text{número de resultados possíveis}}}{\dfrac{\text{número de resultados favoráveis a } B}{\text{número de resultados possíveis}}},$$

verificamos que o numerador não é mais do que a probabilidade de $A \cap B$ enquanto que o denominador é a probabilidade de B. Isto leva-nos à definição seguinte de probabilidade condicionada válida para uma qualquer forma de atribuir probabilidade aos acontecimentos duma experiência aleatória.

Definição de probabilidade condicionada

Se B é um acontecimento com $\mathrm{P}(B) > 0$, a probabilidade condicionada do acontecimento A pelo acontecimento B (ou probabilidade de A sabendo B) é dada por

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

Da fórmula anterior resulta a seguinte **regra da multiplicação das probabilidades**:

P.6) *Para quaisquer acontecimentos A e B com probabilidades positivas vale a igualdade*

$$P(A \cap B) = \mathrm{P}(A|B)\mathrm{P}(B).$$

Na posse da noção de probabilidade condicionada podemos então precisar a noção de independência de dois acontecimentos aleatórios A e B. Diremos que os acontecimentos A e B com probabilidades positivas são **independentes** se

$$\mathrm{P}(A|B) = \mathrm{P}(A).$$

Nestas circunstâncias é fácil verificar que também vale a igualdade

$$\mathrm{P}(B|A) = \mathrm{P}(B).$$

Tendo em conta a propriedade P.6 estamos agora em condições de estabelecer a regra já anunciada para o cálculo da probabilidade da interseção de dois acontecimentos aleatórios a partir exclusivamente da probabilidade de cada um deles.

P.6′) *A probabilidade da interseção de dois acontecimentos independentes A e B, é igual ao produto das suas probabilidade:*

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B).$$

Reparemos que se A e B verificam a igualdade $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$, então A e B são acontecimentos independentes. Com efeito,

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A)\mathrm{P}(B)}{\mathrm{P}(B)} = \mathrm{P}(A).$$

A igualdade expressa na propriedade P.6′ dá-nos assim uma caraterização da independência entre os acontecimentos A e B. Por esta razão, a mesma pode ser usada para definir a independência entre dois acontecimentos.

Exemplo 3.5.1 Numa esfera de extração de números da lotaria foram colocadas 20 bolas idênticas (exceto na cor) numeradas de 1 a 20, sendo as 10 primeiras vermelhas e as restantes 10 azuis. Considere a experiência aleatória que consiste na extração de uma bola da esfera e anotação do seu número, e os acontecimentos aleatórios:

$A =$ "saída de bola vermelha"
$B =$ "saída de bola com múltiplo de 4"
$C =$ "saída de bola com número par".

O espaço dos resultados da experiência aleatória é

$$\Omega = \{1, 2, \ldots, 20\},$$

e os acontecimentos A, B e C são dados por

$$A = \{1, 2, \ldots, 10\},$$

$$B = \{4, 8, 12, 16, 20\}$$

e

$$C = \{2, 4, \ldots, 18, 20\}.$$

Tendo todas as bolas igual possibilidade de serem extraídas, usamos a definição clássica de probabilidade para obter a probabilidade de ocorrência de cada um dos acontecimentos A, B e C:

$$\mathrm{P}(A) = \frac{10}{20} = \frac{1}{2},$$

$$\mathrm{P}(B) = \frac{5}{20} = \frac{1}{4}$$

e

$$\mathrm{P}(C) = \frac{10}{20} = \frac{1}{2}.$$

Admitamos agora que ocorreu o acontecimento C, isto é, a bola que saiu tem um número par. Será que este facto altera a probabilidade de ocorrência de algum dos acontecimentos A ou B? Dos dez resultados agora possíveis, cinco deles são favoráveis a A e também cinco deles são favoráveis a B. Quer isto dizer que:

$$\mathrm{P}(A|C) = \frac{5}{10} = \frac{1}{2}$$

e

$$\mathrm{P}(B|C) = \frac{5}{10} = \frac{1}{2}.$$

Verificamos que A é tão provável como antes, enquanto que B é agora mais provável que antes. Assim, A e C são acontecimento independentes, enquanto que B e C não são acontecimentos independentes.

Reparemos que apenas no primeiro dos casos anteriores, a probabilidade da interseção dos acontecimentos em causa, é igual ao produto das probabilidades respetivas. Com efeito,

$$A \cap C = \{2, 4, 6, 8, 10\}$$

e

$$P(A \cap C) = \frac{5}{20} = \frac{1}{4} = \mathrm{P}(A)\mathrm{P}(C),$$

enquanto que,

$$B \cap C = \{4, 8, 12, 16, 20\}$$

e

$$\mathrm{P}(B \cap C) = \frac{5}{20} = \frac{1}{4} \neq \mathrm{P}(B)\mathrm{P}(C).$$

Suponhamos agora que ocorreu o acontecimento A, isto é, a bola que saiu é vermelha, mas que não conseguimos ver o seu número. Como dos dez resultados agora possíveis, apenas dois são favoráveis a B, o acontecimento B é agora menos provável que antes. A e B não são, por isso, acontecimentos independentes. Mais uma vez, reparemos que a probabilidade da interseção não é igual ao produto das probabilidades. Com efeito,

$$A \cap B = \{4, 8\}$$

e

$$P(A \cap B) = \frac{2}{20} = \frac{1}{10} \neq \frac{1}{8} = \mathrm{P}(A)\mathrm{P}(B).$$

A noção de independência está intimamente relacionada com a primeira das propriedades que enunciámos das experiências aleatórias. Ao dizermos que uma experiência aleatória pode repetir-se nas mesmas condições, estamos implicitamente a dizer que o resultado de uma qualquer das repetições não influencia o resultado de qualquer outra. Os acontecimentos aleatórios associados a cada uma das repetições da experiência são, por isso, independentes. Esta situação é ilustrada no exemplo seguinte.

Exemplo 3.5.2 Uma moeda equilibrada é lançada duas vezes ao ar e é registada a face que fica voltada para cima. Consideremos os acontecimentos:

$A =$ "saída de face portuguesa no 1º lançamento"
$B =$ "saída de face portuguesa no 2º lançamento"

Atendendo a que a ocorrência de A não afeta a probabilidade de ocorrência de B, nem a ocorrência de B afeta a probabilidade de ocorrência de A, estes acontecimentos são independentes. Reparemos uma vez mais que neste caso a probabilidade da interseção $A \cap B$ é igual ao produto das probabilidades de A e de B. Com efeito, neste caso

$$\Omega = \{PP, PE, EP, EE\},$$

$$A = \{PP, PE\},$$

$$B = \{PP, EP\},$$

$$A \cap B = \{PP\},$$

e, da definição clássica de probabilidade, vale a igualdade

$$\mathrm{P}(A \cap B) = \frac{1}{4} = \frac{1}{2}\frac{1}{2} = \mathrm{P}(A)\mathrm{P}(B).$$

Terminamos este parágrafo com um exemplo que reforça a importância da noção de independência para o cálculo da probabilidade de acontecimentos associados a uma experiência aleatória.

Exemplo 3.5.3 Quando uma máquina está a funcionar adequadamente, apenas 0.1% das peças que produz apresentam defeito por razões várias que não podem na totalidade ser controladas. Admitamos que em dois momentos, razoavelmente afastados no tempo, decidimos observar duas peças que acabaram de ser produzidas pela máquina, e que pretendemos saber qual é a probabilidade de nenhuma das peças ser defeituosa.

Neste caso, o conjunto dos resultados da experiência é

$$\Omega = \{00, 01, 10, 11\},$$

onde, por exemplo, 01 significa que a primeira peça observada não é defeituosa mas que a segunda o é. Estamos interessado na probabilidade do acontecimento

$$A = \{00\}.$$

Como os resultados da experiência não são igualmente prováveis não podemos recorrer à definição clássica para calcular a probabilidade de A. Também não temos informação suficiente para usar a definição frequencista. No entanto, tendo em conta que

$$A = A_1 \cap A_2,$$

onde

$$A_1 = \text{“peça defeituosa na primeira observação”},$$

$$A_2 = \text{“peça defeituosa na segunda observação”},$$

e que é razoável admitir que A_1 e A_2 são acontecimentos independentes, uma vez que as duas observações foram realizadas em momentos afastados no tempo, então

$$\mathrm{P}(A) = \mathrm{P}(A_1 \cap A_2) = \mathrm{P}(A_1)\mathrm{P}(A_2) = 0{,}999 \times 0{,}999 = 0{,}99801.$$

3.6 Testes de diagnóstico

Os testes de diagnóstico são procedimentos em que os indivíduos sujeitos a tais testes são classificados como saudáveis ou como tendo determinada doença (teste TOFG (diabetes), teste VIH (SIDA), teste COVID-19, etc). Os testes de diagnóstico são, como tal, imperfeitos, podendo indivíduos saudáveis ser classificados como doentes (ditos falsos positivos) e indivíduos doentes ser classificados como não-doentes (ditos falsos

negativos). Um teste de diagnóstico pode ser avaliado através da sua **sensibilidade** e da sua **especificidade**, definidas pelas seguintes probabilidades condicionais:

$$\text{sensibilidade} = \text{P(teste positivo}|\text{doença presente)} = \text{P}(T^+|D^+)$$

$$\text{especificidade} = \text{P(teste negativo}|\text{doença ausente)} = \text{P}(T^-|D^-).$$

Um teste com grande sensibilidade identificará os doentes com grande precisão, enquanto que um teste com grande especificidade terá grande precisão na identificação dos não-doentes.

Na prática, a sensibilidade e a especificidade de um teste de diagnóstico podem ser estimadas lançando mão da definição frequencista de probabilidade. Para tal, o teste de diagnóstico é aplicado a dois grupos de indivíduos, um deles constituído por indivíduos que sabemos doentes (no caso da sensibilidade), e o outro constituído por indivíduos que sabemos não-doentes (no caso da especificidade). Sendo A e C o número de indivíduos efetivamente doentes classificados como doentes (verdadeiros positivos) e como não-doentes (falsos negativos), respetivamente, a sensibilidade do teste de diagnóstico pode assim ser aproximada por

$$\text{sensibilidade} \approx \frac{\text{A}}{\text{A} + \text{C}}.$$

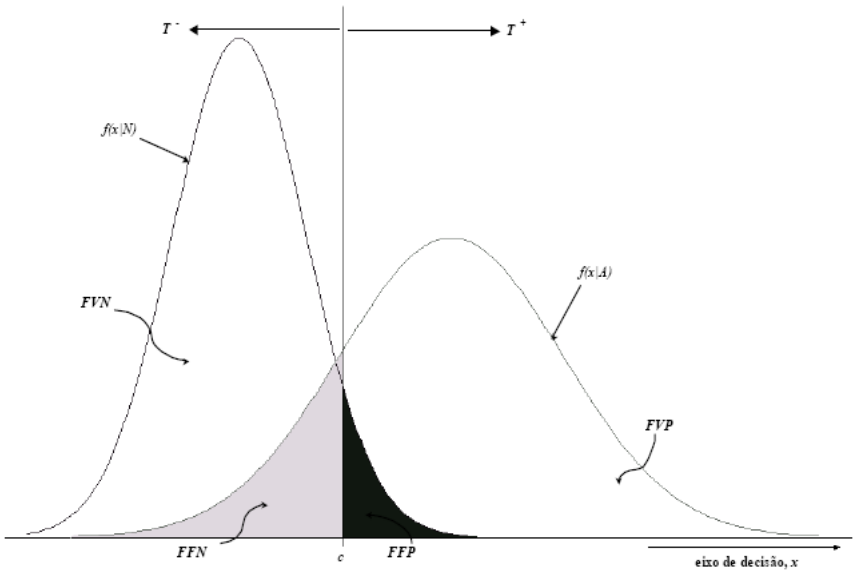
Por outro lado, sendo B e D o número de indivíduos saudáveis que são classificados como doentes (falsos positivos) e como não-doentes (verdadeiros negativos), respectivamente, a especificidade do teste de diagnóstico pode assim ser aproximada por

$$\text{especificidade} \approx \frac{\text{D}}{\text{B} + \text{D}}.$$

Teste	Doença	
	Presente D^+	Ausente D^-
Positivo T^+	Verdadeiros Positivos (A)	Falsos Positivos (B)
Negativo T^-	Falsos Negativos (C)	Verdadeiros Negativos (D)
Total	Doentes (A+C)	Não Doentes (B+D)

Em geral os testes de diagnóstico passam pela medição no indivíduo observado de determinada resposta quantitativa X, sendo o indivíduo classificado como saudável se a variável X não for superior a determinado limiar c, ou como doente, em caso contrário.

Assim, se $X \leq c$ o indivíduo é classificado como não-doente, e se $X > c$ o indivíduo é classificado como doente. Um teste será tanto melhor quanto maior discriminação induzir entre as populações de doentes e saudáveis.



Exemplo 3.6.1 Um grupo de investigação médica deseja avaliar um teste de diagnóstico para a doença de Alzheimer. O teste foi aplicado a amostras aleatórias de 450 pacientes com doença de Alzheimer e de 500 pacientes sem sintomas da doença. As duas amostras foram extraídas de populações de pacientes com 65 anos ou mais. Os resultados observados estão resumidos no quadro seguinte ([1]):

	Doença de Alzheimer	
Teste	Sim	Não
Positivo	436	5
Negativo	14	495
Total	450	500

Estimativas para a sensibilidade e a especificidade do teste de diagnóstico anterior são dadas por:

$$\text{sensibilidade} \approx \frac{436}{450} = 0{,}9689,$$

e

$$\text{especificidade} \approx \frac{495}{500} = 0{,}99.$$

[1] Fonte: Daniel, 2009, p. 82.

Duas importantes caraterísticas de um teste de diagnóstico são os seus **valor preditivo positivo** (VPP) e **valor preditivo negativo** (VPN):

$$\text{VPP} = \text{P(doença presente|teste positivo)} = \text{P}(D^+|T^+)$$

$$\text{VPN} = \text{P(doença ausente|teste negativo)} = \text{P}(D^-|T^-).$$

Estes valores preditivos são fundamentais para determinar o valor do teste de diagnóstico num indivíduo específico. Contrariamente à sensibilidade e à especificidade, que são atributos intrínsecos do teste de diagnóstico, veremos a seguir que os valores preditivos positivo e negativo dependem da sensibilidade e da especificidade do teste, mas também da prevalência, ou taxa de incidência, da doença na população.

Para calcularmos os valores preditivos de um teste de diagnóstico, vamos lançar mão de dois resultados importantes das probabilidades, conhecidos por Teorema da probabilidade total e Teorema de Bayes, que enunciamos a seguir.

Teorema da probabilidade total:

Se A_1 e A_2 são acontecimentos incompatíveis com $A_1 \cup A_2 = \Omega$, então a probabilidade de um acontecimento B pode ser escrita na forma

$$\text{P}(B) = \text{P}(B|A_1)\text{P}(A_1) + \text{P}(B|A_2)\text{P}(A_2).$$

Teorema de Bayes:

Dados dois acontecimentos A e B, e as probabilidades $\text{P}(A)$, $\text{P}(B)$ e $\text{P}(A|B)$, então

$$\text{P}(B|A) = \frac{\text{P}(A|B)\text{P}(B)}{\text{P}(A)}.$$

Representando por Se e Es, a **sensibilidade** e a **especificidade** do teste, e por Pr a **prevalência** da doença na população em causa, os dois resultados anteriores permitem concluir que o **valor preditivo positivo** pode ser obtido a partir da fórmula

$$\begin{aligned} \text{VPP} &= \text{P}(D^+|T^+) \\ &= \frac{\text{P}(T^+|D^+)\text{P}(D^+)}{P(T^+)} \end{aligned}$$

$$\begin{aligned} &= \frac{\mathrm{P}(T^+|D^+)\mathrm{P}(D^+)}{\mathrm{P}(T^+|D^+)\mathrm{P}(D^+) + \mathrm{P}(T^+|D^-)\mathrm{P}(D^-)} \\ &= \frac{\mathrm{P}(T^+|D^+)\mathrm{P}(D^+)}{\mathrm{P}(T^+|D^+)\mathrm{P}(D^+) + (1-\mathrm{P}(T^-|D^-))(1-\mathrm{P}(D^+))} \\ &= \frac{\mathrm{Se}\times\mathrm{Pr}}{\mathrm{Se}\times\mathrm{Pr} + (1-\mathrm{Es})\times(1-\mathrm{Pr})}, \end{aligned}$$

e o **valor preditivo negativo** pode ser obtido a partir da fórmula

$$\begin{aligned} \mathrm{VPN} &= \mathrm{P}(D^-|T^-) \\ &= \frac{\mathrm{P}(T^-|D^-)\mathrm{P}(D^-)}{\mathrm{P}(T^-)} \\ &= \frac{\mathrm{P}(T^-|D^-)(1-\mathrm{P}(D^+))}{\mathrm{P}(T^-|D^+)\mathrm{P}(D^+) + \mathrm{P}(T^-|D^-)\mathrm{P}(D^-)} \\ &= \frac{\mathrm{P}(T^-|D^-)(1-\mathrm{P}(D^+))}{(1-\mathrm{P}(T^+|D^+))\mathrm{P}(D^+) + \mathrm{P}(T^-|D^-)(1-\mathrm{P}(D^+))} \\ &= \frac{\mathrm{Es}\times(1-\mathrm{Pr})}{(1-\mathrm{Se})\times\mathrm{Pr} + \mathrm{Es}\times(1-\mathrm{Pr})}. \end{aligned}$$

Exemplo 3.6.1 (cont.) De acordo com os dados da *Alzheimer Europe* relativos a 2013 ([2]), 178925 indivíduos com mais de 65 anos eram portadores da doença de Alzheimer em Portugal, o que significa que a prevalência da doença na população com mais de 65 anos era de aproximadamente 8,9%. Retomando os dados anteriormente considerados, e admitindo que tal prevalência se mantém na atualidade, os valores preditivos positivo e negativo do teste de diagnóstico para a doença de Alzheimer são dados por:

$$\mathrm{VPP} = \frac{0{,}9689\times 0{,}089}{0{,}9689\times 0{,}089 + (1-0{,}99)\times(1-0{,}089)} = 0{,}904,$$

e

$$\mathrm{VPN} = \frac{0{,}99\times(1-0{,}089)}{(1-0{,}9689)\times 0{,}089 + 0{,}99\times(1-0{,}089)} = 0{,}997.$$

Ambos os valores preditivos são altos.

Como se ilustra no exemplo seguinte, no caso das doenças com muito baixa prevalência, mesmo um teste de diagnóstico com grande sensibilidade e grande especificidade, por ter um valor preditivo positivo muito baixo. Neste caso, o teste de diagnóstico é de grande utilidade no despiste da doença mas não na sua identificação.

Exemplo 3.6.2 Um estudo de grande escala sobre um dos testes standard para diagnosticar a infeção HIV-1 indica uma sensibilidade de 99,8% e uma especificidade de

[2]Fonte: http://alzheimerportugal.org/pt/prevalencia

98,5% para esse teste. Numa população em que a com prevalência da infeção HIV-1 é de 3 por 1000 habitantes, os valores preditivos do teste são dados por:

$$\text{VPP} = \frac{0{,}998 \times 0{,}003}{0{,}998 \times 0{,}003 + 0{,}015 \times 0{,}997} = 0{,}167 \text{ (!!!)},$$

e

$$\text{VPN} = \frac{0{,}985 \times 0{,}997}{0{,}002 \times 0{,}003 + 0{,}985 \times 0{,}997} = 0{,}999.$$

3.7 Bibliografia

Albert, J.H. (2003). College students' conceptions of probability, *The American Statistician*, 57, 37–45.

Daniel, W.W. (2009). *Biostatistics: a foundation for analysis in the health sciences*, Wiley.

Dawson, B., Trapp, R.G. (2004). *Basic & Clinical Biostatistics*, McGraw-Hill.

Graça Martins, M.E., Cerveira, A.G. (1999). *Introdução às Probabilidades e à Estatística*, Universidade Aberta.

Moore, D.S. (1985). *Statistics: concepts and controversies*, W.H. Freeman and Company.

Moore, D.S., McCabe, G.P. (2006). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

4

Distribuição de probabilidade duma variável aleatória

Noção de variável aleatória. Variáveis discretas e contínuas. Distribuição de probabilidade. Histograma de probabilidade. Densidade de probabilidade. Média e variância duma variável aleatória. Propriedades da média e da variância. Lei dos grandes números.

4.1 Noção de variável aleatória

Em cada uma das experiências aleatórias descritas nos capítulos anteriores, estivemos interessados na observação duma variável que, tendo em conta a distinção que fizemos no Capítulo 1, podemos classificar de qualitativa ou de quantitativa. Quer num quer noutro caso, quando uma variável associa um valor numérico a cada resultado duma experiência aleatória, vamos chamar-lhe **variável aleatória**.

Nos parágrafos 1.2 e 1.3 estudámos métodos gráficos e numéricos para descrever a distribuição duma variável a partir de observações efetuadas dessa variável. A noção de distribuição duma variável como sendo o conjunto de valores que a variável toma e também a frequência com que os toma, está intimamente relacionada com as observações realizadas. Dois conjuntos de dados retirados de uma mesma população conduzem normalmente a distribuições de frequências diferentes para determinada variável. Apesar disso, será de esperar que esses dois conjuntos de dados, porque relativos a uma mesma variável e a uma mesma população, comportem informação semelhante no que respeita ao centro, à dispersão e à forma da distribuição dessa variável.

Lançando mão da noção de probabilidade que estudámos no capítulo anterior, vamos precisar um pouco mais a noção de **distribuição duma variável aleatória**, tornando-a, em particular, independente do conjunto de observações realizadas. Vamos chamar-lhe por isso, **distribuição de probabilidade** da variável. Como veremos, a distribuição de probabilidade pode ser interpretada como uma versão idealizada da

distribuição de frequências dessa variável. Distinguiremos os casos das variáveis que tomam um número finito de valores distintos, a que chamamos **discretas**, das variáveis que tomam todos os valores dum determinado intervalo, a que chamamos **contínuas**.

4.2 Distribuição de probabilidade

A **distribuição de probabilidade** duma variável dá-nos conta dos valores que a variável toma e da **probabilidade** com que os toma.

4.2.1 Variáveis aleatórias discretas

Para uma variável aleatória discreta X que toma os valores $x_1, x_2, \ldots, x_k$ com probabilidades $p_1, p_2, \ldots, p_k$, respetivamente, a sua distribuição de probabilidade pode ser apresentada numa tabela do tipo seguinte:

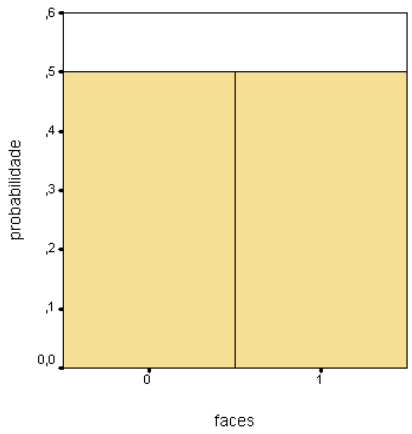
valores de X	x_1	x_2	$\ldots$	x_k
probabilidade	p_1	p_2	$\ldots$	p_k

Como a probabilidade de ocorrência de cada um dos valores x_i é, de acordo com a definição frequencista de probabilidade (cf. §3.3.2), aproximada pela sua frequência relativa obtida a partir dum grande número de observações da variável, há uma relação óbvia entre a tabela anterior e uma tabela de frequências relativas da variável. Neste sentido, é por vezes útil interpretar a distribuição de probabilidade como sendo a verdadeira distribuição da variável ou uma descrição idealizada da distribuição de frequências relativas da variável, sendo esta última distribuição, a que podemos aceder através da observação da variável, uma aproximação da verdadeira distribuição.

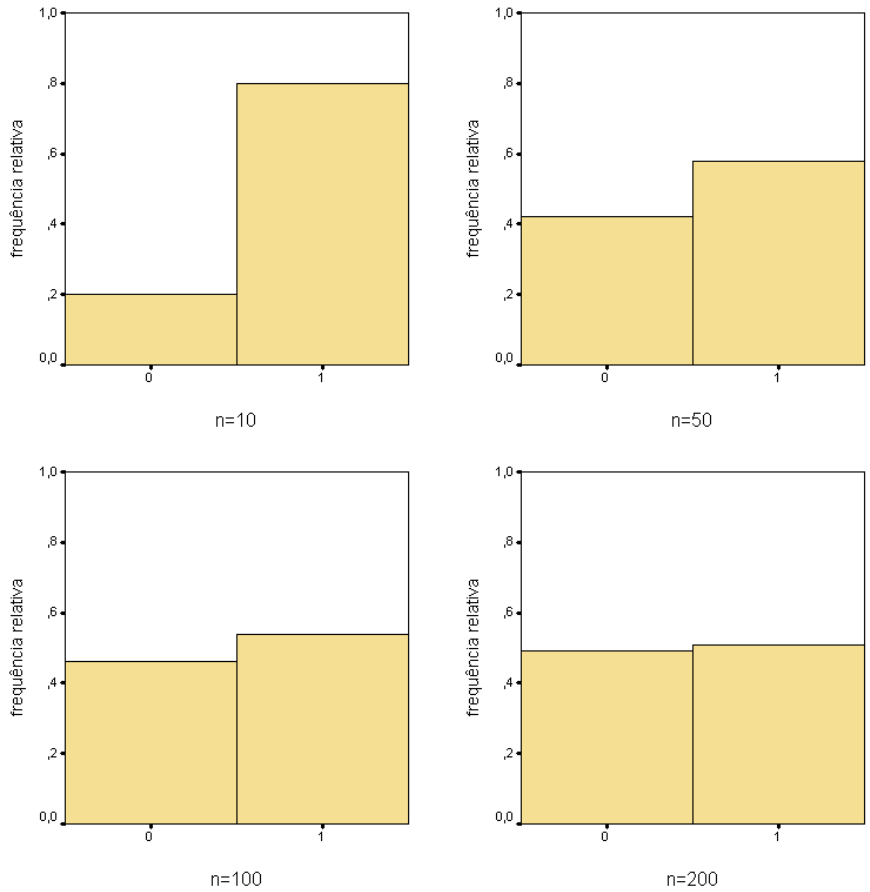
Exemplo 4.2.1 Consideremos o caso do lançamento duma moeda equilibrada de um euro, em que X representa a face que ocorre em cada lançamento. Representando por 0 a ocorrência da face europeia e por 1 a ocorrência da face portuguesa, a distribuição de probabilidade de X é dada por:

valores de X	0	1
probabilidade	1/2	1/2

Esta distribuição pode também ser representada graficamente na forma de histograma, a que chamamos **histograma de probabilidade**:



Sendo a probabilidade de ocorrência de cada uma das faces aproximada pela sua frequência relativa obtida ao longo dum grande número de lançamentos da moeda (definição frequencista de probabilidade), o histograma de probabilidade anterior está naturalmente relacionado com os histogramas de frequências relativas obtidos a partir de vários lançamentos da moeda.

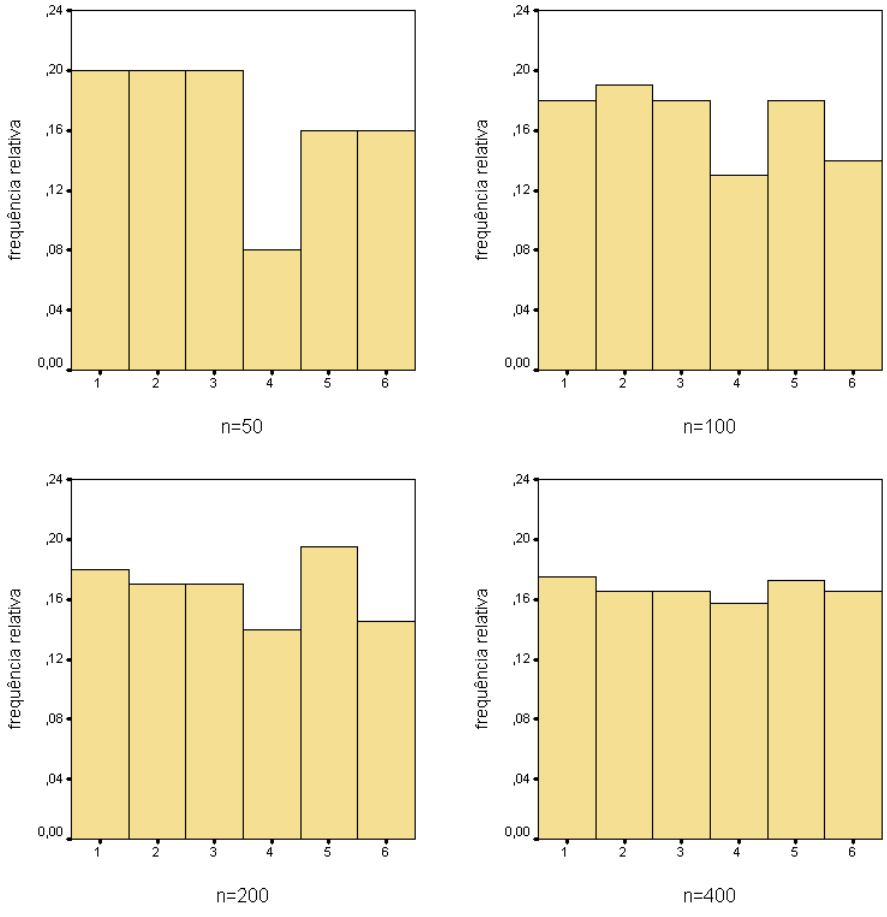


Os histogramas anteriores descrevem a distribuição de frequências de X a partir de 10, 50, 100 e 200 lançamentos da moeda. Cada uma destas representações descreve uma realidade particular. No entanto, quando o número de observações aumenta, os histogramas (ou seja, as respetivas frequências relativas) estabilizam aproximando-se do histograma de probabilidade da variável X (ou seja, das respetivas probabilidades).

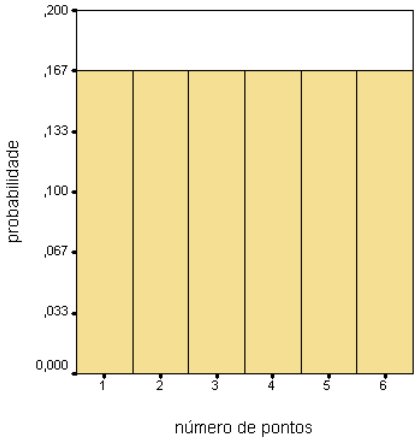
Exemplo 4.2.2 No caso do lançamento dum dado equilibrado, representando por Y o número de pontos obtidos em cada lançamento do dado, a distribuição de probabilidade de Y é dada por

valores de Y	1	2	3	4	5	6
probabilidade	1/6	1/6	1/6	1/6	1/6	1/6

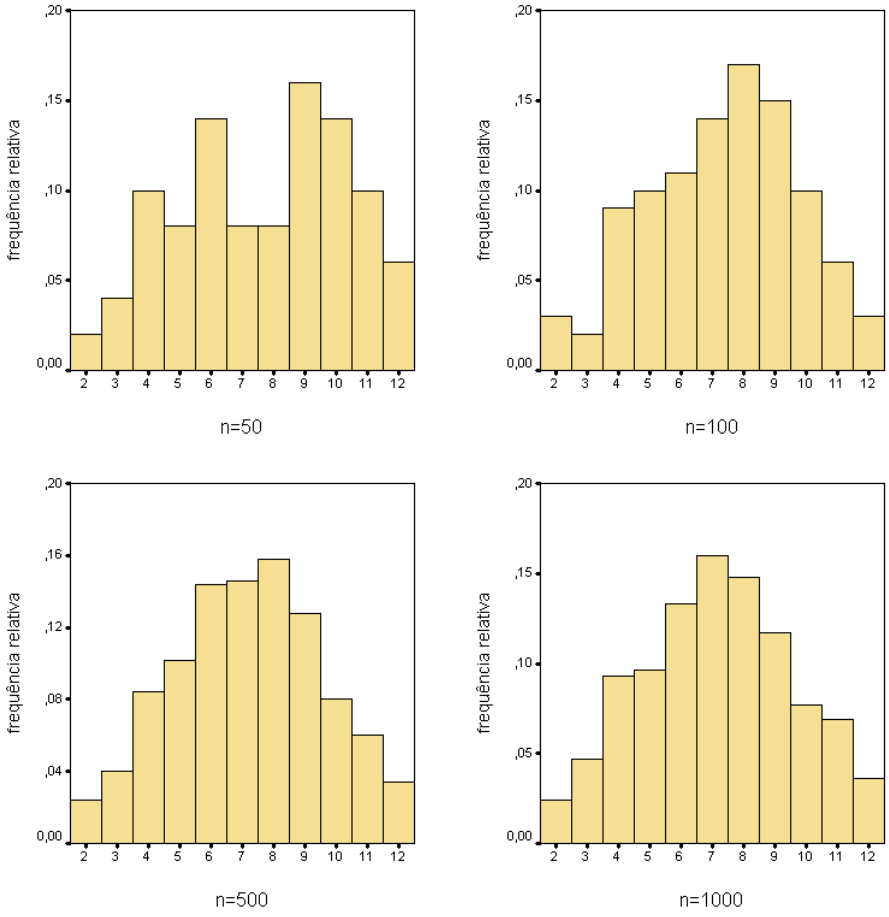
A partir de 50, 100, 200 e 400 lançamento do dado, obtemos os histogramas de frequências relativas seguintes:



Tal como no caso da moeda, à medida que o número de observações aumenta, o histograma de frequências relativas aproxima-se do histograma de probabilidade, que neste caso é dado por

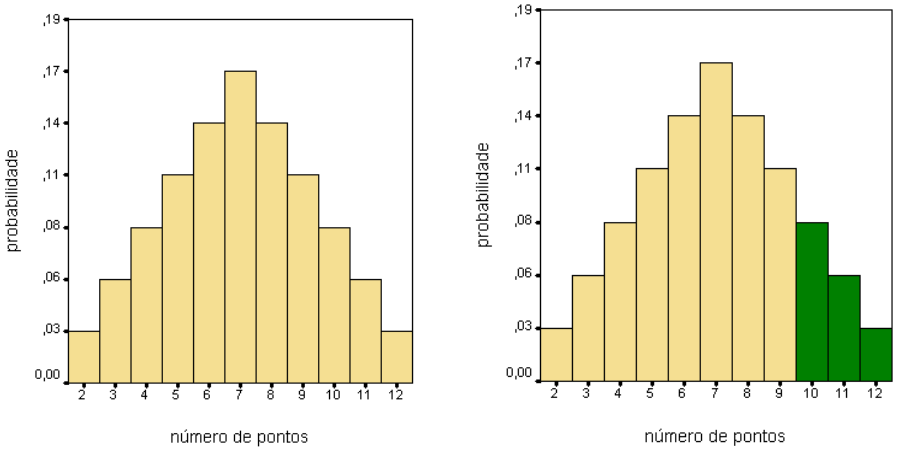


Exemplo 4.2.3 No lançamento de dois dados equilibrados usuais, observaram-se as seguintes frequências relativas para o número total de pontos obtidos nos dois dados em 50, 100, 500 e 1000 lançamentos dos mesmos:



De acordo com a interpretação frequencista de probabilidade estes histogramas aproximar-se-ão do histograma de probabilidade correspondente à variável aleatória S que nos dá a soma dos pontos obtidos em ambos os dados, e cuja distribuição de probabilidade é dada por

valores de S	2	3	4	5	6	7	8	9	10	11	12
probabilidade	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



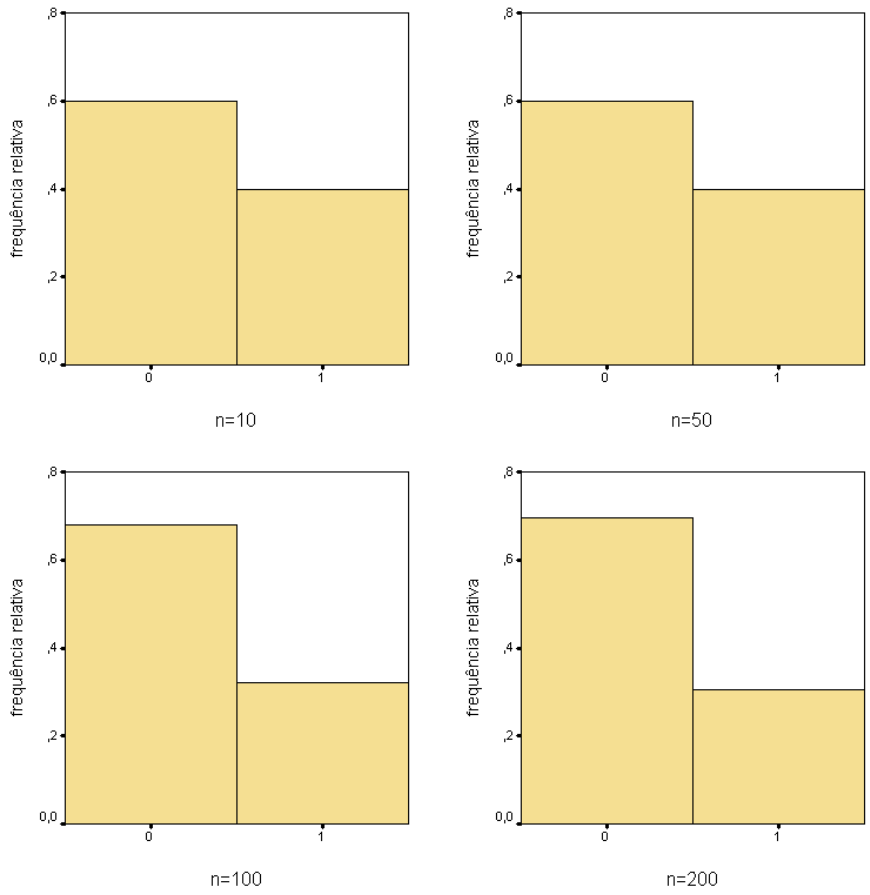
A probabilidade de obter 10 ou mais pontos no lançamento de dois dados equilibrados é igual a

$$\mathrm{P}(S \geq 10) = \mathrm{P}(S = 10) + \mathrm{P}(S = 11) + \mathrm{P}(S = 12) = \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{6}{36} = \frac{1}{6}.$$

Notemos que tal probabilidade não é mais do que a área da região marcada no segundo dos histogramas de probabilidade anteriores.

Os exemplos anteriores, apesar de importantes para motivar a noção de distribuição de probabilidade a partir da noção de distribuição de frequências, são pouco interessantes dum ponto de vista da inferência estatística. Com efeito, nos casos anteriores sabemos tudo sobre a experiência aleatória em causa, isto é, conseguimos, a partir da informação *a priori* sobre a experiência, explicitar a distribuição de probabilidade das variáveis X, Y e S. No exemplo seguinte isso não acontece.

Exemplo 4.2.4 Os gráficos seguintes resumem os resultados obtidos no lançamento duma moeda de um euro, para 10, 50, 100 e 200 lançamentos da moeda, onde por 0 representamos a ocorrência da face europeia e por 1 a ocorrência da face portuguesa.



Achando que os resultados anteriores revelam fortes indícios de que a moeda não é equilibrada, uma vez que os histogramas anteriores não parecem aproximar-se do histograma de probabilidade relativo a uma moeda equilibrada, não podemos explicitar a distribuição de probabilidade da variável Z que representa a face que ocorre em cada lançamento. No entanto, se representarmos por p a probabilidade de ocorrência da face portuguesa, podemos dizer que a distribuição de probabilidade de Z é da forma

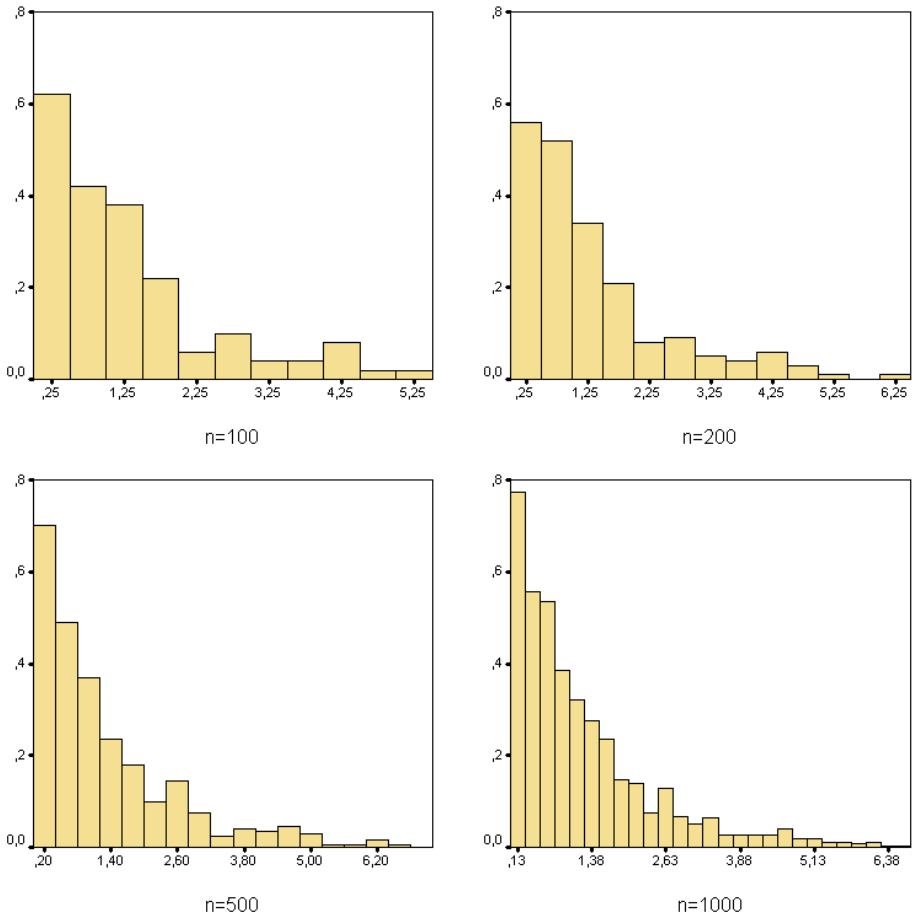
valores de Z	1	0
probabilidade	p	$1-p$

Usando a linguagem dos estudos observacionais por amostragem, p pode ser interpretado como uma caraterística numérica desconhecida da população em estudo, isto é, p é um parâmetro. A inferência sobre o verdadeiro valor do parâmetro p é um problema do interesse da estatística inferencial. Em particular, podemos querer saber se a moeda é equilibrada, isto é, se $p = 1/2$.

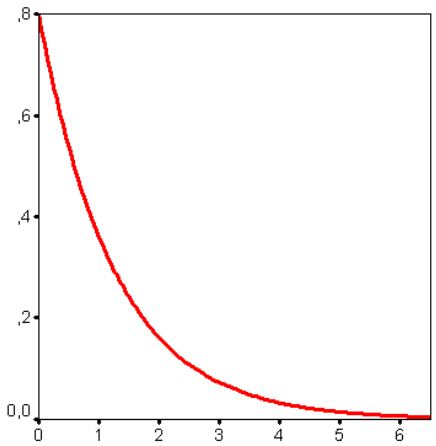
4.2.2 Variáveis aleatórias contínuas

A estabilização do histograma de frequências relativas à medida que o número de observações da variável em estudo aumenta, ocorre não só no caso das variáveis discretas, como também para as variáveis contínuas. Este facto é ilustrado nos dois exemplos seguintes.

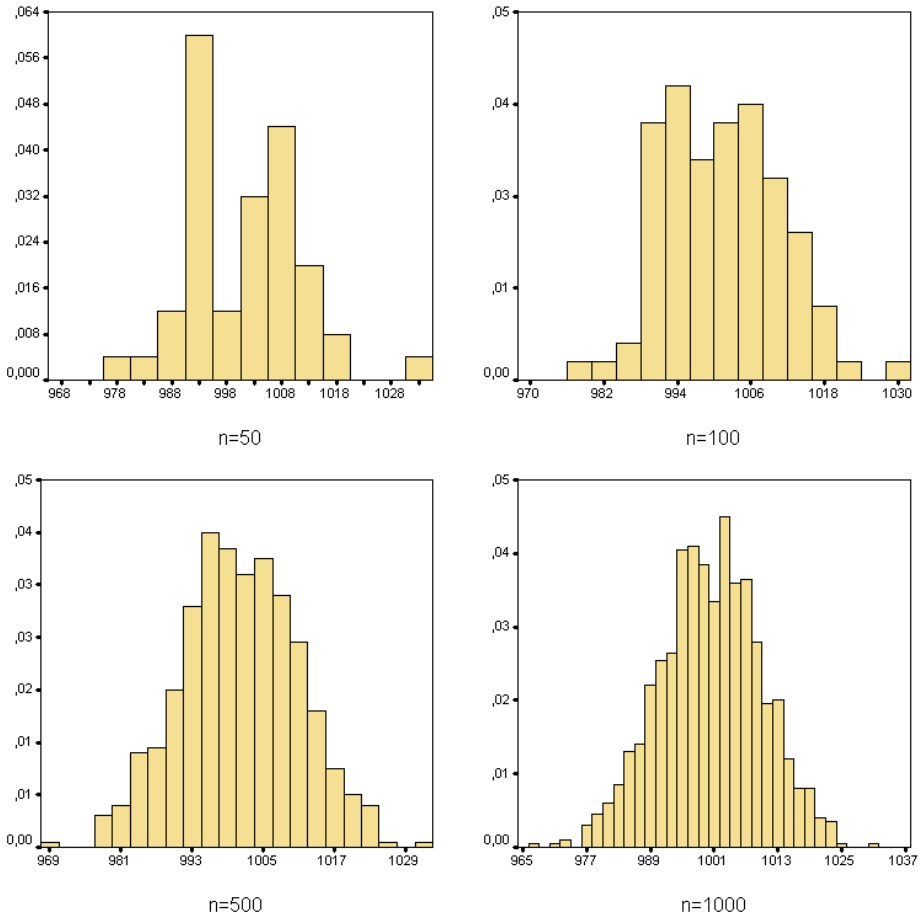
Exemplo 4.2.5 Representemos por X o tempo que medeia a chegada de dois clientes consecutivos a uma caixa de supermercado. Os histogramas seguintes descrevem a distribuição de X a partir de amostras de tamanho 100, 200, 500 e 1000.



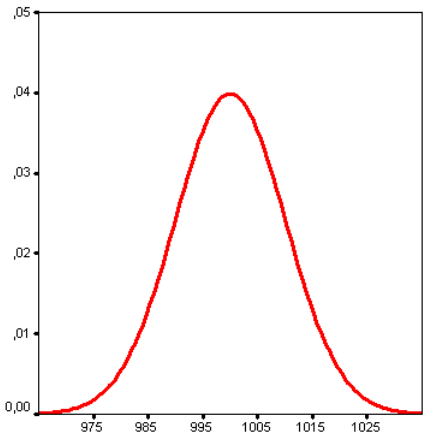
Tal como nos exemplos anteriores, os histogramas tendem a estabilizar quando o número de observações é grande. Pode ainda ser demonstrado que à medida que o número de observações aumenta e o tamanho das classes diminui não muito violentamente, a sua forma aproxima-se duma curva regular. No caso presente, uma tal curva é representada no gráfico seguinte.



Exemplo 4.2.6 Seja agora Y o peso, em gramas, de pacotes de açúcar empacotados por uma máquina. Os histogramas normalizados seguintes descrevem a distribuição de Y para de amostras de tamanho 50, 100, 500 e 1000:



Com o aumento do números de observações e a diminuição da amplitude das classes, os histogramas aproximam-se da curva



Como o aspeto do histograma não depende da escala usada no eixo vertical, vamos restringir a nossa atenção aos histogramas cuja área total é igual a 1 (para construir histogramas com área total igual a 1, devemos marcar no eixo dos yy a frequência relativa dividida pela amplitude de cada classe; os histogramas dos exemplos anteriores foram assim construídos). Neste caso será de esperar que a curva que aproxima o histograma goze das seguintes propriedades que são caraterísticas duma classe de curvas a que chamamos **curvas densidade**. A última das propriedades seguintes é consequência da interpretação frequencista de probabilidade e do facto da frequência relativa de qualquer intervalo que marquemos no eixo dos xx ser aproximadamente igual à área do histograma que tem por base esse intervalo.

Curva densidade:

- é uma curva que está acima do eixo dos xx e em que a área compreendida entre ela e esse eixo é igual a 1;
- é usada para descrever a distribuição duma variável contínua;
- a probabilidade dessa variável tomar valores num qualquer intervalo que marquemos no eixo dos xx é igual à área da região compreendida entre a **curva densidade** e o eixo dos xx que tem por base esse intervalo.

A curva densidade é assim um **modelo matemático** para a distribuição da variável em estudo, sendo, por isso, uma descrição idealizada duma tal distribuição. À curva

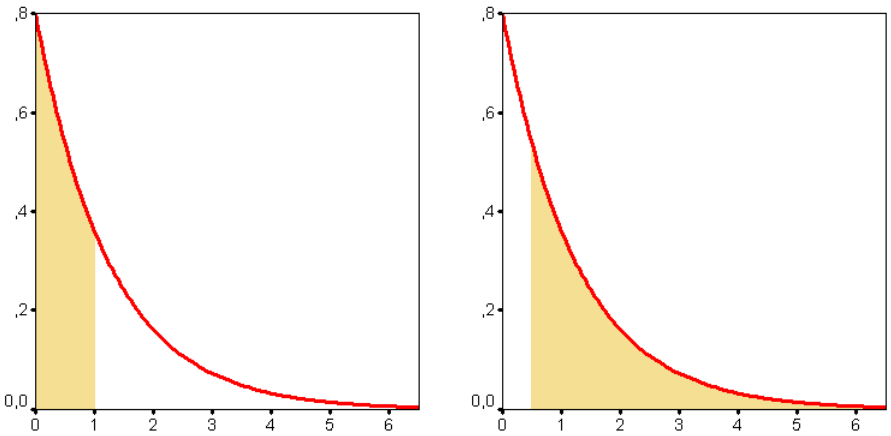
densidade duma variável X chamamos **densidade de probabilidade da variável** X ou apenas **densidade de** X. Como veremos mais tarde, um tal modelo matemático é essencial para o desenvolvimento de muito dos procedimentos estatísticos próprios da **estatística indutiva**.

Interpretando a distribuição de probabilidade como a verdadeira distribuição da variável em estudo, uma vez que esta é obtida a partir dum conjunto idealmente infinito de observações da variável, o conhecimento da densidade de probabilidade duma variável X permite-nos calcular a probabilidade de acontecimentos aleatórios que estão associados à variável X.

Exemplo 4.2.5 (cont.) Tendo em conta que a curva apresentada no Exemplo 4.2.5 descreve a distribuição de probabilidade dos tempos de interchegada (X) de clientes a uma caixa dum hipermercado, pela última das propriedades duma densidade de probabilidade podemos concluir que a probabilidade de cada um dos acontecimentos

$$A = \text{“tempo de interchegada inferior a 1 minuto”} = \{X < 1\}$$
$$B = \text{“tempo de interchegada superior a meio minuto”} = \{X > 0{,}5\},$$

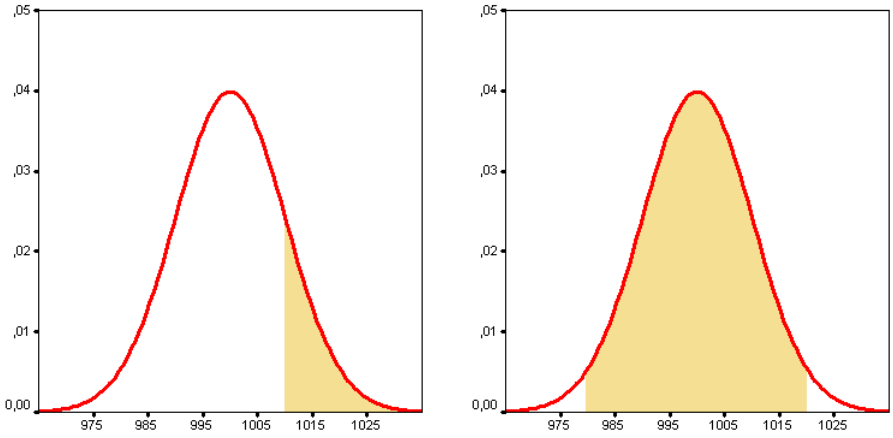
é igual, respetivamente, à área de cada uma das regiões representadas nas figuras seguintes:



Exemplo 4.2.6 (cont.) De igual modo, sendo a distribuição dos pesos de pacotes de açúcar (Y) descrita pela densidade de probabilidade apresentada no Exemplo 4.2.6, a probabilidade de cada um dos acontecimentos

$$A = \text{“peso superior a 1010 gramas”} = \{Y > 1010\}$$
$$B = \text{“peso superior a 980 gramas e inferior a 1020 gramas”} = \{980 < Y < 1020\},$$

é igual à área das regiões seguintes:



Levanta-se agora o problema de saber como calcular cada uma das áreas que associámos aos acontecimentos aleatórios anteriores. Voltaremos a esta questão mais à frente.

4.3 Média e variância duma variável aleatória

No §1.3 vimos como calcular a média $\bar{x}$ e a variância s_x^2 dum conjunto de n observações duma variável X. Se $x_1, x_2, \ldots, x_k$ são os valores distintos que ocorrem nessas observações, e $n_1, n_2, \ldots, n_k$ o número de vezes que cada um deles ocorre, as fórmulas para o cálculo da média e da variância são, respetivamente,

$$\bar{x} = \frac{\sum n_i x_i}{n} = \sum \frac{n_i}{n} x_i$$

e

$$s_x^2 = \frac{\sum n_i (x_i - \bar{x})^2}{n-1} = \sum \frac{n_i}{n-1} (x_i - \bar{x})^2,$$

onde n_i/n é a frequência relativa do valor x_i assumido pela variável X.

A média e a variância assim calculadas, dependem duma distribuição de frequências particular de X. Outro conjunto de observações conduziria a outra distribuição de frequências e, consequentemente, a outros valores para $\bar{x}$ e s_x^2. Para reforçar o facto de $\bar{x}$, s_x^2 variarem de amostra para amostra, $\bar{x}$ e s_x^2 são também ditas **média amostral** e **variância amostral**, respetivamente.

Utilizando a noção de distribuição de probabilidade de X, é fácil introduzir uma noção de **média** e de **variância** da variável aleatória X que não dependa de qualquer conjunto de observações de X. Para as distinguir das média e variância amostrais, vamos denotá-las por μ_X e σ_X^2, ou, simplesmente, por μ e σ^2.

4.3.1 O caso discreto

Para uma **variável aleatória discreta** X com distribuição de probabilidade dada por

valores de X	x_1	x_2	$\ldots$	x_k
probabilidade	p_1	p_2	$\ldots$	p_k

a **média**, μ_X, e a **variância**, σ_X^2, são definidas, respetivamente, por

$$\mu_X = \sum p_i \, x_i$$

e

$$\sigma_X^2 = \sum p_i \, (x_i - \mu_X)^2.$$

À raiz quadrada da variância, σ_X, chamamos **desvio-padrão** da variável aleatória X. Como podemos constatar, estas fórmulas são semelhantes às fórmulas de cálculo das média e variância amostrais. Em vez de utilizarmos uma distribuição de frequências de X, utilizamos a distribuição de probabilidade de X.

Tal como para as caraterísticas amostrais respetivas, a média μ_X é uma medida do centro da distribuição de probabilidade de X, enquanto que a variância σ_X^2, ou o desvio-padrão σ_X, são medidas de dispersão da distribuição de probabilidade de X em torno da média μ_X. Variáveis aleatórias com distribuições concentradas em torno da média têm pequenos desvios-padrão, contrariamente a variáveis mais dispersas em torno da média.

A interpretação geométrica que apresentámos anteriormente para $\bar{x}$ como sendo o ponto do eixo horizontal que "equilibra" o histograma de frequências relativas da variável X, mantém-se para μ_X, mas relativamente ao seu histograma de probabilidade.

Tal como já acontecia com o cálculo da variância amostral, a fórmula anterior não é a mais apropriada para o cálculo de σ_X^2. Para esse efeito é preferível utilizar a fórmula

Cálculo da variância de X:

$$\sigma_X^2 = \sum p_i \, x_i^2 - \mu_X^2.$$

Exemplo 4.3.1 Ilustremos a aplicação das fórmulas anteriores, efetuando o cálculo da média e da variância das variáveis aleatórias X e Y definidas nos Exemplos 4.2.1 e 4.2.2, respetivamente. Para a variável X temos,

$$\mu_X = \frac{1}{2} \times 0 + \frac{1}{2} \times 1 = \frac{1}{2} = 0{,}5$$

e

$$\sigma_X^2 = \frac{1}{2} \times 0^2 + \frac{1}{2} \times 1^2 - 0{,}5^2 = 0{,}25$$

e para Y obtemos

$$\mu_Y = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = 3{,}5$$

e

$$\sigma_Y^2 = \frac{1}{6} \times 1^2 + \frac{1}{6} \times 2^2 + \frac{1}{6} \times 3^2 + \frac{1}{6} \times 4^2 + \frac{1}{6} \times 5^2 + \frac{1}{6} \times 6^2 - 3{,}5^2 \approx 2{,}9167.$$

Tendo em conta a interpretação geométrica da média, reparemos que dos histogramas de probabilidade das variáveis X e Y (ver pág. 99 e 101) poderíamos ter concluído imediatamente, e sem efetuar qualquer cálculo, que

$$\mu_X = 0{,}5 \quad \text{e} \quad \mu_Y = 3{,}5.$$

Reparemos na interpretação simples das médias anteriores como número médio, ou esperado, de faces portuguesas, em cada lançamento da moeda, ou de pontos, em cada lançamento do dado. Com efeito, no caso do lançamento da moeda, sendo ela equilibrada, esperamos, em média, obter uma face portuguesa em cada dois lançamentos, isto é, esperamos obter um ponto em cada dois lançamentos, ou seja, 0,5 pontos por lançamento. No caso do lançamento do dado esperamos, em média, obter cada uma das faces em cada seis lançamentos, isto é, esperamos obter em média $(1+2+3+4+5+6)/6 = 3{,}5$ pontos por lançamento.

Exemplo 4.3.2 Suponhamos agora que um dado equilibrado tem marcados os números 1, em três das faces, 2, em duas das faces, e o número 3 na face restante. Se Z representar o número de pontos obtidos num lançamento do dado, a distribuição de probabilidade de Z é dada por

valores de Z	1	2	3
probabilidade	1/2	1/3	1/6

A média e a variância de Z são dadas por

$$\mu_Z = \frac{1}{2} \times 1 + \frac{1}{3} \times 2 + \frac{1}{6} \times 3 = \frac{5}{3}$$

e

$$\sigma_Z^2 = \frac{1}{2} \times 1^2 + \frac{1}{3} \times 2^2 + \frac{1}{6} \times 3^3 - \left(\frac{5}{3}\right)^2 = \frac{5}{9}.$$

4.3.2 O caso contínuo

No caso da **variável X ser contínua**, vimos já que a sua distribuição de probabilidade é caraterizada pela densidade de probabilidade de X. Neste caso, a **média**, μ_X, e a **variância**, σ_X^2, **da variável contínua** X são definidas à custa da sua densidade de probabilidade. Para efetuar tais cálculos, bem como de outras caraterísticas numéricas duma distribuição como a mediana, a amplitude interquartil e os percentis, há procedimentos matemáticos adequados para o efeito. Devido há complexidade de tais métodos, não os vamos aqui abordar. Ficar-nos-emos apenas pela identificação gráfica da média a partir da densidade de probabilidade. Para o efeito, procedemos de forma análoga ao que fizemos para o histograma da Figura 1.3.2: a média é o ponto do eixo dos xx que mantém a densidade de probabilidade em equilíbrio. A variância não tem, em geral, uma interpretação geométrica simples.

De forma perfeitamente análoga ao que fizemos para os histogramas das Figuras 1.3.5 e 1.3.11, poderíamos também identificar geometricamente outras caraterísticas numéricas duma distribuição como são os casos dos quartis. A mediana é o ponto do eixo dos xx em que as áreas das regiões compreendidas entre a densidade de probabilidade e o eixo dos xx à esquerda e à direita desse ponto são iguais. O primeiro quartil é o ponto do eixo dos xx em que as áreas das regiões compreendidas entre a densidade de probabilidade e o eixo dos xx à esquerda e à direita desse ponto são iguais a 1/4 e a 3/4, respetivamente. Analogamente se identifica o terceiro quartil.

