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

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

Lista de Tabelas

Lista de Figuras

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 razão[1].

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

[1] Já 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 útimo 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 uma 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

[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:
1º - As Rendibilidades divulgadas representam dados passados, não constiuindo garantia de rendibilidade futura, porque o valor das unidades de participação pode aumentar ou diminuir em função do nível de risco que varia entre 1 (risco minimo) e 6 (risco maximo).

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],

[1] Da metereorologia, da medicina, da engenharia...
[1] Apesar 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 rendibilidade 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.