Estatística Aplicada

Parte II

Lino Costa e Pedro Oliveira

2018 (versão 3.0)

Conteúdo

Capítulo 1

Distribuições Amostrais

1.1 Introdução

A Estatística pretende retirar conclusões a partir de conjuntos de medidas ou observações que constituem um subconjunto, amostra, de um conjunto mais vasto designado por população. Assim, uma sondagem de opinião política recolhe uma amostra da população de todos os votantes, da mesma forma que em controlo de qualidade se recolhe uma amostra da população de todas as peças produzidas num determinado período. No entanto, este subconjunto será definido como uma amostra aleatória retirada de uma população, se cada elemento da população tem igual probabilidade de ser selecionado.

Contudo, as n medidas de uma amostra podem ser vistas como n observações de variáveis aleatórias, $x_1, x_2, \ldots, x_n$. Uma estatística é uma medida numérica calculada a partir dos dados amostrais (como, por exemplo, a média aritmética $\bar{x}$ ou a variância). Por outro lado, um parâmetro é uma medida numérica de uma população, por exemplo, a media μ ou a variância σ^2. Como tal, a média aritmética, $\bar{x}$, a variância, s^2, ou outras estatísticas são funções de variáveis aleatórias, que serão usadas para fazer inferências acerca dos parâmetros populacionais. Nesse sentido, a Estatística Inferencial é o conjunto de procedimentos que permitem, a partir de uma amostra, fazer inferências para a população. Por exemplo, a partir de uma amostra de portugueses pretender inferir o rendimento médio da população. Contudo, as medidas ou observações vão variar de amostra para amostra e, como resultado, também a estatística irá variar de acordo. Assim, é necessário determinar ou aproximar a distribuição amostral da estatística.

Definição 1.1.1: Amostra aleatória

Se $x_1, x_2, \ldots, x_n$ são variáveis aleatórias independentes e identicamente distribuídas, então constituem uma amostra aleatória de uma população infinita caracterizada pela sua distribuição conjunta.

Definição 1.1.2: Média e variância amostrais

Se $x_1, x_2, \ldots, x_n$ constituem uma amostra aleatória, então $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ é a média amostral e, $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$ é a variância amostral.

1.2 A distribuição da média

Se $x_1, x_2, \ldots, x_n$ constituem uma amostra aleatória de uma população infinita com média μ e variância σ^2 então

$$E[\bar{x}] = \mu,$$
$$Var[\bar{x}] = \frac{\sigma^2}{n}.$$

Demonstração.

$$E[\bar{x}] = E\left[\frac{\sum_i x_i}{n}\right] = \frac{1}{n}(n\mu) = \mu$$
$$V[\bar{x}] = V\left[\frac{\sum_i x_i}{n}\right] = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

□

Teorema 1.2.1. *Teorema Limite Central*

Se $x_1, x_2, \ldots, x_n$ constituem uma amostra aleatória de uma população com média μ e variância σ^2 finita, a distribuição limite de $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ à medida que $n \to \infty$ é a distribuição normal padrão.

A importância do Teorema Limite Central resulta do facto de ser possível usar a distribuição normal para aproximar a distribuição de $\bar{x}$, desde que a população possua média e variância finita e n seja suficientemente grande. A dimensão de n depende da população em estudo mas, em geral, $n \geq 30$ é usada, independentemente da forma da distribuição da população amostrada. Contudo, se a população amostrada é normal, a distribuição de $\bar{x}$ é normal, qualquer que seja a dimensão de n.

Teorema 1.2.2. *Se $\bar{x}$ é a média de uma amostra aleatória de tamanho n de uma população normal com média μ e variância σ^2, a sua distribuição amostral é normal com média μ e variância σ^2/n .*

De facto, pode ser demonstrado que a distribuição amostral de qualquer combinação linear de variáveis normais, mesmo correlacionadas e com médias e variâncias diferentes, é uma distribuição normal.

Exemplo 1.2.1: Distribuição da média amostral

Suponha que as classificações, a nível nacional, do exame de Geografia, têm uma média de 14.3, com um desvio padrão 2.1. Assumindo que a distribuição é normal, calcule:

a) a probabilidade de que um estudante, selecionado aleatoriamente, tenha uma classificação superior a 16 valores;

b) a probabilidade de que uma amostra aleatória de 10 estudantes tenha um média superior a 16 valores.

Solução

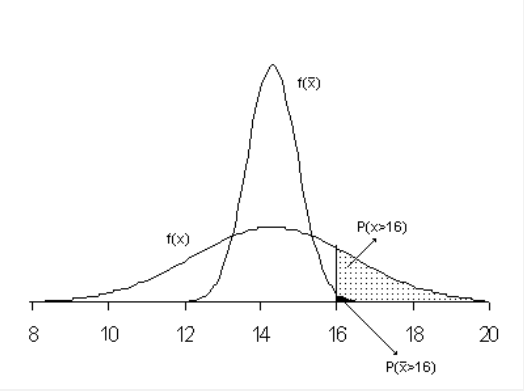
a) Assumindo uma distribuição normal, a probabilidade de um estudante obter uma nota superior a 16 é dada por,

$$P(x > 16) = P\left(z > \frac{16 - 14.3}{2.1}\right)$$
$$P(z > 0.81) = 0.2090$$

b) No caso de uma amostra de 10 estudantes,

$$P(\bar{x} > 16) = P\left(z > \frac{16 - 14.3}{2.1/\sqrt{10}}\right)$$
$$P(z > 2.56) = 0.0052$$

A figura explica, graficamente, os resultados obtidos.



Exemplo 1.2.2: Distribuição da média amostral

Uma máquina de enchimento de açúcar está regulada por forma a que a quantidade em cada pacote seja de 1000 gramas, com um desvio padrão de 50 gramas. Qual a probabilidade de que a média de uma amostra de 36 pacotes seja menor que 980 gramas?

Solução

Como resultado do Teorema Limite Central,

$$\mu_{\bar{x}} = \mu_x = 1000$$
$$\sigma_{\bar{x}} = \sigma_x/\sqrt{n} = \frac{50}{\sqrt{36}} = 8.3$$
$$P(\bar{x} \leq 990) = P(z \leq \frac{980 - 1000}{8.3})$$
$$P(z \leq -2.41) = 0.0080$$

Teorema 1.2.3. *Sejam $a_1, a_2, \ldots, a_n$ constantes e sejam $x_1, x_2, \ldots, x_n$ variáveis aleatórias normais com $E[x_i] = \mu_i$, $Var[x_i] = {\sigma_i}^2$ e $Cov[x_i, x_j] = \sigma_{i,j}$ $(i = 1, 2, \ldots, n)$. Então a distribuição amostral de uma combinação linear das variáveis normais $l = a_1x_1 + a_2x_2 + ... + a_nx_n$ possui uma distribuição normal com média*

$$E[l] = \mu = a_1\mu_1 + a_2\mu_2 + ... + a_n\mu_n$$

e variância

$$Var[l] = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + ... + a_n^2\sigma_n^2 + 2a_1a_2\sigma_{1,2} + 2a_1a_3\sigma_{1,3} + ... + 2a_1a_n\sigma_{1,n}$$
$$+ 2a_2a_3\sigma_{2,3} + ... + 2a_2a_n\sigma_{2,n} + ... + 2a_{n-1}a_n\sigma_{n-1,n}$$

Exemplo 1.2.3: Distribuição amostral

Suponha que amostras aleatórias independentes de dimensão n_1 e n_2, são selecionadas de duas populações normais, respetivamente, com médias e variâncias, μ_1, σ_1^2 e μ_2, σ_2^2. Assumindo que as médias amostrais são, respetivamente, $\bar{x}_1$ e $\bar{x}_2$, encontre a distribuição da sua diferença $(\bar{x}_1 - \bar{x}_2)$.

Solução

$$E[\bar{x}_1] = \mu_1,\ Var[\bar{x}_1] = \frac{\sigma_1^2}{n_1} \text{ e } E[\bar{x}_2] = \mu_2,\ Var[\bar{x}_2] = \frac{\sigma_2^2}{n_2}$$

Então, $l = \bar{x}_1 - \bar{x}_2$, é uma combinação linear de duas variáveis normais. De acordo com as

propriedades da esperança matemática,

$$E[l] = E[\bar{x}_1] - E[\bar{x}_2] = \mu_1 - \mu_2$$
$$Var[l] = (1)^2 Var[\bar{x}_1] + (-1)^2 Var[\bar{x}_2] + 2(1)(-1) Cov[\bar{x}_1, \bar{x}_2]$$

Contudo, uma vez que as amostras são independentes, as variáveis $\bar{x}_1, \bar{x}_2$ são independentes e logo $Cov[\bar{x}_1, \bar{x}_2] = 0$. Assim,

$$Var[l] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

1.3 Aproximação Normal à distribuição binomial

Considere-se uma experiência binomial com n tentativas, cada uma com uma probabilidade de sucesso p. O número de sucessos x em n tentativas é a proporção amostral de sucessos $\hat{p} = x/n$. Esta estatística é um estimador não tendencioso da proporção p, com variância $Var[\hat{p}] = \frac{p(1-p)}{n}$. Pelo Teorema Limite Central, a proporção amostral, à medida que n aumenta, será aproximadamente normal. Logo,

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

terá, aproximadamente, uma distribuição normal padrão para grandes valores de n.

1.4 A distribuição de Qui-Quadrado

A distribuição de Qui-Quadrado é um caso especial da distribuição Gama, e tem uma importância relevante em amostragem de populações normais.

Definição 1.4.1: Distribuição de Qui-Quadrado

Uma variável aleatória x segue a distribuição de Qui-Quadrado χ^2 com ν graus de liberdade, se a sua função densidade de probabilidade é dada por

$$f(x) = \begin{cases} \frac{1}{2^{\nu/2}\,\Gamma(\nu/2)} x^{\frac{\nu-2}{2}} e^{-x/2} & x > 0 \\ 0 & \text{outros valores} \end{cases}$$

A média e a variância da distribuição de Qui-Quadrado com ν graus de liberdade são, respetivamente, ν e 2ν.

A Figura 1.1 mostra a distribuição de Qui-Quadrado para diferentes graus de liberdade.

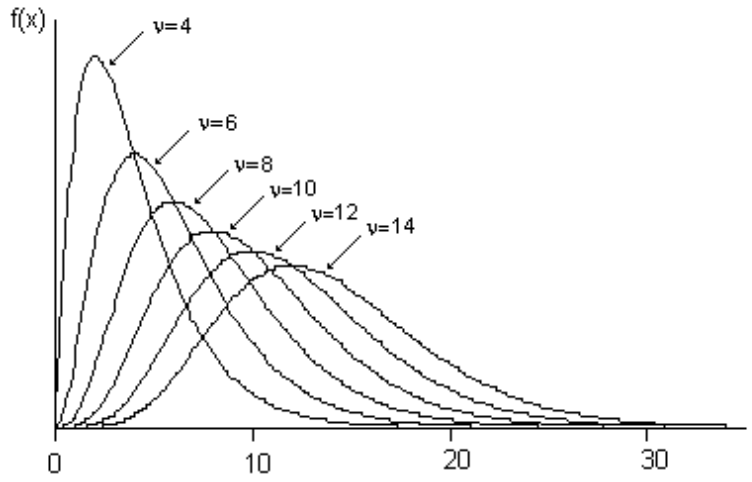


Figura 1.1: Distribuição de Qui-Quadrado.

Teorema 1.4.1. *Se x segue uma distribuição normal padrão, então x^2 tem uma distribuição de Qui-Quadrado com $\upsilon = 1$ grau de liberdade.*

Teorema 1.4.2. *Se $x_1, x_2, \ldots, x_n$ são variáveis aleatórias independentes seguindo distribuições normais padrão, então*

$$y = \sum_{i=1}^{n} x_i^2$$

segue uma distribuição de Qui-Quadrado com $\upsilon = n$ graus de liberdade.

Teorema 1.4.3. *Se $x_1, x_2, \ldots, x_n$ são variáveis aleatórias independentes seguindo distribuições de Qui-Quadrado com $\upsilon_1, \upsilon_2, \ldots, \upsilon_n$ graus de liberdade, então*

$$y = \sum_{i=1}^{n} x_i$$

segue uma distribuição de Qui-Quadrado com $\upsilon_1 + \upsilon_2 + \ldots + \upsilon_n$ graus de liberdade.

Teorema 1.4.4. *Se $\bar{x}$ e s^2 são a média e a variância de uma amostra aleatória de dimensão n de uma população normal com média μ e variância σ^2, então*

1. *$\bar{x}$ e s^2 são independentes;*

2. *a variável aleatória $\frac{(n-1)s^2}{\sigma^2}$ segue uma distribuição de Qui-Quadrado com $n-1$ graus de liberdade.*

1.5 A distribuição de t-Student

Alguns dos resultados apresentados anteriormente baseiam-se na assunção de que a variância da população σ^2 é conhecida. Na prática, contudo, tal não se verifica, donde se torna necessário substituir a variância da população por uma estimativa, em geral, a variância amostral s^2. A distribuição de t-Student vai permitir derivar a distribuição amostral de $\frac{\bar{x}-\mu}{s/\sqrt{n}}$.

Definição 1.5.1: Distribuição de t-Student

Se y e z são variáveis aleatórias independentes, y com uma distribuição de Qui-Quadrado com ν graus de liberdade e z uma distribuição normal padrão, então a distribuição de

$$t = \frac{z}{\sqrt{y/\nu}}$$

é dada por

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad -\infty < t < \infty$$

que é a distribuição de t-Student com ν graus de liberdade.

A Figura 1.2 mostra a distribuição t-Student para diferentes graus de liberdade. Como se pode verificar pela figura, à medida que o número de graus de liberdade aumenta, a distribuição de t-Student vai tender para a distribuição normal.

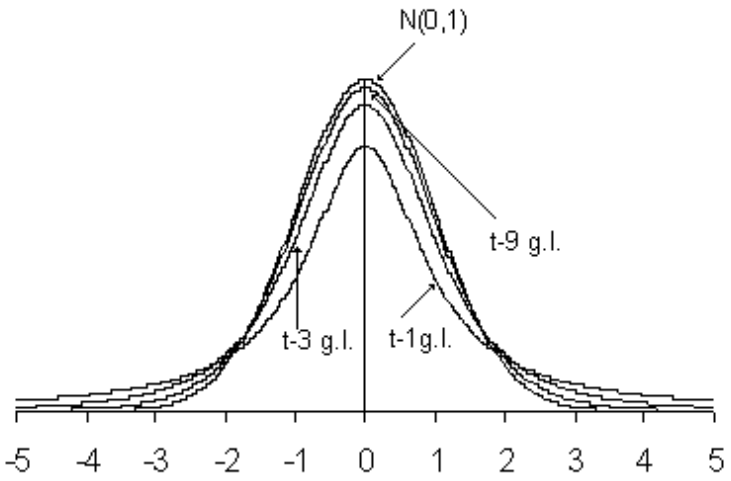


Figura 1.2: Distribuição de t-Student.

1.6 A distribuição F

A distribuição F vai permitir derivar a distribuição da razão de duas variâncias, sendo definida como a razão de duas variáveis de Qui-Quadrado divididas pelos respetivos graus de liberdade.

Definição 1.6.1: Distribuição F

Se y_1 e y_2 são variáveis aleatórias independentes seguindo distribuições de Qui-Quadrado com ν_1 e ν_2 graus de liberdade, então a distribuição de

$$x = \frac{y_1/\nu_1}{y_2/\nu_2}$$

é dada por

$$f(x) = \begin{cases} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right) x^{\frac{\nu_1}{2}-1} \left(1+\frac{\nu_1}{\nu_2}x\right)^{-\frac{1}{2}(\nu_1+\nu_2)} & x > 0 \\ 0 & \text{outros valores} \end{cases}$$

e corresponde à distribuição F com ν_1 e ν_2 graus de liberdade.

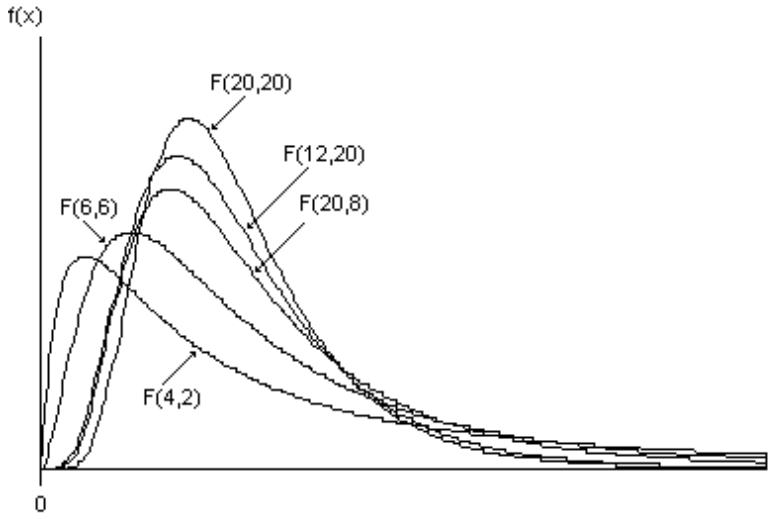


Figura 1.3: Distribuição F.

A Figura 1.3 apresenta as curvas da distribuição F para diversos pares de graus de liberdade. É possível demonstrar que se x tem uma distribuição F com ν_1 e ν_2 graus de liberdade, então $1/x$ tem uma distribuição F com ν_2 e ν_1 graus de liberdade. Assim, e tendo em consideração que

$P\left(x > F_{\alpha,\nu_1,\nu_2}\right) = \alpha$ (Figura 1.4) pode-se mostrar que

$$F_{1-\alpha,\nu_1,\nu_2} = \frac{1}{F_{\alpha,\nu_2,\nu_1}}.$$

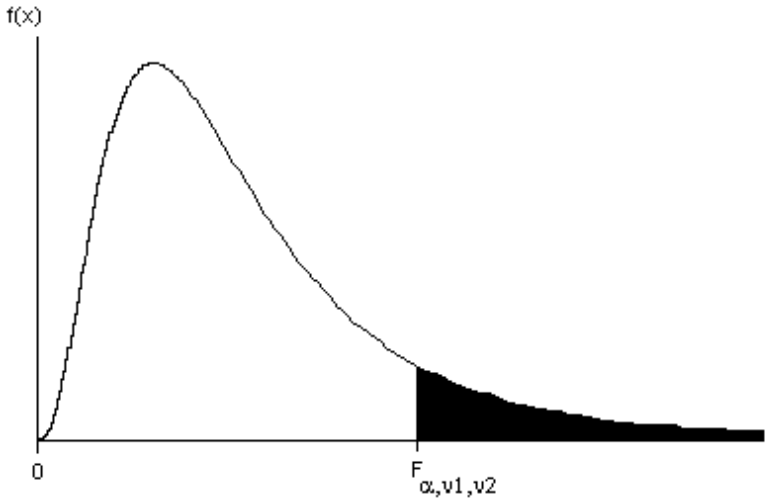


Figura 1.4: Área correspondente a $P\left(x > F_{\alpha,n_1,n_2}\right)$.

Finalmente, a distribuição de t-Student pode ser vista como um caso particular da distribuição F com ν_1=1 e ν_2 graus de liberdade.

Teorema 1.6.1. *Se s_1^2 e s_2^2 são as variâncias de amostras aleatórias independentes de dimensão n_1 e n_2 de populações normais com variâncias σ_1^2 e σ_2^2, então*

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2}$$

segue uma distribuição F com $n_1 - 1$ e $n_2 - 1$ graus de liberdade.

Capítulo 2

Estimadores Pontuais

2.1 Introdução

A inferência estatística é em geral dividida em estimação e testes de hipóteses. Em estimação pretende-se escolher um valor de um parâmetro de um conjunto possível de alternativas. Em geral, uma estatística é usada para estimar um parâmetro populacional e, por isso, constitui uma estimativa pontual do referido parâmetro. Assim, exemplos de estimativas pontuais são a média amostral, a proporção amostral ou a variância amostral, usadas para estimar, respetivamente, a média de uma população, a proporção de uma distribuição binomial ou a variância de uma população. Estas estimativas fornecem um valor pontual para o parâmetro a estimar, sendo também referidas como estimadores. Assim, a média aritmética $\bar{x}$ é um estimador de μ, assim como s^2 é um estimador de σ^2.

Contudo, como um estimador é o resultado de uma amostra aleatória, possui, portanto, uma distribuição amostral. A distribuição da média amostral será aproximadamente normal, centrada em μ. A Figura 2.1 mostra uma possível distribuição da média amostral.

Definição 2.1.1: Estimador não enviesado

Um estimador $\hat{\theta}$ é um estimador não enviesado de θ se e só se

$$E\left[\hat{\theta}\right] = \theta.$$

O enviesamento de um estimador $\hat{\theta}$ é dado pela diferença

$$E\left[\hat{\theta}\right] - \theta.$$

Como os estimadores são variáveis aleatórias, importa estudar as suas propriedades estatísticas,

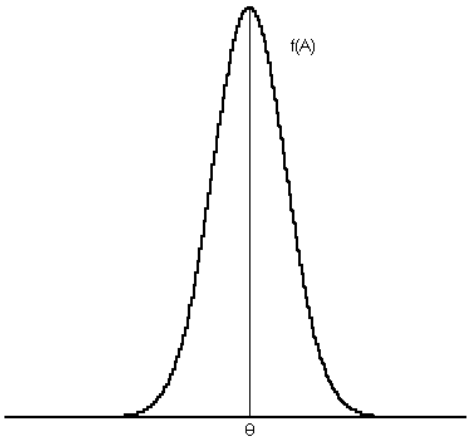


Figura 2.1: Distribuição de um estimador centrado.

por forma a definir com maior certeza quão próxima está a estimativa do parâmetro que se pretende estimar ou, em face de vários estimadores possíveis, qual o melhor. Estas propriedades são as seguintes: não enviesamento, consistência e eficiência relativa.

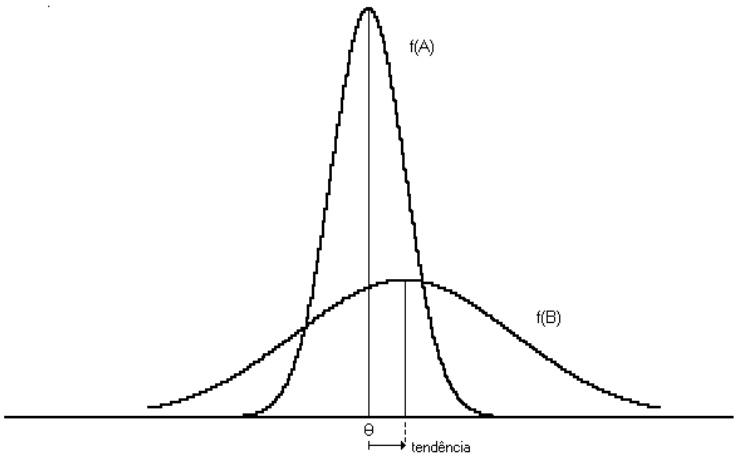


Figura 2.2: Distribuição de dois estimadores, com e sem tendência.

Assim, o Teorema 1.2.1 mostra que a média aritmética $\bar{x}$ é um estimador não enviesado para μ. Um estimador não tendencioso possui, portanto, uma distribuição amostral centrada no parâmetro a ser estimado. Na Figura 2.2, o estimador A é não tendencioso e o estimador B apresenta um enviesamento. Por outro lado, a Figura 2.3 apresenta as distribuições amostrais de dois estimadores não enviesados, mas com diferentes variâncias.

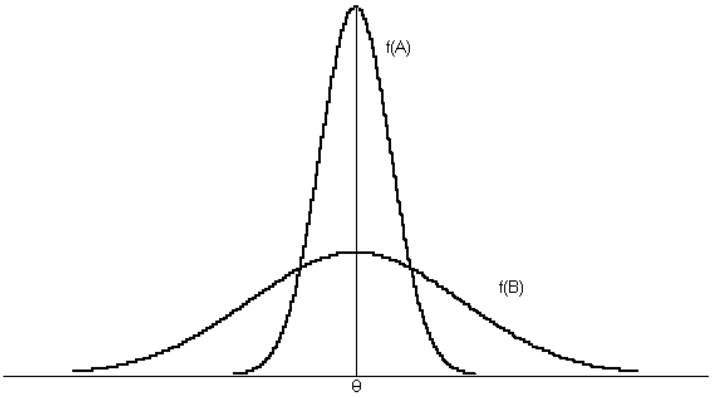


Figura 2.3: Distribuição de dois estimadores não tendenciosos.

Exemplo 2.1.1: Variância do estimador

Seja x uma variável aleatória com distribuição binomial. Mostre que x/n, a proporção observada de sucessos, é um estimador não tendencioso do parâmetro p. Calcule a variância do estimador.

Solução

$$E\left[\frac{x}{n}\right] = \frac{1}{n}E\left[x\right] = \frac{1}{n}np = p$$

$$Var\left[\frac{x}{n}\right] = \frac{1}{n^2}V\left[x\right] = \frac{1}{n^2}np\left(1-p\right) = \frac{p\left(1-p\right)}{n}$$

Exemplo 2.1.2: Estimador tendencioso

Sejam $x_1, x_2, \ldots, x_n$ uma amostra aleatória de uma população normal com média μ e variância σ^2. Mostre que a variância amostral s^2, é um estimador não tendencioso de σ^2.

Solução

$$s^2 = \sum_{i=1}^{n} \frac{\left(x_i - \bar{x}\right)^2}{n-1}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x} + \mu - \mu)^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \mu) - (\bar{x} - \mu)]^2$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left[(x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) + (\bar{x} - \mu)^2\right]$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \mu)^2 - 2(\bar{x} - \mu)\sum_{i=1}^{n}(x_i - \mu) + n(\bar{x} - \mu)^2\right]$$

Contudo, $-2(\bar{x} - \mu)\sum_{i=1}^{n}(x_i - \mu) = -2(\bar{x} - \mu)\, n\left(\frac{\sum_{i=1}^{n} x_i - n\mu}{n}\right) = -2n(\bar{x} - \mu)^2$.

Logo, $s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \mu)^2 - n(\bar{x} - \mu)^2\right]$.

Como cada x_i é um valor selecionado de uma população com média μ e variância σ^2, então

$$E\left[(x_i - \mu)^2\right] = \sigma^2 \quad (i = 1, 2, ...n)$$

$$E\left[(\bar{x} - \mu)^2\right] = \sigma^2_{\bar{x}} = \frac{\sigma^2}{n}.$$

Assim, o valor esperado de s^2 é dado por,

$$E\left[s^2\right] = E\left\{\frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \mu)^2 - n(\bar{x} - \mu)^2\right]\right\}$$
$$= \frac{1}{n-1}\left\{E\left[\sum_{i=1}^{n}(x_i - \mu)^2\right] - E\left[n(\bar{x} - \mu)^2\right]\right\}$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}\sigma^2 - n\left(\frac{\sigma^2}{n}\right)\right]$$
$$= \sigma^2.$$

Fica, assim, demonstrado que s^2 é um estimador não tendencioso de σ^2, qualquer que seja a forma da distribuição da população amostrada, bem como a razão do uso do divisor n-1 na fórmula de cálculo da variância amostral. Convém notar, contudo, que s não é um estimador não tendencioso de σ, dado que, debaixo de transformações funcionais, o não enviesamento de um estimador nem sempre é conservado.

Um estimador não é somente avaliado em termos do enviesamento, mas também com base na sua variância. Nesse sentido, pretende-se que o estimador seja tão concentrado quanto possível à volta do parâmetro a estimar. Um estimador, cujos valores se aproximam do parâmetro a estimar à medida que n aumenta, é dito consistente.

Definição 2.1.2: Estimador consistente

O estimador $\hat{\theta}$ é um estimador consistente do parâmetro θ se e só se, para qualquer constante positiva ε,

$$\lim_{n\to\infty} P\left(\left|\hat{\theta}-\theta\right| < \varepsilon\right) \to 1.$$

Um estimador $\hat{\theta}$ é consistente se verificar as seguintes condições suficientes, mas não necessárias:

1. $\hat{\theta}$ é não tendencioso.
2. $Var\left[\hat{\theta}\right] \to 0$ à medida que $n \to \infty$.

Convém notar que a consistência é uma propriedade assimptótica que não explicita a rapidez da convergência, sendo, no entanto, mais fácil de analisar do que a eficiência.

Exemplo 2.1.3: Estimador consistente

Mostre que a média aritmética $\bar{x}$ é um estimador consistente da média μ.

Solução

A média $\bar{x}$ é um estimador não tendencioso cuja variância

$$Var\left[\bar{x}\right] = \frac{\sigma^2}{n} \to 0$$

quando $n \to \infty$.

Entre dois estimadores não enviesados, como os da Figura 2.3, o estimador A é preferível ao estimador B, porque apresenta uma menor variância. A definição de eficiência permite comparar dois estimadores não tendenciosos $\hat{\theta}_1$ e $\hat{\theta}_2$, através da razão das variâncias,

$$\frac{Var\left[\hat{\theta}_1\right]}{Var\left[\hat{\theta}_2\right]}.$$

Contudo, quando se pretende comparar um estimador tendencioso com um outro não enviesado, ou mesmo um outro tendencioso, é necessário conjugar a tendência com a variância do estimador. Assim, por exemplo, na Figura 2.4 são apresentados três estimadores. Poderá ser justificada a escolha do estimador C em virtude de apresentar uma menor variância, apesar de um grande enviesamento. No entanto, e para que a escolha não seja subjetiva, é possível usar um critério que combina a tendência com a variância. Esse critério, para um estimador $\hat{\theta}$, denominado como Erro Quadrático Médio (EQM), é definido através da seguinte expressão,

$$EQM = E\left[\left(\hat{\theta}-\theta\right)^2\right] = V\left[\hat{\theta}\right] + \left(E\left[\hat{\theta}-\theta\right]\right)^2.$$

Assim, a eficiência de dois quaisquer estimadores pode ser calculada através da razão dos erros quadráticos médios. Para o caso de dois estimadores não tendenciosos, esta razão é equivalente à razão das variâncias.

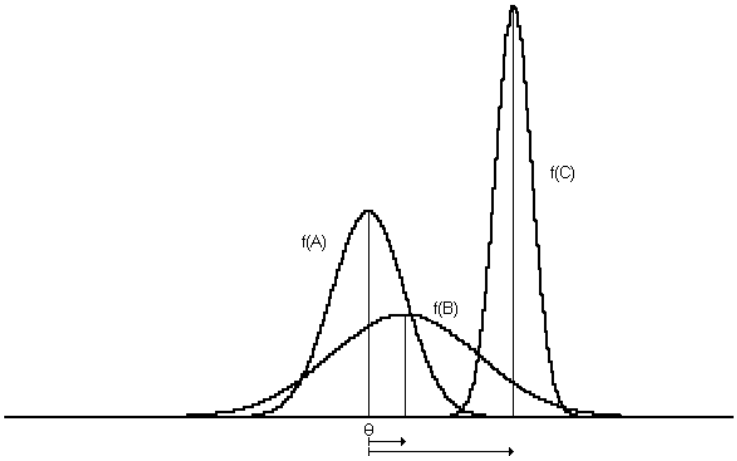


Figura 2.4: Comparação entre estimadores.

2.2 Método dos Momentos

Um dos métodos mais antigos para gerar estimadores pontuais é o chamado método dos momentos, que tem por base o facto de que o momento de ordem k, definido na origem, de uma variável aleatória é $\mu'_k = E\left[X^k\right]$. O momento de ordem k para uma amostra pode ser definido de forma semelhante.

Definição 2.2.1: Momento de ordem k

O momento de ordem k de um conjunto de observações, é a média da potência de ordem k, simbolicamente representada por m'_k,

$$m'_k = \frac{\sum_{i=1}^{n} x_i^k}{n}$$

Assim, para o caso $k = 1$, o primeiro momento populacional é μ e o correspondente momento amostral é $\bar{x}$. Para uma qualquer população definida por p parâmetros, o método dos momentos consiste em resolver um sistema de p equações, $m'_k = \mu'_k \quad k = 1, 2, \ldots, p$.

Exemplo 2.2.1: Momento de ordem k

Considere uma amostra de tamanho n de uma população gama, cuja função densidade de probabilidade é dada por

$$f(x) = \begin{cases} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x > 0, \alpha > 0, \beta > 0 \\ 0 & \text{outros valores} \end{cases}$$

sendo os momentos de ordem k, centrados na origem, dados por

$$\mu'_k = \frac{\beta^k \Gamma(\alpha + k)}{\Gamma(\alpha)}.$$

Use o método dos momentos para estimar os parâmetros α e β.

Solução

A função gama satisfaz a seguinte relação recursiva

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

e os primeiros momentos são

$$E[x] = \alpha\beta$$
$$E\left[x^2\right] = \alpha(\alpha + 1)\beta^2.$$

Logo,

$$\begin{cases} m'_1 = \alpha\beta \\ m''_2 = \alpha(\alpha + 1)\beta^2 \end{cases}$$

e as correspondentes estimativas,

$$\begin{cases} \hat{\alpha} = \frac{(m'_1)^2}{m'_2 - (m'_1)^2} \\ \hat{\beta} = \frac{m'_2 - (m'_1)^2}{m'_1} \end{cases}$$

e como $m'_1 = \bar{x}$ e $m'_2 = \sum_{i=1}^{n} x_i^2/n$

$$\begin{cases} \hat{\alpha} = \frac{n\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n\bar{x}_1} \end{cases}.$$

O método de estimação baseado nos momentos, apesar da sua simplicidade, tem desvantagens quando comparado com o método da Máxima Verosimilhança, já que em alguns casos, as

estimativas produzidas não possuem as propriedades desejáveis de um estimador.

2.3 Método da Máxima Verosimilhança

Um dos melhores métodos para gerar estatísticas pontuais é o chamado método da Máxima Verosimilhança desenvolvido por R. A. Fisher. Entre outras vantagens, Fisher demonstrou que os estimadores gerados por este método eram, em geral, suficientes, não tendenciosos e assimptoticamente de variância mínima.

Para compreender o método, considere-se o seguinte exemplo. Uma urna contém um grande número de bolas vermelhas e negras, na proporção de 3:1. Contudo, não se sabe qual das cores está presente em maioria, se a vermelha se a negra. Para o efeito, uma amostra de 3 bolas é retirada dessa urna. Assim, os resultados possíveis são (n^o de bolas vermelhas, n^o de bolas negras): (3,0); (2,1); (1,2); (0,3). Para um grande número de bolas dentro da urna, as probabilidades podem ser descritas por uma distribuição binomial. Contudo, as probabilidades associadas a cada um dos eventos dependem de qual a cor presente em maioria. Se a cor maioritária for a vermelha, então a probabilidade de retirar uma bola vermelha é $p = 3/4$, caso contrário é $p = 1/4$.

Nº de bolas vermelhas	0	1	2	3
$p = 3/4$	1/64	9/64	27/64	27/64
$p = 1/4$	27/64	27/64	9/64	1/64

A tabela lista as probabilidades de todos os acontecimentos possíveis, para os dois casos, de cor maioritária vermelha ou negra. Se, por exemplo, o resultado observado fosse 2 bolas vermelhas, a maior probabilidade de tal ocorrência resulta da situação em que a cor vermelha é maioritária (27/64 contra 9/64), ou seja, tal favoreceria a escolha de p=3/4; inversamente, a ocorrência de 0 bolas vermelhas, favoreceria a escolha de uma maioria de bolas negras, já que para esta situação a probabilidade é muito maior (27/64 contra 1/64). Neste caso, esta tabela poderá ser vista como uma tabela de decisão, em que o resultado 2 ou 3 favorece a hipótese vermelha, enquanto que um resultado de 0 ou 1 favoreceria a escolha oposta.

Em resumo, com base nos valores observados na amostra aleatória, é escolhido um valor para a estimativa que maximiza a probabilidade de obter aqueles dados. Assim, no caso discreto, uma amostra aleatória de n observações, $x_1, x_2, \ldots, x_n$, com uma função de probabilidade dependente de um parâmetro θ, a probabilidade de observar estes n valores independentes é dada por, $P(x_1, x_2, ..., x_n) = P(x_1)P(x_2)...P(x_n) = f(x_1, x_2, ..., x_n; \theta)$ que corresponde à distribuição de probabilidade conjunta das variáveis aleatórias no ponto amostral $(x_1, x_2, \ldots, x_n)$. Dado que os valores de $x_1, x_2, \ldots, x_n$, são conhecidos, esta função depende de θ, e é referida como função de verosimilhança. No caso contínuo, $f(x_1, x_2, ..., x_n; \theta)$, representa a função de densidade con-

junta no ponto $(x_1, x_2, \ldots, x_n)$. Fisher sugeriu que o valor de θ devia ser escolhido, por forma a maximizar esta função.

Definição 2.3.1: Função de Verosimilhança

Se $x_1, x_2, \ldots, x_n$ são os valores de uma amostra aleatória de uma população com parâmetro θ, a função de Verosimilhança é dada por, $L(\theta) = f(x_1, x_2, ..., x_n; \theta)$ para valores de θ no domínio dado. $f(x_1, x_2, ..., x_n; \theta)$ é o valor da função de probabilidade conjunta ou a função de densidade conjunta das variáveis aleatórias $X_1, X_2, \ldots, X_n$ observadas.

Assim, o método da Máxima Verosimilhança consiste na maximização da função de Verosimilhança e, por via do cálculo diferencial, no caso de um só parâmetro θ, o valor que anula a primeira derivada, corresponde ao máximo.

Exemplo 2.3.1: Método da Máxima Verosimilhança

Seja x uma variável aleatória de Bernoulli. A função de probabilidade é dada por

$$f(x;p) = \begin{cases} p^x(1-p)^{1-x} & x = 0, 1 \\ 0 & \text{outros valores} \end{cases}$$

onde p é o parâmetro a ser estimado.

Solução

$$\begin{aligned} L(p) &= p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}...p^{x_n}(1-p)^{1-x_n} \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \end{aligned}$$

O máximo de $L(p)$ é também o máximo de $\ln L(p)$. Assim,

$$\begin{aligned} \ln L(p) &= \left(\sum_{i=1}^{n} x_i\right)\ln p + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p) \\ \frac{d\ln L(p)}{dp} &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{\left(n - \sum_{i=1}^{n} x_i\right)}{1-p} \\ \hat{p} &= \frac{\sum_{i=1}^{n} x_i}{n}. \end{aligned}$$

Exemplo 2.3.2: Método da Máxima Verosimilhança

Seja X uma variável aleatória exponencial com parâmetro λ. Calcule o estimador de Máxima Verosimilhança para o parâmetro λ, com base numa amostra de tamanho n. Considere, em seguida, uma amostra de $n = 10$ valores respeitantes ao tempo de vida (em horas) de um componente elétrico (8.2, 40.5, 3.9, 7.7, 7.1, 3.3, 4.3, 25.4, 5.2, 1.0). Estime o valor do parâmetro λ com base nestes 10 valores.

Solução

$$L(\lambda) = f(x_1)f(x_2)...f(x_n)$$

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

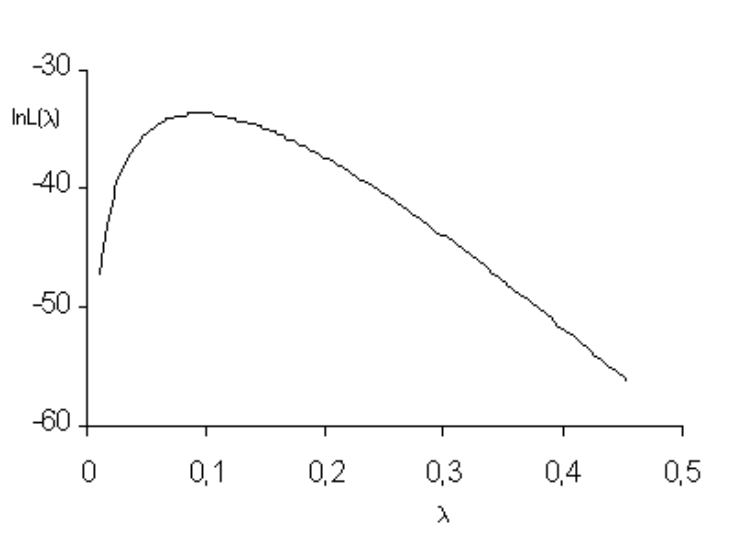
$$\ln L(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

Como $\bar{x} = 10.66$, a estimativa de λ é $\hat{\lambda} = \frac{1}{10.66} = 0.094$.

Os valores apresentados foram gerados a partir de uma distribuição exponencial com $\lambda = 0.1$, e como se pode ver pelo gráfico da figura, o máximo do logaritmo da função de Verosimilhança ocorre para $\hat{\lambda} = 0.094$.



Podem surgir, contudo, algumas situações em que poderá não ser fácil a aplicação do método

da Máxima Verosimilhança, nomeadamente nas situações em que não seja possível obter a derivada da função de Verosimilhança.

Exemplo 2.3.3: Método da Máxima Verosimilhança

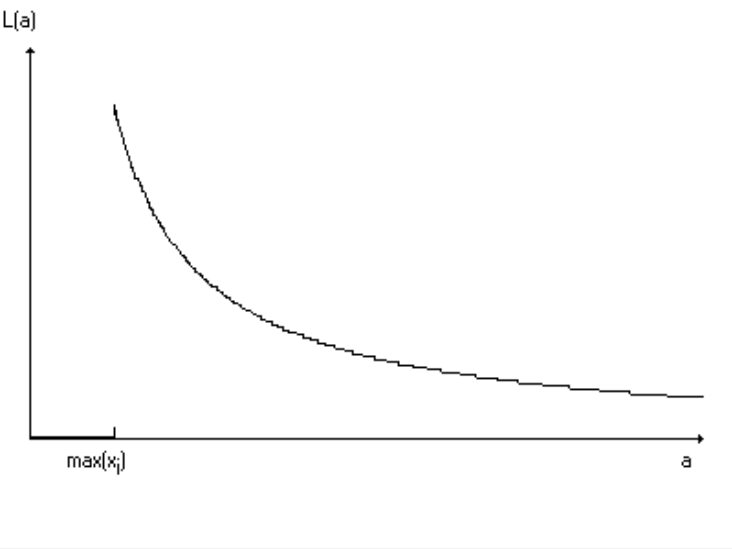
Sejam $x_1, x_2, \ldots, x_n$ os valores de uma amostra de uma distribuição uniforme, com parâmetros $\alpha = 0, \beta = a$. Encontre o estimador de Máxima Verosimilhança para a.

Solução

A função densidade de probabilidade é dada por $f(x; a) = \frac{1}{a}$ e a função de Verosimilhança por

$$L(a) = \prod_{i=1}^{n} f(x_i; a) = \left(\frac{1}{a}\right)^n .$$

O valor da função de Verosimilhança cresce à medida que a decresce. Contudo, para qualquer valor observado, $0 \leq x_i \leq a$, logo a não pode ser menor que qualquer valor da amostra, e a função atinge o seu máximo quando a é igual ao maior dos valores na amostra, isto é, $\hat{a} = \max(x_i)$. Esta situação é ilustrada na figura e, como se pode ver, as regras do cálculo não se podem aplicar nesta situação, já que o máximo ocorre num ponto de descontinuidade.



Finalmente, importa referir, em jeito de resumo, algumas das propriedades mais importantes dos estimadores de Máxima Verosimilhança.

Definição 2.3.2: Propriedades dos estimadores de Máxima Verosimilhança

Em condições muito gerais, quando a dimensão da amostra n é grande e se $\hat{\theta}$ é o estimador de Máxima Verosimilhança do parâmetro θ, então

1. $\hat{\theta}$ é aproximadamente um estimador não tendencioso;
2. a variância de $\hat{\theta}$ é quase tão pequena quanto a variância que poderia ser obtida com qualquer outro estimador;
3. $\hat{\theta}$ tem uma distribuição aproximadamente normal.

Definição 2.3.3: Propriedade da Invariância

Sejam $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ os estimadores de máxima verosimilhança dos parâmetros $\theta_1, \theta_2, \ldots, \theta_k$. Então, o estimador de máxima verosimilhança de qualquer função $h(\theta_1, \theta_2, \ldots, \theta_k)$ destes parâmetros é a mesma função $h\left(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k\right)$ dos estimadores $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$.

Existem outras técnicas de estimação, nomeadamente o método dos mínimos quadrados, que será abordado no capítulo de regressão. Outras técnicas incluem os estimadores robustos, os estimadores "jacknife" e os estimadores bayesianos.

Exercícios

1. Considerando uma amostra aleatória de dimensão n, encontre o estimador de máxima verosimilhança para o parâmetro

$$f(x) = \frac{e^{\lambda}\lambda^{x}}{x!} \qquad x = 0, 1, 2, \ldots$$

da distribuição de probabilidade de Poisson.