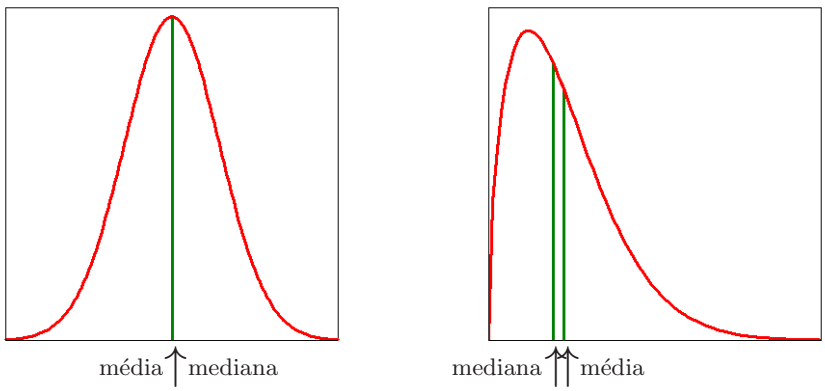


Figura 4.3.3: Localização gráfica da média e da mediana em curvas densidade

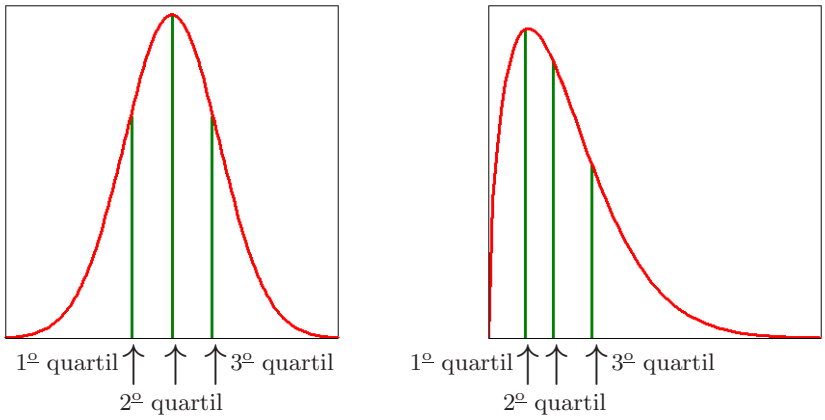


Figura 4.3.4: Localização gráfica dos quartis em curvas densidade

4.4 Propriedades da média e da variância

Por razões análogas às expostas no §1.4, a média, μ, e a variância, σ^2, duma variável aleatória, gozam das propriedades da média e variância amostrais. Mais precisamente, se duas variáveis aleatórias X e Y obedecem a uma relação do tipo

$$Y = a\,X + b,$$

para determinados valores reais a e b, então valem as relações seguintes entre as médias e variâncias de X e Y:

Efeito da transformação linear $Y = a\,X + b$:

- média:

$$\mu_Y = a\,\mu_X + b;$$

- variância e desvio-padrão:

$$\sigma_Y^2 = a^2\,\sigma_X^2, \quad \sigma_Y = a\,\sigma_X.$$

Conhecidas a média e a variância de duas variáveis aleatórias X e Y, é por vezes importante saber como calcular a média e a variância da variável soma $X + Y$, à custa das médias e variâncias de cada uma das variáveis X e Y (nos casos em que tal seja possível). Vejamos um exemplo duma tal situação.

Exemplo 4.4.1 O Abel joga com um adversário o seguinte jogo: cada um deles lança uma moeda portuguesa de um euro; por cada face portuguesa que ocorra nas duas moedas o Abel paga ao adversário 5 euros; por cada face europeia que ocorra nas duas moedas o Abel recebe do adversário 5 euros. Representemos por X o ganho (ou perda) do Abel com a sua moeda em cada lançamento da mesma, e por Y o ganho (ou perda) do Abel devido à moeda do seu adversário. Reparemos que X e Y têm a mesma distribuição de probabilidade que é dada por

valores de X (Y)	-5	5
probabilidade	1/2	1/2

As médias e variâncias de X e Y, que nos dão o ganho médio por partida do Abel com a sua moeda e com a moeda do seu adversário, respetivamente, coincidem, sendo dadas por:

$$\mu_X = \mu_Y = \frac{1}{2} \times (-5) + \frac{1}{2} \times 5 = 0$$

e

$$\sigma_X^2 = \sigma_Y^2 = \frac{1}{2} \times (-5)^2 + \frac{1}{2} \times 5^2 - 0^2 = 25.$$

O ganho total do Abel em cada repetição do jogo é dado pela variável $Z = X + Y$. Como fazer para calcular a média e a variância de Z? Seguindo o procedimento anterior, precisamos de determinar a distribuição de probabilidade de Z:

valores de Z	-10	0	10
probabilidade	1/4	1/2	1/4

Assim

$$\mu_Z = \frac{1}{4} \times (-10) + \frac{1}{2} \times 0 + \frac{1}{4} \times 10 = 0$$

e

$$\sigma_Z^2 = \frac{1}{4} \times (-10)^2 + \frac{1}{2} \times 0^2 + \frac{1}{4} \times 10^2 - 0^2 = 50.$$

No exemplo anterior, valem as igualdades

$$\mu_{X+Y} = \mu_X + \mu_Y$$

e

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2.$$

Serão estas relações válidas em geral? No caso da média, tal é com efeito verdade:

Média da soma $X+Y$:

Para quaisquer duas variáveis aleatórias X e Y, a média da soma $X+Y$ é igual à soma das médias respetivas:

$$\mu_{X+Y} = \mu_X + \mu_Y.$$

Reparemos que esta propriedade é partilhada pela média amostral. Com efeito, se $x_1, \ldots, x_n$ e $y_1, \ldots, y_n$ são os valores observados para duas variáveis X e Y, onde os valores x_i e y_i são observações relativas a um mesmo indivíduo, a média amostral $\bar{z}$ relativa à variável $Z = X + Y$, é dada por

$$\begin{aligned} \bar{z} &= \frac{\sum z_i}{n} = \frac{\sum (x_i + y_i)}{n} \\ &= \frac{x_1 + y_1 + x_2 + y_2 + \ldots + x_n + y_n}{n} \\ &= \frac{x_1 + x_2 + \ldots + x_n}{n} + \frac{y_1 + y_2 + \ldots + y_n}{n} \\ &= \bar{x} + \bar{y}. \end{aligned}$$

Notemos, no entanto, que a variância amostral da soma de duas variáveis não é necessariamente igual à soma das variâncias amostrais de cada uma das variáveis. Com efeito, efetuando mais alguns cálculos chegaríamos à conclusão que a variância amostral s_z^2 de Z era dada por

$$s_z^2 = s_x^2 + s_y^2 + 2\, r\, s_x\, s_y,$$

onde r é um número pertencente ao intervalo $[0, 1]$ a que, mais à frente, chamaremos de coeficiente de correlação linear entre as variáveis X e Y.

Da mesma forma, também para a variância da soma de duas variáveis X e Y, σ^2_{X+Y}, não é em geral igual à soma das variância de X e Y. Tal acontece quando as variáveis X e Y são **independentes**, isto é, quando a ocorrência de qualquer um dos valores de uma das variáveis não afeta a probabilidade de ocorrência de qualquer um dos valores da outra variável:

Variância da soma $X+Y$:

Se X e Y são variáveis aleatórias independentes, a variância da soma $X+Y$ é igual à soma das variâncias respetivas:

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y.$$

Exemplo 4.4.1 (cont.) Tendo em conta as duas propriedades anteriores, e a independência entre as variáveis X e Y (uma vez que o resultado obtido numa moeda não influencia, nem é influenciado, pelo resultado obtido na outra), concluímos que o cálculo da média e da variância da variável $X+Y$, que nos dá o ganho total obtido pelo Abel em cada repetição do jogo, pode ser feito sem ser necessário obter a distribuição de probabilidade de $X+Y$. Basta conhecermos as média e variância de cada uma das variáveis X e Y. Assim

$$\mu_{X+Y} = \mu_X + \mu_Y = 0 + 0 = 0 \quad \text{e} \quad \sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y = 25 + 25 = 50.$$

Exemplo 4.4.2 Relativamente ao Exemplo 4.2.3 (pág. 101), se representarmos por S_1 e S_2 os pontos que saem no primeiro e no segundo dado, respetivamente, a soma, S, dos pontos obtidos nos dois dados é dada por $S = S_1 + S_2$, onde as variáveis S_1 e S_2 são independentes. Como $\mu_{S_1} = \mu_{S_2} = 3{,}5$ e $\sigma^2_{S_1} = \sigma^2_{S_2} \approx 2{,}9167$, obtemos

$$\mu_S = \mu_{S_1} + \mu_{S_2} = 3{,}5 + 3{,}5 = 7$$

e

$$\sigma^2_S = \sigma^2_{S_1} + \sigma^2_{S_2} \approx 2{,}9167 + 2{,}9167 = 5{,}8334.$$

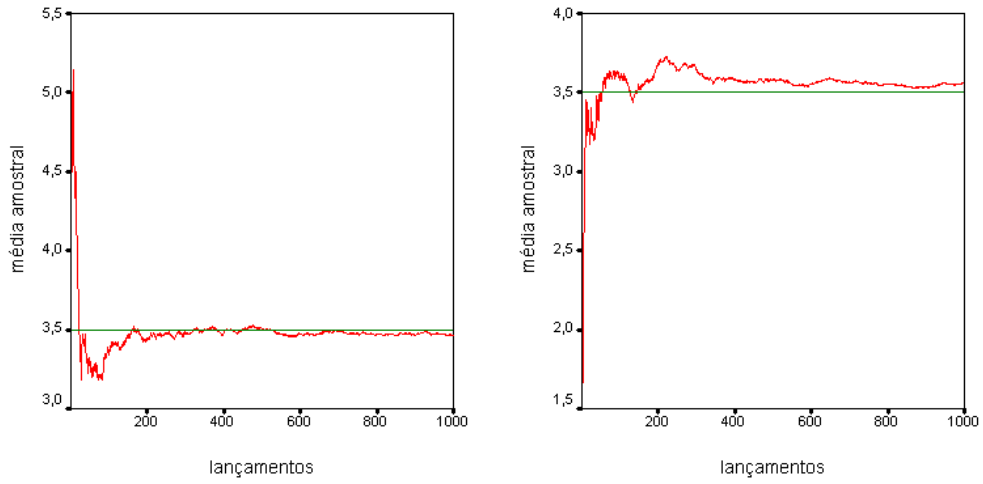
Em lançamentos sucessivos de dois dados equilibrados obtemos assim, em média, 7 pontos por lançamento.

Para reforçar a importância da condição de independência, ou mais precisamente, de ausência de associação linear, para a validade da regra anterior para o cálculo da variância da soma de duas variáveis aleatórias, atentemos no exemplo seguinte.

Exemplo 4.4.3 Para um adulto do sexo masculino medimos o comprimento de ambos os braços. Admitamos que a variável X representa o comprimento do braço esquerdo, enquanto que o simétrico do comprimento do braço direito é representado pela variável Y. Por exemplo, para um adulto com um braço esquerdo com $750\,mm$ e um braço direito com $755\,mm$, X vale 750 e Y vale -755. Como todos temos os braços esquerdo e direito com aproximadamente o mesmo tamanho, será de esperar que a variável $X+Y$ possua uma variabilidade pequena, e por conseguinte uma pequena variância. No entanto, há adultos com braços pequenos, adultos com braços médios e adultos com braços grandes. A variabilidade de cada uma das variáveis X e Y será, por isso, grande. Neste caso, a variância da soma $X+Y$ será seguramente inferior à soma das variâncias de X e de Y. Pelo que vimos atrás, este facto pode ser explicado pela forte dependência existente entre as variáveis X e Y.

4.5 Lei dos grandes números

Que relação existirá entre a média duma variável X e a média amostral $\bar{x}$ calculada a partir de observações da variável X? No caso particular de X representar os pontos obtidos em cada lançamento dum dado equilibrado, o gráficos seguintes sugerem que, à medida que o número de lançamentos aumenta, a média amostral se aproxima da média de X, que como vimos atrás é igual a $\mu = 3{,}5$.



Se recordarmos a definição frequencista de probabilidade enunciada no §3.3.2, sabemos que à medida que o número de observações aumenta, e se essas observações são realizadas aproximadamente nas mesmas condições, isto é, se as várias observações da variável X são independentes, a probabilidade p_i, de ocorrer qualquer um dos valores x_i, pode ser aproximada pela frequência relativa n_i/n desse valor, quando n é grande:

$$\frac{n_i}{n} \approx p_i.$$

Consequentemente,

$$\bar{x} = \sum \frac{n_i}{n}\, x_i \approx \sum p_i\, x_i = \mu_X,$$

isto é, a média amostral aproxima-se da média da variável X.

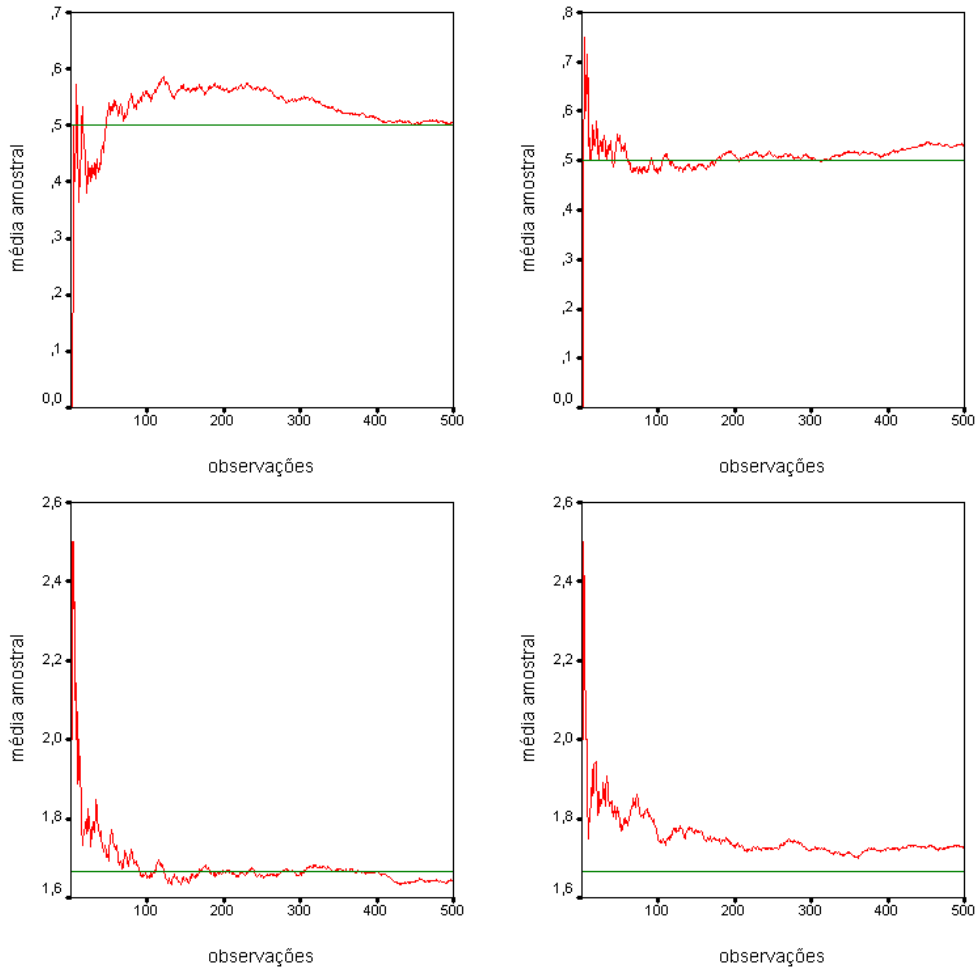
Esta igualdade explica o observado nos gráficos anteriores. A média $\mu = 3{,}5$ pode ser assim interpretada como o número médio de pontos obtidos por lançamento, quando o número de lançamentos é grande.

O facto anterior, válido, como vimos, para variáveis discretas mas também para variáveis contínuas, é conhecido como **lei dos grandes números**:

Lei dos grandes números:

Se as várias observações duma variável X com média μ são independentes, a média amostral $\bar{x}$ aproxima-se, tanto quanto queiramos, de μ, à medida que o número de observações aumenta.

Exemplo 4.5.1 Para cada uma das variáveis X e Z definidas nos Exemplos 4.3.1 e 4.3.2 (pág. 110), respetivamente, relativas aos resultados observados no lançamento duma moeda equilibrada, e dum dado equilibrado que tem marcados os números 1, em três das faces, 2, em duas das faces, e o número 3 na face restante, a lei dos grandes números é ilustrada nos gráficos seguintes que dão conta da evolução das médias amostrais com o aumento das observações, para dois conjuntos de observações de cada uma das variáveis. Notemos que no caso da variável X, $\bar{x}$ não é mais do que a proporção de faces portuguesas nos n primeiros lançamentos da moeda (porquê?).



Vejamos mais um exemplo que reforça a interpretação da média μ_X duma variável X, como o valor do qual se aproxima a média amostral, quando o número de observações aumenta.

Exemplo 4.5.2 No jogo da roleta, a roda da roleta está dividida em 37 partes iguais numeradas de 0 a 36, e um jogador, que à partida aposta num dos números de 1 a 36, recebe em caso de vitória 36 vezes mais do que aquilo que apostou. Admitindo que a aposta do jogador é sempre de 10 euros, ele recebe os 10 euros que apostou mais 350 euros pagos pelo casino se sair o número em que apostou. Caso contrário, perde o que apostou. Representando por X o ganho líquido do jogador em cada partida, X tem como distribuição de probabilidade

valores de X	-10	350
probabilidade	$36/37$	$1/37$

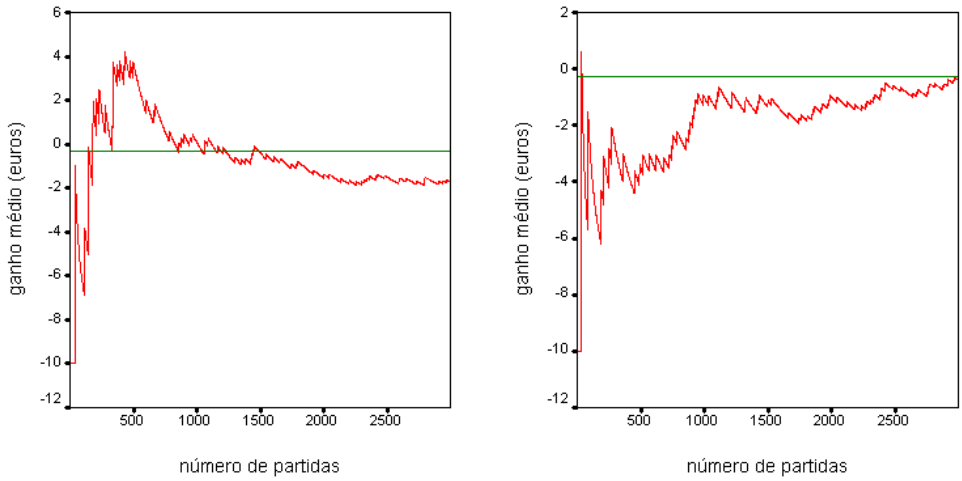
O ganho médio por partida é dado por

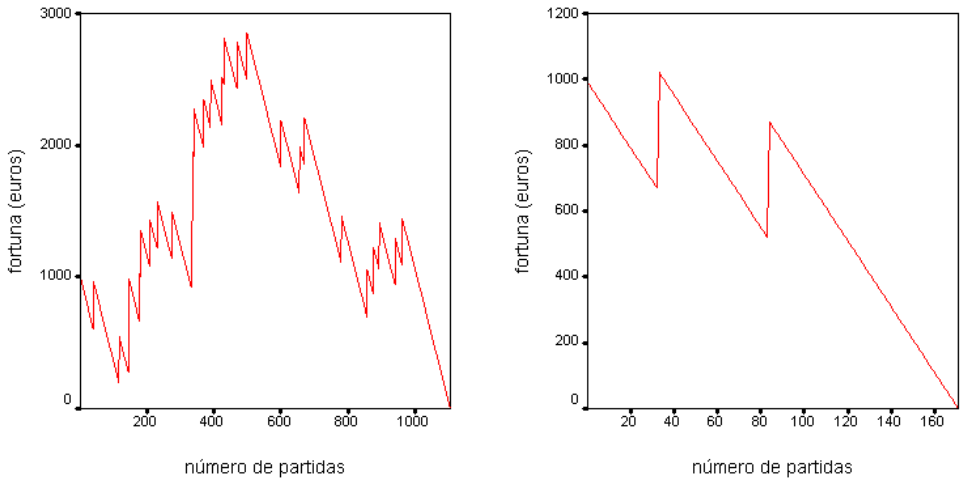
$$\mu_X = \frac{36}{37} \times (-10) + \frac{1}{37} \times 350 = -\frac{10}{37} = -0{,}27,$$

isto é, em cada partida, por cada 10 euros apostados, o jogador perde 27 cêntimos. Atendendo à lei dos grandes números, quer isto dizer que, independentemente do dinheiro que o jogador leva para o casino, ao fim dum grande número de partidas ficará sem dinheiro nenhum. Notemos, no entanto, que devido à grande variabilidade da variável X que é traduzida pela sua grande variância

$$\sigma_X^2 = \frac{36}{37} \times (-10)^2 + \frac{1}{37} \times 350^2 - \left(\frac{10}{37}\right)^2 \approx 3408{,}035,$$

a média amostral aproxima-se lentamente de $-0{,}27$.





Para ilustrar os factos referidos, apresentamos nos gráficos anteriores duas possíveis evoluções da média amostral, ou seja, do ganho médio por partida para um jogador com uma grande fortuna inicial, e também as correspondentes evoluções da fortuna (até ficar sem dinheiro) de um jogador que entra para o casino com 1000 euros para jogar na roleta.

4.6 Lei dos grandes números e inferência estatística

Contrariamente aos exemplos anteriores em que a população de onde recolhemos a amostra pode ser considerada infinita, uma vez que a experiência aleatória pode ser repetida tantas vezes quantas quisermos, num estudo observacional por amostragem a população é finita, sendo a amostra recolhida por métodos aleatórios, por exemplo, por amostragem aleatória simples. Apesar deste método de recolha de amostras não produzir observações independentes (basta pensar que se um indivíduo é observado, não volta a sê-lo), se o tamanho da população é grande relativamente à dimensão da amostra, as observações podem ser consideradas aproximadamente independentes, valendo ainda nesse caso a lei dos grandes números. Tendo em conta a linguagem introduzida quando falámos de estudos por amostragem, a lei dos grandes números permite-nos concluir que quando o parâmetro de interesse é uma média μ, este pode ser aproximado pela estatística $\bar{x}$ quando a dimensão da amostra for grande.

Num estudo observacional por amostragem em que, para uma determinada população de grande dimensão, pretendemos conhecer a proporção p de indivíduos que possuem determinada caraterística, vimos já que a partir da amostra recolhida podemos calcular a **estatística** $\widehat{p}$ associada ao parâmetro de interesse p que, neste caso, não é mais do que a proporção de indivíduos nessa amostra que possuem a caraterística em

estudo. Reparemos que $\widehat{p}$ não é mais do que a média amostral associada à variável aleatória X que toma o valor 1 se o indivíduo observado tem a caraterística em estudo, e 0 se isso não acontece. Como X é (aproximadamente) uma variável aleatória com distribuição de probabilidade

valores de X	1	0
probabilidade	p	$1-p$

a sua média é precisamente o parâmetro p:

$$\mu_X = p \times 1 + (1-p) \times 0 = p.$$

A lei dos grandes números permite concluir que, quando a amostra é grande, $\widehat{p}$ é uma boa aproximação de p. Desta forma justificamos uma afirmação anteriormente feita de que, não havendo enviesamento no que respeita à amostragem, esperávamos que a estatística $\widehat{p}$ nos desse uma boa informação sobre o parâmetro desconhecido p (ver §2.4).

Propriedades semelhantes são válidas para a variância ou para o desvio-padrão amostrais. Como consequência da lei dos grandes números, as estatísticas s_x^2 e s_x aproximam-se, tanto quanto queiramos, de σ_X^2 e σ_X (variância e desvio-padrão populacionais), à medida que o número de observações aumenta.

Os factos anteriores têm grande importância na inferência estatística pois asseguram-nos que na inferência sobre a média populacional μ_X (ou sobre uma proporção p), a estatística $\bar{x}$ que calculamos a partir das observações realizadas é, quando o tamanho da amostra é grande, uma aproximação para μ_X. No entanto, a lei dos grandes números não nos permite, por si só, quantificar a confiança que podemos depositar na estimativa $\bar{x}$ de μ_X. Para tal é fundamental que tenhamos informação sobre a distribuição amostral de $\bar{x}$ (ver §2.4), isto é, informação sobre os valores que a estatística $\bar{x}$ toma para as diferentes amostras bem como a probabilidade com que toma esses valores. Este será um assunto que abordaremos num próximo capítulo.

4.7 Bibliografia

Blume, J.D., Royall, R.M. (2003). Illustrating the law of large numbers, *The American Statistician*, 57, 51–55.

Daniel, W.W. (2009). *Biostatistics: a foundation for analysis in the health sciences*, Wiley.

Dawson, B., Trapp, R.G. (2004). *Basic & Clinical Biostatistics*, McGraw-Hill.

Moore, D.S. (1985). *Statistics: concepts and controversies*, W.H. Freeman and Company.

Moore, D.S., McCabe, G.P. (2006). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

5

As distribuições normal e binomial

Densidade normal e variável aleatória normal. Regra 68-95-99,7. Tabela da distribuição normal standard. Cálculos envolvendo a distribuição normal. Gráficos de quantis normais. Experiência aleatória binomial. Variável aleatória binomial: distribuição de probabilidade, média e variância. Cálculos envolvendo a variável binomial. Aproximação normal para a distribuição binomial.

5.1 Introdução

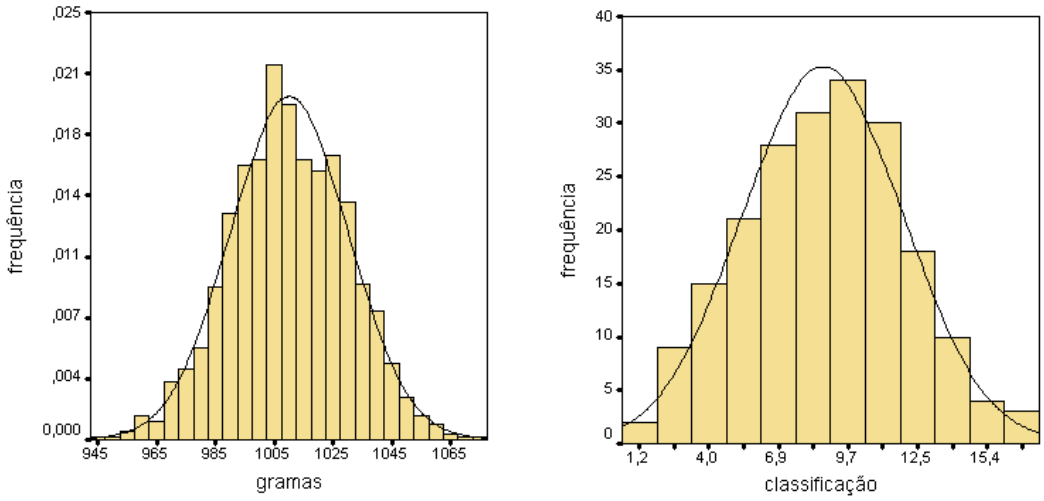
Estudamos neste capítulo duas distribuições de probabilidade, uma contínua e outra discreta, que são das mais usadas na modelização de diversos conjuntos de dados. A primeira, a que chamaremos **distribuição normal**, está associada a várias das experiências aleatórias como as dos Exemplos 1.2.5 (pág. 24) e 1.2.7 (pág. 26), em que o histograma de frequências pode ser razoavelmente aproximado por uma curva densidade simétrica, unimodal e com a forma de um sino. Como veremos no próximo capítulo, a distribuição normal é ainda usada como aproximação das distribuições amostrais de estatísticas como a proporção e a média amostrais tendo, por isso, um papel de destaque na estatística inferencial. A segunda distribuição que estudamos neste capítulo, dita **distribuição binomial**, está relacionada com experiências aleatórias em que contamos as vezes em que determinado acontecimento ocorre quando repetimos uma experiência aleatória um número fixo de vezes.

Apesar da distribuição normal ser contínua e da distribuição binomial ser discreta, veremos que estas duas distribuições de probabilidade estão intimamente relacionadas.

5.2 A distribuição normal

Foram vários os exemplos que apresentámos de variáveis aleatórias contínuas cujo histograma de frequências pode ser mais ou menos aproximado por uma curva densidade **simétrica**, **unimodal** e com a **forma dum sino**. Dois desses exemplos são os casos

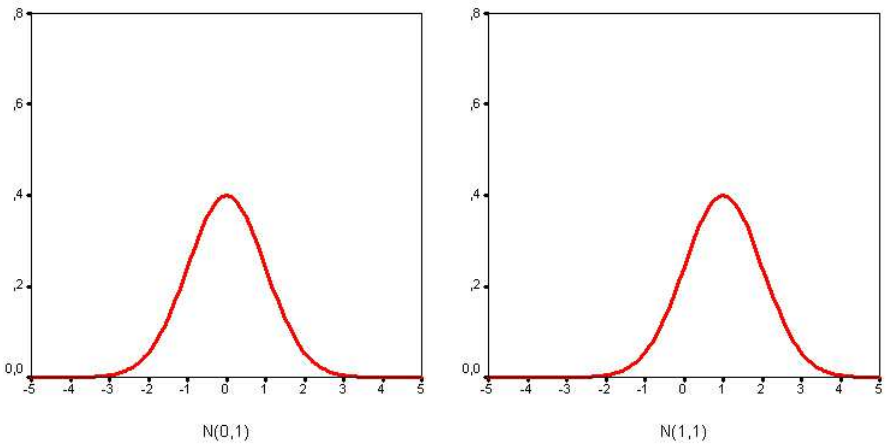
do peso dos pacotes de açúcar de que temos vindo a falar (ver Exemplo 1.2.5, pág. 24) e também o caso da distribuição das classificações de Análise Matemática (ver Exemplo 1.2.7, pág. 26):



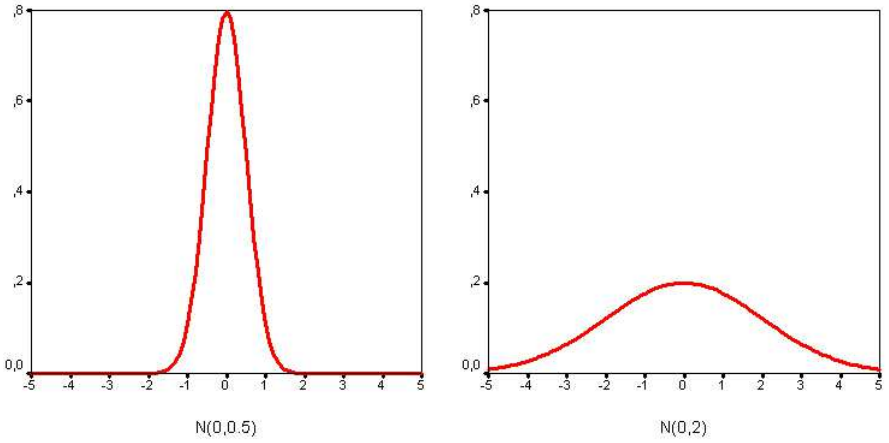
Estas curvas a que chamamos **curvas normais** descrevem distribuições de dados ditas **distribuições normais**. Estas curvas revelam-se muito importantes em estatística. Para justificar parcialmente esta afirmação, referimos o facto de que são várias as distribuições de dados que são bem descritas por curvas normais. Nelas se incluem dados provenientes da cotação de testes ou de medições repetidas duma mesma grandeza (peso, altura, distância). Mais razões para a importância da curva normal surgirão durante o curso. Como veremos, ela surge envolvida em muitos dos procedimentos da estatística inferencial que estudaremos.

Todas as curvas normais têm a mesma forma. São **simétricas**, **unimodais** e têm a **forma dum sino**. Uma curva normal fica completamente determinada pela especificação da sua média μ e do seu desvio-padrão σ. Este facto é claro a partir da expressão analítica que define uma curva normal, em que a cada valor x do eixo das abcissas, corresponde o ponto y do eixo das ordenadas dado por

$$y = \frac{1}{\sigma\sqrt{2\pi}}\, \mathrm{e}^{\,-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$



Curva normal para diferentes valores da média



Curva normal para diferentes valores do desvio-padrão

onde $\pi = 3{,}1415926535\ldots$ é o nosso bem conhecido número Pi da geometria e $e = 2{,}7182818282\ldots$ é o número de Neper.

A média μ duma curva normal está localizada no ponto de simetria da curva e coincide com a mediana. Aumentar μ sem alterar σ corresponde a deslocar horizontalmente a curva para a direita, enquanto que diminuir μ conduz a um deslocamento horizontal da curva para a esquerda. O desvio-padrão σ controla a dispersão da curva normal. Estes factos estão ilustrados nas figuras anteriores.

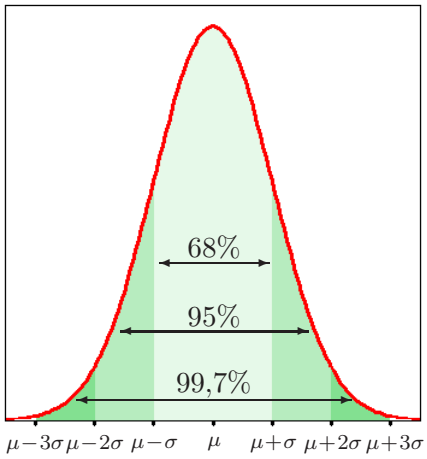
Se representarmos por X a variável que estamos a observar, escrevemos

$$X \sim N(\mu, \sigma)$$

sempre que a distribuição da variável possa ser descrita por uma curva normal com média μ e desvio-padrão σ. Dizemos então que X é uma **variável aleatória normal** com média μ e desvio-padrão σ, ou que X **possui**, ou tem, **uma distribuição normal** de média μ e desvio-padrão σ.

5.2.1 Regra 68-95-99,7

Sabemos já que a probabilidade duma variável contínua tomar valores num qualquer intervalo que marquemos no eixo dos xx é igual à área da região compreendida entre a sua **curva densidade** e o eixo dos xx que tem por base esse intervalo. Fazendo o cálculo das áreas correspondentes aos intervalos $[\mu-\sigma, \mu+\sigma]$, $[\mu-2\sigma, \mu+2\sigma]$ e $[\mu-3\sigma, \mu+3\sigma]$, quando a variável é $N(\mu, \sigma)$ (mais à frente veremos como podemos calcular tais áreas), obtemos para a frequência percentual destes intervalos os valores seguintes:



As propriedades seguintes, conhecidas como **regra 68-95-99,7**, são assim válidas para todas as distribuições normais:

Regra 68-95-99,7:
Numa **distribuição normal** de média μ e desvio-padrão σ:

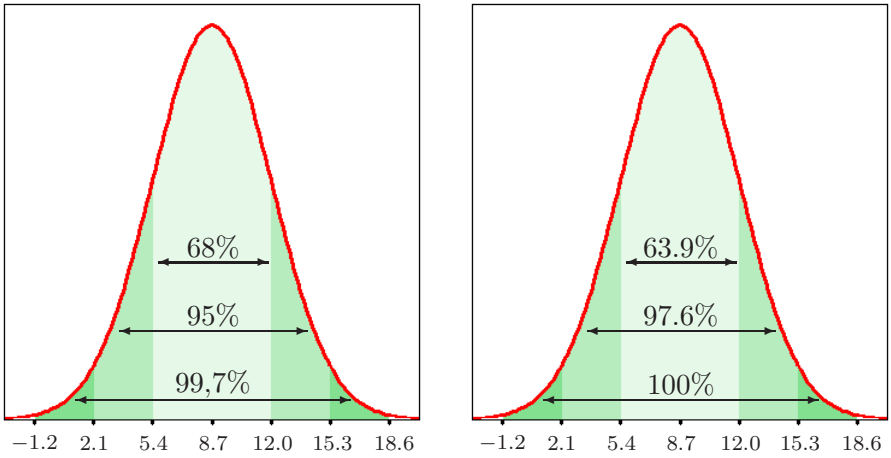
⊙ aproximadamente 68% das observações estão no intervalo $[\mu-\sigma, \mu+\sigma]$;

⊙ aproximadamente 95% das observações estão no intervalo $[\mu-2\sigma, \mu+2\sigma]$;

⊙ aproximadamente 99,7% das observações estão no intervalo $[\mu-3\sigma, \mu+3\sigma]$.

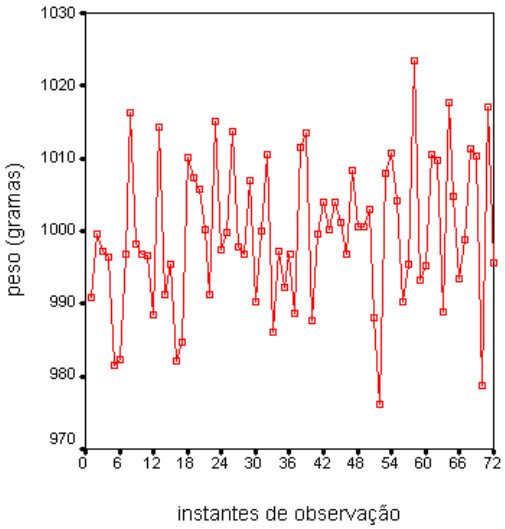
Exemplo 5.2.1 Admitindo que a distribuição das classificações de Análise Matemática, cujo histograma é apresentado no início do §5.2, possui uma distribuição normal $N(8{,}7; 3{,}3)$, onde 8,7 e 3,3 são, respetivamente, aproximações às décimas da média e do

desvio-padrão do conjunto das classificações, mostramos a seguir a distribuição esperada das classificações dada pela regra 68-95-99,7 e a distribuição efetivamente observada. Estes resultados, reforçam a ideia de que a distribuição das classificações de Análise é bem aproximada por uma distribuição normal. Utilizando esta regra podemos concluir que a frequência relativa das classificações superiores a 15,3 é aproximadamente de 2,5%. Reparemos que o valor observado para esta frequência relativa foi de 1,95%, uma vez que 4 dos 205 alunos que realizaram a prova obtiveram nota superior a 15,3 valores.

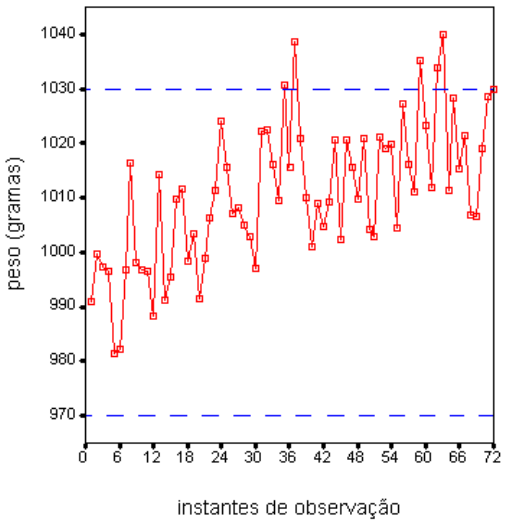


Regra 68-95-99,7 teórica e observada

Exemplo 5.2.2 É por vezes interessante analisar a evolução duma variável com o tempo. Admitamos que a distribuição dos pesos dos pacotes de açúcar produzidos por uma máquina bem calibrada possui uma distribuição normal com 1000 gramas de média e com desvio-padrão de 10 gramas (ver histograma no início do §5.2). Para controlar o processo de empacotamento, de 10 em 10 minutos é recolhido um pacote de açúcar empacotado pela máquina e é registado o seu peso. Este tipo de observações pode ser descrito graficamente por um **gráfico sequencial**, representando os pontos (t, y_t), eventualmente ligados com segmentos de reta, onde t é o instante de observação e y_t o peso observado, num sistema de eixos coordenados. O gráfico sequencial seguinte dá conta dos pesos registados durante um período de 12 horas de funcionamento da máquina. Pela regra 68-95-99,7, sabemos que 99,7% dos pesos registados pertence ao intervalo $[970, 1030]$. Assim, 99,7% dos pontos marcados deve estar entre as retas horizontais $y = 970$ e $y = 1030$. Como podemos verificar, tal acontece com todas as observações anteriores.



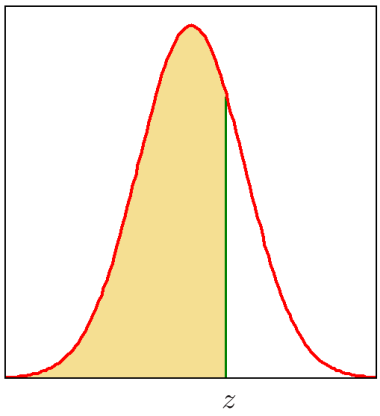
No gráfico seguinte registam-se os pesos de pacotes de açúcar recolhidos, como acima se indicou, durante um outro período de 12 horas de funcionamento da máquina. Nele se põe em evidência uma alteração da distribuição do peso dos pacotes de açúcar. A partir do instante de observação 18 (aproximadamente) é clara uma tendência de aumento do peso dos pacotes observados, que culmina com duas observações, a 35 e a 38, a excederem o limite superior de variação. Significa isto que a máquina ficou descalibrada produzindo pacotes com peso a mais. Se o gráfico for construído, não *a posteriori*, mas à medida que as observações vão sendo feitas, podemos controlar o funcionamento da máquina e proceder a uma imediata calibragem da mesma evitando que durante o resto do período de funcionamento a máquina produza pacotes com peso excessivo. Por estas razões, estes gráficos são, neste contexto, designados por **cartas de controlo**.



5.2.2 Cálculos envolvendo a distribuição normal

Como vimos em §4.2.2, se os dados $x_1, \ldots, x_n$ resultantes da observação duma variável X, puderem ser descritos por uma curva densidade, para calcular a probabilidade de X tomar valor num intervalo que consideremos no eixo dos xx, é importante saber determinar a área da região compreendida entre a curva densidade e o eixo horizontal que tem por base esse intervalo.

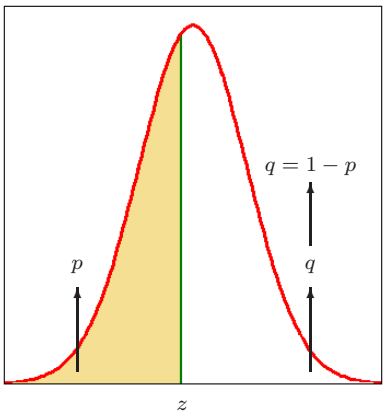
No caso da distribuição de X ser normal com média 0 e desvio-padrão 1, dita **distribuição normal standard** ou distribuição normal **centrada** (por ter média 0) e **reduzida** (por ter desvio-padrão 1), um tal cálculo pode ser feito com a ajuda duma tabela da distribuição normal standard como é o caso da Tabela B (pág. 309). Para cada valor z do eixo do xx, encontramos na Tabela B o valor da área da região compreendida entre a curva normal média 0 e desvio-padrão 1 e o eixo horizontal que está à esquerda de z:



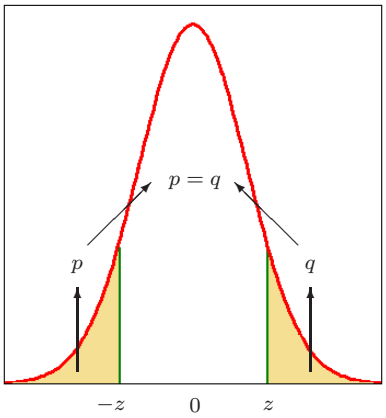
Tal como para a distribuição respetiva, a esta curva normal chamamos **curva normal standard** ou curva normal **centrada** (por ter média 0) e **reduzida** (por ter desvio-padrão 1).

Atendendo a que a área sob uma curva densidade é igual a 1, e que a curva normal standard é simétrica relativamente ao ponto $z = 0$, outras áreas sob a curva normal standard podem ser obtidas a partir das que tiramos diretamente da Tabela B.

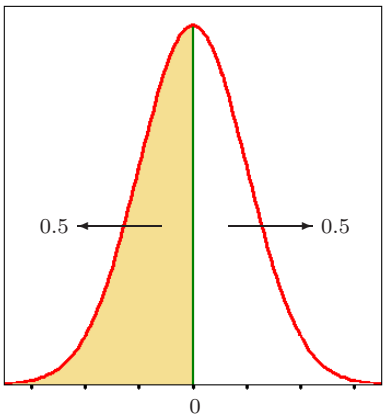
Assim, se a área à esquerda de um ponto z é igual a p, a área à sua direita é igual a $1 - p$:



Pela simetria da curva, as áreas à esquerda de um ponto $-z$ e à direita do seu simétrico z são iguais:

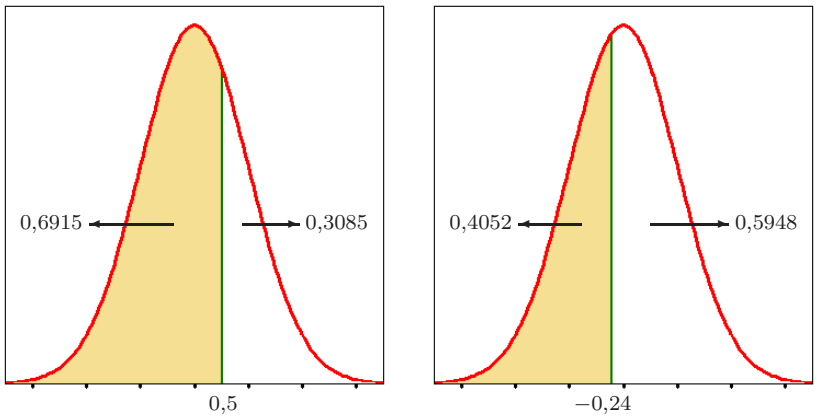


Em particular, as áreas à esquerda e à direita do ponto $z = 0$ são iguais a 0,5:

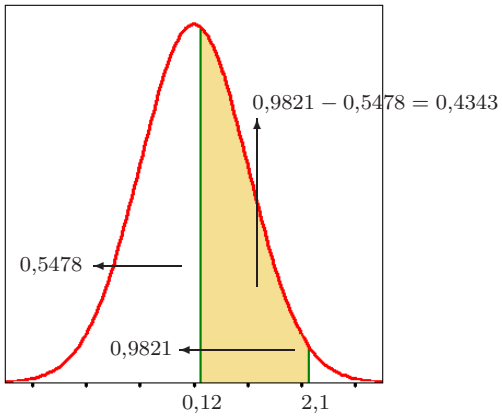


Exemplificamos a seguir a utilização da Tabela B, em alguns casos particulares relevantes.

Exemplo 5.2.3 Para $z = 0{,}50$, obtemos, para área da região compreendida entre a curva e o eixo horizontal que está à esquerda de 0,50, o valor 0,6915. Como a área total sob a curva é igual a 1, a área da região compreendida entre a curva e o eixo horizontal que está à direita de 0,50 é igual a $1 - 0{,}6915 = 0{,}3085$. De forma análoga, as áreas à esquerda e à direita do ponto $z = -0{,}24$ são iguais a 0,4052 e 0,5948, respetivamente.

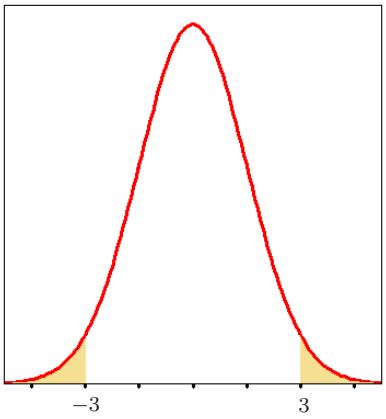


Exemplo 5.2.4 O cálculo de áreas um pouco mais complicadas pode ainda ser feito utilizando a Tabela B. Por exemplo, suponhamos que pretendemos calcular a área da região compreendida entre a curva e o eixo horizontal que está entre os pontos $z = 0{,}12$ e $z = 2{,}10$. O valor desta área pode ser obtido subtraindo ao valor da área à esquerda de $z = 2{,}10$ o valor da área à esquerda de $z = 0{,}12$. Obtemos então o valor $0{,}9821 - 0{,}5478 = 0{,}4343$.



Facilmente se obtém agora a área da região compreendida entre a curva e o eixo horizontal que está à esquerda do ponto $z = 0{,}12$ ou à direita de $z = 2{,}10$: $1 - 0{,}4343 = 0{,}5657$.

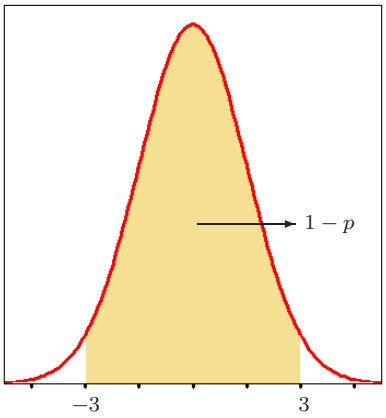
Exemplo 5.2.5 Admitamos que a distribuição dos erros de medida (em milímetros) cometidos por um teodolito nas várias medições de determinada distância, pode ser descrita por uma distribuição normal standard. Representando a variável "erro" por Z, temos então que $Z \sim N(0,1)$. Determinemos a proporção p de medições em que o valor absoluto do erro cometido é superior a 3 milímetros, isto é, a proporção de medições em que $Z < -3$ ou $Z > 3$. O valor pedido pode ser aproximado pela probabilidade da variável Z tomar valores à esquerda de -3 ou à direita de 3, não é mais do que a soma das áreas, sob a curva normal standard, à esquerda de -3 e à direita de 3.



Efetuamos o cálculo de três maneiras diferentes:

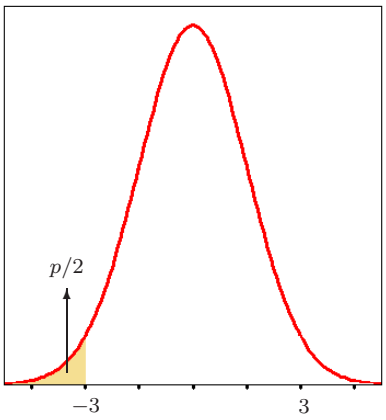
A) De forma direta, somando as áreas, sob a curva normal standard, à esquerda de -3 e à direita de 3, obtemos $p = 0{,}0013 + (1 - 0{,}9987) = 0{,}0026$.

B) Alternativamente, podemos começar por calcular a área da região compreendida entre a curva normal standard e o eixo horizontal que está entre os pontos $z = -3$ e $z = 3$ e que é igual a $1 - p$.



Assim $1 - p = 0{,}9987 - 0{,}0013 = 0{,}9974$, e portanto $p = 0{,}0026$.

C) Podemos ainda começar por observar que a área à esquerda de $z = -3$ é igual a $p/2$, pois esta é igual à área à direita de $z = 3$ (pela simetria da curva normal standard relativamente a $z = 0$).



Como a área à esquerda de $z = -3$ é igual a 0,0013, então $p = 2 \times 0{,}0013 = 0{,}0026$.

Suponhamos agora que a distribuição dos dados $x_1, \ldots, x_n$ resultantes da observação duma variável X, pode ser descrita por uma curva normal de média μ e desvio-padrão σ. Tendo em conta o que estudámos nos parágrafos 1.4 e 5.2, é de esperar que os dados $z_1, \ldots, z_n$ definidos por

$$z_i = \frac{x_i - \mu}{\sigma},$$

correspondentes a uma alteração da unidade de medida, sejam bem descritos por uma curva densidade normal standard (porquê?). Como a variável Z foi obtida da variável X subtraindo-lhe em primeiro lugar a sua média μ e dividindo o resultado obtido pelo seu desvio-padrão σ, dizemos que **padronizámos** a variável X. Como Z tem média 0 e desvio-padrão 1, dizemos também que **centrámos** e **reduzimos** X.

Padronização duma variável normal:
Se

$$X \sim N(\mu, \sigma)$$

então

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

As relações anteriores, exprimem o facto dum problema sobre uma distribuição normal qualquer, poder ser convertido num problema sobre a distribuição normal standard.

Exemplo 5.2.6 Para ilustrar a importância deste facto, retomemos o Exemplo 5.2.1 (pág. 126) e denotemos por X a variável "classificação obtida em Análise Matemática". Admitamos que a sua distribuição é normal com média 8,7 e desvio-padrão 3,3, isto é, $X \sim N(8{,}7; 3{,}3)$. Suponhamos que pretendemos calcular a proporção de classificações inferiores a 8 valores, isto é, queremos calcular a proporção de vezes em que $X < 8$. Sendo tal proporção aproximada pela probabilidade de termos $X < 8$, que indicamos por $\mathrm{P}(X < 8)$, calculemos esta probabilidade. Subtraindo a média e dividindo pelo desvio-padrão, isto é, **centrando e reduzindo** X, vamos converter este problema num problema sobre a distribuição normal standard:

$$\begin{aligned} X &< 8 \\ X - 8{,}7 &< 8 - 8{,}7 \\ (X - 8{,}7)/3{,}3 &< (8 - 8{,}7)/3{,}3 \\ Z &< -0{,}21 \end{aligned}$$

onde $Z \sim N(0, 1)$. Assim, a probabilidade pedida não é mais do que a probabilidade da variável normal standard ser inferior a $-0{,}21$. Sabemos já que esta última probabilidade é dada pela área da região compreendida entre a curva normal standard e o eixo horizontal que está à esquerda do ponto $z = -0{,}21$. Uma tal área é aproximadamente igual a 0,4168:

$$\mathrm{P}(X < 8) = \mathrm{P}(Z < -0{,}21) = 0{,}4168.$$

Como as classificações são sempre positivas, poderíamos também optar por calcular a probabilidade de obter classificações para as quais $0 \leq X < 8$, onde $X \sim N(8{,}7; 3{,}3)$. Procedendo como atrás, obteríamos o valor 0,4127:

$$\mathrm{P}(0 \leq X < 8) = \mathrm{P}(-2{,}64 \leq Z < -0{,}21) = 0{,}4168 - 0{,}0041 = 0{,}4127.$$

Estamos agora em condições de justificar a **regra 68-95-99,7** que afirmámos ser válida para qualquer distribuição normal $N(\mu, \sigma)$. Usando o procedimento anterior, verifiquemos que é de aproximadamente 68% a frequência relativa das observações X para as quais

$$\mu - \sigma \leq X \leq \mu + \sigma$$

quando $X \sim N(\mu, \sigma)$. Calculemos então a probabilidade do acontecimento anterior. Subtraindo a média μ e dividindo pelo desvio-padrão σ obtemos:

$$\begin{aligned} \mu - \sigma \leq \quad & X \quad & \leq \mu + \sigma \\ -\sigma \leq \quad & X - \mu \quad & \leq \sigma \\ -1 \leq \quad & (X - \mu)/\sigma \quad & \leq 1 \\ -1 \leq \quad & Z \quad & \leq 1 \end{aligned}$$

onde $Z \sim N(0,1)$. Assim, usando a tabela da distribuição normal standard concluímos que

$$\begin{aligned} \mathrm{P}(\mu - \sigma \leq X \leq \mu + \sigma) &= \mathrm{P}(-1 \leq Z \leq 1) \\ &= 0{,}8413 - 0{,}1587 \\ &= 0{,}6826. \end{aligned}$$

De igual forma procederíamos para calcular aproximações para as frequências relativas das observações X para as quais $\mu - 2\sigma \leq X \leq \mu + 2\sigma$ e $\mu - 3\sigma \leq X \leq \mu + 3\sigma$.

5.2.3 Julgando a assunção de normalidade

Como vimos, um histograma, ou um gráfico de extremos-e-quartis, pode revelar caraterísticas da distribuição em estudo, como assimetrias e existência de elevado número de observações discordantes, que não são compatíveis com a assunção de normalidade.

Quando o histograma é aproximadamente simétrico e unimodal, revelando uma forma de sino, é importante ter um instrumento sensível para julgar da justeza da assunção de normalidade, uma vez que a decisão de descrever a distribuição das observações por uma curva normal pode determinar passos futuros na análise dos dados. O instrumento gráfico mais útil para julgar a hipótese de normalidade é o chamado **gráfico de quantis normais**.

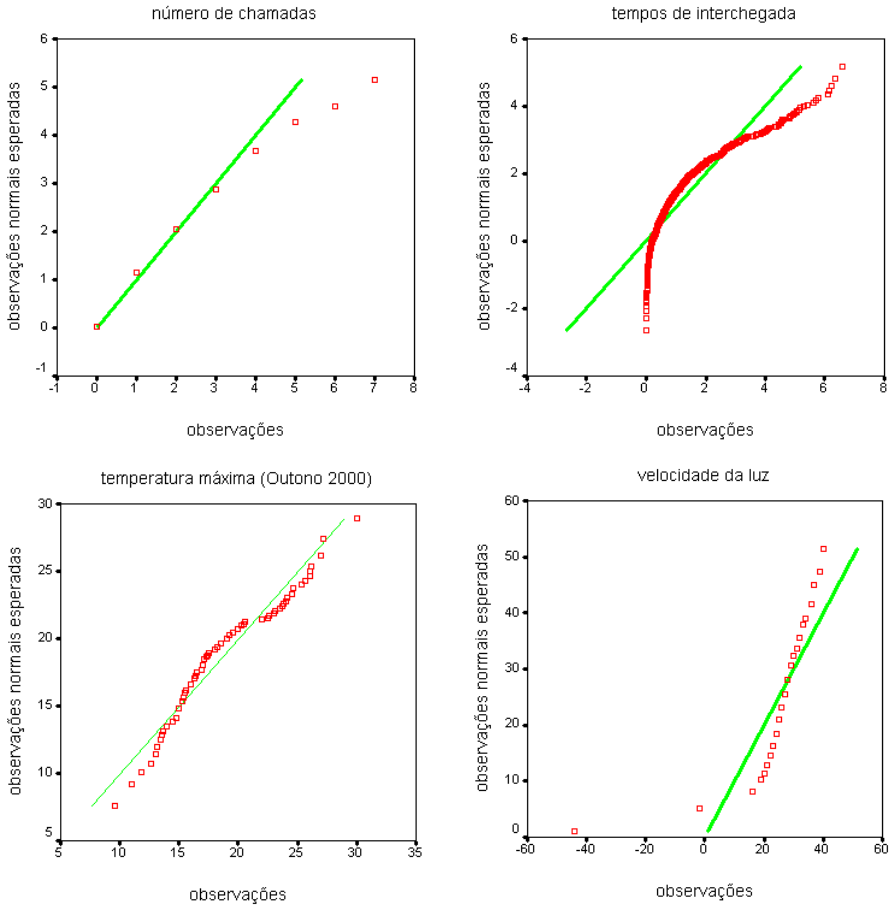
A ideia que está na base deste gráfico, é a comparação dos quantis do conjunto das observações com os quantis correspondentes da distribuição normal que tem por média a média das observações e por desvio-padrão o desvio-padrão das observações. Para cada observação x e para cada quantil z que associamos a x, o ponto (x, z) é marcado num sistema de eixos coordenados.

A **interpretação dum gráfico de quantis normais** é muito simples: se os pontos assim marcados estiverem próximos da reta $x = z$, não apresentando desvios sistemáticos relativamente à reta, o gráfico indica que a distribuição dos dados é normal. Desvios sistemáticos relativamente à reta $x = z$, são indicadores de não normalidade.

Não sendo estes gráficos fáceis de fazer sem auxílio dum computador, vamos limitar-nos no que se segue a analisar alguns gráficos de quantis normais para alguns dos conjuntos de dados que temos vindo a analisar.

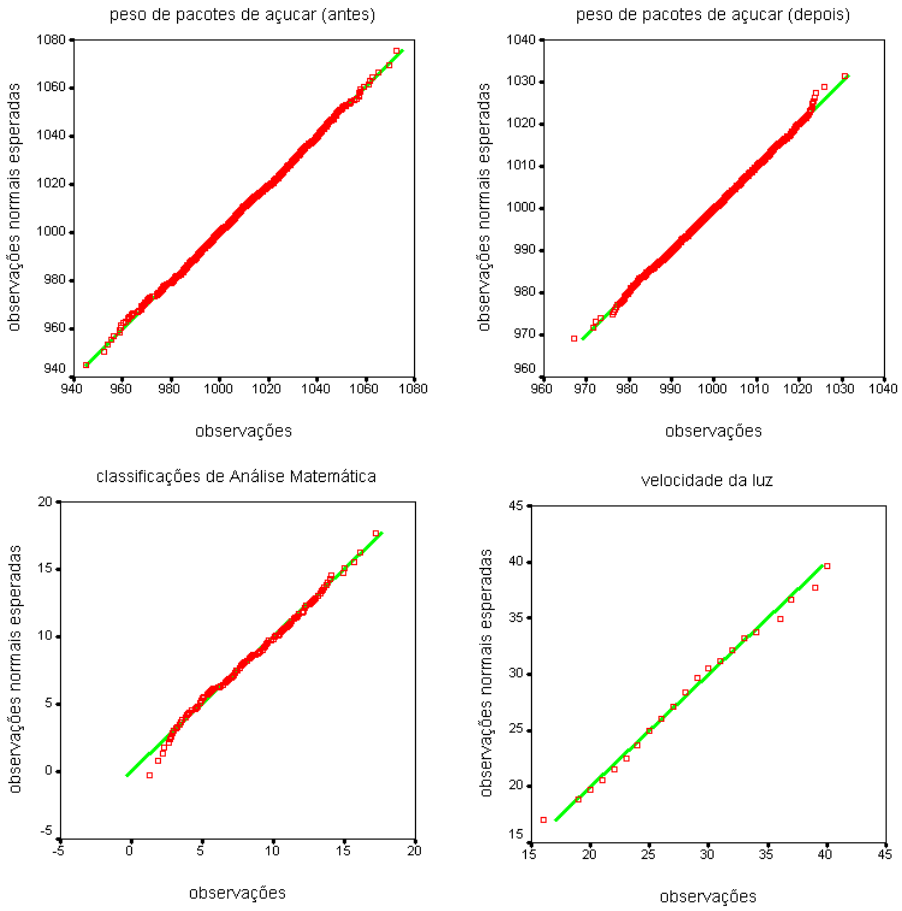
Comecemos pelas distribuições descritas nos Exemplos 1.2.8 (pág. 26), 1.2.9 (pág. 27) e 1.2.6 (pág. 25), cujos histogramas revelam padrões claros de não normalidade. Os gráficos de quantis normais apresentados a seguir confirmam esta ideia. Estes gráficos dão-nos indicações importantes sobre as caudas das distribuições, isto é, sobre os menores e maiores valores da distribuição. Vejamos, por exemplo, o gráfico relativos aos

tempos de interchegada que revela uma cauda esquerda mais curta do que a normal (os pontos da lado esquerdo estão abaixo da reta) e uma cauda direita mais longa que a normal (os pontos do lado direito estão abaixo da reta). Trata-se, por isso, duma distribuição com assimetria positiva. Estas caraterísticas são confirmadas pelo histograma respetivo (ver pág. 26). Reparemos também no facto das duas observações discordantes nos dados relativos à medição da velocidade da luz (ver Exemplo 1.2.6, pág. 25), surgirem fora do padrão comum às restantes observações. Finalmente, e como podemos constatar do primeiro dos gráficos seguintes, reparemos que na execução de gráficos de quantis normais, o SPSS representa com um único ponto observações repetidas.



