

Estatística com R: Aprenda Fazendo

Mais de 100 Exercícios Resolvidos
Aplicações com Dados Reais
1700 Linhas de código em 'R' para resolver
Exercícios e Aplicações

José Dias Curto

<https://diascurto.wixsite.com/sitedc/estatisticar>
dias.curto@iscte-iul.pt

É expressamente proibido reproduzir, no todo ou em parte, sob qualquer forma ou meio, **NOMEADAMENTE FOTOCÓPIA**, esta obra. As transgressões serão passíveis das penalizações previstas na legislação em vigor.

Visite a página do livro:

<https://diascurto.wixsite.com/sitedc/estatisticar>

FICHA TÉCNICA:

Título: Estatística com R: Aprenda Fazendo

Autor: José Joaquim Dias Curto

© José Joaquim Dias Curto

Direção de arte: Maria Inês Lopes

Impressão e acabamentos: GUIDE – Artes Gráficas, Lda

1ª edição, setembro de 2021

Depósito Legal: 485937/21

ISBN: 978-989-33-2076-1

Ser professor dá-me vida! Por isso, dedico este livro a todos os meus alunos. Muito obrigado por me ajudarem a ser feliz.

Índice

Prefácio	xiii
1 Para começar...	1
1.1 Género e Salário: são a mesma coisa?	1
1.1.1 Dados qualitativos	3
1.1.2 Dados quantitativos	4
1.2 Dados de séries cronológicas, seccionais e em painel	4
1.2.1 Variáveis discretas e contínuas	5
1.3 Apresentação dos dados: em tabelas, por classes...	6
1.4 R e RStudio: instalação e primeiros passos na utilização	7
1.5 Ficheiros de dados	13
1.6 Breve caracterização dos funcionários da PIXIES	13
1.7 Vendas da empresa ‘Walkman’	18
1.8 Género, Cidade, Marca de automóvel e Cor dos olhos	21
1.9 Autoavaliação a Estatística e Matemática	26
1.10 À procura de dados para analisar	30
2 Estatística Descritiva	31
2.1 Medidas de Tendência Central	33
2.2 Medidas de Tendência não Central	39
2.3 Medidas de Dispersão	40
2.4 Coeficientes de Assimetria e de Curtose	46
2.5 Representações gráficas	52
2.6 Quanto vale uma empresa de SI em Portugal?	56
2.6.1 Indicação sobre o VNme e RLme por empresa	60
2.6.2 Para destacar as diferenças entre empresas	61
2.6.3 O peso das ‘anormalidades’	62
2.7 Vendas trimestrais da APPLE	63
2.8 Rendibilidade e risco nos mercados financeiros	66

3	Distribuições Teóricas	73
3.1	Não é certo? Mas tem uma probabilidade...	75
3.2	Variáveis aleatórias	78
3.3	Caraterísticas das distribuições de probabilidade	82
3.3.1	Média ou valor esperado	83
3.3.2	Variância	84
3.3.3	Covariância e coeficiente de correlação simples	85
3.4	Distribuições de probabilidade teóricas	87
3.4.1	Distribuições: o que são e para que servem?	87
3.4.2	Todos têm a mesma probabilidade	89
3.4.3	Sucessos/Insucessos: quantos são, quantos são?	90
3.4.4	Com reposição: distribuição hipergeométrica	96
3.4.5	Limitar no espaço ou no tempo	97
3.4.6	Distribuição Normal	98
3.4.7	Distribuição do qui-quadrado	106
3.4.8	Distribuição t de Student	108
3.4.9	Distribuição F de Snedecor	110
3.5	Simulação de vendas	112
3.5.1	Distribuição Uniforme: $U(a, b)$	113
3.5.2	Distribuição Triangular: $T(a, b, c)$	117
3.5.3	Distribuição Exponencial	119
3.5.4	Simulação de vendas da empresa WILDNOTHING	123
3.6	Momentos e parâmetros de ordem de uma distribuição	125
4	Amostragem e Estimação	127
4.1	População e amostra	128
4.2	Métodos de seleção das amostras	130
4.3	Parâmetros e estatísticas	134
4.4	Distribuições por amostragem (DPA)	134
4.4.1	Teorema do Limite Central	137
4.4.2	DPA da média amostral	138
4.4.3	DPA da variância amostral	139
4.4.4	DPA da proporção amostral	139
4.5	Estimação paramétrica	140
4.6	Propriedades em amostras finitas	141
4.7	Propriedades em grandes amostras	145
4.8	Estimação por intervalos	147
4.9	Dimensão de uma amostra aleatória simples	152