2. Considerando uma amostra aleatória de dimensão n, encontre os estimadores de máxima verosimilhança para os parâmetros μ e σ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty \quad \sigma > 0$$

da distribuição normal.

3. Considerando uma amostra aleatória de dimensão n, encontre o estimador de máxima verosimilhança para o parâmetro da seguinte distribuição de probabilidade,

$$f(x) = (\alpha + 1)x^{\alpha} \qquad 0 < x < 1.$$

4. Numa experiência binomial, foram observados sucessos em tentativas. Encontre o estimador de máxima verosimilhança para o parâmetro

$$f(x) = C_x^n p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \ldots n$$

da distribuição binomial. É o estimador tendencioso?

5. Dada uma amostra aleatória de dimensão n, encontre o estimador de máxima verosimilhança para o parâmetro da seguinte distribuição acumulada

$$F(x) = x^\alpha \qquad 0 < x < 1.$$

6. Encontre o estimador de máxima verosimilhança, dada uma amostra de dimensão n, para o parâmetro θ da seguinte função densidade

$$f(x|\theta) = e^{-(x-\theta)}$$

Capítulo 3

Intervalos de Confiança

3.1 Introdução

Uma estimativa pontual, apesar de fornecer um valor para o parâmetro a estimar, em muitas situações pode ser pouco útil, pois é pouco provável que coincida com o valor do parâmetro a estimar. Por outro lado, não fornece uma medida da distância da estimativa ao parâmetro, apesar de esta poder ser aproximada pelo erro padrão da estimativa. Assim, em vez de um valor, pode ser preferível fornecer um intervalo de valores, cuja dimensão tenha em consideração a dispersão dos valores observados. Portanto, para um qualquer parâmetro θ, o objetivo é determinar os dois limites que definem o intervalo $\hat{\theta}_I < \theta < \hat{\theta}_S$, onde $\hat{\theta}_I$ e $\hat{\theta}_S$ dependem da distribuição amostral de $\hat{\theta}$ e são, respetivamente, os limites inferior e superior do intervalo.

Contudo, amostras diferentes irão produzir diferentes valores $\hat{\theta}$, assim como os correspondentes limites $\hat{\theta}_I$ e $\hat{\theta}_S$ e, nesse sentido, estes limites podem ser considerados variáveis aleatórias. Tendo por base a distribuição amostral de $\hat{\theta}$, é então possível construir um intervalo com uma determinada probabilidade de conter o parâmetro θ. Assim,

$$P(\hat{\theta}_I < \theta < \hat{\theta}_S) = 1 - \alpha$$

com $0 < \alpha < 1$, define o intervalo bilateral $\hat{\theta}_I < \theta < \hat{\theta}_S$ com $(1 - \alpha)\,100\%$ de confiança. Deve ser notado, contudo, que os intervalos de confiança para um dado parâmetro não são únicos, já que é possível construir intervalos com diferentes limites, mas a mesma confiança. Além disso, podem ser construídos intervalos unilaterais tais como,

$$P(\hat{\theta}_I < \theta) = 1 - \alpha$$

a que corresponde o intervalo $\left[\hat{\theta}_I, \infty\right)$; da mesma forma que a

$$P\left(\theta < \hat{\theta}_S\right) = 1 - \alpha$$

corresponde o intervalo $\left(-\infty, \hat{\theta}_S\right]$.

Uma forma de construir intervalos de confiança é através do método da variável fulcral, para o que é necessário encontrar uma quantidade fulcral, que seja uma função das medidas amostrais e do parâmetro desconhecido, e cuja distribuição de probabilidade não dependa de θ. Se a distribuição da quantidade fulcral é conhecida, a probabilidade de um evento não é afetada por mudanças de escala ou de translação.

Exemplo 3.1.1: Intervalo de confiança

Com base numa só observação x de uma distribuição exponencial, construa um intervalo de confiança de 90% para o parâmetro θ .

Solução

A função densidade de é dada por

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & x \geq 0 \\ 0 & \text{outros valores} \end{cases}$$

Usando a transformação $u = \frac{x}{\theta}$, que é uma função de x e de θ, mas cuja distribuição não depende de θ, é possível encontrar dois limites u_I e u_S tal que

$$P\left(u_I < u < u_S\right) = 0.90,$$

ou seja,

$$P\left(u < u_I\right) = \int_0^{u_L} e^{-u} du = 0.05$$
$$P\left(u > u_S\right) = \int_{u_S}^{\infty} e^{-u} du = 0.05$$

Calculando os integrais,

$$P\left(0.051 < u < 2.996\right) = P\left(0.051 < \frac{x}{\theta} < 2.996\right) = 0.90$$

Como o objetivo é encontrar um intervalo para θ, manipulando as desigualdade é possível obter

$$P\left(\frac{0.051}{x} < \frac{1}{\theta} < \frac{2.996}{x}\right)$$
$$P\left(\frac{x}{2.996} < \theta < \frac{x}{0.051}\right) = 0.90.$$

Assim, com base num valor observado x e uma confiança de 90%, é possível afirmar que os limites obtidos incluem o parâmetro θ.

3.2 Intervalos de Confiança para as Médias

A média aritmética $\bar{x}$ de uma amostra de dimensão n de uma distribuição normal, possui uma distribuição normal, com média $\mu_{\bar{x}} = \mu$ e uma variância $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$. Com base nestes pressupostos, é possível estabelecer um intervalo de confiança

$$P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha$$

onde

$$z = \frac{\bar{x} - \mu}{\sigma\sqrt{n}}$$

e $-z_{\alpha/2}$ e $z_{\alpha/2}$ são os limites de integração da distribuição normal padrão tal que as áreas $\left(-\infty, -z_{\alpha/2}\right]$ e $\left[z_{\alpha/2}, \infty\right)$ são iguais a $\alpha/2$, como se pode ver pela Figura 3.1.

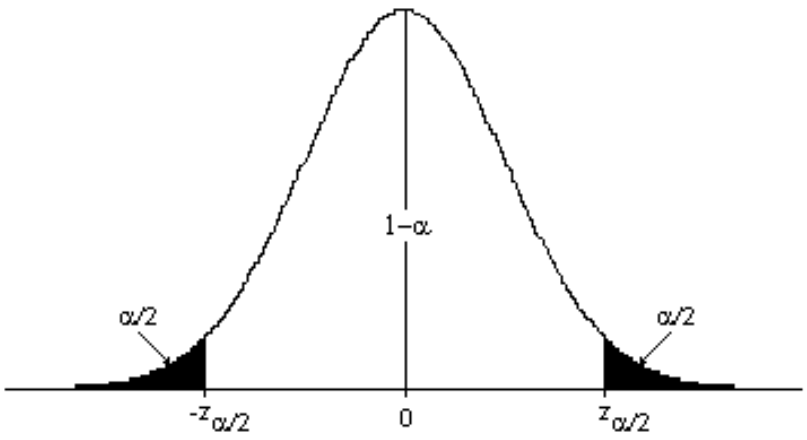


Figura 3.1: Intervalo de confiança de $(1 - \alpha)100\%$.

Assim,

$$P\left(-z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

ou, de uma forma equivalente, por manipulação das desigualdades,

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Daqui resulta que os limites inferior e superior do intervalo, respetivamente,

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad , \quad \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

para um dado valor de α só dependem dos valores observados.

Definição 3.2.1: Intervalo de Confiança de $(1-\alpha)100\%$ para μ

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \approx \bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

onde $z_{\alpha/2}$ é o valor de z que localiza uma área de $\alpha/2$ à direita da curva normal padrão, σ é o desvio padrão da população de onde a amostra foi retirada, n é a dimensão da amostra e $\bar{x}$ é o valor da média amostral.

Para construir um intervalo de confiança quando σ é desconhecido, o desvio padrão amostral s, quando $n \geq 30$, pode ser usado como uma aproximação a σ. Atendendo a que o Teorema Limite Central garante que a distribuição $\bar{x}$ é aproximadamente normal, independentemente da distribuição da população, não são impostas quaisquer condições à aplicabilidade desta fórmula.

Como já foi afirmado, os intervalos de confiança não são únicos, como se pode ver pelo seguinte intervalo, também com $(1-\alpha)100\%$,

$$\bar{x} - z_{2\alpha/3}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/3}\frac{\sigma}{\sqrt{n}}.$$

Para tornar mais explícito o significado da confiança num intervalo, assuma-se que era conhecida a média e o desvio padrão das alturas de todos os estudantes da Universidade do Minho, com 20 anos de idade ($\mu = 170$ cm, $\sigma = 10$ cm). A partir dos diferentes cursos, eram recolhidas amostras aleatórias de 25 estudantes, com 20 anos de idade, e a respetiva média calculada. Os resultados observados foram os seguintes:

Amostra	1	2	3	4	5
Média (cm)	172	168	171	165	172

Admitindo que se pretendia um intervalo com 95% de confiança, o valor correspondente de $z_{\alpha/2} = 1.96$, conduz a um intervalo, definido em função da média,

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
$$\bar{x} - 1.96\frac{10}{\sqrt{25}} < \mu < \bar{x} + 1.96\frac{10}{\sqrt{25}}$$
$$\bar{x} \pm 4\,\text{cm}.$$

Como se pode ver pela Figura 3.2, alguns dos intervalos conterão o parâmetro, enquanto que os limites de outros intervalos não compreendem o valor do parâmetro. Para um intervalo de 95% de confiança, isto quer dizer que 5% desses intervalos não conterão o parâmetro.

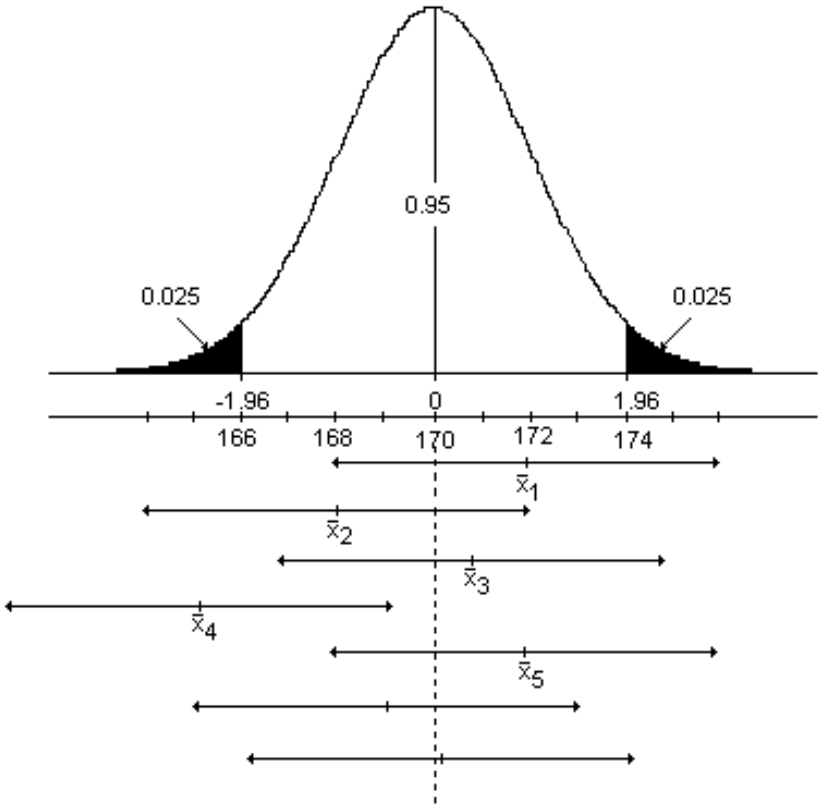


Figura 3.2: Intervalos de confiança a partir de várias amostras.

Exemplo 3.2.1: Intervalo de confiança para μ

O peso ao nascer é uma das variáveis mais importantes na avaliação do bem-estar de um recém-nascido. Ao longo de vários anos, diversos estudos permitiram conhecer a média e o desvio padrão para a população portuguesa. Suponha que o valor do desvio padrão para os bebés de sexo masculino é $\sigma = 562$ gramas. Num determinado centro de saúde, uma amostra de 19 recém nascidos apresentou uma média $\bar{x} = 3222$ gramas. Construa um intervalo de confiança de 95% para média do peso dos bebés.

Solução

Como se pode ver pela Figura 3.2, o valor que define uma área de 95% é $z_{\alpha/2} = 1.96$.

Assim,

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$
$$3222 - 1.96\frac{562}{\sqrt{19}} < \mu < 3222 + 1.96\frac{562}{\sqrt{19}}$$
$$3222 \pm 253g$$
$$2969 < \mu < 3475$$

O resultado obtido permite afirmar, com uma confiança de 95%, que o intervalo contém o verdadeiro valor da média do peso dos recém-nascidos.

Para construir um intervalo para amostras de pequena dimensão $n << 30$, em que σ é desconhecido, a variável

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

segue uma distribuição de t-Student com $n-1$ graus de liberdade. Logo,

$$P\left(-t_{\alpha/2,n-1} < t < t_{\alpha/2,n-1}\right) = 1 - \alpha$$

onde $t_{\alpha/2,n-1}$ corresponde à cauda superior de área $\alpha/2$ da distribuição t-Student com $n-1$ graus de liberdade. Substituindo t nesta expressão,

$$P\left(-t_{\alpha/2,n-1} < \frac{\bar{x} - \mu}{s/n} < t_{\alpha/2,n-1}\right) = 1 - \alpha$$

e resolvendo as desigualdades,

$$P\left(\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Definição 3.2.2: Intervalo de Confiança de $(1-\alpha)100\%$ para μ (amostras de pequena dimensão, σ desconhecido)

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$
$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

onde a distribuição de t-Student é baseada em $n-1$ graus de liberdade. A construção deste intervalo implica que a população de onde a amostra foi retirada segue uma distribuição aproximadamente normal.

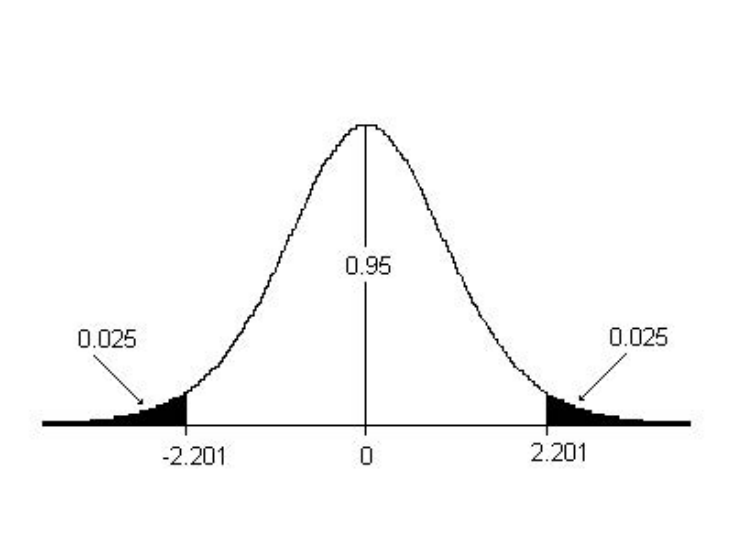
Exemplo 3.2.2: Intervalo de confiança para μ

Numa fábrica, uma amostra de 12 eixos foi retirada da linha de produção. O diâmetro médio encontrado foi de 19.92 cm com um desvio padrão de 0.17 cm. Construa um intervalo de confiança de 95% para o verdadeiro valor do diâmetro médio.

Solução

A figura mostra a distribuição de t-Student, com 11 graus de liberdade, a que corresponde um valor de $t_{0.025,11} = 2.201$. O intervalo de confiança correspondente fica assim definido,

$$19.92 - 2.201\frac{0.17}{\sqrt{12}} < \mu < 19.92 + 2.201\frac{0.17}{\sqrt{12}}$$
$$19.92 \pm 0.108$$
$$19.812 < \mu < 20.028$$



3.3 Intervalos de Confiança para a Diferença entre Médias

Se $\bar{x}_1$ e $\bar{x}_2$ são as médias de amostras aleatórias independentes, de dimensão n_1 e n_2, retiradas de populações normais com médias e variâncias, respetivamente, μ_1 e μ_2, e σ_1 e σ_2, então

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

segue uma distribuição normal padrão. Por via do Teorema Limite Central, z será aproximadamente normal para grandes amostras, independentemente das distribuições das populações amostradas.

Definição 3.3.1: Intervalo de Confiança de $(1-\alpha)100\%$ para $\mu_1 - \mu_2$ (amostras independentes)

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < (\mu_1 - \mu_1) < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx (\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Para grandes amostras, com $n_1 \geq 30$ e $n_2 \geq 30$, as variâncias amostrais podem ser usadas como aproximações aos respetivos parâmetros.

A aplicabilidade desta fórmula exige que as duas amostras sejam selecionadas de uma forma independente das duas populações; além disso, no caso de variâncias desconhecidas, a dimensão das amostras deve ser suficientemente grande ($n_1 \geq 30$, $n_2 \geq 30$) para que o Teorema Limite Central seja aplicável.

Exemplo 3.3.1: Intervalo de Confiança para a Diferença entre Médias

Pretende-se estimar a diferença entre os salários médios dos licenciados em Engenharia e dos licenciados em Relações Internacionais da Universidade do Minho. Para tal efeito, foram selecionados aleatoriamente licenciados com um ano de experiência de vida ativa. Os dados recolhidos foram os seguintes:

Curso	n	Salário médio, $\bar{x}$	Desvio padrão, s
Engenharia	40	1873.50	262.29
Relações Internacionais	30	1489.00	496.33

Calcule um intervalo de confiança de 90%.

Solução

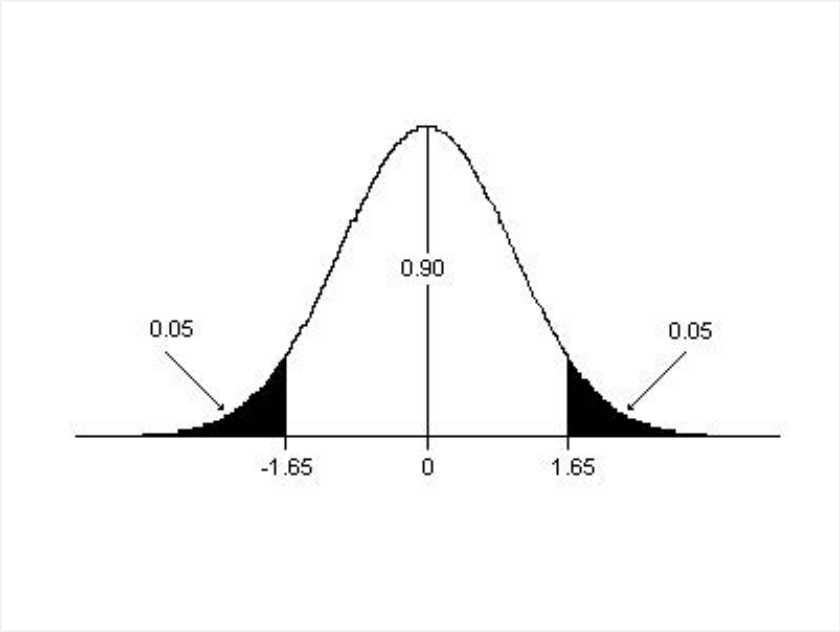
Como se pode ver pela figura, o valor de $z_{\alpha/2} = 1.645 \approx 1.65$ define um intervalo com 90% de confiança. Assim,

$$(1873.50 - 1489.00) \pm 1.65\sqrt{\frac{(262.29)^2}{40} + \frac{(496.33)^2}{30}}$$

$$3845.00 \pm 164.43$$

$$220.07 < \mu_1 - \mu_2 < 548.93$$

Atendendo a que o intervalo não inclui o zero, os resultados obtidos permitem concluir que os licenciados em Engenharia auferem ordenados médios superiores aos licenciados em Relações Internacionais.



Um intervalo de confiança para amostras de pequena dimensão e variâncias desconhecidas é deduzido a partir da distribuição de t-Student. No caso em que as duas populações normais donde são extraídas as amostras, possuem a mesma variância, esta pode ser estimada a partir da seguinte expressão

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

que é um estimador não tendencioso de σ^2.

Por outro lado, as variáveis aleatórias independentes $\frac{(n_1-1)s_1^2}{\sigma^2}$ e $\frac{(n_2-1)s_2^2}{\sigma^2}$ possuem uma distribuição de Qui-Quadrado, respetivamente, com $n_1 - 1$ e $n_2 - 1$ graus de liberdade, e a sua soma

$$y = \frac{(n_1 - 1)\, s_1^2}{\sigma^2} + \frac{(n_2 - 1)\, s_2^2}{\sigma^2} = \frac{(n_1 + n_2 - 2)\, s_p^2}{\sigma^2}$$

segue também uma distribuição de Qui-Quadrado com $n_1 + n_2 - 2$ graus de liberdade. Como as variáveis z e y são independentes, então

$$t = \frac{z}{\sqrt{\frac{y}{n_1+n_2-2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

segue uma distribuição de t-Student com $n_1 + n_2 - 2$ graus de liberdade. Assim, através da expressão

$$P\left(-t_{\alpha/2,n_1+n_2-2} < t < t_{\alpha/2,n_1+n_2-2}\right) = 1 - \alpha$$

é possível chegar ao intervalo de confiança para a diferença entre médias para amostras de pequena dimensão, com variância comum igual e desconhecida.

Definição 3.3.2: Intervalo de Confiança de $(1-\alpha)100\%$ para $\mu_1 - \mu_2$ (amostras independentes, $\sigma_1 = \sigma_2$ desconhecidos)

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2,n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2,n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

onde a distribuição de t-Student é baseada em $n_1 + n_2 - 2$ graus de liberdade.

A aplicação desta fórmula exige que ambas as populações de onde são retiradas as amostras tenham uma distribuição aproximadamente normal, que as variâncias σ_1^2 e σ_2^2 sejam iguais e que as amostras sejam recolhidas de uma forma independente.

Exemplo 3.3.2: Intervalo de Confiança para a Diferença entre Médias

Pretende-se testar duas formulações alimentares no crescimento de frangos de aviário. Os frangos distribuídos, por dois pavilhões A e B, foram alimentados durante cinco semanas com a respetiva ração. No fim do período de crescimento, foram selecionadas duas amostras, que produziram os seguintes resultados:

Grupo	n	Média, $\bar{x}$ (g)	Desvio padrão, s (g)
Ração 1	16	1623.75	192.71
Ração 2	10	1588.00	167.12

Construa um intervalo de confiança de 95%. O que pode concluir?

Solução

Para uma confiança de 95%, o valor aproximado de t-Student é $t_{0.025,24} = 2.06$, donde

$$s_p^2 = \frac{(16-1)(192.7131)^2 + (10-1)(167.1194)^2}{16+10-2} = 33684.7970$$

$$(1623.75 - 1588.00) \pm (2.06)\,(183.5342)\sqrt{\tfrac{1}{16} + \tfrac{1}{10}}$$

$$35.75 \pm 152.4091$$

$$-116.6591 < \mu_1 - \mu_2 < 188.1591$$

Como o intervalo contém o zero, os resultados permitem concluir que não existem diferenças entre os pesos médios dos frangos alimentados com as duas rações.

As condições de aplicabilidade do intervalo de confiança não têm que ser satisfeitas exatamente, sendo admissíveis pequenos desvios da normalidade, bem como variâncias diferentes, desde que

a dimensão das amostras seja igual. No entanto, para dimensões diferentes é ainda possível estabelecer um intervalo de confiança, mas para o qual o número de graus de liberdade associado à distribuição de t-Student é alterado.

Definição 3.3.3: Intervalo de Confiança de $(1-\alpha)100\%$ para $\mu_1 - \mu_2$ (amostras independentes, $\sigma_1 \neq \sigma_2$ desconhecidos, $n_1 = n_2 = n$)

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,2(n-1)} \sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}$$

onde a distribuição de t-Student é baseada em $n_1 - 1 + n_2 - 1 = 2(n-1)$ graus de liberdade.

Definição 3.3.4: Intervalo de Confiança de $(1-\alpha)100\%$ para $\mu_1 - \mu_2$ (amostras independentes, $\sigma_1 \neq \sigma_2$ desconhecidos, $n_1 \neq n_2$)

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,2(n-1)} \sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}$$

onde a distribuição de t-Student é baseada

$$\nu = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1} + \frac{\left(s_2^2/n_2\right)^2}{n_2-1}}$$

graus de liberdade.

A aplicação desta fórmula exige que ambas as populações de onde são retiradas as amostras tenham uma distribuição aproximadamente normal e que as amostras sejam recolhidas de uma forma independente. Pode acontecer que o valor de resultante da fórmula não seja inteiro. Nesse caso, a consulta da tabela de t-Student será feita com base no arredondamento para o inteiro mais próximo.

Os intervalos de confiança apresentados para a diferença entre médias assumiam que as amostras eram recolhidas de uma forma aleatória e independente. No entanto, podem surgir situações em que as amostras não são independentes, e a forma da sua recolha se diz emparelhada. Um exemplo seria testar a presença de cloro nas condutas de água de consumo, em dois pontos diferentes da mesma rede, ao longo de vários meses. Neste caso, as amostras não são independentes já que o teor de cloro vai depender da quantidade de cloro adicionada ao reservatório que fornece a rede. Numa experiência deste tipo, pretende-se determinar a diferença média entre pares de resultados, $\mu_d = (\mu_1 - \mu_2)$. Em geral, este procedimento é aplicado a amostras de pequena dimensão, e uma vez calculada a diferença média, o intervalo de confiança é estabelecido por via da distribuição de t-Student, tal como no caso já apresentado.

Definição 3.3.5: Intervalo de Confiança de $(1-\alpha)100\%$ para $\mu_1 - \mu_2$ (amostras emparelhadas)

Sejam $d_1, d_2, \ldots, d_n$ as diferenças entre observações emparelhadas de uma amostra aleatória de n pares. O intervalo de confiança para a diferença $\mu_d = (\mu_1 - \mu_2)$ é dado por

$$\bar{d} - t_{\alpha/2,n-1}\left(\frac{s_d}{\sqrt{n}}\right) < \mu_d < \bar{d} + t_{\alpha/2,n-1}\left(\frac{s_d}{\sqrt{n}}\right)$$
$$\bar{d} \pm t_{\alpha/2,n-1}\left(\frac{s_d}{\sqrt{n}}\right)$$

onde $\bar{d}$ é a média de diferenças amostrais, s_d o desvio padrão e onde a distribuição de t-Student é baseada em $n-1$ graus de liberdade.

A aplicação desta fórmula exige que as observações emparelhadas sejam recolhidas de uma forma aleatória e que a população das diferenças emparelhadas seja aproximadamente normal.

Exemplo 3.3.3: Intervalo de Confiança para a Diferença entre Médias

A exposição a determinadas substâncias perigosas tais como metais pesados ou dioxinas conduz a acumulações destas substâncias no sangue e no tecido gordo. Uma amostra de dez trabalhadores de uma fábrica onde existe a manipulação de dioxinas foi selecionada aleatoriamente. Nestes trabalhadores foi determinada a concentração (em ppm, partes por milhão) de dioxinas no plasma e no tecido gordo. Construa um intervalo de confiança para a diferença entre as concentrações de dioxina no plasma e no tecido gordo.

Trabalhador	1	2	3	4	5	6	7	8	9	10
Plasma	2.5	3.5	1.8	4.7	7.2	4.1	3.0	3.3	3.1	2.5
Tecido Gordo	4.9	6.9	4.2	4.4	7.7	2.5	5.5	2.9	5.9	2.3

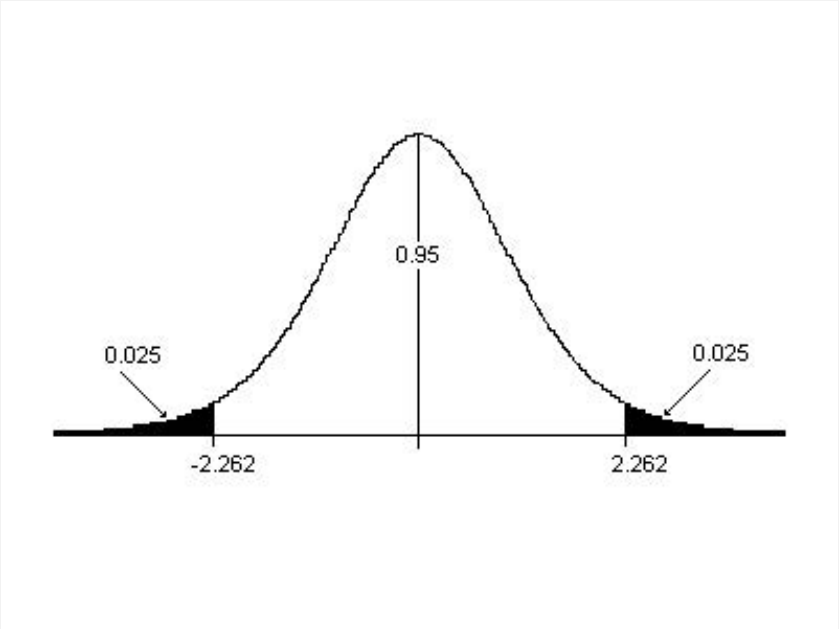
Solução

A diferença média entre a concentração no plasma e no tecido gordo é -1.7600 e o desvio padrão 2.3801. Assim, para uma confiança de 95%, o valor da distribuição de t-Student é $t_{0.025,9} = 2.262$, e o intervalo será definido como

$$\bar{d} = -1.7600$$
$$s_d = 2.3801$$
$$-1.7600 - 2.262\frac{2.3801}{\sqrt{10}} < \mu_d < -1.7600 + 2.262\frac{2.3801}{\sqrt{10}}$$
$$-1.7600 \pm 2.262\frac{2.3801}{\sqrt{10}}$$
$$-1.7600 \pm 1.7025$$
$$-3.4625 < \mu_d < -0.0575$$

Como os limites não incluem o zero, pode concluir-se que a concentração de dioxina se

verifica sobretudo ao nível do tecido gordo.



3.4 Intervalos de Confiança para as Proporções

O Teorema Limite Central permite, à medida que a dimensão da amostra aumenta, assumir que a distribuição de

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

será aproximadamente normal. Assim, para construir um intervalo de confiança $(1-\alpha)100\%$ para proporção p, tem-se que

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) = 1 - \alpha$$

e resolvendo as desigualdades,

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right)$$

Substituindo p por $\hat{p}$ na expressão, obtêm-se um intervalo aproximado.

Definição 3.4.1: Intervalo de Confiança de $(1-\alpha)100\%$ para p

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

onde $\hat{p}$ é a proporção de observações na amostra, com a característica desejada.

Para este intervalo ser aplicável é necessário que a dimensão da amostra seja suficientemente grande para que o Teorema Limite Central seja válido, o que se pode traduzir numa dimensão aproximada, tal que o intervalo $\hat{p} \pm 2\sqrt{\hat{p}\,(1-\hat{p})}$ não inclua o valor zero ou um.

Exemplo 3.4.1: Intervalo de Confiança de para p

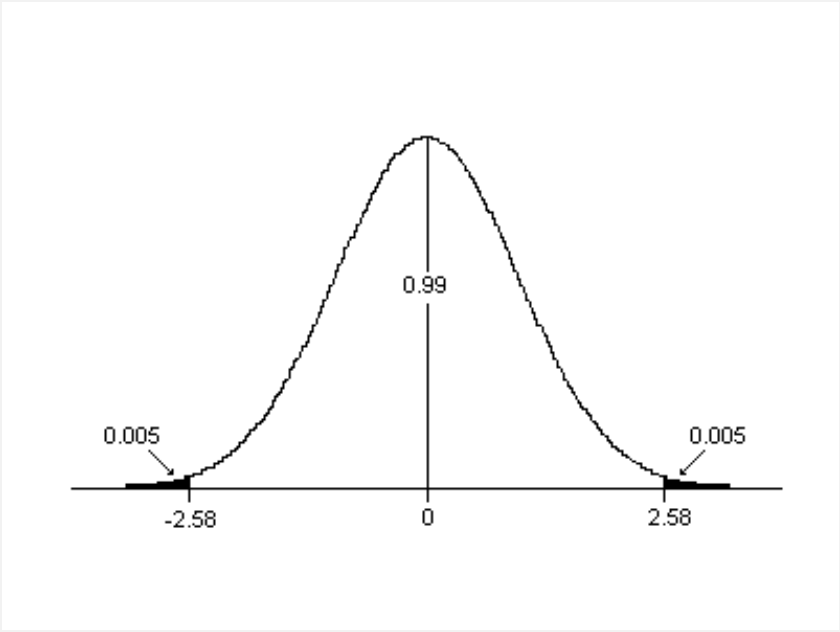
Para determinar a incidência de uma determinada doença genética no Norte de Portugal, foi recolhida uma amostra de gotas de sangue de 500 bebés, nascidos no ano de 1994. As análises permitiram detetar 37 bebés portadores da doença. Estime um intervalo de confiança de 99% para a proporção de portadores da doença.

Solução

Para um intervalo de confiança de 99%, como se pode ver pela Figura 3.6, o valor de $z_{\alpha/2} = z_{0.005} = 2.575 \approx 2.58$. Assim,

$$\hat{p} = \tfrac{37}{500} = 0.074$$
$$0.074 \pm 2.58\sqrt{\tfrac{0.074(1-0.074)}{500}}$$
$$0.074 \pm 0.030$$
$$0.044 < p < 0.104$$

pode-se afirmar que o intervalo 7.4% a 10.4% contém a incidência com uma confiança de 99%.



Por vezes, pode ser necessário estimar a diferença entre duas proporções binomiais, como por exemplo, estimar a diferença entre as proporções, de dois estratos sócio-económicos, dos votos favoráveis à despenalização da interrupção voluntária da gravidez, ou a diferença entre as incidências de uma doença entre duas regiões. Sejam x_1 e x_2 o número de sucessos em duas experiências binomiais com dimensão n_1 e n_2. As proporções $\hat{p}_1 = \frac{x_1}{n_1}$ e $\hat{p}_2 = \frac{x_2}{n_2}$, para grandes valores de n_1 e n_2, têm, aproximadamente, uma distribuição normal. Será de esperar que a distribuição de $(\hat{p}_1 - \hat{p}_2)$ seja, também, aproximadamente normal. Assumindo independência entre as amostras, o valor esperado e a variância para a diferença entre as proporções é dado por,

$$E[\hat{p}_1 - \hat{p}_2] = E[\hat{p}_1] - E[\hat{p}_2] = p_1 - p_2$$
$$Var[\hat{p}_1 - \hat{p}_2] = Var[\hat{p}_1] + Var[\hat{p}_2] = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

e a variável,

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

tem, aproximadamente, uma distribuição normal padrão. Assim, tal como no caso anterior, pode-se escrever

$$P\left(-z_{\alpha/2} < \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} < z_{\alpha/2}\right) = 1 - \alpha$$

o que após a resolução das desigualdades permite definir o intervalo de confiança.

Definição 3.4.2: Intervalo de Confiança de $(1-\alpha)100\%$ para $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

ou

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

onde $\hat{p}_1$ e $\hat{p}_2$ são as proporções das observações com a característica desejada.

Este é um intervalo aproximado, já que no cálculo da variância da diferença foram usadas as estimativas $\hat{p}_1$ e $\hat{p}_2$. Torna-se necessário, também, assumir que as amostras são suficientemente grandes para a aproximação ser válida, o que, em geral, é verificado se os intervalos

$$\hat{p}_1 \pm 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}, \quad \hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

não contiverem os valores zero ou um.

Exemplo 3.4.2: Intervalo de Confiança para a Diferença entre Proporções

Suponha que relativamente ao referendo sobre a regionalização, de uma amostra de 500 pessoas da zona Norte de Portugal e de uma amostra de 1000 pessoas da zona de Lisboa e Vale do Tejo, o número de respostas favoráveis foram, respetivamente, 377 e 223. Estabeleça um intervalo de confiança de 95% para a diferença entre as proporções.

Solução

Para um nível de confiança de 95%, o valor de $z_{0.025} = 1.96$, e o intervalo correspondente é

$$\hat{p}_1 = \tfrac{377}{500} \quad \hat{p}_2 = \tfrac{223}{1000}$$
$$(0.754 - 0.223) \pm 1.96\sqrt{\tfrac{0.754(1-0.754)}{500} + \tfrac{0.223(1-0.223)}{1000}}$$
$$0.531 \pm 0.046$$
$$0.485 < p_1 - p_2 < 0.577$$

donde se pode concluir que este intervalo contém a diferença entre proporções com uma confiança de 95%.

3.5 Intervalos de Confiança para as Variâncias

Ao contrário dos intervalos de confiança para as médias e proporções, os intervalos de confiança para a variância e razão de variâncias, as estatísticas pivô não possuem uma distribuição normal.

Sejam $x_1, x_2, \ldots, x_n$, uma amostra aleatória de dimensão n de uma população normal, com

média μ e variância σ^2. A estatística

$$\frac{(n-1)\,s^2}{\sigma^2}$$

é uma variável aleatória seguindo uma distribuição de Qui-Quadrado com $n-1$ graus de liberdade.

Assim, pode-se escrever que

$$P\left[\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)\,s^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right] = 1-\alpha$$

ou seja,

$$P\left[\frac{(n-1)\,s^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)\,s^2}{\chi^2_{1-\alpha/2,n-1}}\right]$$

o que é descrito pela Figura 3.3.

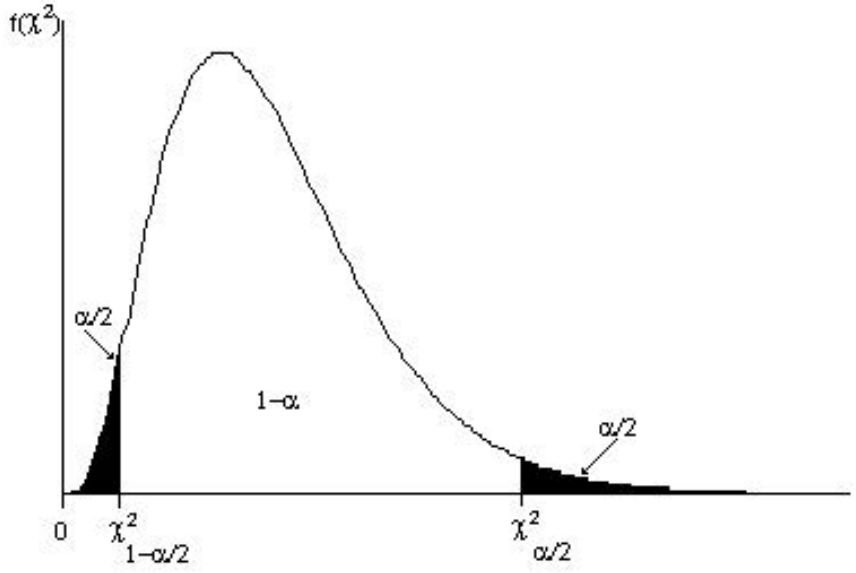


Figura 3.3: Intervalo de confiança com base na distribuição de χ^2.

Definição 3.5.1: Intervalo de Confiança de $(1-\alpha)100\%$ para σ^2

$$\frac{(n-1)\,s^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)\,s^2}{\chi^2_{1-\alpha/2,n-1}}$$

onde $\chi^2_{\alpha/2}$ e $\chi^2_{1-\alpha/2}$, são os valores que localizam uma área de $\alpha/2$, respetivamente, à direita e à esquerda da distribuição de Qui-Quadrado com $n-1$ graus de liberdade.

Este intervalo é válido para amostras retiradas de populações aproximadamente normais.

Exemplo 3.5.1: Intervalo de Confiança para a Variância

Numa fábrica de cimento, o controlador de qualidade sabe que a quantidade exata de cimento em sacos de 15 kg irá variar, uma vez que são vários os fatores incontroláveis que afetam o processo de enchimento. No entanto, para além da quantidade média de enchimento, a variação dessa quantidade é igualmente importante, na medida em que alguns sacos conterão mais, e outros menos, de 15 kg. Para estimar a variação, o controlador de qualidade recolheu uma amostra de 15 sacos de cimento, com uma média de 15.02 kg e um desvio padrão de 0.075 kg. Construa um intervalo de 90% de confiança para a variância do enchimento.

Solução

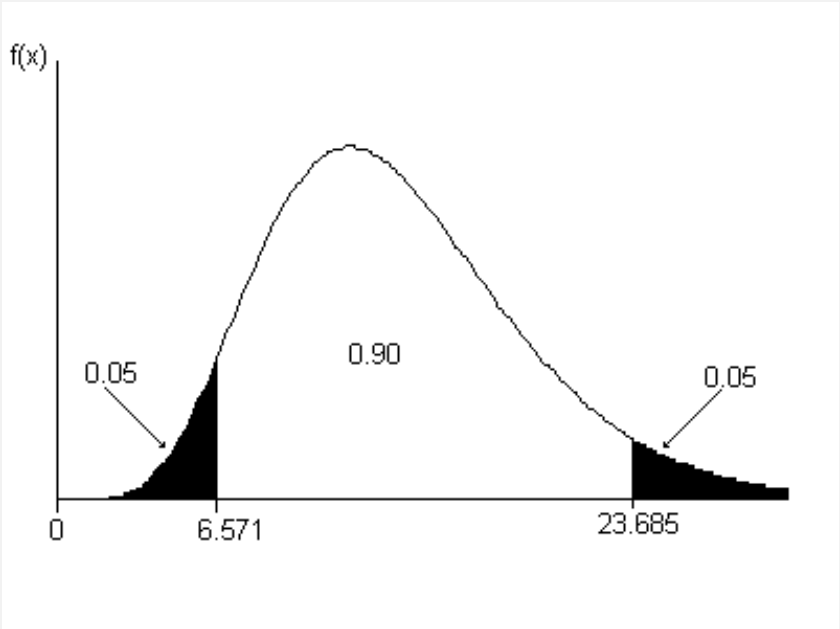
Para um intervalo de confiança de 90%, os valores de Qui-Quadrado correspondentes, com 14 graus de liberdade, são, respetivamente, $\chi^2_{0.95,4}$ e $\chi^2_{0.05,4}$, que definem uma área de 90%, tal como se pode ver na figura.

Assim, o intervalo será definido pelos limites

$$\frac{(15-1)(0.075)^2}{23.685} < \sigma^2 < \frac{(15-1)(0.075)^2}{6.571}$$

ou seja,

$$0.003325 < \sigma^2 < 0.011984$$



Assim, com uma confiança de 90% é possível afirmar que estes dois limites compreendem a variância do processo de enchimento.

3.6 Intervalo de Confiança para a Razão das Variâncias

A comparação entre duas variâncias populacionais é feita recorrendo à sua razão, pois esta possui uma distribuição amostral bem conhecida. Se s_1^2 e s_2^2 são as variâncias de amostras aleatórias independentes de tamanho n_1 e n_2 de populações normais com variâncias σ_1^2 e σ_2^2 , então a distribuição de

$$F = \frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} = \frac{\frac{(n_1-1)s_1^2}{\sigma_1^2}\Big/(n_1-1)}{\frac{(n_2-1)s_2^2}{\sigma_2^2}\Big/(n_2-1)}$$

ou, de outra forma,

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2}$$

possui uma distribuição F com $n_1 - 1$ e $n_2 - 1$ graus de liberdade.

Assim, para um determinado nível de confiança, é possível construir um intervalo de confiança com base nesta estatística, através da expressão

$$P\left(F_{1-\alpha/2,n_1-1,n_2-1} < F < F_{\alpha/2,n_1-1,n_2-1}\right) = 1-\alpha$$

tal como se pode ver na Figura 3.4.

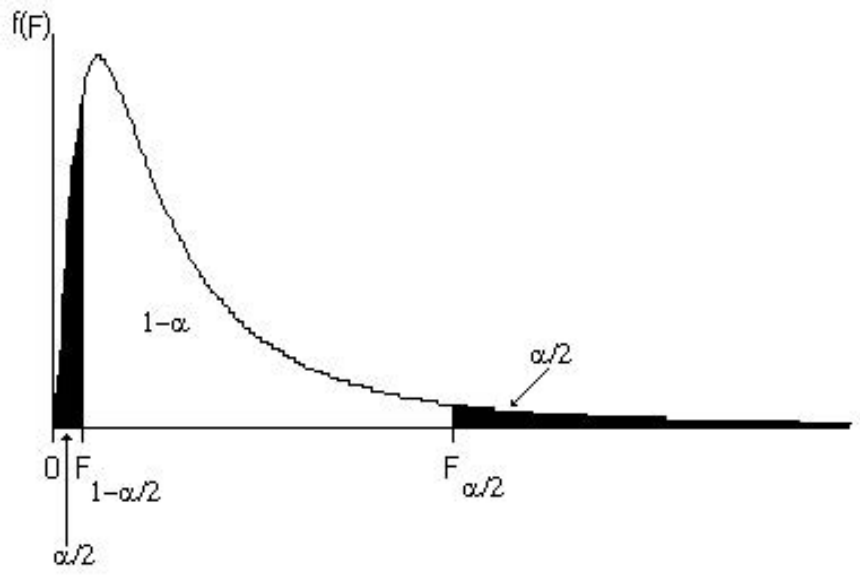


Figura 3.4: Intervalo de confiança com base na Distribuição F.