Distribuições não normais

A assunção de normalidade das distribuições descritas nos Exemplos 1.2.5 (pág. 24) e 1.2.7 (pág. 26), é reforçada pelos gráficos de quantis normais apresentados a seguir. Relativamente à distribuição descrita no Exemplo 1.2.6 (pág. 25), é interessante notar que se excluirmos do conjunto dos dados as duas observações discordantes, as restantes observações podem ser descritas por uma distribuição normal.



Distribuições normais

5.3 A distribuição binomial

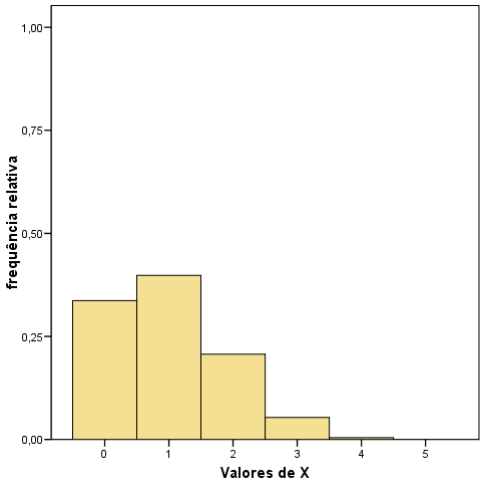
Suponhamos que lançamos 5 vezes consecutivas um dado equilibrado vulgar e que estamos interessados na variável X que nos dá o número de faces 6 que ocorrem nos 5 lançamentos do dado. Indicam-se a seguir vários resultados da experiência bem como o número de faces 6 obtido em cada caso:

resultado		X	resultado		X
1,1,1,1,1	$\longrightarrow$	0	...	...	...
1,1,1,1,2	$\longrightarrow$	0	1,1,1,6,1	$\longrightarrow$	1
...	...	...	...	...	...
1,1,1,1,6	$\longrightarrow$	1	1,1,1,6,6	$\longrightarrow$	2
1,1,1,2,1	$\longrightarrow$	1	...	...	...
...	...	...	6,6,6,6,6	$\longrightarrow$	5

Neste caso X toma os valores $0, 1, 2, 3, 4, 5$ e a questão que pretendemos resolver é a de saber se é possível ter uma ideia sobre a probabilidade com que X toma cada um dos valores anteriores. Atendendo à lei dos grandes números sabemos que se o número de repetições da experiência for grande

$$\mathrm{P}(X = k) \approx \text{Frequência relativa do valor } k.$$

Assim, para obter uma ideia aproximada das probabilidades anteriores podemos repetir muitas vezes a experiência e calcular as frequências relativas dos acontecimentos anteriores.



O histograma anterior descreve a distribuição de frequências da variável X obtida a partir de 2000 repetições da experiência. Trata-se de uma aproximação do histograma de probabilidade de X.

Voltando à questão anterior, reparemos que não será de estranhar que consigamos calcular de forma exata a distribuição de probabilidade de X uma vez que temos muita informação sobre a experiência: a probabilidade de ocorrência da face 6 em cada lançamento do dado é de 1/6 e os sucessivos lançamentos são independentes uns dos outros (o que nos permite usar a propriedade P.6 da probabilidade).

5.3.1 Experiência aleatória binomial

A situação descrita do lançamento dum dado um número n de vezes, possui características que são comuns a muitas outras experiências aleatórias:

Experiência aleatória binomial:

1. São realizadas n observações.
2. As n observações são independentes.
3. Cada observação pode tomar dois valores possíveis, ditos *sucesso* e *insucesso*, que denotamos por 1 e por 0, respetivamente, que ocorrem sempre que o indivíduo observado possui, ou não, respetivamente, a caraterística em estudo.
4. A probabilidade p dum sucesso é a mesma para cada observação.

Quando se verificam as condições anteriores dizemos que estamos na presença duma **experiência aleatória binomial**.

São experiências aleatórias binomiais:

1. O lançamento duma moeda equilibrada de euro 10 vezes consecutivas e a observação do número de vezes em que ocorre a face portuguesa. Neste caso $n = 10$ e $p = 1/2$.
2. A observação do número de vezes que ocorre a face 6 em 20 lançamentos de um dado equilibrado. Neste caso $n = 20$ e $p = 1/6$.
3. A extração sucessiva, com reposição, de 5 cartas escolhidas ao acaso dum baralho vulgar de 52 cartas em que estamos interessados no número de cartas do naipe de paus que ocorrem nessas 5 cartas. Neste caso $n = 5$ e $p = 13/52 = 1/4$. Reparemos que se a extração das 5 cartas é feita por amostragem aleatória simples, a experiência deixa de ser binomial. Perde-se a independência entre as sucessivas observações e a probabilidade de ocorrer paus em cada observação não é sempre a mesma.
4. Para estimar a percentagem de alunos da UC que concordam com o pagamento de propinas, a partir duma listagem dos alunos da UC escolhe-se ao acaso um aluno e regista-se a sua opinião, "sim" ou "não", sobre o pagamento de propinas. Se o processo anterior for repetido 120 vezes tendo por base a mesma listagem permitindo assim que um aluno seja selecionado mais do que uma vez, isto é, se a amostragem for realizada com reposição, a experiência aleatória é uma experiência binomial com $n = 120$ e p é a proporção de alunos da UC que

concorda com o pagamento de propinas. Tal como no exemplo anterior, se a amostra for uma amostra aleatória simples a experiência só aproximadamente pode ser considerada binomial, uma vez que nem as várias observações são independentes, nem a probabilidade de sucesso se mantém igual a p.

5.3.2 Variável aleatória binomial

Numa experiência aleatória binomial estamos interessados na variável X que nos dá o número total de sucessos ocorridos nas n observações. A variável aleatória X toma os valores

$$0, 1, 2, \ldots, n-1, n,$$

e, como veremos a seguir, a probabilidade com que X toma cada um dos valores anteriores depende apenas do número, n, de observações e da probabilidade, p, de obter um sucesso. Chamar-lhe-emos **variável binomial de parâmetros n e p**, e indicamos

$$X \sim B(n, p)$$

quando queremos dizer que X **é uma variável binomial de parâmetros n e p.**

Quando n é pequeno, é fácil calcular as probabilidades $\mathrm{P}(X = k)$, para $k = 0, 1, 2, \ldots, n$. Vejamos o que se passa nos casos em que $n = 2$ e $n = 3$.

• No caso $n = 2$ o espaço dos resultados é

$$\Omega = \{00, 01, 10, 11\},$$

onde, pela independência (reparemos que não podemos usar a definição clássica pois os acontecimentos elementares não são, com exceção do caso $p = 0{,}5$, igualmente prováveis):

$$\begin{aligned}
&\mathrm{P}(\{00\}) = (1-p)(1-p) = (1-p)^2,\\
&\mathrm{P}(\{01\}) = (1-p)p,\\
&\mathrm{P}(\{10\}) = p(1-p)\\
&\mathrm{P}(\{11\} = pp = p^2.
\end{aligned}$$

Assim,

$$\begin{aligned}
&\mathrm{P}(X = 0) = \mathrm{P}(\{00\}) = (1-p)^2,\\
&\mathrm{P}(X = 1) = \mathrm{P}(\{01, 10\}) = \mathrm{P}(\{01\}) + \mathrm{P}(\{10\}) = 2p(1-p),\\
&\mathrm{P}(X = 2) = \mathrm{P}(\{11\}) = p^2.
\end{aligned} \tag{5.3.1}$$

• No caso $n = 3$ o espaço dos resultados é

$$\Omega = \{000, 001, 010, 100, 011, 101, 110, 111\},$$

e, pela independência,

$$\begin{aligned}
&\mathrm{P}(\{000\}) = (1-p)^3,\\
&\mathrm{P}(\{001\}) = \mathrm{P}(\{010\}) = \mathrm{P}(\{100\}) = p(1-p)^2,\\
&\mathrm{P}(\{001\}) = \mathrm{P}(\{101\}) = \mathrm{P}(\{110\}) = p^2(1-p),\\
&\mathrm{P}(\{111\} = p^3.
\end{aligned}$$

Assim,

$$\begin{aligned}
&\mathrm{P}(X=0) = \mathrm{P}(\{000\} = (1-p)^3,\\
&\mathrm{P}(X=1) = \mathrm{P}(\{001,010,100\}) = 3p(1-p)^2,\\
&\mathrm{P}(X=2) = \mathrm{P}(\{001,101,110\} = 3p^2(1-p),\\
&\mathrm{P}(X=3) = \mathrm{P}(\{111\} = p^3.
\end{aligned} \tag{5.3.2}$$

Reparemos que os coeficientes $1,2,1$ e $1,3,3,1$ que surgem nas fórmulas (5.3.1) e (5.3.2), não são mais do que o número de vezes em que como resultado duma experiência binomial não ocorre nenhum sucesso, ocorre 1 sucesso, ocorrem 2 sucessos, e assim sucessivamente, até ao último caso em que ocorrem n sucessos. Estes coeficientes são chamados **coeficientes binomiais**. No caso geral dum qualquer valor de n podemos concluir que o número de vezes em que ocorrem k sucessos, para $k = 0,1,2,\ldots,n$, nos 2^n resultados possíveis duma experiência binomial, é dado pelo **coeficiente binomial** C_k^n definido por

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

onde $n!$ é o **fatorial de** n definido por

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1 \quad \text{e} \quad 0! = 1.$$

Conhecido o coeficiente binomial C_k^n, que para alguns valores de n é dado na Tabela C (pág. 313), é agora fácil calcular a probabilidade de obter k sucessos numa experiência binomial: basta multiplicar o número de vezes em que ocorrem k sucessos nos resultados da experiência binomial, C_k^n, pela probabilidade, $p^k(1-p)^{n-k}$, dum qualquer resultado elementar da experiência em que ocorrem k sucessos.

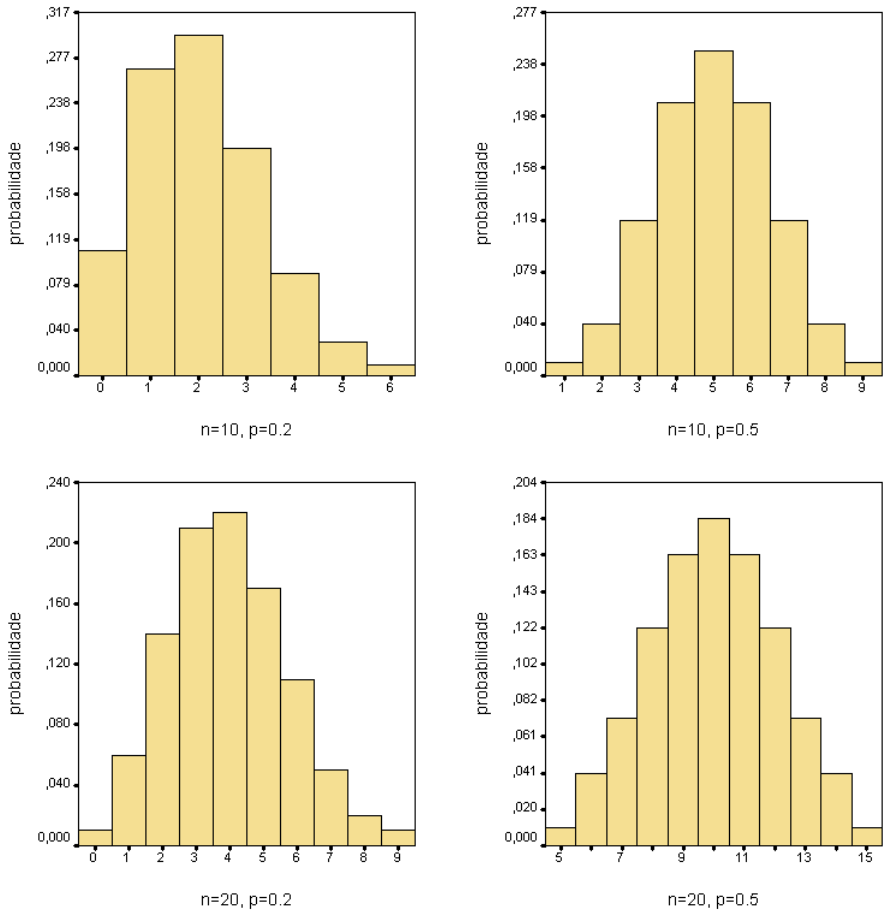
Distribuição de probabilidade duma variável binomial:

Se $X \sim B(n,p)$, então

$$\mathrm{P}(X=k) = C_k^n \; p^k \; (1-p)^{n-k},$$

para $k = 0,1,\ldots,n$.

Nos gráficos seguintes apresentam-se histogramas de probabilidade duma variável binomial para alguns valores de n e p (não estão representados todos os valores da variável binomial). Reparemos na assimetria positiva (resp. negativa) que ocorre para valores pequenos de p (resp. grandes). À medida que p se aproxima de 0,5, a assimetria da distribuição diminui, obtendo-se uma distribuição perfeitamente simétrica quando $p = 0{,}55$.



5.3.3 Média e variância duma variável binomial

Conhecida a distribuição de probabilidade duma variável aleatória, é possível calcular a sua média e a sua variância. No caso duma variável binomial X, devido à forma não muito simples da sua distribuição de probabilidade, é preferível efetuar o cálculo da média μ_X e da variância σ_X^2 utilizando um método alternativo.

Uma variável binomial dá-nos o número de sucessos que ocorrem numa experiência aleatória binomial. Se representarmos por S_i a variável que toma o valor 1 se ocorre sucesso na observação i e 0 se não ocorre sucesso nessa observação, então o número X

de sucessos na experiência é dado por

$$X = S_1 + S_2 + \ldots + S_n. \tag{5.3.3}$$

Como a probabilidade de sucesso numa experiência binomial é p, a distribuição de probabilidade de cada uma das variáveis S_i é dada por

valores de S_i	1	0
probabilidade	p	$1-p$

e a sua média e variância podem ser facilmente calculadas:

$$\mu_{S_i} = 1 \times p + 0 \times (1-p) = p$$

$$\sigma^2_{S_i} = 1^2 \times p + 0^2 \times (1-p)^2 - p^2 = p(1-p).$$

Usando agora a igualdade (5.3.3) e as propriedades já estudadas da média, concluímos que

$$\begin{aligned} \mu_X &= \mu_{S_1} + \mu_{S_2} + \ldots + \mu_{S_n} \\ &= p + p + \ldots + p \\ &= np, \end{aligned}$$

e, pela independência das variáveis $S_1, S_2, \ldots, S_n$,

$$\begin{aligned} \sigma^2_X &= \sigma^2_{S_1} + \sigma^2_{S_2} + \ldots + \sigma^2_{S_n} \\ &= p(1-p) + p(1-p) + \ldots + p(1-p) \\ &= np(1-p). \end{aligned}$$

Média e desvio-padrão duma variável binomial:

Se $X \sim B(n,p)$, então

$$\mu_X = np,$$

$$\sigma_X = \sqrt{np(1-p)}.$$

5.3.4 Cálculos envolvendo a variável binomial

Nos dois exemplos seguintes ilustramos dois casos em que a utilização da noção de variável binomial permite simplificar o cálculo de probabilidades associadas a experiências aleatórias binomiais.

Exemplo 5.3.4 Utilizemos a distribuição de probabilidade duma variável binomial para calcular a probabilidade de no lançamento duma moeda equilibrada de euro 10 vezes consecutivas, observarmos apenas 1 face portuguesa. Neste caso $X \sim B(10; 0{,}5)$, e a probabilidade pedida é dada por

$$\mathrm{P}(X = 1) = C_1^{10}\, 0{,}5^1\, 0{,}5^9 = 10 \times 0{,}5 \times 0{,}5^9 \approx 0{,}009766.$$

A probabilidade de obter mais do que 2 faces portuguesas é dada por

$$\begin{aligned}
\mathrm{P}(X > 2) &= 1 - \mathrm{P}(X \leq 2) \\
&= 1 - (\mathrm{P}(X = 0) + \mathrm{P}(X = 1) + \mathrm{P}(X = 2)) \\
&= 1 - (C_0^{10}\, 0{,}5^0\, 0{,}5^{10} + C_1^{10}\, 0{,}5^1\, 0{,}5^9 + C_2^{10}\, 0{,}5^2\, 0{,}5^8) \\
&= 1 - (1 + 10 + 45) \times 0{,}5^{10} \\
&\approx 0{,}9453.
\end{aligned}$$

Como já referimos, da mesma forma se procede se pretendemos calcular a probabilidade de acontecimentos associados a $\widehat{p}$.

Exemplo 5.3.5 No lançamento de um dado equilibrado 20 vezes consecutivas, calculemos a probabilidade de obter mais que 4% de faces 6, isto é, calculemos $\mathrm{P}(\widehat{p} > 0{,}04)$, onde $\widehat{p} = X/20$ com $X \sim B(20, 1/6)$. Assim,

$$\begin{aligned}
\mathrm{P}(\widehat{p} > 0{,}04) &= \mathrm{P}(X/20 > 0{,}04) \\
&= P(X > 0{,}8) \\
&= \mathrm{P}(X \geq 1) \\
&= 1 - \mathrm{P}(X = 0) \\
&= 1 - C_0^{20}(1/6)^0(1 - 1/6)^{20} \\
&= 0{,}9739
\end{aligned}$$

5.3.5 Aproximação normal para a distribuição binomial

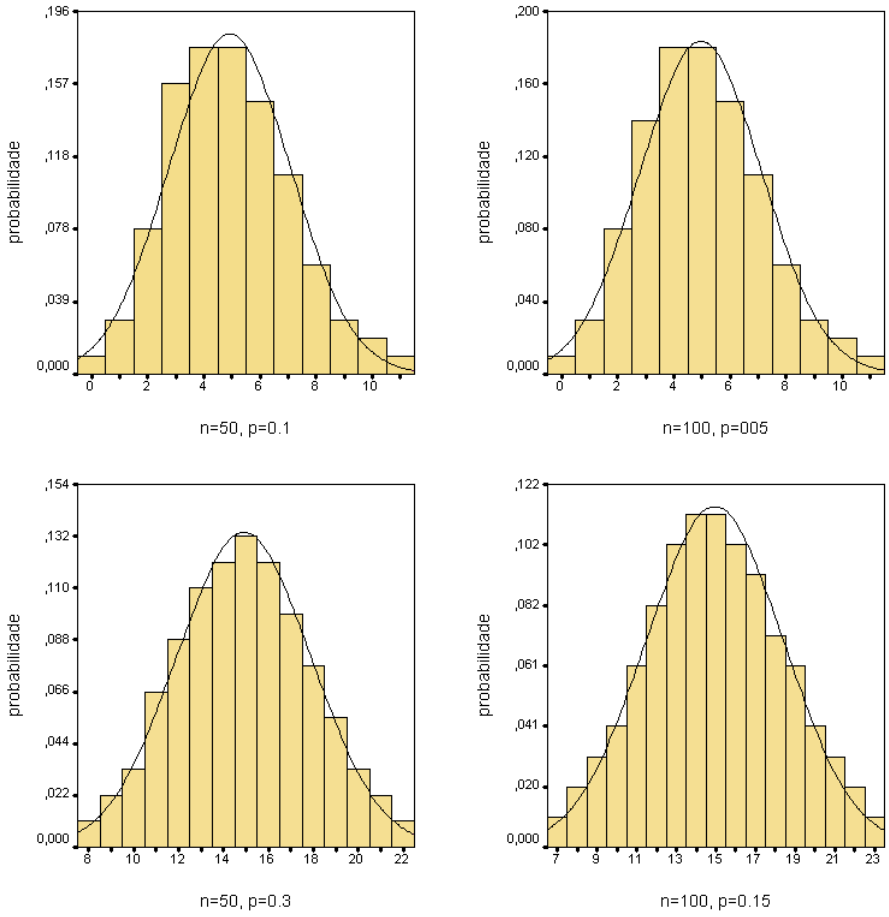
Tal como podem indiciar os histogramas de probabilidade apresentados no final do §5.3.2, a distribuição de probabilidade duma variável binomial $X \sim B(n, p)$ pode ser aproximada por uma curva normal. Tendo em conta o estudo feito no §5.3.3, será natural esperar que uma tal curva normal tenha média np e desvio-padrão $\sqrt{np(1-p)}$.

Aproximação normal para a distribuição binomial:

Se $X \sim B(n,p)$, então

$$X \simeq N\left(np, \sqrt{np(1-p)}\right).$$

Estas aproximações são ilustradas nas figuras seguintes onde, para alguns valores de n e p, com $np = 5, 10$ e 15, se apresentam os histogramas de probabilidade de X e a curva normal respetiva.



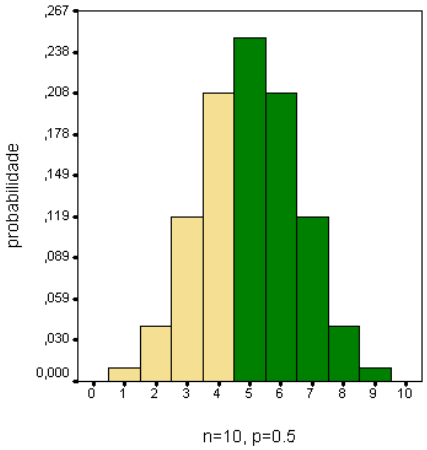
A qualidade da aproximação normal não é a mesma em todos os casos anteriores. Será de esperar que quando p está próximo de 0 ou de 1 (a distribuição binomial é muito assimétrica) a aproximação normal não seja tão boa como no caso em que p está próximo de 0,5 (a distribuição binomial é aproximadamente simétrica). Nos primeiros casos, para que a aproximação seja válida é necessário que n seja grande. Em aplicações

práticas vamos adotar as indicações de Moore e McCabe (2006, p. 344) que consideram que a aproximação normal para as distribuições de X e $\widehat{p}$ é boa se $np \geq 10$ e se $n(1-p) \geq 10$ (regras práticas diferentes das anteriores são consideradas, por exemplo, em Anderson *et al.* , 2002, e McPherson, 1990).

Independentemente do valor de p, a aplicação da regra anterior implica que a aproximação normal seja válida desde que n seja suficientemente grande. Como veremos mais à frente, por detrás deste resultado está o facto da variável X ser, como já vimos no §5.3.3, a soma de variáveis independentes e com a mesma distribuição que no caso da variável binomial tomam o valor 1 se ocorre sucesso e 0 se não ocorre sucesso na i-ésima observação da experiência binomial.

A aproximação normal para a distribuição de X permite simplificar alguns cálculos que seriam complicados de fazer sem o auxílio dum computador.

Exemplo 5.3.6 Suponhamos que pretendemos calcular a probabilidade de no lançamento duma moeda equilibrada de euro 10 vezes consecutivas, observarmos mais do que quatro faces portuguesas. Neste caso $X \sim B(10, 0.5)$, e $\mathrm{P}(X > 4)$ não é mais do que o valor da área representada na figura seguinte:



$$
\begin{aligned}
&\mathrm{P}(X > 4) \\
&= \mathrm{P}(X = 5) + \mathrm{P}(X = 6) + \mathrm{P}(X = 7) + \mathrm{P}(X = 8) + \mathrm{P}(X = 9) + \mathrm{P}(X = 10) \\
&= C_5^{10} 0{,}5^5\, 0{,}5^5 + C_6^{10} 0{,}5^6\, 0{,}5^4 + \ldots + C_9^{10} 0{,}5^9\, 0.5^1 + C_{10}^{10} 0{,}5^{10}\, 0{,}5^0 \\
&= (C_5^{10} + C_6^{10} + C_7^{10} + C_8^{10} + C_9^{10} + C_{10}^{10}) \times 0{,}5^{10} \\
&= (252 + 210 + 120 + 45 + 10 + 1) \times 0{,}5^{10} \\
&= 0{,}623046875.
\end{aligned}
$$

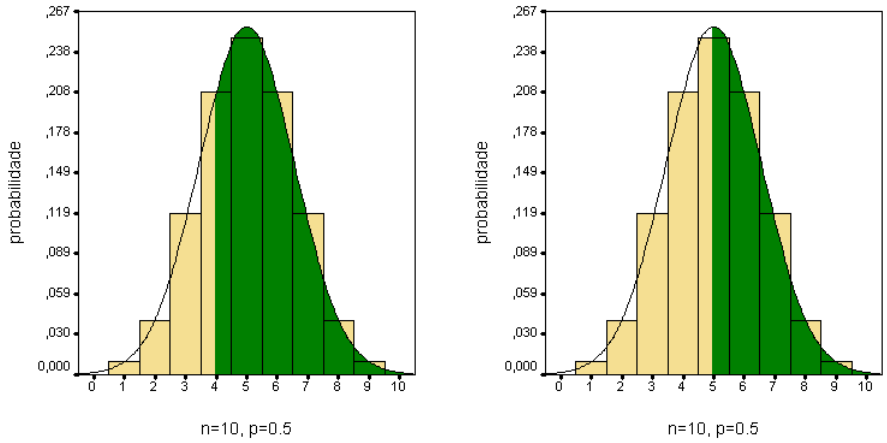
Utilizando a aproximação normal, sabemos que $X \simeq N(5, \sqrt{2{,}5})$. Assim, denotando por Z a variável normal standard, temos

$$\begin{aligned} \mathrm{P}(X > 4) &= \mathrm{P}\left(\frac{X-5}{\sqrt{2{,}5}} > \frac{4-5}{\sqrt{2{,}5}}\right) \\ &\approx \mathrm{P}(Z > -0{,}63) \\ &= 1 - \mathrm{P}(Z \leq -0{,}63) \\ &= 1 - 0{,}2643 = 0{,}7357. \end{aligned}$$

A má qualidade da aproximação pode ser imputada ao facto da condição $np \geq 10$ não ser satisfeita, mas também à forma como utilizámos a variável normal para efetuar a aproximação. Em particular, como $\mathrm{P}(X > 4) = \mathrm{P}(X \geq 5)$ seria também legítimo efetuar a aproximação

$$\begin{aligned} \mathrm{P}(X > 4) &= \mathrm{P}(X \geq 5) \\ &= \mathrm{P}\left(\frac{X-5}{\sqrt{2{,}5}} > \frac{5-5}{\sqrt{2{,}5}}\right) \\ &\approx \mathrm{P}(Z > 0) \\ &= 0{,}5, \end{aligned}$$

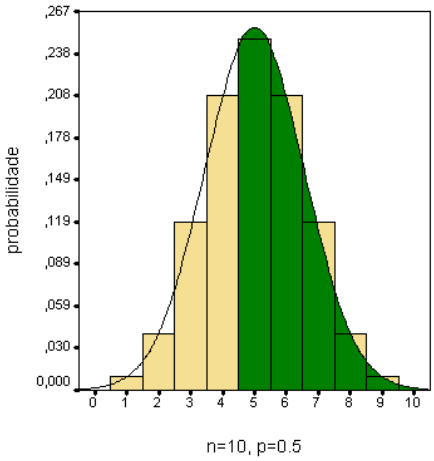
obtendo-se agora uma aproximação por defeito, igualmente fraca, para a probabilidade pretendida.



Quando efetuamos uma aproximação pela normal duma variável binomial, é preferível utilizar não os valores 4 ou 5, mas sim o seu ponto médio, isto é, o valor 4,5. Deste modo obtemos uma aproximação de muito melhor qualidade do que qualquer das aproximações anteriores:

$$\mathrm{P}(X > 4) = \mathrm{P}(X > 4{,}5)$$

$$\begin{aligned} &= \mathrm{P}\left(\frac{X-5}{\sqrt{2{,}5}} > \frac{4{,}5-5}{\sqrt{2{,}5}}\right) \\ &\approx \mathrm{P}(Z > -0{,}32) \\ &= 1 - \mathrm{P}(Z \leq 0{,}32) \\ &= 1 - 0{,}3745 = 0{,}6255. \end{aligned}$$



Esta regra, conhecida como **correção de continuidade**, vale para quaisquer outros valores, e, mais geralmente, sempre que uma variável discreta, que neste exemplo é a variável binomial, seja aproximada por uma variável contínua, que no caso anterior é a variável normal.

5.4 Bibliografia

Dawson, B., Trapp, R.G. (2004). *Basic & Clinical Biostatistics*, McGraw-Hill.

Moore, D.S. (1985). *Statistics: concepts and controversies*, W.H. Freeman and Company.

Gomes, M.I., Barão, M.I. (1999). *Controlo Estatístico de Qualidade*, SPE.

McPherson, G. (1990). *Statistics in Scientific Investigation: its basis, application, and interpretation*, Springer.

Moore, D.S., McCabe, G.P. (2006). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

Pestana, D.D., Velosa, S.F. (2002). *Introdução à Probabilidade e à Estatística*, Vol. I, Fundação Calouste Gulbenkian.

6

Distribuições amostrais para proporções e médias

Distribuição amostral duma estatística. Distribuição amostral de $\widehat{p}$. Média e desvio-padrão de $\widehat{p}$. Aproximação normal para a distribuição de $\widehat{p}$. Distribuição amostral de $\bar{x}$. Média e desvio-padrão de $\bar{x}$. Teorema do limite central.

6.1 Distribuição amostral duma estatística

Quando na realização dum estudo observacional por amostragem pretendemos conhecer a proporção, p, de indivíduos da população em estudo que possuem determinada caraterística (proporção de votantes num determinado partido político, proporção de famílias com baixos rendimentos, etc.), ou a média, μ, de determinada caraterística numérica associada à população (peso médio, altura média, rendimento mensal médio, etc.), a inferência sobre esses parâmetros desconhecidos é baseada numa amostra recolhida dessa população.

Para essa amostra, e com o objetivo de inferir sobre o parâmetro desconhecido de interesse, calculamos normalmente a estatística associada a esse parâmetro: se o parâmetro é uma proporção, calculamos a proporção, $\widehat{p}$, de indivíduos na amostra que possuem a propriedade em estudo; se o parâmetro é uma média, calculamos a média, $\bar{x}$, dos valores da amostra.

Como já referimos no §2.4, diferentes amostras conduzem a diferentes valores para as estatísticas $\widehat{p}$ e $\bar{x}$, facto este a que chamámos variabilidade amostral. Estas estatísticas funcionam assim como variáveis aleatórias: a cada amostra aleatória, que aqui toma o papel de resultado da experiência aleatória, associam um valor numérico.

Assim sendo, faz sentido falar na distribuição de probabilidade de tais estatísticas a que chamamos distribuição amostral da estatística em causa. Uma tal distribuição dá-nos os valores que a estatística toma para as diferentes amostras bem como a probabilidade com que os toma.

Neste capítulo estudaremos a **distribuição amostral** das estatísticas $\widehat{p}$ e $\bar{x}$ que, como veremos, surgem em muitos problemas de inferência estatística. Nos capítulos seguintes, ilustraremos a sua aplicação a dois problemas muito importantes do âmbito da estatística inferencial como são os casos dos intervalos de confiança e dos testes de hipóteses.

6.2 Distribuição amostral de $\widehat{p}$

Suponhamos que lançamos n vezes consecutivas um dado que suspeitamos não ser equilibrado, e que estamos interessados na proporção $\widehat{p}$ de faces 6 que obtemos nos lançamentos realizados. Se representarmos por X o número de faces 6 obtidas nos n lançamentos, $\widehat{p}$ é dada por

$$\widehat{p} = \frac{X}{n},$$

que, pela lei dos grandes números, sabemos ser uma aproximação da probabilidade de ocorrência da face 6, quando n é grande. Denotando por p essa probabilidade (desconhecida), sabemos já que a variável X é uma variável binomial de parâmetros n e p, $X \sim B(n,p)$. Assim, quando X toma o valor k, para algum $k = 0, 1, 2, \ldots, n-1, n$, a variável $\widehat{p}$ toma o valor k/n, sendo por isso iguais as probabilidades com que tais valores ocorrem:

$$\mathrm{P}\left(\widehat{p} = \frac{k}{n}\right) = \mathrm{P}(X = k).$$

A **distribuição amostral** de $\widehat{p}$ pode assim ser obtida a partir da distribuição amostral da variável X que conhecemos já no contexto duma experiência aleatória binomial (ver §5.3.2):

Distribuição de probabilidade de $\widehat{p}$:

Numa experiência binomial temos

$$\mathrm{P}\left(\widehat{p} = \frac{k}{n}\right) = C_k^n \; p^k \; (1-p)^{n-k},$$

para $k = 0, 1, \ldots, n$.

Tendo agora em conta que $\widehat{p} = X/n$, e que conhecemos a média e o desvio-padrão de X (ver §5.3.3), podemos facilmente calcular a média e variância da proporção $\widehat{p}$ numa experiência binomial :

$$\mu_{\widehat{p}} = \frac{\mu_X}{n} = \frac{np}{n} = p$$

e

$$\sigma^2_{\widehat{p}} = \frac{\sigma^2_X}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Média e desvio-padrão da proporção $\widehat{p}$:

Numa experiência binomial temos

$$\mu_{\widehat{p}} = p,$$

$$\sigma_{\widehat{p}} = \sqrt{p(1-p)/n}.$$

Realcemos o significado e a importância de cada uma das igualdades anteriores. Para tal, centremo-nos no exemplo com que começámos esta secção em que um dado (não necessariamente equilibrado) é lançado n vezes e que pretendemos inferir sobre a probabilidade de ocorrência da face 6, probabilidade essa que representámos por p.

Ao dizermos que a média de $\widehat{p}$ é exatamente igual ao valor desconhecido p sobre o qual pretendemos inferir, estamos a dizer que se fizéssemos várias vezes n lançamentos do dado, as várias proporções amostrais que se obteriam teriam uma distribuição com centro em p. Além disso, uma vez que a variabilidade respetiva decresce à medida que n aumenta, essas diversas proporções amostrais estariam mais próximas de p à medida que aumentássemos o número de lançamentos n.

Sabemos também que a distribuição binomial pode ser aproximada pela distribuição normal. Será por isso de esperar que também a distribuição amostral de $\widehat{p}$ possa ser aproximada pela distribuição normal.

Aproximação normal para a distribuição de $\widehat{p}$:

Numa experiência binomial temos

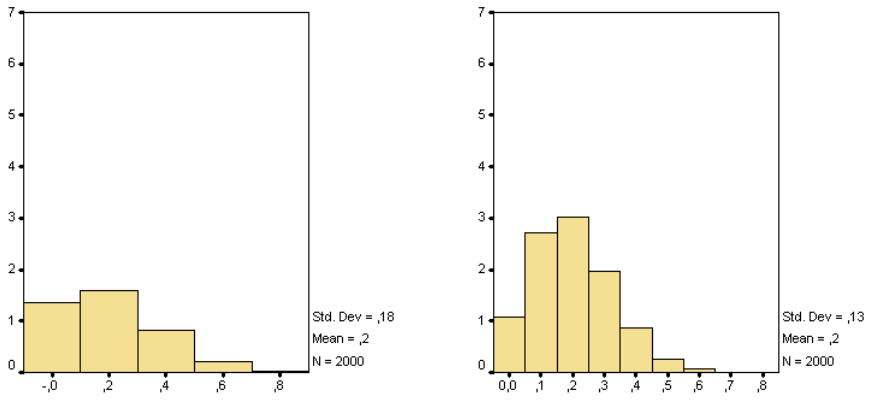
$$\widehat{p} = X/n \simeq N\left(p, \sqrt{p(1-p)/n}\right)$$

Quando a população é finita e a amostra é recolhida por amostragem aleatória simples, as diversas observações não satisfazem as caraterísticas 2. e 4. duma experiência

binomial (ver §5.3.1). No entanto, quando o tamanho da população é grande relativamente à dimensão n da amostra recolhida, podemos ignorar a dependência fraca que existe entre as sucessivas observações e a pequena alteração da probabilidade de ocorrência de sucesso. Assim, quando o tamanho da população é de pelo menos 10 vezes a dimensão da amostra, e a amostra é uma amostra aleatória simples de tamanho n, a distribuição da variável X pode ser considerada aproximadamente binomial $B(n, p)$, onde p é a proporção de sucessos na população.

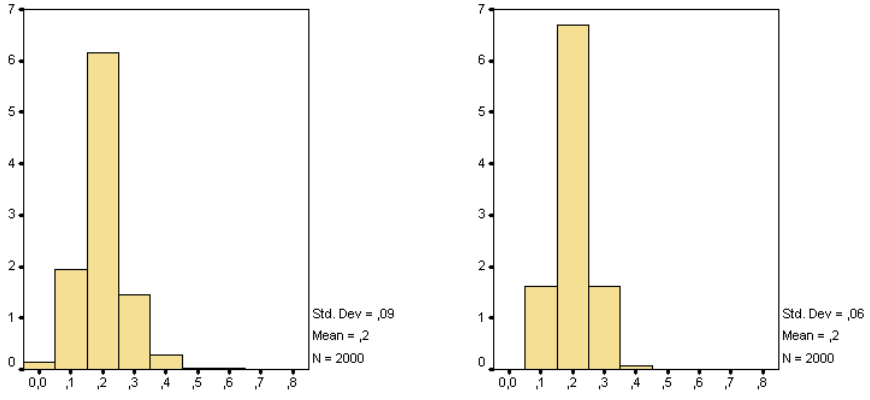
Nos exemplos seguintes exemplificamos cada uma caraterísticas teóricas anteriores sobre a distribuição da proporção amostral.

Exemplo 6.2.1 Suponhamos que numa dada população, apenas uma proporção $p = 0{,}2$ dos indivíduos que a constituem possui determinada caraterística. Os histogramas seguintes relativos aos valores $n = 5, 10, 20$ e 40, descrevem a distribuição de frequências de $\widehat{p}$ obtida a partir de 2000 amostras de dimensão n recolhidas da população referida.



n=5

n=10

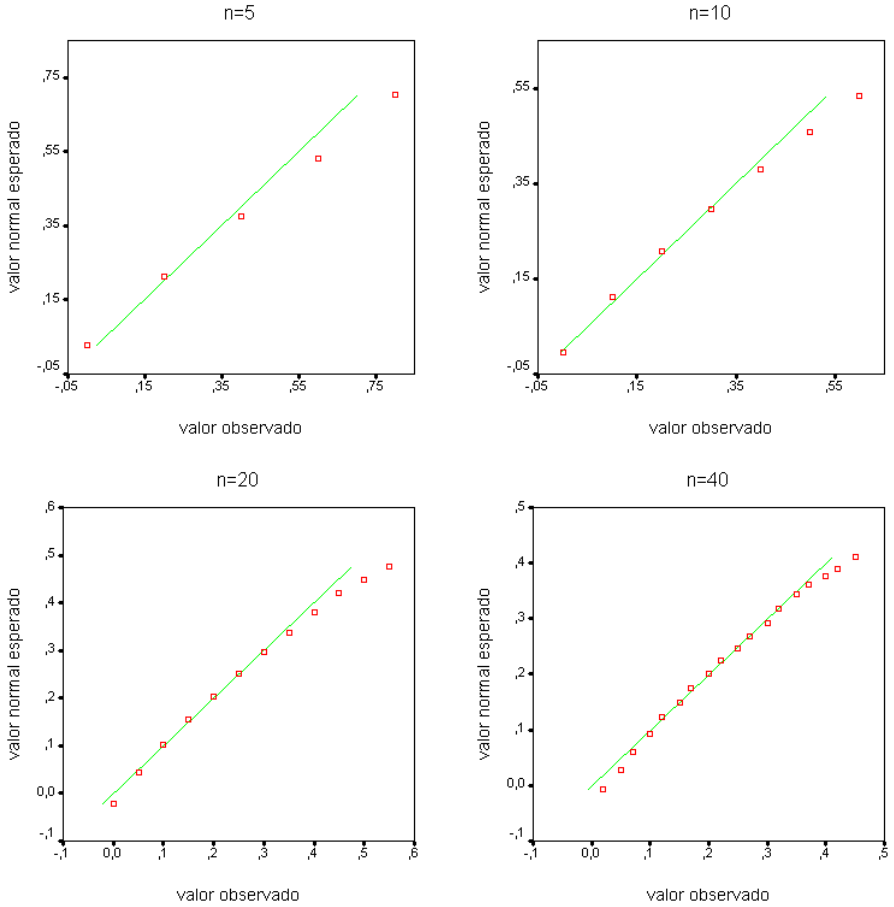


n=20

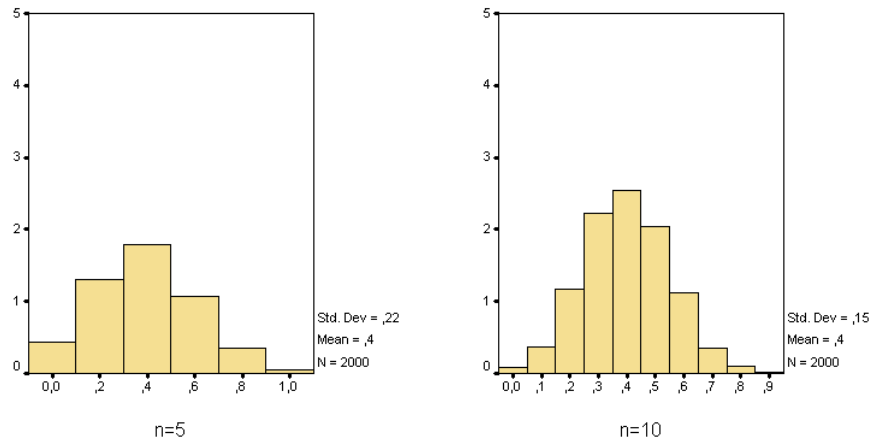
n=40

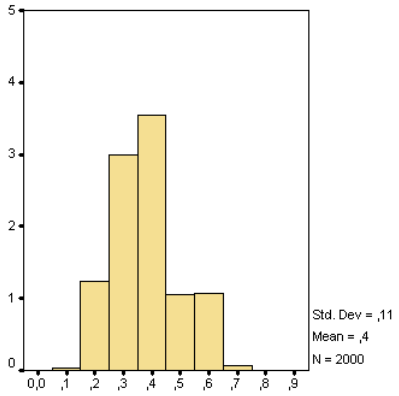
Para pequenos valores de n a distribuição $\widehat{p}$ revela uma assimetria positiva, que já tinhamos identificado na distribuição binomial para valores pequenos de p. Para valores grandes de n, a distribuição de frequências de $\widehat{p}$ torna-se cada vez menos assimétrica,

sendo a sua normalidade aproximada confirmada pelos gráficos de quantis normais seguintes.

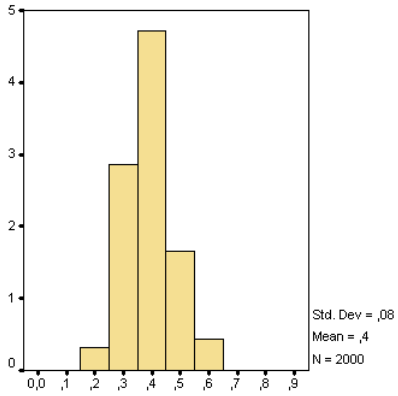


Exemplo 6.2.2 Suponhamos agora que na população em estudo uma proporção $p = 0{,}4$ dos seus membros possui determinada caraterística.



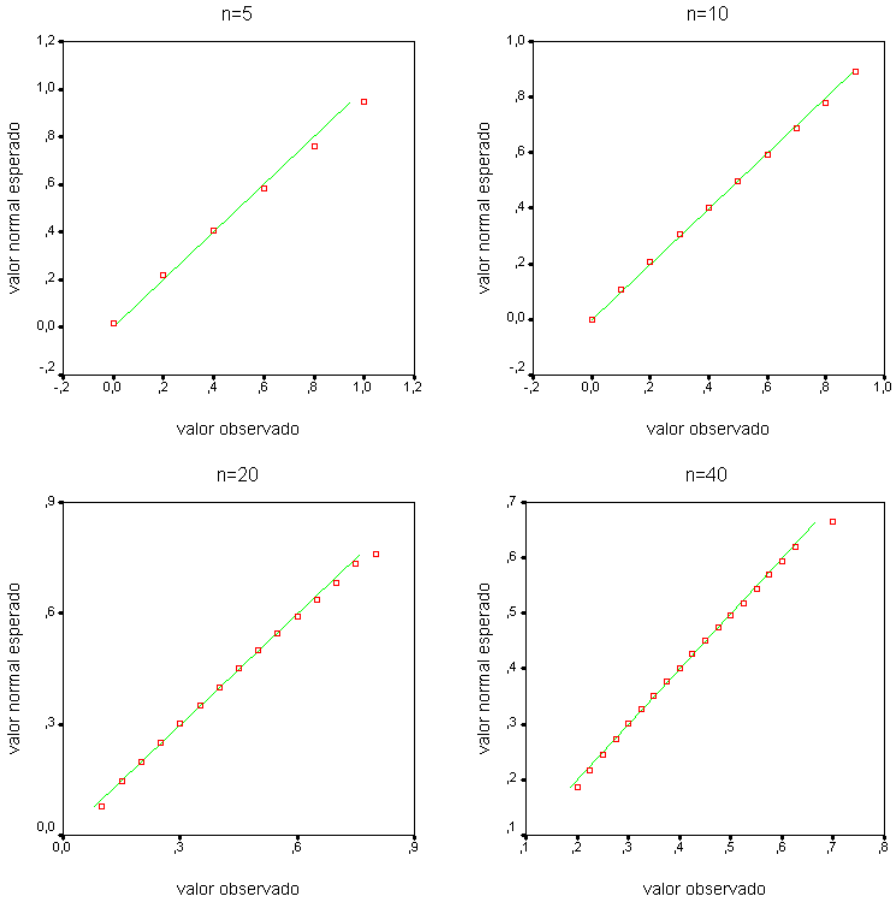


n=20



n=40

Tal como no exemplo anterior, os histogramas anteriores relativos aos valores $n = 5, 10, 20$ e 40, descrevem a distribuição de frequências de $\widehat{p}$ obtida a partir de 2000 amostras de dimensão n recolhidas da população referida.



Como a proporção p é próxima de 0,5, caso em que a distribuição binomial é simétrica, a distribuição de frequências de $\widehat{p}$ revela, para valores pequenos de n, uma

maior simetria que no exemplo anterior. A normalidade aproximada da distribuição de $\widehat{p}$ para valores pequenos e grandes de n é confirmada pelos gráficos de quantis normais seguintes.

Como já esperávamos, nos dois exemplos anteriores o centro das diversas distribuições de frequências de $\widehat{p}$ é aproximadamente igual a p, e a variabilidade respetiva decresce à medida que n aumenta. A normalidade da distribuição amostral de $\widehat{p}$ é mais evidente no caso $p = 0{,}4$ do que no caso $p = 0{,}2$, o que pode ser atribuído à maior assimetria da distribuição binomial $B(n; 0{,}2)$ quando comparada com $B(n; 0{,}4)$. Por outro lado, o aumento de n conduz a uma melhor aproximação da distribuição amostral de $\widehat{p}$ pela distribuição normal.

6.3 Distribuição amostral de $\bar{x}$

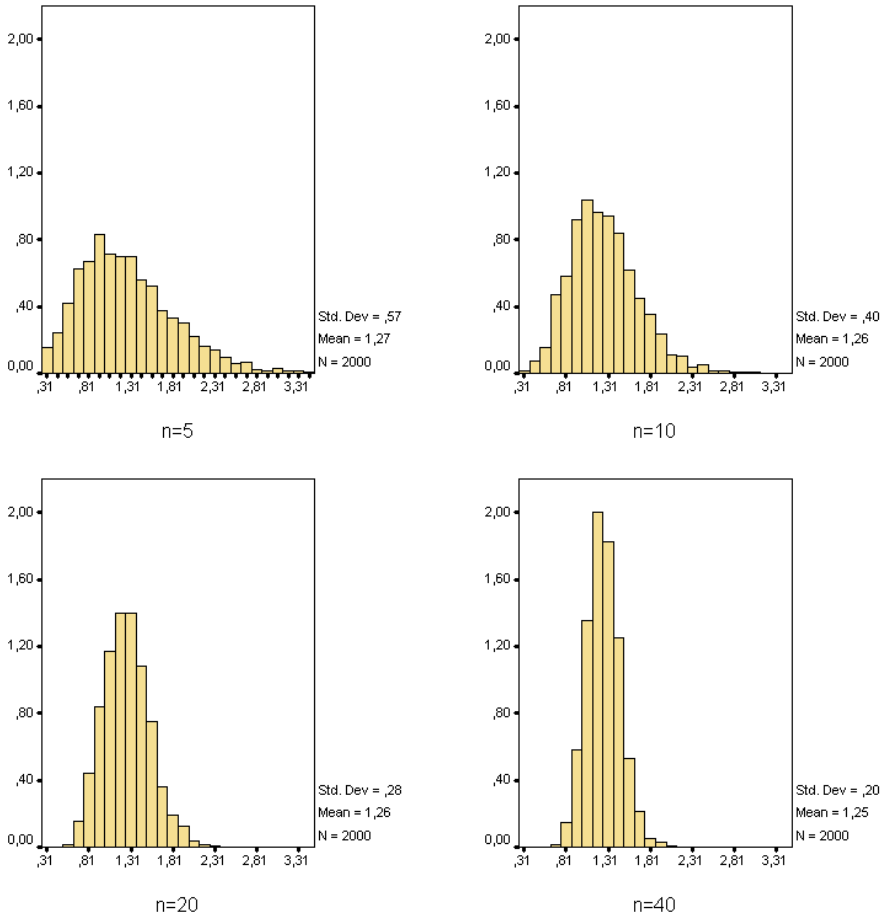
No parágrafo anterior, verificámos que a distribuição amostral da proporção $\widehat{p}$ associada a uma experiência binomial, pode, sob certas condições, ser aproximada por uma distribuição normal. Como já referimos na parte final do §4.5, a proporção amostral $\widehat{p}$ é um caso particular duma média amostral associada à variável aleatória que a cada sucesso numa experiência binomial associa 1 e a cada insucesso associa 0. Com efeito, como as observações $x_1, x_2, \ldots, x_n$ são ou iguais a 1 ou a 0, a proporção de sucessos é precisamente a média dessas observações: $\widehat{p} = \bar{x}$. Neste parágrafo, verificaremos que a aproximação normal de que goza a proporção amostral $\widehat{p}$ não é exclusiva desta estatística. Trata-se duma propriedade geral que é partilhada pela média amostral associada a observações independentes duma qualquer variável aleatória.

6.3.1 Distribuição de frequência de $\bar{x}$: dois exemplos

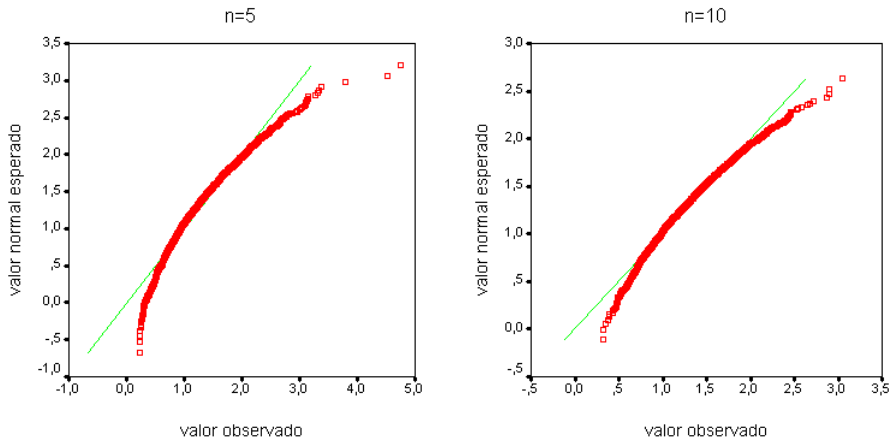
Tal como fizemos atrás, para analisar a distribuição de frequências da média amostral, $\bar{x}$, de duas populações com caraterísticas distintas vamos extrair várias amostras com dimensões iguais, calculando para cada uma delas a média respetiva. Estes vários valores são observações da média amostral $\bar{x}$ que, utilizando os métodos gráficos estudados no Capítulo 1, nos permitem analisar a sua distribuição de frequências, para cada uma das populações consideradas. Uma tal distribuição de frequências dar-nos-á indicações importantes sobre a distribuição de probabilidade da média amostral.

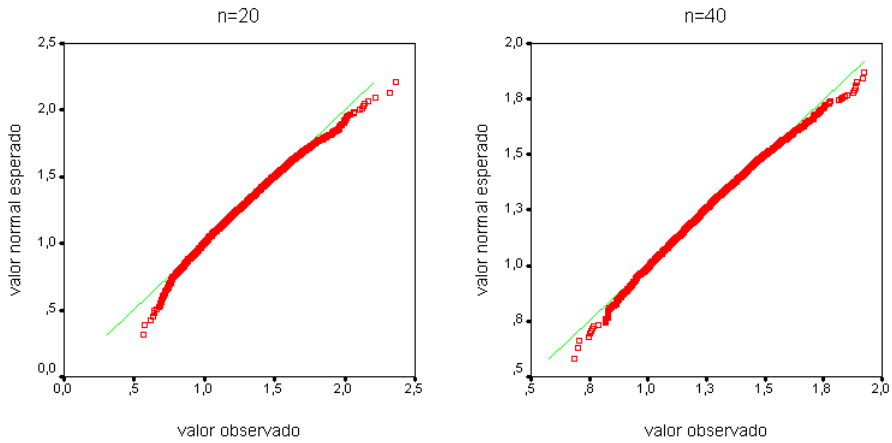
Exemplo 6.3.1 Comecemos por estudar a distribuição de frequências da média amostral $\bar{x}$ relativa à variável aleatória X que dá o tempo que medeia a chegada de dois clientes consecutivos a uma caixa de supermercado (ver Exemplo 4.2.5, pág. 104). Os histogramas que apresentamos relativos aos valores $n = 5, 10, 20$ e 40, descrevem

a distribuição de frequências de $\bar{x}$ obtida a partir de 2000 amostras de dimensão n recolhidas dum conjunto vasto de observações da variável X.



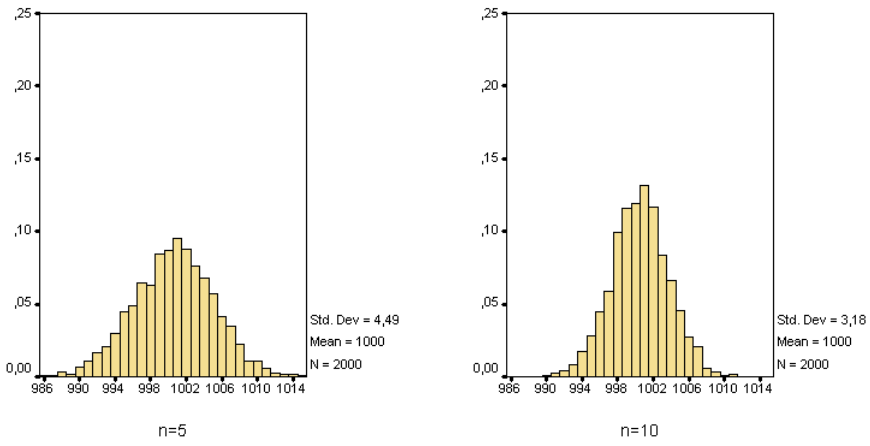
Dos gráficos anteriores constatamos que o centro das diversas distribuições amostrais é aproximadamente o valor 1,2, que podemos interpretar como sendo o tempo médio de

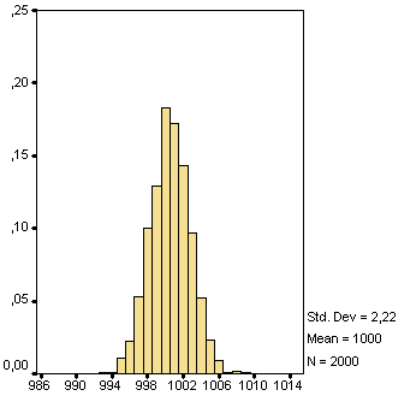




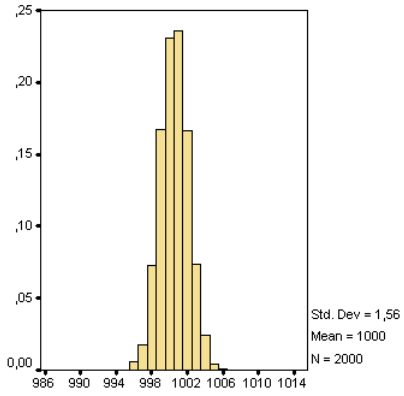
interchegada, e que a variabilidade de $\bar{x}$ diminui com a aumento da dimensão n das Reparemos que quando n é pequeno a distribuição de frequências de $\bar{x}$ revela uma assimetria positiva que é consequência da assimetria positiva marcada da distribuição de probabilidade da variável X (ver Exemplo 4.2.5, pág. 104). Para valores moderados e grandes de n a distribuição de frequência de $\bar{x}$ é aproximadamente simétrica, revelando o histograma de frequências uma forma de sino, própria das distribuições normais. A normalidade aproximada da distribuição de $\bar{x}$ para valores grandes de n é confirmada pelos gráficos de quantis normais respetivos.

Exemplo 6.3.2 Vejamos agora o que se passa com a distribuição de frequências da média amostral da variável aleatória Y relativa ao peso, em gramas, de pacotes de açúcar empacotados por uma máquina (ver Exemplo 4.2.6, pg. 105). Seguindo o procedimento do exemplo anterior, os histogramas seguintes descrevem a distribuição de frequências da média amostral $\bar{y}$ obtida a partir de 2000 amostras de dimensões $n = 5, 10, 20$ e 40, recolhidas dum conjunto vasto de observações da variável Y.



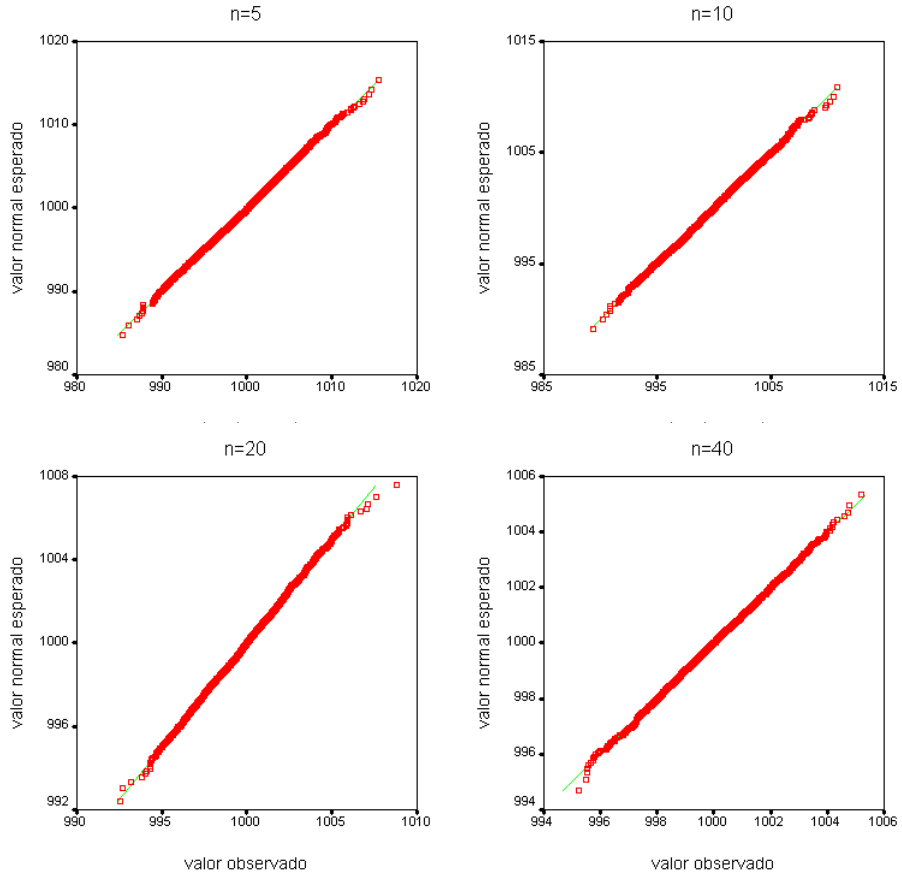


n=20



n=40

Tal como no exemplo anterior, para todos os valores considerados de n, o centro da distribuição de $\bar{y}$ é aproximadamente 1000, que é aproximadamente a média da variável Y, e a sua variabilidade diminui com o aumento de n. Dos gráficos anteriores e dos gráficos de quantis normais seguintes constatamos que, mesmo para pequenos valores de n, a distribuição amostral de $\bar{y}$ é aproximadamente normal. Como veremos, tal acontece pelo facto da variável Y ser ela própria aproximadamente normal.



Em jeito de conclusão, podemos referir três caraterísticas comuns às duas situações anteriores: 1) o centro da distribuição da média amostral parece ser independente de n e é aproximadamente igual à média da variável observada; 2) a variabilidade da distribuição da média amostral diminui com a aumento da dimensão da amostra; e, finalmente, 3) para valores moderados e grandes de n, a distribuição da média amostral é aproximadamente normal. Como caraterística divergente, podemos referir as distribuições das duas médias amostrais para pequenos valores de n.

6.3.2 Média e desvio-padrão de $\bar{x}$

As duas caraterísticas comuns que observámos, nos dois exemplos considerados, sobre o centro e a variabilidade da distribuição de frequências da média amostral, não são especificidades das variáveis aí consideradas. São caraterísticas gerais da média amostral dum conjunto de observações independentes de uma qualquer variável aleatória.

Para justificar esta afirmação, vamos calcular a média e a variância da média amostral

$$\bar{x} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n),$$

onde $X_1, X_2, \ldots, X_n$ representam as várias observações independentes da variável aleatória X com média μ e variância σ^2. Pelas propriedades da média, sabemos que

$$\begin{aligned} \mu_{\bar{x}} &= \frac{1}{n}(\mu_{X_1} + \mu_{X_2} + \ldots + \mu_{X_n}) \\ &= \frac{1}{n}(\mu + \mu + \ldots + \mu) \\ &= \frac{n\,\mu}{n} \\ &= \mu. \end{aligned}$$

Por outro lado, usando a independência entre as várias observações, podemos escrever

$$\begin{aligned} \sigma^2_{\bar{x}} &= \left(\frac{1}{n}\right)^2 (\sigma^2_{X_1} + \sigma^2_{X_n} + \ldots + \sigma^2_{X_n}) \\ &= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \ldots + \sigma^2) \\ &= \frac{n\,\sigma^2}{n^2} \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

Média e desvio-padrão de $\bar{x}$:

Se $X_1, X_2, \ldots, X_n$ são observações independentes da variável aleatória X com média μ e desvio-padrão σ, então

$$\mu_{\bar{x}} = \mu,$$

$$\sigma_{\bar{x}} = \sigma/\sqrt{n}.$$

Constatamos assim que a média de $\bar{x}$ não é mais do que a média da variável aleatória observada, e que o desvio-padrão de $\bar{x}$ é igual a $\sigma/\sqrt{n}$, isto é, decresce proporcionalmente à raiz quadrada do tamanho da amostra. Estas propriedades da média amostral explicam as caraterísticas observadas no parágrafo anterior.