4.9.1	Populações de Bernoulli	153
4.9.2	Populações normais	155
5	Testes de Hipóteses	157
5.1	Como é que funciona um teste de hipóteses?	158
5.1.1	Abordagem do intervalo de confiança	160
5.1.2	A abordagem dos testes (ou ensaios) de hipóteses	161
5.1.3	‘Aceitar’ ou ‘Não Rejeitar’, que terminologia utilizar?	169
5.1.4	Tipo de erros	169
5.2	Testes para comparar médias e variâncias	171
5.2.1	Teste à igualdade de duas médias – amostras independentes	171
5.2.2	Testes à igualdade de variâncias	174
5.2.3	Teste à igualdade de duas médias – amostras emparelhadas	176
5.2.4	Testes não paramétricos	178
5.3	Análise de variância simples (ANOVA)	180
5.3.1	Análise de variância simples não paramétrica	184
5.4	O impacte da formação nas vendas dos lojistas	186
5.5	Testes de aderência ou da bondade do ajustamento	193
5.5.1	Teste do Qui-quadrado	194
5.5.2	Teste de Kolmogorov-Smirnov (KS)	197
5.5.3	Teste de Anderson-Darling (AD)	199
5.5.4	Testes aos coeficientes de assimetria e de curtose	200
5.5.5	Teste de Jarque-Bera (JB)	201
5.5.6	Normalidade das taxas de rendibilidade das ações FATANG	201
5.5.7	EURIBOR a 6 meses	204
5.6	Considerações finais	206
6	Medidas de Correlação e de Associação	207
6.1	Análise de correlação paramétrica	208
6.1.1	Diagrama de dispersão	209
6.1.2	Covariância	211
6.1.3	Coefficiente de correlação simples	212
6.1.4	Teste à significância de r_{XY} quando $\rho_{XY} = 0$	215
6.1.5	Teste à significância de r quando $\rho \neq 0$	217
6.1.6	Limitações da correlação simples	219
6.1.7	Taxas de rendibilidade	220
6.2	Análise de correlação não paramétrica	222
6.2.1	Nível de Escolaridade e Performance	224
6.3	Teste de independência do qui-quadrado	227

6.4	Medidas de Associação	230
6.4.1	Coeficiente C de Pearson	231
6.4.2	Coeficiente T de Tschruprow	231
6.4.3	Coeficiente V de Cramer	231
6.4.4	Género e Marcas de automóveis	232
7	Modelo de Regressão Linear Simples	233
7.1	O modelo de regressão linear simples	235
7.2	Função de regressão	238
7.3	O método dos mínimos quadrados ordinários	244
7.4	Coeficiente de determinação (r^2)	247
7.5	Hipóteses do modelo de regressão linear	249
7.6	Distribuição por amostragem dos estimadores OLS	253
7.7	Inferência Estatística no modelo	257
7.8	Predição no modelo	259
7.9	Consumo de Eletricidade	261
8	Relações não Lineares	267
8.1	Escala e unidades de medida	268
8.2	Modelo logarítmico ou modelo linear-log	270
8.3	Modelo log-linear (ou modelo semilogarítmico)	273
8.4	Modelo duplo-log (ou modelo log-log)	277
8.5	Relações inversas	280
8.6	Polinómios	282
8.7	Comparação dos valores de R^2 entre modelos	282
	Apêndices	289
	Apêndice 8.A Efeitos marginais e Elasticidades	289
	Apêndice 8.B Funções exponencial e logarítmica	290
	8.B.1 Propriedades das funções	292
	8.B.2 Propriedades das derivadas das funções	292
	8.B.3 O conceito de elasticidade	293
9	Modelo de Regressão Linear Múltipla	295
9.1	Os estimadores OLS	296
9.1.1	Coeficientes estandardizados	299
9.1.2	Coeficientes de correlação part e parcial	301
9.2	Coeficiente de determinação corrigido (ou ajustado)	302
9.3	Mais sobre o teste F	304