No entanto, esta expressão pode ser reescrita fazendo uso da relação $F_{1-\alpha/2,n_1-1,n_2-1} = 1/F_{\alpha/2,n_2-1,n_1-1}$.

Definição 3.6.1: Intervalo de Confiança de $(1-\alpha)100\%$ para $\frac{\sigma_1^2}{\sigma_2^2}$

$$\frac{s_1^2}{s_2^2}\frac{1}{F_{\alpha/2,n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}F_{\alpha/2,n_2-1,n_1-1}$$

onde $F_{\alpha/2,n_1-1,n_2-1}$ é o valor que localiza um área de $\frac{\alpha}{2}$ na cauda superior da distribuição F com n_1-1 no numerador e n_2-1 graus de liberdade no denominador e $F_{\alpha/2,n_2-1,n_1-1}$ é o valor que localiza um área de $\frac{\alpha}{2}$ na cauda superior da distribuição F com n_2-1 no numerador e n_1-1 graus de liberdade no denominador.

Este intervalo assume que ambas as populações, de onde são retiradas as amostras aleatórias e independentes, têm distribuições aproximadamente normais.

Exemplo 3.6.1: Intervalo de Confiança para a Razão das Variâncias

Numa fábrica de cimento, de duas máquinas de enchimento de sacos de 15 kg, foram recolhidas duas amostras de 13 e 9 sacos. Estas amostras apresentaram, respetivamente, os seguintes desvios padrão 0.071 e 0.075 kg. Estabeleça um intervalo de confiança de 90% para a razão das variâncias.

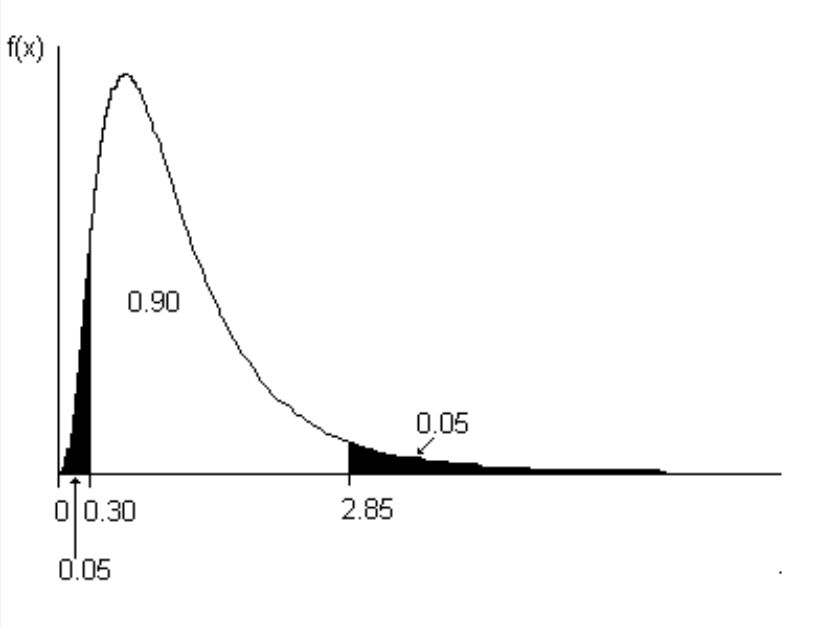
Solução

Para um nível de confiança de 90%, para $\nu_1 = n_1 - 1 = 12$ e $\nu_2 = n_2 - 1 = 8$ graus de liberdade, os valores correspondentes da distribuição são, respetivamente, $F_{0.05,12,8} = 3.28$ e $F_{0.05,8,12} = 2.85$, como se pode ver pela figura. Assim, o intervalo de confiança correspondente será,

$$\frac{(0.071)^2}{(0.075)^2}\frac{1}{3.28} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{(0.071)^2}{(0.075)^2}(2.85)$$

ou seja,

$$(0.87)\,\tfrac{1}{3.28} < \tfrac{\sigma_1^2}{\sigma_2^2} < (0.87)\,(2.85)$$
$$0.27 < \tfrac{\sigma_1^2}{\sigma_2^2} < 2.55$$



Uma vez que este intervalo inclui a possibilidade da razão ser igual a 1, não se pode concluir que as variâncias sejam diferentes.

Exercícios

1. Uma amostra aleatória do peso de jovens do sexo feminino, de tamanho n, é retirada duma população com média μ e desvio padrão σ. A média e o desvio padrão da amostra são $\bar{x} = 55$ kg e $s = 5.7$ kg. Calcule o intervalo de confiança de 95% para μ usando amostras com as seguintes dimensões: a) 30 b) 50 c) 100 Compare as amplitudes dos três intervalos.

2. Considere uma amostra aleatória do peso de jovens do sexo feminino, de tamanho $n = 40$, com uma média e desvio padrão, respetivamente, de $\bar{x} = 55$ kg e $s = 5.7$ kg. Determine os intervalos de confiança de 90%, 95% e 99%. Compare as amplitudes dos três intervalos.

3. Um amostra de 210 bebés do sexo masculino do Hospital Garcia de Orta tem uma média 3272 g e um desvio padrão de 564 g. Construa um intervalo de 95% para o peso médio dos bebés do sexo masculino. Supondo que pretendia um estimativa mais precisa, qual deveria ser a dimensão da amostra se a semi-amplitude do intervalo fosse no máximo 50 g?

4. Um amostra do mesmo Hospital de bebés do sexo feminino com a dimensão $n = 190$ tem uma média de 3148 g e um desvio padrão de 509 g. Pode concluir que os bebés do sexo masculino são mais pesados? Estabeleça um intervalo de confiança de 95% para a diferença

das médias. Assumindo que pretendia um intervalo com uma semi-amplitude inferior a 100 g, considerando amostras de igual dimensão, qual deveria ser o tamanho da amostra global?

5. Numa universidade, uma amostra aleatórias de 12 estudantes do sexo masculino foi selecionada. O comprimento médio da mão encontrado foi de 187.0 mm com um desvio padrão de 5.1 mm.

185	189	188	188	179	180	188	199	189	188	188	183

Construa um intervalo de confiança de 95% para o verdadeiro valor do comprimento médio.

6. A tabela apresenta a dieta calórica média diária (registada durante 10 dias) de um conjunto de 10 mulheres saudáveis. Essas mulheres participaram num programa de educação alimentar no sentido de reduzir a ingestão calórica diária. Verifique se existe uma diferença entre a ingestão diária de calorias antes e após o programa de educação alimentar. O que pode concluir? Construa o gráfico de caixa e bigodes para suportar as suas conclusões.

Antes	3037	2288	2394	2973	2310	2266	2692	1988	2399	2844	1818
Após	2823	2088	2177	2742	2115	2099	2504	1772	2208	2675	1628

7. As normas sanitárias para a água de consumo doméstico (Decreto-Lei 306/2007) impõem uma determinada concentração de cloro livre por forma a garantir que a água seja salubre e isenta de microrganismos patogénicos. Esta concentração, que se deve situar entre 0.2 a 0.6 mg/L, é ajustada na central de bombagem; contudo, é necessário verificar se a concentração se mantém ao longo da rede de distribuição. Em geral, são recolhidas amostras na central de bombagem e na rede em intervalos regulares. A tabela apresenta as concentrações de cloro na central e num ponto da rede, ao longo de 10 semanas. Verifique se existem diferenças significativas nas concentrações de cloro (mg/L) nos dois pontos de amostragem. O que pode concluir? Construa o gráfico de caixa e bigodes para suportar as suas conclusões.

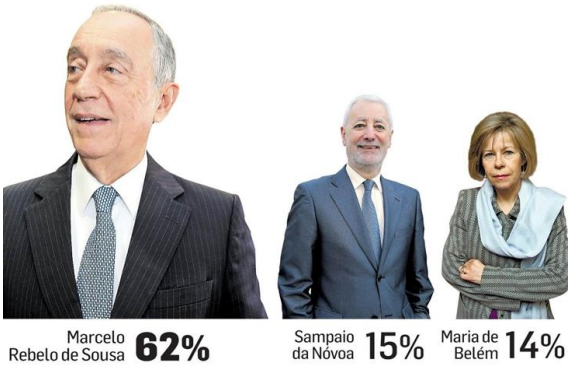
Central	0.52	0.37	0.45	0.54	0.48	0.53	0.55	0.24	0.46	0.52
Rede	0.25	0.36	0.17	0.15	0.10	0.25	0.30	0.43	0.55	0.61

8. O desgaste da cabeça do fémur conduz à implantação de uma cabeça de substituição numa liga metálica leve e resistente. Esta implantação é feita com um cimento especial, que alguns médicos suspeitam que possa diminuir a resistência do osso. No entanto, as opiniões dos ortopedistas dividem-se quanto à necessidade de introdução de um tampão que evite que o cimento se espalhe pelo espaço disponível. Para comparar o efeito do uso do tampão na resistência à flexão, foram efetuados vários implantes em animais de laboratório, tendo sido obtidos os seguintes resultados de resistência (Nm):

C/ Tampão	7.0	6.2	7.1	8.1	5.1	5.6
S/ Tampão	8.9	7.7	5.3	8.6	7.1	4.6

Estabeleça um intervalo de confiança de 95% para a diferença entre as resist encias médias dos dois métodos. O que pode concluir?

9. O Diário de Notícias, na sua edição de 11 de dezembro de 2015 publicou uma sondagem sobre a eleição para Presidente da República de Portugal.



A ficha técnica da sondagem dizia o seguinte: "Esta sondagem foi realizada pelo CESOP-Universidade Católica Portuguesa para a Antena 1, a RTP, o Jornal de Notícias e o Diário de Notícias nos dias 5 e 6 de dezembro de 2015. O universo alvo é composto pelos indivíduos com 18 ou mais anos recenseados eleitoralmente e residentes em Portugal Continental. Foram selecionadas aleatoriamente dezoito freguesias do país, tendo em conta a distribuição da população recenseada eleitoralmente por regiões NUT II e por freguesias com mais e menos de 3200 recenseados. A seleção aleatória das freguesias foi sistematicamente repetida até que os resultados eleitorais das últimas eleições legislativas nesse conjunto de freguesias (ponderado o número de inquéritos a realizar em cada uma) estivessem a menos de 1% dos resultados nacionais dos cinco maiores partidos. Os domicílios em cada freguesia foram selecionados por caminho aleatório e foi inquirido em cada domicílio o próximo aniversariante recenseado eleitoralmente na freguesia. Foram obtidos 1183 inquéritos válidos, sendo 58% dos inquiridos do sexo feminino, 34% da região Norte, 20% do Centro, 34% de Lisboa, 5% do Alentejo e 7% do Algarve. Todos os resultados obtidos foram depois ponderados de acordo com a distribuição de eleitores residentes no Continente por sexo, escalões etários, região e habitat na base dos dados do recenseamento eleitoral e das estimativas do INE. A taxa de resposta foi de 69%. A margem de erro máximo associado a uma amostra aleatória de 1183 inquiridos é de 2,9%, com um nível de confiança de 95%".

A estimativa dos resultados para Marcelo Rebelo de Sousa (MRS), tendo por base somente os 851 eleitores que afirmaram que iriam votar foi $x = 527/(831 = 0.62$. Calcule um intervalo de confiança de 95% para a estimativa da proporção de votantes em MRS. Se fosse pretendido um erro da estimativa inferior a 2%, qual deveria ser a dimensão da amostra?

10. Foi realizado um estudo em que se pretendeu comparar as percentagens de hipertensos na região Norte e na região Lisboa e Vale do Tejo. Um amostra aleatória de 1057 utentes dos Centros de Saúde da Região Norte apresentava 402 utentes hipertensos. A amostra da Região Lisboa e Vale do Tejo, com 1344 utentes dos quais 538 eram hipertensos. Estabeleça um intervalo de confiança de 90% para a diferença entre as duas proporções. O que pode concluir?

11. Uma fábrica de cervejas produz garrafas de 20 cl sendo conhecido que a quantidade em cada garrafa irá variar devido a vários fatores, alguns controláveis e outros não. Por essa razão, para além da quantidade média de enchimento, importa ter noção da variabilidade pois a fábrica perde dinheiro sempre que o volume nominal é ultrapassado e entra em incumprimento sempre que o volume é inferior ao especificado. Assim, a intervalos regulares são recolhidas 10 garrafas e o seu volume determinado. Uma amostra produziu os seguintes resultados:

Vol. (cl)	19.9	20.1	20.0	20.0	19.8	19.7	20.1	20.2	19.9	20.0

a que corresponde uma média $\bar{x} = 19.97$ e um desvio padrão de $s = 0.15$. Construa um intervalo de confiança de 90% para variância do enchimento.

12. No serviço de sangue de um hospital, a preparação das bolsas de transfusão é realizada em duas máquinas. Os requerimentos impõem que cada bolsa de sangue deve ter um valor mínimo de Hemoglobina (Hb) de 43 g. Foram recolhidas duas amostras de cada uma das máquinas que produziram os seguintes resultados:

Amostra 1	57	42	44	59	50	59	50	50	54	54	50		$\bar{x}_1 = 51.73$	$s_1 = 5.57$
Amostra 2	47	45	50	51	63	58	51	43	52	62	41	44	$\bar{x}_2 = 50.58$	$s_2 = 7.25$

Construa um intervalo de confiança de 95% para a razão das variâncias. O que pode concluir?

Soluções

1.

n	L_{Inf}	L_{Sup}	Dif
30	52.96	57.04	4.08
50	53.42	56.58	3.16
100	53.88	56.12	2.23

À medida que n aumenta, a amplitude do intervalo diminui

2.

n	L_{Inf}	L_{Sup}	Dif
90%	53.52	56.48	2.97
95%	53.23	56.77	3.53
99%	52.68	57.32	4.64

À medida que a confiança aumenta, a amplitude do intervalo aumenta

3.

L_{Inf}	L_{Sup}	Dif
3195.72	3348.28	152.57

Semi-amplitude 152.57/2=76.29

n	L_{Inf}	L_{Sup}	Dif
489	3222.01	3321.99	99.98

$n = (z_{\alpha/2} \times s/50)^2$

$n > 488.80; n = 489$

4.

L_{Inf}	L_{Sup}	Dif
18.85	229.15	210.31

Semi-amplitude 210.31/2=105.16

n	L_{Inf}	L_{Sup}	Dif
222	24.06	223.94	199.88

$n = (z_{\alpha/2}/100)^2 \times (s_1^2 + s_2^2)$

$n > 221.73; n = 222$

5. [183.7377, 190.2623]

```
# Intervalo de Confiança para a média, n<30, variância desconhecida
comp=c(185,189,188,188,179,180,188,199,189,188,188,183)
summary(comp) # medidas resumo
mean(comp) # média
sd(comp) # desvio padrão
length(comp)-1 # nº de graus de liberdade
qt(0.975,df=length(comp)-1) #quantil da distribuição t para uma confiança de 95%
error <- qt(0.975,df=length(comp)-1)*sd(comp)/sqrt(length(comp))
# erro da estimativa
```

```
error
L_Inf=mean(comp)-error # Limite Inferior
L_Inf
L_Sup=mean(comp)+error # Limite Superior
L_Sup
t.test(comp, conf.level = 0.95) # método alternativo de cálculo
```

6. $[184.4668, 211.5332]$

```
# Intervalo de Confiança para amostras emparelhadas
Antes=c(3037,2288,2394,2973,2310,2266,2692,1988,2399,2844,1818)
Apos=c(2823,2088,2177,2742,2115,2099,2504,1772,2208,2675,1628)
summary(Antes) # medidas resumo
summary(Apos) # medidas resumo
dif=Antes-Apos # diferenças emparelhadas
dif
mean(dif) # média
sd(dif) # desvio padrão
length(Antes)-1 # nº de graus de liberdade
qt(0.975,df=length(Antes)-1) #quantil da distribuição t para uma confiança de 95%
error <- qt(0.975,df=length(Antes)-1)*sd(dif)/sqrt(length(Antes))
# erro da estimativa
error
L_Inf=mean(dif)-error # Limite Inferior
L_Inf
L_Sup=mean(dif)+error # Limite Superior
L_Sup
t.test(Antes,Apos, conf.level = 0.95, paired = T) # método alternativo de cálculo
```

7. $[-0.00536725, 0.30336725]$

```
# Intervalo de Confiança para amostras emparelhadas
Central=c(0.52,0.37,0.45,0.54,0.48,0.53,0.55,0.24,0.46,0.52)
Rede=c(0.25,0.36,0.17,0.15,0.10,0.25,0.30,0.43,0.55,0.61)
summary(Central) # medidas resumo
```

```
summary(Rede) # medidas resumo
dif=Central-Rede # diferenças emparelhadas
dif
mean(dif) # média
sd(dif) # desvio padrão
length(Central)-1 # nº de graus de liberdade
qt(0.975,df=length(Central)-1) #quantil da distribuição t para uma confiança de 95%
error <- qt(0.975,df=length(Central)-1)*sd(dif)/sqrt(length(Central))
# erro da estimativa
error
L_Inf=mean(dif)-error # Limite Inferior
L_Inf
L_Sup=mean(dif)+error # Limite Superior
L_Sup
t.test(Central, Rede, conf.level = 0.95, paired = T) # método alternativo de cálculo
```

8. $[-2.395965, 1.362632]$

```
# Intervalo de Confiança para amostras independentes
C_T=c(7.0,6.2,7.1,8.1,5.1,5.6)
S_T=c(8.9,7.7,5.3,8.6,7.1,4.6)
summary(C_T) # medidas resumo
summary(S_T) # medidas resumo
mean(C_T) # média
sd(C_T) # desvio padrão
mean(S_T) # média
sd(S_T) # desvio padrão
dif=mean(C_T)-mean(S_T) # diferença das médias
df1=length(C_T)-1 # graus de liberdade
df1
df2=length(S_T)-1 # graus de liberdade
df2
sp2=(df1*var(C_T)+df2*var(S_T))/(df1+df2) #variância pesada
sp=sqrt(sp2) #desvio padrão
qt(0.975,df=df1+df2) #quantil da distribuição t para uma confiança de 95%
error=qt(0.975,df=df1+df2)*sp*sqrt(1/length(C_T)+1/length(S_T))
```

```
L_Inf=dif-error
L_Inf
L_Sup=dif+error
L_Sup
t.test(C_T,S_T, conf.level = 0.95, var.equal = T, paired = F)
```

9. $[0.6014274, 0.666924]$; $n = 2229$

```
# Intervalo de Confiança pata uma proporção
n=831 # nº de respostas
x=527 # nº de sucessos
phat=x/n # proporção de votantes em MRS
phat
SE=sqrt(phat*(1-phat)/n)
# erro padrão, aproximação à normal para grandes amostras
SE
error=qnorm(0.975)*SE # erro da estimativa
error
L_Inf=phat-error # Limite Inferior
L_Inf
L_Sup=phat+error # Limite Superior
L_Sup
prop.test(527,831,conf.level=0.95) # método alternativo de cálculo
n1=(qnorm(0.975)/0.02)^2*phat*(1-phat) # dimensão da amostra para um erro de 0.02
n1
```

10. $[-0.05293808, 0.01298617]$

```
# Intervalo para a diferença de proporções
n1=1057 # nº de respostas
x1=402 # nº de sucessos
phat1=x1/n1 # proporção de hipertensos
phat1
n2=1344 # nº de respostas
x2=538 # nº de sucessos
```

```
phat2=x2/n2 # proporção de hipertensos
phat2
dif=phat1-phat2
SE12=sqrt(phat1*(1-phat1)/n1+phat2*(1-phat2)/n2)
# erro padrão, aproximação à normal para grandes amostras
SE12
error=qnorm(0.95)*SE12 # erro da estimativa
error
L_Inf=dif-error # Limite Inferior
L_Inf
L_Sup=dif+error # Limite Superior
L_Sup
prop.test(c(402,538),c(402+(1057-402),538+(1344-538))) # método alternativo de cálculo
```

11. $[0.01188015, 0.06044908]$

```
# Intervalo de Confiança para a variância
Vol=c(19.9,20.1,20.0,20.0,19.8,19.7,20.1,20.2,19.9,20.0)
summary(Vol)
mean(Vol)
sd(Vol)
length(Vol)-1 # nº de graus de liberdade
qchisq(0.95,length(Vol)-1)
qchisq(0.05,length(Vol)-1)
L_Inf=(length(Vol)-1)*var(Vol)/qchisq(0.95,length(Vol)-1)
L_Inf
L_Sup=(length(Vol)-1)*var(Vol)/qchisq(0.05,length(Vol)-1)
L_Sup
```

12. $[0.2065362, 1.7345105]$

```
# razão de variâncias
S1=c(57, 42, 44, 59, 50, 59, 50, 50, 54, 54, 50)
S2=c(47, 45, 50, 51, 63, 58, 51, 43, 52, 62, 41, 44)
summary(S1)
```

```
mean(S1)
sd(S1)
summary(S2)
mean(S2)
sd(S2)
df1=length(S1)-1 # nº de graus de liberdade
df1
df2=length(S2)-1 # nº de graus de liberdade
df2
qf(0.95,df1,df2)
qf(0.95,df2,df1)
L_Inf=var(S1)/var(S2)*1/qf(0.95,df1,df2)
L_Inf
L_Sup=var(S1)/var(S2)*qf(0.95,df2,df1)
L_Sup
var.test(S1,S2, conf.level = 0.90)
```

Capítulo 4

Testes de Hipóteses

4.1 Introdução

Os intervalos de confiança, tal como foi visto no capítulo anterior, permitem fazer inferências acerca dos parâmetros da população. Outra abordagem possível são os chamados testes de hipóteses, onde o objetivo já não é estabelecer um intervalo, mas tomar uma decisão relativamente a um valor de um parâmetro. De qualquer forma, as duas abordagens estão relacionadas. Assim, suponha que o Serviço de Proteção do Ar pretende determinar se uma fábrica está a emitir um poluente acima da concentração permitida. A concentração máxima admissível na emissão de dióxido de enxofre é $2700\,\mathrm{mg/m^3}$. Aparelhos de registo automático da concentração na chaminé forneceram os seguintes resultados, recolhidos aleatoriamente ao longo de uma semana: $n = 36$, $\bar{x} = 2880\,\mathrm{mg/m^3}$, $s = 540\,\mathrm{mg/m^3}$. Sob o ponto de vista da proteção ambiental, o objetivo é demonstrar que o valor observado está acima do limite legal. Para tal, pode ser construído um intervalo de confiança unilateral,

$$\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} < \mu$$

que permite determinar o limite inferior acima do qual se tem, por exemplo, 95% de confiança de que a verdadeira concentração média estará contida (Figura 4.1).

Assim, para os resultados obtidos,

$$\begin{aligned} &2880 - 1.65\tfrac{540}{\sqrt{36}} < \mu \\ &2731.5 < \mu \end{aligned}$$

poder-se-á afirmar, com uma confiança de 95%, que o valor médio emitido está situado acima de $2731.5\,\mathrm{mg/m^3}$, valor acima do limite legal. Convém notar, contudo, que se o valor médio observado fosse 2800, já não seria possível afirmar que a fábrica estava a operar acima dos limites legais, pois

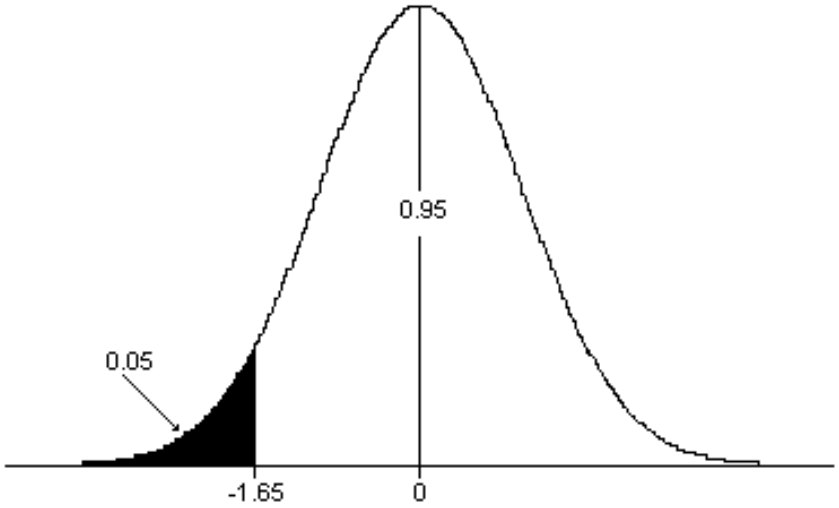


Figura 4.1: Intervalo de confiança unilateral.

o intervalo seria

$$2800 - 1.65\frac{540}{\sqrt{36}} < \mu$$
$$2651.5 < \mu$$

Esta abordagem procura responder à questão se a fábrica está ou não a emitir acima dos limites legais e, em face da evidência recolhida, pretende tomar uma decisão. Um teste de hipóteses procura traduzir este processo de tomada de decisão, pela formulação de duas hipóteses, das quais uma será aceite. Assim, no exemplo 3.4 do intervalo de confiança para a diferença entre os salários médios dos licenciados em Engenharia e em Relações Internacionais, as conclusões retiradas do intervalo, poderiam ser traduzidas pela escolha de uma de duas hipóteses: os salários médios são iguais ou são diferentes. Assim, como se pode ver pelos dois exemplos apresentados, existe uma relação entre estabelecer um intervalo de confiança e um teste de hipóteses. Na maior parte das situações, um teste de hipóteses pretende estabelecer alguma asserção relativa a um parâmetro de uma população, mas também, como veremos mais adiante, podem existir testes relativos à forma da distribuição de onde os dados são retirados, isto é, se a distribuição é normal ou exponencial, ou outra qualquer lei de probabilidade.

Definição 4.1.1: Teste estatístico

Um teste estatístico de hipóteses é uma asserção acerca de uma ou mais variáveis aleatórias. O teste implica a definição de duas hipóteses, referidas como **Hipótese Nula**, H_0, e **Hipótese Alternativa**, H_1. A hipótese é definida como simples se especifica completamente os parâmetros da distribuição; caso tal não se verifique, a hipótese diz-se composta.

No exemplo da emissão de dióxido de enxofre, a Hipótese Nula afirma que a média é $2700\,\mathrm{mg/m^3}$ ($H_0 : \mu = 2700$), e por isso é uma hipótese simples. A Hipótese Alternativa especifica que a média é superior a $2700\,\mathrm{mg/m^3}$, isto é, ($H_1 > 2700$), e como neste caso a média pode tomar qualquer valor acima de $2700\,\mathrm{mg/m^3}$, a hipótese é composta. No exemplo dos salários, a Hipótese Nula simples é que os salários médios dos licenciados são iguais, ou seja, a diferença é nula, isto é, ($H_0 : \mu_1 = \mu_2; \mu_1 - \mu_2 = 0$), e a Hipótese Alternativa composta especifica que a diferença é diferente de zero, isto é, ($H_1 : \mu_1 \neq \mu_2; \mu_1 - \mu_2 \neq 0$). No primeiro exemplo, o teste é unilateral, já que a hipótese alternativa somente considera valores maiores que $2700\,\mathrm{mg/m^3}$ e no segundo caso o teste é bilateral, pois a hipótese alternativa considera tanto diferenças positivas como negativas.

De todo o modo, a aceitação ou rejeição da Hipótese Nula faz com que um teste de hipóteses esteja sujeito a dois tipos de erro. Esta possibilidade de cometer dois tipos de erro, pode ser exemplificada pela decisão que um juiz tem de tomar em face de um réu. O juiz pode condenar um réu inocente ou ilibar um réu culpado. No exemplo de poluição atmosférica, o intervalo de confiança conduz à rejeição de que a fábrica esteja a operar nos limites legais, mas atendendo a que esta afirmação é baseada numa confiança de 95%, existe uma margem de erro, isto é, que em outras repetições da amostragem, o intervalo de confiança não conduzisse à rejeição. Por outro lado, se o valor observado da média tivesse sido $2800\,\mathrm{mg/m^3}$, o intervalo não conduziria à rejeição da hipótese nula (a fábrica opera dentro dos limites legais) e tal dever-se a um acaso resultante da recolha da amostra. A tabela procura explicitar este processo de decisão e os possíveis erros:

	Verdadeiro estado da natureza	
Decisão	A fábrica respeita os limites legais	A fábrica não respeita os limites legais
A fábrica não respeita os limites legais (Rejeitar H_0)	decisão incorreta X	decisão correta V
A companhia respeita os limites legais (Não rejeitar H_0)	decisão correta V	decisão incorreta X

4.2 Construção de um Teste de Hipóteses

A construção de um teste de hipóteses implica a definição de duas hipóteses, ditas nula e alternativa, em que a aceitação de uma das hipóteses implica a rejeição da outra. Este processo de decisão é baseado num valor de uma estatística de teste, para a qual é necessário definir a região de aceitação e de rejeição da Hipótese Nula. Assim, o resultado de um teste pode estar sujeito a dois tipos de erro.

Definição 4.2.1: Erros

Erro de Tipo I - corresponde à rejeição da Hipótese Nula, H_0, quando esta é verdadeira. A probabilidade de cometer um Erro de Tipo I é designada por α.

Erro de Tipo II - corresponde à não rejeição da Hipótese Nula, H_0, quando esta é falsa. A probabilidade de cometer um Erro de Tipo II é designada por β.

	Verdadeiro estado da natureza	
Decisão	H_0 verdadeira	H_0 falsa
Rejeitar H_0	Erro tipo I	Decisão correta
Não rejeitar H_0	Decisão correta	Erro tipo II

Exemplo 4.2.1: Teste de hipóteses

Suponha que pretende verificar qual a percentagem de estudantes que são contra o pagamento de propinas. O governo afirma que 25% dos estudantes se opõem, enquanto que as associações de estudantes afirmam que esta percentagem é muito superior. Numa determinada Universidade, uma amostra de 12 estudantes foi selecionada aleatoriamente. Oito estudantes declararam a sua oposição. Este resultado fornece evidência suficiente para rejeitar a afirmação do governo?

Solução

Cada estudante inquirido escolherá uma de duas alternativas. Assumindo X como uma variável aleatória binomial, p como a proporção de opositores ao projeto, a estatística de teste é x, o número de respostas não favoráveis ao projeto das propinas. Assim, a Hipótese Nula, será $H_0 : p = 0.25$, e a Hipótese Alternativa, será $H_1 : p > 0.25$. Se o número de estudantes que se opõem for muito grande, por exemplo, superior ou igual a 6, a Hipótese Nula será rejeitada em favor da Hipótese Alternativa, caso contrário será aceite. Assim, o

teste estatístico será formulado da seguinte forma:

$$H_0 : p = 0.25$$
$$H_1 : p > 0.25$$

com x como estatística de teste e $x \geq 6$ definindo a região de rejeição da Hipótese Nula, H_0. Como o número de estudantes que declararam a sua oposição é superior a seis, a Hipótese Nula será rejeitada em favor da Hipótese Alternativa.

Exemplo 4.2.2: Teste de hipóteses

No exemplo anterior, qual a probabilidade de tomar uma decisão errada?

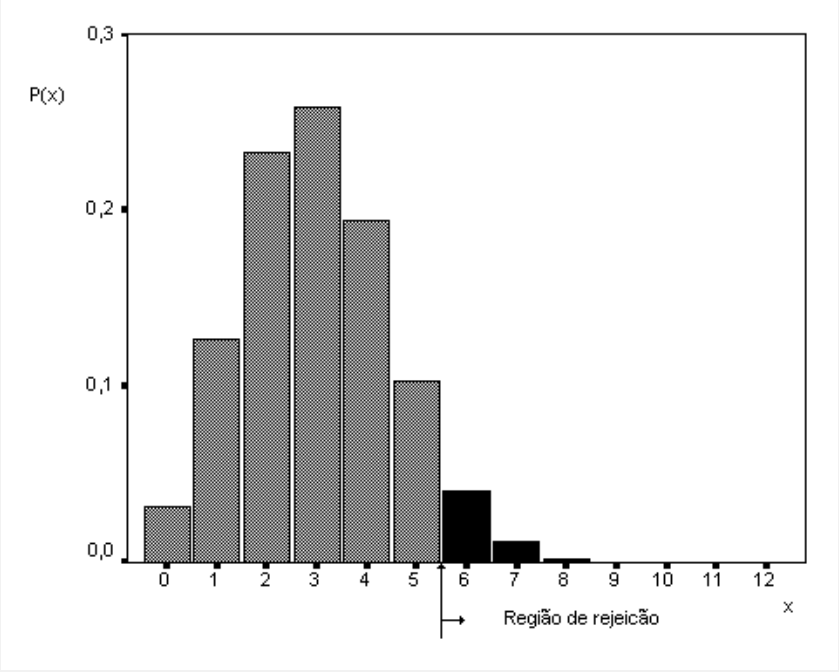
Solução

A figura apresenta o gráfico da distribuição binomial para $n = 12$, $p = 0.25$. Como se pode observar na figura, é possível obter valores iguais ou superiores a 6. A probabilidade de observar um valor igual ou superior a 6 corresponde à probabilidade de cometer um Erro de Tipo I (área sombreada a negro na figura). A tabela lista as probabilidades binomiais para cada valor de x, para $p = 0.25$.

x	$f(x)$	x	$f(x)$
0	0.0317	7	0.0115
1	0.1267	8	0.0024
2	0.2323	9	0.0004
3	0.2581	10	0.0000
4	0.1936	11	0.0000
5	0.1032	12	0.0000
6	0.0401		

Assim, a probabilidade de cometer um Erro de Tipo I é

$$\begin{aligned}\alpha =& P\left(\text{Erro de Tipo I}\right) = P\left(\text{Rejeitar } H_0 \,|H_0\right) \\ =& P\left(x \geq 6 \,|p = 0.25\right) = 0.0544\end{aligned}$$



Exemplo 4.2.3: Teste de Hipóteses

Suponha que a proporção p de estudantes que opõem ao projeto de pagamento de propinas é 0.70. Calcule as probabilidades de cometer um Erro de Tipo I e um Erro de Tipo II, quando a região de rejeição é, respetivamente, $x \geq 6$, $x \geq 5$ e $x \geq 7$.

Solução

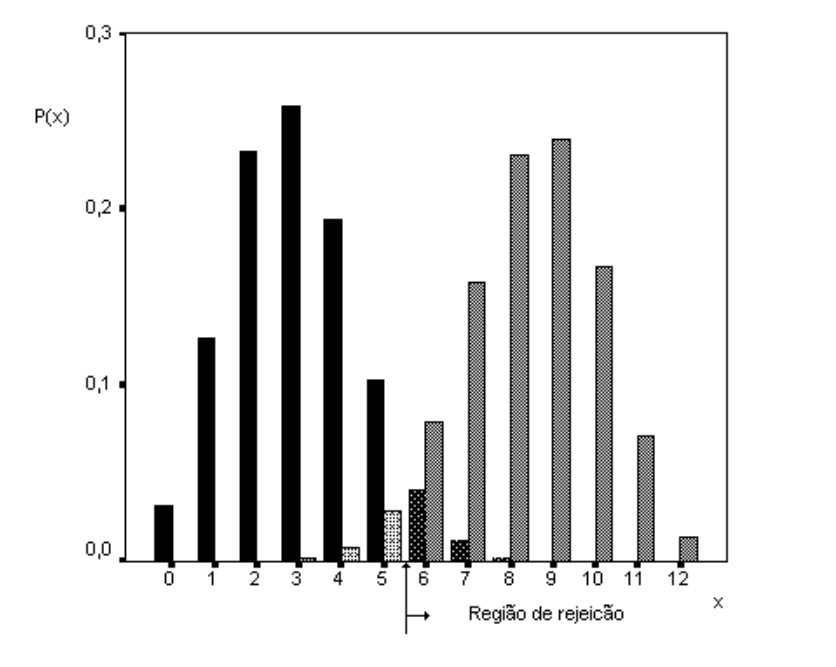
A probabilidade de cometer um Erro de Tipo I, para uma região de rejeição $x \geq 6$ já foi calculada no exemplo anterior. Um Erro de Tipo II corresponde a não rejeitar a Hipótese Nula quando esta é falsa, isto é, observar um valor (número de respostas opondo-se ao pagamento de propinas) inferior a 6, e assim concluir que a proporção p é menor ou igual a 0.25. A tabela lista as probabilidades binomiais para cada valor de x, para $p = 0.70$.

x	$f(x)$	x	$f(x)$
0	0.0000	7	0.1585
1	0.0000	8	0.2311
2	0.0002	9	0.2397
3	0.0015	10	0.1678
4	0.0078	11	0.0712
5	0.0291	12	0.0138
6	0.0792		

Assim, a probabilidade de cometer um Erro de Tipo II é

$$\begin{aligned}\beta =& P\,(\text{Erro de Tipo II}) = P\,(\text{Aceitar } H_0\,|H_1\,)\\ =& P\,(x < 6\,|p = 0.70\,) = 0.0386\end{aligned}$$

A seguinte figura ilustra as duas distribuições para os dois valores da proporção binomial, e assinala as áreas correspondentes aos Erros de Tipo I e de Tipo II.

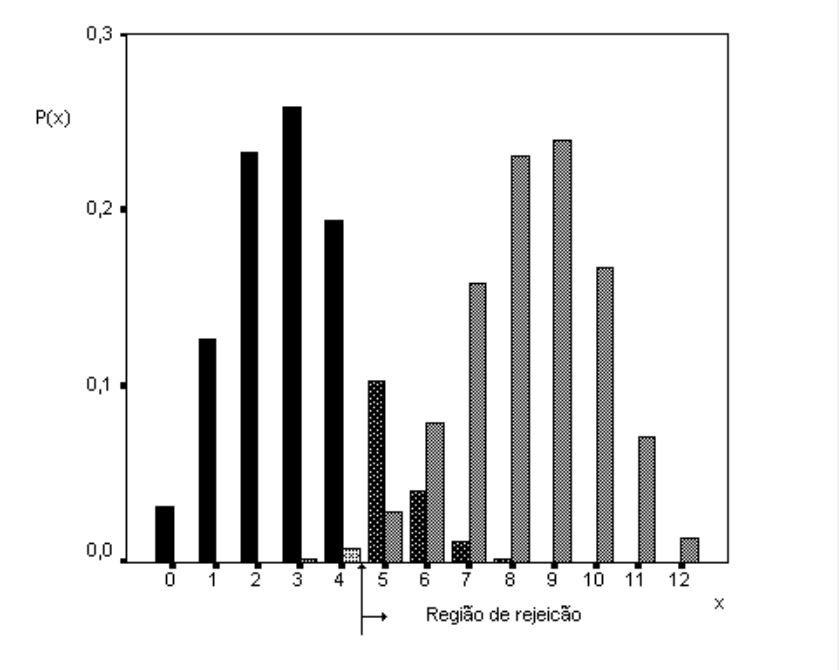


Se a região de rejeição for alterada, os valores correspondentes das probabilidades associadas aos Erros de Tipo I e de Tipo II vão ser alterados. Assim, e com base nas tabelas apresentadas da distribuição binomial,

- Região de Rejeição de H_0, $x \geq 5$

$$\begin{aligned}\alpha =& P\,(\text{Erro de Tipo I}) = P\,(\text{Rejeitar } H_0\,|H_0\,)\\ =& P\,(x \geq 5\,|p = 0.25\,) = 0.1576\end{aligned}$$

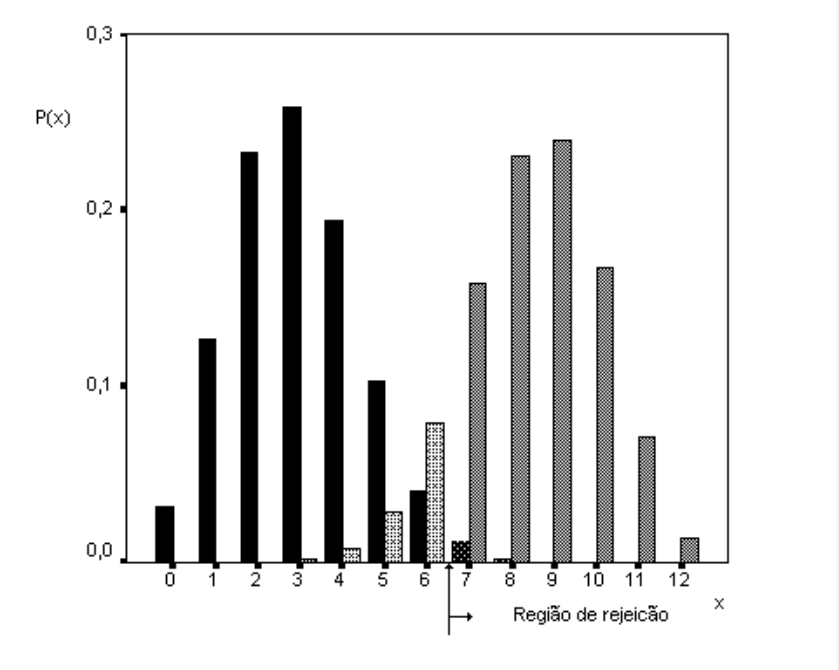
$$\begin{aligned}\beta =& P\,(\text{Erro de Tipo II}) = P\,(\text{Aceitar } H_0\,|H_1\,)\\ =& P\,(x < 5\,|p = 0.70\,) = 0.0095\end{aligned}$$



- Região de Rejeição de H_0, $x \geq 7$

$$\begin{aligned}\alpha =& P\left(\text{Erro de Tipo I}\right) = P\left(\text{Rejeitar } H_0 \,|H_0\right)\\ =& P\left(x \geq 7 \,|p = 0.25\right) = 0.0143\end{aligned}$$

$$\begin{aligned}\beta =& P\left(\text{Erro de Tipo II}\right) = P\left(\text{Aceitar } H_0 \,|H_1\right)\\ =& P\left(x < 7 \,|p = 0.70\right) = 0.1178\end{aligned}$$



As áreas associadas aos Erros de Tipo I e de Tipo II são apresentadas nas figuras para cada uma das regiões de rejeição.

Desta forma é possível verificar a relação entre os dois tipos de erro, isto é, a diminuição da probabilidade de um dos erros conduz ao aumento da probabilidade do outro e vice-versa. Mais adiante, será possível verificar que a maneira de diminuir simultaneamente as probabilidades associadas aos dois erros será através do aumento da dimensão da amostra.

4.3 Função Potência de um Teste

No exemplo anterior, foi possível calcular a probabilidade de cometer um Erro de Tipo II porque foi assumido um valor particular para a proporção p debaixo da Hipótese Alternativa, H_1. Esta assunção transformou a Hipótese Alternativa composta $H_1 : p > 0.25$ numa hipótese simples $H_1 : p = 0.70$. Contudo, na maioria das situações a Hipótese Alternativa será composta e, neste caso, o problema de avaliar o mérito de uma região crítica torna-se muito mais difícil. Nesse sentido, interessa determinar como afastamentos da Hipótese Nula se vão refletir na decisão de a aceitar ou rejeitar. Assim, é especialmente importante calcular a probabilidade de rejeitar a Hipótese Nula quando esta é falsa, o que corresponde à probabilidade de tomar uma decisão correta, à medida que o parâmetro varia. Esta probabilidade pode ser calculada da seguinte forma,

$$\begin{aligned} P\left(\text{Rejeitar } H_0 \,|H_0 \text{ falsa}\right) =& P\left(\text{Rejeitar } H_0 \,|H_1\right) \\ =& 1 - P\left(\text{Aceitar } H_0 \,|H_1\right) \\ =& 1 - P\left(\text{Erro de Tipo II}\right) \\ =& 1 - \beta \end{aligned}$$

Definição 4.3.1: Função potência

A função potência de um teste estatístico de uma Hipótese Nula contra uma Hipótese Alternativa é dada por

$$\pi\left(\theta\right) = \begin{cases} \alpha\left(\theta\right) & \text{para valores de } \theta \text{ debaixo de } H_0 \\ 1 - \beta\left(\theta\right) & \text{para valores de } \theta \text{ debaixo de } H_1 \end{cases}$$

As funções potência são muito importantes na comparação de testes estatísticos, nomeadamente na comparação de regiões de rejeição (regiões críticas) e, por exemplo, na comparação de diferentes escalas de classificação, como sejam as escalas de gravidade clínica. Assim, entre vários testes possíveis, o que correspondesse a uma curva mais próxima da ideal deveria ser escolhido. Se existe um teste para o qual a função potência é sempre superior a outro, isto é, para o qual a probabilidade de cometer um Erro de Tipo II é sempre menor, diz-se que esse teste é uniformemente mais potente. Convém referir, contudo, que a definição da região de rejeição não vai depender somente de critérios probabilísticos, mas também de outras considerações como os custos de uma decisão errada ou os efeitos dessa mesma decisão, como seja o caso de deteção de doenças graves.