6.3.3 O teorema do limite central

Outra caraterística interessante que constatámos sobre a distribuição da média amostral tem a ver com a sua normalidade, que observámos ocorrer, no caso da variável Y para todos os valores de n, e no caso da variável X para valores moderados e grandes de n.

Quando a dimensão da amostra for grande, há um teorema matemático, conhecido como **teorema central do limite** ou **teorema do limite central**, que assegura que, nesse caso, a distribuição da média amostral é aproximadamente normal. A palavra "central" deve-se à importância que este resultado teve na investigação matemática em Probabilidades, nas primeiras décadas do século passado.

Teorema do limite central:

Se $\bar{x}$ é calculada a partir de n observações independentes com média μ e desvio-padrão σ, então

$$\bar{x} \simeq N\left(\mu, \sigma/\sqrt{n}\right)$$

para n grande.

Reparemos que a aproximação normal obtida anteriormente para a distribuição amostral da proporção $\widehat{p}$, é um caso particular do teorema do limite central. Com efeito, usando (5.3.3), $\widehat{p}$ é a média das variáveis $S_1, S_2, \ldots, S_n$,

$$\widehat{p} = \frac{1}{n}(S_1 + S_2 + \ldots + S_n),$$

que como vimos têm média $\mu = p$ e desvio-padrão $\sigma = \sqrt{p(1-p)}$. Pelo teorema do limite central concluímos que

$$\widehat{p} \simeq N\left(p, \sqrt{p(1-p)}/\sqrt{n}\right),$$

ou seja,

$$\widehat{p} \simeq N\left(p, \sqrt{p(1-p)/n}\right),$$

que foi precisamente a aproximação normal dada anteriormente para a distribuição amostral de $\widehat{p}$.

O comportamento da distribuição da média amostral descrito pelo teorema do limite central ocorre também em situações mais gerais do que aquelas que enunciámos. Por exemplo, a aproximação normal para a média amostral é ainda válida em casos em que há dependência entre as diversas observações, ou em casos em que as várias observações não podem ser consideradas realizações de variáveis aleatórias com a mesma distribuição. Em particular, se a amostra é recolhida por amostragem aleatória simples duma população finita, o teorema do limite central é ainda válido.

A qualidade da aproximação da distribuição da média amostral pela distribuição normal depende muito da forma da distribuição de probabilidade subjacente à variável observada. Se uma tal distribuição for próxima da distribuição normal, será de esperar que a aproximação normal para a distribuição da média amostral ocorra para valores de n mais pequenos do que no caso em que a distribuição da variável observada for muito diferente da distribuição normal. Quando a distribuição das observações é exatamente normal a distribuição da média amostral é exatamente normal para qualquer dimensão da amostra. Isto explica os resultados observados no Exemplo 6.3.2.

Distribuição de $\bar{x}$ para observações normais e independentes:

Se $\bar{x}$ é calculada a partir de n observações normais e independentes com média μ e desvio-padrão σ, então

$$\bar{x} \sim N\left(\mu, \sigma/\sqrt{n}\right)$$

para todos os valores de n.

Exemplo 6.3.3 Vimos no Exemplo 5.2.2 como podemos controlar a qualidade dum processo de fabrico através da construção duma carta de controlo. No exemplo que focámos sobre o controlo do peso de pacotes de açucar empacotados por uma máquina, que em condições ideais de funcionamento produz pacotes cuja distribuição dos pesos

possui uma distribuição normal com média 1000 gramas e com desvio-padrão 10 gramas, cada um dos pontos marcado na carta de controlo resultava duma única observação o que introduz no processo de controlo uma variabilidade indesejada. Mais natural é que cada ponto marcado resulte da observação de mais do que um pacote. Admitamos assim que para controlar o processo de empacotamento, de hora a hora é recolhida uma amostra de 5 pacotes, que acabaram de sair da máquina, e é registado o seu peso médio. Como esta média é uma média de observações normais que vamos admitir independentes, o resultados anterior permite concluir que

$$\bar{x} \sim N(1000, 10/\sqrt{5}).$$

Em particular, e atendendo à regra 68-95-99,7, podemos dizer que 99,7% dos pesos médios assim registados pertence ao intervalo $[1000 - 3 \times 10/\sqrt{5}, 1000 - 3 \times 10/\sqrt{5}] = [986{,}6; 1013{,}4]$. Se alguma das médias registadas não pertence a este intervalo, isso pode ser uma indicação de que a máquina está a funcionar mal, necessitando por isso de ser calibrada.

Vejamos dois exemplos simples de utilização do teorema do limite central, no cálculo de probabilidades associadas a uma variável aleatória que se exprime como soma de variáveis aleatórias independentes.

Exemplo 6.3.4 Suponhamos que decidimos lançar um dado equilibrado 100 vezes consecutivas, e que apostamos com um amigo A que vamos obter pelo menos 350 pontos na soma dos pontos obtidos nos vários lançamentos, e com outro amigo B que vamos obter mais do que 400 pontos. Qual é a probabilidade de ganharmos a aposta com cada um dos amigos? Se representarmos por $X_1, X_2, \ldots, X_{100}$ os pontos obtidos em cada um dos 100 lançamentos e por S a sua soma, isto é, $S = X_1 + X_2 + \ldots + X_{100}$, as probabilidades pedidas são dadas por $\mathrm{P}(S \geq 350)$ e $\mathrm{P}(S > 400)$, respetivamente. Como vimos no Exemplo 4.3.1, cada uma das variáveis X_i tem média 3,5 e desvio-padrão $\sqrt{2{,}9167}$. Atendendo ao teorema do limite central, a média amostral

$$\bar{x} = (X_1 + X_2 + \ldots + X_{100})/100 = S/100,$$

é aproximadamente normal com média 3,5 e desvio-padrão $\sqrt{2{,}9167}/\sqrt{100} \approx 0{,}1708$. Para obter resultados mais fidedignos, vamos usar a correção de continuidade no cálculo das duas probabilidades anteriores. Assim, denotando por Z a variável normal standard, temos

$$\begin{aligned} \mathrm{P}(S \geq 350) &= \mathrm{P}(S \geq 349{,}5) \\ &= \mathrm{P}(\bar{x} \geq 3{,}495) \end{aligned}$$

$$\begin{aligned} &= \mathrm{P}\left(\frac{\bar{x}-3{,}5}{0{,}1708} \geq \frac{3{,}495-3{,}5}{0{,}1708}\right) \\ &\approx \mathrm{P}(Z \geq -0{,}029) \\ &= 1-0{,}4884 = 0{,}5116 \end{aligned}$$

e

$$\begin{aligned} \mathrm{P}(S>400) &= \mathrm{P}(S>400{,}5) \\ &= \mathrm{P}(\bar{x}>4{,}005) \\ &= \mathrm{P}\left(\frac{\bar{x}-3{,}5}{0{,}1708} > \frac{4{,}005-3{,}5}{0{,}1708}\right) \\ &\approx \mathrm{P}(Z>2{,}957) \\ &= 1-0{,}9984 = 0{,}0016. \end{aligned}$$

Exemplo 6.3.5 Suponhamos que no jogo da roleta descrito no Exemplo 4.5.2 (pág. 118), o jogador decide jogar 100 partidas numa das suas idas ao casino. Calculemos uma aproximação para a probabilidade dele ganhar mais do que aquilo que perde. Representando por X_i o ganho (ou perda) líquido do jogador na i-ésima partida, o ganho líquido do jogador no fim das 100 partidas é dado por $G = X_1 + X_2 + \ldots + X_{100}$. Estas variáveis já foram por nós estudadas no Exemplo 4.5.2, onde vimos que possuiam média $-0{,}27$ euros e desvio-padrão $\sqrt{3408{,}035} \approx 58{,}3784$ euros. Usando o teorema do limite central, sabemos que a média amostral $\bar{x} = G/100$, pode ser aproximada pela distribuição normal de média $-0{,}27$ e desvio-padrão $58{,}3784/\sqrt{100} = 5{,}83784$. Assim, denotando por Z a variável normal standard, temos (para efetuar a correção de continuidade, devemos ter em conta que G toma valores de 10 em 10)

$$\begin{aligned} \mathrm{P}(G>0) &= \mathrm{P}(G>5) \\ &= \mathrm{P}(\bar{x}>0{,}05) \\ &= \mathrm{P}\left(\frac{\bar{x}-(-0{,}27)}{5{,}83784} > \frac{0{,}05-(-0{,}27)}{5{,}83784}\right) \\ &\approx \mathrm{P}(Z>0{,}055) \\ &= 1-0{,}5219 = 0{,}4781. \end{aligned}$$

Vejamos agora o que acontece à probabilidade anterior, se o jogador decide jogar 1000 partidas em vez de 100. Neste caso, $G = X_1 + X_2 + \ldots + X_{1000}$ e a média amostral, $\bar{x} = G/1000$, pode ser aproximada pela distribuição normal de média $-0{,}27$ e desvio-padrão $58{,}3784/\sqrt{1000} \approx 1{,}8461$, e portanto

$$\begin{aligned} \mathrm{P}(G>0) &= \mathrm{P}(G>5) \\ &= \mathrm{P}(\bar{x}>0{,}005) \end{aligned}$$

$$\begin{aligned} &= \mathrm{P}\left(\frac{\bar{x}-(-0{,}27)}{1{,}8461} > \frac{0{,}005-(-0{,}27)}{1{,}8461}\right) \\ &\approx \mathrm{P}(Z > 0{,}149) \\ &= 1 - 0{,}5592 = 0{,}4408. \end{aligned}$$

Vemos assim, que quantas mais partidas o jogador joga, mais probabilidade tem de sair do casino com menos dinheiro do que quando entrou. Esta conclusão está de acordo com as conclusões a que chegámos através da lei dos grandes números.

6.4 Bibliografia

Daniel, W.W. (2009). *Biostatistics: a foundation for analysis in the health sciences*, Wiley.

Dawson, B., Trapp, R.G. (2004). *Basic & Clinical Biostatistics*, McGraw-Hill.

Moore, D.S. (1985). *Statistics: concepts and controversies*, W.H. Freeman and Company.

McPherson. G. (1990). *Statistics in Scientific Investigation: its basis, application and interpretation*, Springer-Verlag.

Moore, D.S. (1985). *Statistics: concepts and controversies*, W.H. Freeman and Company.

Moore, D.S., McCabe, G.P. (2006). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

7

Intervalos de confiança para proporções e médias

Inferência estatística. Noção de intervalo de confiança. Margem de erro e nível de confiança. Intervalos de confiança para proporções. Intervalos de confiança para médias. O caso das populações normais. A distribuição de Student. Como escolher o tamanho da amostra.

7.1 Inferência estatística

Tão ou mais interessantes do que as aplicações do teorema do limite central com que terminámos o capítulo anterior, são as suas aplicações à inferência estatística que vamos abordar em detalhe neste e no próximo capítulos. O conhecimento das distribuições amostrais das estatísticas $\bar{x}$ e $\widehat{p}$, ou da respetiva aproximação normal, é de importância fundamental na implementação de dois procedimentos de inferência estatística, conhecidos como **intervalos de confiança** e **testes de hipóteses**, cujo objetivo comum é inferir sobre um parâmetro desconhecido da população que estudamos, e que no caso particular das estatísticas $\bar{x}$ e $\widehat{p}$, ou é uma média, μ, ou uma proporção, p, respetivamente.

Exemplo 7.1.1 Para ilustrar o que acabámos de dizer, recordemos o Exemplo 3.3.1 (pág. 75) em que uma moeda portuguesa de um euro foi lançada 50 vezes tendo-se obtido 45 vezes a face europeia e 5 vezes a face portuguesa. A questão que colocámos na altura era a de saber qual era a probabilidade de sair a face europeia. Vimos que a resposta a esta questão poderia depender do nosso conhecimento sobre a experiência em causa, em particular sobre o facto de termos, ou não, razões para admitir que a moeda é equilibrada. Representando por p a probabilidade de ocorrência da face europeia no lançamento desta moeda, sabemos já que estamos na presença duma experiência aleatória binomial de parâmetros $n = 50$ e p, onde p é um parâmetro desconhecido

sobre o qual pretendemos inferir. Atendendo à lei dos grandes números sabemos que a proporção de faces europeias observadas, $\widehat{p} = 45/50 = 0{,}9$, é uma aproximação da probabilidade p de ocorrência da face europeia no lançamento desta moeda.

Se além da estimativa 0,9 (dita estimativa pontual), pretendemos dar indicação sobre a **precisão** da mesma, que será naturalmente dada sob a forma dum intervalo cuja amplitude indicará a precisão da estimativa, estamos caídos num problema de estimação por intervalos de confiança.

Em vez de pretendermos uma aproximação para p, poderemos querer saber se a moeda é, ou não, equilibrada. Por outras palavras, poderemos querer saber se a proporção observada, 0,9, é, ou não, compatível com a hipótese $p = 0{,}5$ da moeda ser equilibrada. Temos neste caso um problema de testes de hipóteses.

Podemos assim dizer, que no caso dos intervalos de confiança, pretende-se estimar o parâmetro de interesse dando indicação da precisão da estimativa apresentada, enquanto que no caso dos testes de hipóteses pretende-se avaliar a adequação das observações realizadas com uma hipótese formulada, *a priori*, sobre o parâmetro de interesse. Em ambos os casos, e é essa caraterística que distingue a estatística inferencial da estatística descritiva, pretende-se quantificar a **confiança** que temos nas conclusões que apresentamos, ou de forma equivalente, quantificar o erro que podemos estar a cometer quando "tomamos a parte (amostra) para inferir sobre o todo (população)". Como veremos a seguir, o conhecimento da distribuição amostral da estatística de interesse, seja ela a média amostral $\bar{x}$ ou a proporção amostral $\widehat{p}$, é essencial para atingirmos estes objetivos.

7.2 Estimação por intervalos de confiança

A estimação por intervalos de confiança é uma técnica do âmbito da estatística inferencial cujo objetivo é o da estimação dum parâmetro (desconhecido) duma população que estudamos. A particularidade desta técnica que a torna diferente da denominada **estimação pontual**, é que para além da estimativa para o parâmetro que se obtém a partir das observações realizadas, própria da estimação pontual, são também indicadas a **precisão** e a **confiança** que temos na estimativa produzida. A precisão da estimativa é definida pela chamada **margem de erro**, que conjuntamente com a estimativa calculada definem um intervalo do tipo

$$\text{estimativa pontual} \pm \text{margem de erro}$$

dito **intervalo de confiança** para o parâmetro de interesse. A **confiança** que temos na estimativa produzida, será avaliada em termos da probabilidade dos intervalos assim

construídos, que são diferentes de amostra para amostra, conterem o verdadeiro valor do parâmetro. Um exemplo, bem nosso conhecido, em que esta técnica estatística é usada, é o das sondagens eleitorais a que fizemos já referência no capítulo introdutório e ao qual voltaremos mais à frente.

Exemplo 7.2.1 Para ilustrar a construção dum intervalo de confiança, retomemos o Exemplo 7.1.1 (pág. 165) do lançamento duma moeda de um euro em que observámos a face europeia em 45 dos 50 lançamentos que efetuámos, e em que pretendemos estimar a probabilidade p de ocorrência da face europeia num lançamento da moeda. Para esta amostra, a proporção de faces europeias ocorridas foi de $\widehat{p} = 0{,}9$. Se repetíssemos a experiência aleatória efetuando mais e mais sucessões de 50 lançamentos da mesma moeda, sabemos que a proporção $\widehat{p}$ possui uma distribuição de probabilidade aproximadamente normal com média

$$\mu_{\widehat{p}} = p,$$

e com desvio-padrão

$$\sigma_{\widehat{p}} = \sqrt{p(1-p)/50} \approx 0{,}14\sqrt{p(1-p)}.$$

Dito de outra maneira, a variável aleatória

$$\frac{\widehat{p} - p}{0{,}14\sqrt{p(1-p)}}$$

é aproximadamente normal com média 0 e desvio-padrão 1. Utilizando a regra 68-95-99,7, sabemos que a probabilidade da variável anterior pertencer ao intervalo $[-2, 2]$ é aproximadamente igual 0,95. Atendendo à interpretação frequencista da noção de probabilidade, isto quer dizer que se repetirmos a experiência aleatória efetuando mais e mais sucessões de 50 lançamentos da moeda, em 95% dessas repetições ter-se-á

$$-2 \leq \frac{\widehat{p} - p}{0{,}14\sqrt{p(1-p)}} \leq 2,$$

ou seja, em 95% dessas repetições $\widehat{p}$ pertencerá ao intervalo

$$\left[\, p - 0{,}28\sqrt{p(1-p)}\,, p + 0{,}28\sqrt{p(1-p)}\,\right].$$

Dizer que a proporção $\widehat{p}$ pertence ao intervalo anterior em 95% das repetições da experiência, é a mesma coisa que dizer que o intervalo

$$\left[\, \widehat{p} - 0{,}28\sqrt{p(1-p)}\,, \widehat{p} + 0{,}28\sqrt{p(1-p)}\,\right],$$

contém a verdadeira probabilidade p em 95% das repetições da experiência.

Este intervalo é ainda de pouca utilidade pois não pode ser calculado exclusivamente a partir das observações realizadas. Ele depende do parâmetro p cujo verdadeiro valor desconhecemos. No entanto, pela lei dos grande números, sabemos que, para valores grandes de n, $\widehat{p}$ está próximo de p, o que nos permite afirmar que o intervalo

$$\left[\,\widehat{p}-0{,}28\sqrt{\widehat{p}(1-\widehat{p})}\,,\widehat{p}+0{,}28\sqrt{\widehat{p}(1-\widehat{p})}\,\right],$$

contém p em aproximadamente 95% das vezes que repetirmos a experiência.

O intervalo anterior diz-se **intervalo de confiança** para p com um **nível de confiança** de aproximadamente 0,95. O nível de confiança é também designado por **grau de confiança**, ou ainda, pelas razões anteriores, por **probabilidade de cobertura** do intervalo de confiança. É também frequente usar a percentagem para exprimir o nível de confiança do intervalo. Neste caso diremos que o intervalo anterior é um intervalo de confiança para p com um **nível de confiança** de 95%.

Atendendo a que para os lançamentos realizados observámos $\widehat{p}=0{,}9$, dizemos também que o intervalo

$$\left[\,0{,}9-0{,}28\sqrt{0{,}9(1-0{,}9)}\,;0{,}9+0{,}28\sqrt{0{,}9(1-0{,}9)}\,\right]=[\,0{,}816;0{,}984\,]$$

é um intervalo de confiança para p com um nível de confiança de 0,95. Apesar do elevado grau de confiança, notemos que nada nos garante que a amostra observada não seja uma daquelas 5% em que os intervalos a partir delas obtidos não contêm o verdadeiro valor de p.

Reparemos que, de forma análoga, podemos utilizar a regra 68-95-99,7 para construir intervalos de confiança com níveis de confiança de 68% e de 99,7%. Atendendo às observações realizadas, concluímos que

$$\left[\,0{,}9-0{,}14\sqrt{0{,}9(1-0{,}9)}\,;0{,}9+0{,}14\sqrt{0{,}9(1-0{,}9)}\,\right]=[\,0{,}858;0{,}942\,]$$

é um intervalo de confiança para p com um nível de confiança de 68%, enquanto que

$$\left[\,0{,}9-0{,}42\sqrt{0{,}9(1-0{,}9)}\,;0{,}9+0{,}42\sqrt{0{,}9(1-0{,}9)}\,\right]=[\,0{,}774;1{,}026\,]$$

é um intervalo de confiança para p com um nível de confiança de 99,7%. Como podemos constatar, o aumento do nível de confiança tem como contrapartida o aumento da margem de erro, ou seja, a diminuição da precisão da estimativa. Como seria fácil de constatar, para aumentarmos a precisão da estimativa para um nível de confiança fixado à partida, seria necessário aumentar o tamanho da amostra.

Como podemos concluir deste exemplo, a quantificação da confiança na estimativa apresentada tem a ver, não com o intervalo de confiança que calculámos a partir das

observações, pois este, ou contém, ou não contém o verdadeiro valor de p, mas sim com o que se passaria se o processo fosse repetido um grande número de vezes. Por outras palavras, a quantificação da confiança tem a ver com o método utilizado para construir o intervalo de confiança.

Os intervalos de confiança para proporções e médias que estudamos neste capítulo, serão apresentados admitindo que as observações são realizações independentes de determinada variável aleatória. Como já referimos a propósito das distribuições amostrais, estes intervalos são ainda válidos sob condições mais gerais. Tal acontece, em particular, quando a amostra é recolhida por amostragem aleatória simples. Tal já não acontece se usarmos outro dos métodos aleatórios de recolha de amostras a que fizemos referência no Capítulo 2.

7.3 Intervalos de confiança para uma proporção

Analisemos agora o caso geral duma qualquer experiência aleatória binomial de parâmetros n e p, onde n representa o número de observações realizadas, e em que pretendemos obter um intervalo de confiança para o parâmetro desconhecido p, com um nível de confiança C, fixado à partida. Como queremos intervalos com um nível de confiança elevado, C é habitualmente um número inferior mas próximo de 1.

Seguindo o método descrito no parágrafo anterior, podemos, sem dificuldades de maior, obter um método geral que permita, a partir da distribuição amostral de $\widehat{p} = X/n$, onde X é o número de sucessos observados, que sabemos ser aproximadamente normal com média

$$\mu_{\widehat{p}} = p,$$

e com desvio-padrão

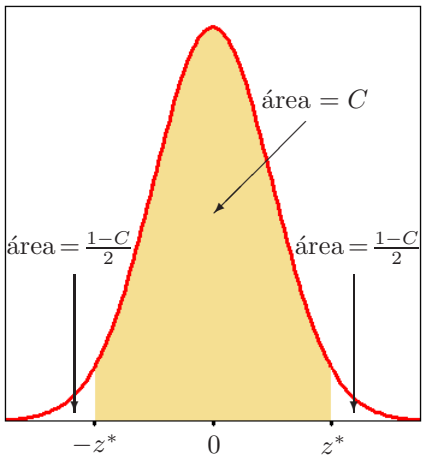
$$\sigma_{\widehat{p}} = \sqrt{p(1-p)/n},$$

construir intervalos centrados em $\widehat{p}$, que, com probabilidade aproximadamente igual a C, contenham o verdadeiro valor de p.

Se pretendemos um intervalo com nível de confiança C, devemos começar por consultar a tabela da distribuição normal para determinar o intervalo da forma $[\,-z^*, z^*\,]$, ao qual pertence uma variável normal standard com uma probabilidade C (ver a figura seguinte). Isto é, começamos por determinar o valor z^* para o qual se tem

$$\mathrm{P}(-z^* \leq Z \leq z^*) = C$$

onde Z tem uma distribuição normal $N(0,1)$.



Nos casos habituais escolhemos para C um dos valores 0,9, 0,95 ou 0,99. Para cada um destes valores de C, obtemos para z^* os valores dados na tabela seguinte:

C	0,90	0,95	0,99
z^*	1,645	1,960	2,576

Reparemos que existem vários intervalos não centrados na origem que têm a propriedade de terem entre as suas extremidades uma área igual a C. No entanto, pode ser demonstrado que são os intervalos centrados na origem que têm uma menor amplitude, conduzindo, por isso, a intervalos de confiança com uma menor margem de erro.

Determinado o valor de z^*, e atendendo a que a variável aleatória

$$\frac{\widehat{p}-p}{\sqrt{p(1-p)/n}} \tag{7.3.1}$$

é aproximadamente normal com média 0 e desvio-padrão 1, podemos dizer que, com probabilidade aproximadamente igual a C, vale a dupla desigualdade

$$-z^* \leq \frac{\widehat{p}-p}{\sqrt{p(1-p)/n}} \leq z^*, \tag{7.3.2}$$

ou ainda, que o intervalo

$$\left[\,\widehat{p}-z^*\sqrt{\widehat{p}(1-\widehat{p})/n}\,,\widehat{p}+z^*\sqrt{\widehat{p}(1-\widehat{p})/n}\,\right]$$

contém p com probabilidade aproximadamente igual a C.

Este intervalo é habitualmente designado como **intervalo de confiança de Wald** para p com nível de confiança C. À variável (7.3.1), que esteve na base da construção do intervalo de confiança, chamamos **variável fulcral**.

O nível de confiança do intervalo de Wald é, devido à aproximação normal para a distribuição amostral de $\widehat{p}$ que utilizámos na sua construção, apenas aproximadamente igual a C. O mesmo acontece com outros intervalos de confiança que sejam construídos a partir duma aproximação para a distribuição amostral de $\widehat{p}$.

Intervalo de confiança de Wald para uma proporção:

Numa experiência aleatória binomial de parâmetros n e p, um intervalo de confiança para p, com nível de confiança aproximadamente igual a C, tem por extremidades

$$\widehat{p} \pm z^* \sqrt{\widehat{p}(1-\widehat{p})/n},$$

onde z^* é tal que

$$\mathrm{P}(-z^* \leq Z \leq z^*) = C$$

e Z tem uma distribuição normal $N(0,1)$.

Quando a probabilidade de sucesso p é próxima de 0 ou de 1, e n é pequeno, o nível de confiança efetivo do intervalo de Wald pode ser muito diferente de C, em consequência da aproximação normal para a distribuição amostral de $\widehat{p}$ não ser de grande qualidade. Além disso, a probabilidade de obter $\widehat{p} = 0$ ou $\widehat{p} = 1$, é grande, levando, nesses casos, aos intervalos de confiança $[0, 0]$ ou $[1, 1]$, que são manifestamente desapropriados como intervalos de confiança para p. Neste sentido vamos usar o intervalo anterior apenas quando $n\widehat{p} \geq 15$ e $n(1-\widehat{p}) \geq 15$, ou seja, quando tivermos pelo menos 15 sucessos e 15 insucesso na amostra (Moore e McCabe, 2006, p. 537).

Mais grave é o facto de que mesmo para valores grandes de n e para p nem muito pequeno nem muito grande, o intervalo de Wald poder ter um **nível de confiança efetivo** muito diferente do que desejamos. Isto quer dizer, que se tirarmos várias amostras e calcularmos para cada uma o intervalo de confiança de nível 0,95 para p, a proporção de intervalos que contêm p pode ser muito diferente de 0,95. Dito de outra forma, a **probabilidade de cobertura** dos intervalos de confiança de Wald pode ser muito diferente de 0,95.

Como vamos ver a seguir, é possível construir intervalos de confiança para uma proporção que não sofram dos problemas que apontámos. Para tal, retomemos novamente a dupla desigualdade (7.3.2) e em vez de substituirmos p por $\widehat{p}$ no denominador da variável (7.3.1), o que deu origem ao intervalo de confiança de Wald, vamos desenvolver a dupla desigualdade. Depois de alguns cálculos, chegamos à conclusão que o

intervalo

$$\left[\widetilde{p} - \frac{z^*}{\widetilde{n}}\sqrt{n\widehat{p}(1-\widehat{p}) + \frac{(z^*)^2}{4}}, \widetilde{p} + \frac{z^*}{\widetilde{n}}\sqrt{n\widehat{p}(1-\widehat{p}) + \frac{(z^*)^2}{4}}\right],$$

onde $\widetilde{p} = \widetilde{X}/\widetilde{n}$, $\widetilde{X} = X + (z^*)^2/2$ e $\widetilde{n} = n + (z^*)^2$, contém p com probabilidade aproximadamente igual a C. Este intervalo é dito **intervalo de confiança de Wilson** para uma proporção.

Intervalo de confiança de Wilson para uma proporção:

Numa experiência aleatória binomial de parâmetros n e p, um intervalo de confiança para p, com nível de confiança aproximadamente igual a C, tem por extremidades

$$\widetilde{p} \pm \frac{z^*}{\widetilde{n}}\sqrt{n\widehat{p}(1-\widehat{p}) + \frac{(z^*)^2}{4}},$$

onde

$$\widetilde{p} = \widetilde{X}/\widetilde{n},$$

com $\widetilde{X} = X + (z^*)^2/2$, $\widetilde{n} = n + (z^*)^2$, e onde z^* é tal que

$$\mathrm{P}(-z^* \leq Z \leq z^*) = C$$

onde Z tem uma distribuição normal $N(0,1)$.

O intervalo de confiança de Wilson é preferível ao intervalo de Wald. A sua probabilidade de cobertura é próxima do nível de confiança fixado à partida, excepto quando a amostra é pequena e p é muito próximo de 0 ou de 1.

A principal desvantagem do intervalo de confiança de Wilson está na complexidade dos cálculos que envolve para quem não tem à mão uma folha de cálculo. Uma forma simples de rodear este inconveniente, obtendo ao mesmo tempo um intervalo de confiança com boas propriedades e que para valores grandes de n ($n \geq 40$) é muito próximo do intervalo de Wilson, é considerar um intervalo de confiança cuja forma é a do intervalo de Wald mas que seja baseado, não na proporção amostral $\widehat{p}$ mas num estimador corrigido que tenha uma forma semelhante à ponto médio $\widetilde{p}$ do intervalo de confiança de Wilson que podemos considerar como uma proporção amostral corrigida uma vez que $\widetilde{X} = X + (z^*)^2/2$ e $\widetilde{n} = n + (z^*)^2$, podem ser interpretadas como correções para o número de sucessos observados e para o número de observações realizadas, respetivamente. Isto é, o ponto médio do intervalo de Wilson sugere que proporção amostral deve ser corrigida juntando à amostra um conjunto de pseudo-observações sendo metade

delas sucessos. A correção que habitualmente se considera pelos seus bons resultados é a de juntar à amostra 4 pseudo-observações sendo 2 delas sucessos, o que dá origem ao intervalo de confiança seguinte, dito **intervalo de confiança de Agresti-Coull**. Assim, o intervalo de confiança de Agresti-Coull é obtido tomando para variável fulcral a variável

$$\frac{\widetilde{p}-p}{\sqrt{\widetilde{p}(1-\widetilde{p})/\widetilde{n}}},$$

onde $\widetilde{p}=\widetilde{X}/\widetilde{n}$, $\widetilde{X}=X+2$ e $\widetilde{n}=n+4$.

Apesar do intervalo de Wilson possuir melhores propriedades que o intervalo de Agresti-Coull para amostras de dimensão $n<40$, vamos, por simplicidade de cálculo, utilizar o intervalo de Agresti-Coull independentemente do tamanho da amostra. Tal como para o intervalo de Wilson, a utilização do intervalo de Agresti-Coull quando a amostra é pequena ($n<10$) e p é muito próximo de 0 ou de 1 pode ser problemática (Moore e McCabe, 2006, p. 539).

Intervalo de confiança de Agresti-Coull para uma proporção:

Numa experiência aleatória binomial de parâmetros n e p, um intervalo de confiança para p, com nível de confiança aproximadamente igual a C, tem por extremidades

$$\widetilde{p}\pm z^*\sqrt{\widetilde{p}(1-\widetilde{p})/\widetilde{n}},$$

onde

$$\widetilde{p}=\widetilde{X}/\widetilde{n},$$

com $\widetilde{X}=X+2$ e $\widetilde{n}=n+4$, onde z^* é tal que

$$\mathrm{P}(-z^*\leq Z\leq z^*)=C$$

e Z tem uma distribuição normal $N(0,1)$.

Exemplo 7.3.3 Para exemplificar a utilização das fórmula anteriores a um exemplo concreto, consideremos a sondagem eleitoral Expresso/Euroexpansão publicada pelo semanário *Expresso* em 16 de Setembro de 1995, relativa às eleições legislativas desse ano que seriam realizadas a 1 de Outubro de 1995. A título de curiosidade refira-se que esta sondagem foi feita após o primeiro frente-a-frente televisivo entre Fernando Nogueira e António Guterres, candidatos a primeiro ministro pelo PSD e PS, respetivamente.

Os resultados da tabela seguinte são os apresentados pelo *Expresso* e, de acordo com o explicado no jornal, resultam da inquirição de 1006 indivíduos depois de se distribuírem os resultados de 188 indecisos pelos diversos partidos de forma proporcional de acordo com as preferências demonstradas pelos não-indecisos (o que é equivalente a dizer que apenas se consideraram as respostas dos não-indecisos).

Partidos	Intenção de voto
CDU	8,8%
PS	41,8%
PSD	33,7%
CDS/PP	9,1%
Outros	6,6%

Ficha técnica:
Universo da sondagem – eleitorado de Portugal Continental;
Amostragem – de tipo aleatório, estratificada segundo a região e o "habitat";
Dimensão da amostra – 1006 indivíduos;
Margem de erro máxima – 3,1%, com um grau de probabilidade de 95%.

Admitindo que a amostragem realizada foi a amostragem aleatória simples, o que não aconteceu como podemos constatar pela informação dada na ficha técnica, determinamos a seguir intervalos de confiança de nível 0,95 para as intenções de voto em cada um dos partidos anteriores. Para tal usamos a tabela seguinte que tenta reproduzir os resultados não tratados dos inquéritos feitos caso a amostra fosse recolhida por amostragem aleatória simples:

Partido	Efectivo
CDU	72
PS	342
PSD	276
CDS/PP	74
Outros	54
Indecisos	188
Total	1006

Vamos limitar-nos ao cálculo dos intervalos de confiança de Wald e de Agresti-Coull. Uma vez que o tamanho da amostra é muito grande, os dois métodos produzem intervalos semelhantes.

Intervalos de confiança de Wald

Atendendo a que $n = 818$ (estamos a excluir os indecisos), temos

$$\widehat{p}_{cdu} = \frac{72}{818} \approx 0{,}08802,$$
$$\widehat{p}_{ps} = \frac{342}{818} \approx 0{,}41809,$$
$$\widehat{p}_{psd} = \frac{276}{818} \approx 0{,}33741,$$
$$\widehat{p}_{cds/pp} = \frac{74}{818} \approx 0{,}09046.$$

o que permite obter os intervalos de confiança

$$\begin{array}{rcl} \text{CDU} & - & 0{,}08802 \pm 0{,}01942 \\ \text{PS} & - & 0{,}41809 \pm 0{,}03380 \\ \text{PSD} & - & 0{,}33741 \pm 0{,}03240 \\ \text{CDS/PP} & - & 0{,}09046 \pm 0{,}19657 \end{array}$$

ou, em termos percentuais

$$\begin{array}{rcr} \text{CDU} & - & 8{,}80 \pm 1{,}94\% \\ \text{PS} & - & 41{,}81 \pm 3{,}38\% \\ \text{PSD} & - & 33{,}74 \pm 3{,}24\% \\ \text{CDS/PP} & - & 9{,}05 \pm 1{,}97\% \end{array}$$

Intervalos de confiança de Agresti-Coull

Tendo em conta as definições de $\widetilde{n}$ e de $\widetilde{p}$ temos então

$$\widetilde{n} = 818 + 4 = 822,$$
$$\widetilde{p}_{cdu} = \frac{72 + 2}{822} \approx 0{,}09002,$$
$$\widetilde{p}_{ps} = \frac{342 + 2}{822} \approx 0{,}41849,$$
$$\widetilde{p}_{psd} = \frac{276 + 2}{822} \approx 0{,}33820,$$
$$\widetilde{p}_{cds/pp} = \frac{74 + 2}{822} \approx 0{,}09246.$$

Os intervalos de confiança de Agresti-Coull são assim

$$\begin{array}{rcl} \text{CDU} & - & 0{,}08995 \pm 0{,}01956 \\ \text{PS} & - & 0{,}41848 \pm 0{,}03373 \\ \text{PSD} & - & 0{,}33817 \pm 0{,}03234 \\ \text{CDS/PP} & - & 0{,}09360 \pm 0{,}02002 \end{array}$$

ou, em termos percentuais:

$$\begin{array}{rcr} \text{CDU} & – & 9{,}00 \pm 1{,}96\% \\ \text{PS} & – & 41{,}85 \pm 3{,}37\% \\ \text{PSD} & – & 33{,}82 \pm 3{,}23\% \\ \text{CDS/PP} & – & 9{,}36 \pm 2{,}00\% \end{array}$$

Reparemos que para cada um dos partidos temos margens de erro diferentes, enquanto que na ficha técnica da sondagem apenas a margem de erro máxima era referida (ver pág. 5). Como podemos concluir da forma geral dum intervalo de confiança para uma proporção, a margem de erro dum intervalo depende da estatística $\widehat{p}$ (resp. $\widetilde{p}$). Mais precisamente, para uma mesma dimensão da amostra, a margem de erro é máxima quando $\widehat{p} = 0{,}5$ (resp. $\widetilde{p} = 0{,}5$), tornando-se cada vez mais pequena à medida que $\widehat{p}$ (resp. $\widetilde{p}$) se afasta, por excesso ou por defeito, de 0,5.

7.4 Intervalos de confiança para uma média

O método que desenvolvemos para a construção de intervalos de confiança para uma proporção, pode ser adaptado, sem alterações significativas, à construção de intervalos de confiança para uma média, μ, a partir de n observações independentes $x_1, x_2, \ldots, x_n$, que vamos interpretar como sendo realizações duma variável aleatória X com média μ e desvio-padrão σ.

Estando agora interessados na estimação duma média, é natural basearmos a construção dos intervalos de confiança na estatística $\bar{x}$ que, pelo teorema do limite central, sabemos ter uma distribuição de probabilidade aproximadamente normal com média

$$\mu_{\bar{x}} = \mu$$

e com desvio-padrão

$$\sigma_{\bar{x}} = \sigma/\sqrt{n},$$

onde σ é o desvio-padrão de variável X.

Procedendo como no parágrafo anterior, será natural que a construção dum intervalo de confiança para μ seja baseada na variável fulcral

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{7.4.1}$$

que é, para n grande, aproximadamente normal com média 0 e desvio-padrão 1.

Fixado o nível de confiança C, devemos começar por determinar um intervalo da forma $[-z^*, z^*]$ ao qual pertence uma variável normal standard com probabilidade C.

Podemos então dizer que a dupla desigualdade

$$-z^* \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z^*$$

ocorrerá com probabilidade aproximadamente igual a C, ou ainda, que o intervalo

$$\left[\bar{x} - z^*\sigma/\sqrt{n}\,, \bar{x} + z^*\sigma/\sqrt{n} \right]$$

conterá μ com probabilidade aproximadamente igual a C, para n grande. Admitindo que **o desvio-padrão σ é conhecido**, o intervalo anterior pode ser calculado exclusivamente a partir das observações, sendo assim um intervalo de confiança para μ, com nível de confiança aproximadamente igual a C.

O facto do intervalo de confiança apresentado ter nível de confiança efetivo apenas aproximadamente igual a C para n grande, deve-se à aproximação normal que estamos a usar para a distribuição de probabilidade da média amostral. Dizemos neste caso que se trata dum **intervalo de confiança aproximado**. Como já referimos, a qualidade desta aproximação depende fortemente da distribuição subjacente às observações realizadas e da dimensão da amostra. Se esta distribuição é próxima da normal, o nível de confiança efetivo é mais próximo do nível anunciado do que se essa distribuição for, por exemplo, fortemente assimétrica. Enquanto que no primeiro caso podemos usar amostras de tamanho pequeno, no segundo caso somos obrigados a usar amostras de dimensões mais elevadas sob pena de obtermos um intervalo de confiança com um nível de confiança efetivo muito diferente do nível desejado. Neste caso, vários autores aconselham o uso de amostras de dimensão superior ou igual a 30 e outros de amostras de dimensão superior ou igual a 40. No caso limite em que a distribuição da variável observada é normal e o seu desvio-padrão σ é conhecido, sabemos que a distribuição amostral de $\bar{x}$ é também normal, o que implica que o intervalo de confiança anterior tenha nível de confiança exatamente igual a C. Dizemos neste caso que se trata dum **intervalo de confiança exato**.

Intervalo de confiança para uma média com σ conhecido:

Se $\bar{x}$ é calculada a partir de n observações independentes com média μ e desvio-padrão σ conhecido, então um intervalo de confiança de nível C para μ tem por extremidades:

$$\bar{x} \pm z^* \sigma/\sqrt{n}$$

onde z^* é tal que

$$\mathrm{P}(-z^* \leq Z \leq z^*) = C$$

e Z tem uma distribuição normal $N(0,1)$. Este intervalo de confiança é exato quando as observações são normais, e é aproximado nos outros casos, quando é n grande.

No caso em que σ **é desconhecido**, que é a situação mais comum na prática, é natural basear a construção dum intervalo de confiança na variável (7.4.1), em que o valor desconhecido σ é substituído pelo desvio-padrão amostral s. No entanto, a nova variável fulcral

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{7.4.2}$$

não goza das mesmas propriedades que a variável (7.4.1). Mesmo no caso em que as observações são normais, esta variável não é normal. No entanto, para observações normais a distribuição de probabilidade da variável anterior é conhecida. Trata-se duma distribuição, a que chamamos **distribuição de Student**, que depende dum parâmetro designado por **grau de liberdade**.

Distribuição t de Student:

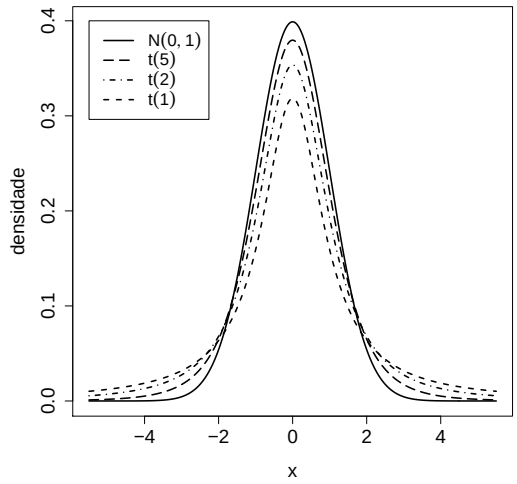
Se $\bar{x}$ e s são calculados a partir de n observações normais e independentes com média μ e desvio-padrão σ, então a variável

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

tem uma distribuição t de Student com $n-1$ graus de liberdade.

A **distribuição de Student** com k graus de liberdade é uma distribuição próxima da normal standard para valores moderados e grandes de k. A distribuição de Student tem média 0 quando $k > 1$, e a sua densidade de probabilidade tem, tal como a normal,

uma forma de sino, sendo simétrica relativamente à origem. Na figura seguinte, para alguns valores de k, apresentamos as densidades de probabilidade da distribuição de Student que denotamos por $t(k)$.



Tal como para a distribuição normal, o cálculo de áreas sob uma curva densidade de Student pode ser feito utilizando a Tabela D (pág. 317) onde estão tabeladas algumas dessas áreas para vários graus de liberdade. Reparemos que a última linha da tabela é precisamente a correspondente à da distribuição normal standard.

Voltemos à questão da construção de intervalos de confiança para a média μ duma população normal, quando o desvio-padrão σ é desconhecido. Fixado um nível de confiança C, começamos por usar a tabela da distribuição de Student para determinar o intervalo da forma $[-t^*, t^*]$ ao qual pertence, com probabilidade C, uma variável de Student com $n-1$ grau de liberdade, onde n é a dimensão da amostra. Isto é, começamos por determinar o valor t^* para o qual se tem

$$\mathrm{P}(-t^* \leq T \leq t^*) = C$$

onde T tem uma distribuição de student $t(n-1)$.

Atendendo à simetria da distribuição de Student, a determinação de t^* é análoga à determinação de z^* para a distribuição normal, mas contrariamente ao caso da distribuição normal, o valor t^* depende de n.

Podemos então dizer que a dupla desigualdade

$$-t^* \leq \frac{\bar{x} - \mu}{s/\sqrt{n}} \leq t^*$$

ocorrerá com probabilidade (exatamente) igual a C, ou ainda, que o intervalo

$$\left[\, \bar{x} - t^* s/\sqrt{n}\,, \bar{x} + t^* s/\sqrt{n} \,\right]$$

conterá μ com probabilidade igual a C, para todo o valor de n.

No caso em que σ **é desconhecido mas as observações não são normais**, a variável (7.4.2) não possui uma distribuição de Student, mas é, para n grande, aproximadamente normal com média 0 e desvio-padrão 1. Como a distribuição $t(n-1)$ de Student é também aproximadamente normal standard quando n é grande, podemos concluir que o intervalo de confiança anterior é ainda um intervalo de confiança, de nível aproximadamente igual a C, para a média duma população não normal com desvio-padrão desconhecido.

Intervalo de confiança para uma média com σ desconhecido:

Se $\bar{x}$ é calculada a partir de n observações independentes com média μ e desvio-padrão σ desconhecido, então um intervalo de confiança de nível C para μ tem por extremidades:

$$\bar{x} \pm t^* s/\sqrt{n}$$

onde t^* é tal que

$$\mathrm{P}(-t^* \leq T \leq t^*) = C$$

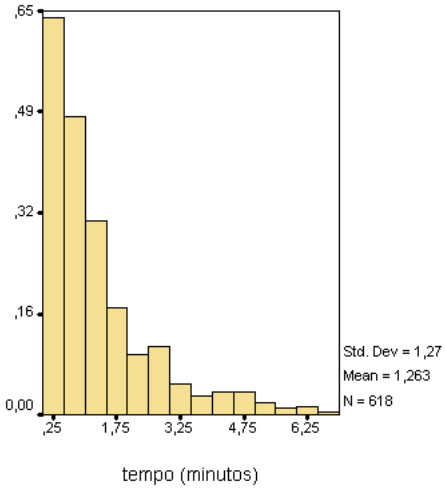
e T tem uma distribuição de Student $t(n-1)$. Este intervalo de confiança é exato quando as observações são normais, e é aproximado nos outros casos, quando é n grande.

Sendo os intervalos de confiança anteriores baseados em medidas de localização e dispersão, que vimos serem pouco robustas contra a presença de observações discordantes, é essencial usar os métodos que já estudámos para identificar e, se for caso disso, eliminar tais observações. Atendendo ao teorema do limite central, vamos considerer que os intervalos de confiança anteriores são **robustos contra a não verificação da hipótese de normalidade** quando o tamanho da amostra satisfaz $n \geq 40$. Quer isto dizer que verificando-se esta condição sobre a dimensão da amostra, os níveis de confiança efetivos dos intervalos apresentados são muito próximos dos anunciados. Para amostras com $15 \leq n < 40$, os intervalos podem ser usados a não ser que haja observações discordantes ou a distribuição das observações seja fortemente assimétrica. Para amostras de dimensão $n < 15$ os intervalos de confiança devem ser usados apenas quando os dados são aproximadamente normais e não haja observações discordantes (Moore e McCabe, 2006, p. 463).

Vejamos três exemplos da determinação de intervalos de confiança para conjuntos

de dados considerados noutros capítulos.

Exemplo 7.4.3 Consideremos as observações descritas no Exemplo 1.2.8 (pág. 26) que a seguir reproduzimos, relativas ao tempo (em minutos) que medeia a chegada de dois clientes consecutivos a uma caixa dum hipermercado.



Determinemos um intervalo de confiança, de nível 0,99, para o tempo médio de interchegada de clientes. O intervalo de confiança que vamos calcular é apenas aproximado uma vez que a distribuição subjacente às observações é fortemente assimétrica, não sendo, por isso, normal. No entanto, atendendo à elevada dimensão da amostra, será de esperar que os intervalos de confiança obtidos pelo método exposto sejam praticamente exatos, isto é, a sua probabilidade de cobertura é próxima do nível de confiança fixado à partida.

Como para a amostra observada $\bar{x} = 1{,}263$ e $s = 1{,}269$ (desvio-padrão populacional desconhecido), da tabela da distribuição $t(617)$ de Student obtemos o seguinte intervalo de confiança

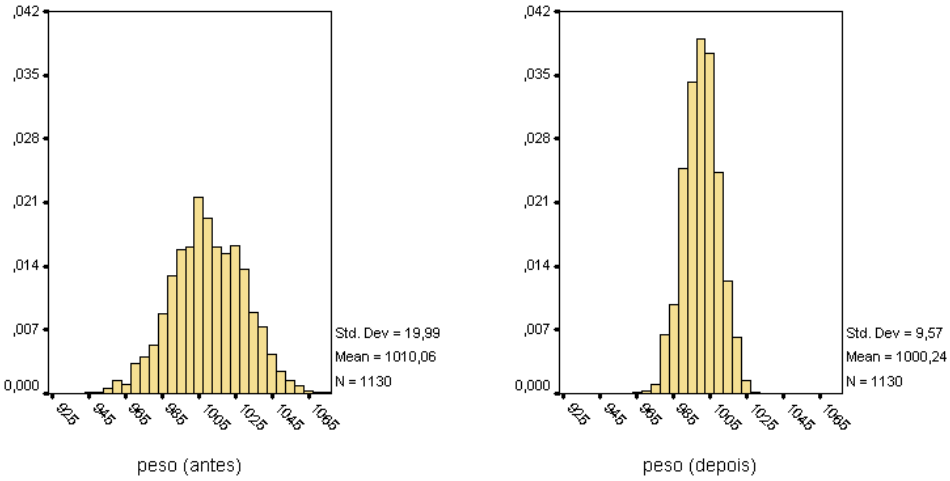
$$1{,}263 \pm 2{,}584 \times 1{,}269/\sqrt{618} \;\longrightarrow\; [1{,}131\,;1{,}395].$$

Usando o SPSS podemos também obter intervalos de confiança para a média apenas no caso em que o desvio-padrão é desconhecido. No caso presente obtemos:

Descriptives

			Statistic
tempo (minutos)	Mean		1,2626
	99% Confidence Interval for Mean	Lower Bound	1,1307
		Upper Bound	1,3945
	Std. Deviation		1,26918

Exemplo 7.4.4 Consideremos agora os dois conjuntos de dados descritos pelos histogramas do Exemplo 1.2.5 (pág. 24), que a seguir reproduzimos, relativos aos pesos (em gramas) de pacotes de açúcar empacotados por uma máquina antes e depois de ter sido calibrada.



Determinemos intervalos de confiança, de nível 0,95, para os pesos médios dos pacotes de açúcar empacotados pela máquina antes e depois de ter sido calibrada. Da tabela da distribuição de Student obtemos os seguintes intervalos de confiança

$$1010{,}06 \pm 1{,}962 \times 19{,}99/\sqrt{1130} \longrightarrow [1008{,}89\,;1011{,}23]$$

e

$$1000{,}24 \pm 1{,}962 \times 9{,}57/\sqrt{1130} \longrightarrow [999{,}68\,;1000{,}80].$$

Usando o SPSS, obtemos os intervalos:

Descriptives

	maquina			Statistic
peso (gramas)	antes	Mean		1010,064
		95% Confidence Interval for Mean	Lower Bound	1008,897
			Upper Bound	1011,231
		Std. Deviation		19,994
	depois	Mean		1000,236
		95% Confidence Interval for Mean	Lower Bound	999,677
			Upper Bound	1000,794
		Std. Deviation		9,573

As diferenças entre os intervalos por nós calculados e os que constam do quadro anterior devem-se unicamente a erros de arredondamento.

Exemplo 7.4.5 A partir das 64 medições não discordantes efetuadas por Newcomb sobre a velocidade da luz (ver Exemplo 1.2.6, pág. 25), e que vimos poderem ser descritos por uma distribuição normal (ver o último dos gráficos de quantis normais da pág. 137), podemos obter o intervalo de confiança seguinte, de nível de confiança 0,95, para a velocidade da luz:

Descriptives

			Statistic
tempo	Mean		27,750
	95% Confidence Interval for Mean	Lower Bound	26,480
		Upper Bound	29,020
	Std. Deviation		5,0834

Trata-se do intervalo centrado em 27,75 com margem de erro de 1,27.

7.5 Como escolher o tamanho da amostra

Vimos nos parágrafos anteriores que a margem de erro dum intervalo de confiança para uma proporção, p, ou para uma média, μ, diminui à medida que o número de observações aumenta. Neste parágrafo discutimos a questão da determinação do tamanho da amostra necessário para obter uma margem de erro inferior ou igual a um valor fixado à partida.

7.5.1 Caso da estimação duma proporção

No caso da estimação duma proporção, e independentemente do intervalo que decidirmos usar (Wald, Wilson ou Agresti-Coull), a escolha do tamanho da amostra será por nós feita a partir da fórmula da margem de erro do intervalo de Wald.

No caso da estimação duma proporção, sendo $\widehat{p}$ aproximadamente igual a p, para n grande, a margem de erro do intervalo de Wald é aproximadamente igual a

$$\text{margem de erro} = z^* \sqrt{p(1-p)/n}.$$

Aumentando o tamanho da amostra podemos reduzir a margem de erro tanto quanto queiramos. Assim, se pretendemos um intervalo de confiança com uma margem de erro inferior ou igual a um valor E fixado à partida,

$$\text{margem de erro} \leq E,$$

devemos escolher n de modo que

$$\begin{aligned} z^* \sqrt{\frac{p(1-p)}{n}} &\leq E \\ z^* \sqrt{p(1-p)} &\leq E\sqrt{n} \\ (z^*)^2 p(1-p) &\leq E^2 n \\ \frac{(z^*)^2 p(1-p)}{E^2} &\leq n \end{aligned}$$

ou seja:

Tamanho da amostra na estimação duma proporção:

$$n \geq (z^*)^2 \, \frac{p(1-p)}{E^2}$$

Sendo p desconhecido, a fórmula anterior só pode ser usada se tivermos uma ideia aproximada sobre o verdadeiro valor de p. Esse valor aproximado pode, por exemplo, ser obtido se tivermos uma estimativa de p obtida num estudo anteriormente realizado, ou se desenvolvermos um estudo preliminar baseado numa amostra de pequena dimensão com o intuito de obter uma estimativa inicial para p.

Outra forma de resolver o problema é tomar na fórmula anterior $p = 0{,}5$, uma vez que o produto $p(1-p)$ é máximo para este valor de p. Neste caso, somos conduzidos à seguinte regra de escolha de n:

Tamanho da amostra na estimação duma proporção, na ausência de qualquer informação sobre p:

$$n \geq \frac{(z^*)^2}{4\,E^2}$$

Ao usarmos esta regra, a dimensão da amostra é por vezes superior ao que seria necessário se conhecessemos uma aproximação, mesmo que grosseira, de p. Por exemplo, se pretendemos um intervalo de nível de confiança 0,95 e soubermos que o verdadeiro valor de p não é superior a 0,2, bastará uma amostra de tamanho 246 para obtermos uma margem de erro inferior ou igual a 0,05. Com efeito,

$$n \geq (1{,}96)^2 \frac{0{,}2(1-0{,}2)}{(0{,}05)^2} = 245{,}86$$

Usando a fórmula anterior, somos levados a recolher uma amostra com dimensão igual ou superior a 385, pois

$$n \geq \frac{(1{,}96)^2}{4(0{,}05)^2} = 384{,}16$$

Quando as observações custam dinheiro, a diferença entre os valores anteriores pode ser importante.

7.5.2 Caso da estimação duma média

No caso da estimação duma média μ, a margem de erro, para n grande, é aproximadamente igual a

$$\text{margem de erro} = z^* \sigma / \sqrt{n}.$$

Tal como atrás, se pretendemos um intervalo de confiança com uma margem de erro inferior ou igual a um valor E fixado à partida, devemos escolher n de modo que

$$z^* \sigma / \sqrt{n} \leq E,$$

ou seja:

Tamanho da amostra na estimação duma média:

$$n \geq \frac{(z^*)^2 \sigma^2}{E^2}$$

Sendo σ conhecido, a fórmula anterior pode ser diretamente utilizada. Sendo σ desconhecido, a fórmula anterior só pode ser usada se tivermos uma ideia aproximada sobre o verdadeiro valor de σ. Esse valor aproximado pode, por exemplo, ser obtido se tivermos uma estimativa de σ obtida num estudo anteriormente realizado, ou num estudo preliminar baseado numa amostra de pequena dimensão. Em alternativa, podemos também ter uma ideia do valor máximo que σ pode assumir na população em causa. Neste caso, a utilização da fórmula anterior conduz a um valor de n superior ao que seria necessário para obter a margem de erro desejada.

7.6 Bibliografia

Brown, L.D., Cai, T.T., DasGupta, A. (2001). Interval estimation for binomial proportion, *Statistical Science*, 16, 101–133 (para informação adicional sobre os intervalos de Wilson e de Agresti-Coull).

Daniel, W.W. (2009). *Biostatistics: a foundation for analysis in the health sciences*, Wiley.

Dawson, B., Trapp, R.G. (2004). *Basic & Clinical Biostatistics*, McGraw-Hill.

McPherson. G. (1990). *Statistics in Scientific Investigation: its basis, application and interpretation*, Springer-Verlag.

Moore, D.S. (1985). *Statistics: concepts and controversies*, W.H. Freeman and Company.

Moore, D.S., McCabe, G.P. (2006). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

8

Testes de hipóteses para proporções e médias

Noção de teste de hipóteses. Hipótese nula e hipótese alternativa. Estatística de teste. p-valor. Nível de significância. Testes de hipóteses para proporções e médias.

8.1 Generalidades sobre testes de hipóteses

Neste capítulo estudamos outro dos procedimentos do âmbito da estatística inferencial, que, conjuntamente com a estimação intervalar, é dos procedimentos mais usados por todos aqueles que utilizam a Estatística. Estamos a referir-nos aos **testes de hipóteses**, cujo objetivo principal é analisar a compatibilidade das observações realizadas com uma hipótese formulada *a priori* sobre a população. Tal como fizemos no capítulo anterior, vamos centrar a nossa atenção nos casos em que essa hipótese, que é traduzida por uma hipótese sobre um parâmetro associado à população, é uma hipótese sobre uma proporção, p, ou sobre uma média, μ.

Vejamos um exemplo típico dum problema de testes de hipóteses.

Exemplo 8.1.1 Suponhamos que ao observarmos alguns lançamentos dum dado aparentemente vulgar, suspeitamos que o dado é viciado, mais precisamente, que a probabilidade de ocorrência da face 6 é superior à dum dado equilibrado. Pretendendo averiguar se há boas razões para acreditar que a probabilidade p de ocorrência da face 6 é efetivamente superior a 1/6, calculámos a proporção $\widehat{p}$ de faces 6 obtidas em 100 lançamentos do dado. Observámos 22 vezes a face 6, isto é, a proporção de faces 6 observada foi de $\widehat{p} = 22/100 = 0{,}22$. A questão que agora se coloca é a de saber se aquilo que observámos é natural ocorrer num dado não viciado, ou, pelo contrário, é pouco usual.

Se o dado não fosse viciado, ou seja, se a probabilidade p de ocorrência da face 6 fosse 1/6, sabemos que a proporção amostral $\widehat{p}$ seria aproximadamente normal com

média

$$\mu_{\widehat{p}} = \frac{1}{6} \approx 0{,}1667$$

e desvio-padrão

$$\sigma_{\widehat{p}} = \sqrt{\frac{1}{6}\left(1-\frac{1}{6}\right)\Big/100} \approx \sqrt{0{,}1667(1-0{,}1667)/100} = 0{,}03727,$$

ou ainda,

$$z = \frac{\widehat{p} - 0{,}1667}{0{,}03727} \simeq N(0,1).$$

Isto permite concluir, que se o dado fosse equilibrado, o acontecimento

$$z < 1{,}645,$$

teria grande probabilidade de ocorrer, mais precisamente

$$\mathrm{P}(z < 1{,}645) \approx 0{,}95,$$

isto é, em aproximadamente 95% das vezes em que efetuássemos 100 lançamentos do dado ocorreria o acontecimento anterior.

Atendendo a que o raciocínio anterior foi efetuado assumindo que o dado não era viciado, podemos concluir que se para a proporção observada se verificar $z \geq 1{,}645$, isso será uma indicação de que as observações realizadas não são compatíveis com a hipótese do dado não ser viciado. Uma vez que se observou $\widehat{p} = 0{,}22$, temos

$$z = \frac{\widehat{p} - 0{,}1667}{0{,}03727} = \frac{0{,}22 - 0{,}1667}{0{,}03727} = 1{,}430,$$

e portanto, as observações realizadas são compatíveis com a hipótese do dado ser equilibrado.

Suponhamos agora que nos 100 lançamentos efetuados tínhamos observado 24 vezes a face 6, isto é, $\widehat{p} = 0{,}24$. Nesse caso,

$$z = \frac{\widehat{p} - 0{,}1667}{0{,}03727} = \frac{0{,}24 - 0{,}1667}{0{,}03727} = 1{,}967,$$

e como $z \geq 1{,}645$ concluiríamos que as observações realizadas não eram compatíveis com a hipótese do dado ser equilibrado.

No exemplo anterior estão todos os ingredientes que podemos encontrar num qualquer problema de testes de hipóteses:

1) Em primeiro lugar, é formulada sobre a população uma hipótese que pretendemos ver testada pois esperamos, ou suspeitamos, que não seja verdadeira. Esta hipótese

traduz normalmente uma afirmação de "ausência de efeito" ou "ausência de diferença". Por oposição a esta hipótese, é formulada uma outra hipótese que suspeitamos ser verdadeira. À primeira hipótese damos o nome de **hipótese nula**, e denota-mo-la por H_0, enquanto que à segunda chamamos **hipótese alternativa** ou **hipótese experimental**, e denota-mo-la por H_a. Ambas as hipóteses são formuladas em termos dum parâmetro populacional.

No exemplo anterior elas são dadas por

$$H_0 : \ p = 1/6 \quad \text{e} \quad H_a : \ p > 1/6,$$

onde p representa a probabilidade de ocorrência da face 6 no lançamento do dado.