9.3.1	Poder explicativo e significância estatística	307
9.3.2	Igualdade de coeficientes	308
9.4	Erros de especificação	309
9.4.1	O problema da multicolinearidade	311
9.4.2	Forma funcional da relação	316
9.4.3	Variáveis explicativas endógenas	318
9.4.4	Teste de Chow	319
9.4.5	Heterocedasticidade	320
9.4.6	Autocorrelação	325
	Bibliografia	335

Lista de Tabelas

1.1	Dados de Trabalhadores	2
1.2	Tabela de Frequências: desempenho, habilit e genero	16
1.3	Tabela de Frequências: idade e salario	16
1.4	Vendas anuais e IPC – Walkman	19
1.5	Variações absoluta, relativa e percentual – Vendas	20
1.6	Frequências absolutas e relativas – Cidade	22
1.7	Marca de carro por Cidade	24
1.8	Teste de Independência do Qui-quadrado	26
1.9	Tabela de frequências – Estatística e Matemática	28
2.1	Volume de Negócios (Vendas: VN), Resultados Líquidos (RL) e Número de Trabalhadores (FT) das 30 Maiores Empresas Portuguesas de Sistemas de Informação (Ano 2011)	57
2.2	Medidas de Estatística Descritiva – Vendas e Resultados Líquidos	60
2.3	Medidas de Estatística Descritiva – Vendas e Resultados Líquidos	65
2.4	Cotações e Rendibilidades diárias do PSI20, SONAE e REN	68
2.5	Medidas de Estatística Descritiva	70
2.6	Análise de <i>outliers</i> : Resultados	71
3.1	Conceito frequencista de probabilidade	77
3.2	Função de distribuição	110
3.3	Quantis	110
6.1	Coefficiente de Spearman: ordenação dos valores	225
6.2	Avaliação e Cidade - Teste de independência	229
7.1	Número e receita dos turistas	233
7.2	Salários por anos de escolaridade	239
7.3	Média condicional, valor estimado, erros e resíduos	242
7.4	Cálculos preliminares para estimar β_1 e β_2	246

7.5	Modelo de regressão linear simples: resultados	262
8.1	Efeitos marginais decrescentes	286
8.A.1	Formas funcionais, efeitos marginais e elasticidades	289
9.1.1	Modelo de regressão linear múltipla: resultados	299
9.4.1	Coefficientes de correlação <i>part</i> e parcial	314
9.4.2	Tolerância e VIF	314
9.4.3	Correlação e Multicolinearidade	315
9.4.4	Modelo de regressão linear múltipla: sem variável rendimento	316
9.4.5	RL <i>vs</i> VN e TRAB	323
9.4.6	Teste de White	324

Lista de Figuras

1	Rendibilidade e Risco de um Fundo de Ações	xv
1.1	Vista inicial do RStudio	8
1.2	Janela de comandos	9
1.3	Executar comandos/instruções em RStudio	10
1.4	Importação de dados ‘indicando o caminho’	12
1.5	Diagrama de barras: Desempenho	17
1.6	Histograma: Salário	18
1.7	Gráfico de barras – Marca do Carro	23
1.8	Gráfico de barras sobrepostas – Marca de carro por Género	24
1.9	Marca de Carro por Cidade	25
1.10	Caixa de bigodes – Boxplot	27
1.11	Gráfico circular – Estatística	28
1.12	Tabela de dupla entrada – Estatística e Matemática	29
2.1	Outliers severos e moderados	45
2.2	Tipos de assimetria	47
2.3	Efeito do ‘peso’ das caudas na assimetria	49
2.4	Tipos de curtose	50
2.5	Histograma	53
2.6	Histograma	54
2.7	Caixa-de-bigodes – Boxplot	55
2.8	Histograma: PSI20	69
3.1	Frequências relativas do PSI20, SONAE e REN	77
3.2	Função densidade de probabilidade	80
3.3	Funções densidade de probabilidade: $f(x)$, e de distribuição: $F(x)$, de uma variável aleatória com distribuição normal standardizada	82
3.4	Distribuição das vendas (milhares de euros)	88