Exemplo 4.3.1: Função potência

Em relação ao exemplo do número de estudantes que se opõem ao pagamento de propinas, considere agora o seguinte teste de hipóteses,

$$\begin{aligned} H_0 &: p \leq 0.25 \\ H_1 &: p > 0.25 \end{aligned}$$

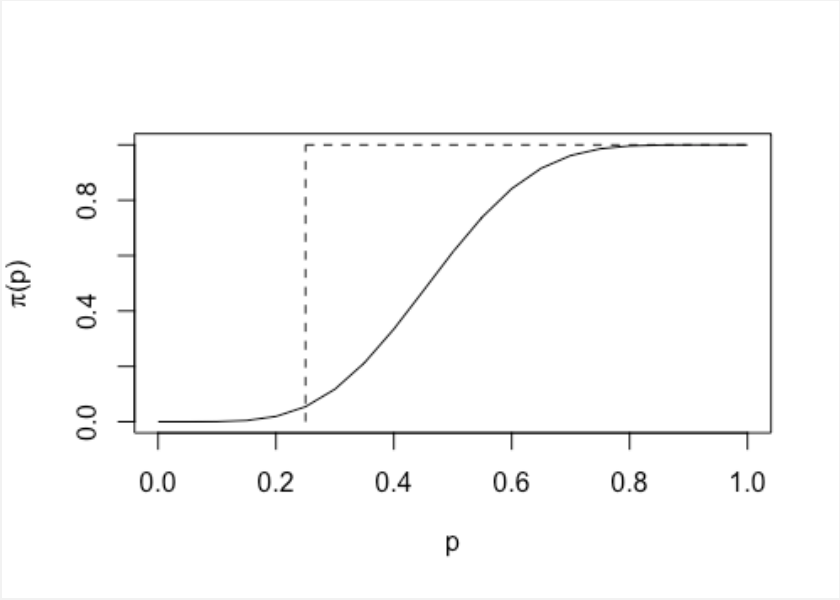
Calcule a função potência do teste.

Solução

As duas hipóteses são agora formuladas como hipóteses compostas. A função potência do teste pode assim ser calculada para diversos valores de p, tendo como região rejeição $x \geq 6$. A tabela lista as probabilidades de cometer cada tipo de erro em função dos diversos valores de p, bem como os valores da função potência.

p	α	β	$\pi(p)$	p	α	β	$\pi(p)$
0.00	0.0000		0.0000	0.55		0.2607	0.7393
0.05	0.0000		0.0000	0.60		0.1582	0.8418
0.10	0.0005		0.0005	0.65		0.0846	0.9154
0.15	0.0046		0.0046	0.70		0.0386	0.9614
0.20	0.0194		0.0194	0.75		0.0143	0.9857
0.25	0.0544		0.0544	0.80		0.0039	0.9961
0.30		0.8822	0.1178	0.85		0.0007	0.9993
0.35		0.7873	0.2127	0.90		0.0001	0.9999
0.40		0.6652	0.3348	0.95		0.0000	1.0000
0.45		0.5269	0.4731	1.00		0.0000	1.0000
0.50		0.3872	0.6128				

A figura apresenta o gráfico da função potência, e assim parece mais claro que à medida que p aumenta, a probabilidade de rejeitar a Hipótese Nula também aumenta. A figura apresenta a tracejado a curva da função potência ideal que permitiria uma decisão correta para todos os valores de p. Assim, entre vários testes possíveis, o que correspondesse a uma curva mais próxima da ideal deveria ser escolhido.



4.4 Razão de Verosimilhanças

O Teorema de Neyman-Pearson permite construir regiões críticas mais potentes em teste de hipóteses simples. Para hipóteses alternativas compostas, o método de construção de regiões críticas é baseado nos chamados testes da Razão de Verosimilhanças. No entanto, estes testes não serão necessariamente uniformemente mais potentes. Suponha que $x_1, x_2, \ldots, x_n$ constituem uma amostra aleatória de dimensão n de uma população com função densidade de probabilidade $f(x;\theta)$. A Hipótese Nula específica que θ pertence a um conjunto de valores possíveis Ω_0; a Hipótese Alternativa define o conjunto de valores possíveis para θ como sendo Ω_1, que não se sobrepõe a Ω_0, ou seja, Ω_1 é o complementar de Ω_0 ($\Omega = \Omega_0 \cap \Omega_1$). Assim, para testar a Hipótese Nula, $H_0 = \theta = \theta_0$, contra a Hipótese Alternativa, $H_1 = \theta > \theta_0$, Ω_0 é constituído pelo único valor θ_0 e Ω_1 pelo conjunto de todos os valores para os quais $\theta > \theta_0$. Seja $L(\hat{\theta_0})$ a função de verosimilhança com todos os parâmetros desconhecidos substituídos pelas suas estimativas de máxima verosimilhança, com a condição de que $\theta \in \Omega_0$, e $L(\hat{\theta})$ a função de verosimilhança obtida da mesma forma tal que $\theta \in \Omega$. O teste da razão das verosimilhanças é então definido pela sua razão.

Definição 4.4.1: Razão da Verosimilhança

Sejam Ω_0 e Ω_1 subconjuntos complementares do espaço paramétrico Ω, e

$$\lambda = \frac{L\left(\hat{\theta}_0\right)}{L\left(\hat{\theta}\right)}$$

onde $L(\hat{\theta_0})$ e $L(\hat{\theta})$ são os máximos das funções de Verosimilhança para todos os valores de θ em Ω_0 e Ω; a região crítica $\lambda \leq k$ onde $0 < k < 1$, define o teste da Razão das Verosimilhanças para testar a Hipótese Nula, $H_0 : \theta \in \Omega_0$, contra a Hipótese Alternativa, $H_1 : \theta \in \Omega_1$.

Os máximos das funções de Verosimilhança são sempre positivos e por isso $\lambda > 0$; por outro lado, uma vez que Ω_0 é um subconjunto de Ω, então $\lambda \leq 1$. Assim, quando a Hipótese Nula é falsa, o valor da razão será próximo de zero e, no caso contrário, os máximos das duas funções serão aproximadamente iguais, e logo λ será próximo de um.

Exemplo 4.4.1: Razão da Verosimilhança

Encontre a região crítica do teste da Razão das Verosimilhanças para o seguinte teste de hipóteses,

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

assumindo uma amostra aleatória de dimensão n de uma população normal com variância conhecida σ^2.

Solução

A Hipótese Nula só contém μ_0, logo $\Omega_0 = \{\mu_0\}$ e a Hipótese Alternativa define o seguinte conjunto $\Omega_1 = \{\mu \neq \mu_0\}$. A união destes dois conjuntos é $\Omega = \{\mu : -\infty < \mu < \infty\}$. A função de verosimilhança para a distribuição normal é dada por

$$L\left(\mu, \sigma^2\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\sum_{i=1}^{n} \frac{\left(x_i - \mu\right)^2}{2\sigma^2}\right]$$

Assim, a função de verosimilhança debaixo da Hipótese Nula será

$$L\left(\Omega_0\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\sum_{i=1}^{n} \frac{\left(x_i - \mu_0\right)^2}{2\sigma^2}\right]$$

e para a Hipótese Alternativa, atendendo a que a estimativa de Máxima Verosimilhança para μ é $\bar{x}$, tomará a seguinte forma,

$$L\left(\Omega\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\sum_{i=1}^{n} \frac{\left(x_i - \bar{x}\right)^2}{2\sigma^2}\right].$$

A Razão das Verosimilhanças será então,

$$\lambda = \frac{e^{-\frac{1}{2\sigma^2}\sum\left(x_i-\mu_0\right)^2}}{e^{-\frac{1}{2\sigma^2}\sum\left(x_i-\bar{x}\right)^2}} = e^{-\frac{1}{2\sigma^2}\left[\sum\left(x_i-\mu_0\right)^2-\sum\left(x_i-\bar{x}\right)^2\right]}$$

que após alguma manipulação pode ser escrito como,

$$\lambda = e^{-\frac{n}{2\sigma^2}\left(\bar{x}-\mu_0\right)^2} \leq k.$$

Aplicando logaritmos é possível obter

$$\left(\bar{x} - \mu_0\right)^2 \geq -\frac{2\sigma^2}{n} \ln k$$
$$\left(\bar{x} - \mu_0\right)^2 \geq K$$

onde o valor de K será determinado por forma a que o tamanho da região crítica seja α. Como $\bar{x}$ tem uma distribuição normal com média μ e variância $\frac{\sigma^2}{n}$, a região crítica desta razão é dada por

$$\left|\left(\bar{x} - \mu_0\right)\right| \geq z_{\alpha/2}\frac{\sigma}{n}$$

ou seja,

$$|z| \geq z_{\alpha/2}$$

com

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

o que corresponde a rejeitar a Hipótese Nula se o valor de z se situar fora do intervalo $-z_{\alpha/2}$, $z_{\alpha/2}$.

Exemplo 4.4.2: Razão de Verosimilhança

Considere uma amostra aleatória $x_1, x_2, \ldots, x_n$ de uma população normal com média e variância desconhecidas. Pretende-se efetuar o seguinte teste estatístico,

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

Encontre o teste da razão de verosimilhanças.

Solução

A Hipótese Nula só contém μ_0, logo $\Omega_0 = \{\mu_0\}$ e a Hipótese Alternativa define o seguinte conjunto $\Omega_1 = \{\mu > \mu_0\}$. A união destes dois conjuntos é $\Omega = \{\mu > \mu_0\}$. A função de verosimilhança para a distribuição normal é dada por

$$L\left(\mu, \sigma^2\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left[-\sum_{i=1}^{n} \frac{\left(x_i - \mu\right)^2}{2\sigma^2}\right]$$

a que corresponde a estimativa para a variância, debaixo da Hipótese Nula,

$$\hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \mu_0\right)^2.$$

Assim,

$$L\left(\Omega_0\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\hat{\sigma}_0^2}\right)^{n/2} e^{-n/2}.$$

O estimador de máxima verosimilhança para μ é $\bar{x}$. Assim, tendo em atenção a restrição $\mu > \mu_0$, o estimador de Máxima Verosimilhança é dado por $\hat{\mu} = \max(\bar{x}, \mu_0)$. O estimador para a variância, debaixo da Hipótese Alternativa, é

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \hat{\mu}\right)^2.$$

A função de Verosimilhança $L(\Omega)$ fica assim definida,

$$L\left(\Omega\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\hat{\sigma}^2}\right)^{n/2} e^{-n/2}$$

e a razão,

$$\lambda = \frac{L\left(\Omega_0\right)}{L\left(\Omega\right)} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} = \begin{cases} \left[\dfrac{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2}{\sum\limits_{i=1}^{n} (x_i - \mu_0)^2}\right] & \bar{x} > \mu_0 \\ 1 & \bar{x} \leq \mu_0 \end{cases}$$

Para valores pequenos de λ, por exemplo, $k < 1$, a Hipótese Nula será rejeitada. Assim,

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\mu_0)^2} < k^{2/n}$$

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu_0)^2} < k^{2/n}$$

$$\frac{1}{1+\frac{n(\bar{x}-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} < k^{2/n}$$

Esta desigualdade pode ser escrita como,

$$\frac{n(\bar{x}-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} > \frac{1}{k^{2/n}} - 1 = k'$$

$$\frac{n(\bar{x}-\mu_0)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} > (n-1)\,k'$$

Tendo em atenção que

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

a desigualdade toma a seguinte forma,

$$\frac{\sqrt{n}(\bar{x}-\mu_0)}{s} > \sqrt{(n-1)\,k'}$$

que corresponde à estatística t derivada anteriormente.

Os testes estatísticos podem ser obtidos através do método da Razão das Verosimilhanças, produzindo, em geral, o melhor teste em termos de potência. Contudo, nem sempre o resultado obtido é uma estatística com distribuição conhecida, como nos exemplos apresentados. No entanto, para grandes amostras, e assumindo condições de regularidade como a existência de derivadas relativamente aos parâmetros, é possível usar o seguinte teorema.

Teorema 4.4.1. *Para grandes amostras, a distribuição de $-2ln\lambda$ aproxima-se da distribuição de Qui-Quadrado com 1 grau de liberdade.*

Para situações em que estejam envolvidos mais do que um parâmetro desconhecido, impondo r restrições, o número de graus de liberdade na aproximação de Qui-Quadrado de $-2ln\lambda$ é igual a r.

Exemplo 4.4.3: Razão das Verosimilhanças

Considerando o exemplo do teste para média da distribuição normal, assumindo uma amostra aleatória de dimensão n, de uma população normal com variância conhecida σ^2, e usando o teorema anterior, deduza a expressão para a estatística de teste.

Solução

A Razão das Verosimilhanças é dada por

$$\lambda = \frac{e^{-\frac{1}{\sigma^2}\sum(x_i-\mu_0)^2}}{e^{-\frac{1}{\sigma^2}\sum(x_i-\bar{x})^2}} = e^{-\frac{1}{\sigma^2}\left[\sum(x_i-\mu_0)^2-\sum(x_i-\bar{x})^2\right]},$$

ou seja,

$$-2\ln\lambda = \frac{n}{\sigma^2}(\bar{x}-\mu_0)^2 = \left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)^2$$

que é o valor de uma variável aleatória de Qui-Quadrado com 1 grau de liberdade.

4.5 Classificação dos Testes

Num teste de hipóteses acerca do valor de um parâmetro, em geral, a Hipótese Nula será formulada da seguinte forma: $H_0 : \theta = \theta_0$. Se a estimativa pontual $\hat{\theta}$ está próxima de θ_0, o teste deverá conduzir à não rejeição da Hipótese Nula. No entanto, se for adotada uma política de rejeição da Hipótese Nula para valores de $\hat{\theta}$ muito maiores ou muito mais pequenos que θ_0, está-se em presença de uma Hipótese Alternativa bilateral, do tipo $\theta \neq \theta_0$. Assim, a região de rejeição deve ser localizada nas duas caudas da distribuição da estatística de teste a $\hat{\theta}$. Por outro lado, se a observação de valores de $\hat{\theta}$ muito baixos conduzisse à rejeição da Hipótese Nula, a Hipótese Alternativa unilateral seria $\theta < \theta_0$, a que corresponderia uma região de rejeição (região crítica) na cauda esquerda da distribuição de $\hat{\theta}$. Da mesma forma, a rejeição da Hipótese Nula para grandes valores definiria uma Hipótese Alternativa unilateral, $\theta > \theta_0$, avaliada na cauda direita da distribuição de $\hat{\theta}$.

No caso do teste bilateral a $H_0 : \mu = \mu_0$, a que corresponde a seguinte Hipótese Alternativa, $H_1 : \mu \neq \mu_0$, a técnica da Razão das Verosimilhanças, conduziu à seguinte região crítica,

$$|\bar{x}-\mu_0| \geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

ou seja,

$$\bar{x} \leq \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

ou

$$\bar{x} \geq \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

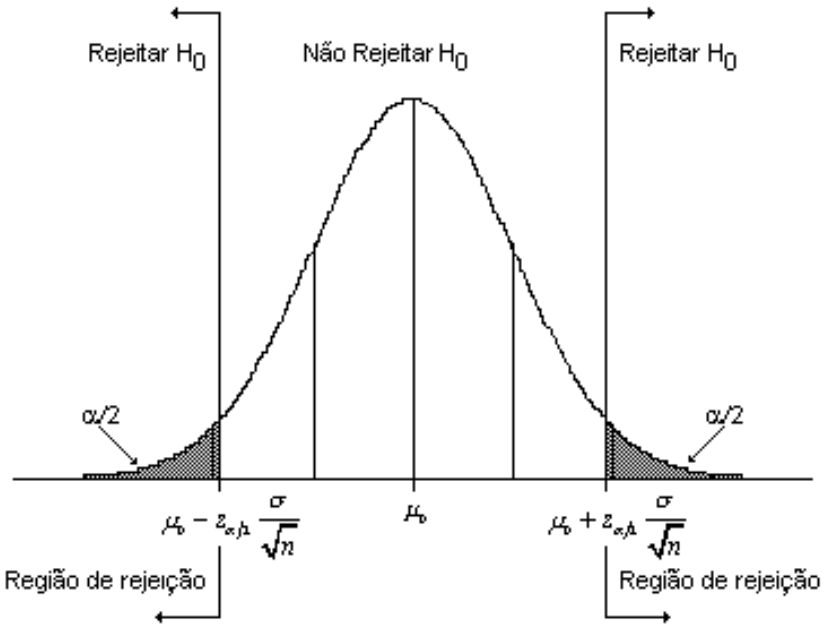


Figura 4.2: Região crítica para a Hipótese Alternativa $H_1 : \mu \neq \mu_0$.

tal como a Figura 4.2 descreve.

Em termos da variável normal padrão, a Hipótese Nula seria rejeitada se $z < -z_{\frac{\alpha}{2}}$ ou $z > z_{\frac{\alpha}{2}}$, onde

$$z = \frac{x - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

De forma semelhante, no caso de um teste unilateral, em que a Hipótese Alternativa fosse definida como $H_1 : \mu < \mu_0$, a região de rejeição seria a correspondente à Figura 4.3. A rejeição da Hipótese Nula, em termos da variável normal padrão, é definida como $z < -z_\alpha$.

A região de rejeição para a Hipótese Alternativa, $H_1 : \mu > \mu_0$, corresponde à assinalada na Figura 4.4, e em termos de variável normal padrão a $z > z_\alpha$.

A realização de um teste de hipóteses pressupõe o seguinte conjunto de passos:

1. Especificação da Hipótese Nula, H_0, acerca de um ou mais parâmetros da população e da Hipótese Alternativa, H_1, apropriada, que será aceite se a Hipótese Nula for rejeitada.

2. Definir a região crítica de tamanho α, tendo por base uma estatística apropriada.

3. Cálculo de uma estatística a partir dos dados da amostra.

4. Decidir, em função do valor observado para a estatística de teste, se a Hipótese Nula, H_0, é aceite ou rejeitada.

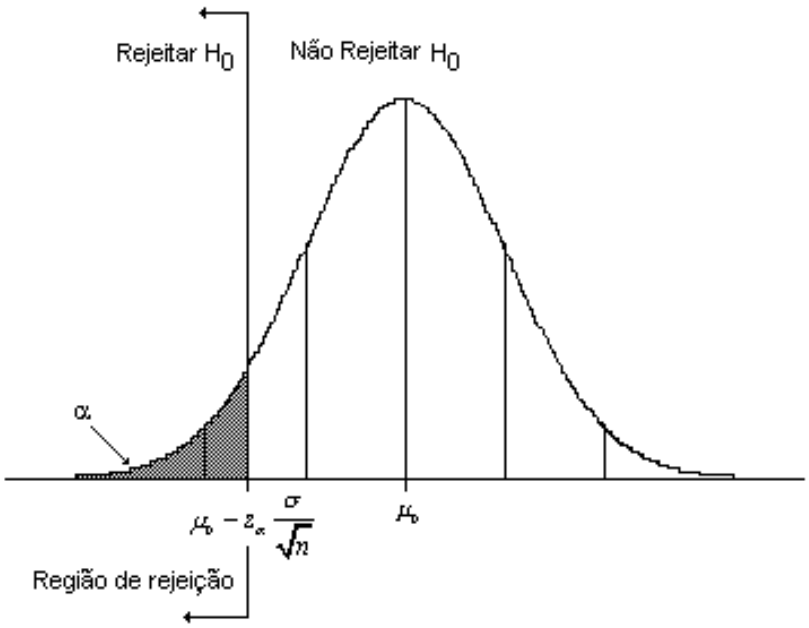


Figura 4.3: Região crítica para a Hipótese Alternativa $H_1 : \mu < \mu_0$.

4.6 Testes de Hipóteses acerca das Médias

Os testes de hipóteses relativos a médias e diferença de médias são baseados na distribuição normal, isto é, assumindo que as amostras são retiradas de populações normais, ou que a dimensão das amostras é suficientemente grande para justificar a aproximação à normal. A técnica da razão das verosimilhanças fornece seguinte estatística

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

baseada na distribuição normal.

O Teorema Limite Central garante que a distribuição de z é aproximadamente normal qualquer que seja a distribuição da população amostrada, pelo que não existem restrições à aplicabilidade deste teste.

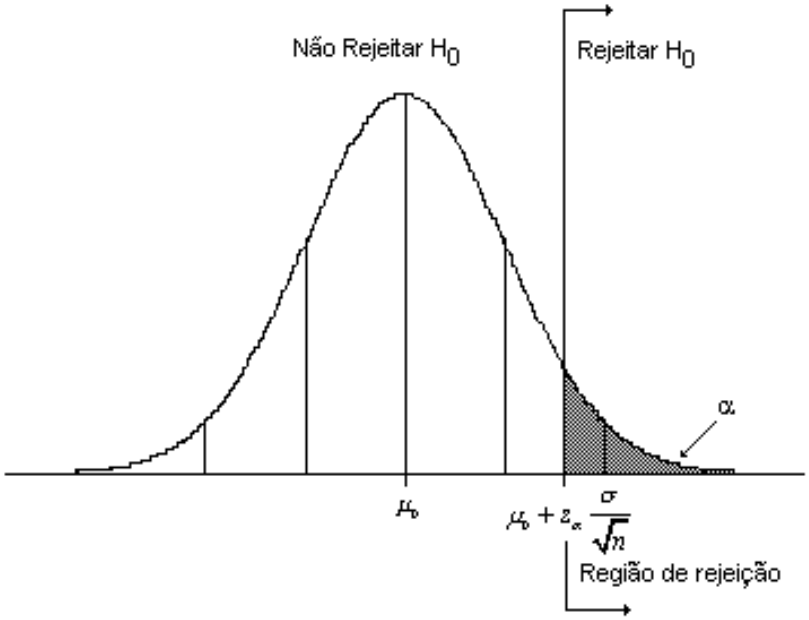


Figura 4.4: Região crítica para a Hipótese Alternativa, $H_1 : \mu > \mu_0$.

Definição 4.6.1: Teste de Hipóteses acerca da Média da População, μ (Grandes Amostras, $n \geq 30$)

Teste Unilateral	Teste Bilateral
$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu > \mu_0$	$H_1 : \mu \neq \mu_0$
$(H_1 : \mu < \mu_0)$	

Estatística

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \approx \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Região de Rejeição	Região de Rejeição
$z > z_\alpha$	$\lvert z \rvert > z_{\alpha/2}$
$(z < -z_\alpha)$	

onde z_α e $z_{\alpha/2}$ são, respetivamente, os valores da distribuição normal padrão tal que $P(z > z_\alpha) = \alpha$ e $P\left(z > z_{\alpha/2}\right) = \alpha/2$.

Exemplo 4.6.1: Teste de Hipóteses acerca da média

Suponha que a Inspeção das Atividades Económicas quer verificar se os sacos de cimento de uma determinada fábrica têm um peso médio de 15 kg. Para tal recolheu uma amostra aleatória de 50 sacos, tendo encontrado uma média de 14.81 kg com um desvio padrão de 0.06 kg. Permitem os dados concluir que a fábrica está a fornecer sacos com um peso inferior ao especificado? Assuma $\alpha = 0.05$.

Solução

1. Formulação das hipóteses

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0$$

2. Região crítica

$$z \leq -z_{0.05} = -1.65$$

3. Teste estatístico

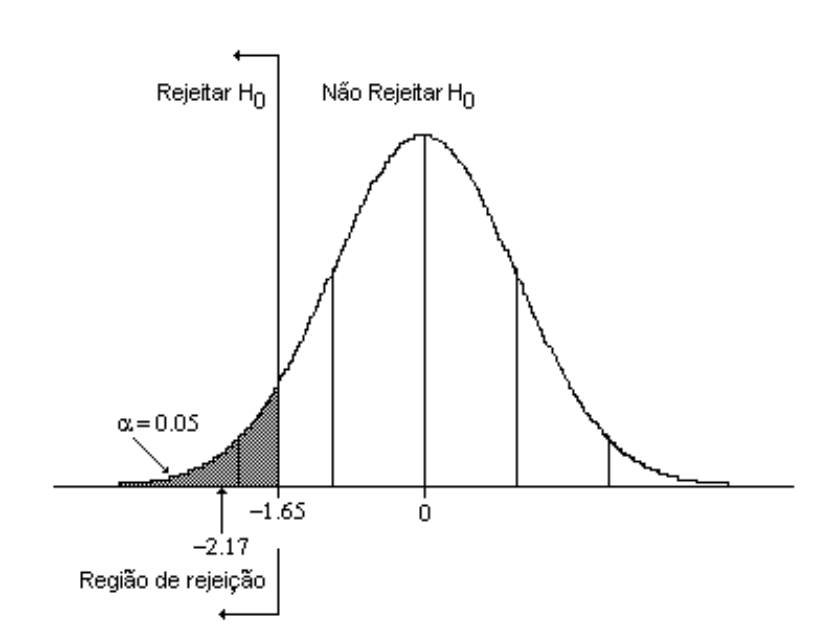
$$z \approx \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{14.81 - 15}{0.62/\sqrt{50}} = -2.17$$

4. Decisão

   Rejeitar a Hipótese Nula, ou seja, os sacos têm, em média, um peso inferior a 15 kg.

A seguinte figura mostra a região de rejeição. Convém notar que assumir que a Hipótese Nula é rejeitada para valores de $z < -1.65$, corresponde a fixar a região de rejeição para valores da média inferiores a

$$\bar{x} < \mu_0 - z_\alpha \frac{s}{\sqrt{n}} = 14.86$$



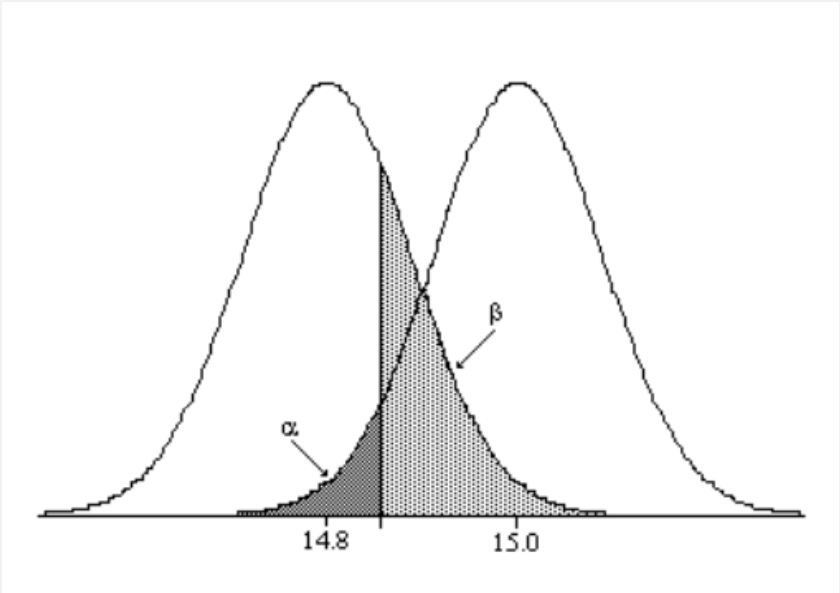
O intervalo de confiança unilateral, com base na mesma amostra, será definido como

$$\mu < \bar{x} + z_\alpha \frac{s}{\sqrt{n}} = 14.81 + 1.65 \frac{0.62}{\sqrt{50}}$$
$$\mu < 14.95$$

o que permite concluir, com 95% de confiança, que a média se encontra no intervalo $(-\infty, 14.95]$, que não inclui o valor de 15 kg.

Convém referir que o teste realizado é equivalente, em termos de região crítica, ao teste em que a Hipótese Nula tivesse sido formulada como $H_0 : \mu \geq \mu_0$ contra a mesma Hipótese Alternativa. Neste caso, contudo, representaria a probabilidade máxima de cometer um Erro de Tipo I, para qualquer valor assumido debaixo da Hipótese Nula.

O teste realizado garante que a probabilidade de cometer um Erro de Tipo I é inferior a 5%. Como a Hipótese Alternativa foi formulada como uma hipótese composta, a determinação do Erro de Tipo II implicaria o cálculo da probabilidade para valores de $\mu < 15$. Assumindo, por exemplo, que o verdadeiro valor da média da população de sacos era 14.8, a relação entre os dois tipos de erro é apresentada na seguinte figura.



Quando a dimensão das amostras é pequena, isto é, $n < 30$, e σ^2 desconhecido, o teste da razão das verosimilhanças dá origem a uma estatística

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

baseada na distribuição t-Student, com $n - 1$ graus de liberdade.

Definição 4.6.2: Teste de Hipóteses acerca da Média da População, μ (Amostras pequenas, $n < 30$)

Teste Unilateral	Teste Bilateral
$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0$
$H_1 : \mu > \mu_0$	$H_1 : \mu \neq \mu_0$
$(H_1 : \mu < \mu_0)$	

Estatística

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Região de Rejeição	Região de Rejeição
$t > t_\alpha$	$\lvert t \rvert > t_{\alpha/2}$
$(t < -t_\alpha)$	

onde t_α e $t_{\alpha/2}$ são, respetivamente, os valores da distribuição t-Student com $n-1$ graus de liberdade, tal que $P(t > t_\alpha) = \alpha$ e $P\left(t > t_{\alpha/2}\right) = \alpha/2$.

Este teste requer que a distribuição de frequências de onde a amostra foi retirada seja aproximadamente normal. Se a amostra se desvia muito da normalidade, este teste pode conduzir a conclusões erradas, sendo então preferível usar um teste não paramétrico.

Exemplo 4.6.2: Teste de Hipóteses acerca da média

Uma máquina produz parafusos com 2.5 cm de comprimento. No entanto, se os parafusos forem demasiado curtos ou longos, serão rejeitados. Neste caso a máquina necessita de ser ajustada. Para tal, uma amostra de parafusos é recolhida, a intervalos regulares, para verificar se os parafusos estão a ser produzidos com o comprimento médio de 2.5 cm. Suponha que foi recolhida uma amostra de 16 parafusos, com uma média $\bar{x} = 2.52$ cm e um desvio padrão $s = 0.04$ cm. Há evidência suficiente para assumir que a máquina não está a produzir segundo a especificação, isto é, que a máquina está fora de controlo? Use $\alpha = 0.05$.

Solução

1. Formulação das hipóteses
$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

2. Região crítica
$$|t| \geq t_{0.005} = 2.947$$

3. Teste estatístico
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2.52 - 2.5}{0.04/\sqrt{16}} = 2.00$$

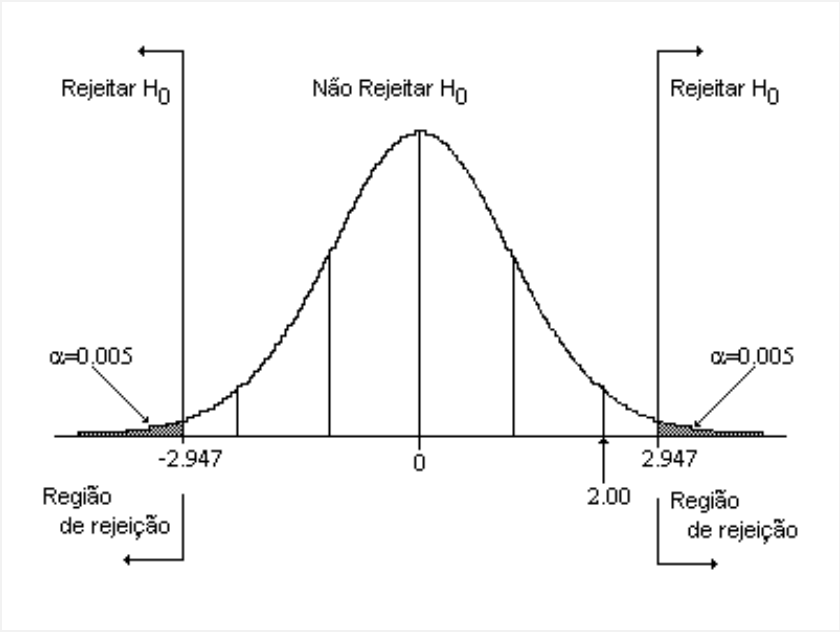
4. Decisão

   Não rejeitar a Hipótese Nula, ou seja, os parafusos têm um comprimento médio de 2.5 cm.

A figura mostra a região de rejeição para o exemplo 4.6.2. Como se pode ver, a Hipótese Nula seria rejeitada para valores de $|t| > 2.947$, ou seja para valores do comprimento superiores a 2.529 ou inferiores a 2.471. Para a amostra recolhida, o intervalo de confiança correspondente é

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$
$$2.52 \pm 2.947\frac{0.04}{\sqrt{16}}$$
$$2.491 < \mu < 2.549$$

que compreende o valor 2.5, logo conduzindo à mesma conclusão do teste.



4.7 Teste de Hipóteses acerca da Diferença entre Médias

Em muitos dos problemas o objetivo é testar se existem diferenças entre as médias de duas populações. Assumindo duas amostras independentes de dimensão n_1 e n_2, recolhidas de duas populações normais com médias μ_1 e μ_2, e variâncias σ_1^2 e σ_2^2, o teste da Razão das Verosimilhanças, associado à Hipótese Nula $H_0 : \mu_1 - \mu_2 = D_0$, contra a Hipótese Alternativa bilateral ou unilateral, produz a seguinte estatística,

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

Quando as variâncias são desconhecidas, mas as amostras são de grande dimensão, a estatística anterior pode ser usada, substituindo as variâncias pelas respetivas variâncias amostrais, s_1^2 e s_2^2.

Definição 4.7.1: Teste de Hipóteses acerca da Diferença entre Médias, $\mu_1 - \mu_2$ (Amostras independentes)

Teste Unilateral	Teste Bilateral
$H_0 : \mu_1 - \mu_2 = D_0$	$H_0 : \mu_1 - \mu_2 = D_0$
$H_1 : \mu_1 - \mu_2 > D_0$	$H_1 : \mu_1 - \mu_2 \neq D_0$
$(H_1 : \mu_1 - \mu_2 < D_0)$	

Estatística

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Região de Rejeição	Região de Rejeição
$z > z_\alpha$	$\lvert z \rvert > z_{\alpha/2}$
$(z < -z_\alpha)$	

onde z_α e $z_{\alpha/2}$ são, respetivamente, os valores da distribuição normal padrão tal que $P(z > z_\alpha) = \alpha$ e $P\left(z > z_{\alpha/2}\right) = \alpha/2$. D_0 é o valor especificado para a diferença $\mu_1 - \mu_2$; em geral, na maioria das situações, pretende-se testar a não existência de diferenças entre as médias das duas populações e, nesse caso, $D_0 = 0$.

A aplicação deste teste requer que os tamanhos das amostras sejam suficientemente grandes, isto é, $n_1 \geq 30$ e $n_2 \geq 30$, e que as amostras sejam, recolhidas de uma forma independente.

Exemplo 4.7.1: Teste de Hipóteses acerca da Diferença entre Médias

Uma fábrica de calçado possui duas linhas de montagem. O engenheiro de produção pretende testar uma nova organização da sequência das operações de montagem. Para tal, reorganizou a linha 2 e, ao fim de um mês, registou o tempo médio de montagem de um determinado modelo em cada uma das linhas. Os resultados obtidos foram os seguintes:

Linha 1	$n_1 = 36$	$\bar{x}_1 = 10.9$ min	$s_1 = 0.5$ min
Linha 2	$n_2 = 40$	$\bar{x}_2 = 10.9$ min	$s_2 = 0.4$ min

Permitem os dados concluir que a nova organização é mais eficiente?

Solução

1. Formulação das hipóteses

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 > 0$$

2. Região crítica

$$z \geq z_{0.01} = 2.33$$

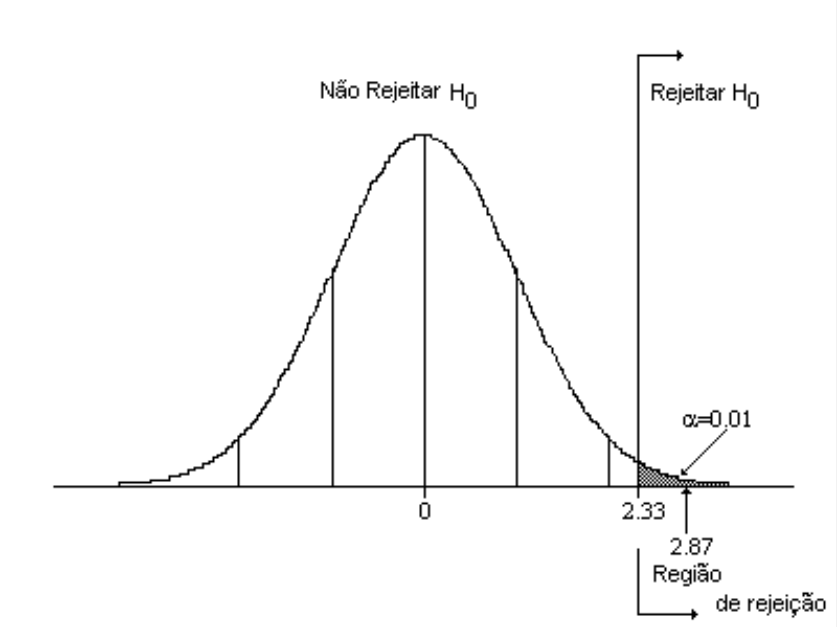
3. Teste estatístico

$$z \approx \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(10.9 - 10.6) - 0}{\sqrt{\frac{0.5^2}{36} + \frac{0.4^2}{40}}} = 2.87$$

4. Decisão

   Rejeitar a Hipótese Nula, ou seja, a nova organização da linha 2 é mais eficiente.

A Hipótese Nula de igualdade das médias dos tempos de montagem é rejeitada já que o valor observado para a estatística cai na região de rejeição. Assim, pode-se concluir que a montagem na linha 2 é efetuada com um menor tempo médio, sendo, por isso, mais eficiente. A seguinte figura mostra a região de rejeição.



Se as variâncias populacionais são desconhecidas e as amostras são de pequena dimensão, é possível ainda testar a existência de diferenças entre as médias, com base na seguinte estatística,

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{com} \quad S_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

obtida pela técnica da razão das verosimilhanças e assumindo uma variância desconhecida mas igual para as duas populações de onde as amostras são retiradas. Esta estatística segue a distribuição de t-Student com $n_1 + n_2 - 2$ graus de liberdade.

Definição 4.7.2: Teste de Hipóteses acerca da Diferença entre Médias, $\mu_1 - \mu_2$ (Amostras independentes e de pequena dimensão)

Teste Unilateral	Teste Bilateral
$H_0 : \mu_1 - \mu_2 = D_0$	$H_0 : \mu_1 - \mu_2 = D_0$
$H_1 : \mu_1 - \mu_2 > D_0$	$H_1 : \mu_1 - \mu_2 \neq D_0$
$(H_1 : \mu_1 - \mu_2 < D_0)$	

Estatística

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Região de Rejeição	Região de Rejeição
$t > t_\alpha$	$\lvert t \rvert > t_{\alpha/2}$
$(t < -t_\alpha)$	

com

$$S_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

e onde t_α e $t_{\alpha/2}$ são, respetivamente, os valores da distribuição t-Student com $n_1 + n_2 - 2$ graus de liberdade, tal que $P(t > t_\alpha) = \alpha$ e $P\left(t > t_{\alpha/2}\right) = \alpha/2$. D_0 é o valor especificado para a diferença $(\mu_1 - \mu_2)$; em geral, na maioria das situações, pretende-se testar a não existência de diferenças entre as médias das duas populações e, nesse caso, $D_0 = 0$.

A aplicabilidade do teste exige que as populações de onde as amostras são retiradas sejam aproximadamente normais, com variâncias populacionais iguais, $\sigma_1^2 = \sigma_2^2$. Além disso, as amostras devem ser recolhidas de uma forma independente.

Se a assunção da normalidade não é verificada, é aconselhável o uso de um teste não paramétrico.

Exemplo 4.7.2: Teste de Hipóteses acerca da Diferença entre Médias

O desgaste da cabeça do fémur conduz à implantação de uma cabeça de substituição composta por uma liga metálica leve e resistente. Esta implantação é feita com um cimento especial, que alguns médicos suspeitam que possa diminuir a resistência do osso. No entanto, as opiniões dos ortopedistas dividem-se quanto à necessidade de introdução de um tampão que evite que o cimento se espalhe pelo espaço disponível. Para comparar o efeito do uso do tampão na resistência à flexão, foram efetuados vários implantes em animais de laboratório, tendo sido obtidos os seguintes resultados de resistência (Nm):

C/ Tampão	7.0	6.2	7.1	8.1	5.1	5.6
S/ Tampão	8.9	7.7	5.3	8.6	7.1	4.6

O que pode concluir acerca do efeito do tampão na resistência à flexão?

Solução

$$\bar{x}_1 = 6.517 \qquad s_1 = 1.098$$
$$\bar{x}_2 = 7.033 \qquad s_2 = 1.750$$

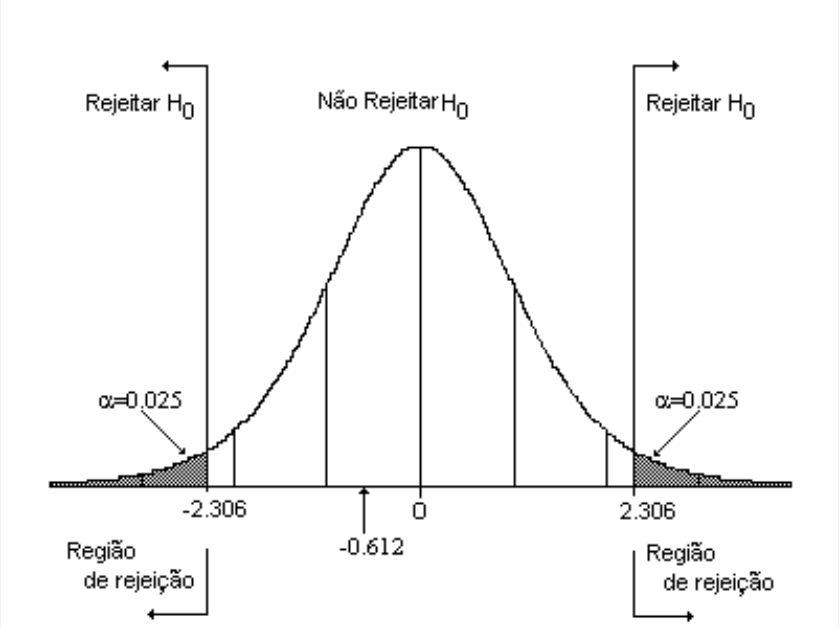
1. Formulação das hipóteses
$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

2. Região crítica
$$|t| \geq t_{0.025} = 2.306$$

3. Teste estatístico
$$S_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} = \frac{5\,(1.098)^2 + 5\,(1.750)^2}{10} = 2.134$$
$$z = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(6.517 - 7.033) - 0}{1.461\sqrt{\frac{1}{6} + \frac{1}{6}}} = -0.612$$

4. Decisão Não rejeitar a Hipótese Nula, ou seja, não existe uma diferença estatisticamente significativa entre as médias da resistência à flexão entre os dois tipos de implantes, com e sem tampão.

Como se pode ver pela figura, o valor da estatística não se encontra na região de rejeição, pelo que a conclusão é da não existência de diferenças estatisticamente significativas.



Nas situações em que a igualdade das variâncias não pode ser assumida, é possível ainda testar a diferença entre médias, ajustando, contudo, o número de graus de liberdade.

Definição 4.7.3: Teste de Hipóteses acerca da Diferença entre Médias, $\mu_1 - \mu_2$ (Amostras independentes e $\sigma_1^2 \neq \sigma_2^2$)

Teste Unilateral	Teste Bilateral
$H_0 : \mu_1 - \mu_2 = D_0$	$H_0 : \mu_1 - \mu_2 = D_0$
$H_1 : \mu_1 - \mu_2 > D_0$	$H_1 : \mu_1 - \mu_2 \neq D_0$
$(H_1 : \mu_1 - \mu_2 < D_0)$	

Estatística

$n_1 = n_2 = n$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{1}{n}\left(s_1^2 + s_2^2\right)}}$$

$n_1 \neq n_2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Região de Rejeição	Região de Rejeição
$t > t_\alpha$	$\lvert t \rvert > t_{\alpha/2}$
$(t < -t_\alpha)$	

e onde t_α e $t_{\alpha/2}$ são, respetivamente, os valores da distribuição t-Student com graus de

liberdade, tal que $P(t > t_\alpha) = \alpha$ e $P\left(t > t_{\alpha/2}\right) = \alpha/2$. O número de graus de liberdade é calculado, para o caso em que $n_1 = n_2 = n$, por

$$\nu = n_1 + n_2 - 2 = 2(n-1),$$

e no caso em que $(n_1 \neq n_2)$, por

$$\nu = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}.$$

Neste último caso, o valor de ν não será, em geral, inteiro. Contudo, para efeitos da consulta da tabela de t-Student, considera-se o arredondamento para o inteiro mais próximo.