Um teste de hipóteses surge assim como um procedimento estatístico que nos permite medir, em termos de probabilidade, a evidência que os dados comportam contra a hipótese nula. A hipótese alternativa indica-nos quais os valores do parâmetro que devemos considerar contra a hipótese nula. Quer uma, quer outra das hipóteses em confronto, deve ser formulada antes de recolhermos os dados que vamos utilizar para efetuar o teste.

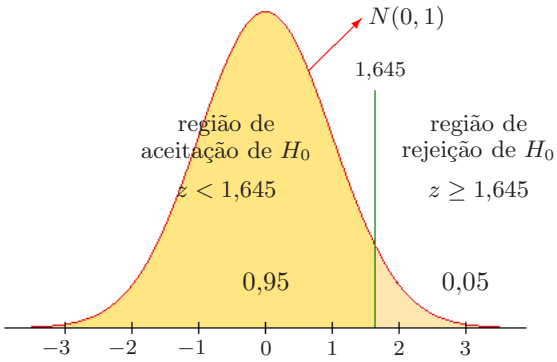
2) Em segundo lugar, para medir a evidência que os dados comportam contra a hipótese nula, lançamos mão da denominada **estatística de teste**, que no exemplo anterior é dada por

$$z = \frac{\widehat{p} - 0{,}1667}{0{,}03727}.$$

Esta estatística mede a compatibilidade entre a hipótese nula e as observações realizadas.

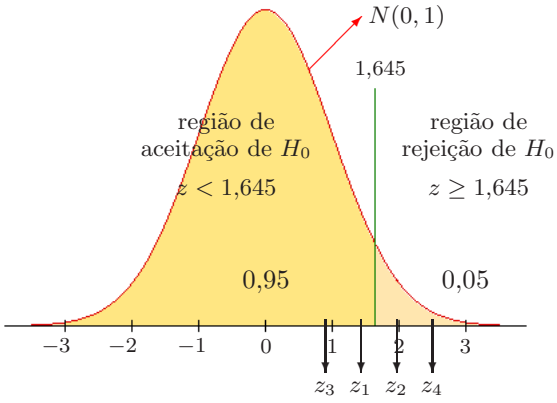
3) Finalmente, a distribuição de probabilidade da estatística de teste, ou uma sua aproximação, é usada para calcular o "ponto de corte" 1,645, que nos permite avaliar se as observações realizadas são, ou não, compatíveis com a hipótese nula, o que acontece quando $z < 1{,}645$ ou $z \geq 1{,}645$, respetivamente. O ponto de corte anterior depende naturalmente da probabilidade 0,95 considerada no exemplo anterior. Outras probabilidades conduziriam a outros pontos de corte.

A figura seguinte ilustra o que acabámos de dizer, apresentando as regiões que conduzem a aceitação e à rejeição da hipótese nula. Esta última região é também denominada **região crítica do teste**.



8.2 Noção de p-valor

No exemplo que considerámos atrás, vimos que os valores $z_1 = 1{,}430$ e $z_2 = 1{,}967$ da estatística de teste, associados às proporções amostrais $\widehat{p} = 0{,}22$ e $\widehat{p} = 0{,}24$, conduziam à aceitação e à rejeição da hipótese nula, respetivamente. O mesmo aconteceria para os valores $z_3 = 0{,}893$ e $z_4 = 2{,}503$ da estatística de teste, associados às proporções amostrais $\widehat{p} = 0{,}2$ e $\widehat{p} = 0{,}26$, respetivamente. No entanto, sendo $\widehat{p} = 0{,}2$ ainda menor do que $\widehat{p} = 0{,}22$, e $\widehat{p} = 0{,}26$ ainda maior do que $\widehat{p} = 0{,}24$, o valor z_3 da estatística de teste mostraria ainda mais compatibilidade com a hipótese nula do que o valor z_1, e o valor z_4 da estatística de teste mostraria mais evidência contra a hipótese nula do que o valor z_2.

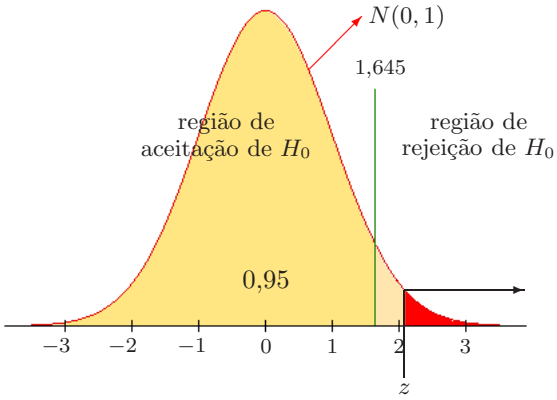


As observações anteriores revelam que para além de sabermos se determinado conjunto de observações revela, ou não, evidência contra a hipótese nula, é também importante quantificar essa evidência, ou seja, quantificar a maior ou menor compatibilidade das observações com a hipótese nula. Tal pode ser conseguido lançando mão da noção

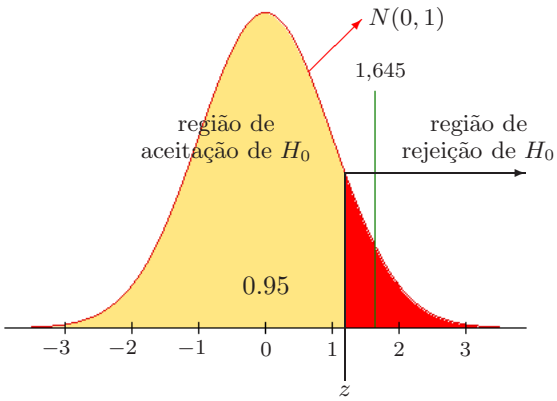
de p-**valor**. O p-valor do teste associado à observação realizada é definido como a probabilidade da estatística de teste tomar um valor que favoreça H_a e que seja tão ou mais extremo do que aquele que foi efetivamente observado.

No caso do exemplo que temos vindo a considerar, sendo z o valor da estatística de teste associado às observações realizadas indicamos nas figuras seguintes o respetivo p-valor, que não é mais do que a probabilidade $\mathrm{P}(Z \geq z)$.

Dos gráficos é claro que quando o p-valor é inferior ou igual a 0,05 a estatística de teste toma valor na região de rejeição, e, portanto, temos evidência contra a hipótese nula. Quando o p-valor é maior que 0,05, a estatística de teste toma valor na região de aceitação, sendo, por isso, os dados compatíveis com a hipótese nula.



p-valor $\leq 0{,}05 \longrightarrow$ rejeição de H_0



p-valor $> 0{,}05 \longrightarrow$ aceitação de H_0

Para cada um dos valores $z_1 = 1{,}430$, $z_2 = 1{,}967$, $z_3 = 0{,}893$ e $z_4 = 2{,}503$, considerados atrás podemos facilmente calcular os respetivos p-valores lançando mão da tabela da distribuição normal standard:

– $\mathrm{P}(Z \geq z_1) = \mathrm{P}(Z \geq 1{,}430) = 0{,}0764$;

– $\mathrm{P}(Z \geq z_2) = \mathrm{P}(Z \geq 1{,}967) = 0{,}0246$;

– $\mathrm{P}(Z \geq z_3) = \mathrm{P}(Z \geq 0{,}893) = 0{,}1859$;

– $\mathrm{P}(Z \geq z_4) = \mathrm{P}(Z \geq 2{,}503) = 0{,}0062$.

Quanto maior é o p-valor, mais forte é a evidência fornecida pelos dados a favor da hipótese nula. Por outro lado, quanto mais pequeno for o p-valor, mais forte é a evidência fornecida pelos dados contra a hipótese nula. O p-valor pode ser assim interpretado como uma medida da credibilidade da hipótese nula, tendo em conta as observações realizadas.

Como vimos no exemplo anterior, a decisão em favor de H_0 acontece quando o p-valor não é muito pequeno, enquanto que a decisão em favor de H_a ocorre quando o p-valor é pequeno. Para transformar esta ideia num verdadeiro procedimento de decisão, é necessário estabelecer à partida um "valor de corte" para o p-valor. Esse valor de corte é habitualmente denotado pela letra grega α a que chamamos **nível de significância do teste**. Assim, se p-valor $\leq \alpha$, decidimos em favor de H_a, e se p-valor $> \alpha$, decidimos em favor de H_0. Um conjunto de observações ou resultado que conduza à aceitação da hipótese H_a, é dito **significativo ao nível** α.

No exemplo anterior $\alpha = 0{,}05$, e os valores z_1 e z_3 conduziram à aceitação de H_0, enquanto que os valores z_2 e z_4 conduziram à sua rejeição. No entanto, se tomarmos $\alpha = 0{,}01$, além dos valores z_1 e z_3, também o valor z_2 conduz à aceitação de H_0, enquanto que o valor z_4 continuará a conduzir à sua rejeição.

Ao escolhermos um teste de nível de significância α, estamos a dizer que, no caso da hipótese H_0 ser verdadeira, rejeitamos H_0 (ou seja, erramos) se o resultado efetivamente observado, ou outro mais extremo, ocorrer não mais do que em $100\alpha\%$ das vezes que repetirmos o processo de amostragem. O valor α pode ser assim interpretado como um limite superior para a probabilidade de incorretamente rejeitarmos a hipótese nula quando ela é verdadeira. Por exemplo, para $\alpha = 0{,}01$, e sendo H_0 verdadeira, rejeitamos H_0 em não mais do que 1% das vezes que repetirmos o processo de amostragem. Quanto mais pequeno for o nível de significância, mais exigentes estamos a ser na evidência que as observações têm que apresentar em favor de H_a, ou equivalentemente, contra H_0.

Tal como fizemos para os intervalos de confiança, os testes de hipóteses para proporções e médias que estudaremos neste capítulo serão apresentados para observações independentes de determinada variável aleatória. Questões relacionadas com observações que não satisfaçam de forma estrita estas condições, ou com a robustez das estatísticas em que basearemos tais testes, foram por nós já abordadas no final do §7.2 e mantêm-se válidas no contexto presente.

8.3 Testes de hipóteses para proporções

Generalizemos o que fizemos no parágrafo anterior ao caso duma qualquer experiência binomial em que efetuamos n observações e pretendemos testar as hipóteses

$$H_0 : p = p_0 \quad \text{contra} \quad H_a : p > p_0 \tag{8.3.1}$$

onde p é a probabilidade de sucesso e p_0 é um valor conhecido e fixo à partida.

Tal como atrás, o teste deverá ser baseado na proporção amostral $\widehat{p}$, cuja distribuição de probabilidade é, sendo a hipótese nula verdadeira, aproximadamente normal com média

$$\mu_{\widehat{p}} = p_0$$

e desvio-padrão

$$\sigma_{\widehat{p}} = \sqrt{p_0(1-p_0)/n}.$$

Obtemos assim a estatística de teste

$$z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

cuja distribuição de probabilidade é aproximadamente normal standard desde que sejam verificadas as condições $np_0 \geq 10$ e $n(1-p_0) \geq 10$ (ver §5.3.5).

Sendo z o valor observado para a estatística de teste, podemos usar a variável normal standard Z para efetuar o cálculo do p-valor associado à observação feita, que, como vimos, é dado por

$$\mathrm{P}(Z \geq z).$$

De forma análoga se procede para testar as hipóteses

$$H_0 : p = p_0 \quad \text{contra} \quad H_a : p < p_0 \tag{8.3.2}$$

ou

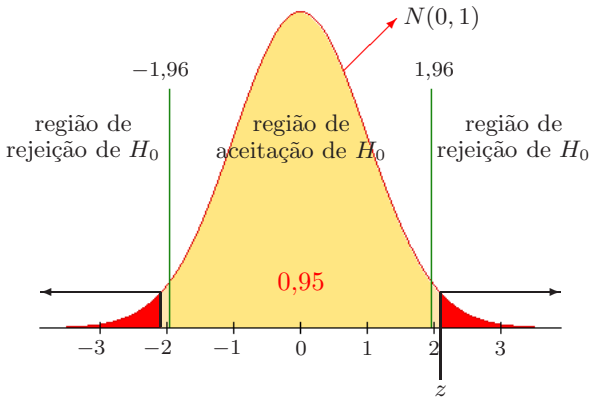
$$H_0 : p = p_0 \quad \text{contra} \quad H_a : p \neq p_0. \tag{8.3.3}$$

A única diferença relativamente ao caso anterior, está no cálculo do p-valor uma vez que, para as hipóteses anteriores, os valores do parâmetro p que são favoráveis a H_a são, no caso (8.3.2), os inferiores a p_0, sendo o p-valor dado por

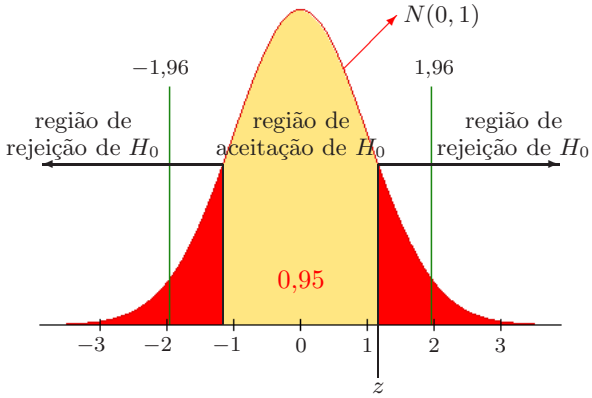
$$\mathrm{P}(Z \leq z),$$

e no caso (8.3.3), os inferiores ou superiores a p_0, sendo o p-valor dado por

$$\mathrm{P}(Z \leq -|z|) + \mathrm{P}(Z \geq |z|) = 2\mathrm{P}(Z \geq |z|).$$



p-valor $\leq 0{,}05 \longrightarrow$ rejeição de H_0



p-valor $> 0{,}05 \longrightarrow$ aceitação de H_0

Nos testes anteriores a hipótese nula $p = p_0$ diz-se **simples** pois diz respeito apenas a um valor do parâmetro. Por oposição, cada uma das hipóteses alternativas consideradas é **composta**. Além disso, as hipóteses alternativas $H_a : p > p_0$ e $H_a : p < p_0$ dizem-se **hipóteses alternativas unilaterais**, enquanto que a hipótese $H_a : p \neq p_0$ diz-se **hipótese alternativa bilateral**.

Há também situações em que interessa considerar testes de **hipótese nula composta unilateral** que poderão ter uma das formas $H_0 : p \leq p_0$ ou $H_0 : p \geq p_0$, tomando as hipóteses alternativas a forma $H_a : p > p_0$ ou $H_a : p < p_0$, respetivamente. Nestes casos procedemos de forma análoga ao que fizemos para os testes das hipóteses (8.3.1) e (8.3.2), respetivamente.

As fórmulas dadas para o cálculo do p-valor do teste usam a aproximação normal para a distribuição de probabilidade da estatística de teste, sendo, por isso, aproximações do verdadeiro p-valor do teste. Neste sentido, para que tais aproximações sejam credíveis é essencial que a dimensão da amostra recolhida verifique as condições

$np_0 \geq 10$ e $n(1-p_0) \geq 10$ (ver §5.3.5).

Testes de hipóteses para uma proporção:

Numa experiência aleatória binomial de parâmetros n e p, para testar a hipótese $H_0 : p = p_0$ (resp. $H_0 : p \leq p_0$, $H_0 : p \geq p_0$), use as observações para calcular

$$z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}},$$

e, de acordo com a hipótese alternativa H_a em causa, calcule a aproximação para o p-valor respetivo dado por uma das fórmulas seguintes, onde Z tem uma distribuição normal $N(0,1)$:

- $H_a : p > p_0$, $\quad \mathrm{P}(Z \geq z)$;
- $H_a : p < p_0$, $\quad \mathrm{P}(Z \leq z)$;
- $H_a : p \neq p_0$, $\quad 2\mathrm{P}(Z \geq |z|)$.

Exemplo 8.3.4 No Exemplo 7.1.1 (pág. 165) colocámos a questão de saber se a moeda na qual observámos 45 vezes a faces europeia em 50 lançamentos da mesma, era ou não equilibrada. Esta questão pode ser formalizada através do teste das hipóteses

$$H_0 : p = 0{,}5 \quad \text{contra} \quad H_a : p \neq 0{,}5$$

onde p denota a probabilidade de ocorrência da face europeia na moeda. (Apesar de podermos assumir que os resultados obtidos indiciam que a face europeia ocorre mais vezes do que seria de esperar numa moeda equilibrada, tal não é tido em conta nas hipóteses formuladas.)

Como referimos atrás, os dados que nos levaram a formular as hipóteses a testar não podem ser usados para efetuar o teste. Neste sentido, suponhamos que efetuamos mais 50 lançamentos da moeda e que desta vez observamos 40 vezes a face europeia.

Seguindo o procedimento descrito atrás, e tendo em conta que $\widehat{p} = 40/50 = 0{,}8$, começamos por calcular

$$z = \frac{0{,}8 - 0{,}5}{\sqrt{0{,}5(1-0{,}5)/50}} \approx 4{,}243$$

sendo o p-valor associado à observação feita dado aproximadamente por (como $np_0 = 50 \times 0{,}5 \geq 10$, é de esperar que esta aproximação seja boa)

$$p\text{-valor} = 2\mathrm{P}(Z \geq |4{,}243|) = 2\mathrm{P}(Z \geq 4{,}243) = 2\mathrm{P}(Z \leq -4{,}243).$$

Usando a Tabela B, concluímos que

$$p\text{-valor} < 2 \times 0{,}0002 = 0{,}0004,$$

o que revela fortíssimos indícios de que a moeda não é equilibrada. Usando uma aplicação estatística ou uma calculadora adequada, podemos mesmo verificar que o p-valor anterior é igual a 0,0000221 o que significa que mesmo para um nível de significância tão pequeno como $\alpha = 0{,}00005$, seríamos levados a rejeitar a hipótese nula.

Reparemos que se tivéssemos observado apenas 10 vezes a face europeia, o resultado do teste seria exatamente o mesmo pois neste caso $\widehat{p} = 0{,}2$ e

$$z = \frac{0{,}2 - 0{,}5}{\sqrt{0{,}5(1-0{,}5)/50}} \approx -4{,}243,$$

sendo o p-valor igual ao que calculámos acima:

$$p\text{-valor} = 2\mathrm{P}(Z \geq |-4{,}243|) = 2\mathrm{P}(Z \leq -4{,}243).$$

Usando o SPSS podemos executar o teste binomial que calcula o p–valor anterior usando a distribuição binomial:

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
moeda	Group 1	europeia	40	,80	,50	,0000239
	Group 2	portuguesa	10	,20		
	Total		50	1,00		

Como podemos verificar o p-valor exacto não difere muito do p–valor calculado atrás (dito, também, assintótico), o que pode ser explicado por se verificarem, no caso presente, as condições $np_0 = n(1-p_0) = 50 \times 0{,}5 = 25 \geq 10$.

Exemplo 8.3.5 Um supermercado compra laranjas a uma cooperativa que afirma que nos frutos que fornece a percentagem de frutos impróprios para consumo não excede 6%. Tendo em conta os últimos lotes de laranjas fornecidos pela cooperativa, o gerente do supermercado suspeita que a percentagem de frutos impróprios para consumo excede 6%, e deseja testar as hipóteses

$$H_0 : p \leq 0{,}06 \quad \text{contra} \quad H_a : p > 0{,}06$$

onde p representa a verdadeira proporção de frutos impróprios para consumo que a cooperativa fornece. Escolhe ainda para nível de significância do teste $\alpha = 0{,}05$. Se o teste conduzir à aceitação de H_a, o gerente reclamará junto da cooperativa.

Sabendo que é importante que a condição $np_0 = n \times 0{,}06 \geq 10$ seja verificada, recolheu-se, por um método aleatório, uma amostra de tamanho 200 do lote em causa. Verificou-se que 15 laranjas estavam impróprias para consumo, ou seja, $\widehat{p} = 15/200 = 0{,}075$. Assim, como

$$z = \frac{0{,}075 - 0{,}06}{\sqrt{0{,}06(1-0{,}06)/200}} \approx 0{,}893,$$

o p-valor associado à observação feita é dado aproximadamente por

$$\mathrm{P}(Z \geq 0{,}893) = \mathrm{P}(Z \leq -0{,}893) = 0{,}1859 > \alpha = 0{,}05.$$

Significa isto que valores tão ou mais extremos do que os que observámos ocorrem em mais de 5% das possíveis repetições do processo de amostragem no caso de H_0 ser verdadeira. A proporção observada de laranjas impróprias para consumo não pode, por isso, ser considerada significativa ao nível $\alpha = 0{,}05$. A este nível de significância não há assim evidência de que as suspeitas do gerente do supermercado tenham fundamento.

Usando o SPSS para executar o teste binomial que calcula o p-valor anterior usando a distribuição binomial, verificamos que o p-valor assintótico calculado atrás não é tão próximo do p-valor exato como acontecia no exemplo anterior. Tal pode ser uma consequência do facto da condição $np_0 = 200 \times 0{,}06 = 12 \geq 10$, apesar de ser verificada, é-o por uma pequena margem.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
laranjas	Group 1	imprópria	15	,075	,06	,222
	Group 2	própria	185	,925		
	Total		200	1,00		

Para terminar reparemos os p-valores anteriores, não nos dão apenas a possibilidade de decidir por uma das duas hipóteses em confronto. Além disso, e principalmente, eles permitem-nos obter uma quantificação da evidência que as observações realizadas nos dão contra a hipótese nula. Esta situação é particularmente clara no primeiro dos exemplos anteriores. Atendendo ao p-valor calculado, sabemos que, se a hipótese nula fosse verdadeira, valores tão ou mais extremos do que os observados ocorreriam em menos de 0,000001% das vezes em que efetuássemos 50 lançamentos da moeda. Temos assim uma fortíssima evidência contra a hipótese da moeda ser equilibrada.

8.4 Testes de hipóteses para médias

O método apresentado nos parágrafos anteriores para testar uma hipótese sobre

uma proporção, pode ser adaptado à construção de testes para a hipótese $H_0 : \mu = \mu_0$ (resp. $H_0 : \mu \leq \mu_0$, $H_0 : \mu \geq \mu_0$) a partir de n observações independentes $x_1, x_2, \ldots, x_n$, que vamos interpretar como sendo realizações duma variável aleatória X com média μ e desvio-padrão σ. Tal como fizemos para os intervalos de confiança, vamos distinguir as situações em que conhecemos, ou não, o desvio-padrão σ da população.

No caso em que o **desvio-padrão σ é conhecido**, é natural basear o teste da hipótese $H_0 : \mu = \mu_0$ na estatística

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

que, sob a hipótese nula, sabemos ter uma distribuição normal standard quando as observações são normais, e ser aproximadamente normal se as observações não são normais e n é grande (ver §6.3).

Por analogia com o que fizemos no parágrafo anterior, somos conduzidos ao procedimento descrito no primeiro dos quadros seguintes para testar uma hipótese sobre uma média duma população no caso do desvio-padrão populacional ser conhecido.

Testes de hipóteses para uma média com σ conhecido:

Para testar a hipótese $H_0 : \mu = \mu_0$ (resp. $H_0 : \mu \leq \mu_0$, $H_0 : \mu \geq \mu_0$), a partir de n observações independentes com média μ e desvio-padrão σ conhecido, calcule

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

e, de acordo com a hipótese alternativa H_a em causa, o p-valor respetivo é dado por uma das fórmulas seguintes, onde Z tem uma distribuição normal $N(0, 1)$:

- $H_a : \mu > \mu_0$, $\quad \mathrm{P}(Z \geq z)$;
- $H_a : \mu < \mu_0$, $\quad \mathrm{P}(Z \leq z)$;
- $H_a : \mu \neq \mu_0$, $\quad 2\mathrm{P}(Z \geq |z|)$.

Estes p-valores são exatos se as observações são normais, e são aproximados nos outros casos quando é n grande.

Quando o **desvio-padrão σ é desconhecido**, é natural basear o teste da hipótese $H_0 : \mu = \mu_0$ na estatística

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

que, sob a hipótese nula, sabemos ter uma distribuição $t(n-1)$ de Student quando as observações são normais, e ser aproximadamente normal se as observações não são

normais e n é grande (ver §7.4). Como a distribuição $t(n-1)$ de Student é também aproximadamente normal standard quando n é grande, o procedimento descrito no segundo dos quadros seguintes permite testar uma hipótese sobre uma média duma população no caso do desvio-padrão populacional ser desconhecido.

Testes de hipóteses para uma média com σ desconhecido:

Para testar a hipótese $H_0 : \mu = \mu_0$ (resp. $H_0 : \mu \leq \mu_0$, $H_0 : \mu \geq \mu_0$), a partir de n observações independentes com média μ e desvio-padrão σ desconhecido, calcule

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

e, de acordo com a hipótese alternativa H_a em causa, o p-valor respetivo é dado por uma das fórmulas seguintes, onde T tem uma distribuição $t(n-1)$ de Student:

- $H_a : \mu > \mu_0$, $\quad \mathrm{P}(T \geq t)$;
- $H_a : \mu < \mu_0$, $\quad \mathrm{P}(T \leq t)$;
- $H_a : \mu \neq \mu_0$, $\quad 2\mathrm{P}(T \geq |t|)$.

Estes p-valores são exatos se as observações são normais, e são aproximados nos outros casos quando é n grande.

Atendendo ao teorema do limite central, verifica-se que os p-valores anteriores são robustos contra a não verificação da hipótese de normalidade quando o tamanho da amostra satisfaz $n \geq 40$. Para amostras com $15 \leq n < 40$, os p-valores podem ser usados a não ser que haja observações discordantes ou a distribuição das observações seja fortemente assimétrica. Para amostras de dimensão $n < 15$ os p-valores devem ser usados apenas quando os dados são aproximadamente normais e não haja observações discordantes (Moore e McCabe, 2006, p. 463).

Exemplo 8.4.1 A partir dos dados sobre pesos (em gramas) de pacotes de açúcar empacotados por uma máquina, antes e depois desta ter sido calibrada, descritos no Exemplo 7.4.4 (pág. 182), testemos, ao nível 0,01, a hipótese do peso médio dos pacotes de açúcar ser de

$$H_0 : \mu = 1000 \quad \text{contra a hipótese} \quad H_a : \mu \neq 1000.$$

Para os pesos dos pacotes antes da calibragem da máquina temos

$$t = \frac{1010{,}06 - 1000}{19{,}99/\sqrt{1130}} \approx 16{,}92,$$

com p-valor de

$$2\mathrm{P}(T \geq |16{,}917|) = 2\mathrm{P}(T \geq 16{,}917),$$

onde T tem uma distribuição de Student $t(1130-1) = t(1129)$. Usando a Tabela D não podemos calcular a probabilidade anterior. No entanto, podemos dizer que é inferior a $2 \times 0{,}001 = 0{,}002$, o que significa que rejeitamos a hipótese da máquina estar bem calibrada ao nível 0,01 (e também ao nível 0,002).

Depois de calibrada, temos

$$t = \frac{1000{,}24 - 1000}{9{,}57/\sqrt{1130}} \approx 0{,}84,$$

sendo o p-valor dado por

$$2\mathrm{P}(T \geq |0{,}843|) = 2\mathrm{P}(T \geq 0{,}843) > 2 \times 0{,}1 = 0{,}2,$$

o conduz à aceitação, ao nível 0,01, da hipótese da máquina estar calibrada.

Estes testes podem ser feitos a partir do SPSS. No quadro seguinte, são dados os valores das estatísticas de teste, os graus de liberdade a considerar e os p-valores:

One-Sample Test

Test Value = 1000

	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
peso antes (gramas)	16,461	1083	,00000	9,99385	8,8026	11,1851
peso depois (gramas)	,828	1129	,40798	,23573	-,3230	,7945

Ficamos a saber que o p-valor que tínhamos concluído ser superior a 0,2 é igual a 0,408. O valor indicado para o outro p-valor é de 0,000. Quer num quer noutro caso, tratam-se de aproximações às milésimas dos verdadeiros p-valores. Se pretendermos aproximações com mais casas decimais, também as podemos obter facilmente. No caso do p-valor indicado como 0,000, continuam a surgir zeros nas casas decimais seguintes o que significa que as observações revelam uma fortíssima evidência contra a hipótese nula.

No caso em que as amostras têm dimensão elevada, o facto dum resultado ser considerado estatisticamente significativo pode não querer dizer que a diferença detetada seja relevante, isto é, tenha **significância prática**. Isto acontece devido ao facto de grandes amostras permitirem, em geral, detetar diferenças muito pequenas, que podem não ter significado prático no contexto do problema em análise. Será neste caso sensato apresentar um intervalo de confiança para o parâmetro em que estamos interessados.

No exemplo que estamos a analisar, poderia fazer sentido apresentar um intervalo de confiança para o peso médio do peso dos pacotes de açúcar antes da máquina ter sido calibrada, ou para a diferença $\mu - 1000$. Da tabela anterior concluímos que um intervalo de confiança com nível 95% para $\mu - 1000$ é dado por

$$]8{,}8026\,;\,11{,}1851[.$$

Caberá agora aos entendidos avaliar se a diferença observada tem significado prático. Aparentemente, assim foi entendido uma vez que se procedeu à calibragem da máquina.

Exemplo 8.4.2 O aumento médio do peso dum pinto alimentado com uma ração vulgar é de 360 gramas às três semanas de vida. Usando os dados apresentados no Exemplo 1.2.3 (pág. 17) relativos ao peso de pintos com três semanas aos quais foi ministrada uma nova ração, vamos testar, ao nível 0,05, a hipótese da nova ração ser melhor que a ração habitualmente usada. Trata-se dum teste sobre o peso médio μ de pintos alimentados com a nova ração, cujas hipóteses nula e alternativa são

$$H_0 : \mu = 360$$

(corresponde à situação de não alteração), e

$$H_a : \mu > 360$$

(corresponde à situação que queremos ver testada), respetivamente.

Como $\bar{x} = 403{,}2$ e $s = 43{,}42$, temos então

$$t = \frac{403{,}2 - 360}{43{,}42/\sqrt{20}} \approx 4{,}45,$$

sendo o p-valor dado por $\mathrm{P}(T \geq 4{,}45)$ onde T tem um distribuição $t(19)$. Da Tabela D concluímos que

$$\mathrm{P}(T \geq 4{,}45) < 0{,}001,$$

ou seja, os resultados obtidos são significativos ao nível 0,05.

Usando o SPSS obtemos o quadro seguinte onde apenas é apresentado o p-valor para o teste de hipótese alternativa bilateral, ou seja,

$$\mathrm{P}(|T| \geq 4{,}45) = \mathrm{P}(T \leq -4{,}45) + \mathrm{P}(T \geq 4{,}45) = 0{,}000275.$$

One-Sample Test

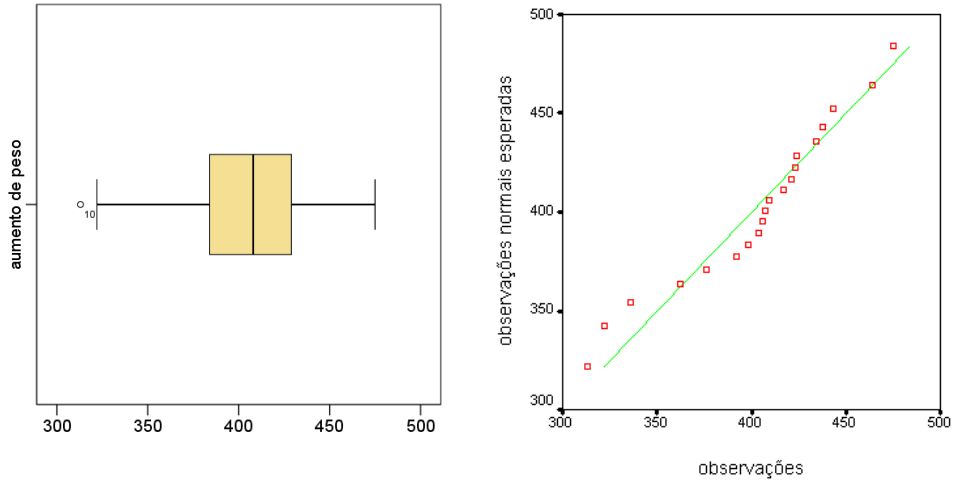
Test Value = 360

	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
aumento de peso	4,450	19	,000275	43,200	22,88	63,52

Atendendo à simetria da distribuição de Student, sabemos que $\mathrm{P}(T \leq -4{,}45) = \mathrm{P}(T \geq 4{,}45)$, de onde obtemos facilmente o p-valor para o nosso caso

$$\mathrm{P}(T \geq 4{,}45) = 0{,}000275/2 = 0{,}0001375.$$

Apesar do gráfico de quantis normais relativo aos dados em análise revelar que a distribuição dos dados possa não ser bem descrita por uma distribuição normal, o gráfico de extremos-e-quartis revela que a distribuição dos dados, apesar de assimétrica negativa, não apresenta uma assimetria muito marcada.



Neste gráfico podemos identificar uma observação discordante cuja influência será importante analisar. A tabela seguinte é produzida retirando a observação discordante do conjunto de dados. Passamos assim a trabalhar com uma amostra de tamanho 19.

One-Sample Test

	Test Value = 360					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
aumento de peso	5,371	18	,000042	47,947	29,19	66,70

Como podemos constatar, o resultado da aplicação do teste não se altera. O resultado observado continua a ser significativo ao nível 0,05. Não sendo a observação influente, podemos decidir mantê-la no estudo, especialmente por não ser uma observações fortemente discordante.

Concluímos assim que o p-valor anterior pode ser considerado fidedigno (dimensão da amostra entre 15 e 40, e distribuição dos dados não apresenta assimetria negativa muito marcada, apesar de haver uma observação discordante) revelando os dados uma forte evidência contra a hipótese nula.

8.5 Bibliografia

Daniel, W.W. (2009). *Biostatistics: a foundation for analysis in the health sciences*, Wiley.

Dawson, B., Trapp, R.G. (2004). *Basic & Clinical Biostatistics*, McGraw-Hill.

Moore, D.S. (1985). *Statistics: concepts and controversies*, W.H. Freeman and Company.

Moore, D.S., McCabe, G.P. (2006). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

Pestana, D.D., Velosa, S.F. (2002). *Introdução à Probabilidade e à Estatística*, Vol. I, Fundação Calouste Gulbenkian.

9

Comparando proporções e médias de duas populações

Comparação de duas proporções e de duas médias para amostras independentes. Comparação de duas proporções e de duas médias em amostras emparelhas.

9.1 Comparando duas proporções

Neste parágrafo estudamos o problema da comparação de proporções relativas a dois grupos de indivíduos a que chamamos população 1 e população 2. Representando por p_1 e p_2, as proporções de indivíduos de cada uma das populações que possui determinada caraterística em estudo, pretendemos comparar as proporções p_1 e p_2. Essa comparação pode ser feita através dum intervalo de confiança para a diferença $p_1 - p_2$, ou através de um teste da hipótese $H_0 : p_1 = p_2$.

9.1.1 O caso das amostras independentes

Começamos por considerar a situação em que efetuamos n_1 observações independentes da população 1, e n_2 observações independentes da população 2, e que estas duas amostras de dimensões n_1 e n_2 são entre si independentes. Denotando por $\widehat{p_1}$ e por $\widehat{p_2}$ as proporções de indivíduos de cada uma das amostras com a caraterística em estudo, será natural basearmos o intervalo de confiança ou o teste da hipótese na diferença

$$\widehat{p}_1 - \widehat{p}_2,$$

que sabemos possuir, para n grande, uma distribuição aproximadamente normal com média

$$\mu_{\widehat{p}_1 - \widehat{p}_2} = p_1 - p_2$$

e cuja variância é, atendendo à independência das estatísticas $\widehat{p}_1$ e $\widehat{p}_2$, igual à soma das

variância de $\widehat{p}_1$ e $\widehat{p}_2$:

$$\sigma^2_{\widehat{p}_1-\widehat{p}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

Esta variância depende das proporções p_1 e p_2 que desconhecemos, mas quem para n_1 e n_2 grandes, pode ser aproximada por

$$\sigma^2_{\widehat{p}_1-\widehat{p}_2} = \frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}.$$

Procedendo como no Capítulo 7, será natural que a construção dum intervalo de confiança para $p_1 - p_2$ seja baseada na variável fulcral

$$\frac{\widehat{p}_1 - \widehat{p}_2 - (p_1 - p_2)}{\sqrt{\widehat{p}_1(1-\widehat{p}_1)/n_1 + \widehat{p}_2(1-\widehat{p}_2)/n_2}},$$

que possui uma distribuição aproximadamente normal standard. Tal conduz ao chamado intervalo de confiança de Wald para a diferença de duas proporções:

Intervalo de Wald para a diferença de duas proporções (amostras independentes):

Um intervalo de confiança para a diferença de proporções $p_1 - p_2$, com nível de confiança aproximadamente igual a C, tem por extremidades

$$\widehat{p}_1 - \widehat{p}_2 \pm z^* \sqrt{\widehat{p}_1(1-\widehat{p}_1)/n_1 + \widehat{p}_2(1-\widehat{p}_2)/n_2},$$

onde z^* é tal que

$$\mathrm{P}(-z^* \leq Z \leq z^*) = C$$

e Z tem uma distribuição normal $N(0,1)$.

De forma análoga ao que referimos para o intervalo de confiança de Wald para uma proporção, o intervalo de Wald para a diferença de duas proporções pode apresentar uma probabilidade de cobertura muito inferior ao nível de confiança desejado C, quando o tamanho das amostras é pequeno, ou quando uma das proporções é próxima de 0 ou de 1. Neste sentido, para os níveis de confiança habituais $C = 0{,}9, 0{,}95, 0{,}99$, o intervalo de Wald deve ser usado apenas quando o número de sucessos e de insucessos em cada uma das amostras for pelos menos de 10, isto é, $n_i\widehat{p}_i \geq 10$ e $n_i(1-\widehat{p}_i) \geq 10$, para $i = 1, 2$ (Moore e McCabe, 2006, p. 556).

Tal como fizemos para o intervalo de confiança de Wald para uma proporção, vamos corrigir o intervalo anterior juntando a cada uma das amostras 2 pseudo-observações, 1 sucesso e 1 insucesso, o que no total das duas amostras corresponde, tal como no

intervalo para uma proporção, a juntar 4 pseudo-observações, sendo duas sucessos e as outras duas insucessos. Obtemos assim o intervalo de confiança de Agresti-Coull para a diferença de duas proporções:

Intervalo de Agresti-Coull para a diferença de duas proporções (amostras independentes):

Um intervalo de confiança para a diferença de proporções $p_1 - p_2$, com nível de confiança aproximadamente igual a C, tem por extremidades

$$\widetilde{p}_1 - \widetilde{p}_2 \pm z^* \sqrt{\widetilde{p}_1(1-\widetilde{p}_1)/\widetilde{n}_1 + \widetilde{p}_2(1-\widetilde{p}_2)/\widetilde{n}_2},$$

onde $\widetilde{p}_i = \widetilde{X}_i/\widetilde{n}_i$, com $\widetilde{X}_i = X_i + 1$ e $\widetilde{n}_i = n_i + 2$, e z^* é tal que

$$\mathrm{P}(-z^* \leq Z \leq z^*) = C$$

e Z tem uma distribuição normal $N(0,1)$.

Para os níveis de confiança habituais $C = 0{,}9, 0{,}95$ e $0{,}99$, recomenda-se a utilização do intervalo de Agresti-Coull desde que o tamanho de ambas as amostras seja de pelo menos 5 ($n_1, n_2 \geq 5$) (Moore e McCabe, 2006, p. 559).

Se em vez de construir um intervalo de confiança, pretendermos testar a hipótese $H_0 : p_1 = p_2$, sabemos já que devemos começar por estudar a distribuição da diferença $\widehat{p}_1 - \widehat{p}_2$ sob a hipótese nula. Assim, assumindo que a hipótese nula é verdadeira, isto é, que $p_1 = p_2 = p$, onde p é a proporção de indivíduos com a caraterística em estudo em ambas as populações (desconhecida), a média anteriormente calculada de $\widehat{p}_1 - \widehat{p}_2$ é agora igual a zero

$$\mu_{\widehat{p}_1 - \widehat{p}_2} = p_1 - p_2 = p - p = 0,$$

e a sua variância pode ser escrita na forma

$$\sigma^2_{\widehat{p}_1 - \widehat{p}_2} = p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

Uma forma simples de estimar a variância anterior é estimar p a partir da proporção $\widehat{p}$ de indivíduos nas duas amostras que possuem a caraterística em estudo:

$$\widehat{p} = \frac{X_1 + X_2}{n_1 + n_2}, \tag{9.1.1}$$

onde X_1 e X_2 são o número de sucessos em cada uma das amostras.

Concluímos assim, que, sendo H_0 verdadeira, a estatística

$$\frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

é aproximadamente normal standard, sendo esta a estatística de teste que usaremos para testar a hipótese da igualdade das duas proporções.

Teste de comparação de duas proporções (amostras independentes):

Para testar a hipótese $H_0 : p_1 = p_2$ (resp. $H_0 : p_1 \leq p_2$, $H_0 : p_1 \geq p_2$), use as observações para calcular

$$z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}},$$

onde $\widehat{p}$ é dado por (9.1.1), e, de acordo com a hipótese alternativa H_a em causa, calcule a aproximação para o p-valor respetivo dado por uma das fórmulas seguintes, onde Z tem uma distribuição normal $N(0,1)$:

- $H_a : p_1 > p_2$, $\quad \mathrm{P}(Z \geq z)$;
- $H_a : p_1 < p_2$, $\quad \mathrm{P}(Z \leq z)$;
- $H_a : p_1 \neq p_2$, $\quad 2\mathrm{P}(Z \geq |z|)$.

Para que as aproximações dadas para os p-valores do teste anterior sejam credíveis, é essencial que em cada amostra haja pelo menos cinco sucessos e cinco insucessos (Moore e McCabe, 2006, p. 559).

Exemplo 9.1.2 Numa sondagem, publicada pelo semanário *Expresso*, em 28 de Fevereiro de 2004, sobre o posicionamento político dos portugueses, nas áreas metropolitanas de Lisboa e do Porto foram recolhidas amostras aleatórias simples com base na lista telefónica, de dimensões 278 e 145, respetivamente, tendo-se obtido os seguintes resultados:

	Lisboa	Porto
Esquerda	147	71
Direita	103	58
Nenhum	28	16
Total	278	145

Será que com base nos resultados anteriores podemos concluir que a percentagem de eleitores de direita e de esquerda são significativamente diferentes em Lisboa e no Porto?

Comecemos por testar, ao nível de significância 0,05, a hipótese da proporção de eleitores de esquerda ser a mesma em Lisboa (população 1) e no Porto (população 2). Por outras palavras, denotando por p_1 e p_2, respetivamente, tais proporções, pretendemos testar

$$H_0 : p_1 = p_2 \quad \text{contra} \quad H_a : p_1 \neq p_2.$$

Temos

$$\widehat{p} = \frac{147 + 71}{278 + 145} \approx 0{,}5154$$

e

$$z = \frac{0{,}5288 - 0{,}4897}{\sqrt{0{,}5154(1 - 0{,}5154)\left(\frac{1}{278} + \frac{1}{145}\right)}} \approx 0{,}764.$$

O p-valor associado a esta observação é

$$2\mathrm{P}(Z \geq |0{,}764|) = 2 \times 0{,}2224 = 0{,}4448,$$

o que não é significativo ao nível 0,05.

Para os eleitores de direita, temos

$$\widehat{p} = \frac{103 + 58}{278 + 145} \approx 0{,}3806$$

e

$$z = \frac{0{,}3705 - 0{,}4}{\sqrt{0{,}3806(1 - 0{,}3806)\left(\frac{1}{278} + \frac{1}{145}\right)}} \approx -0{,}593.$$

O p-valor associado a esta observação é

$$2\mathrm{P}(Z \geq |-0{,}593|) = 2 \times 0{,}2766 = 0{,}5532,$$

o que também não é significativo ao nível 0,05.

Em geral, os *softwares* estatísticos implementam o teste anterior introduzindo uma correção de continuidade na estatística de teste z, por forma a melhorar a qualidade do p-valor anterior, e reportando o valor de z^2 e não o de z. Quando tal acontece, a

distribuição estatística utilizada para calcular o p-valor bilateral é a **distribuição do qui-quadrado** com um grau de liberdade, que, como veremos no início do próximo capítulo, não é mais do que o quadrado da distribuição normal standard Z^2.

Usando o SPSS obtemos para os eleitores de esquerda:

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	,584[a]	1	,445
Continuity Correction[b]	,438	1	,508
N of Valid Cases	423		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 70,27.

b. Computed only for a 2x2 table

Desta tabela tiramos então que $|z| = \sqrt{0{,}584} = 0{,}764$ (sem correção de continuidade), e $|z| = \sqrt{0{,}438} = 0{,}662$ (com correção de continuidade), não havendo muita diferença entre os dois p-valores reportados.

Se em vez de pretendermos saber se a percentagem de eleitores de esquerda (ou de direita) é significativamente diferente em Lisboa e no Porto, quisessemos estimar a diferença entre as duas percentagens, seríamos conduzidos a construir um intervalo de confiança para a diferença $p_1 - p_2$. No caso dos eleitores de esquerda, o intervalo de confiança de Agresti-Coull para esta diferença, com nível de confiança de 95%, é dado por

$$]-0{,}06097\,;\,0{,}13852\,[.$$

Assim, o intervalo de confiança para a diferença das duas percentagens é

$$]-6{,}097\,;\,13{,}852\,[.$$

Como seria de esperar na sequência do teste de hipóteses realizado, o valor 0 pertence ao intervalo anterior o que significa que a diferença entre as duas percentagens não é significativa.

Exemplo 9.1.3 No Exemplo 3.3.5 (pág. 80) vimos que a probabilidade p_1 de ocorrer a soma 9 no lançamento de três dados equilibrados é menor do que a probabilidade p_2 de ocorrer a soma 10. No quadro seguinte indicam-se as frequências absolutas das somas 9 e 10 em 100, 1000, 10000 e 20000 lançamentos de 3 dados equilibrados. Para garantir a independência entre as duas proporções amostrais, foram simuladas duas séries de 20000 lançamentos.

soma \ n	100	1000	10000	20000
9	9	128	1166	2287
10	10	126	1239	2493

Vejamos para que valores de n mostram os resultados anteriores evidência contra a hipótese de igualdade das duas probabilidades, e em favor da hipótese da probabilidade p_1 ser menor que p_2. Para cada um dos valores anteriores, calculemos os p-valores relativos ao teste da hipótese

$$H_0 : p_1 = p_2 \quad \text{contra} \quad H_1 : p_1 < p_2.$$

Para $n = 100$ temos

$$\widehat{p} = \frac{9 + 10}{100 + 100} = 0{,}095$$

e

$$z = \frac{0{,}09 - 0{,}10}{\sqrt{0{,}095(1 - 0{,}095)\left(\frac{1}{100} + \frac{1}{100}\right)}} \approx -0{,}241.$$

O p-valor é dado por

$$\mathrm{P}(Z \leq -0{,}241) = 0{,}4048,$$

o que não revela evidência contra a hipótese nula.

Para $n = 1000$ é claro que a evidência revelada será ainda menor do que a obtida para $n = 100$ (porquê?). Para $n = 10000$ temos

$$\widehat{p} = \frac{1166 + 1239}{10000 + 10000} = 0{,}12025$$

e

$$z = \frac{0{,}1166 - 0{,}1239}{\sqrt{0{,}12025(1 - 0{,}12025)\left(\frac{1}{10000} + \frac{1}{10000}\right)}} \approx -1{,}587.$$

O p-valor é dado por

$$\mathrm{P}(Z \leq -1{,}587) = 0{,}056,$$

o que revela evidência mais forte contra a hipótese nula.

Para $n = 20000$ temos

$$\widehat{p} = \frac{2287 + 2493}{20000 + 20000} = 0{,}1195$$

e

$$z = \frac{0{,}2287 - 0{,}2493}{\sqrt{0{,}1195(1 - 0{,}1195)\left(\frac{1}{20000} + \frac{1}{20000}\right)}} \approx -6{,}351$$

Usando a Tabela B, podemos afirmar que o p-valor, que é dado por $\mathrm{P}(Z \leq -6{,}351)$, é inferior a 0,0002, o que revela ainda maior evidência contra a hipótese nula. Pode no entanto verificar-se que

$$\mathrm{P}(Z \leq -6{,}351) = 1{,}0696 \times 10^{-10},$$

o que demonstra uma fortíssima evidência contra a hipótese nula.

9.1.2 O caso das amostras emparelhadas

Estamos agora interessados na comparação de duas proporções quando as duas amostras extraídas de cada uma das populações são **amostras são emparelhadas**. Recordando o que já dissemos em §2.2, esta situação ocorre, por exemplo, na comparação de dois tratamentos, quando é possível aplicar ambos os tratamentos num mesmo indivíduo ou em dois indivíduos que sejam semelhantes relativamente a variáveis influentes na variável resposta. No primeiro caso, e quando possível, os dois tratamentos são aplicados ao indivíduo por ordem aleatória, enquanto que no segundo caso os indivíduos emparelhados são afetados a um ou a outro dos grupos de forma aleatória.

Tal como fizemos para o teste de igualdade de duas proporções para amostras independentes que estudámos em §9.1.1, estamos interessados na comparação das proporções p_1 e p_2, onde representamos por p_1 e p_2, as proporções de indivíduos de cada uma das populações que possuem a caraterística em estudo.

Exemplo 9.1.1 Um investigador pretende averiguar se um medicamento tem um efeito positivo no tratamento de uma doença particular. Um total de 314 indivíduos é envolvido no estudo. Antes da administração do medicamento os sujeitos são classificados como doentes (+) ou não doentes (–). O mesmo é feito depois da administração do medicamento. Na tabela seguinte são apresentadas as contagens dos indivíduos com e sem a doença (+/–) antes e depois do tratamento ([1]):

Antes \ Depois	+	−	total
+	131	91	222
−	59	33	92
total	190	124	314

Representando por p_1 e por p_2 a proporção de indivíduos na população que são classificados doentes, antes e depois de lhes ser ministrado o medicamento, pretendemos inferir acerca da diferença $p_1 - p_2$, quer seja construindo um intervalo de confiança para esta diferença, quer testando a hipótese $H_0 : p_1 = p_2$, que traduz a hipótese

[1] Fonte: http://en.wikipedia.org/wiki/McNemar-Test

do medicamento não ter qualquer efeito no tratamento da doença, contra a hipótese alternativa $H_a : p_1 > p_2$, que traduz a hipótese do medicamento ter um efeito positivo no tratamento da doença.

Tal como acontece no exemplo anterior, os dados associados a este tipo de problemas podem ser organizados numa tabela de duas entradas que toma a forma

Antes \ Depois	Sucesso	Insucesso	total
Sucesso	a	b	$a+b$
Insucesso	c	d	$c+d$
total	$a+c$	$b+d$	n

onde, no exemplo anterior, sucesso ou insucesso, designam os indivíduos classificados como doentes, ou não doentes, respetivamente.

Sendo natural basear o teste da hipótese anterior na diferença $\widehat{p}_1 - \widehat{p}_2$, esta é dada por

$$\widehat{p}_1 - \widehat{p}_2 = \frac{a+b}{n} - \frac{a+c}{n} = \frac{b-c}{n},$$

dependendo apenas dos indivíduos que mudaram de estado, passando de sucesso a insucesso e de insucesso a sucesso.

A distribuição de $\widehat{p}_1 - \widehat{p}_2$ é aproximadamente normal com

$$\mu_{\widehat{p}_1 - \widehat{p}_2} = p_1 - p_2$$

e a sua variância não é necessariamente igual à soma das variâncias de $\widehat{p}_1$ e $\widehat{p}_2$, uma vez que as proporções amostrais $\widehat{p}_1$ e $\widehat{p}_2$ não são agora independentes. É possível mostrar que a variância de $\widehat{p}_1 - \widehat{p}_2$ é dada por

$$\sigma^2_{\widehat{p}_1 - \widehat{p}_2} = \frac{1}{n}\Big(p_{10} + p_{01} - (p_1 - p_2)^2\Big),$$

onde p_{10} é a probabilidade dum indivíduo passar de sucesso a insucesso, e p_{01} é a probabilidade dum indivíduo passar de insucesso a sucesso.

Atendendo à lei dos grandes números, sabemos que quando o tamanho da amostra é grande, a probabilidade $p_{10} + p_{01}$ pode ser aproximada por

$$p_{10} + p_{01} \approx \frac{b+c}{n},$$

o que implica que a variância anterior possa ser aproximada por

$$\sigma^2 \approx \frac{1}{n}\left(\Big(\frac{b+c}{n}\Big) - \Big(\frac{b-c}{n}\Big)^2\right).$$

Assim tomando como variável fulcral a variável

$$z = \frac{\frac{b-c}{n} - (p_1 - p_2)}{\sqrt{\frac{1}{n}\left(\left(\frac{b+c}{n}\right) - \left(\frac{b-c}{n}\right)^2\right)}},$$

somos conduzidos ao intervalo de Wald para a diferença de duas proporções em amostras emparelhadas.

Intervalo de Wald para a diferença de duas proporções (amostras emparelhadas):

Um intervalo de confiança para a diferença de proporções $p_1 - p_2$, com nível de confiança aproximadamente igual a C, tem por extremidades

$$\frac{b-c}{n} \pm z^* \sqrt{\frac{1}{n}\left(\left(\frac{b+c}{n}\right) - \left(\frac{b-c}{n}\right)^2\right)},$$

onde z^* é tal que

$$\mathrm{P}(-z^* \leq Z \leq z^*) = C$$

e Z tem uma distribuição normal $N(0,1)$.

O intervalo anterior deve ser usado apenas quando $n \geq 40$. Especialmente quando o tamanho da amostra é pequeno, a probabilidade de cobertura do intervalo de Wald pode ser bastante inferior ao nível de confiança desejado. Apesar de haver alternativas ao intervalo de Wald quando o tamanho da amostra é pequeno, não as vamos considerar neste curso.

Pretendendo agora testar a hipótese $H_0 : p_1 = p_2$, comecemos por estudar a distribuição da diferença $\widehat{p}_1 - \widehat{p}_2$ sob a hipótese nula. Assumindo que a hipótese nula, isto é, que $p_1 = p_2$, sabemos já que, para n grande, a diferença $\widehat{p}_1 - \widehat{p}_2$ possui uma distribuição aproximadamente normal com média

$$\mu_{\widehat{p}_1 - \widehat{p}_2} = p_1 - p_2 = 0$$

e cuja variância é dada por

$$\sigma^2_{\widehat{p}_1 - \widehat{p}_2} = \frac{1}{n}\Big(p_{10} + p_{01}\Big),$$

onde p_{10} é a probabilidade dum indivíduo passar de sucesso a insucesso, e p_{01} é a probabilidade dum indivíduo passar de insucesso a sucesso. Uma vez que esta variância

pode ser aproximada por

$$\sigma^2_{\widehat{p}_1-\widehat{p}_2} \approx \frac{1}{n}\Big(\frac{b+c}{n}\Big),$$

somos assim conduzidos à estatística de teste

$$z = \frac{\dfrac{b}{n} - \dfrac{c}{n}}{\sqrt{\dfrac{1}{n}\Big(\dfrac{b+c}{n}\Big)}} = \frac{b-c}{\sqrt{b+c}},$$

que, sob a hipótese nula e para n grande, possui uma distribuição aproximadamente normal standard.

Teste de McNemar de comparação de duas proporções (amostras emparelhadas):

Para testar a hipótese $H_0 : p_1 = p_2$ (resp. $H_0 : p_1 \leq p_2$, $H_0 : p_1 \geq p_2$) quando as amostras são emparelhadas, use as observações para calcular

$$z = \frac{b-c}{\sqrt{b+c}},$$

e, de acordo com a hipótese alternativa H_a em causa, calcule a aproximação para o p-valor respetivo dado por uma das fórmulas seguintes, onde Z tem uma distribuição normal $N(0,1)$:

- $H_a : p_1 > p_2$, $\quad \mathrm{P}(Z \geq z)$;
- $H_a : p_1 < p_2$, $\quad \mathrm{P}(Z \leq z)$;
- $H_a : p_1 \neq p_2$, $\quad 2\mathrm{P}(Z \geq |z|)$.

Exemplo 9.1.1 (cont.) A partir da tabela que apresenta as contagens dos indivíduos com e sem a doença, antes e depois do tratamento, a estatística de teste é dada por

$$z = \frac{91-59}{\sqrt{91+59}} = 2{,}6128,$$

sendo o p-valor associado dado por

$$p\text{-valor} = \mathrm{P}(Z \geq 2{,}6128) = 0{,}00449,$$

o que revela forte evidência contra a hipótese nula.

Quando usamos um *software* estatístico, o teste anterior é implementado introduzindo uma correção de continuidade na estatística de teste z, por forma a melhorar a qualidade do p-valor anterior, e reportando o valor de z^2 e não o de z. O p-valor indicado é, no caso do SPSS, o p-valor bilateral:

Test Statistics[a]

	antes & depois
N	314
Chi-Square[b]	6,407
Asymp. Sig.	,011

a. McNemar Test

b. Continuity Corrected

De acordo com esta tabela $|z| = \sqrt{6{,}407} = 2{,}531$, valor este que é ligeiramente diferente do valor que obtivemos acima uma vez que o *software* está a calcular a estatística de teste com correção de continuidade. Como já referimos, o p-valor que se indica na tabela anterior é o p-valor para a hipótese alternativa bilateral, ou seja,

$$2\mathrm{P}(Z \geq 2{,}531) = 0{,}011.$$

Isto significa que o p-valor para a hipótese alternativa $H_a : p_1 > p_2$ é dado por

$$p\text{-valor} = \mathrm{P}(Z \geq 2{,}531) = 0{,}011/2 = 0{,}0055,$$

um valor que não é muito diferente do valor que calculámos atrás.

9.2 Comparando duas médias

Neste parágrafo estudamos o problema da comparação das médias, relativas a dois grupos de indivíduos a que chamamos população 1 e população 2, que representamos por μ_1 e μ_2. Essa comparação pode ser feita através dum intervalo de confiança para a diferença $\mu_1 - \mu_2$, ou através de um teste da hipótese $H_0 : \mu_1 = \mu_2$.

9.2.1 O caso das amostras independentes

Começamos por considerar a situação em que efetuamos n_1 observações independentes da população 1, e n_2 observações independentes da população 2, e que estas duas amostras de dimensões n_1 e n_2 são entre si independentes. Denotando por $\bar{x}_1$ e $\bar{x}_2$ as médias relativas a cada uma das amostras, será natural basearmos o intervalo de confiança ou o teste da hipótese na estatística

$$\bar{x}_1 - \bar{x}_2,$$

que, atendendo à independência das duas amostras, é, para n_1 e n_2 grandes, aproximadamente normal com média

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

e com variância

$$\sigma^2_{\bar{x}_1-\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}, \tag{9.2.1}$$

onde σ_1^2 e σ_2^2 são as variâncias das populações 1 e 2, respetivamente. Assim, tendo em conta que a variável

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

é aproximadamente normal standard, podemos proceder como em §7.4 e §8.4, para construir um intervalo de confiança para a diferença $\mu_1 - \mu_2$, e um teste de hipóteses para a hipótese $H_0 : \mu_1 = \mu_2$, caso as variâncias σ_1^2 e σ_2^2 sejam conhecidas.

Intervalo de confiança para a diferença de duas médias com σ_1 e σ_2 conhecidos:

Um intervalo de confiança para a diferença de médias $\mu_1 - \mu_2$, com nível de confiança C, tem por extremidades

$$\bar{x}_1 - \bar{x}_2 \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

onde z^* é tal que

$$\mathrm{P}(-z^* \leq Z \leq z^*) = C$$

e Z tem uma distribuição normal $N(0,1)$. Este intervalo de confiança é exato quando as observações são normais, e é aproximado nos outros casos, quando é n_1 e n_2 são grandes.

Teste de comparação de duas médias com σ_1 e σ_2 conhecidos:

Para testar a hipótese $H_0 : \mu_1 = \mu_2$ (resp. $H_0 : \mu_1 \leq \mu_2$, $H_0 : \mu_1 \geq \mu_2$), de igualdade das médias de duas populações normais com variâncias iguais mas desconhecidas, use as observações para calcular

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

e, de acordo com a hipótese alternativa H_a em causa, o p-valor respetivo é dado por uma das fórmulas seguintes, onde Z tem uma distribuição de normal standard:

- $H_a : \mu_1 > \mu_2$, $\quad \mathrm{P}(Z \geq z)$;
- $H_a : \mu_1 < \mu_2$, $\quad \mathrm{P}(Z \leq z)$;
- $H_a : \mu_1 \neq \mu_2$, $\quad 2\mathrm{P}(Z \geq |z|)$.

Estes p-valores são exatos se as observações são normais, e são aproximados nos outros casos quando n_1 e n_2 são grande.

A situação mais comum na prática é aquela em que as variâncias σ_1^2 e σ_2^2 não são conhecidas. Nesse caso, é natural basear a construção do intervalo de confiança para $\mu_1 - \mu_2$ na variável fulcral

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

que se obtém da anterior substituindo as variâncias desconhecidas pelas variâncias amostrais relativas a cada uma das amostras.

Mesmo no caso em que as duas populações são normalmente distribuídas, esta estatística não possui uma distribuição de Student. No entanto, sendo as duas populações normais, é possível aproximar a distribuição amostral da variável anterior por uma distribuição de Student $t(k)$ onde o número k de graus de liberdade, que pode não ser um número inteiro, é calculado a partir das observações realizadas sendo dado por

$$k = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}. \tag{9.2.2}$$

Intervalo de confiança para a diferença de duas médias:

Um intervalo de confiança para a diferença de médias $\mu_1 - \mu_2$, com nível de confiança C, tem por extremidades

$$\bar{x}_1 - \bar{x}_2 \pm z^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

onde t^* é tal que

$$\mathrm{P}(-t^* \leq T \leq t^*) = C$$

e T tem uma distribuição de Student $t(k)$ com k dado pela fórmula (9.2.2). Este intervalo de confiança é praticamente exato quando as observações são normais, e é aproximado nos outros casos, quando n_1 e n_2 são grandes.