3.5	Função densidade de probabilidade (ou curva de Gauss) de uma variável aleatória com distribuição normal	99
3.6	RETPSI20 – Histograma com sobreposição da curva Normal	106
3.7	Distribuição qui-quadrado: função densidade de probabilidade	107
3.8	Distribuição t de Student: função densidade de probabilidade para diferentes graus de liberdade	109
3.9	Distribuição F : função densidade de probabilidade para diferentes graus de liberdade	111
3.10	Função densidade de probabilidade $U(2,4)$	114
3.11	Histograma: distribuição $U(0,1)$	116
3.12	Histograma: distribuição $N(0,1)$	117
3.13	Função densidade de probabilidade $T(0, 12, 4)$	118
3.14	Histograma: distribuição $T(0, 50, 10)$	120
3.15	Função densidade de probabilidade $\text{Exp}(\lambda = 0.5)$	120
3.16	Histograma: distribuição $\text{Exp}(0.5)$	122
4.1	Histograma com sobreposição da curva normal	136
4.2	Histograma com sobreposição da curva normal	137
5.1	Regiões de Não Rejeição (RNR) e de Rejeição (RR)	162
5.2	Testes Unilateral Esquerdo e Unilateral Direito	162
5.3	Regiões de Não Rejeição (RNR) e de Rejeição (RR)	165
5.4	Cálculo da probabilidade associada ao valor de t	166
5.5	Região de Não Rejeição (RNR) e de Rejeição (RR)	168
5.6	Índices de cumprimento antes, durante e depois da formação	186
6.1	Diagramas de dispersão	210
6.2	Diagrama de dispersão	211
6.3	Diagrama de dispersão	219
6.4	Exemplo de correlação espúria	220
6.5	Diagrama de dispersão e coeficiente de correlação simples	221
7.1	Evolução da Receita e do N° de Turistas	234
7.2	Diagrama de dispersão	235
7.3	Retas ‘alternativas’	236
7.4	Distribuição condicional dos salários por anos de escolaridade	240
7.5	Funções de regressão da população e da amostra	243
7.6	Média condicional dos erros	250
7.7	Variância condicional dos erros (homocedasticidade)	251
7.8	Variância condicional dos erros (heterocedasticidade)	252

7.9	Ausência de autocorrelação	253
7.10	Resíduos e Normalidade	266
8.1	yields - maturidade	273
8.2	PIB Norte-Americano	276
8.3	Modelo log-log	279
8.4	Risco de uma carteira de títulos	281
8.5	Funções Linear e Quadrática	285
8.B.1	Função Exponencial	291
8.B.2	Função Logarítmica	291
9.4.1	Resíduos ao quadrado <i>vs</i> valores estimados de RL	323
9.4.2	Autocorrelação dos erros	327
9.4.3	Regiões da estatística de Durbin-Watson	329

Prefácio

“*Sem estatísticas andamos perdidos na sociedade em que vivemos.*” (Maria João Valente, diretora da PORDATA, (ECO, 08/06/2017).

Este livro é sobre Estatística (medidas, gráficos, amostra, população, testes de hipóteses, correlação, regressão, etc.) e tem uma orientação clara para as aplicações (são mais de 100 exercícios resolvidos e várias aplicações com dados reais). No início de cada capítulo/secção explicam-se resumidamente os conceitos e ‘parte-se’ de imediato para as aplicações, recorrendo diretamente ao R/RStudio para se obterem os resultados. De seguida procede-se à interpretação dos mesmos, destacando a importância da Estatística na análise de dados. Portanto, ‘o que é?’ (resumidamente e ‘sem muitas fórmulas’), ‘para que serve?’, ‘como implementar em R/RStudio?’ e ‘que conclusões?’, são estas as perguntas a que o livro pretende dar resposta.