Existem situações em que as amostras não são recolhidas de uma forma independente, mas de uma forma emparelhada. A técnica da razão das verosimilhanças produz a seguinte estatística

$$t = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}}$$

que segue uma distribuição de t-Student, com $n-1$ graus de liberdade, em que s_d é o desvio padrão das diferenças emparelhadas.

Definição 4.7.4: Teste de Hipóteses acerca da Diferença entre Médias, $\mu_1 - \mu_2$ (Amostras emparelhadas)

Teste Unilateral	Teste Bilateral
$H_0 : \mu_1 - \mu_2 = D_0$	$H_0 : \mu_1 - \mu_2 = D_0$
$H_1 : \mu_1 - \mu_2 > D_0$	$H_1 : \mu_1 - \mu_2 \neq D_0$
$(H_1 : \mu_1 - \mu_2 < D_0)$	

Estatística

$$t = \frac{\bar{d} - D_0}{\frac{s_d}{\sqrt{n}}}$$

Região de Rejeição	Região de Rejeição
$t > t_\alpha$	$\lvert t \rvert > t_{\alpha/2}$
$(t < -t_\alpha)$	

e onde t_α e $t_{\alpha/2}$ são, respetivamente, os valores da distribuição t-Student com graus de liberdade, tal que $P(t > t_\alpha) = \alpha$ e $P\left(t > t_{\alpha/2}\right) = \alpha/2$. D_0 é o valor especificado para a diferença $\mu_1 - \mu_2$; em geral, na maioria das situações, pretende-se testar a não existência de diferenças entre as médias das duas populações e, nesse caso, $D_0 = 0$. $\bar{d}$ é a média das

diferenças emparelhadas e s_d o respetivo desvio padrão.

Exemplo 4.7.3: Teste de Hipóteses acerca da Diferença entre Médias, amostras emparelhadas

As normas sanitárias para a água de consumo doméstico impõem uma determinada concentração de cloro. Esta concentração é ajustada na central de bombagem; contudo, é necessário verificar se a concentração se mantém ao longo da rede de distribuição. Em geral, são recolhidas amostras na central de bombagem e na rede em intervalos regulares. Estas amostras não são independentes, na medida em que uma alta concentração na central deverá originar também maiores concentrações na rede. A tabela apresenta as concentrações de cloro na central e num ponto da rede, ao longo de 10 semanas. Verifique se existem diferenças significativas nas concentrações de cloro nos dois pontos de amostragem.

Central	2.3	1.9	2.0	1.8	1.8	2.2	2.2	2.1	2.1	1.9
Rede	1.9	2.0	2.0	1.9	1.7	1.7	2.0	2.2	2.0	2.0

Solução

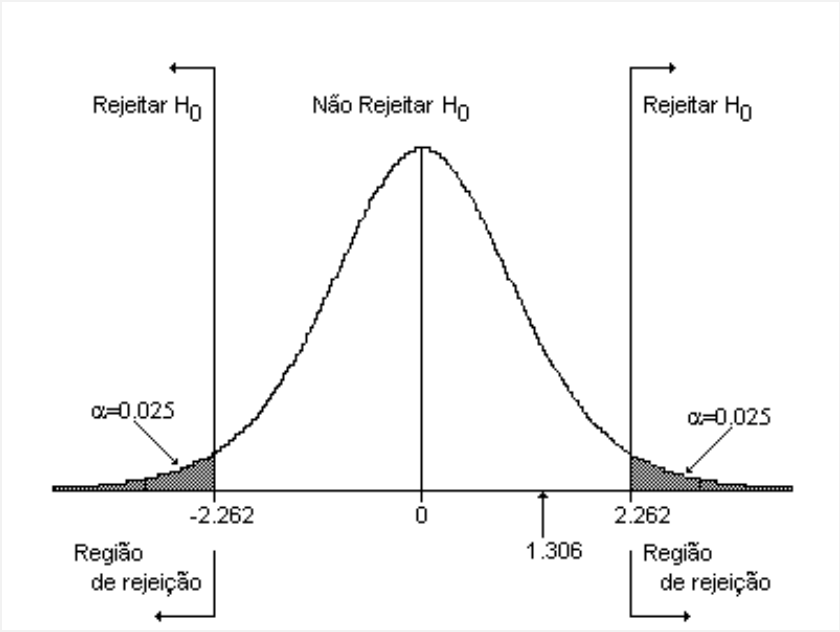
1. Formulação das hipóteses
$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

2. Região crítica
$$|t| \geq t_{0.025} = 2.262$$

3. Teste estatístico
$$\bar{d} = 0.09 \qquad s_d = 0.218$$
$$t = \frac{\bar{d} - D_0}{s_d \sqrt{\frac{1}{n}}} = \frac{0.09 - 0}{0.218\sqrt{\frac{1}{10}}} = 1.306$$

4. Decisão Não rejeitar a Hipótese Nula, ou seja, não existem diferenças entre as concentrações médias de cloro na central e na rede.

A estatística de teste não está situada na região de rejeição, o que implica a não rejeição da Hipótese Nula, tal como se pode ver pela figura.



4.8 Teste de Hipóteses acerca de Proporções

Por vezes, pretende-se testar uma proporção, que pode representar o número de pessoas infetadas com uma determinada doença, ou a proporção de eleitores que apoiam um determinado candidato. O procedimento de teste para amostras de grande dimensão baseia-se na aproximação normal à distribuição binomial, em que

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}}$$

segue uma distribuição normal padrão.

Definição 4.8.1: Teste de Hipóteses acerca de uma Proporção, p

Teste Unilateral	Teste Bilateral
$H_0 : p = p_0$	$H_0 : p = p_0$
$H_1 : p > p_0$	$H_1 : p \neq p_0$
$(H_1 : p < p_0)$	

Estatística

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}}$$

Região de Rejeição	Região de Rejeição
$z > z_\alpha$	$\lvert z\rvert > z_{\alpha/2}$
$(z < -z_\alpha)$	

onde z_α e $z_{\alpha/2}$ são, respetivamente, os valores da distribuição normal padrão tal que $P\left(z > z_\alpha\right) = \alpha$ e $P\left(z > z_{\alpha/2}\right) = \alpha/2$.

A aplicação deste teste requer que o intervalo $\hat{p} \pm 2\sqrt{\hat{p}\,(1-\hat{p})/n}$ não contenha o valor 0 ou 1.

Exemplo 4.8.1: Teste de Hipóteses acerca de uma Proporção

Uma empresa de detergentes sabe que aproximadamente 2 em cada 10 clientes potenciais usam o seu produto. Após uma campanha publicitária, 200 consumidores selecionados aleatoriamente foram entrevistados, tendo 53 expresso a sua preferência pela marca. Permitem os dados concluir que houve um aumento da aceitação do produto?

Solução

1. Formulação das hipóteses

$$H_0 : p = 0.20$$
$$H_1 : p > 0.20$$

2. Região crítica

$$z \geq z_{0.05} = 1.65$$

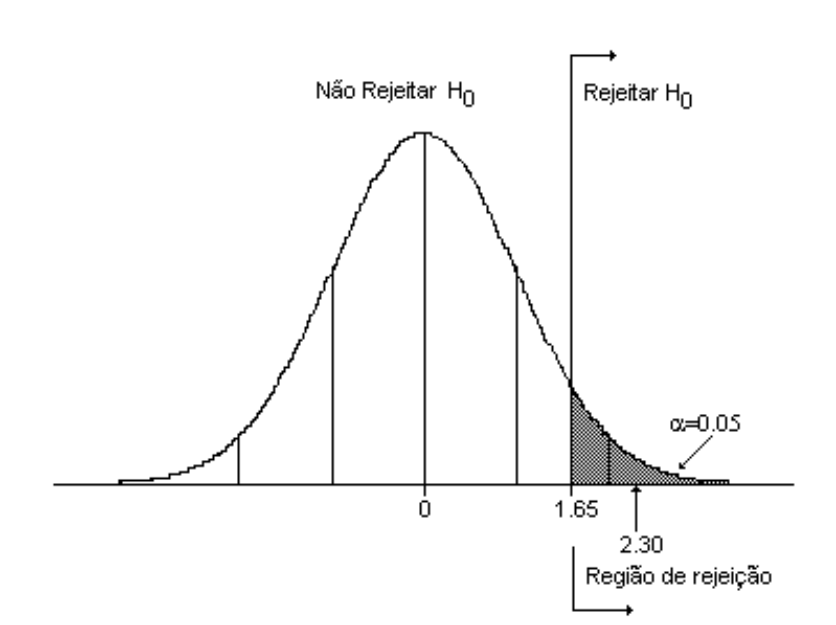
3. Teste estatístico

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.265 - 0.2}{\sqrt{\frac{0.2(1-0.2)}{200}}} \approx 2.30$$

4. Decisão

   Rejeitar a Hipótese Nula, ou seja, a campanha publicitária teve um efeito no aumento dos potenciais consumidores.

Atendendo a que o valor da estatística é superior ao valor que define a região crítica, tal como se pode ver na figura, a Hipótese Nula é rejeitada.



Quando se pretende comparar duas proporções, como sejam as proporções dos alunos do ensino superior público e privado, que estão de acordo com o pagamento de propinas, ou as proporções de peças defeituosas produzidas por duas máquinas, mais uma vez se faz uso da aproximação normal à distribuição binomial. A estatística para testar a diferença entre duas proporções é dada por,

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sigma_{(\hat{p}_1 - \hat{p}_2)}}$$

e segue uma distribuição normal padrão, onde D_0 representa a diferença específica que se pretende testar.

Definição 4.8.2: Teste de Hipóteses acerca da Diferença entre duas Proporções, $p_1 - p_2$ (Amostras independentes)

Teste Unilateral	Teste Bilateral
$H_0 : p_1 - p_2 = D_0$	$H_0 : p_1 - p_2 = D_0$
$H_1 : p_1 - p_2 > D_0$	$H_1 : p_1 - p_2 \neq D_0$
$(H_1 : p_1 - p_2 < D_0)$	

Estatística

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sigma_{(\hat{p}_1 - \hat{p}_2)}}$$

Região de Rejeição	Região de Rejeição
$z > z_\alpha$	$\lvert z \rvert > z_{\alpha/2}$
$(z < -z_\alpha)$	

onde z_α e $z_{\alpha/2}$ são, respetivamente, os valores da distribuição normal padrão tal que $P(z > z_\alpha) = \alpha$ e $P\left(z > z_{\alpha/2}\right) = \alpha/2$. D_0 é o valor especificado para a diferença $p_1 - p_2$. Se $D_0 \neq 0$,

$$\sigma_{(\hat{p}_1 - \hat{p}_2)} \approx \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Quando $D_0 = 0$,

$$\sigma_{(\hat{p}_1 - \hat{p}_2)} \approx \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

onde o número total de sucessos é $x_1 + x_2$ e

$$\hat{p}_1 = \hat{p}_2 = \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

A aplicação deste teste de hipóteses requer que os intervalos

$$\hat{p}_1 \pm 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}$$
$$\hat{p}_2 \pm 2\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

não contenham 0 ou 1.

Exemplo 4.8.2: Teste de Hipóteses acerca da Diferença entre duas Proporções

As estatísticas sobre a população portuguesa, relativas ao ano 1998, mostram que, na região do Vale do Tejo e na região Norte, o número de nascimentos e o número de óbitos com menos de um ano de idade foram os seguintes:

Região	Nascimentos	Óbitos
Lisboa e Vale do Tejo	37695	221
Norte	43469	279

Verifique se existem diferenças nas taxas de mortalidade infantil (número de óbitos/número de nascimentos) das duas regiões.

Solução

$$\hat{p}_1 = \frac{221}{37695} = 0.0059 \qquad \hat{p}_2 = \frac{279}{43469} = 0.0064 \qquad \hat{p} = \frac{221 + 279}{37695 + 43469} = 0.0062$$

1. Formulação das hipóteses

$$H_0 : p_1 - p_2 = 0$$
$$H_1 : p_1 - p_2 \neq 0$$

2. Região crítica

$$|z| \geq z_{0.025} = 1.96$$

3. Teste estatístico

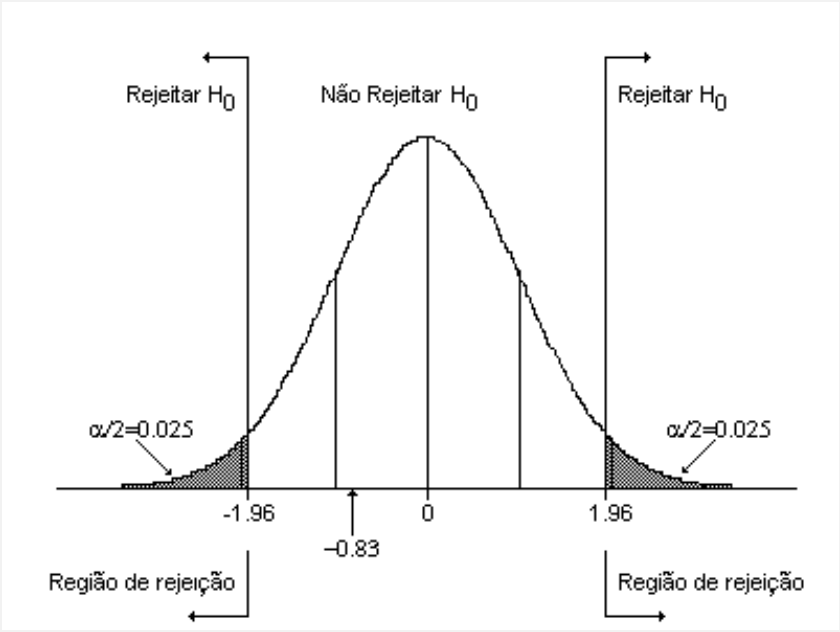
$$\sigma_{(\hat{p}_1 - \hat{p}_2)} \approx \sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\sigma_{(\hat{p}_1 - \hat{p}_2)} \approx \sqrt{(0.0062)\,(1-0.0062)\left(\frac{1}{37695} + \frac{1}{43469}\right)} = 0.0006$$

$$z = \frac{(p_1 - p_2) - D_0}{\sigma_{(\hat{p}_1 - \hat{p}_2)}} = \frac{(0.0059 - 0.0064) - 0}{0.0006} = -0.8333$$

4. Decisão Não rejeitar a Hipótese Nula, ou seja, as taxas de mortalidade infantil são equivalentes.

A figura mostra a região de rejeição. Como se pode ver, o valor da estatística não se encontra na região de rejeição. Convém referir que as taxas de mortalidade são, em geral, expressas em permilagem, o que no caso apresentado seria traduzido em taxas de mortalidade de 5.9 e 6.4. A média nacional da taxa de mortalidade é de 6.0, verificando-se uma percentagem de óbitos masculinos de 57%; a taxa de mortalidade mais elevada situa-se na Madeira (10.4) e a mais baixa na região Centro (4.4) (Instituto Nacional de Estatística).



4.9 Teste de Hipóteses acerca das Variâncias

Em Controlo de Qualidade o interesse reside não só na dimensão de fabrico, como sejam o diâmetro de uma peça, mas também na variabilidade. Em particular, a produção tem que obedecer a especificações que determinam os limites dessa variabilidade. Assim, dada uma amostra de dimensão n pretende-se testar a Hipótese Nula se $\sigma^2 = \sigma_0^2$, contra as alternativas unilaterais ou bilateral. Para tal, a estatística

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma_0^2}$$

que segue a distribuição de Qui-Quadrado, com $n-1$ graus de liberdade, será usada como estatística de teste.

Definição 4.9.1: Teste de Hipóteses acerca da Variância Populacional, σ^2

Teste Unilateral	Teste Bilateral
$H_0 : \sigma^2 = \sigma_0^2$	$H_0 : \sigma^2 = \sigma_0^2$
$H_1 : \sigma^2 > \sigma_0^2$	$H_1 : \sigma^2 \neq \sigma_0^2$
$\left(H_1 : \sigma^2 < \sigma_0^2\right)$	

Estatística

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma_0^2}$$

Região de Rejeição	Região de Rejeição
$\chi^2 > \chi^2_{\alpha}$	$\chi^2 < \chi^2_{1-\alpha/2}$ ou $\chi^2 > \chi^2_{\alpha/2}$
$\left(\chi^2 < \chi^2_{1-\alpha}\right)$	

onde $\chi^2 < \chi^2_{1-\alpha}$ e $\chi^2 < \chi^2_{1-\alpha}$ são, respetivamente, os valores de χ^2 que localizam uma área α à direita e à esquerda da distribuição de Qui-Quadrado com $n-1$ graus de liberdade.

A aplicação deste teste requer que a população de onde a amostra aleatória foi retirada, seja aproximadamente normal.

Exemplo 4.9.1: Teste de Hipóteses acerca da Variância

As garrafas de refrigerantes contêm um volume aproximado de 33 cl. O produtor perderá dinheiro se as garrafas contiverem muito mais do que o volume especificado, e correrá o risco de ser multado, se o volume for bastante inferior. Assim, é necessário controlar a variação do volume de enchimento das garrafas. Se a variância for superior a 0.25 o processo está fora de controlo, e a máquina de enchimento deve ser ajustada Para tal, um controlador de qualidade recolhe uma amostra de 15 garrafas, com um enchimento médio de 33.15 cl e um desvio padrão de 0.71 cl. Face a estes resultados pode-se concluir que o processo está controlado?

Solução

1. Formulação das hipóteses

$$H_0 : \sigma^2 = 0.25$$
$$H_1 : \sigma^2 > 0.25$$

2. Região crítica

$$\chi^2 \geq \chi^2_{0.05} = 23.685$$

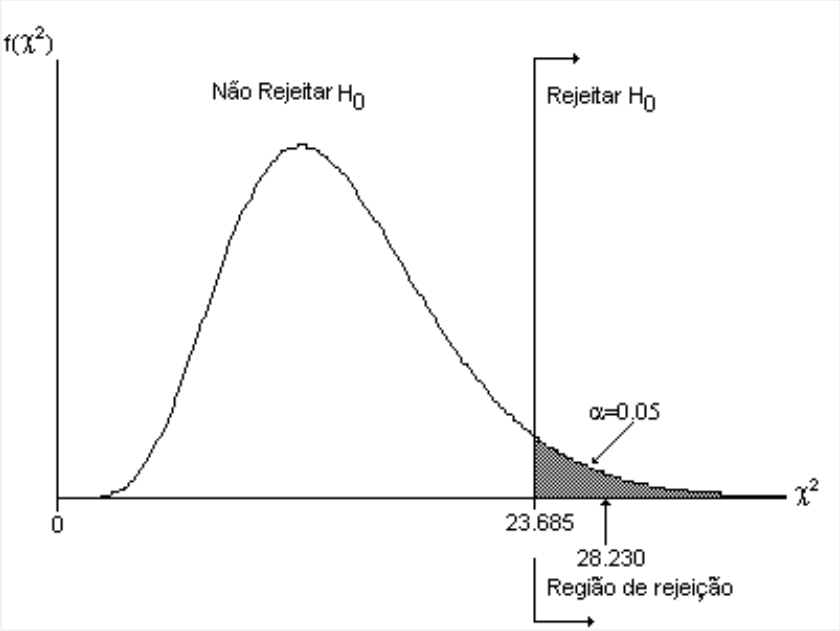
3. Teste estatístico

$$\chi^2 = \frac{(n-1)\, s^2}{\sigma_0^2} = \frac{14\,(0.71)^2}{0.25} = 28.230$$

4. Decisão

   Rejeitar a Hipótese Nula, ou seja, o processo de enchimento está fora de controlo.

A figura mostra a região de rejeição para o exemplo apresentado. Convém referir que se o nível de significância do teste fosse 0.01, a que corresponde o valor Qui-Quadrado de 29.141, a Hipótese Nula não seria rejeitada. Assim, não é correto procurar o nível de significância que conduz à rejeição da Hipótese Nula e, por isso, o nível de significância deve ser especificado a priori.



4.10 Teste de Hipóteses acerca das Variâncias

Dadas duas amostras independentes de dimensão n_1 e n_2 , respetivamente, o teste à igualdade das variâncias faz uso da seguinte estatística

$$F = \frac{s_1^2}{s_2^2}$$

obtida a partir da técnica da razão das verosimilhanças, e que segue uma distribuição F.

Definição 4.10.1: Teste de Hipóteses para a Razão de duas Variâncias Populacionais, $\frac{\sigma_1^2}{\sigma_2^2}$

Teste Unilateral	Teste Bilateral
$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \Rightarrow \left(\sigma_1^2 = \sigma_2^2\right)$	$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \Rightarrow \left(\sigma_1^2 = \sigma_2^2\right)$
$H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1 \Rightarrow \left(\sigma_1^2 > \sigma_2^2\right)$	$H_1 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \Rightarrow \left(\sigma_1^2 = \sigma_2^2\right)$
$\left(H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1 \Rightarrow \left(\sigma_1^2 < \sigma_2^2\right)\right)$	

Estatística

$$F = \frac{s_1^2}{s_2^2} \quad \left(F = \frac{s_2^2}{s_1^2}\right) \qquad F = \begin{cases} \frac{s_1^2}{s_2^2} & s_1^2 > s_2^2 \\ \frac{s_2^2}{s_1^2} & s_1^2 < s_2^2 \end{cases}$$

Região de Rejeição	Região de Rejeição
$F > F_\alpha$	$F > F_{\alpha/2}$

onde F_α e $F_{\alpha/2}$ são, respetivamente, os valores da distribuição F que localizam uma área α e $\alpha/2$ na cauda superior, com ν_1 graus de liberdade no numerador e ν_2 graus de liberdade no denominador.

As condições de aplicação do teste requerem que as populações de onde as amostras são recolhidas, de forma independente, sejam aproximadamente normais.

Exemplo 4.10.1: Teste de Hipóteses para a Razão de duas Variâncias

Uma das características importantes num vinho de qualidade é a constância do gosto no sabor e no aroma. A variabilidade no sabor depende do processo de vinificação, que pode incluir o controlo de variáveis como a temperatura, fermentação, qualidade das leveduras, etc. Um empresa quer ensaiar um novo processo de vinificação com o objetivo de reduzir a variabilidade no gosto, medido por um índice. Duas amostras aleatórias, respetivamente de 25 e 15 copos, retiradas da produção dos dois processos de vinificação foram avaliadas por um painel de provadores, numa escala de 0 a 10, produzindo os seguintes resultados:

$$\bar{x}_1 = 7.2 \quad s_1 = 1.22$$
$$\bar{x}_2 = 7.0 \quad s_2 = 0.72$$

Permitem os dados concluir que a variabilidade no segundo processo de vinificação é menor?

Solução

1. Formulação das hipóteses

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$
$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$$

2. Região crítica

$$F \geq F_{0.01,25,15} = 2.35$$

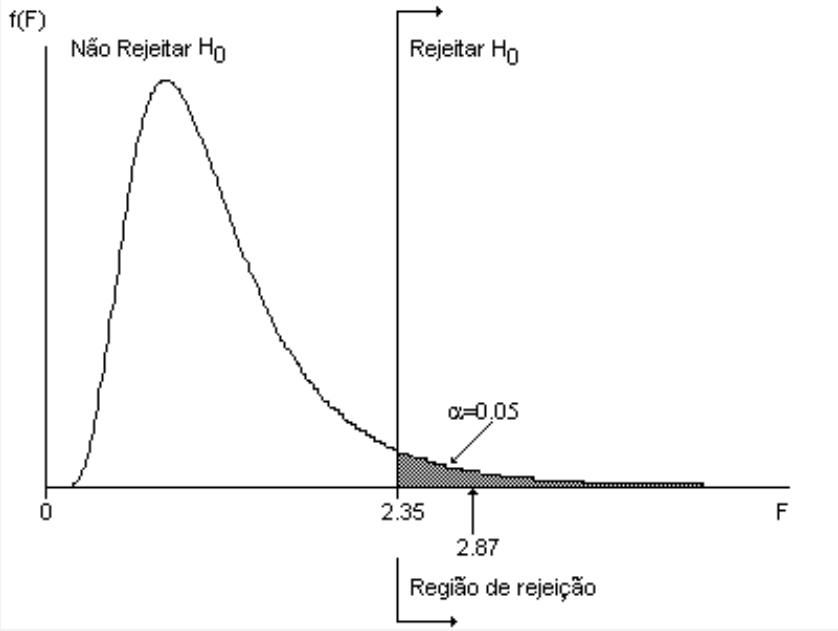
3. Teste estatístico

$$F = \frac{s_1^2}{s_2^2} = \frac{(1.22)^2}{(0.72)^2} = 2.87$$

4. Decisão

   Rejeitar a Hipótese Nula, ou seja, a variabilidade no segundo processo de vinificação é menor.

A figura mostra a região de rejeição. Assim, como o valor da estatística se encontra na região de rejeição, pode-se concluir que os dois processos de vinificação não apresentam a mesma variabilidade, com o segundo processo exibindo uma menor variabilidade.



Exercícios

1. Suponha que pretende determinar a produção média diária de leite de uma vacaria. Para esse efeito, recolheu amostras da produção de 50 dias, tendo encontrado um desvio padrão de 25 litros e uma média 672 litros. Assumindo que a produção segue uma distribuição aproximadamente normal,

   a) Teste a hipótese de que produção média diária é de 680 litros ($\alpha = 0.05$).

b) Calcule o valor de prova.

c) Calcule o poder do teste quando a média da produção é $\mu = 670$, isto é, debaixo da Hipótese alternativa.

d) Calcule a curva de potência para valores da μ, debaixo da hipótese alternativa entre 660 e 700 litros.

2. Suponha que o produtor de leite resolve usar uma nova ração alimentar se esta lhe garantir um aumento na produção diária de 30 litros. Quão grande deve ser a amostra para detetar esse aumento de 20 litros com uma probabilidade de 90% $(1 - \beta)$, assumindo uma desvio padrão de 25 litros? Use $\alpha = 0.05$.

3. Tendo por base um determinado conjunto de dados, a hipótese nula foi rejeitada ao nível de significância de 0.05. Seria também rejeitada ao nível de significância de

a) 0.01;

b) 0.10.

4. Num determinado teste de hipótese, o valor de prova p correspondente à estatística é de 0.0316. Pode a hipótese nula ser rejeitada ao nível de significância α de

a) 0.01;

b) 0.05;

c) 0.10.

5. Uma fábrica de cimento tem duas linhas de produção, tecnicamente diferentes. Com o objetivo de comparar a produção, foram recolhidas amostras das duas linhas. Assim, na primeira linha, um amostra de 44 dias, produziu uma média de 1045 kg com um desvio padrão de 225 e, da segunda linha, uma média de 1255, com um desvio padrão 200, com base numa amostra de 54 dias. Pode-se concluir que as duas linhas são diferentes na produção média diária?

6. Numa universidade, uma amostra aleatórias de 12 estudantes do sexo feminino foi selecionada. O comprimento médio da mão encontrado foi de 165.4 mm com um desvio padrão de 4.5 mm.

163	168	167	166	159	159	166	176	167	166	166	162

Teste a hipótese, para um nível de significância de 5%, de que o comprimento médio da mão das meninas é 169 mm. Determine também o intervalo de confiança de 95%. O que pode concluir?

7. A tabela apresenta os valores dos comprimentos das mãos de 12 estudantes de cada sexo.

Masculino	185	189	188	188	179	180	188	199	189	188	188	183
Feminino	163	168	167	166	159	159	166	176	167	166	166	162

Teste a hipótese de que o comprimento médio da mão das estudantes é menor que dos rapazes. Use $\alpha = 0.05$.

8. Os dados registam o número médio de horas-homem perdidas devidas a acidentes em 10 fábricas antes e depois dum programa de higiene e segurança ter sido implementado:

Fábrica	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
Antes	45	73	46	124	33	57	83	34	26	17
Depois	36	60	44	119	35	51	77	29	24	11

O que pode concluir acerca da eficácia do programa? Use $\alpha = 0.05$.

9. A Inspeção de Atividades Económicas selecionou uma bomba de gasolina para averiguar se o volume debitado numa bomba correspondia ao valor nominal cobrado ao cliente. Para esse efeito, recolheu num recipiente, 10 amostras de 5 litros que produziram os seguintes resultados. Verifique se a variabilidade da bomba está de acordo com o valor nominal de 0.1 litros. Use $\alpha = 0.05$.

5.0	4.9	5.1	4.9	5.0	4.8	4.9	5.1	5.1	5.1

O que pode concluir?

10. Um fábrica pretende adquiri lâmpadas LED para incorporar num equipamento que produz. Para esse efeito contacta dois produtores de lâmpadas que lhe forneceram cada um 10 lâmpadas para teste. Estas lâmpadas foram ligadas até falharem tendo sido obtidos os seguintes resultados.

Fornecedor A	25091	24996	24967	24957	24777	24788	24962	25196	24978	24962
Fornecedor B	24946	24814	25328	24860	24928	24714	24756	25162	25318	25393

Os dados indicam uma diferença na variabilidade do tempo de vidas das lâmpadas dos dois produtores? Use $\alpha = 0.10$.

11. A proporção de hipertensos na população portuguesa é aproximadamente de 40%. Num amostra de 2542 utentes dos serviços de saúde do concelho de Guimarães, encontraram-se 1076 hipertensos. Teste a hipótese de que a proporção de hipertensos em Guimarães é superior à proporção do país.

12. A tabela apresenta a proporção de hipertensos entre homens e mulheres dos centros de saúde de Guimarães e Vizela, com mais de 50 anos de idade.

	Não HT	HT	
Feminino	146	447	593
Masculino	136	389	525
	283	836	1118

Teste a hipótese de que não há diferenças nas proporções de hipertensos entre homens e mulheres.

Soluções

1. a) $[673.0704, 686.9296]$

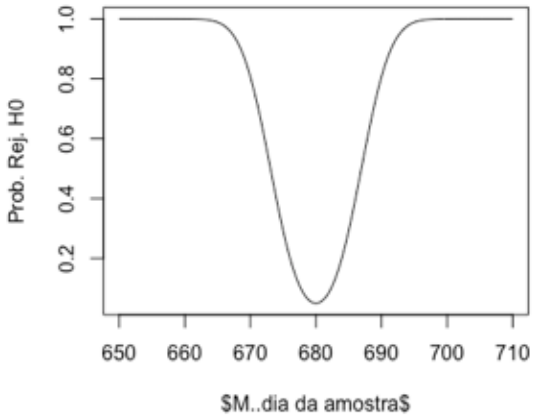
b) $x = 672$; Estatística de teste $Z = -2.262742$, Rejeitar H0 $H_0 : \mu = 680 \quad H_1 : \mu \neq 680$

c) valor $p = 0.02365162$

d) $\beta = 0.1925794$ $Pot = 1 - \beta = 0.8074206$

e)

μ	β	Pot	μ	β	Pot
661	0.0003	0.9997	681	0.941	0.059
662	0.0009	0.9991	682	0.912	0.088
663	0.0022	0.9978	683	0.864	0.136
664	0.0051	0.9949	684	0.7957	0.2043
665	0.0113	0.9887	685	0.7084	0.2916
666	0.0228	0.9772	686	0.6025	0.3975
667	0.0427	0.9573	687	0.492	0.508
668	0.0764	0.9236	688	0.3821	0.6179
669	0.1251	0.8749	689	0.2776	0.7224
670	0.1922	0.8078	690	0.1922	0.8078
671	0.2776	0.7224	691	0.1251	0.8749
672	0.3821	0.6179	692	0.0764	0.9236
673	0.492	0.508	693	0.0427	0.9573
674	0.6025	0.3975	694	0.0228	0.9772
675	0.7084	0.2916	695	0.0113	0.9887
676	0.7957	0.2043	696	0.0051	0.9949
677	0.864	0.136	697	0.0022	0.9978
678	0.912	0.088	698	0.0009	0.9991
679	0.941	0.059	699	0.0003	0.9997
680	0.95	0.05	700	0.0001	0.9999



2. $n = 54$

```
# teste unilateral
delta=10
#H0:mu0=680
#H1:mu0=680+delta
alpha=0.05
beta=0.10
sigma=25
#P(Z>zalpha)=0.05
z_a=qnorm(0.95)
z_a
#P(Z<zbeta)=0.1
z_b=qnorm(0.1)
z_b
#(c-mu0)/(sigma/sqrt(n))=za; erro de tipo I
#(c-mu0)/(sigma/sqrt(n))=zb; erro tipo II
#igualando e resolvendo para n
n=sigma^2*(z_a-z_b)^2/delta^2
n
################################
# limite de corte, c
c1=z_a*sigma/sqrt(n)+H0
c1
c1=z_b*sigma/sqrt(n)+H0+delta
c1
```

```
##############################
#gráfico
H0=680
x1=seq(665,705,by=0.5)
z_1=(x1-(H0))/(sigma/sqrt(n))
hx_1=dnorm(z_1)
z_2=(x1-(H0+delta))/(sigma/sqrt(n))
hx_2=dnorm(z_2)

plot(x1,hx_1,"l",xlim=c(660,705),xlab="", ylab="")
par(new=TRUE)
plot(x1,hx_2,"l",xlim=c(660,705),xlab="Prod. (l)", ylab="Densidade")
abline(v=c1)
text(680, y = 0.05, labels = "beta")
text(690, y = 0.05, labels = "alpha")
```

3. a) Depende do valor de prova. O valor de prova é menor que 0.05, daí decorrendo a rejeição de H0 ao nível de 5%. Se o valor de prova for maior que 1% não se rejeitaria mas se fosse menor, tal conduziria à rejeição de H_0

4. a) Não, pois valor $p > 0.01$; Sim, pois valor $p < 0.05$; Sim, pois valor $p < 0.10$

5. $H_0 : \mu_1 - \mu_2 = 0 \quad H_1 : \mu_1 - \mu_2 \neq 0$

   $|Z| > 1.96$ Rej. H0, valor $p < 0.05$; $Z = -4.828787$

```
# teste a duas amostras independentes de grande dimensão
n1=44
m1=1045
s1=225
n2=54
m2=1255
s2=200
s_error=sqrt(s1^2/n1+s2^2/n2)
s_error
Z=(m1-m2)/s_error # estatística de teste
Z
qnorm(0.975) # valor crítico bilateral
```

6. Rejeitar H_0, valor $p < 0.05$;$t = -2.7451$, $df = 11$

   $[162.5436, 168.2897]$

```
# teste a uma amostra de pequena dimensão n<30
comp_f=c(163, 168, 167, 166, 159, 159, 166, 176, 167, 166, 166, 162)
summary(comp_f) # medidas resumo
mean(comp_f) # média
sd(comp_f) # desvio padrão
T=((mean(comp_f)-169)/(sd(comp_f)/sqrt(12))) # estatística T
T
qt(0.975,length(comp_f)-1) # valor crítico
pt(T,length(comp_f)-1)*2 # valor de prova bilateral
t.test(comp_f,
       alternative = c("two.sided"),
       mu = 169,
       conf.level = 0.95) # teste t e intervalo de confiança
```

7. Teste unilateral. Rejeitar H_0, valor $p < 0.05$; $t = 10.928$, $df = 22$

```
#teste a duas amostras independentes de pequena dimensão
comp_m=c(185, 189, 188, 188, 179, 180, 188, 199, 189, 188, 188, 183)
comp_f=c(163, 168, 167, 166, 159, 159, 166, 176, 167, 166, 166, 162)
summary(comp_m) # medidas resumo
mean(comp_m) # média
sd(comp_m) # desvio padrão
summary(comp_f) # medidas resumo
mean(comp_f) # média
sd(comp_f) # desvio padrão
dif=mean(comp_f)-mean(comp_m)
dif
df1=length(comp_m)-1 # graus de liberdade
df1
df2=length(comp_f)-1 # graus de liberdade
df2
sp2=(df1*var(comp_m)+df2*var(comp_f))/(df1+df2) #variância pesada
sp=sqrt(sp2) #desvio padrão
```

```
sp
s_error=sp*sqrt(1/length(comp_m)+1/length(comp_f))
s_error
T=(dif-0)/s_error
T
qt(0.95,df1+df2)*(-1) # valor crítico teste unilateral à esquerda
pt(T,df1+df2) # valor de prova unilateral
t.test(comp_f, comp_m,
       alternative = c("less"),
       mu = 0,
       var.equal = T, # assumindo igualdade de variãncias
       conf.level = 0.95) # teste t e intervalo de confiança
```

8. Amostras emparelhadas. Rejeitar H_0, valor $p < 0.05$; $t = 4.0333$, $df = 9$

```
#teste a duas amostras emparelhadas
c_ant=c(45, 73, 46, 124, 33, 57, 83, 34, 26, 17)
c_dep=c(36, 60, 44, 119, 35, 51, 77, 29, 24, 11)
summary(c_ant) # medidas resumo
mean(c_ant) # média
sd(c_ant) # desvio padrão
summary(c_dep) # medidas resumo
mean(c_dep) # média
sd(c_dep) # desvio padrão
dif=c_ant-c_dep # diferenças emparelhadas
dif
mean(dif) # média
sd(dif) # desvio padrão
length(dif)-1 # nº de graus de liberdade
qt(0.95,df=length(dif)-1)
#quantil da distribuição t para uma confiança de 95%
s_error <- sd(dif)/sqrt(length(dif)) # erro da estimativa
error
T=(mean(dif)-0)/s_error
T
1-pt(T,length(dif)-1) # valor de prova unilateral
```

```
t.test(c_ant, c_dep, # teste t e intervalo de confiança
       alternative = c("greater"),
       mu = 0,
       conf.level = 0.95,
       paired = T)
```

9. Não rejeitar H_0; $Q = 10.9$; $1 - 0.7173742 > 0.025$

```
# teste a uma variância
c_gas=c(5.0, 4.9, 5.1, 4.9, 5.0, 4.8, 4.9, 5.1, 5.1, 5.1)
summary(c_gas)
mean(c_gas)
mean
sd(c_gas)
sd
var(c_gas)
sigma_0=0.1
Q=((length(c_gas)-1)*var(c_gas))/sigma_0^2; Q # valor da estatística
Q
pchisq(Q,length(Vol)-1)
qchisq(0.025,length(Vol)-1)
qchisq(0.975,length(Vol)-1)
```

10. Rejeitar H_0, valor $p < 0.05$; $F = 4.2915$, $numdf = 9, denomdf = 9$

11. Rejeitar H_0; $Z = 2.396781$, valor $p < 0.05$ $\chi^2 = 5.7446$, $df = 1$

```
# teste a uma proporção
n=2542 # nº de respostas
x=1076 # nº de sucessos
p0=0.4
phat=x/n # proporção de votantes em MRS
phat
SE=sqrt(p0*(1-p0)/n) # erro padrão, aproximação à normal para grandes amostras
SE
Z=(phat-p0)/SE; Z # estatística de teste
```

```
1-pnorm(Z) #valor de prova
prop.test(x, n, p=0.4, # teste a uma proporção aproximação à normal; chisq=Z^2
          alternative = c("greater"),
          conf.level = 0.95, correct = FALSE) # sem correcção para a continuidade
```

12. Não rejeitar H_0; $Z = 0.4934339$, valor $p > 0.05$ $\chi^2 = 0.24348$, $df = 1$

```
# Teste para a diferença de proporções
n1=593 # nº de respostas
x1=447 # nº de sucessos
phat1=x1/n1 # proporção de hipertensos
phat1
n2=525 # nº de respostas
x2=389 # nº de sucessos
phat2=x2/n2 # proporção de hipertensos
phat2
phat=(x1+x2)/(n1+n2)
s_error=sqrt(phat*(1-phat)*(1/n1+1/n2))
# erro padrão, aproximação à normal para grandes amostras
s_error
Z=((phat1-phat2)-0)/s_error; Z # estatística de teste
2*(1-pnorm(Z)) # valor de prova bilateral
prop.test(c(x1,x2),c(x1+(n1-x1),x2+(n2-x2)),
          #p=0.4, # teste a uma proporção aproximação à normal; chisq=Z^2
          alternative = c("two.sided"),
          conf.level = 0.95, correct = FALSE) # sem correcção para a continuidade
```

Capítulo 5

Planeamento Experimental

5.1 Introdução

A experimentação constitui a base de muitos dos processos de decisão em engenharia, agricultura, medicina, sociologia e nas ciências aplicadas. Em geral, o objetivo é investigar os efeitos de uma ou mais variáveis numa outra variável de interesse. Assim, por exemplo, o objetivo pode ser investigar a resistência de provetes de betão, para diversos tipos de cimento. Um outro exemplo é o estudo do rendimento agrícola em função de diversas quantidades de fertilizante usadas. As variáveis independentes são referidas como fatores, sendo a sua intensidade referida como nível. Assim, no caso dos provetes o fator é uma variável qualitativa, sendo os seus níveis definidos pelo tipo de cimento usado e, no segundo exemplo, o fator é uma variável quantitativa, cujos níveis definem a quantidade de fertilizante usada por unidade de área. Existem, contudo, experiências em que mais do que um fator está presente, como sejam o caso de experiências em que se estudam o efeito da temperatura, pressão e alimentação do reator. Nestes casos, cada combinação de fatores, num determinado nível, define um tratamento ao qual a unidade experimental é sujeita. Em geral, a realização de experiências pode ter vários objetivos entre os quais se salientam o aumento do rendimento do processo, a redução de custos de operação, a redução da variabilidade do processo ou a redução do tempo de projeto e implementação.

Definição 5.1.1: Fatores

Numa experiência, as variáveis independentes que influenciam uma variável resposta são referidas como fatores.

Definição 5.1.2: Nível

A intensidade de um fator é definida como nível.

Definição 5.1.3: Tratamento

Um tratamento é uma combinação particular de níveis dos fatores envolvidos na experiência.

Uma experiência, tal como um teste estatístico, requer a definição dos seguintes pontos:

- Conjetura - a hipótese que motiva a experiência.
- Experiência - o teste para investigar a conjetura.
- Análise - a análise estatística dos dados experimentais.
- Conclusão - verificação, rejeição ou revisão da conjetura que motivou a experiência.

A realização de uma experiência exige um planeamento cuidado que envolve os seguintes passos:

1. Seleção dos fatores e identificação dos parâmetros que são objeto do estudo;
2. Decisão sobre a magnitude dos erros padrão pretendidos;
3. Escolha dos tratamentos (combinações de níveis de fatores) a serem incluídos na experiência, bem como o número de observações em cada tratamento;
4. Atribuição dos tratamentos às unidades experimentais.

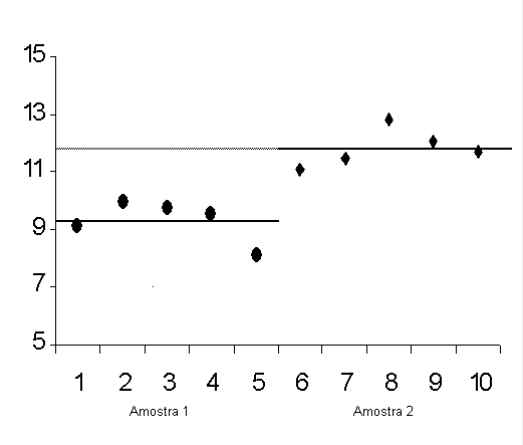
Num certo sentido, poder-se-á dizer que todas as experiências são planeadas; contudo, algumas são mal planeadas pois não permitem o uso de métodos estatísticos na sua análise e, por isso, conduzem a um desperdício de recursos e a um aumento de custos.

Relativamente ao teste de hipóteses acerca de duas médias, a Análise da Variância permite efetuar inferências acerca de três ou mais médias associadas aos vários tratamentos. No entanto, a lógica subjacente pode ser compreendida através do seguinte exemplo para dois tratamentos.

Exemplo 5.1.1: Análise da variância

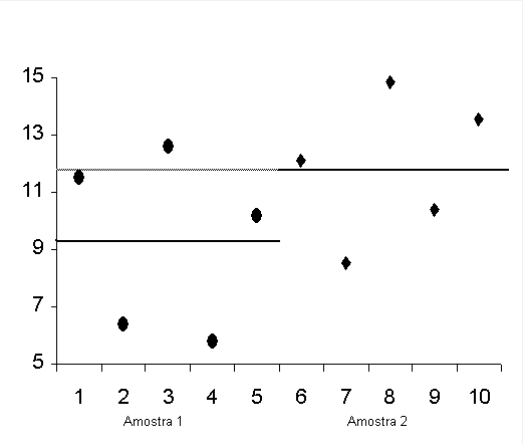
Os dados da tabela apresentam as observações de duas amostras aleatórias independentes, de dois processos produtivos, em kg/h. Pretende-se verificar se existem diferenças entre os dois tratamentos. A respetiva figura apresenta as observações pela ordem em que foram registadas.

Amostra 1	Amostra 2
9.1	11.1
10.0	11.5
9.7	12.8
9.6	12.1
8.1	11.7
$\bar{x_1} = 9.3$	$\bar{x_2} = 11.8$
$s_1 = 0.75$	$s_2 = 0.65$



Suponha, contudo, que os valores registados eram os da tabela abaixo, apresentando a figura a respetiva representação das observações.