Teste de Welch para a comparação de duas médias:

Para testar a hipótese $H_0 : \mu_1 = \mu_2$ (resp. $H_0 : \mu_1 \leq \mu_2$, $H_0 : \mu_1 \geq \mu_2$), de igualdade das médias de duas populações, use as observações para calcular

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

e, de acordo com a hipótese alternativa H_a em causa, calcule a aproximação para o p-valor respetivo dado por uma das fórmulas seguintes, onde T tem uma distribuição de Student $t(k)$ com k dado pela fórmula (9.2.2):

- $H_a : \mu_1 > \mu_2$, $\quad \mathrm{P}(T \geq t)$;
- $H_a : \mu_1 < \mu_2$, $\quad \mathrm{P}(T \leq t)$;
- $H_a : \mu_1 \neq \mu_2$, $\quad 2\mathrm{P}(T \geq |t|)$.

Estes p-valores são praticamente exatos se as observações são normais, e são aproximados nos outros casos quando n_1 e n_2 são grandes.

No caso das duas populações serem normais e das duas variâncias σ_1^2 e σ_2^2 serem iguais, apesar de desconhecidas, é possível construir uma estatística de teste que possua

uma distribuição de Student. Se $\sigma_1^2 = \sigma_2^2 = \sigma_0^2$, a variância (9.2.1) escreve-se na forma,

$$\sigma^2 = \sigma_0^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

onde σ_0^2, que é a variância comum às duas populações, pode ser estimada combinando as variâncias amostrais s_1^2 e s_2^2 da forma seguinte

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Nestas condições verifica-se que, sendo a hipótese nula verdadeira, a estatística

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

possui uma distribuição de Student, $t(n_1 + n_2 - 2)$, o que dá origem ao seguinte teste.

Teste de comparação de duas médias (variâncias iguais):

Para testar a hipótese $H_0 : \mu_1 = \mu_2$ (resp. $H_0 : \mu_1 \leq \mu_2$, $H_0 : \mu_1 \geq \mu_2$), de igualdade das médias de duas populações normais com variâncias iguais mas desconhecidas, use as observações para calcular

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

e, de acordo com a hipótese alternativa H_a em causa, o p-valor respetivo é dado por uma das fórmulas seguintes, onde T tem uma distribuição de Student $t(n_1 + n_2 - 2)$:

- $H_a : \mu_1 > \mu_2$, $\quad \mathrm{P}(T \geq t)$;
- $H_a : \mu_1 < \mu_2$, $\quad \mathrm{P}(T \leq t)$;
- $H_a : \mu_1 \neq \mu_2$, $\quad 2\mathrm{P}(T \geq |t|)$.

Estes p-valores são exatos se as observações são normais, e são aproximados nos outros casos quando n_1 e n_2 são grandes.

Se as populações não são normais mas os tamanhos das amostras são iguais, isto é, $n_1 = n_2$, ou aproximadamente iguais, verifica-se que os intervalos de confiança e os testes anteriores são robustos contra a não verificação da hipótese de normalidade. Se as duas populações têm formas semelhantes tal acontece deste que as amostras tenham

dimensões pelo menos iguais a 5. Quando as duas populações têm formas distintas, é necessário usar amostras de dimensões mais elevadas. Neste caso aconselhe-se o uso das regras dadas em §8.4 com $n_1 + n_2$ no lugar de n (Moore e McCabe, 2006, p. 493).

Quando planeamos um estudo deste género é importante que as amostras tenham iguais dimensões, ou pelo menos, tenham dimensões semelhantes, uma vez que tal aumenta a **robustez** do procedimento estatístico. Quando $n_1 = n_2$ reparemos ainda que as estatísticas de teste anteriores são iguais.

Exemplo 9.2.3 Retomemos os dados do Exemplo 1.2.3 (pág. 17), e comparemos as duas farinhas através da comparação dos aumentos médios dos pesos verificados nos pintos de ambos os grupos. Representando por μ_1 e μ_2 os aumentos médios dos pesos dos pintos do grupo de controlo e do grupo experimental, respetivamente, pretendemos testar

$$H_0 : \mu_1 = \mu_2 \quad \text{contra} \quad H_a : \mu_1 < \mu_2$$

(pois esperamos que a nova farinha seja melhor que a antiga).

Descriptives

	grupo		Statistic
aumento de peso	controlo	Mean	366,65
		Variance	2577,713
		Std. Deviation	50,771
	experimental	Mean	403,20
		Variance	1885,221
		Std. Deviation	43,419

Não havendo razões para pensar que as variâncias populacionais respectivas sejam iguais, vamos usar o primeiro dos testes anteriores. Sendo as duas amostras de dimensão 20, temos

$$t = \frac{366{,}65 - 403{,}20}{\sqrt{2577{,}713/20 + 1885{,}221/20}} \approx -2{,}447$$

sendo o p-valor respetivo dado por

$$p\text{-valor} = \mathrm{P}(T \leq -2{,}447),$$

onde T tem uma distribuição de Student $t(37{,}107)$, uma vez que da fórmula (9.2.2) tiramos que $k = 37{,}107$. Usando a Tabela D da distribuição de Student e a simetria da distribuição não podemos fazer muito mais do que dizer que

$$p\text{-valor} = \mathrm{P}(T \geq 2{,}447) < 0{,}02,$$

sendo o resultado obtido significativo ao nível 0,02.

Atendendo aos cálculos envolvidos na determinação do número de graus de liberdade k, usaremos habitualmente o quadro seguinte produzida pelo SPSS para executar o teste anterior:

Independent Samples Test

		t-test for Equality of Means					
						95% Confidence Interval of the Difference	
		t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
aumento	Equal variances assumed	-2,447	38	,01915	-36,550	-66,791	-6,309
	Equal variances not assumed	-2,447	37,107	,01927	-36,550	-66,815	-6,285

O p-valor que surge no quadro é o relativo à hipótese alternativa bilateral, o que significa que

$$\mathrm{P}(|T| \geq 2{,}447) = \mathrm{P}(T \leq -2{,}447) + \mathrm{P}(T \geq 2{,}447) = 0{,}01927.$$

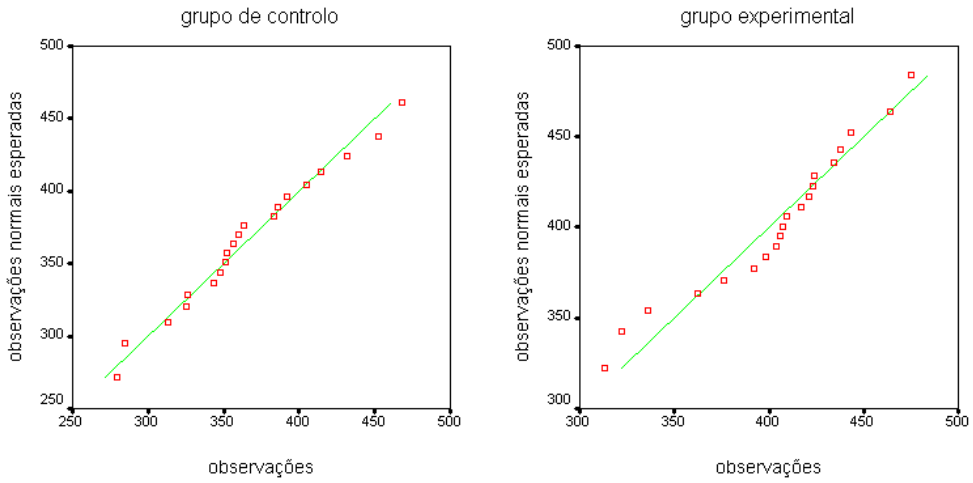
Pela simetria da distribuição de Student concluímos que

$$p\text{-valor} = \mathrm{P}(T \geq 2{,}447) = 0{,}01927/2 = 0{,}009635,$$

revelando este p-valor uma forte evidência contra a hipótese nula.

Reparemos que se pudéssemos admitir que as variâncias populacionais eram iguais, a aplicação do segundo dos testes anteriores conduziria ao mesmo valor para a estatística de teste (pois as dimensões das duas amostras são iguais), apenas se alterando o número de graus de liberdade da distribuição de Student, que passaria a $20 + 20 - 2 = 38$.

Finalmente notemos que apesar dos gráficos seguintes revelarem desvios ligeiros relativamente à hipótese de normalidade no caso do grupo experimental, como as dimensões das duas amostras são iguais, os p-valores calculados podem ser considerados fidedignos.



Mais interessante do que o teste da hipótese de igualdade de médias, que nos permitiu concluir haver evidência de que a nova farinha é melhor que a usual, é estimar o aumento médio $\mu_2 - \mu_1$. Do quadro anterior podemos obter o intervalo de confiança de nível 95% para a diferença $\mu_2 - \mu_1$:

$$]6{,}285\,;\ 66{,}815\,[.$$

Ou seja, o intervalo terá por extremidades

$$36{,}550 \pm 30{,}265.$$

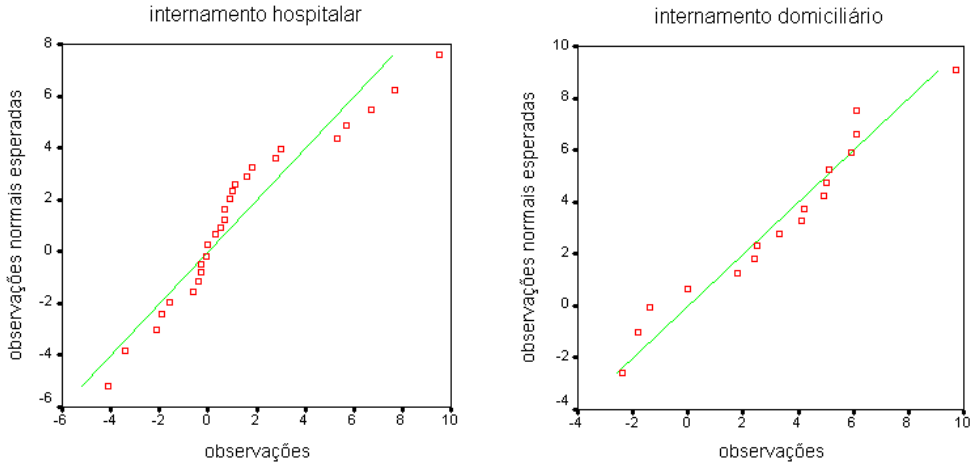
Exemplo 9.2.4 Os gráficos de extremos-e-quartis apresentados no Exemplo 1.3.20 (pág. 46), vieram em apoio da ideia, formulada *a priori*, de que o tratamento de jovens anoréxicas descrito no Exemplo 0.2.1 (pág. 3), poderia ser mais eficaz em internamento domiciliário do que hospitalar. Pretendendo confirmar, ou não, esta ideia, propomo-nos comparar as médias dos aumentos de pesos (peso final-peso inicial) de ambos os grupos. Denotando por μ_1 e μ_2, respetivamente, as médias dos aumentos dos pesos das jovens em internamento hospitalar e domiciliário, pretendemos testar a hipótese

$$H_0 : \mu_1 = \mu_2 \quad \text{contra} \quad H_1 : \mu_1 < \mu_2.$$

Resumos numéricos dos resultados obtidos e gráficos de quantis-normais de cada um dos grupos são apresentados a seguir:

Descriptives

	local		Statistic
aumento de peso	Hospital	Mean	1,2138
		Variance	9,979
		Std. Deviation	3,15897
	Família	Mean	3,2647
		Variance	10,544
		Std. Deviation	3,24710



Apesar de nada sabermos sobre a forma como as jovens foram divididas pelos dois grupos de tratamento, vamos admitir que a afectação a cada um dos grupo foi feita por métodos aleatórios. Tendo em conta o que dissemos atrás, teria sido melhor planear a experiência de modo que as dimensões dos dois grupos fossem semelhantes. Tal não acontece neste caso, o que pode implicar menor precisão no cálculo dos p-valores. Esta observação é reforçada pelos gráficos anteriores que revelam desvios relativamente à hipótese de normalidade. Recordemos que já tinhamos visto que a distribuição das diferenças dos pesos para as jovens em internamento hospitalar era positivamente assimétrica.

Sendo de 17 e 29 as dimensões das amostras consideradas em tratamento domiciliário e hospitalar, respetivamente, do quadro anterior obtemos,

$$t = \frac{1{,}214 - 3{,}265}{\sqrt{9{,}979/29 + 10{,}544/17}} \approx -2{,}089$$

sendo o p-valor respetivo dado por

$$p\text{-valor} = \mathrm{P}(T \leq -2{,}089) = \mathrm{P}(T \geq 2{,}089)$$

onde T tem uma distribuição de Student $t(32{,}893)$. Usando a tabela da distribuição de

Student obtemos

$$p\text{-valor} < 0{,}05,$$

sendo assim o resultado obtido significativo ao nível 0,05.

Usando o quadro seguinte produzido pelo SPSS concluímos que o p-valor para o teste de hipótese alternativa unilateral é $0{,}0446/2 = 0{,}0223$. Atendendo às observações anteriores sobre a precisão do p-valor calculado, devemos ser cautelosos na aceitação destes resultados como indicador claro de que o tratamento tem melhores resultados em regime domiciliário.

Independent Samples Test

		t-test for Equality of Means					
						95% Confidence Interval of the Difference	
		t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
aumento de peso (Kg)	Equal variances assumed	-2,104	44	,04114	-2,05091	-4,01553	-,08630
	Equal variances not assumed	-2,089	32,89	,04457	-2,05091	-4,04905	-,05277

Do quadro anterior podemos também obter o intervalo de confiança de nível 95% para a diferença $\mu_1 - \mu_2$:

$$]-4,015527\,;\ -0,052774\,[.$$

Também no caso do intervalo de confiança, devemos colocar a possibilidade de que a probabilidade dos intervalos de confiança produzidos pelo método anterior possa ser inferior a 95%.

Havendo razões para admitir que as variâncias populacionais são iguais, a aplicação do teste respetivo conduz a um valor para a estatística de teste diferente do anterior, pois neste caso as amostras recolhidas em ambos os grupos têm dimensões diferentes. No entanto os p-valores associados a ambos os testes são semelhantes.

9.2.2 Comparação de médias em amostras emparelhadas

A situação mais simples de comparação das médias μ_1 e μ_2 de duas populações, ocorre quando as amostras seleccionadas das duas populações são **emparelhadas**. Esta situação é comum na comparação de dois tratamentos, quando é possível aplicar ambos os tratamentos num mesmo indivíduo ou em dois indivíduos (emparelhados) que sejam semelhantes relativamente a variáveis influentes na variável resposta.

Denotando por $\bar{x}_1$ e $\bar{x}_2$ as médias relativas a cada uma das amostras, será natural basearmos o intervalo de confiança para a diferença $\mu_1 - \mu_2$ ou o teste da hipótese

$H_0 : \mu_1 = \mu_2$ na estatística

$$\bar{x}_1 - \bar{x}_2.$$

Sendo emparelhadas as amostras de tamanho n seleccionadas em cada uma das populações, não são independentes as médias amostrais $\bar{x}_1$ e $\bar{x}_2$ relativas a cada uma das amostras, independência essa que foi essencial na avaliação da variância da diferença $\bar{x}_1 - \bar{x}_2$. Para contornar este problema, basta notar que a diferença $\bar{x}_1 - \bar{x}_2$ não é mais do que a média das diferenças $x_{1i} - x_{2i}$, onde x_{1i} e x_{2i} representam os valores da variável em estudo que foram observados no i-ésimo indivíduo da amostra (ou de ambas as amostras caso os indivíduos estejam emparelhados). Assim sendo, a variância de $\bar{x}_1 - \bar{x}_2$ pode ser estimada pela variância amostral dessas diferenças que vamos vamos denotar por $s^2_{x_1 - x_2}$.

A construção dum intervalo de para a diferença $\mu_1 - \mu_2$ será assim baseado na variável fulcral

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_{x1-x2}/\sqrt{n}}$$

Intervalo de confiança para a diferença de duas médias (amostras emparelhadas):

Um intervalo de confiança para a diferença de médias $\mu_1 - \mu_2$, com nível de confiança C, tem por extremidades

$$\bar{x}_1 - \bar{x}_2 \pm t^* s_{x_1 - x_2}/\sqrt{n},$$

onde t^* é tal que

$$\mathrm{P}(-t^* \leq T \leq t^*) = C$$

e T tem uma distribuição de student $t(n-1)$. Este intervalo de confiança é exato quando as diferenças $x_{1i} - x_{2i}$ são normais, e é aproximado quando n é grande.

e o teste da hipótese de igualdade das duas médias será baseado na estatística de teste

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{x1-x2}/\sqrt{n}}.$$

Teste de comparação de médias em amostras emparelhadas:

Para testar a hipótese $H_0 : \mu_1 = \mu_2$ (resp. $H_0 : \mu_1 \leq \mu_2$, $H_0 : \mu_1 \geq \mu_2$), de igualdade das médias a partir de amostras emparelhadas, use as observações para calcular

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{x_1 - x_2}/\sqrt{n}},$$

e, de acordo com a hipótese alternativa H_a em causa, o p-valor respetivo é dado por uma das fórmulas seguintes, onde T tem uma distribuição de Student $t(n-1)$:

- $H_a : \mu_1 > \mu_2$, $\quad \mathrm{P}(T \geq t)$;
- $H_a : \mu_1 < \mu_2$, $\quad \mathrm{P}(T \leq t)$;
- $H_a : \mu_1 \neq \mu_2$, $\quad 2\mathrm{P}(T \geq |t|)$.

Estes p-valores são exatos as diferenças $x_{1i} - x_{2i}$ são normais, e são aproximados nos outros casos quando é n grande.

No caso das diferenças $x_{1i} - x_{2i}$ não serem normais, e de forma a aumentar a robustez do intervalo de confiança anterior, devem ser seguidas as orientações dadas em §7.4 sobre o intervalo de confiança para uma média quando a variância populacional é desconhecida.

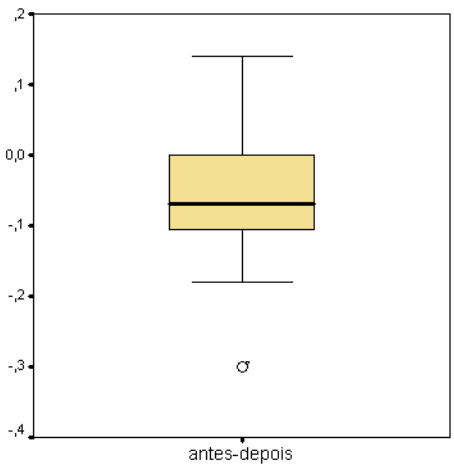
No caso das diferenças $x_{1i} - x_{2i}$ não serem normais, as fórmulas anteriores para o cálculo dos p-valores devem ser utilizadas com cuidado. De forma a aumentar a robustez do procedimento, devem ser seguidas as recomendações feitas em §8.4 a propósito do teste de Student para uma média com variância desconhecida.

Exemplo 9.2.1 Uma empresa farmacêutica realizou uma experiência para verificar se se confirmavam as suspeitas de que determinado medicamento aumentava o tempo de reação a determinados estímulos. Se for esse o caso, essa observação deve ser incluída na literatura que acompanha o medicamento. Para tal, seleccionaram-se ao acaso 36 indivíduos de um grupo mais vasto de indivíduos que tomavam o medicamento, e registou-se o seu tempo de reação (em centésimos de segundo) a determinado estímulo, antes e depois de tomar o medicamento.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	antes	,9056	36	,12311	,02052
	depois	,9608	36	,09376	,01563

O gráfico de extremos-e-quartis para a diferença dos tempos de reação antes e depois de tomar o medicamento (antes-depois), revela a presença duma possível observação discordante. Admitamos que tal observação foi confirmada e que decidimos mantê-la como observação válida. Apesar da assimetria positiva da distribuição, reparemos no facto do seu terceiro quartil ser muito próximo de zero, o que constitui um indício forte de que o tempo de reação aumenta com a utilização do medicamento.



Pretendendo quantificar os indícios anteriores, e verificar se o que foi observado pode ser considerado natural no caso de não haver alteração do tempo de reação, vamos testar a hipótese

$$H_0 : \mu_1 = \mu_2 \quad \text{contra a hipótese} \quad H_1 : \mu_1 < \mu_2,$$

onde μ_1 e μ_2, representam os tempos médios de reação antes e depois do medicamento ser ministrado. Reparemos que a hipótese alternativa foi fixada tendo em conta o que se pretendia inicialmente testar, tendo sido fixada antes de recolhermos os dados em que baseamos o estudo.

Usando o SPSS obtemos:

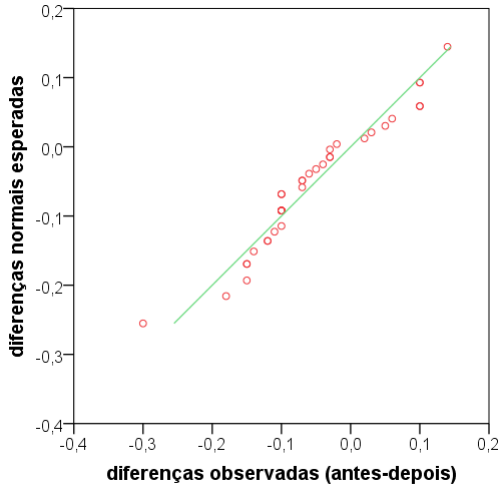
Paired Samples Test

	Paired Differences						
			95% Confidence Interval of the Difference				
	Mean	Std. Deviation	Lower	Upper	t	df	Sig. (2-tailed)
antes - depois	-,05528	,09455	-,08727	-,02329	-3,508	35	,00126

Sendo o nosso teste um teste de hipótese alternativa unilateral, o p-valor é dado por

$$p\text{-valor} = \mathrm{P}(T \leq -3{,}508) = 0{,}00126/2 = 0{,}00063,$$

o que revela forte evidência contra a hipótese nula. Atendendo ao tamanho da amostra, mesmo na presença duma distribuição ligeiramente assimétrica, este p-valor pode ser considerado fidedigno. Esta conclusão é ainda suportada pelo gráfico de quantis normais seguinte que revela que a distribuição das diferenças dos tempos de reação é aproximadamente normal.



O quadro anterior produzido pelo SPSS fornece ainda o intervalo de confiança de nível 95% para a diferença $\mu_1 - \mu_2$:

$$]-0,087269\,;\ -0,023286\,[.$$

9.3 Bibliografia

Daniel, W.W. (2009). *Biostatistics: a foundation for analysis in the health sciences*, Wiley.

Dawson, B., Trapp, R.G. (2004). *Basic & Clinical Biostatistics*, McGraw-Hill.

Moore, D.S. (1985). *Statistics: concepts and controversies*, W.H. Freeman and Company.

Moore, D.S., McCabe, G.P. (2006). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

Pestana, D.D., Velosa, S.F. (2002). *Introdução à Probabilidade e à Estatística*, Vol. I, Fundação Calouste Gulbenkian.

10

Análise de frequências: testes do qui-quadrado

Estatística do qui-quadrado. A distribuição do qui-quadrado. Testes do qui-quadrado de homogeneidade e de independência. Teste de ajustamento do qui-quadrado.

10.1 Estatística e distribuição do qui-quadrado

Neste capítulo vamos estudar procedimentos de teste para testar a independência entre duas variáveis X e Y, testar a homogeneidade da distribuição de X relativamente a diversas populações, e também o ajustamento da distribuição duma variável X a uma distribuição teórica fixa à partida. As respetivas estatísticas de teste, conhecidas por **estatísticas do qui-quadrado**, são baseadas na comparação entre as frequências observadas na amostra recolhida e as frequências esperadas caso as hipóteses nulas respetivas fossem verdadeiras. Como veremos, em todos estes casos, os p-valores associados às observações realizadas podem ser calculados usando uma nova distribuição de probabilidade, dita **distribuição do qui-quadrado**.

O exemplo seguinte motiva e permite compreender melhor esta ideia.

Exemplo 10.1.1 Nos final da década de 1940 pretendia-se estabelecer uma ligação entre o cancro do pulmão e os hábitos tabágicos. Reuniram-se dois grupos de 709 pessoas cada um. O primeiro era constituído por pessoas com cancro do pulmão, enquanto que o segundo era constituído por pessoas que sofriam de outras doenças. Os resultados observados são apresentados na **tabela de contingência** de duas entradas seguinte ([1]):

[1] Fonte: Oliveira, P.E., *Apontamentos de Estatística (Ciências Farmacêuticas)*, 2007, Coimbra.

	Com cancro	Sem cancro
Fumador	688	650
Não fumador	21	59
Total	709	709

Será que os resultados apresentados permitem concluir que não há associação entre o cancro do pulmão e os hábitos tabágicos?

Em cada uma das populações consideradas (pessoas com cancro do pulmão e sem cancro do pulmão mas com outras doenças), observámos a variável X com dois níveis (fumador e não fumador). Tomando cada uma destas populações como níveis (com cancro e sem cancro) de uma outra variável Y, a pergunta que queremos ver respondida é a de saber se os resultados observados indicam, ou não, existir uma associação entre as variáveis X e Y.

Atendendo à forma como ambas as amostras foram recolhidas, a primeira na população de pessoas com cancro do pulmão e a segunda na população de pessoas sem cancro do pulmão mas com outras doenças, testar a hipótese de independência das variáveis X e Y não é mais do que testar se a variável X se distribui de igual forma nas duas populações. Esta hipótese é habitualmente conhecida como **hipótese de homogeneidade** da distribuição de X relativamente às populações envolvidas.

Assim, representando por p_1 e p_2 as proporções de fumadores em ambas as populações, testar se X se distribui de igual forma nas duas populações é equivalente a testar

$$H_0 : p_1 = p_2 \quad \text{contra} \quad H_a : p_1 \neq p_2$$

Este problema já foi por nós estudado no Capítulo 9. Para testar a hipótese anterior lançámos mão da estatística de teste (ver §9.1.1)

$$z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

que pode ser interpretada como uma medida da compatibilidade das observações com a hipótese nula. Tal é conseguido através da comparação das proporções amostrais de fumadores nas duas amostras. Valores pequenos de $|z|$ indicam compatibilidade com H_0 enquanto que quanto maior for $|z|$ maior é evidência que os dados comportam contra H_0.

Alguns cálculos revelam que esta estatística, ou melhor, o seu quadrado, pode ter uma interpretação alternativa que nos será bastante útil ao longo deste capítulo. Com efeito, é possível mostrar que o quadrado da estatística de teste z pode ser escrito na

forma seguinte

$$\begin{aligned} z^2 &= \frac{(N_{11} - N_{1+}n_1/N)^2}{N_{1+}n_1/N} + \frac{(N_{12} - N_{1+}n_2/n)^2}{N_{1+}n_2/N} \\ &\quad + \frac{(N_{21} - N_{2+}n_1/N)^2}{N_{2+}n_1/N} + \frac{(N_{22} - N_{2+}n_2/n)^2}{N_{2+}n_2/N} \\ &= \sum \frac{(N_{ij} - N_{i+}n_j/N)^2}{N_{i+}n_j/N}, \end{aligned}$$

onde os valores N_{ij} representam as frequências observadas em cada uma das células da tabela anterior, n_j são os tamanhos das duas amostras observadas, N_{i+} são os totais de cada uma das linhas da tabela e $N = n_1 + n_2$:

	Com cancro	Sem cancro	Total
Fumador	N_{11}	N_{12}	N_{1+}
Não fumador	N_{21}	N_{22}	N_{2+}
Total	n_1	n_2	N

No caso da hipótese nula ser verdadeira, a frequência de observações que devemos esperar para a célula $(1, 1)$ da tabela (linha 1 e coluna 1 da tabela) é de

$$\begin{aligned} n_1\, p_1 &= n_1 \mathrm{P}(\text{“ser fumador”}) \quad \text{(homogeneidade)} \\ &\approx n_1 \frac{N_{1+}}{N} \\ &= \frac{N_{1+}n_1}{N} = E_{11}. \end{aligned}$$

O mesmo se passa para as outras células da tabela. No caso do hipótese H_0 ser verdadeira, a frequência que devemos esperar na célula (i, j) é assim de

$$E_{ij} = \frac{N_{i+}n_j}{N} = \frac{\text{total linha } i \times \text{total coluna } j}{N}.$$

Atendendo à expressão alternativa obtida atrás para z^2, concluímos que a estatística de teste z^2 não é mais do que uma medida da discrepância entre as frequências observadas em cada uma das células da tabela e as frequências que seriam de esperar nessas células caso a hipótese nula fosse verdadeira. Esta estatística recebe o nome de **estatística do qui-quadrado** e será representada por

$$Q^2 = \sum \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

onde o somatório indica que a soma é relativa a todas as células da tabela.

Exercício 10.1.1 (cont.) Retomando os dados da tabela de frequências, é simples verificar que $z^2 \simeq (4{,}374)^2 = 19{,}129$. Sendo as frequência observadas e os totais de linhas e colunas dados por

N_{ij}	Com cancro	Sem cancro	Total
Fumador	688	650	1338
Não fumador	21	59	80
Total	709	709	1418

as frequências esperadas E_{ij} são dadas por

E_{ij}	Com cancro	Sem cancro
Fumador	669	669
Não fumador	40	40

A estatística do qui-quadrado é então dada por

$$Q^2 = \frac{(688-669)^2}{669} + \frac{(650-669)^2}{669} + \frac{(21-40)^2}{40} + \frac{(59-40)^2}{40} \approx 19{,}129$$

o que, como tínhamos referido, coincide com o valor obtido para z^2.

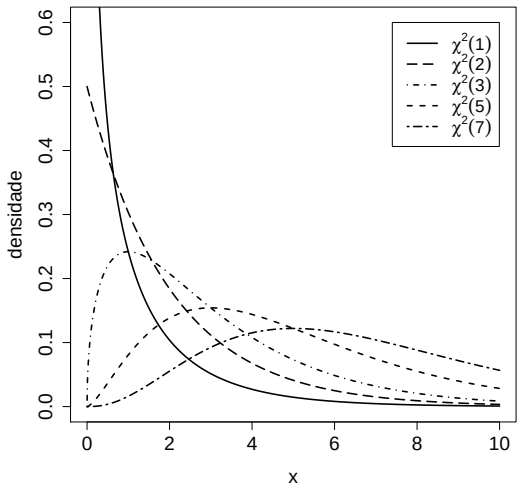
Quando usávamos o teste de comparação de duas proporções, após o cálculo da estatística z recorríamos à tabela da distribuição normal standard Z para calcularmos uma aproximação para o p-valor associado às observações realizadas:

$$p\text{-valor} = 2\mathrm{P}(Z \geq |z|) = 2\mathrm{P}(Z \geq 4{,}374) < 2 \times 0{,}0002 = 0{,}0004.$$

Como devemos agora calcular o p-valor associado à estatística Q^2? Claro que podemos continuar a usar a tabela da distribuição normal standard, mas podemos também usar a tabela da **distribuição do qui-quadrado** com um grau de liberdade e representa-mo-la por $\chi^2(1)$, distribuição esta que não é mais do que a distribuição do quadrado Z^2 da distribuição normal standard. Para tal basta ter em conta que o p-valor anterior pode ser escrito em termos de Q^2 e do quadrado da variável normal standard:

$$p\text{-valor} = \mathrm{P}(Z \leq -|z|) + \mathrm{P}(Z \geq |z|) = \mathrm{P}(Z^2 \geq |z|^2) = \mathrm{P}(Z^2 \geq Q^2).$$

A distribuição do qui-quadrado com um grau de liberdade, $\chi^2(1)$, pertence a uma família mais vasta de distribuições, que tal como a das distribuições de Student depende dum parâmetro designado por **grau de liberdade**. De uma forma geral, se $Z_1, \ldots, Z_k$ são variáveis independentes com distribuições normais standard, a soma dos seus quadrados, $Z_1^2 + \ldots + Z_k^2$, possui uma distribuição do **qui-quadrado com k graus de liberdade**, que representamos por $\chi^2(k)$. Todas estas distribuições são assimétricas positivas e a suas densidades de probabilidade têm a forma seguinte:



O cálculo de áreas sob uma curva densidade do qui-quadrado pode ser feito utilizando a Tabela E onde estão tabeladas algumas dessas áreas para vários graus de liberdade.

Voltando ao nosso exemplo e tendo então em conta que $Z^2 \sim \chi^2(1)$, a partir da primeira linha da Tabela E concluímos que

$$p\text{-valor} = \mathrm{P}(Z^2 \geq Q^2) = \mathrm{P}(\chi^2(1) \geq 19{,}129) < 0{,}001.$$

Usando a Tabela E não podemos ser mais precisos relativamente ao cálculo do p-valor. O valor da probabilidade anterior surge no quadro seguinte produzido pelo SPSS, que nos dá também conta do valor da estatística de teste. O quadro apresenta ainda o valor da estatística de teste com uma correção de continuidade, dita correcção de Yates, com o objetivo de melhorar a aproximação da distribuição da estatística de teste pela distribuição $\chi^2(1)$.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	19,129[a]	1	,0000122
Continuity Correction[b]	18,136	1	,0000206
N of Valid Cases	1418		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 40,00.

b. Computed only for a 2x2 table

Como veremos nos parágrafos seguintes, a estatística Q^2 é particularmente útil

quando a variável X tem mais do que dois níveis, ou quando há mais do que duas populações envolvidas. Este assunto será discutido em §10.2 e a este tipo de teste chamamos **teste de homogeneidade do qui-quadrado**. A estatística Q^2 será também muito útil quando a recolha da amostra for feita não em populações distintas mas de uma só população em que para cada sujeito observado são registadas as duas variáveis X e Y, a primeira com r níveis e a segunda com s níveis. Neste caso, pretendendo-se testar a independência entre as duas variáveis observadas, o teste baseado em Q^2 recebe o nome de **teste de independência do qui-quadrado** e será abordado em §10.3.

10.2 Teste de homogeneidade do qui-quadrado

Vamos neste parágrafo generalizar a estatística do qui-quadrado ao caso em que uma variável X que tem r níveis diferentes, é observada em s populações. O nosso objectivo é testar a hipótese de homogeneidade da distribuição de X relativamente às populações consideradas, isto é, pretendemos testar

H_0: A distribuição de X não depende da população

contra a hipótese alternativa

H_a: A distribuição de X depende da população

Como já referimos, quando as s populações são níveis de uma variável Y as hipóteses anteriores são equivalentes a H_0: X e Y são independentes e H_a: X e Y não são independentes.

As observações dão agora origem a uma tabela de contingência com $r \times s$ células onde N_{ij} é o número de observações na população j com $X = i$, n_j é o tamanho da amostra recolhida na população j, e N_{i+} representa os total da linha i da tabela:

	Populações				
X	1	2	...	s	Total
1	N_{11}	N_{12}	...	N_{1s}	N_{1+}
2	N_{21}	N_{22}	...	N_{2s}	N_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	N_{r1}	N_{r2}	...	N_{rs}	N_{r+}
Total	n_1	n_2	...	n_s	N

A estatística do qui-quadrado, que mede a discrepância entre as frequências observadas e as frequências esperadas em cada uma das células da tabela de contingência, é

dada por

$$Q^2 = \sum \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

onde a soma é relativa a todas as células da tabela e as frequências esperadas são, como vimos, dadas por

$$E_{ij} = \frac{N_{i+}n_j}{N} = \frac{\text{total linha } i \times \text{total coluna } j}{N}.$$

Valores grandes de Q^2 conduzem à rejeição da hipótese nula. Após o cálculo da estatística Q^2 o p-valor associado às observações pode ser aproximado usando a distribuição χ^2 com $(r-1)\times(s-1)$ graus de liberdade, uma vez que se pode mostrar que quando o tamanho das amostras é grande a distribuição amostral de Q^2 pode ser aproximada por uma tal distribuição.

Teste de homogeneidade do qui-quadrado:

A partir da tabela de contingência calcule a estatística do qui-quadrado Q^2 e obtenha o p-valor (aproximado) associado às observações realizadas através da fórmula

$$\mathrm{P}(\chi^2 \geq Q^2),$$

onde χ^2 possui uma distribuição do qui-quadrado com $(r-1)\times(s-1)$ graus de liberdade.

Para tabelas de contingência 2×2 a aproximação é considerada boa se para todas as frequências esperadas são superiores ou iguais a 5, isto é, $E_{ij} \geq 5$. Para as restantes tabelas considera-se que a aproximação é boa se a média das frequências esperadas é igual ou superior a 5 e se para todas as frequências esperadas se tem $E_{ij} \geq 1$.

Como já referimos em §10.1, no caso duma tabela 2×2, caso em que a variável X é dicotómica e temos apenas duas populações, testar a homogeneidade da distribuição de X relativamente às duas populações é equivalente a testar

$$H_0 : \ p_1 = p_2 \quad \text{contra} \quad H_a : \ p_1 \neq p_2,$$

onde p_1 e p_2 representam as probabilidades de sucesso em ambas as populações. Além disso, o teste do qui-quadrado é equivalente ao teste z de igualdade de duas proporções.

No caso geral em que temos mais que duas populações envolvidas, e em que as probabilidades de sucesso nas diversas populações são representadas por $p_1, \ldots, p_s$ (s

populações envolvidas), testar a hipótese de homogeneidade da distribuição de X relativamente às diversas populações é testar a hipótese de igualdade das diversas proporções (probabilidades)

$$H_0 : \ p_1 = \ldots = p_s$$

contra a hipótese alternativa

$$H_a : \ p_i \neq p_j, \ \text{ para algum par } i, j.$$

Exemplo 10.2.1 A tabela de contingência seguinte corresponde a um estudo feito a partir de 353 amostras de água do mar classificadas segundo dois factores: distância à costa a que foram recolhidas e nível de mercúrio detectado. Para cada uma das distâncias consideradas, foram recolhidas e analisadas amostras com tamanhos semelhantes ([2]):

	Distância à costa			
Níveis de exposição	Menos de 5 km	Entre 5 km e 15 km	Mais de 15 km	Total
Irrelevante	23	29	32	84
Sem perigosidade	47	44	45	136
Perigoso	53	41	39	133
Total	123	114	116	353

Face a estes dados será que podemos concluir que os níveis de mercúrio na água dependem da proximidade da costa (ao nível de significância 0,05)?

Depois de algum trabalho de cálculo verificamos que $Q^2 \approx 3{,}729$. Usando agora a distribuição do qui-quadrado com $(3-1) \times (3-1) = 4$ graus de liberdade concluímos da Tabela E que

$$p\text{-valor} = P(\chi^2(4) \geq 3{,}729) > 0{,}2.$$

O resultado obtido não é significativo ao nível 0,05 o que nos leva a aceitar a hipótese nula de independência entre proximidade à costa e os níveis de mercúrio presentes na água. De forma equivalente, podemos também dizer que a distribuição do nível de mercúrio é análoga para cada uma das distâncias consideradas (populações).

Uma melhor aproximação para a probabilidade anterior é dada no quadro seguinte:

[2]Fonte: Mendes, M.G.T., *Notas de Estatística (Mestrado Integrado em Ciências Farmacêuticas)*, 2008, Coimbra.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	3,729[a]	4	,444
N of Valid Cases	353		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 27,13.

Apesar de não nos ser dada qualquer informação sobre a forma como são medidas as duas variáveis envolvidas, é natural pensar que quer o nível de mercúrio quer a distância à costa possam ter sido inicialmente registados numa escala contínua. Quando tal acontece, é necessário proceder a uma categorização das variáveis envolvidas para posteriormente se usar o teste do qui-quadrado.

10.3 Teste de independência do qui-quadrado

Vamos supor que n observações são realizadas por amostragem aleatória simples numa população e classificadas segundo dois factores X e Y. Admitamos que X e Y têm r e s níveis, respetivamente, que representamos por $1, 2, \ldots, r$ e $1, 2, \ldots, s$. As n observações dão origem a uma tabela de contingência de duas entradas onde N_{ij} é o número de observações com $X = i$ e $Y = j$:

X/Y	1	2	$\ldots$	s	Total
1	N_{11}	N_{12}	$\ldots$	N_{1s}	N_{1+}
2	N_{21}	N_{22}	$\ldots$	N_{2s}	N_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	N_{r1}	N_{r2}	$\ldots$	N_{rs}	N_{r+}
Total	N_{+1}	N_{+2}	$\ldots$	N_{+s}	n

A partir dos dados anteriores pretendemos testar as hipóteses

H_0: As variáveis X e Y são independentes

contra a hipótese alternativa

H_a: As variáveis X e Y não são independentes

Mais uma vez, a ideia subjacente ao teste do qui-quadrado é a da comparação entre as frequências observadas (N_{ij}) em cada uma das células da tabela de contingência e as frequências esperadas (E_{ij}) na hipótese das variáveis X e Y serem independentes. Tal comparação é feita através da **estatística do qui-quadrado** Q^2 definida no parágrafo

anterior, onde a frequência esperada na célula (i, j) é agora dada por:

$$\begin{aligned} n\mathrm{P}(X=i, Y=j) &= n\,\mathrm{P}(X=i)\,\mathrm{P}(Y=j) \quad \text{(independência)} \\ &\approx n\,\frac{N_{i+}}{n}\,\frac{N_{+j}}{n} \\ &= \frac{N_{i+}N_{+j}}{n} \\ &= E_{ij}. \end{aligned}$$

Tal como no teste de homogeneidade, a frequência esperada é dada pela fórmula

$$E_{ij} = \frac{\text{total linha } i \times \text{total coluna } j}{n}.$$

A estatística do qui-quadrado para o teste de independência é assim obtida a partir da tabela de contingência da mesma forma que a estatística do qui-quadrado para o teste de homogeneidade. Também a distribuição a usar no cálculo do p-valor é a distribuição do qui-quadrado com $(r-1)\times(s-1)$ graus de liberdade.

Teste de independência do qui-quadrado:

A partir da tabela de contingência calcule a estatística do qui-quadrado Q^2 e obtenha o p-valor (aproximado) associado às observações realizadas através da fórmula

$$\mathrm{P}(\chi^2 \geq Q^2),$$

onde χ^2 possui uma distribuição do qui-quadrado com $(r-1)\times(s-1)$ graus de liberdade.

Os comentários feitos no caso do teste de homogeneidade sobre a qualidade da aproximação da distribuição amostral da estatística de teste pela distribuição do qui-quadrado, valem também aqui.

Exemplo 10.3.1 A tabela seguinte apresenta os resultados de um estudo de amostragem no qual participaram 300 adultos residindo numa determinada área metropolitana ([3]). A cada um foi pedido para indicarem, entre três possíveis políticas, qual a que ele preferiam relativamente a fumar em locais públicos .

[3] Fonte: Daniel, 2009, pp. 622–623.

Nível de Educação	Política preferida: Sem restrições	Fumar apenas em certas áreas	Interdito fumar	Sem opinião	Total
Licenciado	5	44	23	3	75
3.º Ciclo	15	100	30	5	150
2.º Ciclo	15	40	10	10	75
Total	35	184	63	18	300

Será que a partir destes dados podemos concluir que existe uma relação entre o nível de educação e a atitude perante o fumar em locais públicos?

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	22,502[a]	6	,00098
N of Valid Cases	300		

a. 2 cells (16,7%) have expected count less than 5. The minimum expected count is 4,50.

Atendendo a que o p-valor = 0,00098 é muito pequeno, concluímos que os dados revelam uma forte evidência da existência de uma associação entre o nível de educação e a atitude perante o fumar em locais públicos. Atendendo a que todas as frequências esperadas são superiores ou iguais a 1 e a sua média é superior ou igual a 5, o p-valor anterior é credível.

Exemplo 10.3.2 Retomemos agora os dados considerados no Exemplo 9.1.1 (pág. 212), e suponhamos que estamos interessados em testar a independência entre as variáveis X e Y, onde X nos dá o estado do indivíduo antes de tomar o medicamento (doente ou não-doente) e Y o estado do mesmo indivíduo depois de tomar o medicamento (doente ou não-doente).

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	,714[a]	1	,398
Continuity Correction[b]	,516	1	,473
N of Valid Cases	314		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 36,33.

b. Computed only for a 2x2 table

Atendendo ao p-valor = 0,473, somos levados a concluir que as observações são compatíveis com a hipótese nula das variáveis X e Y serem independentes. Isto não significa que o medicamento não tenha um efeito positivo no tratamento da doença. Atendendo ao conceito de independência, apenas quer dizer que os dados não mostram evidência de que a probabilidade dum indivíduo ser não doente depois de tomar o medicamento não se altera se soubermos que ele era classificado como doente ou como não doente antes de tomar o medicamento. Se o queremos é verificar se o medicamento tem um efeito positivo no tratamento da doença, devemos executar o teste de McNemar de comparação de proporções, como fizemos no Exemplo 9.1.1.

10.4 Teste de ajustamento do qui-quadrado

Como vimos nos parágrafos anteriores, os testes do qui-quadrado são procedimentos estatísticos que são usados para avaliar a discrepância entre as frequências observadas e as frequências teóricas esperadas no caso da hipótese nula ser verdadeira. A mesma ideia pode ser adaptada para desenvolver procedimentos formais para testar se determinado conjunto de dados possui uma determinada distribuição fixada à partida, como a distribuição normal. O procedimento passará por dividir as observações em classes, contar as frequências absolutas observadas nessas classes, e finalmente comparar, através de uma estatística de tipo qui-quadrado, tais frequências com as frequências esperadas no caso da distribuição normal com média e desvio-padrão iguais à média e desvio-padrão amostrais.

Neste parágrafo vamos limitar-nos à situação em que, na presença de n observações de uma variável X que possui r níveis diferentes $1, 2, \ldots, r$, pretendemos testar a hipótese das probabilidades de X tomar cada um dos valores $1, 2, \ldots, r$, serem iguais a valores $p_1, p_2, \ldots, p_r$, com $p_1 + p_2 + \ldots + p_r = 1$, conhecidos à partida. As frequências de cada um dos níveis é registado numa tabela de frequências

	Níveis de X				
	1	2	...	r	Total
Frequências observadas	N_1	N_2	...	N_r	n

e pretendemos testar a hipótese

$$H_0 \ : \ \mathrm{P}(X = i) = p_i \text{ para todo o } i \in \{1, \ldots, r\},$$

contra a hipótese alternativa

$$H_a \ : \ \mathrm{P}(X = i) \neq p_i \text{ para algum } i \in \{1, \ldots, r\}.$$

Dizemos assim que temos um teste de ajustamento à distribuição $p_1, \ldots, p_r$.

Como referimos atrás, o teste do qui-quadrado para testar a hipótese anterior tem por base a comparação entre as frequências observadas e as frequências esperadas sob a hipótese nula:

	Níveis de X				
	1	2	...	r	Total
Frequências observadas	N_1	N_2	...	N_r	n
Frequências esperadas	np_1	np_2	...	np_r	

A comparação entre as frequências observadas e as esperadas é feito através da estatística do qui-quadrado

$$Q^2 = \sum \frac{(N_i - np_i)^2}{np_i},$$

em que valores grandes de Q^2 dão evidência contra a hipótese nula.

Teste de ajustamento do qui-quadrado:

A partir da tabela de frequências calcule a estatística do qui-quadrado Q^2 e obtenha o p-valor (aproximado) associado às observações realizadas através da fórmula

$$\mathrm{P}(\chi^2 \geq Q^2),$$

onde χ^2 possui uma distribuição do qui-quadrado com $r-1$ graus de liberdade.

Exemplo 10.4.1 Nas experiências que Mendel realizou com ervilhas observou 315 redondas e amarelas, 108 redondas e verdes, 101 enrugadas e amarelas e 32 enrugadas e verdes. De acordo com a sua teoria de hereditariedade, estes números deveriam estar na proporção 9:3:3:1. Com um nível de significância de 0,02, poderemos afirmar que estes dados dão evidência estatística à teoria de Mendel?

Representando por X a variável que nos indica o tipos de ervilha observados, esta variável tem quatro níveis, que vamos representar por $1, 2, 3$ e 4, onde $1 =$"redonda e amarela", $2 =$"redonda e verde", $3 =$"enrugada e amarela" e $4 =$"enrugada e verde".

	Níveis de X				
	1	2	3	4	Total
Freq. observadas (N_i)	315	108	101	32	556

Pretendemos testar a hipótese nula

$$H_0 \ : \ \mathrm{P}(X=1)=9/16,\ \mathrm{P}(X=2)=3/16,\ \mathrm{P}(X=3)=3/16,\ \mathrm{P}(X=4)=1/16,$$

contra a hipótese alternativa

$$H_a : \mathrm{P}(X=1)\neq 9/16 \ \text{ ou } \ \mathrm{P}(X=2)\neq 3/16 \ \text{ ou } \ \mathrm{P}(X=3)\neq 3/16 \ \text{ ou } \ \mathrm{P}(X=4)\neq 1/16.$$

As frequência esperadas são dadas por

	Níveis de X				
	R-A	R-V	E-A	E-V	Total
Freq. observadas (N_i)	315	108	101	32	556
Freq. esperadas (np_i)	312,75	104,25	104,25	34,75	

e a estatística do qui-quadrado é dada por

$$Q^2=\sum \frac{(N_i-np_i)^2}{np_i}=0{,}0162+0{,}135+0{,}101+0{,}218=0{,}470.$$

Usando a Tabele E, concluímos que o p-valor é tal que

$$p\text{-valor} \approx \mathrm{P}(\chi^2(3)\geq 0{,}470) > 0{,}2.$$

Como p-valor $> \alpha = 0{,}02$, concluímos que, ao nível 0,02, os resultados observados não revelam evidência contra a teoria de Mendel.

Os cálculos anteriores podem ser feitos facilmente com o SPSS que produz quadros seguintes com as frequências esperadas e o cálculo da estatística do qui-quadrado e do respetivo p-valor.

ervilha

	Observed N	Expected N	Residual
redonda e amarela	315	312,75	2,25
redonda e verde	108	104,25	3,75
enrugada e amarela	101	104,25	-3,25
enrugada e verde	32	34,75	-2,75
Total	556		

Test Statistics

	ervilha
Chi-Square	,470[a]
df	3
Asymp. Sig.	,925

a. 0 cells (0,0%) have expected frequencies less than 5. The minimum expected cell frequency is 34,8.

10.5 Bibliografia

Daniel, W.W. (2009). *Biostatistics: a foundation for analysis in the health sciences*, Wiley.

Moore, D.S., McCabe, G.P. (2006). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

11

Análise da variância: ANOVA

A análise da variância ou ANOVA. Dois estimadores da variância. Distribuições amostrais de MQD e MQE. O teste ANOVA. Comparações múltiplas. Os métodos de Bonferroni, Tukey e Dunnet.

11.1 Introdução

Neste capítulo estudamos o problema da comparação das médias relativas a dois ou mais grupos de indivíduos a que chamamos populações. Um problema semelhante, mas relativo apenas a duas populações foi estudado na Secção 9.2.1. Admitindo que temos $p \geq 2$ populações com médias $\mu_1, \ldots, \mu_p$, e amostras de tamanhos $n_1, \ldots, n_p$ retiradas de cada uma das populações, o nosso objetivo é testar a hipótese de todas as populações terem a mesma média, isto é,

$$H_0: \quad \mu_1 = \ldots = \mu_p$$

contra a hipótese alternativa de haver pelo menos duas populações com médias diferentes, ou seja,

$$H_a: \quad \mu_i \neq \mu_j, \text{ para algum par } i, j = 1, \ldots, p, i \neq j.$$

Exemplo 11.1.1 Para ilustrar esta situação, consideramos um estudo sobre a compreensão da leitura em crianças em que se comparam três métodos de ensino ([1]). 66 crianças foram divididas aleatoriamente em três grupos com 22 crianças cada, em cada um dos quais é usado um dos métodos de ensino. Para uma das variáveis quantitativas observadas, que mede um dos tipos de compreensão da leitura, pretendemos saber se há diferenças entre as três estratégias de ensino usadas. Mais precisamente, estamos interessados em saber se as três populações têm ou não a mesma média. Será natural

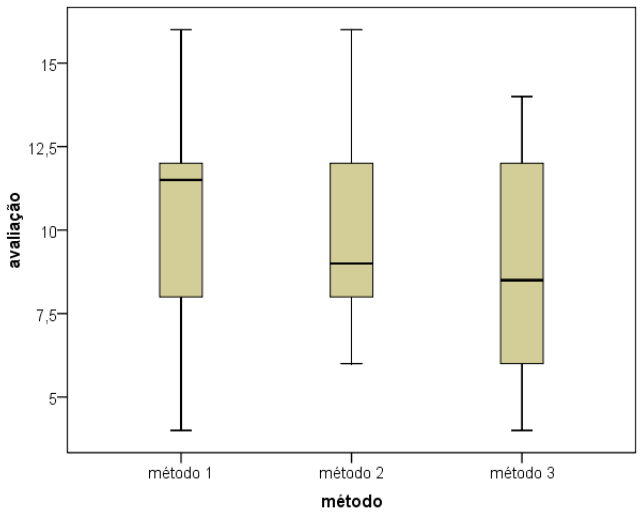
[1] Fonte: Moore e Mccabe, 2003, p. 756.

responder a esta questão comparando as médias amostrais correspondentes. As médias e desvios-padrão observados em cada um dos grupos são dados na tabela seguinte:

	método 1	método 2	método 3
média	10,50	9,73	9,14
desvio-padrão	2,97	2,69	3,34

Mesmo que as médias das três populações sejam iguais, não será de esperar que as médias amostrais sejam iguais. A questão é saber se as diferenças observadas entre as diversas médias amostrais são estatisticamente significativas, isto é, se tais diferenças indiciam que pelo menos duas das médias populacionais sejam distintas.

O gráfico de extremos-e-quartis seguinte dá-nos uma ideia clara sobre as distribuição dos resultados observados em cada um dos grupos, revelando, em particular, uma elevada variabilidade em cada um dos grupos.



No que se segue, vamos assumir que temos observações independentes provenientes de $p \geq 2$ populações com médias $\mu_1, \ldots, \mu_p$ e com iguais variâncias $\sigma_1^2 = \ldots = \sigma_p^2 = \sigma^2$. Além disso, vamos supor que as diferentes populações são normalmente distribuídas. Representando por x_{ij} a j-ésima observação da i-ésima população, podemos dizer que

$$x_{ij} = \mu_i + \epsilon_{ij}$$

para $i = 1, \ldots, p$ e $j = 1, \ldots, n_i$, onde n_i é o número de observações realizadas na população i, e

$$\epsilon_{ij} \sim N(0, \sigma),$$

são variáveis normais independentes.

A análise da variância ou ANOVA (ANalysis Of VAriance) é um método estatístico que serve para comparar as médias de duas ou mais populações. Este método deve o seu nome ao facto da comparação das diferentes médias ser feita através da análise da variância associada às observações. Mais precisamente, a ANOVA é baseada na comparação se dois estimadores da variância σ^2 comum às diversas populações. Um deles depende da veracidade de H_0, enquanto que o outro é válido independentemente de qual das hipóteses, H_0 ou H_a, é a verdadeira. Se os dois estimadores derem estimativas muito diferentes um do outro, isso dar-nos-á evidência contra a hipótese de igualdade das diversas médias.

11.2 Dois estimadores de σ^2

Para simplificar a exposição, iremos assumir num primeiro momento que as amostra recolhidas das diferentes populações têm tamanho n, isto é, $n_1 = \ldots = n_p = n$.

O primeiro estimador de σ^2 que consideramos não é mais do que a média dos estimadores de σ^2 que podemos obter com as observações recolhidas em cada uma das populações. Assim, se

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$$

e

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2,$$

representarem a média e a variância amostrais relativas à amostra recolhida da i-ésima população, atendendo a que as diferentes populações possuem iguais variâncias, um estimador natural de σ^2 é dado por

$$\text{MQD} = \frac{1}{p}\sum_{i=1}^{p} s_i^2 = \frac{1}{p(n-1)}\,\text{SQD}.$$

onde

$$\text{SQD} = \sum_{i=1}^{p}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2.$$

A SQD chamamos **soma de quadrados dentro das amostras** e a MQD **média de quadrados dentro das amostras**. Neste caso, como $\mu_{s_i^2}$ é um estimador sem viés de σ_i^2, isto é,

$$\mu_{s_i^2} = \sigma_i^2 = \sigma^2,$$

para $i = 1, \ldots, p$, a média da distribuição amostral de MQD é exatamente igual a σ^2, independentemente da hipótese nula ser, ou não, verdadeira:

$$\mu_{\text{MQD}} = \frac{1}{p}\sum_{i=1}^{p}\mu_{s_i^2} = \frac{1}{p}\sum_{i=1}^{p}\sigma^2 = \sigma^2.$$

Por outras palavras, sob H_0 ou sob H_a, MQD é um estimador sem viés de σ^2.

Admitamos agora que a hipótese nula é verdadeira, isto é, $\mu_1 = \ldots = \mu_p = \mu$. Neste caso, as p populações são idênticas em distribuição possuindo uma distribuição normal com média μ e variância σ^2. A média amostral $\bar{x}$ obtida a partir de uma qualquer amostra de tamanho n recolhida do conjunto das p populações é um estimador sem viés de μ com variância σ^2/n, isto é,

$$\mu_{\bar{x}} = \mu$$

e

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}.$$

Atendendo à assunção de normalidade podemos ainda dizer que

$$\bar{x} \sim N\left(\mu, \sigma/\sqrt{n}\right).$$

A segunda das igualdades anteriores sugere que é possível estimar a variância σ^2 desde que consigamos estimar a variância de $\bar{x}$ uma vez que

$$\sigma^2 = n\,\sigma^2_{\bar{x}}.$$

Como bem sabemos, tal é possível se dispusermos de uma amostra de realizações de $\bar{x}$. Tal acontece na presente situação uma vez que as médias amostrais $\bar{x}_1, \ldots, \bar{x}_p$, que podemos calcular a partir das amostras recolhidas de cada população, são realizações independentes e identicamente distribuídas de $\bar{x}$ com

$$\bar{x}_i \sim N\left(\mu, \sigma/\sqrt{n}\right).$$

Assim, a variância de $\bar{x}$ pode ser estimada pela variância amostral obtida a partir de $\bar{x}_1, \ldots, \bar{x}_p$, que vamos denotar por

$$s^2_{\bar{x}} = \frac{1}{p-1}\sum_{i=1}^{p}(\bar{x}_i - \bar{\bar{x}})^2,$$

onde

$$\bar{\bar{x}} = \frac{1}{pn}\sum_{i=1}^{p}\sum_{j=1}^{n}x_{ij} = \frac{1}{p}\sum_{i=1}^{p}\bar{x}_i.$$

Finalmente o segundo estimador de σ^2 que vamos considerar é dado por

$$\text{MQE} = \frac{1}{p-1}\,\text{SQE}$$

onde

$$\text{SQE} = n\sum_{i=1}^{p}(\bar{x}_i - \bar{\bar{x}})^2.$$

A SQE chamamos **soma de quadrados entre amostras** e a MQE **média de quadrados entre amostras**.

Sob H_0, podemos então concluir que MQE é um estimador sem viés de σ^2:

$$\mu_{\text{MQE}} = n\mu_{s^2_{\bar{y}}} = n\sigma^2_{\bar{y}} = n\frac{\sigma^2}{n} = \sigma^2.$$

Se H_a é verdadeira podemos mostrar que

$$\mu_{\text{MQE}} = \sigma^2 + \frac{n}{p-1}\sum_{i=1}^{p}(\mu_i - \bar{\mu})^2,$$

onde

$$\bar{\mu} = \frac{1}{p}\sum_{i=1}^{p}\mu_i.$$

Verificamos assim que MQE é um estimador sem viés de σ^2 sob a hipótese nula, mas não o é sob H_a.

11.3 Distribuições amostrais de MQD e MQE

A distribuição amostral das estatísticas MQD e MQE envolve a distribuição do qui-quadrado a que já fizemos referência em §10.1. Como vimos, a distribuição do qui-quadrado está relacionada com a distribuição normal. Assim se $x_1, \ldots, x_k$ são variáveis aleatórias independentes com distribuição $N(0,1)$ então

$$x_1^2 + \ldots + x_k^2 \sim \chi^2(k).$$

Em particular, se $x_1, \ldots, x_k$ são variáveis aleatórias independentes com distribuição $N(\mu, \sigma)$ concluímos que

$$\frac{1}{\sigma^2}\left((x_1-\mu)^2 + \ldots + (x_k-\mu)^2\right) \sim \chi^2(k).$$

É interessante notar que se a verdadeira média μ for substituída pela média amostral $\bar{x}$, a distribuição da variável anterior é ainda um qui-quadrado mas com menos um grau de liberdade, isto é,

$$\frac{1}{\sigma^2}\Big((x_1-\bar{x})^2 + \ldots + (x_k-\bar{x})^2\Big) \sim \chi^2(k-1).$$

Neste caso, as variáveis $y_i = x_i - \bar{x}$ que surgem nas várias parcelas, não são já independentes, estando relacionadas através da igualdade $\sum_{i=1}^{k} y_i = 0$. Significa que conhecendo os valores de $k-1$ dessas variáveis, fica determinado o valor da restante variável. O número de graus de liberdade passa assim de k para $k-1$.

Da definição de MQE e MQD concluímos que sob H_0,

$$\frac{p(n-1)}{\sigma^2}\,\text{MQD} \sim \chi^2(p(n-1))$$

e

$$\frac{(p-1)}{\sigma^2}\,\text{MQE} \sim \chi^2(p-1).$$

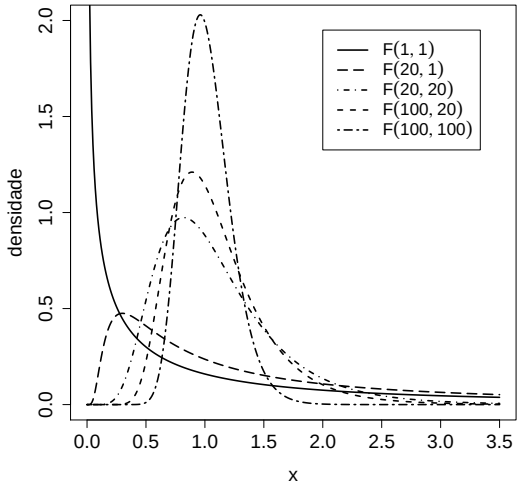
Além disso, é ainda possível mostrar que MQD e MQE são variáveis independentes.