Qual a importância deste livro e o que é que o diferencia dos demais? Há muitos livros sobre Estatística, mas não abundam os livros (em língua Portuguesa) de Estatística com utilização do R/RStudio. Portanto, este é o principal contributo: ‘ensinar’ Estatística com recurso permanente ao R/RStudio. Um outro fator diferenciador é o foco do livro: como fazer e como interpretar, tentando dar resposta a mais de 100 exercícios/aplicações que se resolvem ao longo dos vários capítulos.

E agora um pouco de contexto... Todos os dias recorremos à Matemática e à Estatística, mesmo que não nos demos conta. No supermercado utilizamos a *soma* para saber quanto gastámos e a *diferença* para calcular o troco. Na bomba de gasolina *multiplicamos* o preço pelos litros de combustível para saber quanto é que vão custar os próximos quilómetros a andar de carro. Em casa *dividem-se* as 8 pastilhas pelos quatro filhos do casal (apesar do mais velho reclamar uma quantidade *maior*...).

Mas a Matemática não se esgota nestas coisas simples da vida. Muito do ‘bem bom’ que temos hoje, e que aumentou de forma considerável nos últimos anos, deve-se à Matemática. Sim, não tenham dúvidas, mas como não estamos aqui para destacar as ‘benfeitorias’ da Matemática, sugiro que leiam os livros de Guillen (1995) e Stewart (2003) para me darem razão¹.

Apesar de não ser fácil separar a Estatística da Matemática, o nosso enfoque é a Estatística e o grande objetivo é evidenciar como ela ‘faz parte das nossas vidas’.

Começamos pelo Totoloto (apesar do Euromilhões estar na moda). “Joguei outra vez esta semana e... nada! Não tenho sorte nenhuma.” Pois... a *probabilidade* de sair o totoloto, por cada aposta jogada, é de 1 em 13983816, ou seja, 0.00000007151, e ‘é quase

¹Já Galileu (1564-1642) no distante século XVII acreditava que a Matemática seria a chave para a compreensão do universo. Newton, Bernoulli, Faraday, Clausius e Einstein, entre tantos outros, viriam a dar-lhe razão. Para Galileu, o grande livro do universo estava escrito em linguagem Matemática.

impossível' ser 'bafejado' pela sorte (a Santa Casa, e os que dela dependem, que não nos ouçam). Aliás, há mesmo quem diga que é maior a *probabilidade* de ser atropelado por um carro do que sair o totoloto. Mas há também quem diga "que ele sai, sai!" Portanto, não é um *acontecimento impossível*.

Quando vamos ao futebol queremos que o nosso clube ganhe. Mas nunca sabemos antes do jogo se tal vai acontecer. Para a vitória ser *certa*, sempre se pode 'comprar' o árbitro... Mas, independentemente do árbitro e da capacidade dos nossos jogadores, há sempre fatores *aleatórios* (imprevisíveis, mas que ainda assim têm uma probabilidade de acontecer) que influenciam o resultado (por exemplo, o nosso melhor avançado lesionou-se e a equipa não marcou golos). Portanto, quando se entra no estádio a *esperança* é grande mas a *incerteza* no resultado ainda é maior.

Probabilidade, aleatório e acontecimento são termos comuns que têm fundamento na Estatística, como se explica mais adiante. *Certo, incerto e esperança* também não lhes são estranhos. Mas não se fica por aqui...

Quando pergunta ao filho (ou ao amigo) a nota que teve no exame de física e ele lhe responde: 12, não há razões para grande entusiasmo. Mas se ele disser que a *média* da turma foi 6, aí a 'coisa' muda de figura e até lhe pode dar um abraço de parabéns!

A *média* é talvez o melhor exemplo da nossa 'convivência' com a Estatística pela frequência com que é utilizada. "Quantos quilos tem este saco de laranjas?" "O fornecedor disse que o peso médio de cada saco são dois quilos." "Em média quanto tempo demoras a chegar à escola?" "Cerca de 30 minutos." Até o banco regista o saldo médio da sua conta bancária...