Amostra 1	Amostra 2
11.5	12.1
6.4	8.5
12.6	14.8
5.8	10.4
10.2	13.4
$\bar{x_1} = 9.3$	$\bar{x_2} = 11.8$
$s_1 = 3.05$	$s_2 = 2.74$



As linhas horizontais a cheio representam a média das duas amostras. De notar que as médias das duas amostras, quer no caso da primeira figura quer no caso da segunda figura, são iguais. No entanto, parece evidente que as médias das populações são diferentes no caso da primeira figura e não no caso da segunda figura. A razão reside no facto de a avaliação ter como base a comparação da distância entre as médias com a variação dos valores observados para cada uma das amostras. No primeiro caso, a diferença entre as médias parece ser grande relativamente variação dentro de cada amostra.

5.2 Planeamento Completamente Casual

O Planeamento Completamente Casual é o planeamento experimental mais simples pois destina-se a avaliar a influência de um único fator. Assim, supondo que se dispunham de unidades experimentais e um fator com k níveis, n/k unidades experimentais são distribuídas aleatoriamente, isto é, com igual probabilidade, a cada nível do fator. Esta atribuição é feita, em geral, pelo recurso

a números aleatórios. A aleatorização é extremamente importante pois constitui a única proteção contra alguma interferência que possa influenciar o resultado da experiência. Por exemplo, suponha que se pretende estudar o efeito da velocidade de rotação do parafuso num processo de extrusão de polímeros. Como o processo de extrusão depende da temperatura de aquecimento, se todas as experiências forem conduzidas por ordem crescente da velocidade de rotação, pode haver uma diferença nas respostas observadas que seja devida ao efeito do aquecimento.

Exemplo 5.2.1: Planeamento completamente casual

A resistência de uma fibra têxtil depende da concentração de algodão. Para testar a influência da percentagem de algodão na resistência da fibra foi conduzida uma experiência em que o fator - percentagem de algodão - foi testado em quatro níveis. Para cada nível do fator foram realizadas cinco observações. A tabela apresenta os resultados obtidos.

Percentagem de Algodão	1	2	3	4	5	Total	Média
5%	7,1	8.0	7.7	7.6	6.1	36.5	7.3
10%	8.0	8.3	10.4	10.7	9.4	46.8	9.36
15%	10.7	12.4	12.2	11.0	13.1	59.4	11.88
20%	11.5	13.1	13.0	10.9	13.8	62.3	12.46

No Planeamento Completamente Casual as observações podem ser descritas, genericamente, pela seguinte tabela:

	Observações					
Tratamento	1	2	...	n	Total	Média
1	y_{11}	y_{12}	...	y_{1n}	$T_{1.}$	$\bar{y_{1.}}$
2	y_{21}	y_{21}	...	y_{2n}	$T_{2.}$	$\bar{y_{2.}}$
...	...	...	...	...	...	...
k	y_{k1}	y_{k2}	...	y_{kn}	$T_{k.}$	$\bar{y_{k.}}$

O modelo linear subjacente é descrito pela seguinte equação:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, k \\ j = 1, 2, \ldots, n \end{cases}$$

onde Y_{ij} é uma variável aleatória que corresponde à observação (ij), μ é a média global comum a todas as observações, α_i é o efeito associado com o tratamento i, com a condição $\sum \alpha_i = 0$, e ε_{ij} é o erro aleatório. Contudo, a média associada a cada tratamento pode ser relacionada com a média global através de,

$$\mu_i = \mu + \alpha_i$$

Assim, cada tratamento pode ser compreendido como definindo uma média, constituída pela média global mais o efeito associado ao tratamento. Os erros serão assumidos como normais, com média nula e variância comum σ^2. Esta assunção implica que cada tratamento pode ser compreendido como uma população normal com média μ_i e variância σ^2. Relativamente aos níveis dos fatores, é necessário referir que quando estes são fixados *a priori*, as conclusões que se podem retirar da experiência dizem respeito somente a estes tratamentos e a mais nenhum que não tenha sido considerado. Neste caso, o planeamento é referido como de efeitos fixos. Contudo, se os níveis considerados são uma amostra aleatória de uma população de tratamentos, é possível tirar conclusões para todos os tratamentos na população, quer tenham ou não sido explicitamente considerados. Nesta situação, o planeamento diz-se de efeitos aleatórios, que não será abordado neste texto.

Definição 5.2.1: Planeamento Completamente Casual

Um Planeamento Completamente Casual é um planeamento onde se pretende comparar k tratamentos que foram aleatoriamente atribuídos às unidades experimentais, ou no qual amostras aleatórias independentes foram retiradas de k populações.

A Análise da Variância divide a variabilidade total dos dados em duas componentes, uma associada aos tratamentos e outra associada aos erros. Para explicitar esta partição é necessário definir as expressões para a média global, soma total das observações, média dos tratamentos e soma das observações em cada tratamento. Assim,

$$T_{..} = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{..} = T_{..}/kn$$

$$T_{i.} = \sum_{j=1}^{n} y_{ij} \qquad \bar{y}_{i.} = T_{i.}/n$$

Para uma dada observação é possível escrever a igualdade

$$(y_{ij} - \bar{y}_{..}) = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})\,,$$

e considerando somatórios e quadrados, a partição pode assim ser expressa como,

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2 = n\sum_{i=1}^{k}(y_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2.$$

Assim, a variabilidade total dos dados pode ser expressa em termos da Soma Total dos Quadrados (STQ),

$$STQ = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2,$$

Soma dos Quadrados dos Tratamentos (SQT),

$$SQT = n\sum_{i=1}^{k}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$$

e Soma dos Quadrados dos Resíduos (SQR),

$$SQR = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2.$$

Demonstração.

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left[\left(\bar{y}_{i.} - \bar{y}_{..}\right) + \left(y_{ij} - \bar{y}_{i.}\right)\right]^2$$

O segundo termo pode ser desenvolvido, originando

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n}\left(\bar{y}_{i.} - \bar{y}_{..}\right)\left(y_{ij} - \bar{y}_{i.}\right).$$

Como a primeira parcela não depende de j, é possível escrevê-la tal como a expressão de SQT. Assim, é só necessário demonstrar que a terceira parcela é nula. Desenvolvendo,

$$2\sum_{i=1}^{k}\left(\bar{y}_{i.} - \bar{y}_{..}\right)\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)$$

mostra-se que

$$\sum_{i=1}^{k}\left(\bar{y}_{i.} - \bar{y}_{..}\right) = \sum_{i=1}^{k}\bar{y}_{i.} - k\bar{y}_{..}$$

e tendo em atenção que

$$\bar{y}_{i.} = \frac{\sum_{j=1}^{n} y_{ij}}{n},$$

vem que

$$\sum_{i=1}^{k}\frac{\sum_{j=1}^{n} y_{ij}}{n} - k\frac{\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}}{kn} = 0.$$

□

A análise dos termos devidos aos tratamentos (SQT) e aos resíduos (SQR) permite explicitar de uma forma mais clara o modo como a análise de variância extrai conclusões a partir dos dados observados. Em primeiro lugar, considere-se o valor esperado de SQT,

$$E\left[SQT\right] = E\left[n\sum_{i=1}^{k}\left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2\right].$$

Atendendo a que o modelo subjacente

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

as médias podem ser expressas como,

$$\bar{Y}_{i.} = \mu + \alpha_i + \bar{\varepsilon}_{i.}$$

e

$$\bar{Y}_{..} = \mu + \bar{\varepsilon}_{..}$$

já que $\sum_{i=1}^{k} = \alpha_i = 0$. Assim, a expressão para o valor esperado,

$$\begin{aligned} E\left[SQT\right] &= E\left[n\sum_{i=1}^{k}\left(\alpha_i + \bar{\varepsilon}_i - \bar{\varepsilon}_{..}\right)^2\right] \\ &= \left[n\sum_{i=1}^{k}\alpha_i^2 + n\sum_{i=1}^{k}\bar{\varepsilon}_{i.}^2 - kn\bar{\varepsilon}_{..}^2 + 2n\sum_{i=1}^{k}\alpha_i\bar{\varepsilon}_{i.} - 2n\bar{\varepsilon}_{..}\sum_{i=1}^{k}\alpha_i - 2n\bar{\varepsilon}_{..}\sum_{i=1}^{k}\bar{\varepsilon}_{i.}\right] \end{aligned}$$

e tendo em atenção que os erros ε_{ij} são independentes com média zero e variância comum σ^2,

$$E\left[\bar{\varepsilon}_{i.}^2\right] = \frac{\sigma^2}{n}, \quad E\left[\bar{\varepsilon}_{..}^2\right] = \frac{\sigma^2}{kn}, \quad E\left[\bar{\varepsilon}_{i.}\right] = 0$$

$$E\left[SQT\right] = n\sum_{i=1}^{k}\alpha_i^2 + k\sigma^2 - \sigma^2 = \left(k-1\right)\sigma^2 + n\sum_{i=1}^{k}\alpha_i^2.$$

Debaixo da Hipótese Nula, $H_0 : \alpha_i = 0$, donde,

$$E\left[\frac{SQT}{k-1}\right] = \sigma^2,$$

e debaixo da Hipótese Alternativa,

$$E\left[\frac{SQT}{k-1}\right] = \sigma^2 + \frac{n\sum_{i=1}^{k}\alpha_i^2}{k-1}.$$

Relativamente ao valor esperado da Soma dos Quadrados dos Resíduos,

$$E\left[SQR\right] = E\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_{i.}\right)^2\right]$$

e tendo em consideração as relações,

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \bar{Y}_{i.} = \mu + \alpha_i + \bar{\varepsilon}_{i.}$$

$$E\left[SQR\right] = E\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\left(\varepsilon_{ij} - \bar{\varepsilon}_{i.}\right)^2\right] = E\left[\sum_{i=1}^{k}\sum_{j=1}^{n}\varepsilon_{ij}^2 - 2\sum_{i=1}^{k}\bar{\varepsilon}_{i.}\sum_{j=1}^{n}\varepsilon_{ij} + n\sum_{i=1}^{k}\bar{\varepsilon}_{i.}^2\right],$$

dado que $n\bar{\varepsilon}_{i.} = \sum_{j=1}^{n}\varepsilon_{ij}$ e

$$E\left[\bar{\varepsilon}_{i.}^2\right] = \frac{\sigma^2}{n}, \quad E\left[\varepsilon_{ij}^2\right] = \sigma^2$$

assim,

$$E\left[SQR\right] = kn\sigma^2 - k\sigma^2 = k\left(n-1\right)\sigma^2.$$

O número de graus de liberdade também pode ser particionado. O número total de observações é kn pelo que a STQ terá $kn-1$ graus de liberdade; por outro lado, existem k tratamentos e, assim, o número de graus de liberdade associados à SQT será $k-1$; finalmente, dentro de cada tratamento existem n replicações que definem $n-1$ graus de liberdade para a estimativa do erro experimental, o que para os k tratamentos definem $k\,(n-1)$ graus de liberdade associados à SQR.

A Média dos Quadrados Resíduos (MQR) é uma estimativa independente e não enviesada de σ^2, quer a Hipótese Nula seja ou não verdadeira. Assim, debaixo da Hipótese Nula também a Média dos Quadrados dos Tratamentos (MQT) será uma estimativa independente e não enviesada de σ^2, mas caso não seja verdadeira, será maior que σ^2. Por isso, debaixo da Hipótese Alternativa, a razão entre as duas médias será muito maior que a unidade, o que implica um teste unilateral. Por isso, a Hipótese Nula será rejeitada se o valor desta razão exceder o valor crítico de $F_{\alpha,k,k(n-1)}$.

O cálculo das Somas dos Quadrados pode ser efetuado de uma forma mais eficiente através do recurso às seguintes fórmulas, deduzidas a partir das respetivas somas:

$$STQ = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{T_{\cdot\cdot}^2}{kn} \qquad SQT = \frac{1}{n}\sum_{i=1}^{k} T_{i\cdot}^2 - \frac{T_{\cdot\cdot}^2}{kn} \qquad SQR = STQ - SQT.$$

Os cálculos efetuados podem ser resumidos na chamada Tabela ANOVA (ANalysis Of VAriance):

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	F
Tratamentos	SQT	$k-1$	MQT	$F = MQT/MQR$
Resíduos	SQR	$k(n-1)$	MQR	
Total	STQ	$kn-1$		

Por vezes, não é possível obter o mesmo número de observações por tratamento e, nesse caso, diz-se que o planeamento não é balanceado. Neste caso, é necessário fazer algumas modificações às fórmulas apresentadas. Assim, e considerando que o número total de observações é $N = \sum_{i=1}^{k} n_i$, em que n_i é o número de observações no tratamento i,

$$STQ = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{T_{\cdot\cdot}^2}{N} \qquad SQT = \sum_{i=1}^{k} \frac{T_{i\cdot}^2}{n_i} - \frac{T_{\cdot\cdot}^2}{N} \qquad SQR = STQ - SQT$$

Contudo, deve ser realçado que um planeamento balanceado maximiza a potência do teste e é relativamente insensível a pequenos desvios da assunção de igualdade das variâncias.

5.2.1 Análise de Resíduos

O Planeamento Completamente Casual assume que as observações são independentes, normalmente distribuídas e com variância comum igual para cada tratamento. A verificação destas

assunções é feita com base na análise dos resíduos. O resíduo é definido como a diferença entre o valor observado e o valor ajustado pelo modelo, isto é,

$$e_{ij} = y_{ij} - \hat{y}_{ij}.$$

No Planeamento Completamente Casual, o valor estimado é $\bar{y}_i$, a média individual de cada tratamento. Assim, os resíduos são definidos como a diferença entre cada observação e a média do tratamento correspondente. Esta diferença,

$$e_{ij} = y_{ij} - \bar{y}_{i.},$$

ao remover o efeito do tratamento, contém a informação acerca da variabilidade não explicada.

A verificação das assunções sobre a normalidade, independência e homogeneidade das variâncias pode ser feita através de técnicas gráficas. Em particular, a representação dos resíduos em função dos tratamentos através de um gráfico de caixa e bigodes permite verificar se a variabilidade em cada tratamento é aproximadamente igual. Um gráfico dos resíduos em função das médias dos tratamentos também pode permitir verificar se a variabilidade depende das médias. A independência pode ser verificada através da representação dos resíduos observados em função da ordem em que as experiências foram conduzidas. A sequência de resíduos positivos e negativos deve ser aleatória e o aparecimento de algum padrão põe em causa a independência entre resíduos. A normalidade pode ser verificada através de um gráfico de probabilidade normal.

5.2.2 Intervalos de Confiança

Por vezes, pode ser necessário estabelecer intervalos de confiança para as médias dos tratamentos. O Intervalo de Confiança para a média de um tratamento é dado pela seguinte expressão,

$$\bar{y}_{i.} - t_{\alpha/2,k(n-1)}\sqrt{\frac{MQR}{n}} \leq \mu_i \leq \bar{y}_{i.} + t_{\alpha/2,k(n-1)}\sqrt{\frac{MQR}{n}}$$

De forma semelhante, também é possível definir um Intervalo de Confiança para a diferença entre as médias de dois tratamentos,

$$(\bar{y}_{i.} - \bar{y}_{i.}) - t_{\alpha/2,k(n-1)}\sqrt{\frac{2MQR}{n}} \leq \mu_i - \mu_i \leq (\bar{y}_{i.} - \bar{y}_{i.}) + t_{\alpha/2,k(n-1)}\sqrt{\frac{2MQR}{n}}$$

Exemplo 5.2.2: Planeamento completamente casual

A resistência de uma fibra têxtil depende da concentração de algodão. Para testar a influência da percentagem de algodão na resistência da fibra foi conduzida uma experiência em que o fator - percentagem de algodão - foi testado em quatro níveis. Para cada nível do fator foram realizadas cinco observações. A tabela apresenta os resultados obtidos.

Percentagem de Algodão	1	2	3	4	5	Total	Média
5%	7.1	8.0	7.7	7.6	6.1	36.5	7.3
10%	8.0	8.3	10.4	10.7	9.4	46.8	9.36
15%	10.7	12.4	12.2	11.0	13.1	59.4	11.88
20%	11.5	13.1	13.0	10.9	13.8	62.3	12.46

Solução

O modelo subjacente

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \begin{cases} i = 1, 2, \ldots, k \\ j = 1, 2, \ldots, n \end{cases} \qquad \varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

1. Formulação das hipóteses

$$H_{01} : \alpha_i = 0 \quad i = 1, 2, \ldots, k$$
$$H_{11} : \alpha_i \neq 0 \quad \text{para pelo menos um valor de } i$$

2. Região crítica

$$F \geq F_{0.05,3,16} = 3.24$$

3. Teste estatístico

$$STQ = 2204.38 - \frac{(205)^2}{20} = 103.13$$

$$SQT = \frac{1}{5}\left[(36.5)^2 + (46.8)^2 + (59.4)^2 + (62.3)\right] - \frac{(205)^2}{20} = 85.178$$

$$SQR = 103.13 - 85.178 = 17.952$$

Tabela ANOVA

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	F
Tratamentos	85.178	3	28.393	F=25.303
Resíduos	17.952	16	1.122	
Total	103.13	19		

4. Decisão

Rejeitar a hipótese nula, isto é, há diferenças entre as médias das resistências em função do teor de algodão.

As conclusões da Análise de Variância podem ser confirmadas pelo gráfico de caixa e bigodes da resistência em função do teor de algodão. Relativamente aos resíduos as figuras apresentam a distribuição em função dos níveis de teor de algodão e o gráfico de Probabilidades Normal.

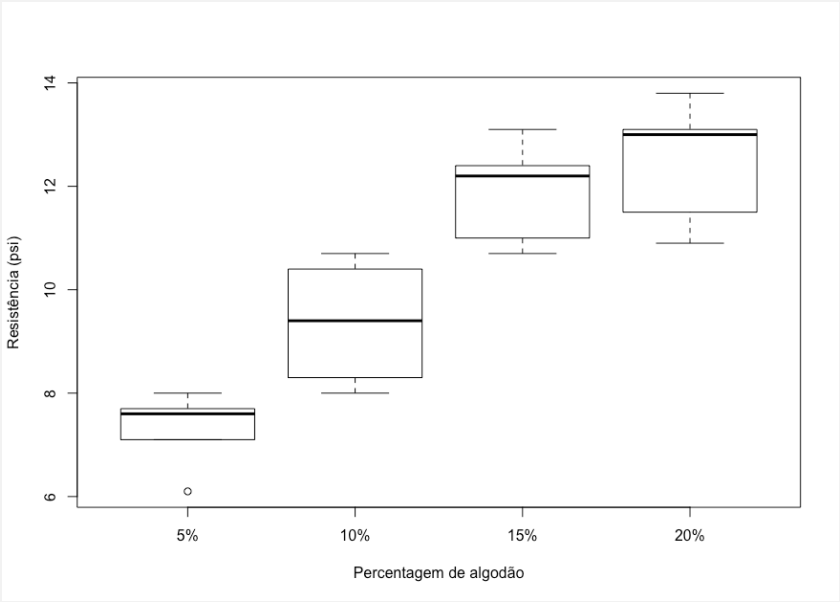


Gráfico de caixa e bigodes da resistência em função dos teores de algodão.

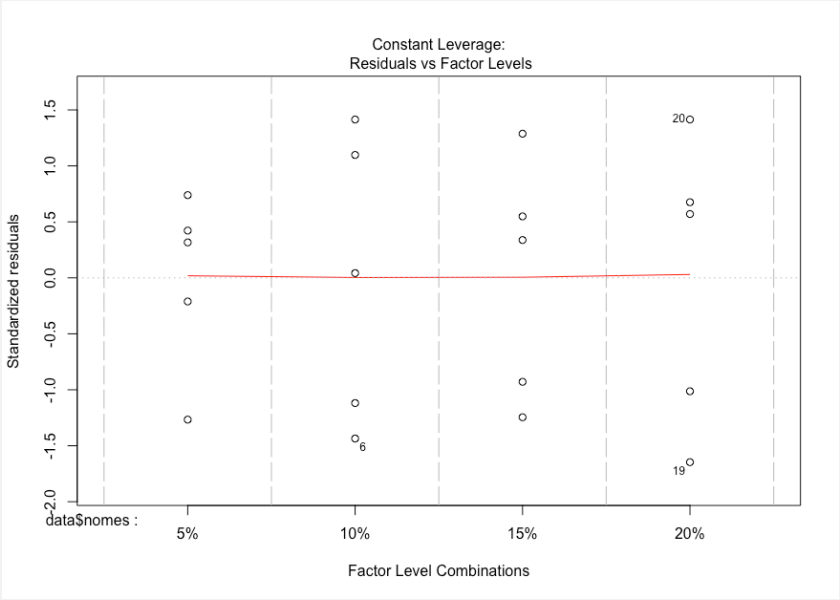


Gráfico dos resíduos em função dos teores de algodão.

5.3 Planeamento com Blocos Aleatórios

Em muitas experiências, a avaliação da influência de um fator pode requerer o controlo de um eventual fator não controlável experimentalmente, mas que influencia os resultados experimentais. Por exemplo, a comparação dos orçamentos efetuados por três engenheiros pode ser perturbada

pelos tipos de obras, pelo que a forma de controlar este fator é obter as estimativas orçamentais de cada um dos engenheiros para um número específico de tipos de obras. Num outro exemplo, a comparação do tempo de secagem de várias tintas pode ser perturbada pela humidade relativa do ar, pelo que a humidade relativa observada nos dias de experimentação deve ser tida em consideração, funcionando como um bloco.

Definição 5.3.1: Planeamento com Blocos Aleatórios

Um Planeamento com Blocos Aleatórios é um planeamento onde se pretende comparar k tratamentos aleatoriamente distribuídos por n unidades experimentais homogéneas designadas por blocos.

O Planeamento com Blocos Aleatórios tem subjacente o seguinte modelo:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad \begin{cases} i = 1, 2, \ldots, k \\ j = 1, 2, \ldots, n \end{cases}$$

em que μ é a média global, α_i o efeito devido ao tratamento i, β_j o efeito devido ao bloco j e ε_{ij} o termo de erro, cuja distribuição é assumida como normal, com média zero e variância comum σ^2. As observações podem ser apresentadas em forma tabular, com a definição dos totais, de linha e coluna, e respetivas médias:

Tratamento	Bloco 1	Bloco 2	...	Bloco n	Total	Média
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$T_{1.}$	$\bar{y_{1.}}$
2	y_{21}	y_{21}	$\cdots$	y_{2n}	$T_{2.}$	$\bar{y_{2.}}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kn}	$T_{k.}$	$\bar{y_{k.}}$
Total	$T_{.1}$	$T_{.2}$	$\cdots$	$T_{.n}$	$T_{..}$	
Média	$\bar{y_{.1}}$	$\bar{y_{.2}}$	$\cdots$	$\bar{y_{.n}}$		$\bar{y_{..}}$

Neste modelo os efeitos dos tratamentos e dos blocos verificam as seguintes condições, $\sum_{i=1}^{k} \alpha_i = 0$ e $\sum_{j=1}^{n} \beta_j = 0$. Neste planeamento, duas hipóteses nulas são testadas, nomeadamente que não existem diferenças entre os tratamentos bem como entre os blocos, ou seja,

$$H_{01} : \alpha_i = 0 \quad i = 1, 2, \ldots, k$$
$$H_{11} : \alpha_i \neq 0 \quad \text{para pelo menos um valor de } i$$

e

$$H_{02} : \beta_j = 0 \quad j = 1, 2, \ldots, n$$
$$H_{12} : \beta_j \neq 0 \quad \text{para pelo menos um valor de } j$$

A partir da igualdade

$$(y_{ij} - \bar{y}_{..}) = (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$$

considerando os quadrados de ambos os termos da equação e aplicando somatórios chega-se à seguinte partição da soma dos quadrados:

$$\sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2 = n \sum_{i=1}^{k} (y_{i.} - \bar{y}_{..})^2 + k \sum_{j=1}^{n} (y_{.j} - \bar{y}_{..})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.} - y_{.j} + \bar{y}_{..})^2.$$

É possível demonstrar que os produtos cruzados, resultantes da consideração do quadrado do segundo termo, são nulos. Assim, a variabilidade total dos dados pode ser expressa em termos da Soma Total dos Quadrados (STQ),

$$STQ = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{..})^2,$$

Soma dos Quadrados dos Tratamentos (SQT),

$$SQT = n \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})^2$$

Soma dos Quadrados dos Blocos (SQB)

$$SQB = k \sum_{j=1}^{n} (\bar{y}_{.j} - \bar{y}_{..})^2$$

e Soma dos Quadrados dos Resíduos (SQR),

$$SQR = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.} - y_{.j} + \bar{y}_{..})^2.$$

De uma forma correspondente, o número de graus de liberdade pode ser dividido como

$$kn - 1 = (k - 1) + (n - 1) + (k - 1)(n - 1).$$

As Médias dos Quadrados são:

$$MQT = \frac{SQT}{k-1} \qquad MQB = \frac{SQB}{n-1} \qquad MQR = \frac{SQR}{(k-1)(n-1)}$$

Os valores esperados podem ser deduzidos de uma forma análoga à dedução para o Planeamento Completamente Casual. Assim,

$$E[MQT] = \sigma^2 + \frac{n \sum_{i=1}^{k} \alpha_i^2}{k-1} \qquad E[MQB] = \sigma^2 + \frac{k \sum_{j=1}^{n} \beta_j^2}{n-1} \qquad E[MQR] = \sigma^2.$$

Portanto, se a Hipótese Nula H_{01} relativa aos tratamentos é verdadeira, os efeitos de todos os tratamentos são nulos e a MQT é uma estimativa não enviesada da variância. Da mesma forma,

se a Hipótese Nula H_{02} relativa aos blocos é verdadeira, também os efeitos dos blocos serão nulos e a MQB será uma estimativa não tendenciosa da variância. Atendendo a que a MQR é sempre uma estimativa não enviesada da variância, para testar as hipóteses relativas aos tratamentos e blocos é suficiente considerar as seguintes razões,

$$F_1 = \frac{MQT}{MQR} \qquad F_2 = \frac{MQB}{MQR}$$

que seguem uma distribuição F com, respetivamente, $k-1, (k-1)(n-1)$ e $n-1, (k-1)(n-1)$ graus de liberdade.

Os cálculos podem ser resumidos através da Tabela ANOVA, muitas vezes referida como tabela de duas entradas pela consideração adicional da linha referente aos blocos:

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	F
Tratamentos	SQT	$k-1$	MQT	$F_1 = MQT/MQR$
Blocos	SQB	$n-1$	MQB	$F_2 = MQB/MQR$
Resíduos	SQR	$(k-1)(n-1)$	MQR	
Total	STQ	$kn-1$		

As fórmulas baseadas nos totais podem ser deduzidas a partir das somas de quadrados, conduzindo às seguintes expressões:

$$STQ = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2 - \frac{T_{\cdot\cdot}^2}{kn} \qquad SQT = \frac{1}{n}\sum_{i=1}^{k} T_{i\cdot}^2 - \frac{T_{\cdot\cdot}^2}{kn}$$

$$SQB = \frac{1}{k}\sum_{j=1}^{n} T_{\cdot j}^2 - \frac{T_{\cdot\cdot}^2}{kn} \qquad SQR = STQ - SQT - SQB.$$

Por vezes pode ser difícil decidir sobre o uso ou não de blocos no planeamento experimental. Na pior das situações pode acontecer que os blocos foram introduzidos sem serem necessários. As consequências fazem-se sentir ao nível dos graus de liberdade e, por reflexo, no cálculo da Média dos Quadrados dos Resíduos. Se os blocos não são necessários e, nesse caso, estar-se-ia em presença de um Planeamento Completamente Casual, o número de graus de liberdade associados aos resíduos vêm reduzidos de $(n-1)$, o que, em geral, representa uma pequena perda que pode ser compensada pelo ganho em informação se os blocos são realmente importantes. De qualquer forma, o uso de um pacote estatístico permite facilmente comparar os resultados, assumindo ou não a existência de blocos.

5.3.1 Análise de Resíduos

A análise de resíduos permite verificar se as assunções relativas aos erros são ou não verificadas. Os resíduos são definidos como a diferença entre os valores observados e estimados,

$$e_{ij} = y_{ij} - \hat{y}_{ij},$$

sendo que no Planeamento com Blocos Aleatórios o valor estimado é dado por,

$$\hat{y}_{ij} = \bar{y}_{i.} + \bar{y}_{.j} - \bar{y}_{..},$$

que corresponde à resposta média quando um tratamento particular é considerado num determinado bloco. A representação gráfica dos resíduos, tal como foi referido no Planeamento Completamente Casual, através de gráficos de caixa e bigodes e do gráfico de probabilidade normal permite verificar a variabilidade e normalidade dos resíduos.

5.3.2 Intervalos de Confiança para a diferença entre dois tratamentos ou blocos

Quando a Análise da Variância conduz à rejeição da igualdade entre as médias dos tratamentos ou dos blocos, podem ser estabelecidos intervalos de confiança de acordo coma as formulas a seguir apresentadas,

- Diferença entre médias de tratamentos

$$(\bar{y}_{i.} - \bar{y}_{i.}) \pm t_{\alpha/2,(k-1)(n-1)} s \sqrt{\frac{2}{n}}$$

- Diferença entre médias de blocos

$$(\bar{y}_{.j} - \bar{y}_{.j}) \pm t_{\alpha/2,(k-1)(n-1)} s \sqrt{\frac{2}{k}}$$

com $s = \sqrt{MQR}$.

Exemplo 5.3.1: Planeamento com blocos aleatórios

Uma Federação Desportiva pretende determinar se a ingestão de um medicamento pode gerar níveis, acima do limite legal, de uma substância dopante. Para tal, três cobaias foram injetadas com três diferentes concentrações do medicamento.
De cada cobaia foram retiradas quatro amostras de urina enviadas para análise em quatro laboratórios diferentes.

	Laboratório			
	1	2	3	4
Injeção 1	0.8	1.6	0.9	1.0
Injeção 2	1.7	2.6	2.1	2.1
Injeção 3	3.0	3.4	3.2	3.0

A Federação pretende determinar não só se existem diferenças de concentração na urina, mas também se existem diferenças entre os quatro laboratórios.

Solução

O modelo subjacente

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \begin{cases} i = 1, 2, \ldots, k \\ j = 1, 2, \ldots, n \end{cases} \qquad \varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

1. Formulação das hipóteses

$$\begin{aligned} &H_{01} : \alpha_i = 0 \quad i = 1, 2, \ldots, k \\ &H_{11} : \alpha_i \neq 0 \quad \text{para pelo menos um valor de } i \end{aligned}$$

$$\begin{aligned} &H_{02} : \beta_j = 0 \quad j = 1, 2, \ldots, n \\ &H_{12} : \beta_j \neq 0 \quad \text{para pelo menos um valor de } j \end{aligned}$$

2. Região crítica

$$\begin{aligned} F_1 &\geq F_{0.05,2,6} = 5.14 \\ F_2 &\geq F_{0.05,3,6} = 4.76 \end{aligned}$$

3. Teste estatístico

	Laboratório					
	1	2	3	4	Total	Média
Injeção 1	0.8	1.6	0.9	1.0	4.3	1.1
Injeção 2	1.7	2.6	2.1	2.1	8.5	2.1
Injeção 3	3.0	3.4	3.2	3.0	12.6	3.2
Total	5.5	7.6	6.2	6.1	25.4	
Média	1.8	2.5	2.1	2.0		2.1

$$\begin{aligned} STQ &= 63.3 - \tfrac{(25.4)^2}{12} = 9.517 \\ SQT &= \tfrac{1}{4}\left[(4.3)^2 + (8.5)^2 + (12.6)^2\right] - \tfrac{(25.4)^2}{12} = 8.612 \\ SQB &= \tfrac{1}{3}\left[(5.5)^2 + (7.6)^2 + (6.2)^2 + (6.1)^2\right] - \tfrac{(25.4)^2}{12} = 0.790 \\ SQR &= 9.517 - 8.612 - 0.790 = 0.115 \end{aligned}$$

Tabela ANOVA

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	F
Tratamentos	8.612	2	4.306	$F_1 = 224.652$
Blocos	0.790	3	0.263	$F_2 = 13.739$
Resíduos	0.115	6	0.019	
Total	9.517	11		

x

4. Decisão Rejeitar H_{01} e H_{02}, isto é, existem diferenças entre os efeitos das injeções (tratamentos) e entre as concentrações detetadas pelos laboratórios (blocos).

A análise gráfica permite corroborar as conclusões da Análise da Variância. As figuras apresentam o gráfico de caixa e bigodes da concentração em função da quantidade injetada e laboratórios. Claramente se verifica que as concentrações detetadas dependem da quantidade injetada.

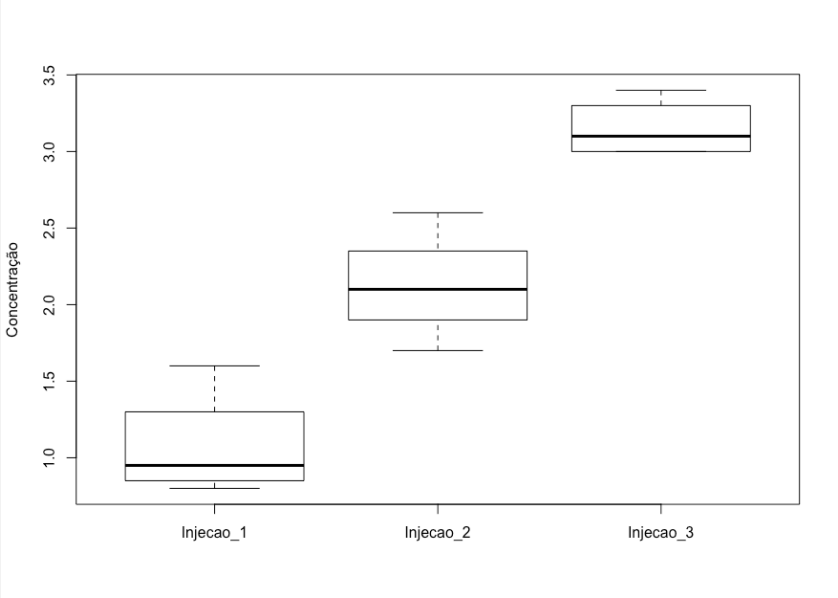


Gráfico de caixa e bigodes do laboratório em função da quantidade injetada.

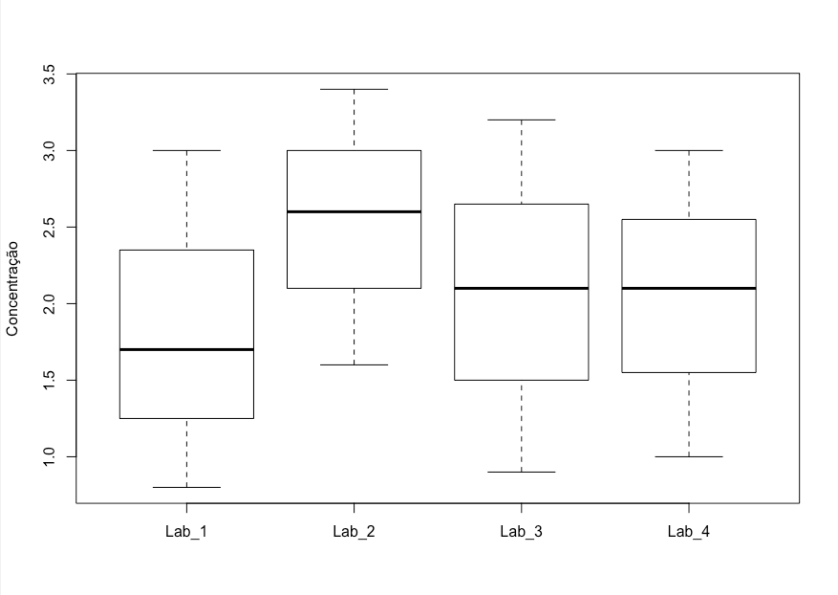


Gráfico de caixa e bigodes da concentração detetada em função do laboratório.

A tabela dos resíduos permite realizar as análises gráficas que permitem confirmar as assunções estabelecidas.

Tabela de Resíduos

	Laboratório 1	Laboratório 2	Laboratório 3	Laboratório 4
Injeção 1	0	0.1	-0.2	0
Injeção 2	-0.1	0.1	0	0.1
Injeção 3	0.1	-0.2	0	-0.1

Os gráficos de caixa e bigodes para os resíduos em função das quantidades injetadas (tratamentos), em função das concentrações detetadas pelos laboratórios (blocos) e o gráfico de Probabilidade Normal dos resíduos são apresentados nas figuras seguintes.

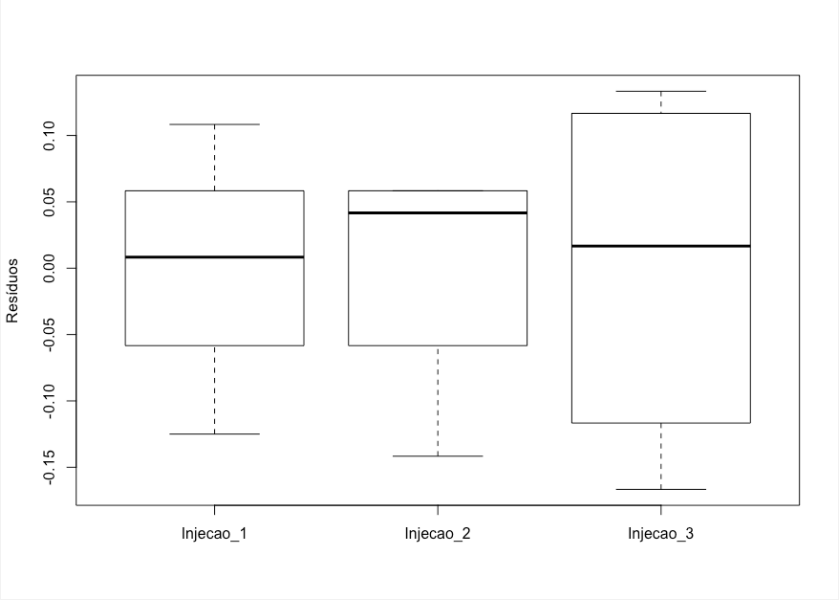


Gráfico de caixa e bigodes dos resíduos em função da quantidade injetada.

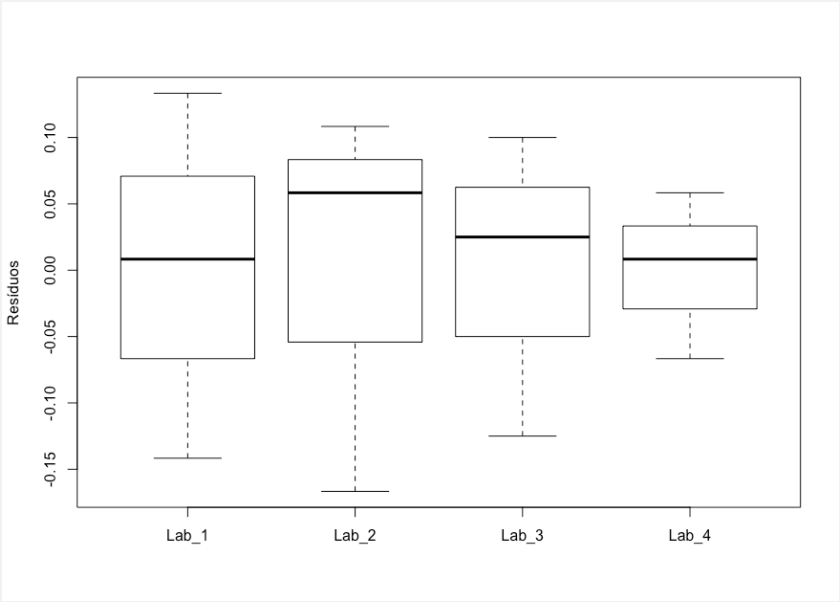


Gráfico de caixa e bigodes dos resíduos em função do laboratório.

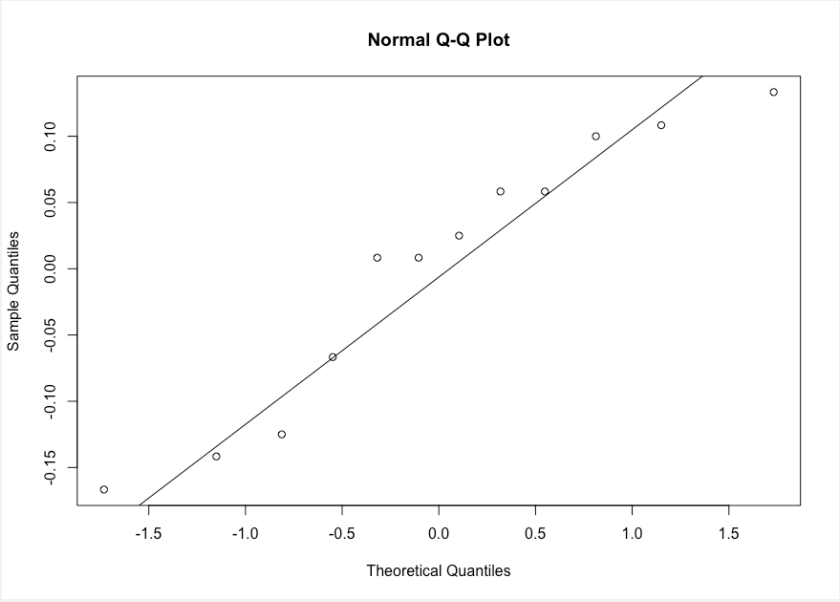


Gráfico de Probabilidade Normal dos resíduos.

5.4 Planeamento Fatorial

A realização de uma experiência é uma forma de conhecer e compreender um sistema. Em larga medida, a realização de experiências está na base do método científico associado ao progresso da ciência e da engenharia. Contudo, a validade das conclusões de uma experiência depende de

maneira essencial da forma como a experiência foi conduzida, isto é, das condições experimentais e do próprio planeamento.

Em geral, numa experiência pretende-se determinar como diferentes variáveis independentes estão relacionadas com (influenciam) a variável resposta. Assim, numa experiência em que se pretende estudar a influência da temperatura e da concentração numa reação química, os fatores são a temperatura e a concentração, e os níveis são as diferentes temperaturas e concentrações a que experiência vai ser realizada, por exemplo, 150 ^{o}C ou 200 ^{o}C e 10 g/l ou 20 g/l. Um tratamento seria, por exemplo, a realização de uma experiência com o fator temperatura no nível 200 ^{o}C e o outro fator concentração no nível 10 g/l.

Quando se pretende estudar a influência de vários fatores numa experiência, a dificuldade reside na forma de planear a experiência. No Planeamento Fatorial, para cada combinação de níveis dos vários fatores são realizadas várias experiências, também referidas como replicações experimentais. Assim, por exemplo, em presença de dois fatores A e B, cada um com, respetivamente, a e b níveis, seriam realizadas replicações para cada uma das combinações. O objetivo de uma experiência fatorial é determinar o efeito de um fator na variável resposta, isto é, como a variável resposta se altera à medida que os níveis dos fatores são variados. Para além da influência de cada fator na variável resposta, um outro objetivo de uma experiência fatorial é determinar a eventual existência de interação entre fatores. A interação existe quando a diferença na variável resposta para os níveis de um fator é diferente a outros níveis dos outros fatores ou, dito de outra forma, quando a resposta não é igual à soma dos contributos de cada um dos fatores.

Definição 5.4.1: Planeamento Fatorial

O Planeamento Fatorial é um planeamento em que os tratamentos incluem todas as combinações possíveis dos níveis dos fatores, sendo as experiências correspondentes aos diversos tratamentos, realizadas segundo uma ordem aleatória.

Definição 5.4.2: Interação

No Planeamento Fatorial a interação entre fatores existe quando a diferença entre os níveis de um fator depende dos níveis de outro.

Exemplo 5.4.1: Planeamento fatorial

Considere-se uma experiência em que se pretende determinar o rendimento de uma reação química, medido em gramas de produto, em função da concentração e da temperatura,

respetivamente, com níveis de 10 e 20 g/l, e 150 °C e 200 °C. A tabela apresenta os resultados da experiência.