11.4 O teste ANOVA

O cálculo do p-valor associado à estatística de teste que iremos definir a seguir para testar a hipótese de igualdade das diversas médias, usa uma nova distribuição de probabilidade. Esta distribuição está intimamente relacionada com a distribuição do qui-quadrado. Assim, se x_1 e x_2 são variáveis independentes com $x_1 \sim \chi^2(k_1)$ e $x_2 \sim \chi^2(k_2)$, o quociente

$$x = \frac{x_1/k_1}{x_2/k_2}$$

possui uma **distribuição de Fisher-Snedecor** com k_1 e k_2 graus de liberdade, e escrevemos $x \sim F(k_1, k_2)$. O primeiro parâmetro k_1 é o número de graus de liberdade do numerador, enquanto que o segundo parâmetro k_2 indica o número de graus de liberdade do denominador. A densidade de probabilidade de x para vários valores de k_1 e k_2 é dada na figura seguinte.



Tendo em conta os resultados anteriores, é assim natural que a estatística de teste que utilizamos para testar a hipótese de igualdade das diversas médias seja dada por

$$F = \frac{\text{MQE}}{\text{MQD}}.$$

Sob a hipótese H_0, podemos então concluir que

$$F \sim F(p-1, p(n-1)) = F(p-1, pn-p),$$

onde p é o número de populações envolvidas e pn é o número total de observações. Sob a hipótese H_a a média da distribuição amostral de MQE é maior que a de MQD, sendo de esperar que a estatística F tome valores superiores a 1.

No caso em que o número de observações em cada população não é o mesmo, o resultado anterior mantém-se válido desde que modifiquemos as estatísticas MQE e MQE de forma apropriada. Assim, definindo

$$\text{SQD} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2, \quad \text{SQE} = \sum_{i=1}^{p} n_i (\bar{x}_i - \bar{\bar{x}})^2,$$

e

$$\text{MQD} = \frac{1}{N-p}\,\text{SQD}, \quad \text{MQE} = \frac{1}{p-1}\,\text{SQE},$$

onde

$$N = n_1 + \ldots + n_p,$$

é o número total de observações, podemos mostrar que MQD e MQE são independentes com

$$\frac{N-p}{\sigma^2}\,\text{MQD} \sim \chi^2(N-p) \quad \text{e} \quad \frac{(p-1)}{\sigma^2}\,\text{MQE} \sim \chi^2(p-1).$$

Assim, concluímos que

$$F = \frac{\text{MQE}}{\text{MQD}} \sim F(p-1, N-p).$$

Reparemos que MQD não é mais do que o estimador da variância populacional que combina os estimadores $s_1^2, \ldots, s_p^2$ definido por

$$\text{MQD} = \frac{(n_1-1)s_1^2 + \ldots + (n_p-1)s_p^2}{n_1 + \ldots + n_p - p}.$$

Teste ANOVA:

Para testar a hipótese H_0 de igualdade das médias contra a hipótese H_a, use as observações para calcular

$$F = \frac{\text{MQE}}{\text{MQD}}$$

e obtenha o p–valor associado às observações realizadas através da fórmula

$$\text{P}(X \geq F),$$

onde X possui uma distribuição de Fisher-Snedecor com $p-1$ e $N-p$ graus de liberdade.

O cálculo da estatística de teste e do p-valor associado às observações é habitualmente realizado utilizando um software estatístico. Os resultados são apresentados numa tabela ANOVA, que no caso do SPSS toma a forma:

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	**SQE**	**p-1**	**MQE**	**$\frac{\text{MQE}}{\text{MQD}}$**	**p-valor**
Within Groups	**SQD**	**N-p**	**MQD**		
Total	**SQT**	**N-1**			

A quantidade SQT que surge na tabela, definida por

$$\text{SQT} = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})^2,$$

dá-nos uma ideia da variabilidade global das observações relativamente à média global. A partir da igualdade

$$x_{ij} - \bar{\bar{x}} = (x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{\bar{x}}),$$

podemos provar que, tal como é sugerido na tabela, as quantidades SQT, SQD e SQE estão relacionadas por

$$\text{SQT} = \text{SQD} + \text{SQE}.$$

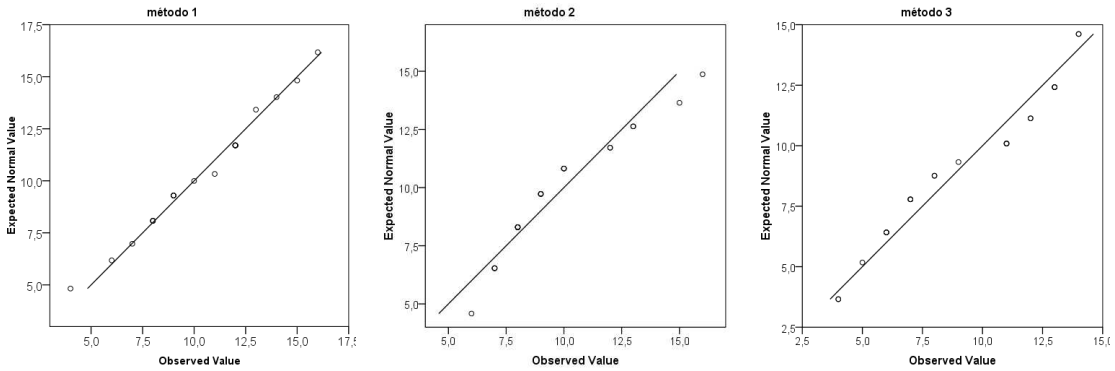
Assim, a variabilidade total associada às observações é decomposta em duas parcelas. A primeira, SQD, é relativa à soma das variabilidades dentro de cada uma das amostras visto estarmos a considerar desvios das observações relativamente à média de cada uma

das amostras. A segunda, SQE, é relativa à soma das variabilidades de cada uma das amostras, representadas pela respectiva média $\bar{x}_i$, relativamente relativamente à média global $\bar{\bar{x}}$.

Não podemos esquecer que o procedimento de teste anterior é baseado num conjunto de hipóteses teóricas que é necessário analisar antes de o pôr em prática. A primeira hipótese é relativa à normalidade das observações em cada uma das populações. A validade desta hipótese pode ser analisada a partir de gráficos de quantis normais para cada uma das amostras. A segunda hipótese é relativa à igualdade das variâncias das diversas populações. Uma primeira ideia sobre a variabilidade das populações pode ser conseguida a partir de gráfico de extremos-e-quartis paralelos. Em vez da utilização de um teste estatístico formal para testar a homogeneidade das variâncias, vamos adotar uma abordagem mais empírica ([2]), considerando que a hipótese de igualdade das variâncias pode ser aceite sempre que

$$\frac{\max(s_1, \ldots, s_p)}{\min(s_1, \ldots, s_p)} \leq 2.$$

Exemplo 11.4.1 Retomando o Exemplo 11.1.1 com que iniciámos o capítulo, comecemos por analisar as hipóteses subjacentes à utilização do procedimento ANOVA. Os gráficos de quantis normais relativos a cada um dos conjuntos de observações não revelam grandes desvios relativamente à hipótese de normalidade de cada uma das populações.



Por outro lado, temos

$$\frac{\max(s_1, s_2, s_3)}{\min(s_1, s_2, s_3)} = \frac{3{,}34}{2{,}69} = 1{,}24 \leq 2,$$

o que nos leva a considerar como válida a hipótese relativa à igualdade das variâncias das três populações.

[2] A este propósito, ver Moore e Mccabe, 2006, p. 729.

Assim, não havendo evidências de que as hipóteses subjacentes à utilização do procedimento ANOVA possam ser postas em causa, o p-valor (= 0,329) dado no quadro seguinte pode ser considerado credível. Ao nível 0,05 somos levados à aceitação da hipótese nula de igualdade das médias da variável observada nas três populações em causa, isto é, os dados não mostram evidência que haja diferenças entre os três métodos de ensino utilizados.

ANOVA

avaliação

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	20,576	2	10,288	1,132	,329
Within Groups	572,455	63	9,087		
Total	593,030	65			

11.5 O caso $p = 2$

No caso de apenas termos duas populações, o teste ANOVA pode ser usado para testar

$$H_0: \quad \mu_1 = \mu_2 \quad \text{contra} \quad H_a: \quad \mu_1 \neq \mu_2,$$

problema este que já estudámos no âmbito dos teste de comparação das médias de duas populações independentes. Tomando $p = 2$ nas fórmulas anteriores, MQD não é mais do que o estimador da variância populacional s^2, considerado em §9.2.1, que combina os estimadores s_1^2 e s_2^2, e

$$\text{MQE} = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^2.$$

Assim,

$$F = t^2 \sim F(1, n_1 + n_2 - 2) = t(n_1 + n_2 - 2)$$

onde

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{11.5.1}$$

é a estatística do teste t de Student de igualdade de duas médias quando assumimos que as variância de ambas as populações são iguais (apesar de desconhecidas). Assim, o teste ANOVA para duas populações e o teste de Student (bilateral) baseado na estatística anterior são equivalentes.

11.6 Comparações múltiplas

O teste ANOVA foi por nós utilizado para testar a hipótese de igualdade de todas as médias populacionais contra a hipótese alternativa de existirem pelo menos duas que são diferentes. A rejeição da hipótese nula não permite identificar quais dos pares de médias são significativamente diferentes. Se o nosso objetivo principal é o de identificar tais diferenças, a melhor abordagem talvez seja a de executar um procedimento de teste múltiplo baseado em testes parciais que permitam identificar tais diferenças, e que, simultaneamente, a rejeição da hipótese nula (de igualdade de todas as médias) seja compatível com a identificação de tais diferenças.

Sabendo nós testar a igualdade de qualquer par de médias, μ_i e μ_j, através dum teste de Student para amostras independentes, podemos pensar em testar as hipóteses anteriores, H_0 e H_a, testando

$$H_0^{ij}: \quad \mu_i = \mu_j \quad \text{contra} \quad H_a^{ij}: \quad \mu_i \neq \mu_j$$

para todos os pares de médias, e rejeitando a hipótese H_0 se pelo menos uma das hipóteses H_0^{ij} for rejeitada. Para p populações temos $k = p(p-1)/2$ destes testes. Se cada um dos testes parciais anteriores for efetuado ao nível $\alpha = 0{,}05$, o procedimento múltiplo descrito que conduz à rejeição de H_0 quando pelo menos uma das hipóteses H_0^{ij} seja rejeitada, tem, em geral, nível de significância superior a $\alpha = 0{,}05$. Tal não é admissível uma vez que o nível de significância fixado pelo utilizador permite ter controlo sobre o erro de primeira espécie associado ao procedimento de teste.

O método de Bonferroni que expomos a seguir permite contornar esta dificuldade através duma adequada alteração do nível a que cada um dos testes parciais é realizado. Outros métodos, como os de Tukey ou de Dunnett, podem ser também usados, com vantagem, em alternativa ao procedimento de Bonferroni. O método de Tukey, ou de Tukey-Kramer, é utilizado quando se pretende comparar todas as condições experimentais entre si, e o método de Dunnett é usado para comparar diversas condições experimentais com um controlo. A vantagem destes procedimentos sobre o procedimento ANOVA é a de permitirem identificar quais os pares de médias que são significativamente diferentes. A sua grande desvantagem está no facto de, caso a hipótese nula H_0 seja falsa, a probabilidade dos procedimento múltiplos rejeitarem H_0 ser em geral inferior à probabilidade do teste ANOVA rejeitar H_0. Qualquer um destes procedimentos múltiplos pode ser usado em alternativa ao procedimento ANOVA levando à rejeição de H_0 quando pelo menos uma das hipóteses H_a^{ij} for rejeitada.

O método de Bonferroni consiste em realizar cada um dos k testes parciais anteriores ao nível α/k. Quando usamos o procedimento de Bonferroni os testes parciais são

realizados usando as estatísticas de teste

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{\text{MQD}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}, \tag{11.6.1}$$

para $i, j = 1, \ldots, p$ e $i < j$. Relativamente à estatística de teste (11.5.1) estamos a substituir o estimador da variância populacional baseado em s_i^2 e s_j^2 pelo estimador combinado MQD que usa informação proveniente de todas as amostras. Sob a hipótese H_0 podemos provar que t_{ij} possui uma distribuição de Student com $N - p$ graus de liberdade, onde $N = n_1 + \ldots + n_p$. O p-valor ajustado (significa que deve ser comparado com o nível α fixado à partida e não com α/k) deste teste é dado por

$$p\text{-valor} = \min\left(2k\text{P}(T > |t_{ij}|)\,,\, 1\right), \tag{11.6.2}$$

onde T é uma variável de Student com $N - p$ graus de liberdade. Finalmente, o p-valor do teste múltiplo de Bonferroni quando testamos H_0 contra H_a é o menor dos p-valores ajustados anteriores.

Exemplo 11.6.3 Aplicando aos dados do Exemplo 11.1.1 o procedimento múltiplo de Bonferroni, o p-valor é de 0,416, e portanto somos conduzidos à mesma conclusão que obtivemos com a utilização do procedimento ANOVA. Relativamente aos p-valores que permitem testar $H_0^{12} : \mu_1 = \mu_2$ contra $H_a^{12} : \mu_1 \neq \mu_2$, $H_0^{13} : \mu_1 = \mu_3$ contra $H_a^{13} : \mu_1 \neq \mu_3$, e $H_0^{23} : \mu_2 = \mu_3$ contra $H_a^{23} : \mu_2 \neq \mu_3$, estes são dados por 1,000, 0,416 e 1,000, respetivamente, e portanto os resultados observados em cada uma destes casos não são significativos ao nível 0,05.

Multiple Comparisons

Dependent Variable: avaliação

Bonferroni

(I) método	(J) método	Mean Difference (I-J)	Std. Error	Sig.
método 1	método 2	,7727	,9089	1,000
método 2	método 3	,5909	,9089	1,000
método 3	método 1	-1,3636	,9089	,416

Uma vez que estamos interessados em efetuar comparações de todas as condições experimentais entre si, poderíamos ter também usado o teste de Tukey cujos resultados são apresentados na tabela seguinte. Nenhum dos resultados observados é significativo ao nível 0,05.

Multiple Comparisons

Dependent Variable: avaliação

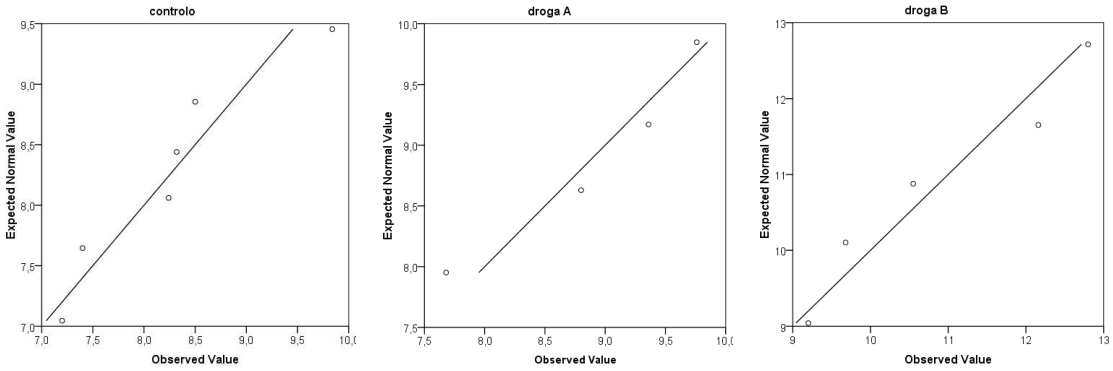
Tukey HSD

(I) método	(J) método	Mean Difference (I-J)	Std. Error	Sig.
método 1	método 2	,7727	,9089	,673
método 2	método 3	,5909	,9089	,793
método 3	método 1	-1,3636	,9089	,298

Exemplo 11.6.4 Os dados seguintes dizem respeito à contagem de células sanguíneas (milhões de células por milímetro cúbico) realizada através dum hemograma em três grupos de animais, um dos quais serviu como controlo, enquanto os outros dois foram tratados com duas drogas ([3]). Devido a perdas acidentais o número de animais em cada grupo não é o mesmo.

Grupos							médias	desvios-padrão
Controlo	7,40	8,50	7,20	8,24	9,84	8,32	8,25	0,94
Droga A	9,76	8,80	7,68	9,36			8,90	0,90
Droga B	12,80	9,68	12,16	9,20	10,55		10,88	1,56

O reduzido número de observações em cada um dos grupos não facilita a análise das condições teóricas subjacentes à utilização do procedimento ANOVA. No entanto, atendendo aos gráficos de quantis normais seguintes, não parece haver fortes indícios de que a distribuição das observações dos três grupos não seja normal.



Por outro lado, temos

$$\frac{\max(s_1, s_2, s_3)}{\min(s_1, s_2, s_3)} = \frac{1{,}56}{0{,}90} = 1{,}73 \leq 2,$$

o que nos leva a considerar como válida a hipótese relativa à igualdade das variâncias das três populações.

[3]Fonte: Dunnett, 1955.

Assim, não havendo evidências de que as hipóteses subjacentes à utilização do procedimento ANOVA possam ser postas em causa, o p-valor = 0,0091 dado no quadro seguinte pode ser considerado credível. Ao nível 0,05 somos levados à rejeição da hipótese nula de igualdade das médias da variável observada nas três populações em causa.

ANOVA

número de células

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	19,705	2	9,853	7,137	,0091
Within Groups	16,566	12	1,381		
Total	36,272	14			

O procedimento múltiplo de Bonferroni implementado no SPSS não permite obter os p-valores quando apenas estamos interessados nas duas comparações controlo vs. droga A e controlo vs. droga B. Os valores que constam da tabela seguinte são relativos às três comparações possíveis.

Multiple Comparisons

Dependent Variable: número de células

Bonferroni

(I) grupo	(J) grupo	Mean Difference (I-J)	Std. Error	Sig.
controlo	droga A	-,6500	,7584	1,0000
droga A	droga B	-1,9780	,7882	,0823
droga B	controlo	2,6280	,7115	,0092

No entanto, utilizando a fórmula (11.6.2) com $k = 2$ (e não com $k = 3$), obtemos os p-valores 0,8164 (controlo vs. droga A) e 0,0061 (controlo vs. droga B). Ao nível 0,05 o resultado observado é significativo quando comparamos o grupo de controlo com o grupo dos animais tratados com a droga B, mas não é significativo quando comparamos o grupo de controlo com o grupo dos animais tratados com a droga A. Assim, a utilização do procedimento múltiplo de Bonferroni conduz à rejeição da hipótese nula de igualdade das médias da variável observada nas três populações, mas permite também identificar como significativo apenas o resultado de comparação do grupo de controlo com o grupo dos animais tratados com a droga B.

Em alternativa ao procedimento de Bonferroni poderíamos também usar o método de Dunnett uma vez que estamos interessados em comparar o grupo de controlo com diversas condições experimentais. Apesar dos p-valores serem diferentes dos obtidos pelo método de Bonferroni, ao nível 0,05 as conclusões a retirar das observações realizadas são as mesmas.

Multiple Comparisons

Dependent Variable: número de células

Dunnett t (2-sided)[a]

(I) grupo	(J) grupo	Mean Difference (I-J)	Std. Error	Sig.
droga A	controlo	,6500	,7584	,6201
droga B	controlo	2,6280*	,7115	,0058

*. The mean difference is significant at the 0.05 level.

a. Dunnett t-tests treat one group as a control, and compare all other groups against it.

11.7 Bibliografia

Daniel, W.W. (2009). *Biostatistics: a foundation for analysis in the health sciences*, Wiley.

Dawson, B., Trapp, R.G. (2004). *Basic & Clinical Biostatistics*, McGraw-Hill.

Dunnett, C.W. (1955). A multiple comparison procedure for comparing several treatments with a control, *J. Amer. Statist. Assoc.* 50, 1096–1121.

Moore, D.S., McCabe, G.P. (2003). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

Moore, D.S., McCabe, G.P. (2006). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

12

Associação e regressão linear

Associação entre variáveis e gráfico de dispersão. Associação positiva e associação negativa. Associação linear e não-linear. Variável dependente e variável independente. Coeficiente de correlação linear. Reta de regressão e previsão pontual. Coeficiente de determinação. Gráfico de resíduos. Observações discordantes e observações influentes. Teste de validação do modelo de regressão. Intervalo de previsão para uma observação futura. Intervalo de confiança para a média de Y quando $X = x$.

12.1 Gráfico de dispersão

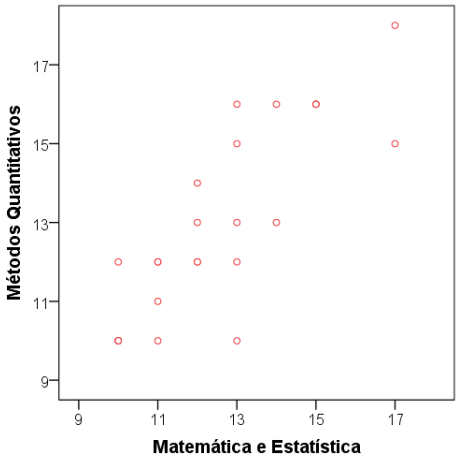
Em alguns dos conjuntos de dados que considerámos anteriormente, para cada um dos indivíduos observados, são registadas várias das suas caraterísticas. Apesar disso, as variáveis que lhes estão associadas foram por nós estudadas separadamente umas das outras. Esse foi, por exemplo, o caso dos dados apresentados na Figura 1.1.1, em que analisámos algumas das variáveis em que os dados estavam organizados. No entanto, poderia ser interessante analisar possíveis **relações** entre essas variáveis. Por exemplo, relações entre as variáveis "número de filhos" e "rendimento", ou entre as variáveis "sexo" e "rendimento".

Neste capítulo, no âmbito da análise exploratória de dados, desenvolveremos métodos gráficos e quantitativos para estudar a relação entre duas variáveis. Mais precisamente, para duas variáveis observadas num **mesmo conjunto de indivíduos**, interessamo-nos por identificar uma possível **associação** entre essas variáveis, isto é, se alguns valores assumidos por uma das variáveis tendem a ocorrer mais frequentemente com uns do que com outros dos valores assumidos pela outra variável.

Uma forma simples de explorar a possível associação entre duas **variáveis quantitativas**, X e Y, a partir de n observações de cada uma delas em que as observações x_i e y_i dizem respeito ao i-ésimo indivíduo observado, é representar estas observações num diagrama ou **gráfico de dispersão** onde cada um dos pontos (x_i, y_i) é marcado num

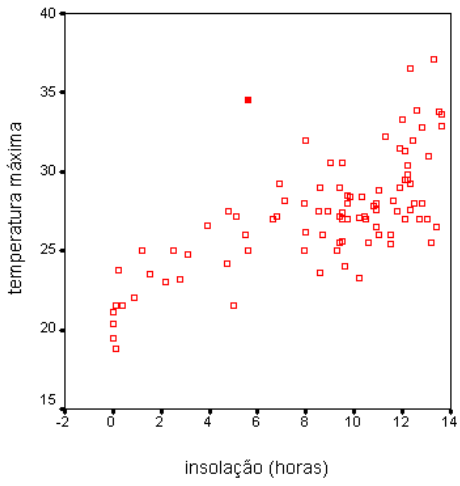
sistema de eixos coordenados. Este tipo de gráfico permite analisar o padrão geral das observações bem como desvios a esse padrão geral. O tipo de relação subjacente, no caso desta existir, e a sua intensidade, isto é, se se trata duma relação fraca, moderada ou forte, são ainda conclusões que podemos tirar deste tipo de gráfico.

Exemplo 12.1.1 Ilustremos o que acabámos de dizer considerando o gráfico de dispersão relativo às classificações obtidas por um grupo de alunos das disciplinas de Matemática e Estatística (1º ano), variável X, e de Métodos Quantitativos (2º ano), variável Y, da antiga licenciatura em Administração Pública da FDUC.



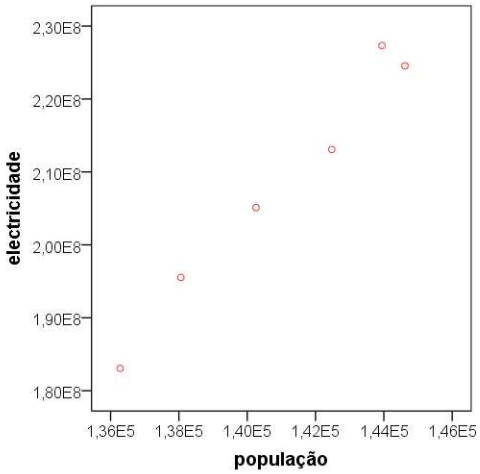
O gráfico anterior revela uma **associação positiva** entre as variáveis X e Y, pois aos menores e aos maiores valores de cada uma das variáveis correspondem, respetivamente, os menores e os maiores valores da outra variável.

Exemplo 12.1.2 O mesmo tipo de associação é revelado pelo gráfico de dispersão seguinte, relativo às horas de insolação e à temperatura máxima diárias observadas



em Coimbra entre 16 de Junho de 2002 e 15 de Setembro de 2002 (dados do Instituto Geofísico da Universidade de Coimbra). Cada um dos pontos do gráfico corresponde a um dos dias do período observado. Reparemos que o gráfico de dispersão põe em evidência a presença duma observação que está em desacordo com o padrão global revelado pelo gráfico (observação marcada a cheio). Por razões análogas ao que fizemos no capítulo anterior dizemos que se trata duma **observação discordante**.

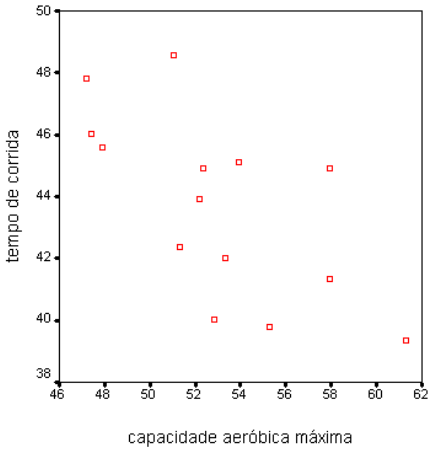
Exemplo 12.1.3 Uma associação positiva bastante mais forte do que a revelada em qualquer dos exemplos anteriores, é aquela que existe entre o consumo doméstico de eletricidade em Coimbra (quilowatt-hora) e a população aí residente durante o período 2009 a 2014 como podemos constatar do gráfico seguinte ([1]):



No caso de aos menores valores de cada uma das variáveis correspondem os maiores valores da outra variável, dizemos que o gráfico exibe uma **associação negativa** entre as duas variáveis. Um exemplo duma tal situação é apresentada a seguir.

Exemplo 12.1.4 Para 14 corredoras, registaram-se a capacidade aeróbica máxima (ml Kg^{-1} min^{-1}) e o tempo gasto para percorrerem determinada distância (min) ([2]). O gráfico de dispersão sugere que quanto maior for a capacidade aeróbica máxima, menor é, em geral, o tempo de corrida.

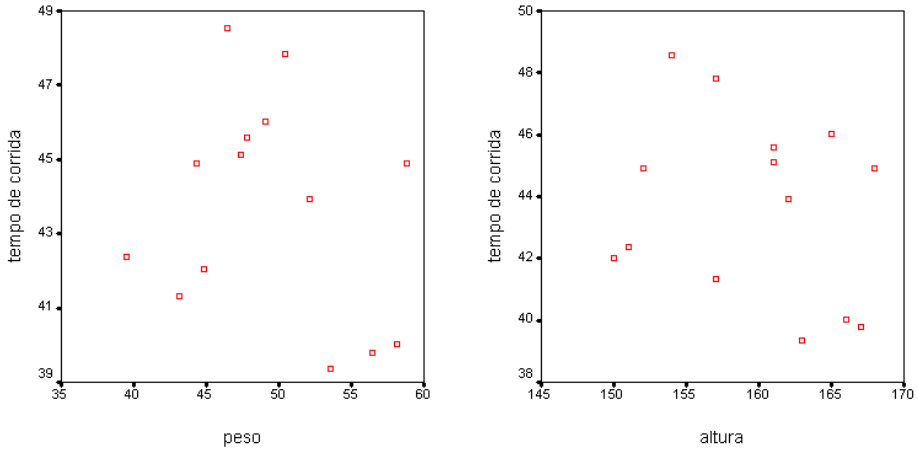
[1]Fonte: PORDATA (http://www.pordata.pt/)
[2]Fonte: Abraham e Ledolter, 1983, pág. 15.



Nos exemplos anteriores, a forma da relação exibida pelas variáveis em estudo é aproximadamente **linear**. Dizemos neste caso que se trata duma **associação linear**. Com isto queremos fazer referência ao facto dos pontos do gráfico se disporem para um e outro lado duma linha reta que podemos imaginar atravessar a nuvem de pontos marcados. A **associação linear** será tanto mais forte ou marcada quanto mais próximos dessa linha reta se dispuserem os pontos do gráfico.

Apresentamos a seguir dois exemplos de não associação. Os gráficos de dispersão respetivos não revelam qualquer padrão. A nuvem de pontos não exibe qualquer direção privilegiada.

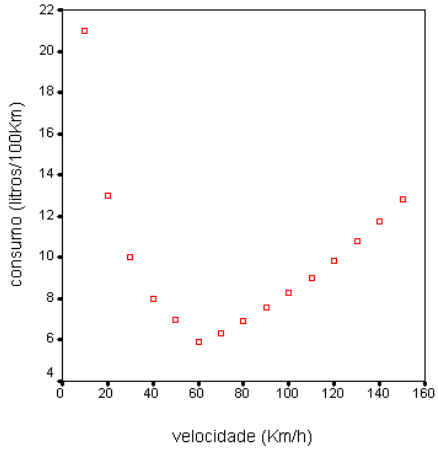
Exemplo 12.1.4 (cont.) Para as 14 corredoras registaram-se também os seus pesos (Kg) e alturas (cm). Os gráficos seguintes não revelam qualquer tipo de associação entre qualquer uma destas variáveis e o tempo de corrida.



As relações entre duas variáveis podem ser dos mais diversos tipos. Apresentamos

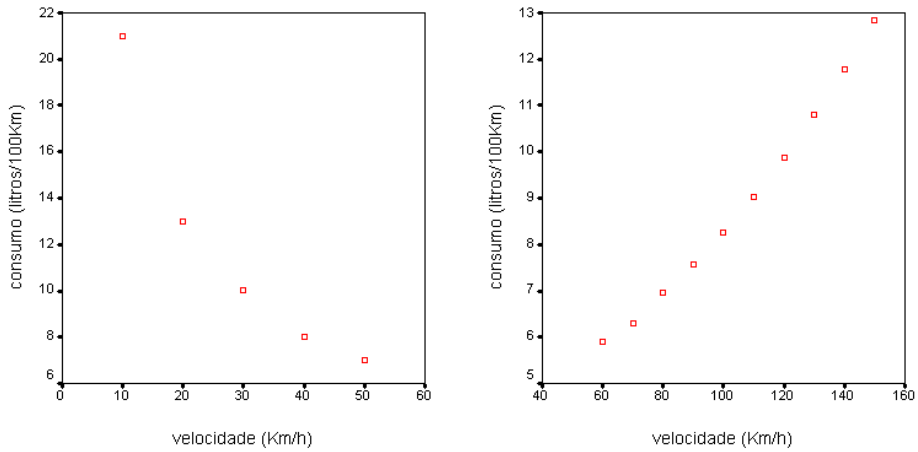
a seguir dois exemplos em que apesar de existir uma associação clara entre as variáveis em presença, esta não pode ser classificada de positiva ou negativa.

Exemplo 12.1.6 No gráfico de dispersão seguinte, registam-se os consumo efetuados por um automóvel (litro/100Km) a diferentes velocidades (Km/h) ([3]):



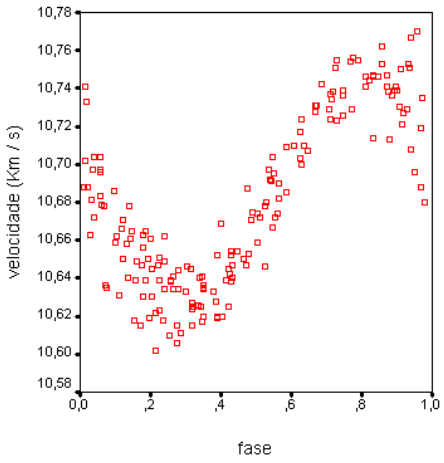
Apesar da forte associação, de tipo não-linear, exibida entre as duas variáveis, esta não pode ser qualificada de positiva nem de negativa pois, quer a valores baixos, quer a valores altos da velocidade, correspondem elevados níveis de consumo.

Reparemos que se nos restringirmos às velocidades inferiores ou iguais a 50Km/h, próprias de circuitos urbanos, ou às velocidades superiores a 50Km/h, habituais em circuitos de estrada, a associação entre as duas variáveis é aproximadamente de tipo linear, sendo negativa no primeiro caso e positiva no segundo. Além disso, a associação linear é mais forte no segundo caso do que no primeiro.



[3]Fonte: Moore e McCabe, 2006, pág. 119.

Exemplo 12.1.7 Observações da componente radial da velocidade de uma estrela (velocidade da estrela relativamente à Terra na direção da linha reta que une os seus centros de massa) que está a aproximadamente 60 anos luz da Terra revelam variações periódicas dessa velocidade com um período de aproximadamente 24,4 dias. As 181 observações realizadas são representadas a seguir em função da sua fase, isto é, em função da proporção de tempo decorrido desde o início do período em que a observação se insere ([4]). Também aqui é evidente uma forte associação de tipo não-linear entre as variáveis fase e velocidade.



Nos exemplos anteriores, estivemos unicamente interessados em explorar uma possível associação entre as variáveis em presença. Ao pormos em evidência uma tal relação, não estamos, necessariamente, a tentar explicar a variação observada numa das variáveis através da variação da outra. No entanto, ao explorarmos uma tal relação, podemos pensar que uma das variáveis, digamos X, pode **explicar** ou mesmo **causar** as variações observadas na outra variável Y. A variável Y diz-se então **variável resposta** ou **variável dependente**. Por oposição, à variável X chamamos **variável explicativa** ou **variável independente**. Mais à frente veremos que é possível quantificar o grau de explicação que a variável independente comporta sobre a variável dependente.

No exemplo sobre uma possível relação entre a temperatura máxima diária e o tempo de insolação diário, podemos colocar a possibilidade desta última variável poder explicar a primeira. No último dos exemplos anteriores, ao observarmos o consumo do automóvel para diferentes velocidades, poderá ser razoável pensar que a variação da variável "consumo" possa ser explicada a partir da variação da variável "velocidade", ou mesmo que a variação desta última seja a causa para a variação da primeira. Nessas

[4]Fonte: Santos, N.C. et al. (2003). The CORALIE survey for southern extra-solar planets, *Astronomy & Astrophysics*, 406, 373–381.

circunstâncias, as variáveis "consumo" e "temperatura máxima diária" são as **variáveis dependentes** enquanto que as variáveis "velocidade" e "tempo de insolação diário" são as **variáveis independentes**.

Sempre que estivermos em presença de variáveis com estas caraterísticas, na construção do gráfico de dispersão devemos colocar no eixo horizontal a variável independente e no eixo vertical a variável dependente.

Antes de terminar este parágrafo frisemos que ao estabelecermos a associação entre duas variáveis, uma dependente e a outra independente, não podemos em geral concluir pela **causalidade** duma delas relativamente à outra. Por outras palavras, não podemos concluir, sem mais, que a causa para a variação da variável dependente seja a variação presente na variável independente.

Um exemplo claro disso é-nos dado no Exemplo 12.1.1. Pensando na classificação de Métodos Quantitativos como variável dependente (cadeira do 2º ano) e na classificação de Matemática e Estatística como variável independente (cadeira do 1º ano), não podemos deduzir uma relação de causa-efeito entre estas duas variáveis. É mais razoável pensar que associação positiva observada se deve, por exemplo, ao facto de ambas as disciplinas exigirem conhecimentos na área da Matemática.

O Exemplo 12.1.2 é outro caso em que sem uma análise mais profunda não podemos dizer que a temperatura máxima é determinada pela insolação. Possivelmente haverá outra ou outras variáveis que não estamos a considerar, que, conjuntamente com a insolação, determinam a temperatura máxima. Também no Exemplo 12.1.7 a associação exibida não pode ser atribuída ao facto da fase ser a causa para a variabilidade observada para velocidade. Os autores do trabalho de onde foram retiradas estas observações defendem que em volta desta estrela orbita um planeta gigante, do tipo de Júpiter ou Saturno, pois de outro modo a componente radial da velocidade não variaria de forma sistemática com a fase.

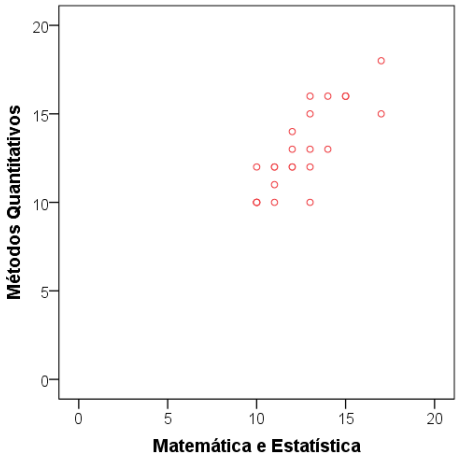
Dos exemplos anteriormente considerados, o Exemplo 12.1.6 é o que mais próximo está duma relação de causalidade. No entanto, seria importante saber mais sobre a experiência realizada, em particular, se não poderá haver mais variáveis que possam influenciar o consumo além da velocidade. Em caso afirmativo, será que essas variáveis estão controladas nas várias observações feitas a diferentes velocidades?

12.2 Coeficiente de correlação linear

Um gráfico de dispersão permite pôr em evidência a **forma**, a **direção** e a **intensidade** da relação entre duas variáveis quantitativas. A **relação linear** entre duas variáveis é, também pela sua simplicidade, particularmente importante.

Na secção anterior, qualificámos a associação linear entre duas variáveis de acordo com a sua intensidade. Usámos as palavras forte, moderada e fraca, para exprimir o facto dos pontos marcados no gráfico de dispersão estarem mais ou menos próximos duma reta imaginária que atravessa a nuvem dos pontos marcados. Apesar de bastante intuitiva, é por vezes difícil dizer quando é que um par de variáveis revela uma maior associação que outro par de variáveis. Por exemplo, não é fácil ordenar, relativamente ao grau de associação exibido, os pares de variáveis consideradas nos Exemplos 12.1.1, 12.1.2 e 12.1.4. Por outro lado, esta análise é bastante subjectiva dependendo, em particular, da escala usada no gráfico de dispersão.

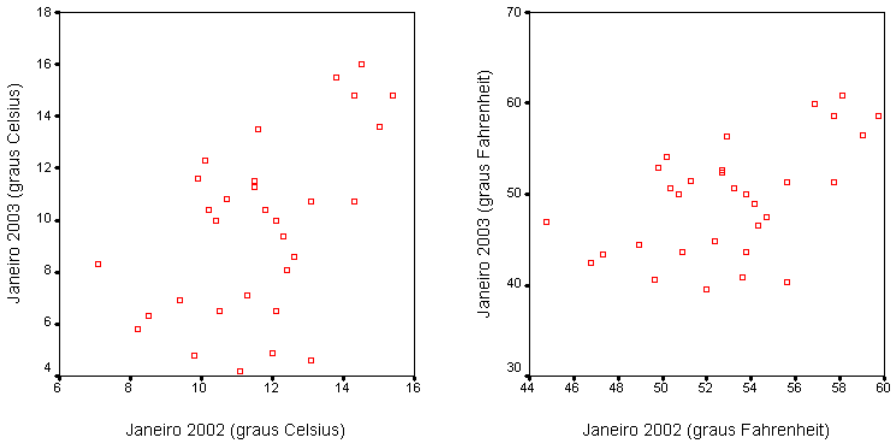
Exemplo 12.2.1 Para o exemplificar, apresentamos a seguir um gráfico de dispersão relativo aos dados do Exemplo 12.1.1 mas onde tomámos em cada um os eixos, valores de 0 a 20. Tendo classificado de moderada a associação positiva entre estas variáveis, também agora a devemos classificar do mesmo modo. No entanto, é claro que, em termos absolutos, neste gráfico os pontos estão mais próximos duma reta imaginária que atravessa o conjunto dos pontos marcados do que no gráfico do Exemplo 12.1.1.



Uma situação análoga pode ser observada, se alterarmos a unidade da medida que utilizamos para registar os dados, e ao mesmo tempo não tivermos o cuidado de, da mesma forma, alterar os intervalos de variação das variáveis em cada um dos eixos.

Exemplo 12.2.2 Os gráfico de dispersão seguintes são relativos às temperaturas médias diárias observadas em Coimbra nos meses de Janeiro de 2002 e 2003. Cada um dos pontos do gráfico corresponde a um dos dias do mês. O segundo gráfico parece revelar uma associação mais forte que o primeiro.

Estas considerações tornam clara a necessidade de quantificar a relação entre as variáveis em estudo. Uma forma simples de quantificar a associação linear entre duas variáveis quantitativas, é através do chamado **coeficiente de correlação linear**.



Denotando por $\bar{x}$, $\bar{y}$, s_x e s_y, a média e o desvio-padrão das observações $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_n$, respetivamente, o **coeficiente de correlação linear** entre as duas variáveis, que denotamos pela letra r, é definido por

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right),$$

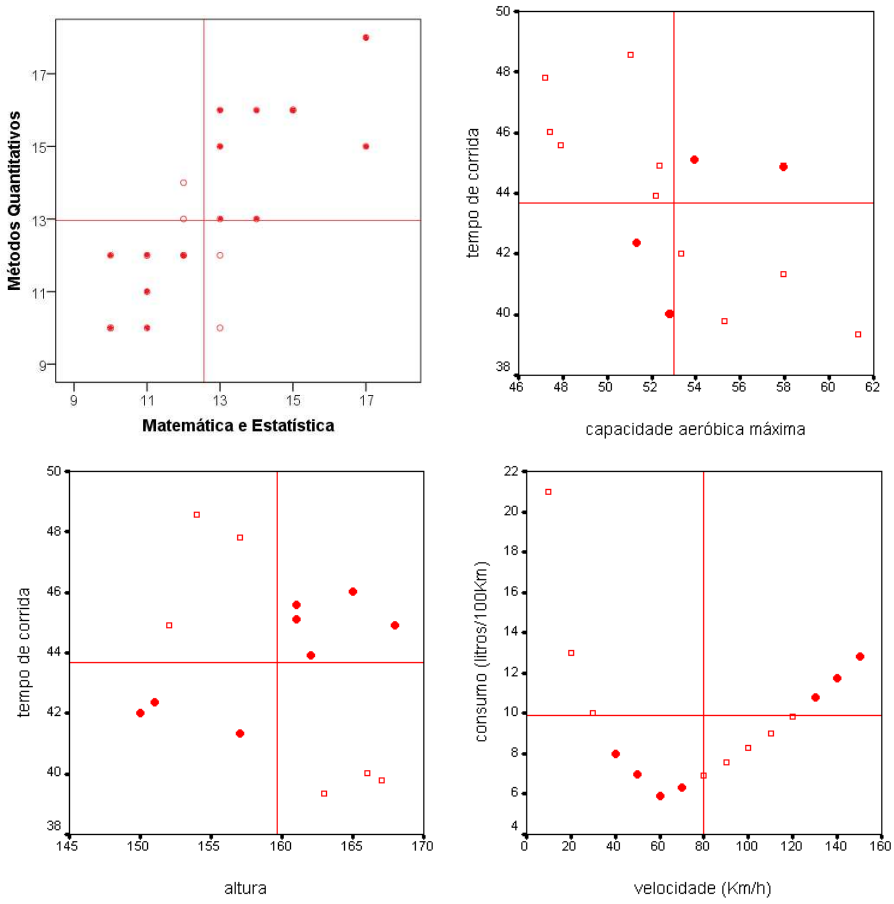
Na fórmula anterior intervêm as observações **padronizadas** ou **estandardizadas** associadas a x_i e a y_i que são definidos, respetivamente, por

$$\frac{x_i - \bar{x}}{s_x} \quad \text{e} \quad \frac{y_i - \bar{y}}{s_y}.$$

Estes valores dão-nos o número de desvios-padrão que cada um dos valores x_i e y_i, se afasta da média respetiva. Se, por exemplo, a observação x_i está à direita de $\bar{x}$ e a observação y_i está à esquerda de $\bar{y}$, o primeiro dos valores anteriores é positivo e o segundo é negativo. De uma forma geral, se x_i e y_i são simultaneamente "grandes" ou simultaneamente "pequenos", o produto dos seus valores padronizados é positivo e estas observações contribuem positivamente para coeficiente de correlação linear. Se x_i é "pequeno" quando y_i é "grande", ou x_i é "grande" quando y_i é "pequeno", o produto dos seus valores padronizados é negativo e estas observações contribuem negativamente para coeficiente de correlação linear. No primeiro caso, o ponto (x_i, y_i) está no primeiro ou no terceiro quadrante dum sistema de eixos coordenados com origem em $(\bar{x}, \bar{y})$. No

segundo caso, o ponto (x_i, y_i) está no segundo ou no quarto quadrante dum sistema de eixos coordenados com origem em $(\bar{x}, \bar{y})$.

Nos gráficos seguintes, identificam-se as observações que contribuem positivamente (marcas circulares) e negativamente (marcas quadradas) para o coeficiente de correlação linear dos pares de variáveis indicadas que considerámos nos Exemplos 12.1.1, 12.1.4 e 12.1.6. Reparemos uma observação (x_i, y_i) contribui mais para r, quer positivamente, quer negativamente, quanto mais distantes de $\bar{x}$ e $\bar{y}$ estejam, x_i e y_i, respetivamente (porquê?).



Contribuições negativas e positivas para r

De seguida enumeramos propriedades importantes do coeficiente de correlação linear.

Propriedades do coeficiente de correlação linear r:

⊙ r é uma medida da associação linear entre duas variáveis quantitativas; r não descreve associações não-lineares;

⊙ r não depende das unidades em que as variáveis estão expressas, isto é, r é invariante para alterações da unidade de medida;

⊙ r toma valores entre -1 e 1;

⊙ valores positivos de r indicam uma associação positiva, sendo esta associação tanto maior quanto mais r estiver próximo de 1; no caso limite $r = 1$, todas as observações estão sobre uma mesma reta com declive positivo;

⊙ valores negativos de r indicam uma associação negativa, sendo esta associação tanto maior quanto mais r estiver próximo de -1; no caso limite $r = -1$, todas as observações estão sobre uma mesma reta com declive negativo;

⊙ valores de r próximos de zero indicam uma fraca associação linear;

⊙ como se baseia no cálculo de médias e desvios-padrão, r é sensível a observações discordantes; deve por isso ser usado com cuidado quando o gráfico de dispersão sugerir a presença de observações discordantes.

A fórmula anteriormente dada para definir r, apesar de adequada para uma fácil interpretação do coeficiente de correlação linear, não é adequada para o seu cálculo. Em alternativa, este pode ser feito a partir de uma das fórmulas seguintes:

Cálculo do coeficiente de correlação linear:

$$r = \frac{\sum x_i\, y_i - n\,\bar{x}\,\bar{y}}{(n-1)\,s_x\,s_y} = \frac{\sum x_i\, y_i - \sum x_i \sum y_i/n}{(n-1)\,s_x\,s_y}.$$

Exemplo 12.2.3 Exemplifiquemos a utilização da fórmula anterior, efetuando o cálculo do coeficiente de correlação linear entre as variáveis "velocidade" (X) e "consumo" (Y) do Exemplo 12.1.6 (pág. 267). Da tabela seguinte, concluímos que

$$\bar{x} = 1200/15 = 80,$$

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	10	21,00	100	441,00	210,00
2	20	13,00	400	169,00	260,00
3	30	10,00	900	100,00	300,00
4	40	8,00	1600	64,00	320,00
5	50	7,00	2500	49,00	350,00
6	60	5,90	3600	34,81	354,00
7	70	6,30	4900	39,69	441,00
8	80	6,95	6400	48,30	556,00
9	90	7,57	8100	57,30	681,30
10	100	8,27	10000	68,39	827,00
11	110	9,03	12100	81,54	993,30
12	120	9,87	14400	97,42	1184,40
13	130	10,79	16900	116,42	1402,70
14	140	11,77	19600	138,53	1647,80
15	150	12,83	22500	164,61	1924,50
Σ	1200	148,28	124000	1670,01	11452,00

$$s_x = \sqrt{\frac{124000 - 15 \times 80^2}{14}} \approx 44{,}72,$$

$$\bar{y} = 148{,}28/15 \approx 9{,}89,$$

$$s_x = \sqrt{\frac{1670{,}01 - 15 \times (148{,}28/15)^2}{14}} \approx 3{,}82,$$

e

$$r \approx \frac{11452 - 15 \times 80 \times (148{,}28/15)}{14 \times 44{,}72 \times 3{,}82} \approx -0{,}17.$$

Sendo o valor de r próximo de zero, concluímos não haver associação linear entre as variáveis. Como podemos constatar através do primeiro gráfico do Exemplo 12.1.6, existe, contudo, uma relação não-linear entre as duas variáveis.

Exemplo 12.2.4 O cálculo do coeficiente de correlação linear pode ser feito de forma rápida utilizando o SPSS. Na tabela seguinte indica-se o coeficiente de correlação linear, conhecido também por coeficiente de correlação de Pearson, entre o consumo doméstico de eletricidade em Coimbra e a população aí residente no período entre 2009 e 2015, dados que considerámos no Exemplo 12.1.3 (pág. 265):

Correlations

		população
electricidade Kwh	Pearson Correlation	,988
	N	6

Tal como já tínhamos constatado no Exemplo 12.1.3 a associação linear positiva entre estas duas variáveis é bastante forte.

Os gráficos de dispersão seguintes ilustram a maior ou menor associação linear em função do coeficiente de correlação linear. Para que a comparação dos vários gráficos seja possível, os desvios-padrão de ambas as variáveis são iguais e as escalas de ambos os eixos são as mesmas.

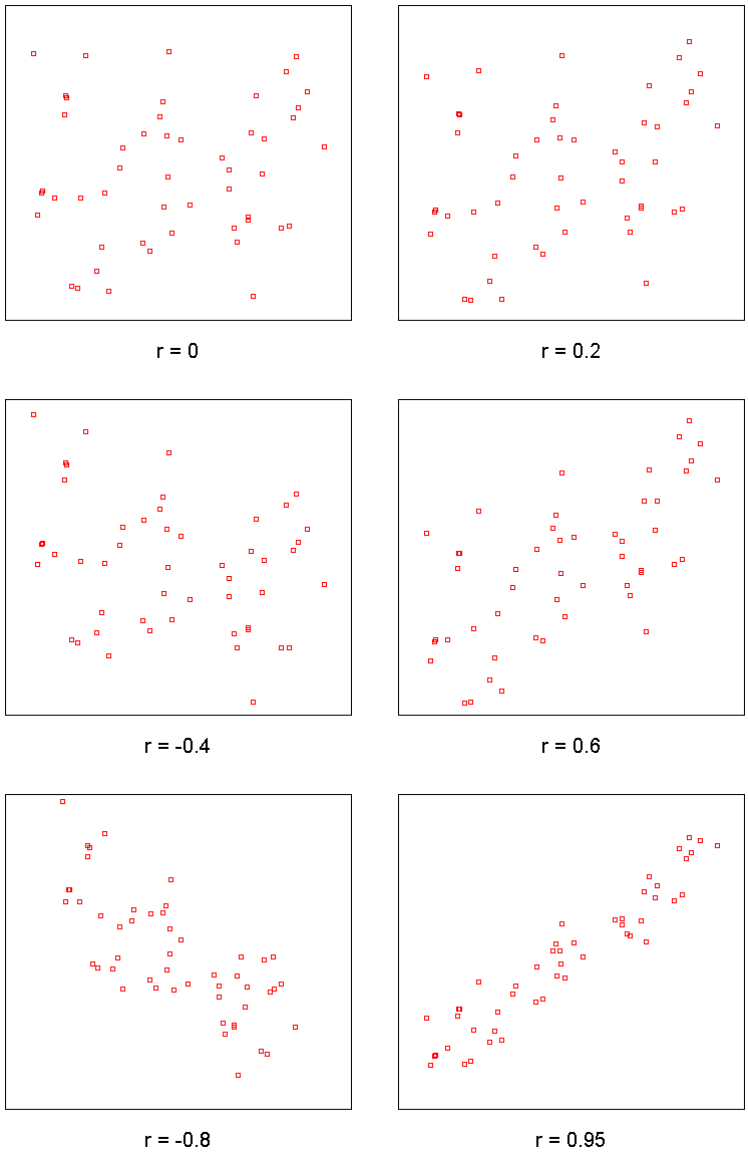


Figura 12.2.5: Associação linear em função de r

Para facilitar a interpretação do coeficiente de correlação linear, alguns autores consideram que a associação linear é forte quando $0{,}75 < |r| \leq 1$, moderada quando

$0{,}5 < |r| \leq 0{,}75$, fraca quando $0{,}25 < |r| \leq 0{,}5$ e que é residual ou inexistente quando $0 \leq |r| \leq 0{,}25$.

12.3 Modelo de regressão linear

Quando pretendemos resumir a informação contida num gráfico de dispersão que revela uma associação linear entre as duas variáveis em presença, é natural tentar ajustar aos dados uma reta. Essa reta resumirá tanto melhor a informação contida nos dados quanto maior for a associação, quer negativa, quer positiva, existente entre as variáveis em estudo. No caso dessa associação ser elevada, a reta poderia ainda ser utilizada para inferir o valor y duma das variáveis a partir do valor x da outra.

Mesmo que a associação linear entre as variáveis em estudo seja muito elevada, não faz sentido admitir que as observações realizadas $(x_1, y_1), \ldots, (x_n, y_n)$ pertencem todas a uma qualquer reta, que como sabemos pode ser representada por uma equação do tipo $y = a\,x + b$, onde os coeficientes a e b são, respetivamente, o **declive da reta** e a sua **ordenada na origem**. Será mais razoável admitir que as observações satisfazem a uma relação da forma

$$y_i = a\,x_i + b + \epsilon_i, \tag{12.3.1}$$

onde ϵ_i é um termo aleatório – positivo ou negativo consoante o ponto (x_i, y_i) esteja acima ou abaixo da reta, respetivamente –, que igual à distância com sinal entre o valor observado y_i e a reta de equação

$$y = a\,x + b, \tag{12.3.2}$$

isto é,

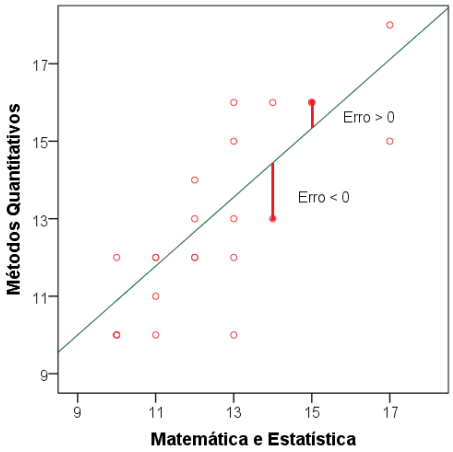
$$\epsilon_i = y_i - (a\,x_i + b). \tag{12.3.3}$$

A este valor chamamos **erro** ou **perturbação** associado à observação (x_i, y_i).

Além da relação (12.3.1) entre a variável dependente ou resposta Y e a variável independente ou explicativa X, é também habitual assumir que os diversos erros ϵ_i são independentes possuindo uma distribuição normal com média zero, o que significa que os pontos estão dispostos para um e outro lado da reta $y = ax + b$, e variância σ^2:

$$\epsilon_i \sim N(0, \sigma). \tag{12.3.4}$$

Exemplo 12.3.5 Para duas das observações consideradas no Exemplo 12.1.1, identificamos no gráfico seguinte os erros ϵ_i que lhes estão associados relativamente à reta marcada no gráfico.



As expressões (12.3.1) e (12.3.4) definem um modelo matemático a que chamamos **modelo de regressão linear**. Além do papel distinto que as variáveis Y e X desempenham num modelo de regressão linear, Y como **variável dependente** e X como **variável explicativa**, as duas variáveis têm natureza diferentes no contexto do modelo. Sem querer entrar em detalhes técnicos sobre esta questão diremos apenas que, contrariamente à variável Y, os valores assumidos pela variável X são considerados fixos à partida — ou seja, há um conjunto de valores pré-selecionados para X, a cada um dos quais corresponde um ou mais valores para Y —, hipótese esta que não é satisfeita na grande parte dos exemplos práticos que analisaremos. Assumiremos que este facto não acarretará erros significativos à análise estatística efetuada. Na prática os valores de a, b e $\sigma > 0$, ditos **parâmetros do modelo**, são desconhecidos estando associados à população de todos os possíveis pares (x_i, y_i) descritos pelo modelo de regressão linear.

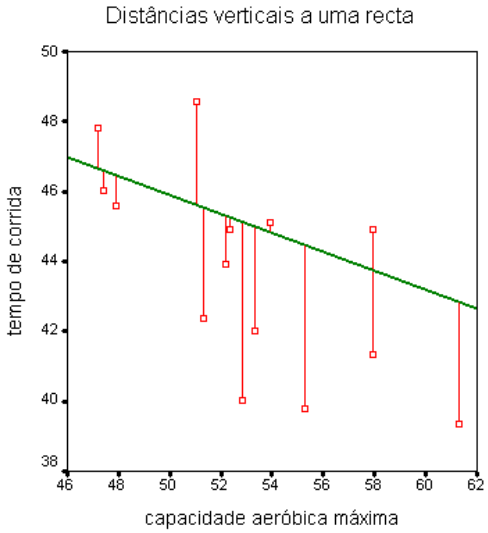
12.4 Reta de regressão

Se os parâmetros a e b do modelo de regressão linear fossem conhecidos, poderíamos facilmente usar a equação (12.3.2) para a partir do conhecimento de um valor particular da variável X inferir sobre o valor da variável Y que lhe está associado: sendo x o valor observado para X, uma aproximação $\widehat{Y}$ para o valor observado para Y seria dada por $\widehat{Y} = ax + b$. No entanto, como referimos atrás, os valores de a e b são desconhecidos e o melhor que podemos fazer é tentar obter aproximações para tais valores a partir das observações realizadas $(x_1, y_1), \ldots, (x_n, y_n)$. Sendo, como vimos, a e b o declive e a ordenada da origem da reta (12.3.2), encontrar aproximações para tais parâmetros a partir das observações realizadas é equivalente a encontrar uma aproximação para a reta anterior, ou ainda, como consequência da condição (12.3.4), equivalente à determinação

de uma reta que se "ajuste" bem às observações realizadas.

Desejando determinar uma reta que se "ajuste às observações realizadas", surge naturalmente o problema de saber o que isto significa. A ideia intuitiva, é a de que uma tal reta deve estar próxima, num certo sentido, de todos os pontos do gráfico de dispersão, ou, inversamente, todos os pontos do gráfico devem estar próximos, num certo sentido, da reta em causa. Usando um critério matemático conhecido como dos **mínimos quadrados**, que consiste em determinar a reta para a qual a soma dos quadrados das distâncias verticais entre a reta e os pontos $(x_1, y_1), \ldots, (x_n, y_n)$ é a mais pequena possível, podemos determinar com facilidade essa reta a que chamamos **reta dos mínimos quadrados** ou **reta de regressão** de Y sobre X.

No gráfico seguinte, que reproduz o primeiro gráfico de dispersão do Exemplo 12.1.4, e para uma reta nele desenhada, marcamos as distâncias verticais entre essa reta e cada um dos pontos do gráfico.



Tal como fizemos para a média e para o desvio-padrão, em que usámos símbolos distintos quando nos referíamos à média e ao desvio-padrão amostrais ($\bar{x}$ e s_x) ou à média e ao desvio-padrão populacionais (μ_X e σ_X), também aqui vamos guardar as letras a e b para representarem o declive e a ordenada na origem da reta $y = ax + b$ que surge na definição do modelo de regressão linear (que representa a população da qual recolhemos a amostra), e vamos representar por $\widehat{a}$ e $\widehat{b}$ o declive e a ordenada na origem da reta associada à amostra observada a que, como dissemos, chamamos **reta de regressão** de Y sobre X:

$$y = \widehat{a}\, x + \widehat{b}.$$

O declive, $\widehat{a}$, e a ordenada na origem, $\widehat{b}$, da **reta de regressão** associada à amostra $(x_1, y_1), \ldots, (x_n, y_n)$ são dados pelas fórmulas seguintes:

Reta de regressão de Y sobre X:

$\odot$ declive:

$$\widehat{a} = r\frac{s_y}{s_x} = \frac{\sum x_i\, y_i - n\,\bar{x}\,\bar{y}}{(n-1)\,s_x^2}$$

$\odot$ ordenada na origem:

$$\widehat{b} = \bar{y} - \widehat{a}\,\bar{x}.$$

Sendo $\widehat{a} = r\, s_y/s_x$ o declive da reta de regressão, podemos concluir que uma variação de um desvio-padrão em x corresponde a uma variação de r desvios-padrão em y. Além disso, a reta de regressão tem a propriedade de passar sempre no ponto $(\bar{x}, \bar{y})$.

Tendo calculado a reta de regressão, e sendo x o valor observado para X, podemos obter uma aproximação $\widehat{Y}$, dita **previsão**, para o valor observado para a variável Y dada por

$$\widehat{Y} = \widehat{a}\,x + \widehat{b}.$$

Notemos que, contrariamente ao coeficiente de correlação linear que não usa o facto de uma das variáveis poder ajudar a explicar ou a prever a outra, a **reta de regressão** necessita que tenhamos uma variável dependente (Y) e uma variável independente (X). Em particular, a reta de regressão de Y sobre X não coincide com a reta de regressão de X sobre Y.