E já que falamos de bancos, o que vai fazer com os 5000 euros que poupou no último ano? Se não os gastar, sempre pode fazer uma aplicação que lhe proporcione algum rendimento. Na hora de decidir revela-se o quanto é afoito e destemido (aquilo que em Finanças se designa por *aversão ao risco*). Se optar por um depósito a prazo, o banco dá-lhe 2% ao fim de um ano (que representam 100 euros). Se comprar obrigações da empresa BAUHAUS, a taxa do cupão é 4% (o dobro em euros: 200) e pode também comprar, à cotação de 5 euros, 1000 unidades de participação (UP) do fundo de ações CURE de um banco a operar em Portugal (a informação que consta do prospeto do fundo, e que nos interessa, é apresentada na Figura 1).

Neste caso não é possível antecipar o valor do rendimento pois depende da cotação das UP no momento da venda: pode ganhar e ter uma *mais-valia*, perder e incorrer numa *menos-valia* ou 'ficar em casa' se a cotação das UP estiver acima, abaixo ou nos 5 euros, respetivamente. Se a cotação estiver nos 6 euros, por exemplo, tem uma rendibilidade de 20%. Fantástico, vai ganhar 1000 euros!... Mas também pode acontecer o contrário e o preço baixar para os 4 euros. Neste caso perde 1000 euros e, ao final de um ano, dos 5000 euros iniciais restam apenas 4000. Ou seja, do depósito a prazo para as ações o *risco* do seu investimento vai aumentando.

O termo *risco* pode até nem lhe ser familiar, mas percebe, de certeza, que a possibilidade de ter um rendimento elevado é maior nas ações. Aliás, se atentar na informação do prospeto, a rendibilidade do fundo foi de 27.34% e 8.47% nos últimos 12 e 24 meses, respetivamente, bem acima dos 2% e 4% oferecidos pelo depósito bancário e pelas obrigações. Mas leia também a nota que se apresenta: *as rendibilidades divulgadas representam dados passados, não constituindo garantia de rendibilidade futura*. Aliás,

Figura 1: Rendibilidade e Risco de um Fundo de Ações

Data Cotação:	Cotação UP:	Rentab. 12M ² : 27,34 %	Rentab. 24M ² : 8,47 %
Data Rentab. e Risco:		Risco 12M: 13,55 %	Risco 24M: 17,26 %

¹ Classe de Risco referente a 12 meses:

Desvio-padrão anualizado (%)	Classe de risco	Escalão de risco
[0 ; 1,5[1	risco baixo
[1,5 ; 5[2	risco médio baixo
[5 ; 10[3	risco médio
[10 ; 15[4	risco médio alto
[15 ; 20[5	risco alto
>=20	6	risco muito alto

² Estas rentabilidades não consideram a retenção de IRS que existe no resgate

NOTAS:

¹ - As Rendibilidades divulgadas representam dados passados, não constituindo garantia de rentabilidade futura, porque o valor das unidades de participação pode aumentar ou diminuir em função do nível de risco que varia entre 1 (risco mínimo) e 6 (risco máximo).

também está escrito que o valor das UP pode cair e... neste caso incorre numa perda. Portanto, é mais ‘arriscado’ aplicar os 5000 euros no fundo de ações, pois com o depósito a prazo ou as obrigações estão garantidos o capital inicial mais 100 ou 200 euros de juros, respetivamente (admitindo que o banco e a empresa BAUHAUS não vão à falência). Os financeiros chamam *risco* à possibilidade ‘da coisa correr mal’ e de perder parte (ou até mesmo a totalidade) do capital investido. Nos casos do BES e do BANIF, por exemplo, os acionistas perderam tudo...