Solução

		Fator B Temperatura	
		B1-150 °C	B2-200 °C
Fator A	A1-10 g/l	40	45
Concentração	A2-20 g/l	47	52

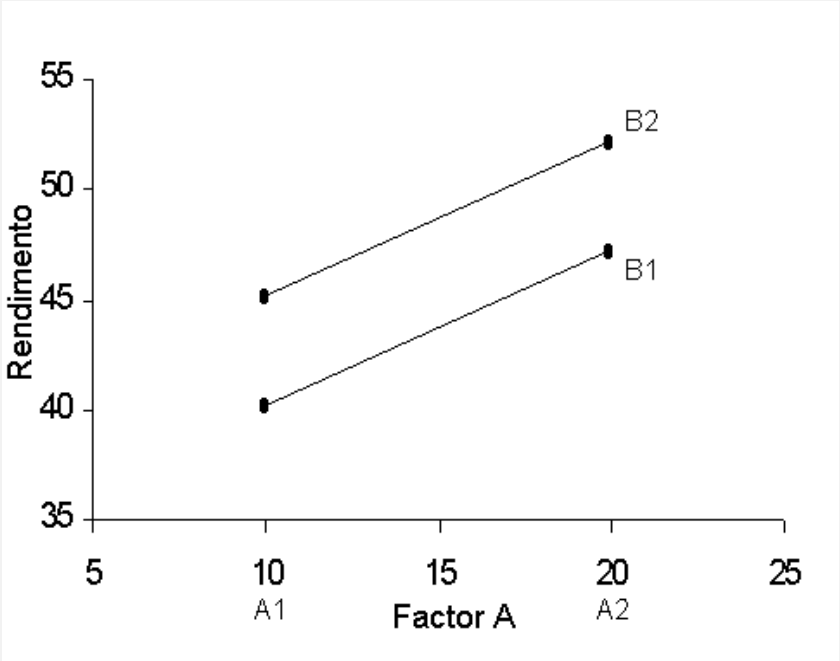
Assim, o efeito do Fator A pode ser estimado como a média da diferença entre os rendimentos nos dois níveis do Fator B, isto é,

$$\text{Efeito A} = \frac{1}{2}\left[(47-40)+(52-45)\right] = 7$$

De forma semelhante, o efeito do Fator B é dado por,

$$\text{Efeito B} = \frac{1}{2}\left[(45-40)+(52-47)\right] = 7$$

Assim, verifica-se que os efeitos dos Fatores A e B são aditivos, o que graficamente se traduz em duas linhas paralelas, como se pode ver na seguinte figura.



No entanto, se os resultados da experiência fossem os observados na seguinte tabela,

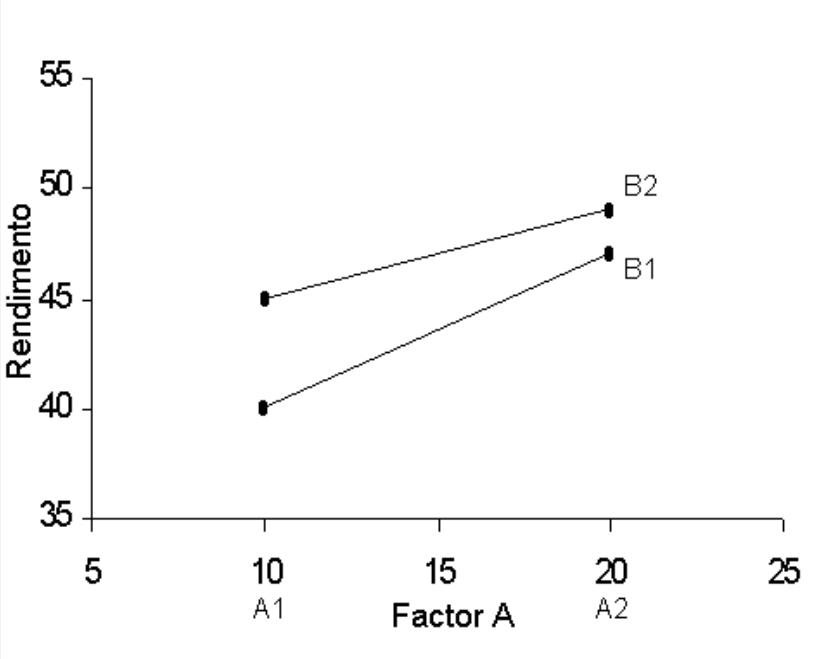
		Fator B Temperatura	
		B1-150 °C	B2-200 °C
Fator A	A1-10 g/l	40	45
Concentração	A2-20 g/l	47	48

os efeitos dos Fatores A e B, seriam,

$$\text{Efeito A} = \frac{1}{2}\left[(47-40)+(48-45)\right] = 5$$

$$\text{Efeito B} = \frac{1}{2}\left[(45-40)+(48-47)\right] = 3$$

Contudo, verifica-se que o efeito de A depende do nível do Fator B, o que denuncia a existência de interação entre os fatores. Graficamente, tal facto, traduz-se por linhas não paralelas, como se pode ver na seguinte figura.



O efeito da interação entre os dois fatores pode ser estimado como a média da diferença entre os efeitos do Fator A em cada nível do Fator B, ou vice-versa, isto é, como a média da diferença entre os efeitos do Fator B em cada nível do Fator A.

$$\text{Efeito AB} = \frac{1}{2}\left[(48-45)-(47-40)\right] = \frac{1}{2}\left[(48-47)-(45-40)\right] = -2$$

O cálculo da interação para a primeira situação conduziria a um resultado nulo, isto é,

$$\text{Efeito AB} = \frac{1}{2}\left[(52-45)-(47-40)\right] = \frac{1}{2}\left[(52-47)-(45-40)\right] = 0$$

Se a interação é significativa, os efeitos principais de cada fator são pouco relevantes pois, como se pode ver pela segunda figura deste exemplo, o efeito de do Fator A depende do nível do Fator B. Portanto, na presença de interação entre fatores, os efeitos principais devidos a cada fator deixam ter grande importância dado que o resultado de um efeito não é proporcional ao seu valor.

Existe um outro planeamento experimental, muito divulgado, em que só um fator é variado de cada vez, isto é, com referência ao exemplo, seriam só realizadas três experiências, para os seguintes tratamentos (A1, B1), (A1, B2) e (A2, B1). Comparativamente com o planeamento fatorial este planeamento requer a realização de menos uma experiência. Contudo, este planeamento pressupõe que não existem interações, pois admite que o rendimento para o tratamento (A2, B2) resultaria dos efeitos de A e de B determinados com base nos três tratamentos, o que só conduziria à resposta correta quando os efeitos dos fatores são aditivos.

Por outro lado, o planeamento fatorial é mais eficiente que o planeamento baseado na variação de um fator de cada vez dado que, no primeiro, as estimativas dos efeitos são baseadas em quatro observações enquanto que, no segundo, somente em duas, tal como se pode ver no quadro abaixo.

	Um fator de cada vez	Completamente Fatorial
Experiências	(A1, B1), (A1, B2), (A2, B1)	(A1, B1), (A1, B2), (A2, B1), (A2,B2)
Efeito do Fator A	(A2,B1)-(A1,B1)	$\frac{1}{2}$[(A2,B1)-(A1,B1)+(A2,B2)-(A1,B2)]
Efeito do Fator B	(A1,B2)-(A1,B1)	$\frac{1}{2}$[(A1,B2)-(A1,B1)+(A2,B2)-(A2,B1)]

As experiências fatoriais requerem também que a ordem de execução das experiências para cada tratamento seja aleatória por forma a evitar a influência de fatores externos não controláveis.

5.4.1 Planeamento Fatorial com dois fatores

O Planeamento Fatorial mais simples envolve somente dois fatores. Para o efeito, considerem-se dois fatores A e B, respetivamente, com a e b níveis. Para cada combinação de níveis de fatores, isto é, para cada tratamento, a experiência é reproduzida n vezes (n replicações). Assim, no total existirão abn resultados experimentais que devem ser recolhidos de uma forma aleatória. A replicação da combinação do Fator A, no nível i, com o nível j do Fator B, a observação y_{ijk} será descrita pelo seguinte modelo,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots r \end{cases} \quad \varepsilon_{ijk} \sim N\left(0, \sigma^2\right)$$

onde μ é a média global, α_i o efeito devido ao nível i do Fator A, β_j o efeito devido ao nível j do Fator B, γ_{ij} o efeito devido à interação entre os Fatores A e B, e ε_{ijk} o termo de erro, com uma distribuição normal com média zero e variância comum σ^2. Neste planeamento, três hipóteses são testadas, nomeadamente a não existência de efeitos devidos ao Fator A, Fator B e à interação AB.

		Fator B					
		1	2	...	b	Total	Média
Fator A	1	$y_{111}, y_{112},$ $\ldots, y_{11r}$	$y_{121}, y_{122},$ $\ldots, y_{12r}$	$\cdots$	$y_{1b1}, y_{1b2},$ $\ldots, y_{1br}$	$T_{1..}$	$\bar{y}_{1..}$
	2	$y_{211}, y_{212},$ $\ldots, y_{21r}$	$y_{221}, y_{222},$ $\ldots, y_{22r}$	$\cdots$	$y_{2b1}, y_{2b2},$ $\ldots, y_{2br}$	$T_{2..}$	$\bar{y}_{2..}$
	...	...	...	...	...	...	...
	a	$y_{a11}, y_{a12},$ $\ldots, y_{a1r}$	$y_{a21}, y_{a22},$ $\ldots, y_{a2r}$	$\cdots$	$y_{ab1}, y_{ab2},$ $\ldots, y_{abr}$	$T_{a..}$	$\bar{y}_{a..}$
	Total	$T_{.1.}$	$T_{.2.}$	...	$T_{.n.}$	$T_{...}$	
	Média	$\bar{y_{.1.}}$	$\bar{y_{.2.}}$	...	$\bar{y_{.n.}}$		$\bar{y_{...}}$

A decomposição da variabilidade nas várias componentes, a partir da igualdade

$$(y_{ijk} - \bar{y}_{...}) = (\bar{y}_{i..} - \bar{y}_{...}) + (\bar{y}_{.j.} - \bar{y}_{...}) + (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}) + (y_{ijk} - \bar{y}_{ij.})$$

e considerando o quadrado de ambos os termos, estendidos aos diversos somatórios, permite definir a seguinte identidade,

$$\begin{aligned}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{...})^2 = \quad & bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2 + an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2 + \\ & n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2\end{aligned}$$

já que todos os seis produtos cruzados são nulos. Assim, a variabilidade total pode ser expressa nos seguintes termos,

$$STQ = SQF_A + SQF_B + SQI_{AB} + SQR$$

ou seja, Soma Total dos Quadrados (STQ),

$$STQ = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{...})^2,$$

Soma dos Quadrados do Fator A ($SQFA$),

$$SQF_A = bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2,$$

Soma dos Quadrados do Fator B ($SQFB$)

$$SQF_B = an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2,$$

Soma dos Quadrados da Interação AB ($SQIAB$)

$$SQI_{AB} = n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2,$$

e Soma dos Quadrados dos Resíduos (SQR),

$$SQR = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(y_{ijk} - \bar{y}_{ij.}\right)^2.$$

De forma semelhante, o número total de graus de liberdade poder ser particionado, de acordo com cada uma das parcelas indicadas acima, isto é,

$$abn - 1 = (a-1) + (b-1) + (a-1)(b-1) + ab(n-1)$$

As Médias dos Quadrados obtêm-se pela divisão da Soma dos Quadrados pelo respetivo número de graus de liberdade, sendo que o valor esperado é dado por,

$$E\left[MQF_A\right] = E\left[\frac{SQF_A}{a-1}\right] = \sigma^2 + \frac{bn\sum_{i=1}^{a}\alpha_i^2}{a-1}$$

$$E\left[MQF_B\right] = E\left[\frac{SQF_B}{b-1}\right] = \sigma^2 + \frac{an\sum_{i=1}^{a}\beta_i^2}{b-1}$$

$$E\left[MQI_{AB}\right] = E\left[\frac{SQI_{AB}}{(a-1)(b-1)}\right] = \sigma^2 + \frac{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ji}^2}{(a-1)(b-1)}$$

$$E\left[MQR\right] = E\left[\frac{SQR}{ab(n-1)}\right] = \sigma^2.$$

Como foi mencionado, três hipóteses podem ser testadas, relativamente aos efeitos do Fator A, Fator B e Interação AB, cuja formulação será, respetivamente,

$$\begin{aligned} &H_{01}: \alpha_i = 0 \quad i = 1, 2, \ldots, a \\ &H_{11}: \alpha_i \neq 0 \quad \text{para pelo menos um valor de } i \end{aligned}$$

$$\begin{aligned} &H_{02}: \beta_j = 0 \quad j = 1, 2, \ldots, b \\ &H_{12}: \beta_j \neq 0 \quad \text{para pelo menos um valor de } j \end{aligned}$$

$$\begin{aligned} &H_{03}: \gamma_{ij} = 0 \quad i = 1, 2, \ldots, a; j = 1, 2, \ldots, b \\ &H_{13}: \gamma_{ij} \neq 0 \quad \text{para pelo menos um par } ij \end{aligned}$$

Se todas as hipóteses nulas forem verdadeiras, as Médias dos Quadrados do Fator A, Fator B e Interação AB, serão estimativas não tendenciosas da variância σ^2. Por outro lado, a Média dos Quadrados dos Resíduos é sempre uma estimativa não enviesada de σ^2, pelo que para testar as três hipóteses nulas basta considerar as seguintes razões,

$$F_1 = \frac{MQF_A}{MQR} \qquad F_2 = \frac{MQF_B}{MQR} \qquad F_3 = \frac{MQI_{AB}}{MQR}$$

que seguem uma distribuição F com, respetivamente, $a-1, ab(n-1)$, $b-1, ab(n-1)$ e $(a-1)(b-1), ab(n-1)$ graus de liberdade.

Os cálculos podem ser apresentados de forma resumida na Tabela ANOVA,

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	F
Fator A	SQF_A	$a-1$	MQF_A	$F_1 = MQF_A/MQR$
Fator B	SQF_B	$b-1$	MQF_B	$F_2 = MQF_B/MQR$
Interação AB	SQI_{AB}	$(a-1)(b-1)$	MQI_{AB}	$F_3 = MQF_IAB/MQR$
Resíduos	SQR	$(ab-1)(n-1)$	MQR	
Total	STQ	$kn-1$		

O cálculo das Somas dos Quadrados pode ser efetuado por recurso às seguintes expressões:

$$STQ = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{T_{...}^2}{abn}$$

$$SQF_A = \frac{1}{bn}\sum_{i=1}^{k} T_{i..}^2 - \frac{T_{...}^2}{abn} \qquad SQF_B = \frac{1}{an}\sum_{j=1}^{n} T_{.j.}^2 - \frac{T_{...}^2}{abn}$$

$$SQR = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 - \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} T_{ij.}^2}{n} \qquad SQI = STQ - SQF_A - SQF_B - SQR,$$

com,

$$T_{ij.} = \sum_{k=1}^{n} y_{ijk}.$$

Análise dos Resíduos

A verificação da adequação do modelo é feita com base na análise dos resíduos. Para um Planeamento Fatorial com dois fatores, os resíduos são,

$$\varepsilon_{ijk} = y_{ijk} - \bar{y}_{ij.},$$

ou seja, a diferença entre os valores observados em cada célula e a respetiva média. A representação gráfica dos resíduos no gráfico de Probabilidade Normal, dos gráficos de caixa e bigodes dos resíduos em função de cada um dos fatores e em função dos valores previstos, permite verificar se os resíduos seguem uma distribuição normal com variância comum σ^2.

Intervalos de Confiança

A construção de intervalos de confiança é semelhante ao procedimento enunciado para o planeamento com blocos. No entanto uma ressalva deve ser feita. Se não existirem interações entre os fatores, as comparações podem ser feitas usando as médias associadas às linhas ou às colunas. Contudo, se existir um interação significativa, a comparação entre os níveis de um fator pode ser perturbada. Neste caso, a comparação entre os níveis de um fator deve ser feita para um nível fixo do outro fator.

Exemplo 5.4.2: Planeamento fatorial

O débito de uma extrusora depende da velocidade de rotação do parafuso e da temperatura do cilindro. Para determinar as melhores condições de operação, foram realizadas experiências em que os dois fatores, velocidade e temperatura, foram estudados, respetivamente, em três e em dois níveis: velocidade de rotação (rpm), 10, 30 e 50, temperatura do cilindro, 160 °C e 200 °C.

Para cada tratamento, isto é, para cada combinação de níveis de fatores, foram recolhidas 3 observações do débito (Kg/h).

Os resultados obtidos são apresentados na tabela seguinte:

		Velocidade de Rotação (rpm)		
		10	30	50
Temperatura	160 °C	2.6; 2.5; 2.8	4.8; 4.2; 5.1	3.6; 3.8; 3.5
	200 °C	2.4; 2.2; 2.5	5.5; 6.2; 6.4	1.9; 2.2; 1.8

Solução

O modelo subjacente

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \begin{cases} i = 1, 2, \ldots, a \\ j = 1, 2, \ldots, b \\ k = 1, 2, \ldots r \end{cases} \qquad \varepsilon_{ijk} \sim N\left(0, \sigma^2\right)$$

1. Formulação das hipóteses

$$H_{01} : \alpha_i = 0 \quad i = 1, 2, \ldots, a$$
$$H_{11} : \alpha_i \neq 0 \quad \text{para pelo menos um valor de } i$$

$$H_{02} : \beta_j = 0 \quad j = 1, 2, \ldots, b$$
$$H_{12} : \beta_j \neq 0 \quad \text{para pelo menos um valor de } j$$

$$H_{03} : \gamma_{ij} = 0 \quad i = 1, 2, \ldots, a; j = 1, 2, \ldots, b$$
$$H_{13} : \gamma_{ij} \neq 0 \quad \text{para pelo menos um par } ij$$

2. Região crítica

$$F_1 \geq F_{0.05,1,12} = 4.75$$
$$F_2 \geq F_{0.05,2,12} = 3.89$$
$$F_2 \geq F_{0.05,2,12} = 3.89$$

3. Teste estatístico

		Velocidade de Rotação (rpm)			
		10	30	50	Total
Temperatura	160 °C	$T_{11.}$=7.9	$T_{12.}$=14.1	$1T_{13.}$=0.9	$T_{1..}$=32.9
	200 °C	$T_{21.}$=7.1	$T_{22.}$=18.1	$T_{23.}$=5.9	$T_{2..}$=31.1
	Total	$T_{.1.}$=15.0	$T_{.2.}$=32.2	$T_{.3.}$=16.8	$T_{...}$=64.0

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}^2 = 265.38$$

$$STQ = 265.38 - \frac{(64.0)^2}{18} = 37.824$$

$$SQF_A = \frac{1}{9}\left(32.9^2 + 31.1^2\right) - \frac{(64.0)^2}{18} = 0.180$$

$$SQF_B = \frac{1}{6}\left(15.0^2 + 32.2^2 + 16.8^2\right) - \frac{(64.0)^2}{18} = 29.791$$

$$SQR = 265.38 - \frac{\left(7.9^2 + 14.1^2 + 10.9^2 + 7.1^2 + 18.1^2 + 5.9^2\right)}{3} = 1.093$$

$$SQI = 37.824 - 0.180 - 27.791 - 1.091 = 6.760$$

Tabela ANOVA

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	F
Fator A	0.180	1	0.180	F_1=1.976
Fator B	29.791	2	14.896	F_2=163.488
Interação AB	6.760	2	3.380	F_3=37.098
Resíduos	1.091	12	0.911	
Total	37.824	17		

4. Decisão Rejeitar H_{02} e H_{03} isto é, existem diferenças entre as velocidades de rotação e existe uma interação entre a velocidade de rotação e a temperatura. O facto de existir uma interação significativa implica que a importância de ambos os fatores deve ser considerada, independentemente de existirem diferenças significativas associadas aos dois ou a um só dos fatores.

O gráfico do rendimento médio em função das velocidades de rotação mostra claramente a existência de interação devido ao facto de as linhas não serem paralelas. Para além disso, é claro que a melhor combinação de níveis de fatores resulta da velocidade de rotação de 30 rpm com a temperatura a 200 °C. As comparações entre as médias individuais

de cada fator, se não existirem interações podem ser obtidas através de uma técnica de comparação múltipla. Contudo, na presença de interação, a comparação não é óbvia, e deve ser executada para um nível fixo do outro fator.

A tabela apresenta os resíduos para a experiência de extrusão. A figura apresenta o gráfico de probabilidade normal onde, com exceção de alguns pontos numa das caudas, a maioria se distribui ao longo de uma linha reta.

		Velocidade de Rotação (rpm)		
		10	30	50
Temperatura	160 °C	-0.033; -0.133; 0.167	0.100; -0.500; 0.400	-0.033; 0.167; -0.133
	200 °C	0.033; -0.167; 0.133	-0.533; 0.167; 0.367	-0.067; 0.233; -0.167

FALTA FIGURA Figura 5.12 - Gráfico do rendimento médio em função da velocidade de rotação.

As Figuras 5.14 a 5.16 mostram os gráficos relativos aos resíduos.

FALTA FIGURA Figura 5.13 - Gráfico de probabilidade normal dos resíduos.

FALTA FIGURA Figura 5.14 - Gráfico dos resíduos em função do valor previsto

FALTA FIGURA Figura 5.15 - Gráfico dos resíduos em função da velocidade de rotação.

FALTA FIGURA Figura 5.16 - Gráfico dos resíduos em função da temperatura.

5.5 Planeamento Fatorial 2^k

Um tipo particular de planeamento fatorial é o chamado planeamento 2^k em que cada fator é avaliado a dois níveis. Este planeamento é usado em fases exploratórias, onde se pretende investigar a relevância de muitos fatores, constituindo o planeamento fatorial com o menor número de tratamentos.

5.5.1 Planeamento 2^2

O planeamento 2^2 constitui a forma mais simples de estudar o efeito de dois Fatores A e B, cada um a dois níveis. Portanto, existem quatro tratamentos correspondentes às combinações possíveis dos níveis dos dois fatores. Assumindo que cada fator é representado no segundo nível pelo sinal (+) e no primeiro nível pelo sinal (-), numa forma tabular o contributo de cada fator pode ser avaliado da seguinte forma

	Média	A	B	AB
(1)	+	-	-	+
a	+	+	-	-
b	+	-	+	-
ab	+	+	+	+
	1/4	1/2	1/2	1/2

Assim, a letra a corresponde ao tratamento com o Fator A no segundo nível e o Fator B no primeiro nível; de forma similar, b corresponde ao segundo nível do Fator B e primeiro do Fator A; ab representa o tratamento correspondente aos segundos níveis dos dois Fatores A e B. A combinação correspondente aos primeiros níveis dos dois fatores é representada pelo símbolo (1). As colunas da matriz são ortogonais e a coluna respeitante à interação pode ser determinada pelo produto, linha a linha, dos sinais das colunas A e B. A Figura 5.1 mostra uma representação geométrica deste planeamento.

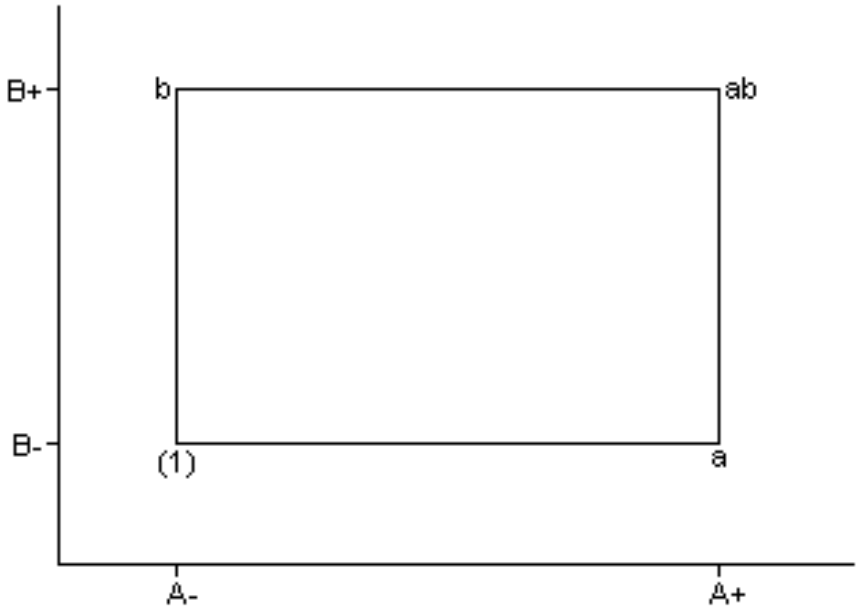


Figura 5.1: Planeamento 2^2.

O efeito do Fator A pode ser estimado como a média da diferença $[a-(1)]$ com a diferença $[ab-b]$. Assim, os efeitos do Fator A e do Fator B podem ser estimados por,

- Efeito de A,

$$e_A = \frac{1}{2}\left[ab - b + a - (1)\right]$$

- Efeito de B,

$$e_B = \frac{1}{2}\left[ab + b - a - (1)\right].$$

O efeito da interação AB resulta da média da diferença entre o efeito de B no segundo nível de A, com o efeito de B no primeiro nível de A, ou da média da diferença de do efeito de A no segundo nível de B, com o efeito de A no primeiro nível de B, isto é,

- Efeito AB,

$$e_{AB} = \frac{1}{2}\left[(ab - a) - (b - (1))\right] = \frac{1}{2}\left[(ab - b) - (a - (1))\right] = \frac{1}{2}\left[ab - b - a + (1)\right].$$

A Soma Total dos Quadrados, a Soma dos Quadrados dos Tratamentos e a Soma dos Quadrados dos Resíduos podem ser estimadas a partir dos efeitos calculados. Assim, para uma experiência com n replicações, a tabela ANOVA correspondente seria,

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	F
Fator A	SQF_A	1	MQF_A	$F_1 = MQF_A/MQR$
Fator B	SQF_B	1	MQF_B	$F_2 = MQF_B/MQR$
Interação AB	SQI_{AB}	1	MQI_{AB}	$F_3 = MQF_IAB/MQR$
Resíduos	SQR	$4(n-1)$	MQR	
Total	STQ	$4n-1$		

A Soma dos Quadrados de um dado Fator é obtida através da seguinte relação

$$SQF = \frac{e_F^2}{n}.$$

A Soma Total dos Quadrados (STQ), é obtida pela mesma relação para o planeamento fatorial, isto é,

$$STQ = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{...})^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{n} y_{ijk}^2 - 4n\bar{y}_{...}^2.$$

As expressões para as restantes somas são: Soma dos Quadrados do Fator A (SQF_A),

$$SQF_A = \frac{e_A^2}{n} = \frac{\left[ab - b + a - (1)\right]^2}{4n},$$

Soma dos Quadrados do Fator B (SQF_B)

$$SQF_B = \frac{e_B^2}{n} = \frac{\left[ab + b - a - (1)\right]^2}{4n},$$

Soma dos Quadrados da Interação AB ($SQIAB$)

$$SQI_{AB} = \frac{e_{AB}^2}{n} = \frac{\left[ab - b - a + (1)\right]^2}{4n},$$

Soma dos Quadrados dos Resíduos (SQR),

$$SQR = SQT - SQF_A - SQF_B - SQI_{AB}.$$

Exemplo 5.5.1: Planeamento 2^2

Pretende-se estudar o efeito de um reagente (Fator A) e de um reator (Fator B) no rendimento de um processo químico. O reagente será aplicado em duas concentrações (20% e 30%) e dois reatores diferentes (R1 e R2) serão empregues. Cada combinação de fatores foi ensaiada três vezes. Os resultados observados representam o rendimento em termos da quantidade produzida (kg/h).

Fator A	Fator B	Rendimento
A-	B-	70
A+	B-	90
A-	B+	45
A+	B+	77
A-	B-	63
A+	B-	82
A-	B+	47
A+	B+	75
A-	B-	68
A+	B-	83
A-	B+	58
A+	B+	73

O que pode concluir?

Solução

O modelo subjacente

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \begin{cases} i = 1,2 \\ j = 1,2 \\ k = 1,2,3 \end{cases} \qquad \varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

1. Formulação das hipóteses

$$H_{01} : \alpha_i = 0 \quad i = 1,2$$
$$H_{11} : \alpha_i \neq 0 \quad \text{para pelo menos um valor de } i$$

$$H_{02} : \beta_j = 0 \quad j = 1,2$$
$$H_{12} : \beta_j \neq 0 \quad \text{para pelo menos um valor de } j$$

$$H_{03} : \gamma_{ij} = 0 \quad i = 1,2; j = 1,2$$
$$H_{13} : \gamma_{ij} \neq 0 \quad \text{para pelo menos um par } ij$$

2. Região crítica

$$F_1, F_2, F_3 \geq F_{0.05,1,8} = 5.32$$

3. Teste estatístico A tabela apresenta os cálculos dos efeitos de cada fator.

	A	B	AB	Resultados	Total	Média
(1)	-1	-1	1	70; 63; 68	201	67
a	1	-1	-1	90; 82; 83	255	85
b	-1	1	-1	45; 47, 58	150	50
ab	1	1	1	77; 75; 73	225	75
					831	69.25

Os efeitos dos fatores são,

- Efeito de A,

$$e_A = \frac{1}{2}\left[ab - b + a - (1)\right] = \frac{[75 - 50 + 85 - 67]}{2} = 21.5$$

- Efeito de B,

$$e_B = \frac{1}{2}\left[ab + b - a - (1)\right] = \frac{[75 + 50 - 85 - 67]}{2} = -13.5$$

- Efeito AB,

$$e_{AB} = \frac{1}{2}\left[ab - b - a + (1)\right] = \frac{[75 - 80 - 85 + 67]}{2} = 3.5$$

As respetivas Somas de Quadrados são

- Soma dos Quadrados do Fator A (SQF_A),

$$SQF_A = \frac{e_A^2}{n} = \frac{(64.5)^2}{3} = 1386.75$$

- Soma dos Quadrados do Fator B (SQF_B)

$$SQF_B = \frac{e_B^2}{n} = \frac{(-40.5)^2}{3} = 546.75$$

- Soma dos Quadrados da Interação AB (SQI_{AB})

$$SQI_{AB} = \frac{e_{AB}^2}{n} = \frac{(10.5)^2}{3} = 36.75$$

- Soma Total dos Quadrados (STQ),

$$STQ = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{n} y_{ijk}^2 - 4n\bar{y}_{...}^2 = 59687 - 12\left(\frac{831}{12}\right)^2$$

- Soma dos Quadrados dos Resíduos (SQR)

$$\begin{aligned} SQR &= STQ - SQF_A - SQF_B - SQI_{AB} \\ &= 2140.25 - 1386.75 - 546.75 - 36.75 = 170.00 \end{aligned}$$

A tabela ANOVA pode então ser preenchida,

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	F
Fator A	1386.75	1	1386.75	$F_1 = 65.259$
Fator B	546.75	1	546.75	$F_2 = 25.729$
Interação AB	36.75	1	36.75	$F_3 = 1.729$
Resíduos	170.00	8	21.25	
Total	2140.25	11		

4. Decisão

   Rejeitar H_{01} e H_{02} isto é, existem diferenças entre as concentrações e entre os reatores, não existindo uma interação entre os dois fatores. A figura mostra que os resíduos são aproximadamente normais.

FALTA FIGURA Figura 5.18 - Gráfico de probabilidade normal dos resíduos.

5.5.2 Planeamento 2^3

O planeamento 2^3 permite estudar o efeito de três Fatores A, B e C, cada um a dois níveis. A análise é muito semelhante ao planeamento 2^2 . Neste caso existem oito tratamentos correspondentes às combinações possíveis dos níveis dos três fatores. Assumindo que cada fator é representado no segundo nível pelo sinal (+) e no primeiro nível pelo sinal (-), numa forma tabular o contributo de cada fator pode ser avaliado da seguinte forma

	Média	A	B	AB	C	AC	BC	ABC
(1)	+	-	-	+	-	+	+	-
a	+	+	-	-	-	-	+	+
b	+	-	+	-	-	+	-	+
ab	+	+	+	+	-	-	-	-
c	+	-	-	+	+	-	-	+
ac	+	+	-	-	+	+	-	-
bc	+	-	+	-	+	-	+	-
abc	+	+	+	+	+	+	+	+
	1/8	1/4	1/4	1/4	1/4	1/4	1/4	1/4

O significado de cada uma das letras é semelhante ao do planeamento 2^2. As colunas desta matriz, tal como no caso anterior, são também ortogonais. As interações são determinadas pelos produtos das colunas respetivas à interação. A Figura 5.2 mostra uma representação geométrica deste planeamento.

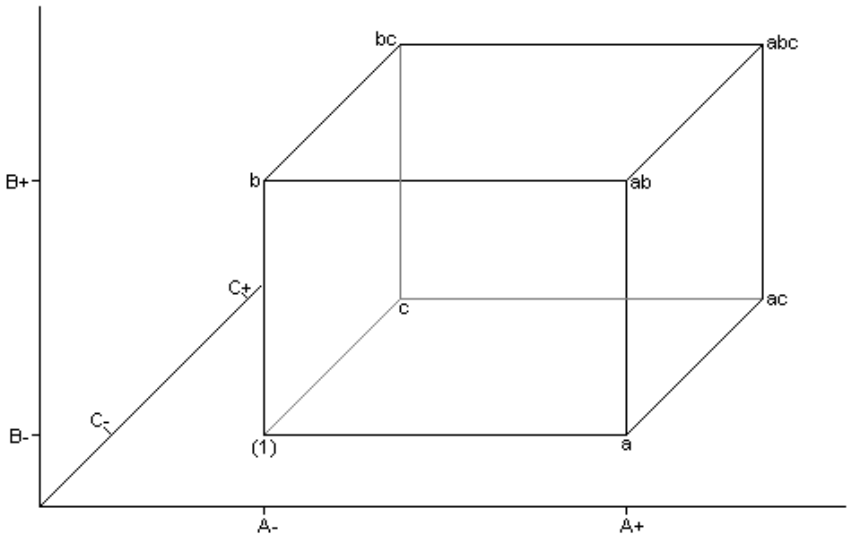


Figura 5.2: Planeamento 2^3.

O efeito de cada fator pode ser estimado como a média de quatro diferenças, correspondentes a cada uma das arestas do paralelogramo da Figura5.2. Portanto, os efeitos para cada fator são dados por,

- Efeito de A,

$$e_A = \frac{1}{4}\left[abc - bc + ac - c + ab - b + a - (1)\right]$$

- Efeito de B,

$$e_B = \frac{1}{4}\left[ab + bc - ac - c + ab + b - a - (1)\right]$$

- Efeito de C,

$$e_C = \frac{1}{4}\left[ab + bc + ac + c - ab - b - a - (1)\right]$$

- Efeito de AB,

$$e_{AB} = \frac{1}{4}\left[ab - bc - ac + c + ab - b - a - (1)\right]$$

- Efeito de AC,

$$e_{AC} = \frac{1}{4}\left[ab - bc + ac - c - ab + b - a - (1)\right]$$

- Efeito de BC,

$$e_{BC} = \frac{1}{4}\left[ab + bc - ac - c - ab - b + a + (1)\right]$$

- Efeito de ABC,

$$e_{ABC} = \frac{1}{4}\left[ab - bc - ac + c - ab + b + a - (1)\right].$$

A Soma Total dos Quadrados, a Soma dos Quadrados dos Tratamentos e a Soma dos Quadrados dos Resíduos podem ser estimadas a partir dos efeitos calculados. Assim, para uma experiência com n replicações, a tabela ANOVA correspondente seria,

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	F
Fator A	SQF_A	1	MQF_A	$F_1 = MQF_A/MQR$
Fator B	SQF_B	1	MQF_B	$F_2 = MQF_B/MQR$
Fator C	SQF_C	1	MQF_C	$F_3 = MQF_C/MQR$
Interação AB	SQI_{AB}	1	MQI_{AB}	$F_4 = MQF_IAB/MQR$
Interação AC	SQI_{AC}	1	MQI_{AC}	$F_5 = MQF_IAV/MQR$
Interação BC	SQI_{BC}	1	MQI_{BC}	$F_6 = MQF_IBC/MQR$
Interação ABC	SQI_{ABC}	1	MQI_{ABC}	$F_7 = MQF_IABC/MQR$
Resíduos	SQR	$8(n-1)$	MQR	
Total	STQ	$8n-1$		

A Soma Total dos Quadrados (STQ) é obtida pela mesma relação para o planeamento fatorial, isto é,

$$STQ = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{l=1}^{2}\sum_{k=1}^{n}\left(y_{ijlk} - \bar{y}_{...}\right)^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{l=1}^{2}\sum_{k=1}^{n} y_{ijlk}^2 - 8n\bar{y}_{....}^2.$$

As expressões para as restantes somas são,

- Soma dos Quadrados do Fator A (SQF_A),

$$SQF_A = \frac{e_A^2}{n},$$

Soma dos Quadrados do Fator B (SQF_B)

$$SQF_B = \frac{e_B^2}{n},$$

- Soma dos Quadrados do Fator C (SQF_C)

$$SQF_C = \frac{e_C^2}{n},$$

- Soma dos Quadrados da Interação AB (SQI_{AB})

$$SQI_{AB} = \frac{e_{AB}^2}{n},$$

- Soma dos Quadrados da Interação AC (SQI_{AC})

$$SQI_{AC} = \frac{e_{AC}^2}{n},$$

- Soma dos Quadrados da Interação BC (SQI_{BC})

$$SQI_{BC} = \frac{e_{BC}^2}{n},$$

- Soma dos Quadrados da Interação AB (SQI_{AB})

$$SQI_{ABC} = \frac{e_{ABC}^2}{n},$$

- Soma dos Quadrados dos Resíduos (SQR),

$$STR = SQT - SQF_A - SQF_B - SQF_C - SQI_{AB} - SQI_{AC} - SQI_{BC} - SQI_{ABC}.$$

Exercícios

1. A tabela apresenta o número de crises que doentes com nevralgia do trigémeo. Estes doentes estão divididos em 3 grupos experimentais: doentes tratados com Ropivicaína (ROP), doentes tratados com Gabapentina (GBP) e doentes tratados com Gabenpentina e Ropivacaína (GBP+ROP). Pretende-se verificar se existem diferenças entre os tratamentos. Os dados apresentam o número de crises no início do tratamento e ao fim e um mês. O que pode concluir?

Início	
GBP	13; 11; 6; 11; 10; 8; 11; 11; 9; 12; 11; 13
GBP+ROP	12; 9; 12; 11; 8; 9; 10; 8; 11; 8; 9; 11
ROP	11; 9; 7; 10; 9
Mês	
GBP	14; 3; 2; 4; 5; 5; 5; 5; 6; 7; 6; 5
GBP+ROP	3; 3, 2; 1, 2, 3; 2, 2; 2; 3: 0; 1
ROP	7; 7; 6; 8; 6

L. Lemos et al.(2008). Gabapentin supplemented with ropivacain block of trigger points improves pain control and quality of life in trigeminal neuralgia patients when compared with gabapentin alone, Clinical Journal of Pain, 24(1), 64-75

2. A tabela apresenta os tempos que ratinhos de laboratório aguentaram no aparelho ROTA-ROD. Os ratinhos são de quatro estirpes genéticas e foram sujeitos à experiência ao fim de 16 semanas de crescimento. Verifique se existem diferenças entre os tempos médios de cada estirpe.

HSF1 +/+; MJD -/-	5.5; 2.3; 9.5; 14.4; 6.1; 9.4; 2.6; 29.9; 2.1; 3.8
HSF1 +/+; MJD -/-	6.7; 8.2; 2.8; 7.3; 11.7; 9.0; 7.9; 28.1; 5.6; 13.9; 11.7
HSF1 +/+; MJD +/-	2.0; 3.4; 23.4; 10.2; 17.6; 2.7; 4.0; 10.1; 2.5; 1.6; 2.0
HSF1 +/-; MJD +/-	5.7; 7.0; 12.9; 3.5; 6.7; 7.2; 10.2; 3.3; 3.7; 6.1; 1.8

3. Os dados da Pressão Sistólica (mm Hg) de uma consulta hospitalar onde, com base numa amostra que contém 433 mulheres, agrupadas por idade (<=50, 51-70, >=71), se apresentam a dimensão, a média e o desvio padrão de cada grupo etário.

Idade	n	$\bar{y}$	s
<= 50	120	145.33	21.841
51 - 70	228	151.80	22.584
>= 71	85	152.09	21.270
Total	433	150.06	22.271

Com base nestes dados, construa a tabela ANOVA correspondente. O que pode concluir?

4. A tabela apresenta a ligação galactose para três grupos: doentes de Crhon, doentes com colite ulcerosa e controlos.

Chron	1343; 1393; 1420; 1641; 1897; 2160; 2169; 2279; 2890
Colite	1264; 1314; 1399; 1605; 2385; 2511; 2514; 2767; 2827; 2895; 3011; 3013; 3355
Controlos	1809; 1926; 2283; 2384; 2447; 2479; 2495; 2525; 2541; 2769 2850; 2964; 2973; 3171; 3257; 3271; 3288; 3358; 3643; 3657

Verifique se há diferenças entre os três grupos.

Altman, D.G., Bland, J.M. (1996). Statistical Notes: Comparing sevral groups using Analysis of Variance, BMJ, 312;1472-1473

5. O controlo anti dopagem pretende testar 3 laboratórios para efeitos de certificação e comparação. Para esse efeito, quatro cobaias são injetadas com com a mesma concentração de um medicamento. De cada cobaia foram retiradas amostras de urina que foram enviadas para 3 laboratórios diferentes.

Cobaia	Laboratório	Conc. (mg/L)
1	1	20.03
1	2	18.24
1	3	23.03
2	1	19.47
2	2	19.97
2	3	22.73
3	1	19.57
3	2	19.36
3	3	21.39
4	1	18.04
4	2	19.99
4	3	22.12

O que pode concluir quanto à avaliação da concentração na urina (mg/L) e diferenças entre os laboratórios?

6. O rendimento (g) de uma reação depende de 3 fatores, a saber, temperatura, concentração de reagente e velocidade de rotação da pá do reator. Para tentar compreender o efeito destes três fatores no rendimento da reação, foi desenhada uma experiência em que cada um destes três fatores podia tomar dois níveis: temperatura (200ºC e 220ºC), concentração (75g/L e 90g/L) e velocidade (30 rpm e 60 rpm). A tabela apresenta os dados para as diversas combinações de níveis dos três fatores para três replicações.

Temp	Conc	Rotação	Rendimento
200	75	30	33; 52; 34
220	75	30	39; 47; 46
200	90	30	66; 49; 47
220	90	30	68; 56; 34
200	75	60	55; 60; 64
220	75	60	68; 64; 65
200	90	60	44; 53; 62
220	90	60	59; 55; 65

O que pode concluir?

7. Uma investigação conduzida por Eysenck (1974) pretendeu estudar a memorização como função do nível de processamento. Para tal, 50 indivíduos, entre 55 e 65 anos, foram aleatoriamente distribuídos por cinco grupos ? quatro grupos de aprendizagem fortuita e um de aprendizagem intencional. O grupo de contagem leu uma lista de palavras e contou o número de letras em cada palavra. O grupo de rima procurou palavras que rimassem com cada palavrada lista. O grupo de adjetivo processou as palavras até encontrar um adjetivo que pudesse ser usado para modificar cada palavra na lista. O grupo de imagem foi instruído para construir imagens para cada palavra. A nenhum destes grupos foi dito que mais tarde lhes seria pedido para lembrarem as palavras. Finalmente, ao grupo intencional foi pedido para memorizar as palavras. Depois de lerem a lista de 27 palavras por três vezes, foi pedido aos sujeitos para escreverem todas as palavras que conseguissem lembrar. Se a aprendizagem depende do nível de processamento, deve haver diferenças entre os diversos grupo. Os resultados obtidos são apresentados na tabela. O que pode concluir? Memorização ? 55 a 65 anos

Contagem	Rima	Adjetivo	Imagem	Intencional
6	12	15	7	14
10	6	10	13	8
5	6	10	19	12
6	7	12	15	20
5	8	10	7	8
5	7	9	5	11
7	9	14	13	11
9	8	10	5	14
6	7	7	12	11
5	6	10	18	15

Eysenck, M.D. (1974). Age Differences in Incidental Learning, Developmental Psychology, Vol 10, No. 6, 936-941

8. O estudo de Eysenck pretendia também verificar como a idade poderia afetar a memorização. Para tal, foram selecionados aleatoriamente mais 50 indivíduos na faixa etária 18 a 30 anos. Os dados relativos a este grupo são apresentados na tabela seguinte.

Contagem	Rima	Adjetivo	Imagem	Intencional
10	5	12	17	21
5	6	17	20	20
5	7	14	17	22
5	9	18	18	18
6	12	11	18	18
3	6	14	15	20
8	7	18	16	20
6	9	16	18	18
7	7	18	18	20
8	8	13	16	17

Verifique se existem diferenças devidas ao grupo etário.