Exemplo 12.4.1 Para os dados descritos no Exemplo 12.1.4 (pág. 265) e considerando o tempo de corrida como variável dependente (Y) e a capacidade aeróbica máxima como variável independente (X), determinemos a equação da reta de regressão de Y sobre X. Tendo em conta a tabela seguinte temos

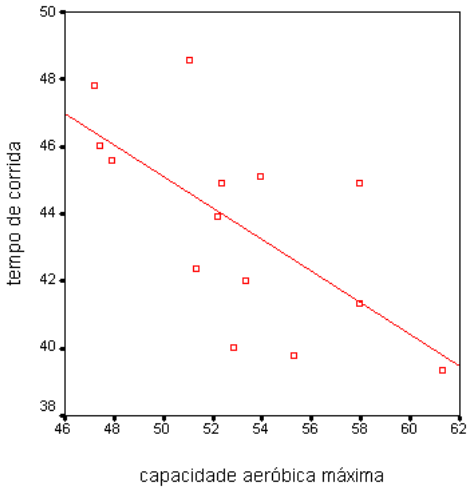
$$\bar{x} \approx 52{,}994, \quad s_x \approx 4{,}143, \quad \bar{y} \approx 43{,}699, \quad s_y \approx 2{,}938,$$

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	61,32	39,37	3760,1424	1549,9969	2414,1684
2	55,29	39,80	3056,9841	1584,0400	2200,5420
3	52,83	40,03	2791,0089	1602,4009	2114,7849
4	57,94	41,32	3357,0436	1707,3424	2394,0808
5	53,31	42,03	2841,9561	1766,5209	2240,6193
6	51,32	42,37	2633,7424	1795,2169	2174,4284
7	52,18	43,93	2722,7524	1929,8449	2292,2674
8	52,37	44,90	2742,6169	2016,0100	2351,4130
9	57,91	44,90	3353,5681	2016,0100	2600,1590
10	53,93	45,12	2908,4449	2035,8144	2433,3216
11	47,88	45,60	2292,4944	2079,3600	2183,3280
12	47,41	46,03	2247,7081	2118,7609	2182,2823
13	47,17	47,83	2225,0089	2287,7089	2256,1411
14	51,05	48,55	2606,1025	2357,1025	2478,4775
Σ	741,91	611,78	39539,5737	26846,1296	32316,0137

$$r \approx -0{,}660, \quad \widehat{a} \approx -0{,}468, \quad \widehat{b} \approx 68{,}500,$$

e a reta de regressão de Y sobre X, que traçamos na figura seguinte, tem por equação

$$y = -0{,}468\,x + 68{,}5.$$



Utilizando o SPSS podemos obter os quadros seguintes que contêm, entre outras coisas, o coeficiente de correlação linear ($r = -0{,}660$), o declive da reta de regressão ($\widehat{a} = -0{,}468$) e a ordenada na origem ($\widehat{b} = 68{,}494$).

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	68,494	8,176		8,377	,000
	capacidade aeróbica máxima	-,468	,154	-,660	-3,041	,010

a. Dependent Variable: tempo de corrida

Como já referimos, a reta de regressão é habitualmente utilizada para **inferir** o valor da variável dependente a partir do valor da variável independente. Por exemplo, para uma corredora com uma capacidade aeróbica máxima de 50, a reta de regressão anterior permite obter a **previsão** $\widehat{Y}$ para o seu tempo de corrida:

$$\widehat{Y} = -0{,}468 \times 50 + 68{,}494 = 45{,}094 \text{ minutos.}$$

12.5 Coeficiente de determinação

Além do coeficiente de correlação linear, também o seu quadrado, r^2, a que chamamos **coeficiente de determinação**, tem um papel importante na quantificação do poder explicativo do modelo de regressão linear. É possível provar matematicamente que r^2 é a fração da variabilidade da variável dependente Y que é explicada pela regressão da variável dependente sobre a variável independente. r^2 é assim uma **medida da qualidade da regressão**. Expresso em percentagem, o coeficiente de determinação indica a percentagem da variabilidade da variável Y que é explicada pela regressão linear de Y sobre X.

Exemplo 12.5.1 Retomando o Exemplo 12.4.1 (pág. 279) com que terminámos o parágrafo anterior, vemos que

$$r^2 = 0{,}435.$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,660[a]	,435	,388	2,29805

a. Predictors: (Constant), capacidade aeróbica máxima

Podemos assim concluir que através do modelo de regressão linear a capacidade aeróbica máxima explica 43,5% da variação observada no tempo de corrida. O valor 0,660 indicado neste quadro é o do módulo do coeficiente de correlação linear. Já vimos que o coeficiente de correlação linear é negativo e igual a $r = -0{,}660$.

Exemplo 12.5.2 Para os dados descritos no Exemplo 12.1.2 (pág. 264) e tomando a variável "horas de insolação diárias" como variável independente Y e a "temperatura máxima diária" como variável dependente X, concluímos dos quadros seguintes que 48,1% ($r^2 = 0{,}481$) da temperatura máxima diária é explicada pelo número de horas de insolação diárias através do modelo de regressão linear. Além disso, a reta de regressão de Y sobre X tem por equação

$$y = 0{,}619\,x + 22{,}002.$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,694[a]	,481	,476	2,58953

a. Predictors: (Constant), insolação (horas)

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	22,002	,651		33,811	,000
	insolação (horas)	,619	,068	,694	9,140	,000

a. Dependent Variable: temperatura máxima

Exemplo 12.5.3 Aproveitando a associação linear forte existente entre o consumo doméstico de eletricidade e a população residente em Coimbra no período 2009 a 2014 (ver Exemplos 12.1.3 e 12.2.4, pág. 265 e 274), a reta de regressão do consumo doméstico de eletricidade Kwh (Y) sobre a população residente (X) em Coimbra, permitir-nos-á apresentar uma previsão para o consumo doméstico de eletricidade em Coimbra no ano de 2015 a partir da população residente em Coimbra que nesse ano foi de 135085 habitantes.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,988[a]	,976	,970	2958399,403

a. Predictors: (Constant), população

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-506168285,600	56024028,120		-9,035	,00083
	população	5068,033	397,415	,988	12,752	,00022

a. Dependent Variable: electricidade

Do primeiro dos quadros anteriores verificamos o poder explicativo do modelo de regressão linear é elevado uma vez que $r^2 = 0{,}976$, ou seja, 97,6% da variabilidade observada no consumo doméstico de eletricidade é explicada pela população residente através do modelo de regressão linear considerado.

Do segundo quadro obtemos a equação da reta de regressão

$$y = 5068{,}033\,x - 506168285{,}6.$$

Tomando na equação anterior $x = 135085$ obtemos a estimativa de 178446952,21 Kwh para o consumo doméstico de eletricidade em Coimbra para 2015.

12.6 Gráfico de resíduos

Como **modelo matemático** que é, o **modelo de regressão linear**, e em particular a reta de regressão, descreve ou resume o padrão global da associação linear entre duas variáveis Y e X, mas não descreve desvios sistemáticos a esse padrão global. A análise dos desvios das observações relativamente ao modelo matemático, permite avaliar a adequação desse modelo às observações e identificar **observações discordantes**, que no contexto da associação entre duas variáveis podem ser discordantes em qualquer uma das direcções x ou y.

Sendo os coeficientes da reta de regressão baseados no cálculo de médias e desvios-padrão, será de esperar que tais coeficientes sejam **pouco robustos**, isto é, muito sensíveis a observações muito maiores ou menores que as restantes observações. É assim importante perceber a **influência** que essas observações discordantes têm, por si só, no cálculo da reta de regressão. Não seria razoável que a reta de regressão, que deverá descrever o padrão global das observações, seja determinada por observações que se desviam desse padrão global.

Uma forma de medirmos o desvio duma observação (x_i, y_i) relativamente ao padrão global, que assumimos resumido pela reta de regressão, é considerar o **resíduo** $\widehat{\epsilon}_i$ associado a essa observação que não é mais do a diferença entre a observação y_i e a **previsão**

$$\widehat{y}_i = \widehat{a}\,x_i + \widehat{b}$$

obtida a partir da reta de regressão:

$$\widehat{\epsilon}_i = y_i - (\widehat{a}\,x_i + \widehat{b}).$$

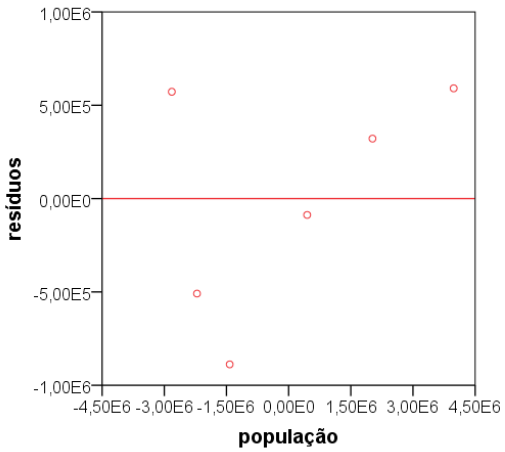
resíduo $\widehat{\epsilon}$ = observação Y - previsão $\widehat{Y}$

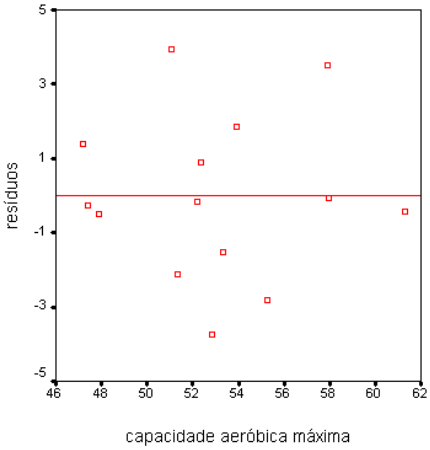
Reparemos que o resíduo $\widehat{\epsilon_i}$ não é mais do que a distância vertical com sinal entre o ponto (x_i, y_i) e a reta de regressão, sendo por isso, como podemos deduzir da equação (12.3.3), uma aproximação para o termo de erro ϵ_i que surge no modelo de regressão linear. O resíduo $\widehat{\epsilon_i}$ será positivo ou negativo consoante o ponto (x_i, y_i) esteja acima ou abaixo, respetivamente, da reta de regressão.

Para analisarmos os diversos resíduos vamos representá-los graficamente no chamado **gráfico de resíduos**. Um gráfico de resíduos é um gráfico de dispersão dos resíduos versus a variável independente. Neste gráfico é habitualmente marcada a reta horizontal correspondente às observações que não exibem qualquer desvio relativamente à reta de regressão. Acima e abaixo desta reta horizontal estão as observações que se encontram acima e abaixo, respetivamente, da reta de regressão. A distância de cada ponto à reta horizontal é precisamente a distância vertical, observada no gráfico de dispersão, entre a correspondente observação e a reta de regressão.

Se a reta de regressão descreve bem o padrão geral das observações, o gráfico de resíduos não deve apresentar nenhum padrão especial. Nesse caso, os resíduos têm a interessante propriedade de terem média zero, e os pontos marcados dispõem-se para um e outro lado da reta horizontal marcada no gráfico.

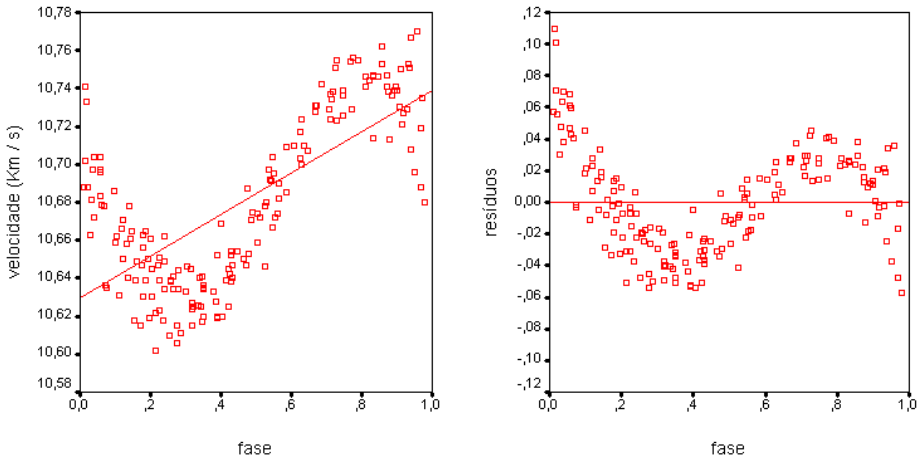
Exemplo 12.6.1 Os gráficos de resíduos seguintes, relativos aos dados dos Exemplos 12.1.3 (pág. 265, 274) e 12.1.4 (pág. 265), são exemplos de uma tal situação. A ausência de padrões nos gráficos de resíduos é indicador de que a reta de regressão descreve bem a associação existente entre as duas variáveis em causa.



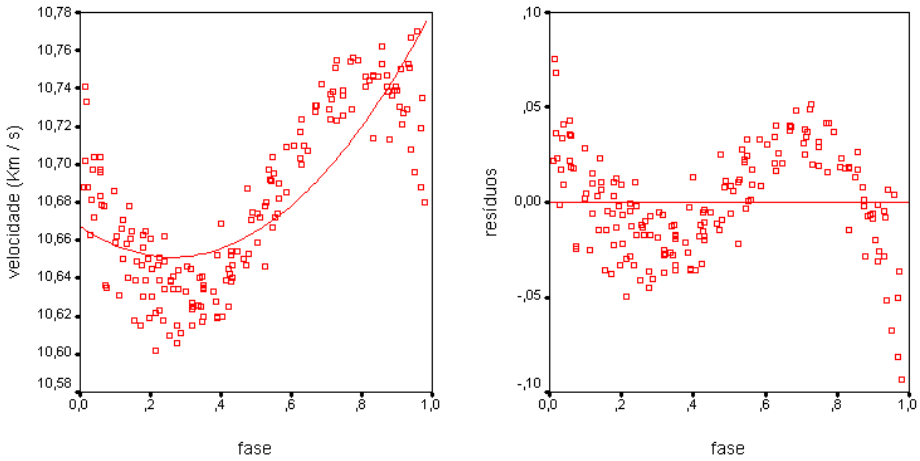


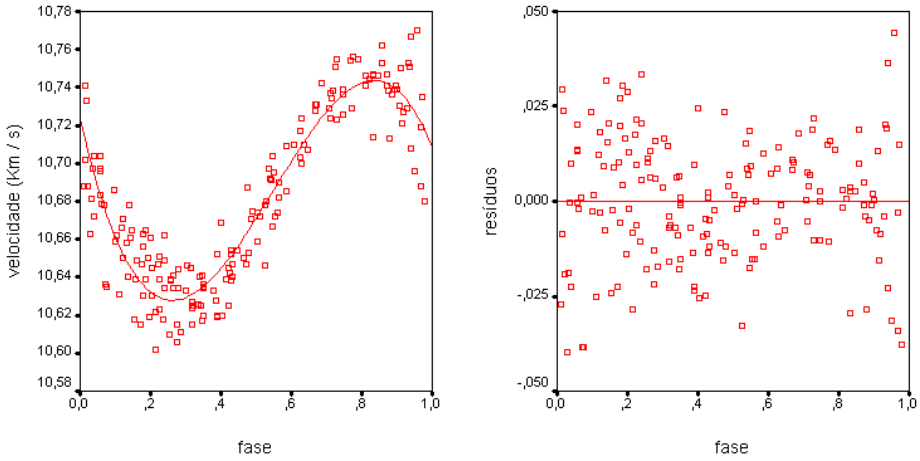
Se o gráfico de resíduos apresenta um padrão bem definido, podemos concluir que o modelo de regressão linear não descreve convenientemente os dados. Em particular, o padrão revelado pelo gráfico de resíduos é relativo à parte do padrão de associação entre as duas variáveis que não foi apreendida pela reta de regressão.

Exemplo 12.6.2 Um exemplo desta situação ocorreria se descrevêssemos através duma reta as observações, que analisámos no Exemplo 12.1.7 (pág. 268), sobre da componente da velocidade radial da estrela e da fase em que as observações foram realizadas. O gráfico de resíduos correspondente, revelaria uma forma sinusoidal que não é captada pela reta de regressão:



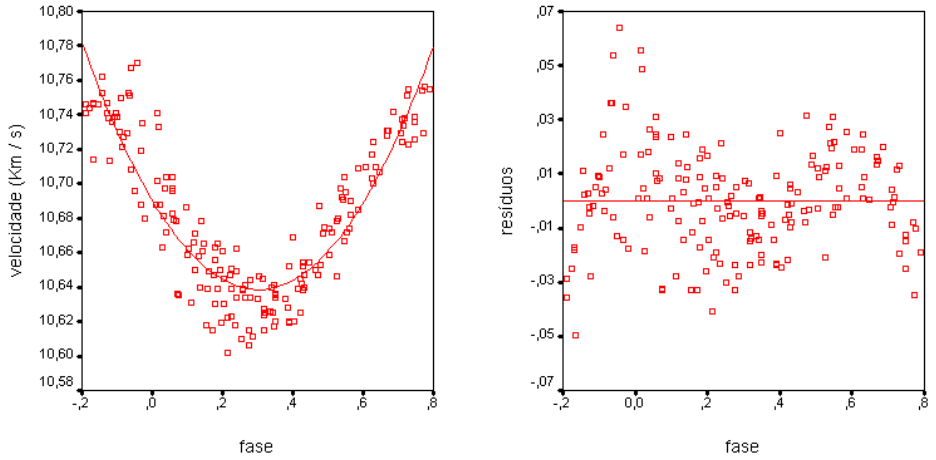
Há no entanto outros modelos matemáticos que poderiam descrever melhor o padrão revelado pelos dados anteriores. Sem entrar em detalhes sobre tais modelos, vejamos os resultados da utilização dum modelo de **regressão quadrática** e dum modelo de **regressão cúbica**.



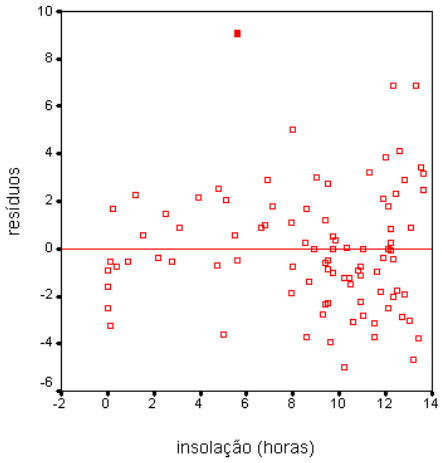


Como os próprios nomes indicam, no primeiro caso os dados são descritos por uma curva de equação $y = a\,x^2 + b\,x + c$, enquanto que no segundo caso é usada uma curva de equação $y = a\,x^3 + b\,x^2 + c\,x + d$. Como podemos concluir dos gráficos anteriores, dos modelos considerados apenas o modelo de regressão cúbica descreve os dados convenientemente.

Tratando-se no entanto de observações periódicas, se alterarmos o instante a partir do qual marcamos o tempo, é possível ajustar aos dados anteriores um modelo de regressão quadrática. Da análise dos gráficos de resíduos parece-nos que este modelo não descreve os dados tão bem como o modelo de regressão cúbica considerado atrás.

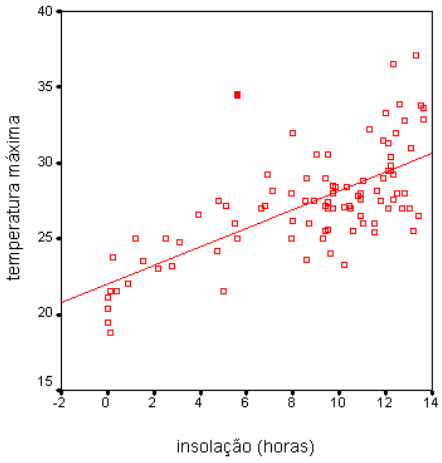


Exemplo 12.6.3 Relativamente aos dados do Exemplo 12.1.2 (pág. 264), o gráfico de resíduos seguinte põe claramente em evidência a observação discordante que tínhamos identificado a partir do gráfico de dispersão.



Como podemos constatar, trata-se duma observação discordante na direção do eixo dos yy. O gráfico revela ainda que maiores resíduos estão, em geral, associados a valores elevados ou muito pequenos de insolação. As previsões para a temperatura máxima a partir da reta de regressão calculada atrás, são assim menos exatas para esses valores de insolação.

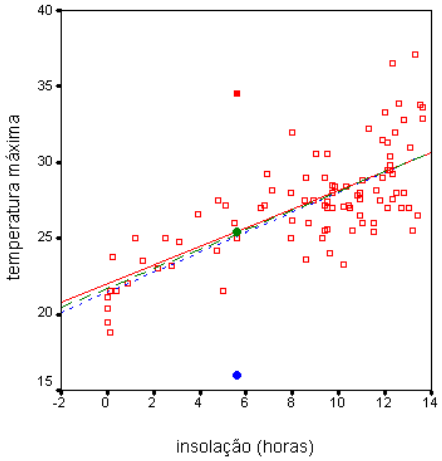
É interessante notar, que esta observação discordante na direção y, não é discordante quando considerada como observação da variável Y. Relativamente a esta variável podemos identificar, pelos métodos que já estudámos, quatro possíveis observações discordantes: duas por defeito e duas por excesso. Como podemos confirmar pelo gráfico seguinte, nenhuma das observações discordantes por excesso é a observação que identificámos como discordante na direção y.



Contrariamente ao que vimos no capítulo anterior em que uma observação discordante influenciava, só por si, o cálculo da média e do desvio-padrão, no contexto da

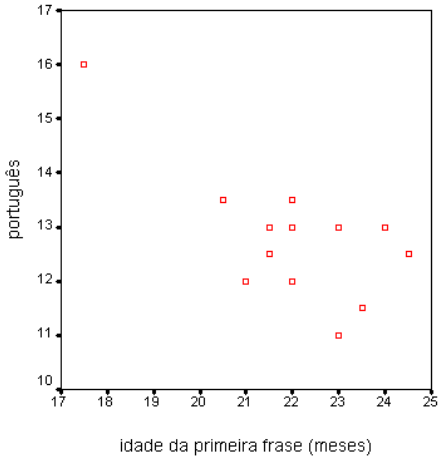
regressão uma observação discordante na direção y, apesar de ter um resíduo grande, não é necessariamente uma observação **influente**.

Na figura seguinte, ilustra-se esta situação deslocando verticalmente a observação discordante identificada no exemplo anterior, colocando-a primeiramente em cima da reta de regressão e depois em baixo desta. Como podemos verificar, a reta de regressão não sofreu uma alteração significativa.



Uma situação completamente diferente ocorre quando o gráfico de dispersão apresenta uma nuvem de pontos muito concentrada e um ponto afastado. Este ponto tem normalmente uma grande influência na reta de regressão.

Exemplo 12.6.4 Os dados apresentados no gráfico de dispersão seguinte dizem respeito a treze adolescentes para os quais foram registadas a idade em que disseram a primeira frase (em meses) e as classificações obtidas numa prova de aferição das suas capacidades em língua portuguesa.



Tomando a variável classificação em português como variável resposta Y e a variável idade da primeira frase como variável explicativa X, obtemos os resultados seguintes. Em particular, concluímos que a variável X explica 50% da variabilidade da variável Y.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,707[a]	,500	,455	,8985

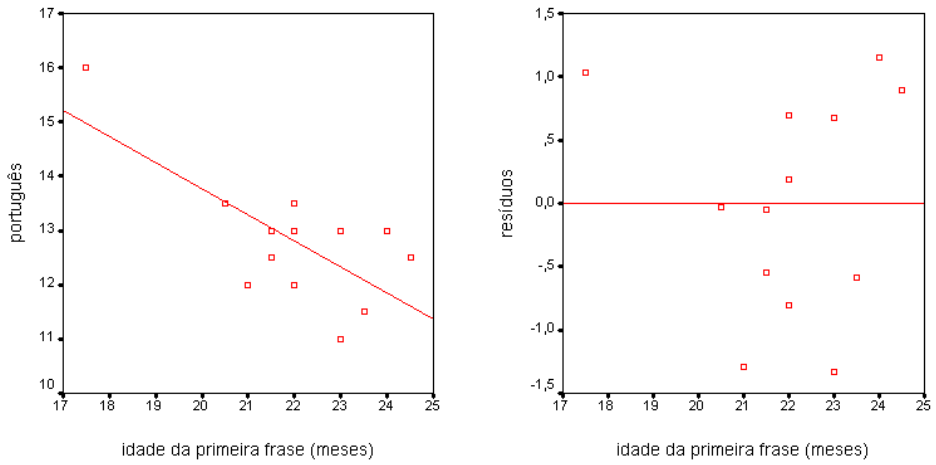
a. Predictors: (Constant), idade da primeira frase (meses)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	23,379	3,195		7,317	,000
	idade da primeira frase (meses)	-,481	,145	-,707	-3,318	,007

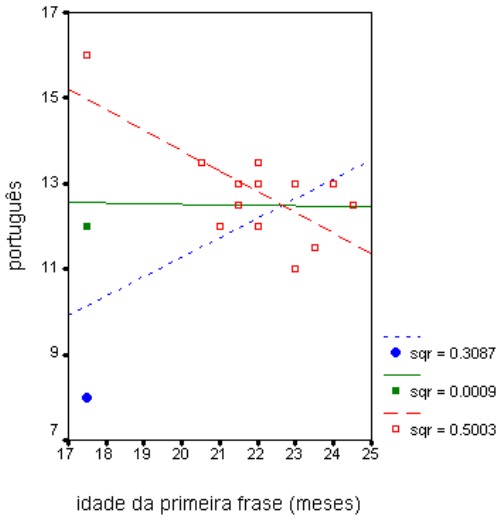
a. Dependent Variable: português

O gráfico de resíduos põe em evidência a presença duma observação discordante na direção x mas não na direção y, como poderia transparecer do gráfico de dispersão anterior. Este facto pode ser facilmente entendido se tivermos em conta a posição da reta de regressão.



Como já referimos, esta observação, além de **discordante**, é também uma observação muito **influente**. As conclusões que possamos tirar dos dados anteriores, dependem de forma determinante desta observação. Tal é claro a partir da figura seguinte. Reparemos também nas alterações significativas do coeficiente de determinação.

Tratando-se de uma verdadeira observação incorretamente registada ou de uma falsa observação, ela deve ser corrigida ou eliminada. No entanto, se a observação estiver correta, é necessário recolher mais informação se pretendemos chegar a alguma conclusão válida. Tendo em conta que a observação influente corresponde a um adolescente que pronunciou a primeira frase precocemente, essa informação adicional deve incidir sobre este tipo de adolescentes.



12.7 Um teste de validação do modelo de regressão

O coeficiente de determinação que considerámos em §12.5 deu-nos uma primeira medida da qualidade do modelo de regressão linear. Outra forma de avaliar a qualidade do modelo de regressão é testar a hipótese de nulidade do coeficiente a (declive da reta) do modelo de regressão, isto é, testar

$$H_0 : a = 0 \quad \text{contra} \quad H_a : a \neq 0.$$

No caso do teste levar à aceitação da hipótese nula, isso significa que as variáveis em estudo satisfazem a uma relação do tipo $y_i = b + \epsilon_i$, o que nos indica que a variável independente X será de pouca utilidade para a partir dela efetuarmos previsões sobre a variável dependente Y.

O teste estatístico que descrevemos a seguir é um procedimento que usa fortemente (no cálculo do p-valor) a hipótese feita em §12.3 sobre a distribuição do termo de erro que intervém no modelo de regressão linear e que assumimos ser normal de média zero e desvio-padrão σ: $\epsilon_i \sim N(0, \sigma)$. Na prática, a validade desta hipótese pode ser verificada lançando mão de um gráfico de quantis normais construído a partir dos

resíduos do modelo de regressão linear, pois já vimos que estes nos dão aproximações dos erros associados à observações realizadas.

Teste da hipótese $a = 0$:

Nas condições gerais do modelos de regressão linear, para testar a hipótese $H_0 : a = 0$ contra $H_a : a \neq 0$, calcule a estatística de teste t

$$t = \frac{\widehat{a}}{\widehat{\sigma}_{\widehat{a}}},$$

onde

$$\widehat{\sigma}_{\widehat{a}} = \frac{s_{\widehat{\epsilon}}}{\sqrt{(n-1)s_x^2}},$$

e o p-valor respetivo é dado por

$$p\text{-valor} = 2\mathrm{P}(T \geq |t|),$$

onde T tem uma distribuição $t(n-2)$ de Student.

As quantidades $s_{\widehat{\epsilon}}$ e $\widehat{\sigma}_{\widehat{a}}$ que surgem no teste anterior são, respetivamente, aproximações para o desvio-padrão σ do erro do modelo de regressão linear e para o desvio-padrão da distribuição amostral de $\widehat{a}$, que, sob a hipótese nula anterior, possui uma distribuição normal

$$\widehat{a} \sim N\left(0, \frac{\sigma}{\sqrt{(n-1)s_x^2}}\right).$$

$s_{\widehat{\epsilon}}$ e $\widehat{\sigma}_{\widehat{a}}$ são definidas por

$$s_{\widehat{\epsilon}}^2 = \frac{1}{n-2}\sum \widehat{\epsilon}_i^2$$

e

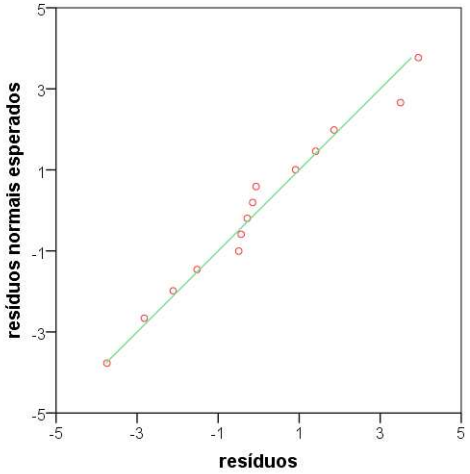
$$\widehat{\sigma}_{\widehat{a}} = \frac{s_{\widehat{\epsilon}}}{\sqrt{(n-1)s_x^2}},$$

mas, como veremos a seguir, estas quantidades podem ser obtidas facilmente das tabelas produzidas pelo SPSS.

Exemplo 12.7.1 Retomemos o modelo linear considerado no Exemplo 12.4.1 (pág. 279), que exprime o tempo de corrida (Y) em função da capacidade aeróbica máxima (X), cuja reta de regressão tinha por equação $y = -0{,}468\,x + 68{,}494$.

Pretendendo validar o modelo através do teste anterior, comecemos por avaliar a normalidade dos resíduos do modelo de regressão linear através do gráfico de quantis normais seguinte, que não apresenta desvios sistemáticos dos pontos relativamente à

reta marcada no gráfico. Concluímos assim que os resíduos do modelo possuem uma distribuição aproximadamente normal.



Passemos então à realização do teste da hipótese $H_0 : a = 0$ contra a hipótese $H_a, : a \neq 0$. Para tal vamos lançar dos quadro produzidos pelo SPSS:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,660[a]	,435	,388	2,29805

a. Predictors: (Constant), capacidade aeróbica máxima

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	68,494	8,176		8,377	,000
	capacidade aeróbica máxima	-,468	,154	-,660	-3,041	,010

a. Dependent Variable: tempo de corrida

Pretendendo obter o p-valor associado às observações realizados, esse valor surge indicado no segundo dos quadros anteriores:

$$p\text{-valor} = 0{,}01.$$

Isto significa que se tomarmos como nível de significância do teste 0,05, o resultado anterior conduziria à rejeição do hipótese nula, ou seja, concluiríamos que o modelo de regressão pode ser usado para efetuar previsões.

Pretendendo-se também obter o valor da estatística de teste a mesma surge também indicada no segundo dos quadros anteriores:

$$t = -3{,}041.$$

Claro que tendo o valor da estatística de teste, podemos obter uma aproximação para o p-valor a partir da fórmula de cálculo do p-valor e da tabela de Student. Com efeito, sabemos que

$$p\text{-valor} = 2\mathrm{P}(T \geq |-3{,}041|) = 2\mathrm{P}(T \geq 3{,}041),$$

onde T tem uma distribuição $t(n-2) = t(12)$. Recorrendo à Tabela D da distribuição de Student, encontramos que

$$0{,}005 < \mathrm{P}(T \geq 3{,}041) < 0{,}01,$$

o que nos leva à conclusão que

$$0{,}01 < p\text{-valor} < 0{,}02.$$

Além de ser mais morosa do que usarmos o quadro produzido pelo *software*, a utilização da tabela de Student não permite obter uma aproximação tão precisa para o p-valor associado às observações realizadas.

O valor da estatística de teste pode também ser calculado usando a fórmula indicada atrás. Para tal precisamos dos valores de $\widehat{a}$ e $\widehat{\sigma}_{\widehat{a}}$ que surgem indicados no segundo dos quadros anteriores:

$$t = \frac{\widehat{a}}{\widehat{\sigma}_{\widehat{a}}} = \frac{-0{,}468}{0{,}154} = -3{,}038.$$

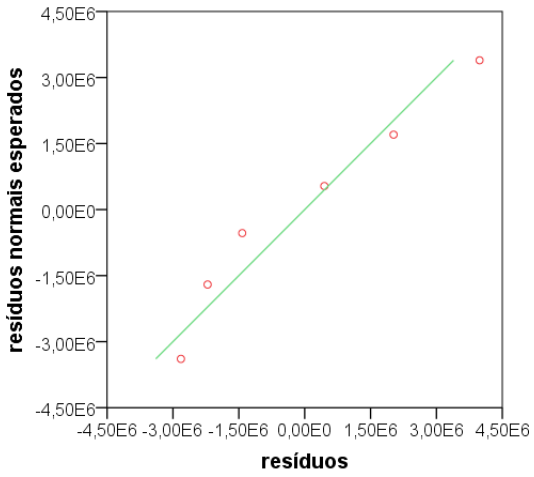
A pequena diferença relativamente ao valor anteriormente obtido para t é devida a erros de arredondamento.

É também possível calcular $\widehat{\sigma}_{\widehat{a}}$ usando a fórmula indicada atrás. Para tal necessitamos ainda de conhecer a variância amostral s_x^2, que neste caso é igual a $s_x^2 = 17{,}163$, e também o desvio-padrão $s_{\widehat{\epsilon}}$, valor que surge no primeiro dos quadros anteriores:

$$\widehat{\sigma}_{\widehat{a}} = \frac{s_{\widehat{\epsilon}}}{\sqrt{(n-1)s_x^2}} = \frac{2{,}298}{\sqrt{(14-1)\times 17{,}163}} = 0{,}1538.$$

Exemplo 12.7.2 Retomemos agora o modelo considerado no Exemplo 12.5.3 (pág. 282) que exprime o consumo doméstico de eletricidade (Kwh) em Coimbra (Y) em função da população residente em Coimbra (X), e cuja reta de regressão tinha por equação $y = 5068{,}033\,x - 506168285{,}6$. Pretendendo validar o modelo através do teste anterior, comecemos por avaliar a normalidade dos resíduos do modelo de regressão linear

através do gráfico de quantis normais seguinte. Tal como no exemplo anterior, também aqui não há desvios sistemáticos dos pontos relativamente à reta marcada no gráfico, o que permite concluir que os resíduos do modelo possuem uma distribuição normal.



A partir dos quadros apresentados no Exemplo 12.5.3 (pág. 282) podemos validar o modelo de regressão linear através da realização do teste da hipótese $H_0 : a = 0$ contra a hipótese $H_a : a \neq 0$. Neste caso encontramos no segundo quadro indicado o p-valor= 0,0022. Sendo este valor muito pequeno, somos conduzidos, para qualquer um dos habituais níveis de significância 0,05 ou 0,01, à rejeição do hipótese nula. Concluímos assim que o modelo de regressão pode ser usado para prever o consumo anual doméstico de eletricidade em Coimbra a partir da população residente em Coimbra nesse ano.

12.8 Intervalo de previsão para uma observação futura

Vimos em §12.4 que a principal utilidade da reta de regressão é permitir inferir o valor da variável dependente Y a partir da observação de um novo valor da variável independente X. Recuperando a terminologia que usámos no Capítulo 7 quando falámos de intervalos de confiança, a reta de regressão apenas nos tem permitido apresentar uma previsão pontual

$$\widehat{Y} = \widehat{a}x + \widehat{b},$$

para a variável Y, quando x é o valor observado para a variável X, não nos possibilitando fornecer qualquer informação sobre a precisão e a confiança que podemos ter nessa previsão. Tal como fizemos aquando da estimação de proporções e médias, que abordámos no Capítulo 7, além da previsão pontual interessa-nos também apresentar um intervalo de confiança centrado nessa previsão pontual, intervalo esse a que

chamamos, no presente contexto, de **intervalo de previsão**.

Nesta fase do processo de previsão é essencial a hipótese feita sobre a distribuição do termo de erro que intervém no modelo de regressão linear e que assumimos ser normal, com média zero e desvio-padrão σ, desconhecido: $\epsilon_i \sim N(0, \sigma)$. Como já referimos no parágrafo anterior, esta hipótese pode ser avaliada na prática lançando mão de um gráfico de quantis normais construído a partir dos resíduos do modelo de regressão linear. Sob a hipótese de normalidade anterior é possível mostrar que a variável $\widehat{Y} - Y$ possui uma distribuição normal com média zero,

$$\mu_{\widehat{Y}-Y} = 0,$$

e com desvio-padrão

$$\sigma_{\widehat{Y}-Y} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}.$$

Tal como fizemos para a construção dum intervalos de confiança para uma média quando o desvio-padrão da população era conhecido (ver pág. 176), caso o desvio-padrão σ, desvio-padrão do termo de erro do modelo de regressão, fosse conhecido, poderíamos usar a variável fulcral

$$z = \frac{\widehat{Y} - Y}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}} \sim N(0,1),$$

para construir um intervalo de previsão para Y. No entanto, sendo o desvio-padrão do termo de erro desconhecido, podemos proceder como nos intervalos de confiança para uma média (ver pág. 178), substituindo σ por $s_{\hat{\epsilon}}$ que, como vimos atrás, é uma aproximação para σ calculada a partir dos resíduos do modelo de regressão linear. Ao fazermos esta substituição e passarmos a usar a variável fulcral

$$t = \frac{\widehat{Y} - Y}{s_{\hat{\epsilon}}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}},$$

esta deixa de possuir uma distribuição normal standard para passar a ter uma distribuição de Student com $n-2$ graus de liberdade. Este facto permite obter o seguinte intervalo de previsão para o valor da variável Y quando a variável X toma o valor x.

Intervalo de previsão para Y observado $X = x$:

Nas condições gerais do modelos de regressão linear, um intervalo de previsão com nível de confiança C para Y observado $X = x$, tem por extremidades:

$$\widehat{Y} \pm t^* \, s_{\hat{\epsilon}} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2}},$$

onde

$$\widehat{Y} = \widehat{a}x + \widehat{b},$$

t^* é tal que

$$\mathrm{P}(-t^* \leq T \leq t^*) = C$$

e T tem uma distribuição de Student $t(n-2)$.

Como já referimos atrás, a quantidade $s_{\hat{\epsilon}}$ que surge no intervalo de previsão anterior pode ser obtida facilmente das tabelas produzidas pelo SPSS.

Exemplo 12.8.1 No modelo linear considerado no Exemplo 12.7.1 (pág. 291), que exprime o tempo de corrida (Y) em função da capacidade aeróbica máxima (X), usámos a reta de regressão $y = -0{,}468\,x + 68{,}494$, para obter a previsão seguinte do tempo de corrida de uma atleta que tem uma capacidade aeróbica máxima de 50:

$$\widehat{Y} = -0{,}468 \times 50 + 68{,}494 = 45{,}094 \text{ minutos.}$$

Complementemos agora esta previsão apresentando um intervalo de previsão de nível de confiança 95% para o tempo de corrida da mesma atleta. Comecemos pelo cálculo do valor t^* que satisfaz a igualdade

$$\mathrm{P}(-t^* \leq T \leq t^*) = 0{,}95,$$

onde T tem uma distribuição de Student $t(14-2) = t(12)$. Da tabela distribuição de Student concluímos que

$$t^* = 2{,}179.$$

Atendendo a que $\bar{x} = 52{,}994$ e $s_x^2 = 17{,}163$, das tabela apresentadas no Exemplo 12.7.1 (pág. 291) deduzimos que a margem de erro do intervalo de previsão é dada por

$$\text{margem de erro} = 2{,}179 \times 2{,}298 \sqrt{1 + \frac{1}{14} + \frac{(50 - 52{,}994)^2}{(14-1) \times 17{,}163}} = 5{,}279 \text{ minutos.}$$

Assim o intervalo de previsão de nível de confiança 95% para o tempo de corrida de uma atleta com uma capacidade aeróbica máxima de 50 é de

$$45{,}094 \pm 5{,}279 = [39{,}815\,;\ 50{,}373].$$

Exemplo 12.8.2 No modelo considerado no Exemplo 12.5.3 (pág. 282) que exprime o consumo doméstico de eletricidade (Kwh) em Coimbra (Y) em função da população residente em Coimbra (X), usámos a reta de regressão $y = 5068{,}033\,x - 506168285{,}6$ para obter a previsão de 178446952,3 Kwh para o consumo de eletricidade em Coimbra para 2015, ano em que a população residente em Coimbra foi de 135085 habitantes.

Pretendendo agora obter um intervalo de previsão com nível de confiança 95% para o consumo doméstico de eletricidade em Coimbra em 2015, comecemos pelo cálculo do valor t^* que satisfaz

$$\mathrm{P}(-t^* \leq T \leq t^*) = 0{,}95,$$

onde T tem uma distribuição de Student $t(6-2) = t(4)$. Da tabela distribuição de Student tiramos

$$t^* = 2{,}776.$$

Um vez que $\bar{x} = 140102{,}0$ e $s_x^2 = 14129986{,}667$, das tabela apresentadas no Exemplo 12.7.1 (pág. 291) deduzimos que

$$\text{margem de erro} = 2{,}776 \times 2958399{,}403 \sqrt{1 + \frac{1}{6} + \frac{(135085 - 140102)^2}{(6-1) \times 14129986{,}667}} = 10134839{,}16.$$

Assim o intervalo de previsão de nível 95% para o consumo doméstico de eletricidade em Coimbra em 2015 é de (em Kwh)

$$178446952{,}21 \pm 10134839{,}16 = [168312113{,}05\,;\ 188581791{,}37].$$

12.9 Intervalo de confiança para a média de Y quando $X = x$

Vimos no parágrafo anterior que, quando x é o valor observado para a variável X, podemos obter um intervalo de previsão para o valor tomado pela variável Y a partir da quantidade

$$\widehat{Y} = \widehat{a}x + \widehat{b}.$$

Atendendo a que, quando x é o valor assumido para a variável X, o modelo de regressão linear assume que a variável Y obedece à igualdade

$$Y = ax + b + \epsilon,$$

onde $\epsilon \sim N(0, \sigma)$, deduzimos, das propriedades da média estudadas em §4.4, que a média da variável Y, quando x é o valor assumido para a variável X, que vamos representar por μ_Y^x, para não confundir com a média μ_Y da variável Y, é dada por

$$\mu_Y^x = ax + b,$$

A expressão anterior leva-nos a pensar que a aproximação $\widehat{Y} = \widehat{a}x + \widehat{b}$ possa ser usada, não só como aproximação de Y, mas também como aproximação da média μ_Y^x. Com efeito, sob a hipótese de normalidade sobre do termo de erro ϵ, é possível mostrar que $\widehat{Y}$ possui uma distribuição normal com média μ_Y^x, e com desvio-padrão

$$\sigma_{\widehat{Y}} = \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}},$$

de onde deduzimos que

$$z = \frac{\widehat{Y} - \mu_Y^x}{\sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}} \sim N(0,1).$$

Tal como vimos para a construção do intervalo de previsão feita no parágrafo anterior, sendo o desvio-padrão do termo de erro desconhecido, podemos proceder como nos intervalos de confiança para uma média (ver pág. 178), substituindo σ por $s_{\hat{\epsilon}}$ que, como vimos atrás, é uma aproximação para σ calculada a partir dos resíduos do modelo de regressão linear. Ao fazermos esta substituição e passarmos a usar a variável fulcral

$$t = \frac{\widehat{Y} - \mu_Y^x}{s_{\hat{\epsilon}}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}},$$

esta deixa de possuir uma distribuição normal standard, para passar a ter uma distribuição de Student com $n-2$ graus de liberdade. Este facto permite obter o seguinte intervalo de confiança para o valor médio da variável Y, quando $X = x$.

Intervalo de confiança para a média de Y quando $X = x$:

Nas condições gerais do modelos de regressão linear, um intervalo de confiança, com nível de confiança C, para a média de Y quando $X = x$, tem por extremidades:

$$\widehat{Y} \pm t^* \, s_{\hat{\epsilon}} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2}},$$

onde

$$\widehat{Y} = \widehat{a}x + \widehat{b},$$

t^* é tal que

$$\mathrm{P}(-t^* \leq T \leq t^*) = C$$

e T tem uma distribuição de Student $t(n-2)$.

Exemplo 12.9.1 Retomando o Exemplo 12.8.1 (pág. 296), vejamos como obter um intervalo de confiança de nível de confiança 95% para o tempo médio de corrida de atletas cuja capacidade aeróbica máxima é de 50 unidades (reparar que deixamos de estar interessados no que acontece com uma atleta em particular, para nos interessarmos pela população de todas as atletas cuja capacidade aeróbica máxima é de 50 unidades).

Atendendo ao que fizemos no Exemplo 12.8.1, sabemos que $\widehat{Y} = 45{,}094$ minutos, $t^* = 2{,}179$, $\bar{x} = 52{,}994$, $s_x^2 = 17{,}163$ e $s_{\hat{\epsilon}} = 2{,}298$. Assim, a margem de erro do intervalo de confiança pedido é dada por

$$\text{margem de erro} = 2{,}179 \times 2{,}298 \sqrt{\frac{1}{14} + \frac{(50 - 52{,}994)^2}{(14-1) \times 17{,}163}} = 1{,}6728 \text{ minutos.}$$

O intervalo de confiança, de nível 95%, para o tempo de corrida de atletas com uma capacidade aeróbica máxima de 50 é dado por

$$45{,}094 \pm 1{,}6728 = [43{,}421\,;\, 46{,}767].$$

Exemplo 12.9.2 Retomando o Exemplo 12.8.2 (pág. 297), vejamos como obter um intervalo de confiança de nível de confiança 95%, para o valor médio do consumo doméstico de eletricidade (Kwh) em Coimbra, sempre que a população residente em Coimbra seja de 125000 habitantes.

Como $X = 125000$, obtemos

$$\widehat{Y} = 5068{,}033 \times 125000 - 506168285{,}6 = 127335839{,}40.$$

Um vez que $t^* = 2{,}776$, $\bar{x} = 140102{,}0$, $s_x^2 = 14129986{,}667$ e $s_{\hat{\epsilon}} = 2958399{,}403$ (ver Exemplo 12.8.2), a margem de erro do intervalo de confiança pedido é igual a

$$\text{margem de erro} = 2{,}776 \times 2958399{,}403 \sqrt{\frac{1}{6} + \frac{(135085 - 140102)^2}{(6-1) \times 14129986{,}667}} = 5938815{,}84,$$

sendo o intervalo de confiança dado por

$$127335839{,}40 \pm 5938815{,}84 = [121397023{,}56\,;\; 133274655{,}24]\,.$$

12.10 Bibliografia

Abraham, B., Ledolter, J. (1983). *Statistical Methods for Forecasting*, Wiley.

Daniel, W.W. (2009). *Biostatistics: a foundation for analysis in the health sciences*, Wiley.

Dawson, B., Trapp, R.G. (2004). *Basic & Clinical Biostatistics*, McGraw-Hill.

Moore, D.S., McCabe, G.P. (2006). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

Murteira, B.J.F. (1993). *Análise Exploratória de Dados. Estatística Descritiva*, McGraw-Hill.

Pestana, D.D., Velosa, S.F. (2002). *Introdução à Probabilidade e à Estatística*, Vol. I, Fundação Calouste Gulbenkian.

Tabelas

Tabela A: Números aleatórios

Tabela B: Distribuição normal standard

Tabela C: Coeficientes binomiais

Tabela D: Distribuição de Student

Tabela E: Distribuição do qui-quadrado

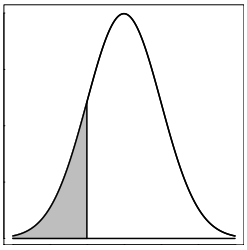
Tabela A

Números aleatórios

Linha / Coluna	01–05	06–10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
01	75965	99218	67035	41041	24104	10997	36233	83214	17221	86381
02	41562	08397	03436	82004	52285	36165	31697	99529	33287	96007
03	45963	59075	50719	41803	84602	40840	58654	90498	04876	81772
04	15685	82676	73539	93042	84761	44222	53712	30497	16060	96390
05	68434	58980	14495	62512	33939	93623	78112	67166	18067	63925
06	25819	80242	61173	44151	87993	75768	29300	29053	31549	89404
07	58147	51328	15756	87583	81546	53593	05210	96239	04489	06755
08	14768	31191	49670	76790	60509	51526	14080	92201	45112	28997
09	96757	60822	54662	75406	64175	40440	24609	47929	27493	66916
10	58166	71615	63933	87079	09758	12503	25232	81453	91695	07215
11	22507	41992	32756	04749	03173	68090	62222	06406	64830	54428
12	43260	52862	82414	72112	77661	72514	36852	80576	76928	49051
13	33800	07259	78554	84532	21914	85491	52543	54189	06862	34688
14	91092	47886	04958	21339	34520	75544	44643	88177	88283	11444
15	38990	07661	10674	06314	10238	54909	44733	84050	26184	58190
16	51084	84984	92876	16021	14823	28248	29927	41140	63241	96709
17	82364	79393	11314	15629	69393	49019	99136	60590	81122	63836
18	70509	84995	98511	17277	26948	10194	77428	41330	92843	06123
19	94988	80990	36878	61994	08783	18920	31530	31020	16693	55555
20	05374	78371	40393	32545	55225	58014	61008	47774	81511	95349
21	90513	07118	57057	90568	05056	14259	14966	26448	39535	33689
22	49444	15940	85582	20874	70424	05764	71326	84178	26384	61458
23	92944	35011	59391	97515	43182	54309	20115	55067	22651	74935
24	47965	22062	98300	86583	21586	66169	70777	39936	22453	44903
25	23471	04820	89156	88682	20475	72972	29677	87269	34959	64822
26	15252	83580	71048	82618	65250	21413	72998	17165	24638	52013
27	97822	94436	71870	78895	41015	86797	09591	78612	29316	48528
28	07908	59429	23872	05167	49670	32985	79270	02955	98886	08124
29	86354	48647	12649	65260	75953	56179	65590	04968	68033	09826
30	90286	53370	23683	78875	63477	77650	34053	71618	73242	62049
31	61308	07573	60875	14675	55980	15220	18148	94651	01289	79347
32	90298	30452	47152	05761	96314	29463	15444	37573	81097	39306
33	49370	80926	33287	71529	80090	82012	66194	27410	15333	11563
34	44453	53555	82291	71913	03937	34881	23578	93248	82102	37429
35	83376	01190	22389	17331	22432	76018	90227	83902	92421	44878
36	09053	22934	82405	28819	10263	31719	51967	28912	39489	00891
37	71297	63934	89685	30432	67115	12591	77207	06090	58026	66610
38	87709	31191	36957	76485	54366	02363	45115	04723	95080	85623
39	28764	46683	02814	41923	31840	92665	98375	82141	44436	87789
40	12106	19976	47485	06811	96639	22701	71381	99186	73322	92974
41	16070	00380	45273	47256	93035	22829	23631	74102	25753	19035
42	73876	40923	94658	82203	42828	13727	39117	85878	27383	17547
43	99071	89561	39140	92680	50789	09663	35333	42208	43757	85953
44	90378	37893	23956	20950	79345	12007	40788	61540	97382	01296
45	37561	17428	16994	75530	62701	01230	96853	96138	95495	97140
46	11391	49272	36911	21734	63012	60975	09638	78895	12204	32516
47	08909	26924	42306	18507	11032	47508	10611	63855	20851	57917
48	62975	12262	18289	34210	84079	13714	57645	16743	95114	05837
49	91109	12383	74149	66530	93604	00094	79689	50199	46360	22786
50	77948	45855	32491	35154	10046	17986	56351	20615	24863	99815

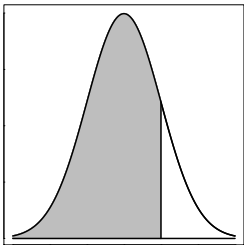
Tabela B

Distribuição normal standard



z

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
-3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
-3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
-3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
-3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
-3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
-3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641



z

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
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

Tabela C

Coeficientes binomiais C_k^n

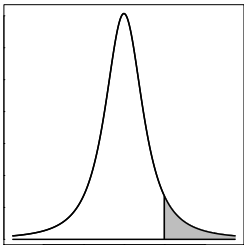
Coeficientes binomiais C_k^n

$k \setminus n$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2		1	3	6	10	15	21	28	36	45	55	66	78	91	105
3			1	4	10	20	35	56	84	120	165	220	286	364	455
4				1	5	15	35	70	126	210	330	495	715	1001	1365
5					1	6	21	56	126	252	462	792	1287	2002	3003
6						1	7	28	84	210	462	924	1716	3003	5005
7							1	8	36	120	330	792	1716	3432	6435
8								1	9	45	165	495	1287	3003	6435
9									1	10	55	220	715	2002	5005
10										1	11	66	286	1001	3003
11											1	12	78	364	1365
12												1	13	91	455
13													1	14	105
14														1	15
15															1

$k \setminus n$	16	17	18	19	20	21	22	23	24	25
0	1	1	1	1	1	1	1	1	1	1
1	16	17	18	19	20	21	22	23	24	25
2	120	136	153	171	190	210	231	253	276	300
3	560	680	816	969	1140	1330	1540	1771	2024	2300
4	1820	2380	3060	3876	4845	5985	7315	8855	10626	12650
5	4368	6188	8568	11628	15504	20349	26334	33649	42504	53130
6	8008	12376	18564	27132	38760	54264	74613	100947	134596	177100
7	11440	19448	31824	50388	77520	116280	170544	245157	346104	480700
8	12870	24310	43758	75582	125970	203490	319770	490314	735471	1081575
9	11440	24310	48620	92378	167960	293930	497420	817190	1307504	2042975
10	8008	19448	43758	92378	184756	352716	646646	1144066	1961256	3268760
11	4368	12376	31824	75582	167960	352716	705432	1352078	2496144	4457400
12	1820	6188	18564	50388	125970	293930	646646	1352078	2704156	5200300
13	560	2380	8568	27132	77520	203490	497420	1144066	2496144	5200300
14	120	680	3060	11628	38760	116280	319770	817190	1961256	4457400
15	16	136	816	3876	15504	54264	170544	490314	1307504	3268760
16	1	17	153	969	4845	20349	74613	245157	735471	2042975
17		1	18	171	1140	5985	26334	100947	346104	1081575
18			1	19	190	1330	7315	33649	134596	480700
19				1	20	210	1540	8855	42504	177100
20					1	21	231	1771	10626	53130
21						1	22	253	2024	12650
22							1	23	276	2300
23								1	24	300
24									1	25
25										1

Tabela D

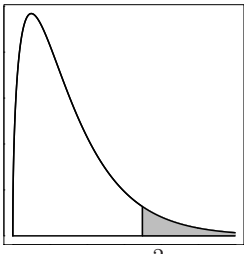
Distribuição de Student



k\área	0,2	0,1	0,05	0,025	0,02	0,01	0,005	0,001
1	1,376	3,078	6,314	12,706	15,895	31,821	63,657	318,309
2	1,061	1,886	2,920	4,303	4,849	6,965	9,925	22,327
3	0,978	1,638	2,353	3,182	3,482	4,541	5,841	10,215
4	0,941	1,533	2,132	2,776	2,999	3,747	4,604	7,173
5	0,920	1,476	2,015	2,571	2,757	3,365	4,032	5,893
6	0,906	1,440	1,943	2,447	2,612	3,143	3,707	5,208
7	0,896	1,415	1,895	2,365	2,517	2,998	3,499	4,785
8	0,889	1,397	1,860	2,306	2,449	2,896	3,355	4,501
9	0,883	1,383	1,833	2,262	2,398	2,821	3,250	4,297
10	0,879	1,372	1,812	2,228	2,359	2,764	3,169	4,144
11	0,876	1,363	1,796	2,201	2,328	2,718	3,106	4,025
12	0,873	1,356	1,782	2,179	2,303	2,681	3,055	3,930
13	0,870	1,350	1,771	2,160	2,282	2,650	3,012	3,852
14	0,868	1,345	1,761	2,145	2,264	2,624	2,977	3,787
15	0,866	1,341	1,753	2,131	2,249	2,602	2,947	3,733
16	0,865	1,337	1,746	2,120	2,235	2,583	2,921	3,686
17	0,863	1,333	1,740	2,110	2,224	2,567	2,898	3,646
18	0,862	1,330	1,734	2,101	2,214	2,552	2,878	3,610
19	0,861	1,328	1,729	2,093	2,205	2,539	2,861	3,579
20	0,860	1,325	1,725	2,086	2,197	2,528	2,845	3,552
21	0,859	1,323	1,721	2,080	2,189	2,518	2,831	3,527
22	0,858	1,321	1,717	2,074	2,183	2,508	2,819	3,505
23	0,858	1,319	1,714	2,069	2,177	2,500	2,807	3,485
24	0,857	1,318	1,711	2,064	2,172	2,492	2,797	3,467
25	0,856	1,316	1,708	2,060	2,167	2,485	2,787	3,450
26	0,856	1,315	1,706	2,056	2,162	2,479	2,779	3,435
27	0,855	1,314	1,703	2,052	2,158	2,473	2,771	3,421
28	0,855	1,313	1,701	2,048	2,154	2,467	2,763	3,408
29	0,854	1,311	1,699	2,045	2,150	2,462	2,756	3,396
30	0,854	1,310	1,697	2,042	2,147	2,457	2,750	3,385
40	0,851	1,303	1,684	2,021	2,123	2,423	2,704	3,307
50	0,849	1,299	1,676	2,009	2,109	2,403	2,678	3,261
60	0,848	1,296	1,671	2,000	2,099	2,390	2,660	3,232
70	0,847	1,294	1,667	1,994	2,093	2,381	2,648	3,211
80	0,846	1,292	1,664	1,990	2,088	2,374	2,639	3,195
90	0,846	1,291	1,662	1,987	2,084	2,368	2,632	3,183
100	0,845	1,290	1,660	1,984	2,081	2,364	2,626	3,174
150	0,844	1,287	1,655	1,976	2,072	2,351	2,609	3,145
200	0,843	1,285	1,653	1,972	2,067	2,345	2,601	3,131
500	0,842	1,283	1,648	1,965	2,059	2,334	2,586	3,107
1000	0,842	1,282	1,646	1,962	2,056	2,330	2,581	3,098
z	0,842	1,282	1,645	1,960	2,054	2,326	2,576	3,090

Tabela E

Distribuição do qui-quadrado



k\área	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.001
1	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.828
2	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.816
3	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.266
4	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.467
5	7.289	9.236	11.070	12.833	13.388	15.086	16.750	20.515
6	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.458
7	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.124
9	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	16.985	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.123
15	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.315
21	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.797
22	27.301	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	28.429	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	29.553	33.196	36.415	39.364	40.270	42.980	45.559	51.179
25	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.620
26	31.795	35.563	38.885	41.923	42.856	45.642	48.290	54.052
27	32.912	36.741	40.113	43.195	44.140	46.963	49.645	55.476
28	34.027	37.916	41.337	44.461	45.419	48.278	50.993	56.892
29	35.139	39.087	42.557	45.722	46.693	49.588	52.336	58.301
30	36.250	40.256	43.773	46.979	47.962	50.892	53.672	59.703
31	37.359	41.422	44.985	48.232	49.226	52.191	55.003	61.098
32	38.466	42.585	46.194	49.480	50.487	53.486	56.328	62.487
33	39.572	43.745	47.400	50.725	51.743	54.776	57.648	63.870
34	40.676	44.903	48.602	51.966	52.995	56.061	58.964	65.247
35	41.778	46.059	49.802	53.203	54.244	57.342	60.275	66.619
36	42.879	47.212	50.998	54.437	55.489	58.619	61.581	67.985
37	43.978	48.363	52.192	55.668	56.730	59.893	62.883	69.346
38	45.076	49.513	53.384	56.896	57.969	61.162	64.181	70.703
39	46.173	50.660	54.572	58.120	59.204	62.428	65.476	72.055
40	47.269	51.805	55.758	59.342	60.436	63.691	66.766	73.402

Referências bibliográficas

Abraham, B., Ledolter, J. (1983). *Statistical Methods for Forecasting*, Wiley.

Daniel, W.W. (2009). *Biostatistics: a foundation for analysis in the health sciences*, Wiley.

Dawson, B., Trapp, R.G. (2004). *Basic & Clinical Biostatistics*, McGraw-Hill.

Gomes, M.I., Barão, M.I. (1999). *Controlo Estatístico de Qualidade*, SPE.

Graça Martins, M.E., Cerveira, A.G. (1999). *Introdução às Probabilidades e à Estatística*, Universidade Aberta.

Levy, P. (1999). *Sampling of Populations: methods and applications*, Wiley.

Martins, M.E.G., Cerveira, A.G. (2000). *Introdução às Probabilidades e à Estatística*, Universidade Aberta.

McPherson, G. (1990). *Statistics in Scientific Investigation: its basis, application, and interpretation*, Springer.

Moore, D.S. (1985). *Statistics: concepts and controversies*, W.H. Freeman and Company.

Moore, D.S., McCabe, G.P. (2003). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

Moore, D.S., McCabe, G.P. (2006). *Introduction to the Practice of Statistics*, W.H. Freeman and Company.

Murteira, B.J.F. (1993). *Análise Exploratória de Dados. Estatística Descritiva*, McGraw-Hill.

Pestana, D.D., Velosa, S.F. (2002). *Introdução à Probabilidade e à Estatística*, Vol. I, Fundação Calouste Gulbenkian.

Tenreiro, C. (2009). *Estatística: notas de apoio às aulas*, Coimbra (disponível em www.mat.uc.pt/∼tenreiro/).

Vicente, P., Reis, E., Ferrão, F. (2001). *Sondagens: a amostragem como factor decisivo de qualidade*, Edições Sílabo.

Índice Remissivo