Ora bem, o risco ‘é das Finanças’¹ mas ‘são da Estatística’ as formas de o quantificar. A medida mais utilizada é o *desvio-padrão*, como se apresenta mais adiante, e antes de ser uma medida de risco é uma medida de *dispersão*, designação dada pela Estatística a um conjunto de medidas para aferir sobre a variação do fenómeno em causa. Se ‘olhar’ novamente para a informação do prospeto, e mesmo que não saiba o que é o desvio-padrão (mas que ficará a saber mais adiante), as colunas *Classe de risco* e *Escalão de risco* são bem claras sobre o ‘perigo’ que é investir no fundo de ações. Nos últimos 2 anos o risco foi superior a 17% (é como se pudéssemos dizer, apesar de não ser totalmente correto, que o valor das UP tanto pode subir como descer 17%) e está na penúltima classe de risco (*risco elevado*). Para melhor contextualizar os 17%, e admitindo que o banco do depósito a prazo e a empresa BAUHAUS não vão à falência, ao fim de um ano receberá os 5000 euros mais os juros vencidos. Portanto, o risco é 0%, que compara com os 17% do fundo de ações.

Lá estão o risco e a Estatística por detrás de uma decisão importante que pode tomar: comprar ou não as unidades de participação daquele fundo.

E para não me tornar chato ao ‘defender a minha dama’ deixo apenas mais três exemplos da nossa familiaridade com a Estatística. “No dia 9 de agosto a temperatura em Vilamoura atingiu os 40º, coisa muito rara nos últimos 20 anos”. Em Estatística estes valores pouco frequentes merecem uma atenção especial e designam-se por *outliers*¹,

¹Da meteorologia, da medicina, da engenharia...

¹Apesar de ser um termo em inglês, e por aparecer de forma recorrente no texto, a partir de agora a palavra

se quisermos utilizar o termo em inglês, ou valores extremos. “A amplitude térmica (AT) diária na Beira Baixa é muito elevada”. A AT (não confunda com Autoridade Tributária...) é a diferença entre as temperaturas máxima e mínima diárias e na Estatística esta diferença constitui também uma medida de dispersão designada por *intervalo de variação*. “Na minha empresa cerca de 95% dos trabalhadores ganham entre 1200 e 3800 euros”. O que quer dizer que o salário dos restantes 5% é inferior a 1200 ou superior a 3800 euros. Não sei se o fez, mas para chegar à conclusão este ‘patrão’ pode ter recorrido à *distribuição normal*, mais um ‘instrumento’ muito importante da Estatística.

E podíamos continuar... mas penso que já os convenci da ‘convivência’ e da ‘importância’ da Estatística no nosso dia a dia. E prometo que vai descobrir muitas outras ‘coisas’ onde a Estatística já é, ou virá a ser, importante para a sua tomada de decisão. A aplicação dos 5000 euros já foi um princípio, mas para já ‘vamos’ à estrutura deste livro.

No próximo capítulo começa-se (de forma ligeira) a falar dos dados, o objeto da Estatística: classificação e formas de apresentação. Esclarece-se sobre os vários tipos de variáveis, instalam-se e preparam-se o R e o RStudio para utilização futura. Calculam-se e interpretam-se as frequências absolutas e relativas, o valor da moda e constrói-se o primeiro histograma.

No capítulo 2 introduzem-se e aplicam-se as medidas de Estatística Descritiva para analisar as vendas e os resultados líquidos da Apple e das 30 maiores empresas de sistemas de informação em Portugal e para avaliar o risco e a rentabilidade do índice PSI20 e das ações das empresas SONAE e REN.

No capítulo 3 discutem-se e aplicam-se os instrumentos estatísticos para lidar com os fenómenos de natureza aleatória, introduzindo os conceitos de probabilidade, variável aleatória e distribuição de probabilidade. A seguir descrevem-se com algum detalhe e aplicam-se as distribuições teóricas mais importantes: Binomial, Poisson, Normal, t-Student, F-Snedecor, etc. O capítulo termina com a simulação das vendas de uma empresa considerando-se as distribuições Triangular, Normal e Exponencial para modelo probabilístico teórico das quantidades vendidas de três produtos.

A distinção entre amostra e população constitui o mote do capítulo 4. Se não se consegue chegar a toda a população, pelo menos pode recolher-se uma amostra, o mais representativa possível daquela população. Explica-se também a diferença entre parâmetro e estatística e fala-se de distribuições por amostragem das estatísticas mais importantes: média, variância e proporção amostrais, não esquecendo o teorema do limite central (que ajuda que ele nos dá!...). A partir das distribuições por amostragem é possível deduzir intervalos de confiança para os parâmetros e calcular também a dimensão de uma amostra aleatória simples tendo em conta a dimensão da população (se conhecida), o nível de confiança e o erro amostral em que se pretende incorrer.

O capítulo 5 trata dos testes de hipóteses, começando por apresentar as abordagens do intervalo de confiança e dos testes de hipóteses na realização de inferências sobre determinada população. Distinguem-se ensaios de significância de testes de hipóteses e explica-se o significado dos erros tipo I e II na tomada de decisão. Por último, procede-se à descrição e aplicação de alguns testes, nomeadamente o teste *t* para a diferença de médias com o propósito de avaliar o efeito das ações de formação nas vendas dos lojistas

outlier vai deixar de ser apresentada em itálico.

de uma empresa portuguesa, e os testes mais populares (KS: Kolmogorov-Smirnov e JB: Jarque-Bera) para aferir sobre a normalidade da distribuição de uma variável aleatória. As aplicações baseiam-se nas taxas de rendibilidade das 6 ações sob o acrónimo FATANG e na taxa de juro Euribor a 6 meses.

Nos quatro capítulos seguintes ‘andamos às voltas’ com a relação entre variáveis. Se o preço de um bem aumentar, qual é o efeito expectável sobre a quantidade vendida? Bem, se não for um daqueles bens mesmo essenciais (pão, por exemplo), as vendas devem baixar (que se cuidem as empresas fornecedoras). No capítulo 6 apresentam-se e aplicam-se as análises de correlação paramétrica e não paramétrica e as medidas de associação. Com a primeira ‘consegue-se’ avaliar o tipo (linear ou não linear), o sentido (direta ou inversa) e a intensidade (forte ou fraca) da relação entre variáveis quantitativas. Para isso costuma recorrer-se a uma representação gráfica (diagrama de dispersão) e a uma medida (coeficiente de correlação simples). A análise de correlação não paramétrica ‘visa’ a relação entre variáveis que admitem como nível de medida mais restrito a escala ordinal. As medidas de associação são utilizadas para concluir sobre a relação entre variáveis qualitativas nominais.

O modelo de regressão linear simples é discutido e apresentado no Capítulo 7. A análise de regressão linear, quando comparada com a análise de correlação paramétrica, e por ser mais completa, permite avaliar como é que a variação de uma variável impacta na variação da outra (se o preço do pão aumentar 10 cêntimos, qual é a redução expectável no valor da vendas). Neste capítulo discute-se ainda o método dos mínimos quadrados ordinários, as propriedades dos estimadores, os pressupostos do modelo de regressão linear simples e a Inferência Estatística através dos testes t e F .

Uma vez que as relações não lineares são também bastante comuns entre variáveis de natureza económica e financeira, no Capítulo 8 são analisadas outras funções de regressão que, apesar de constituírem relações não lineares na sua forma original, podem ser convertidas em funções lineares nos parâmetros através de transformações adequadas (a transformação logarítmica é a mais utilizada). Para isso são propostas formas funcionais alternativas para a função de regressão conhecidas vulgarmente por lin-lin, log-lin, lin-log e log-log. Também se analisam relações polinomiais e inversas bem como a alteração na escala dos dados e o seu impacte nas estimativas dos mínimos quadrados ordinários.

E o modelo de regressão linear múltipla (capítulo 9) põe fim a esta pequena ‘odisseia’... É um modelo mais completo (com mais variáveis explicativas) para ‘responder’ melhor à complexidade da vida real. Fala-se outra vez do método dos mínimos quadrados ordinários, do R^2 , dos testes t e F e introduzem-se novas medidas: coeficiente de determinação ajustado, coeficientes estandardizados, coeficientes de correlação parcial, etc. Por fim, avaliam-se de forma mais formal os pressupostos do modelo de regressão linear através dos procedimentos estatísticos mais comuns (quase sempre baseados em testes de hipóteses).

E mesmo mesmo a acabar... faz-se um apanhado das referências bibliográficas. Espero não me ter esquecido de nenhuma, mas se tal acontecer, as minhas sinceras desculpas a quem ficou injustamente de fora.