Soluções

1. 1.A

```
              Df Sum Sq Mean Sq F value Pr(>F)
crises$TREAT  2   6.53   3.267   1.069  0.358
Residuals    26  79.47   3.056
```

Não rejeitar H_0

```
crises = read.table("CRISES.csv", header=T)
crises = read.csv("CRISES.csv", header=T)
names(crises)[1]="C_0"
crises$TREAT=as.factor(crises$TREAT)
levels(crises$TREAT)=c("GBP","GBP+ROP","ROP")
oneway.test(crises$C_0~crises$TREAT, var.equal=T)
boxplot(crises$C_0~crises$TREAT)
stripchart(crises$C_0~crises$TREAT,cex=0.5, vertical=T, pch=1)
fligner.test(crises$C_0~crises$TREAT)# teste homogeneidade das var.
```

```
bartlett.test(crises$C_0,crises$TREAT) # teste homogeneidade das var.
oneway.test(crises$C_0~crises$TREAT, var.equal=T)
anova(lm(crises$C_0~crises$TREAT))
model=aov(crises$C_0~crises$TREAT)
summary(model)
TukeyHSD(aov(crises$C_0~crises$TREAT))# compara??es m?ltiplas
plot(model,1:2)
```

1.B

```
             Df Sum Sq Mean Sq F value   Pr(>F)
crises$TREAT  2  93.92   46.96   36.94 2.52e-08 ***
Residuals    26  33.05    1.27
```

Rejeitar H_0

```
             diff        lwr       upr     p adj
GBP+ROP-GBP -2.75 -3.8937504 -1.606250 0.0000077
ROP-GBP      2.05  0.5587319  3.541268 0.0057522
ROP-GBP+ROP  4.80  3.3087319  6.291268 0.0000001
```

```
crises = read.table("CRISES.csv", header=T)
crises = read.csv("CRISES.csv", header=T)
names(crises)[1]="C_0"
crises$TREAT=as.factor(crises$TREAT)
levels(crises$TREAT)=c("GBP","GBP+ROP","ROP")
oneway.test(crises$C_1~crises$TREAT, var.equal=T)
boxplot(crises$C_1~crises$TREAT)
stripchart(crises$C_1~crises$TREAT,cex=0.5, vertical=T, pch=1)
fligner.test(crises$C_1~crises$TREAT) # teste homogeneidade das var.
bartlett.test(crises$C_1,crises$TREAT) # teste homogeneidade das var.
oneway.test(crises$C_1~crises$TREAT, var.equal=T)
anova(lm(crises$C_1~crises$TREAT))
model=aov(crises$C_1~crises$TREAT)
summary(model)
TukeyHSD(aov(crises$C_1~crises$TREAT))# compara??es m?ltiplas
plot(model,1:2)
```

2.

```
             Df Sum Sq Mean Sq F value Pr(>F)
rota$grupo    3    102   33.99   0.763  0.521
Residuals    39   1736   44.52
```

Não rejeitar H_0

```
rota = read.csv("rota_rod.csv", header=T)
names(rota)[1]="rota"
rota$grupo=as.factor(rota$grupo)
levels(rota$grupo)=c("HSF1 +/+; MJD -/-","HSF1 +/-; MJD -/-",
"HSF1 +/-; MJD +/+","HSF1 +/+; MJD +/+")
oneway.test(rota$rota~rota$grupo, var.equal=T)
boxplot(rota$rota~rota$grupo)
stripchart(rota$rota~rota$grupo,cex=0.5, vertical=T, pch=1)
fligner.test(rota$rota~rota$grupo) # teste homogeneidade das var.
bartlett.test(rota$rota~rota$grupo) # teste homogeneidade das var.
oneway.test(rota$rota~rota$grupo, var.equal=T)
anova(lm(rota$rota~rota$grupo))
model=aov(rota$rota~rota$grupo)
summary(model)
TukeyHSD(aov(rota$rota~rota$grupo))# compara??es m?ltiplas
plot(model,1:2)
```

3. $SQT = 120(145.33 - 150.06)^2 + 228(151.8 - 150.06)^2 + 85(152.08 - 150.06)^2 = 4387.8$

$SQR = 119(21.481)^2 + 227(22.584)^2 + 84(21.270)^2 = 208686.6$

```
Trat 4387.8 2 2193.9 4.52
Res 208686.6 430 485.3
Total 213074.4 432
```

$F(0.05, 2, 430) = 3.017$

Rejeitar H_0

4.

```
             Df   Sum Sq Mean Sq F value  Pr(>F)
gala$grupo    2  5174310 2587155   7.341 0.00197 **
Residuals    39 13743776  352405
```

Rejeitar H0

```
                    diff        lwr      upr     p adj
Colite-Crohn   463.6239 -163.52587 1090.774 0.1825546
Control-Crohn  894.2778  313.75986 1474.796 0.0016137
Control-Colite 430.6538  -84.60229  945.910 0.1169907
```

Doentes de Crohn com valor médio significativamente diferente do valor médio dos controlos.

```
gala = read.csv("gala.csv", header=T)
names(gala)[1]="grupo"
gala
gala$grupo=as.factor(gala$grupo)
levels(gala$grupo)=c("Crohn","Colite","Control")
boxplot(gala$liga~gala$grupo, data=gala)
model_1=aov(lm(gala$liga~gala$grupo))
summary(model_1)
TukeyHSD(model_1)
plot(model_1,1:2)
```

5.

```
             Df Sum Sq Mean Sq F value  Pr(>F)
dopa$cob      3  0.881   0.294   0.357 0.78648
dopa$lab      2 23.766  11.883  14.444 0.00509 **
Residuals     6  4.936   0.823
```

Rejeitar H0, relativa aos laboratórios

```
          diff        lwr      upr     p adj
L_2-L_1 0.1125 -1.8553827 2.080383 0.9832283
L_3-L_1 3.0400  1.0721173 5.007883 0.0076189
L_3-L_2 2.9275  0.9596173 4.895383 0.0091213
```

Diferença estatisticamente significativa de L_3 para L_1 e L_2

```
dopa = read.csv("dopagem.csv", header=T)
names(dopa)[1]="cob"
dopa
dopa$cob=as.factor(dopa$cob)
dopa$lab=as.factor(dopa$lab)
levels(dopa$cob)=c("C_1","C_2","C_3","C_4")
levels(dopa$lab)=c("L_1","L_2","L_3")
```

```
boxplot(dopa$conc~dopa$cob)
boxplot(dopa$conc~dopa$lab)
fligner.test(dopa$conc~dopa$cob)
fligner.test(dopa$conc~dopa$lab)
anova(lm(dopa$conc~dopa$cob+dopa$lab))
model=aov(lm(dopa$conc~dopa$cob+dopa$lab))
summary(model)
TukeyHSD(model)
plot(model,1:2)
```

6.

```
                              Df Sum Sq Mean Sq F value  Pr(>F)
rend$temp                      1   92.0    92.0   1.098 0.31030
rend$conc                      1   40.0    40.0   0.478 0.49940
rend$rot                       1  852.0   852.0  10.164 0.00572 **
rend$temp:rend$conc            1    9.4     9.4   0.112 0.74241
rend$temp:rend$rot             1   35.0    35.0   0.418 0.52711
rend$conc:rend$rot             1  477.0   477.0   5.690 0.02977 *
rend$temp:rend$conc:rend$rot   1   15.0    15.0   0.179 0.67751
Residuals                     16 1341.3    83.8
```

Interação significativa entre concentração e rotação e, portanto, ambos os efeitos são importantes. Por essa razão, o efeito da temperatura não influencia o resultado da experiência.

```
rend = read.csv("rend.csv", header=T)
names(rend)[1]="temp"
rend
rend$temp=as.factor(rend$temp)
rend$conc=as.factor(rend$conc)
rend$rot=as.factor(rend$rot)
boxplot(rend$rend~rend$temp)
boxplot(rend$rend~rend$conc)
boxplot(rend$rend~rend$rot)
model=aov(lm(rend$rend~rend$temp+rend$conc+rend$rot+rend$temp:rend$conc+
               rend$temp:rend$rot+rend$conc:rend$rot+rend$temp:rend$conc:rend$rot))
summary(model)
TukeyHSD(model)
plot(model,1:2)
model_1=aov(lm(rend$rend~rend$temp*rend$conc*rend$rot))
```

```
summary(model_1)
TukeyHSD(model_1)
plot(model_1,1:2)
```

7. Análise de variância com um só factor (grupo), faixa 55-65 anos

```
               Df Sum Sq Mean Sq F value   Pr(>F)
memo_old$grupo  4  264.8    66.2   6.424 0.000354 ***
Residuals      45  463.7    10.3
```

Rejeitar H_0

```
           diff        lwr       upr     p adj
Rima-Cont.  1.2 -2.8791277  5.279128 0.9179748
Adj.-Cont.  4.3  0.2208723  8.379128 0.0342591
Imag-Cont.  5.0  0.9208723  9.079128 0.0093720
Int-Cont.   6.0  1.9208723 10.079128 0.0012004
Adj.-Rima   3.1 -0.9791277  7.179128 0.2138377
Imag-Rima   3.8 -0.2791277  7.879128 0.0786793
Int-Rima    4.8  0.7208723  8.879128 0.0137629
Imag-Adj.   0.7 -3.3791277  4.779128 0.9881280
Int-Adj.    1.7 -2.3791277  5.779128 0.7602482
Int-Imag    1.0 -3.0791277  5.079128 0.9561942
```

8. Análise de variância com um só factor (grupo), faixa 18-30 anos

```
               Df Sum Sq Mean Sq F value Pr(>F)
memo_new$grupo  4 1378.1   344.5   88.14 <2e-16 ***
Residuals      45  175.9     3.9
```

Rejeitar H_0

```
           diff        lwr       upr     p adj
Rima-Cont.  1.3 -1.2123589  3.812359 0.5866065
Adj.-Cont.  8.8  6.2876411 11.312359 0.0000000
Imag-Cont. 11.0  8.4876411 13.512359 0.0000000
Int-Cont.  13.1 10.5876411 15.612359 0.0000000
Adj.-Rima   7.5  4.9876411 10.012359 0.0000000
```

```
Imag-Rima   9.7  7.1876411 12.212359 0.0000000
Int-Rima   11.8  9.2876411 14.312359 0.0000000
Imag-Adj.   2.2 -0.3123589  4.712359 0.1114716
Int-Adj.    4.3  1.7876411  6.812359 0.0001362
Int-Imag    2.1 -0.4123589  4.612359 0.1410193
```

9. Análise da variância com dois factores, grupo e faixa

```
                            Df Sum Sq Mean Sq F value   Pr(>F)
memo$grupo                   4 1422.9   355.7  50.054  < 2e-16 ***
memo$faixa                   1  295.8   295.8  41.629 5.42e-09 ***
memo$grupo:memo$faixa        4  220.1    55.0   7.741 2.08e-05 ***
Residuals                   90  639.6     7.1
```

Há uma interação significativa entre faixa e grupo pelo que ambos os factores, em conjunto, são importantes na explicação dos resultados. O efeito de um factor depende do nível do outro, ou seja, a memorização depende da faixa etária.

```
memo = read.csv("memo.csv", header=T)
names(memo)[1]="pal"
memo
memo$grupo=as.factor(memo$grupo)
memo$faixa=as.factor(memo$faixa)
levels(memo$grupo)=c("Cont.","Rima","Adj.","Imag","Int")
levels(memo$faixa)=c("55-65","18-30")
boxplot(memo$pal~memo$grupo)
boxplot(memo$pal~memo$faixa)
boxplot(memo$pal~memo$grupo*memo$faixa)

boxplot(memo$pal~memo$grupo*memo$faixa, names="")
box()
axis(at=2)
line=2
axis(1, at=1:10,labels=F)
nnames<-c("Cont.","Rima","Adj.","Imag","Int", "Cont.","Rima","Adj.","Imag","Int")
mtext(nnames,1,at=1:10,line=rep(c(1,2),5))
mtext(c("55-65","18-30"),1,line=4,at=c(3,8))
```

```
segments(c(1,4),c(-3.5,-3.5),c(4,5),c(-3.5,-3.5),xpd=NA,lwd=2,col="black")
segments(c(6,6),c(-3.5,-3.5),c(6,10),c(-3.5,-3.5),xpd=NA,lwd=2,col="black")

memo_old=memo[which(memo$faixa=="55-65"),] #selecionar casos faixa 55-65
boxplot(memo_old$pal~memo_old$grupo)
model_1=aov(lm(memo_old$pal~memo_old$grupo))
summary(model_1)
TukeyHSD(model_1)
plot(model_1,1:2)

memo_new=memo[which(memo$faixa=="18-30"),] #selecionar casos faixa 18-30
boxplot(memo_new$pal~memo_new$grupo)
model_2=aov(lm(memo_new$pal~memo_new$grupo))
summary(model_2)
TukeyHSD(model_2)
plot(model_2,1:2)

model_3=aov(lm(memo$pal~memo$grupo*memo$faixa))
#análise com dois factores grupo e faixa
summary(model_3)
TukeyHSD(model_3)
plot(model_3,1:2)
```

Capítulo 6

Análise de Dados Categóricos

6.1 Introdução

Em geral, as respostas a inquéritos, resultam em dados classificados em diferentes categorias. Assim, é possível classificar as pessoas de uma determinada região de acordo com as categorias do rendimento económico, ou de acordo com a religião professada. Por vezes, as respostas podem permitir classificar as preferências dos consumidores, ou classificar os produtos como aceitáveis, de segunda qualidade e rejeitados. Nestes exemplos o objetivo é estabelecer inferências acerca das probabilidades desconhecidas de uma experiência multinomial. Os dados de uma experiência desse tipo podem ser apresentados de uma forma tabular, como por exemplo, o número de sapatos com defeito produzidos em 3 linhas de produção:

Linha 1	Linha 2	Linha 3
22	53	19

Definição 6.1.1: Experiência Multinomial

1. A experiência consiste em n tentativas idênticas.
2. Existem k resultados possíveis, ou seja, cada resultado pode ser classificado numa de k classes.
3. A probabilidade de cada resultado permanece a mesma em cada tentativa, em que
$$0 \leq p_i \leq 1 \quad \forall i$$
$$p_1 + p_2 + p_3 + \cdots + p_k = 1.$$
4. As tentativas são independentes.

5. As variáveis aleatórias de interesse são o número de casos, $x_1, x_2, x_3, \cdots, x_k$ em cada das k categorias, em que

$$x_1 + x_2 + x_3 + \cdots + x_k = n.$$

Uma tabela de uma entrada, com k categorias, tem a seguinte forma,

1	2	3	...	k
x_1	x_2	x_3	...	x_k

Em geral, pretende-se testar a hipótese de que as probabilidades de cada célula são iguais a determinados valores especificados na Hipótese Nula. Tal teste poderá ser baseado na diferença entre as proporções observadas e as esperadas. Para testar uma só proporção $H_0 : p = p_0$, a estatística a usar seria

$$z = \frac{x_1/n - p}{\sqrt{p\,(1-p)/n}} = \frac{x_1 - np}{\sqrt{np\,(1-p)}}$$

que segue uma distribuição normal padrão, e em que x_1 é o número de sucessos numa amostra de dimensão n. Por outro lado, z^2 segue uma distribuição de Qui-Quadrado com 1 grau de liberdade. No entanto, tendo em atenção que para $n = x_1 + x_2$ tentativas, das quais x_1 representam sucessos e x_2 insucessos, com probabilidades, respetivamente, de p e $q = (1-p)$, a expressão para z^2 pode ser reescrita da seguinte forma,

$$z^2 = \frac{(x_1 - np)^2}{np} + \frac{(x_2 - nq)^2}{nq}$$

K. Pearson, em 1900, propôs a seguinte generalização desta fórmula,

$$Q = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i}$$

que é uma função do quadrado dos desvios entre os valores observados e os valores esperados $E[x_i] = np_i$, ajustados pelo recíproco do valor esperado. Para grandes valores de n , esta soma segue aproximadamente uma distribuição de Qui-Quadrado, com $k-1$ graus de liberdade. O número de graus de liberdade pode também ser determinado como o número de células menos um grau de liberdade para cada restrição linear imposta nas contagens observadas. No caso de uma tabela de uma entrada existe uma só restrição linear,

$$x_1 + x_2 + x_3 + \cdots + x_k = n.$$

A região de rejeição é definida como num teste unilateral, já que grandes desvios entre as frequências observadas e esperadas conduzem à rejeição da Hipótese Nula e a grandes valores de χ^2; portanto, os valores da cauda superior da distribuição de Qui-Quadrado serão usados para definir a região de rejeição, tal como é descrito na Figura 6.1.

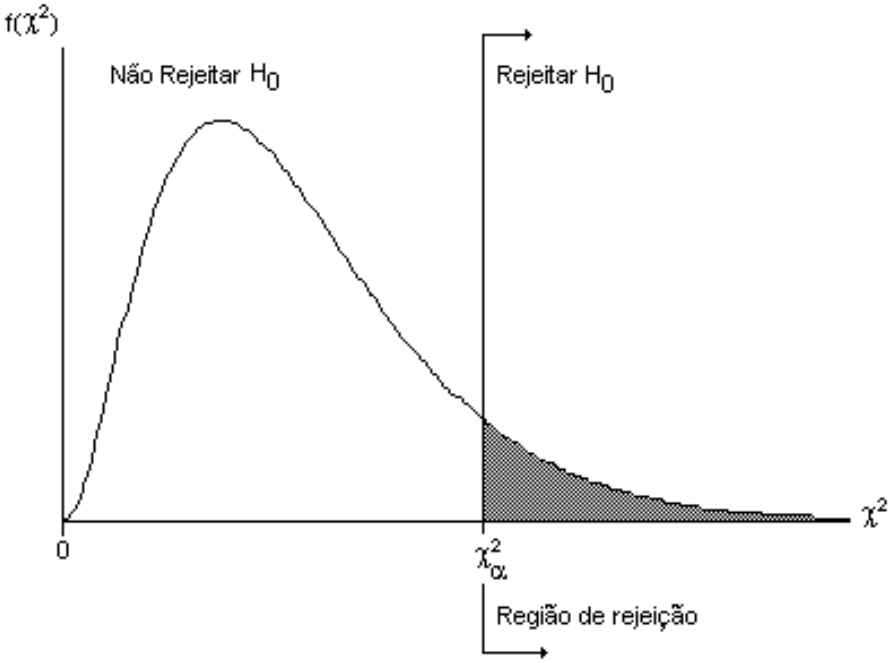


Figura 6.1: Região de rejeição para o teste de Qui-Quadrado.

Definição 6.1.2: Teste de Hipóteses acerca de Probabilidades Multinomiais (Tabela de uma entrada)

- Teste

  $H_0 : p_1 = p_{1,0}; p_2 = p_{2,0}; \cdots ; p_k = p_{k,0}$

  H_1 : Pelo menos um dos valores de probabilidade é diferente do postulado na H_0

- Estatística

$$Q = \sum_{i=1}^{k} \frac{(x_i - np_i)^2}{np_i}$$

- Região de Rejeição

$$Q > \chi^2_\alpha$$

  onde χ^2_α é o valor de χ^2 que localiza uma área α à direita da distribuição de Qui-Quadrado com $n-1$ graus de liberdade.

A aplicação deste teste aconselha que as frequências esperadas, $E[x_i] = np_i$, em cada célula, sejam maiores ou iguais a cinco, $E[x_i] = np_i \geq 5$.

Exemplo 6.1.1: Tabela de uma entrada

Suponha que pretende testar a Hipótese Nula de que as proporções de acidentes mortais por afogamento são iguais em todas as estações, isto é, $p_1 = p_2 = p_3 = p_4 = 0.25$, contra a Hipótese Alternativa de que pelo menos duas proporções são diferentes.
Entre 2007 e 2013, o número de vítimas mortais de acidentes com ondas por estação do ano foi o seguinte (Jornal Público, 22/12/2013):

Primavera	Verão	Outono	Inverno	Total
7	5	11	17	40

Solução

1. Formulação das hipóteses

   $H_0 : p_1 = p_2 = p_3 = p_4 = 0.25$

   H_1 : Pelo menos um dos valores de probabilidade é diferente do postulado na H_0

2. Região crítica

$$\chi^2 \geq \chi^2_{0.05,3} = 7.81$$

3. Teste estatístico

$$Q = \sum_{i=i}^{k} \frac{(x_i - np_i)^2}{np_i} = \frac{(7-10)^2}{10} + \frac{(5-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(17-10)^2}{10} = 8.4$$

4. Decisão

   Rejeitar a Hipótese Nula, ou seja, a proporção de mortes não é igual ao longo do ano.

6.2 Tabelas de Contingência - Teste de Independência

Por vezes, ao contrário do exemplo anterior, pode ser de interesse classificar os dados em função de duas variáveis qualitativas. O objetivo é determinar se as duas direções de classificação são dependentes. A análise destas tabelas, referidas como tabelas de contingência, é baseada no critério de Qui-Quadrado.

Exemplo 6.2.1: Teste de independência

A tabela apresenta o número de mortes para doenças cerebrovasculares por sexo e faixa etária, para o ano de 2015 (dados do Instituto Nacional de Estatística).

Faixa etária	35-44	45-54	55-64	65-74	Total
Homens	45	126	328	764	1263
Mulheres	18	70	137	560	785
Total	63	196	465	1324	2048

Generalizando, uma tabela de contingência $(l \times c)$ tem a seguinte forma, com c colunas representando as diferentes categorias de uma variável, $A_1, A_2, \ldots, A_c$, e l linhas representando as diferentes categorias da outra variável, $B_1, B_2, \ldots, B_l$. As contagens em cada célula, ou seja, as frequências observadas são representadas por f_{ij}, e os totais de linha e coluna, respetivamente, $f_{i\bullet} = \sum_{j=1}^{c} f_{ij}$ e $f_{\bullet j} = \sum_{i=1}^{l} f_{ij}$, são denominados como frequências marginais; $f_{\bullet\bullet}$ é a soma das frequências de todas as células.

	A_1	A_2	$\ldots$	A_c	
B_1	f_{11}	f_{12}	$\ldots$	f_{1c}	$f_{1\bullet}$
B_2	f_{21}	f_{22}	$\ldots$	f_{2c}	$f_{2\bullet}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
B_l	f_{l1}	f_{l2}	$\ldots$	f_{lc}	$f_{l\bullet}$
	$f_{\bullet 1}$	$f_{\bullet 2}$	$\ldots$	$f_{\bullet c}$	$f_{\bullet\bullet}$

O objetivo é determinar se as duas variáveis são independentes, isto é, se a faixa etária é independente do sexo. Seja p_{ij} a probabilidade de que uma resposta seja classificada na célula ij, $p_{i\bullet}$, a probabilidade de uma resposta na linha i, e, $p_{\bullet j}$, a probabilidade na coluna j.

	35-44	45-54	55-64	65-74	Total
Homens	f_{11}	f_{12}	f_{13}	f_{14}	$f_{1\bullet}$
Mulheres	f_{21}	f_{22}	f_{23}	f_{24}	$f_{2\bullet}$
Total	$f_{\bullet 1}$	$f_{\bullet 2}$	$f_{\bullet 3}$	$f_{\bullet 4}$	$f_{\bullet\bullet}$

	35-44	45-54	55-64	65-74	Total
Homens	p_{11}	p_{12}	p_{13}	p_{14}	$p_{1\bullet}$
Mulheres	p_{21}	p_{22}	p_{23}	p_{24}	$p_{2\bullet}$
Total	$p_{\bullet 1}$	$p_{\bullet 2}$	$p_{\bullet 3}$	$p_{\bullet 4}$	$p_{\bullet\bullet}$

Assim, a Hipótese Nula especifica que

$$p_{ij} = (p_{i\bullet})\,(p_{\bullet j})\,,$$

isto é, se as duas classificações são independentes, a probabilidade de uma célula é igual ao produto das probabilidades marginais. A Hipótese Alternativa especifica que as classificações não

são independentes e, portanto, as probabilidades de cada célula não são iguais ao produto das probabilidades marginais.

As melhores estimativas das probabilidades marginais são dadas por,

$$\hat{p}_{i\bullet} = \frac{f_{i\bullet}}{f_{\bullet\bullet}} \qquad \hat{p}_{\bullet j} = \frac{f_{\bullet j}}{f_{\bullet\bullet}}.$$

Assim, a frequência esperada para cada célula ij,

$$e_{ij} = f_{\bullet\bullet}\left(p_{i\bullet}\right)\left(p_{\bullet j}\right) = f_{\bullet\bullet}\frac{f_{i\bullet}}{f_{\bullet\bullet}}\frac{f_{\bullet j}}{f_{\bullet\bullet}} = \frac{\left(f_{i\bullet}\right)\left(f_{\bullet j}\right)}{f_{\bullet\bullet}}$$

e a estatística de teste será,

$$Q = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{\left(f_{ij} - e_{ij}\right)^2}{e_{ij}}$$

O número de graus de liberdade é igual ao número de células (lc) menos o número de restrições lineares independentes impostas sobre as frequências observadas. As restrições impostas são as seguintes:

1. O número total de células é (lc), e a soma das frequências observadas tem que ser igual a $f_{\bullet\bullet}$,

   $$f_{11} + f_{12} + f_{13} + \cdots + f_{lc} = f_{\bullet\bullet}$$

   pelo que o número de graus de liberdade será $lc - 1$.

2. As frequências observadas serão usadas para estimar as probabilidades marginais,

   $$p_{\bullet 1} + p_{\bullet 2} + p_{\bullet 3} + \cdots + p_{\bullet c} = 1 \qquad p_{1\bullet} + p_{2\bullet} + p_{3\bullet} + \cdots + p_{l\bullet} = 1$$

   conduzindo a uma perda, para cada estimativa das probabilidades marginais, de um grau de liberdade; contudo, como por outro lado, a soma das probabilidades marginais tem que ser igual à unidade, o número total de graus de liberdade perdidos será, respetivamente, $c-1$ e $l-1$.

Portanto, o número total de graus de liberdade será dado por,

$$lc - 1 - (c-1) - (l-1) = (l-1)\,(c-1)\,,$$

ou seja, o número de graus de liberdade associado ao teste de Qui-Quadrado, numa tabela de contingência com l linhas e c colunas é igual a $(l-1)(c-1)$.

Exemplo 6.2.2: Teste de independência

No exemplo anterior, o número de graus de liberdade seria de $(2-1)4-1) = 3$ graus de liberdade. A tabela das frequências esperadas permite indiciar a rejeição da Hipótese Nula,

Faixa etária	35-44	45-54	55-64	65-74
Homens	38.85	120.87	286.77	816.51
Mulheres	24.15	75.13	178.23	507.49

a que corresponde um valor da estatística $Q = 27.38$.

Definição 6.2.1: Tabela de Contingência - Teste de Independência

- Teste

$$H_0 : p_{ij} = p_{i\bullet}p_{\bullet j} \quad \text{(As duas classificações são independentes)}$$
$$H_1 : p_{ij} \neq p_{i\bullet}p_{\bullet j} \quad \text{(As duas classificações são dependentes)}$$

- Estatística

$$Q = \sum_{j=1}^{c}\sum_{i=1}^{l}\frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

- Região de Rejeição

$$Q > \chi^2_\alpha$$

onde χ^2_α é o valor de χ^2 que localiza uma área α à direita da distribuição de Qui-Quadrado com $(l-1)(c-1)$ graus de liberdade.

As n observações são uma amostra aleatória da população de interesse e, para que a aproximação de Qui-Quadrado seja válida, é aconselhável que as frequências esperadas, em cada célula, sejam $e_{ij} \geq 5$. Existem, contudo, aplicações onde esta exigência é relaxada de forma a que não mais de 20% das células apresentem frequências esperadas $e_{ij} < 5$. Por outro lado, existem testes exatos para tabelas de contingência de (2×2) ou $(n \times 2)$.

Exemplo 6.2.3: Teste de independência

Uma idade gestacional baixa (menos que 37 semanas), em geral, é um indicador de morbilidade e mortalidade infantil. Nesse sentido, uma gravidez que termine antes de 37 semanas é definida como pré-termo. Sabe-se, contudo, que variáveis sócio-económicas, como o rendimento ou a escolaridade, influenciam o peso ao nascer e a idade gestacional. A tabela apresenta o número de bebés de termo e pré termo, em função do nível educacional da mãe, avaliado em classes, conforme o número de anos de escolaridade: 1 - menos de 4 anos; 2 - entre 5 a 6 anos; 3 - entre 7 a 9 anos; 4 - entre 10 e 12 anos; 5 - mais de 12 anos. Um inquérito às parturientes do Hospital Garcia de Orta, entre 1995 e 1998, produziu os seguintes resultados:

Escolaridade	1	2	3	4	5	Total
Pré termo	236	157	256	284	156	1089
Termo	1367	1021	2056	2218	1138	7800
Total	1603	1178	2312	2502	1294	8889

Solução

1. Formulação das hipóteses

$$H_0 : p_{ij} = p_{i\bullet}p_{\bullet j} \quad \text{(As duas classificações são independentes)}$$
$$H_1 : p_{ij} \neq p_{i\bullet}p_{\bullet j} \quad \text{(As duas classificações são dependentes)}$$

2. Região crítica

$$\chi^2 \geq \chi^2_{0.05,4} = 9.48$$

3. Teste estatístico

Tabela de frequências esperadas

	1	2	3	4	5	Total
Pré termo	196.39	144.32	283.25	306.52	158.53	1089
Termo	1406.61	1033.68	2028.75	2195.48	1135.47	7800
Total	1603	1178	2312	2502	1294	8889

$$\begin{aligned} Q &= \sum_{j=1}^{c}\sum_{i=1}^{l}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \\ &= \frac{(236-196.39)^2}{196.39} + \frac{(157-144.32)^2}{144.32} + \cdots + \frac{(1138-1135.47)^2}{1135.47} = 15.30 \end{aligned}$$

4. Decisão

Rejeitar a Hipótese Nula, ou seja, uma gravidez de termo não é independente da educação escolar, assumida neste exemplo como uma aproximação ao bem estar económico da família.

6.3 Tabelas de Contingência - Teste de Homogeneidade

Nos exemplos anteriores, as frequências marginais não estavam fixas *a priori*, constituindo assim variáveis aleatórias. No entanto, existem situações em que pode ser de interesse fixar uma das frequências marginais, como por exemplo, num inquérito em que uma das categorias pode conter um pequeno conjunto de pessoas e, portanto, podendo eventualmente não estar representada na

amostra. Por exemplo, imagine que se pretende estudar a incidência de uma doença genética em Portugal, suspeitando-se que a sua distribuição seja diferente nas diversas regiões do país; um estudo baseado numa amostragem dos bebés nascidos em Portugal, pode eventualmente falhar a inclusão de indivíduos de regiões com muito baixa natalidade como o Alentejo. Assim, uma estratégia possível seria incluir um número fixo de indivíduos por região, por forma poder estabelecer as incidências por região. Situação semelhante poderia ser encontrada numa amostragem baseada em rendimentos, que poderia não recolher um número suficiente de indivíduos nos extremos da distribuição de rendimentos.

De qualquer forma, tal facto não terá qualquer efeito na análise dos dados como se verá adiante. Assim, na situação em que os totais de linha são fixos, os dados representam os resultados de l experiências multinomiais, cada com c células, e o teste de independência entre linhas e colunas é equivalente a testar a igualdade das c probabilidades em todas as l experiências multinomiais. Por esta razão, um teste de deste tipo é por vezes referido como teste de homogeneidade de duas ou mais experiências multinomiais.

Assumindo que as frequências marginais relativas às linhas são fixas, a probabilidade marginal relativa a cada coluna é dada por,

$$\hat{p}_{\bullet j} = \frac{f_{\bullet j}}{n}$$

	A_1	A_2	$\ldots$	A_c	
B_1	f_{11}	f_{12}	$\ldots$	f_{1c}	n_1
B_2	f_{21}	f_{22}	$\ldots$	f_{2c}	n_2
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
B_l	f_{l1}	f_{l2}	$\ldots$	f_{lc}	n_l
	$f_{\bullet 1}$	$f_{\bullet 2}$	$\ldots$	$f_{\bullet c}$	n

O valor esperado em cada célula resulta então do produto desta probabilidade marginal pelo valor fixado da respetiva linha, ou seja,

$$e_{ij} = \hat{p}_{\bullet j} n_i = \frac{f_{j\bullet}}{n} n_i = \frac{f_{j\bullet} n_i}{n}$$

tal como no teste de independência.

Definição 6.3.1: Tabela de Contingência - Teste de Homogeneidade

- Teste

  $H_0 : p_{1j} = p_{2j} = \cdots = p_{lj} \quad$ com $j = 1, \ldots, c$

  (A distribuição das observações em cada coluna é a mesma para cada linha, $p_{ij} = p_{kj} \quad \forall_{i,j}$)

  $H_1 : p_{ij} \neq p_{kj}$

  (A distribuição das observações em cada coluna difere em pelo menos duas das linhas)

- Estatística

$$Q = \sum_{j=1}^{c} \sum_{i=1}^{l} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

- Região de Rejeição

$$Q > \chi^2_\alpha$$

  onde χ^2_α é o valor de χ^2 que localiza uma área α à direita da distribuição de Qui-Quadrado com $(l-1)(c-1)$ graus de liberdade.

As amostras aleatórias de cada população são selecionadas de uma forma independente e, para que a aproximação de Qui-Quadrado seja válida, é aconselhável que as frequências esperadas, em cada célula, sejam $e_{ij} \geq 5$.

Exemplo 6.3.1: Teste de homogeneidade

Suponha que uma empresa de montagem de computadores possui três turnos, operando em períodos de 8 horas. Pretende verificar se o número de computadores com defeito é igual nos três turnos.

Para tal, uma amostra aleatória de 150 computadores montados em cada turno é recolhida.

A tabela apresenta o número de computadores com defeito encontrado.

Verifique se existem diferenças na proporção de computadores com defeito produzidos em cada linha.

Use $\alpha = 0.05$.

	Turno 1 (0-8 horas)	Turno 2 (8-16 horas)	Turno 3 (16-24 horas)	Totais
Com defeito	23	10	11	44
Sem defeito	127	140	139	406
Totais	150	150	150	450

Solução

1. Formulação das hipóteses

   $H_0 : p_{i1} = p_{i2} = p_{i3}$

   (A distribuição das observações em cada coluna é a mesma para cada linha,

   $p_{ij} = p_{kj} \quad \forall_{i,j}$)

   $H_1 : p_{ij} \neq p_{kj}$

   (A distribuição das observações em cada coluna difere em pelo menos duas das linhas)

2. Região crítica

$$Q \geq \chi^2_{0.05,2} = 5.99$$

3. Teste estatístico

   Tabela de frequências esperadas

	Turno 1	Turno 2	Turno 3
Com defeito	14.67	14.67	14.67
Sem defeito	135.33	135.33	135.33

$$Q = \sum_{j=1}^{c} \sum_{i=1}^{l} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = \frac{(23 - 14.67)^2}{14.67} + \frac{(10 - 14.67)^2}{14.67} + \cdots + \frac{(139 - 135.33)^2}{135.33} = 7.91$$

4. Decisão

   Rejeitar a Hipótese Nula, ou seja, a proporção de computadores com defeito produzidos nos três turnos não é igual.

6.4 Testes de Bom Ajuste

Um teste de bom ajuste, ou de aderência, é usado quando o objetivo é determinar se um conjunto de dados é consistente com a hipótese de constituir uma amostra aleatória de uma população com uma dada distribuição. Assim, o exemplo **??** pode ser visto como um teste de bom ajuste a um modelo em que todas as células são igualmente prováveis.

Definição 6.4.1: Teste de Bom Ajuste de Qui-Quadrado (Grandes Amostras)

- Teste

$$H_0 : F(x) = G(x), \quad \text{os dados seguem uma distribuição especificada.}$$
$$H_1 : F(x) \neq G(x), \quad \text{os dados não seguem a distribuição especificada.}$$

- Estatística

$$Q = \sum_{i=1}^{c} \frac{(f_i - e_i)^2}{e_i}$$

- Região de Rejeição

$$Q > \chi^2_\alpha$$

onde χ^2_α é o valor de χ^2 que localiza uma área α à direita da distribuição de Qui-Quadrado com $(c-t-1)$ graus de liberdade, onde c é o número de termos no somatório e t o número de parâmetros estimados.

Exemplo 6.4.1: Teste de Bom Ajuste de Qui-Quadrado

Em geral, admite-se que o número de defeitos observados em rolos de tecido segue uma distribuição de Poisson. A tabela apresenta o número de rolos com defeitos encontrados da produção de um mês.

Nº de defeitos por rolo	0	1	2	3	4	5	6	7	8	9	10
Frequência observada	3	4	8	12	15	13	10	8	4	3	1

Verifique se os dados se ajustam a uma distribuição de Poisson.

Solução

1. Formulação das hipóteses

H_0 : A distribuição dos defeitos segue uma distribuição de Poisson.

H_1 : A distribuição dos defeitos não segue uma distribuição de Poisson.

2. Região crítica

$$\chi^2 \geq \chi^2_{0.05,6} = 12.592$$

3. Teste estatístico

Estimativa da média

$$\hat{\lambda} = \frac{\sum_i f_i x_i}{\sum_i f_i} = \frac{0\,(3) + 1\,(4) + \cdots + 9\,(3) + 10\,(1)}{3 + 4 + \cdots + 3 + 1} = \frac{366}{81} \approx 4.5$$

Nº de defeitos	Frequência observada	Probabilidades	Frequência esperada
0	3	0.0111	0.9
1	4	0.05	4.05
2	8	0.1125	9.11
3	12	0.1687	13.66
4	15	0.1898	15.37
5	13	0.1708	13.83
6	10	0.1281	10.38
7	8	0.0824	6.67
8	4	0.0463	3.75
9	3	0.0232	1.88
10	1	0.1744	1.39

Agrupando as frequências esperadas menores que 5 nos extremos da tabela,

Nº de defeitos	Frequência observada	Frequência esperada	$\frac{(f_i - e_i)^2}{e_i}$
$1 \leq$	7	4.95	0.85
2	8	9.11	0.14
3	12	13.66	0.2
4	15	15.37	0.01
5	13	13.83	0.05
6	10	10.38	0.01
7	8	6.67	0.26
≥ 8	8	7.01	0.14

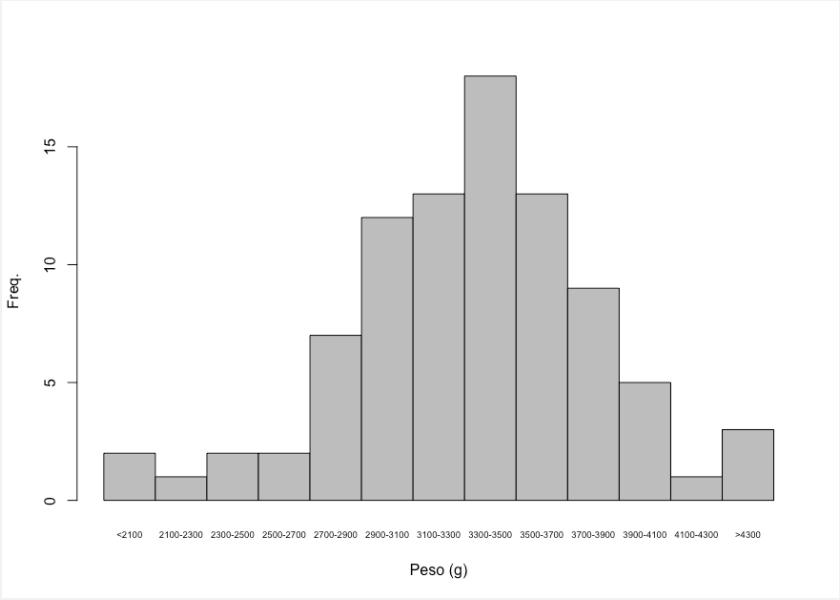
$$Q = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} = \frac{(7 - 4.95)^2}{4.95} + \frac{(8 - 9.11)^2}{9.11} + \cdots + \frac{(8 - 7.01)^2}{7.01} = 1.66$$

4. Decisão

   Não rejeitar a Hipótese Nula, ou seja, a distribuição dos defeitos segue uma distribuição de Poisson.

Exemplo 6.4.2: Teste de Bom Ajuste de Qui-Quadrado

Uma amostra aleatória de 88 bebés, do sexo masculino, nascidos no Hospital Garcia de Orta produziu o seguinte histograma de frequências dos pesos ao nascer:



A tabela apresenta a distribuição dos pesos dos 88 recém nascidos por classes:

Peso (g)	Frequência
$<$2100	2
2100-2300	1
2300-2500	2
2500-2700	2
2700-2900	7
2900-3100	12
3100-3300	13
3300-3500	18
3500-3700	13
3700-3900	9
3900-4100	5
4100-4300	1
$>$4300	3

Verifique se os pesos dos recém-nascidos seguem uma distribuição normal.

Solução

1. Formulação das hipóteses

   H_0 : A distribuição dos pesos dos recém-nascidos é normal.

   H_1 : A distribuição dos pesos dos recém-nascidos não é normal.

2. Região crítica

$$\chi^2 \geq \chi^2_{0.05,7} = 14.067$$

3. Teste estatístico

   Parâmetros estimados: $\bar{x} = 3317.4$ g; $s = 508.77$g.

Peso (g)	Frequência observada	P(a<X<b)	Frequência esperada
<2100	2	0.0084	0.74
2100-2300	1	0.0144	1.27
2300-2500	2	0.0313	2.75
2500-2700	2	0.0584	5.14
2700-2900	7	0.0935	8.23
2900-3100	12	0.1286	11.32
3100-3300	13	0.1518	13.36
3300-3500	18	0.1538	13.54
3500-3700	13	0.1338	11.78
3700-3900	9	0.0999	8.79
3900-4100	5	0.0641	5.64
4100-4300	1	0.0353	3.1
>4300	3	0.0267	2.35

   Atendendo a que existem células, nos extremos da tabela, cujas frequências esperadas são inferiores a 5, a tabela tomará agora a seguinte forma:

Peso(g)	Frequência observada	Frequência esperada	$\frac{(f_i-e_i)^2}{e_i}$
<2700	7	9.9	0.85
2700-2900	7	8,23	0.18
2900-3100	12	11,32	0.04
3100-3300	13	13,36	0.01
3300-3500	18	13,54	1.47
3500-3700	13	11,78	0.13
3700-3900	9	8,79	0.01
3900-4100	5	5,64	0.07
>4100	4	5.45	0.44

$$Q = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} = \frac{(7 - 9.90)^2}{9.90} + \frac{(7 - 8.23)^2}{8.23} + \cdots + \frac{(4 - 5.45)^2}{5.45} = 3.2$$

4. Decisão

   Não rejeitar a Hipótese Nula, ou seja, a distribuição dos pesos dos bebés aos nascer segue uma distribuição normal.

Exercícios

1. A tabela apresenta a classificação de 475 mulheres encarceradas, de acordo com a positividade para o teste de HIV e o consumo anterior de drogas intravenosas.

	Sim	Não	Total
Pos.	61	27	88
Neg.	75	312	387
Total	136	339	475

   Teste a hipótese nula de que não há associação entre a história de consumo de drogas intravenosas e a positividade no teste de HIV.

   Smith,PF, Mikl,J, Truman,BI, Lessner, L, Lehman, JS, Stevens, RW, Lord, EA, Broadlus, RK, Morse, DL (1991). HIV infection Among Women Entering the New York State Correctional System, American Journal of Public Health, 81, 35-40

2. Alguns médicos fornecem aconselhamento genético para prevenirem o nascimento de bebés com doenças genéticas. Contudo, o aconselhamento genético tem encontrado alguma resistência entre médicos e pais. Pretendeu-se determinar se existe uma relação entre a opinião do médico relativamente ao aconselhamento genético e a sua religião. A tabela apresenta as respostas de 287 médicos selecionados aleatoriamente.

Acon. Genético	Hebraica	Protestante	Católica	Total
Sim	21	36	10	67
Não	26	142	52	220
Total	47	178	62	287

   Teste a hipótese nula de que a opinião sobre o aconselhamento genético é independente da religião professada.

   Rose Weitz (2010) Barriers to acceptance of genetic counseling among primary care physicians, Social Biology, 26:3, 189-197, DOI: 10.1080/19485565.1979.9988377

3. Um aviário pretende verificar se duas formulações alimentares conduzem a crescimentos diferenciados em dois bandos de frangos. Esta diferenciação pode ser avaliada pela distribuição em classes de peso. A tabela apresenta a distribuição de pesos para os dois bandos (controlo e experimental). Verifique se existem diferenças nas distribuições dos pesos. Use $\alpha = 0.05$.

	<1000	1000-1200	1200-1400	>1400	Total
Controlo	69	290	285	747	1391
Experimental	368	760	706	1308	3052
Total	447	1050	891	2055	4443

4. Um hospital resolveu avaliar a qualidade dos seus serviços. Para o efeito selecionou 100 utentes de cada um dos seguintes serviços de Pneumologia, Cardiologia e Pediatria. A tabela apresenta as respostas obtidas, numa escala de 1 a 5, relativamente à satisfação com o serviço. O que pode concluir?

Satisfação	Pneumologia	Cardiologia	Pediatria	Total
Discordo totalmente	1	59	14	74
Discordo parcialmente	10	54	31	95
Não concordo, nem discordo	42	25	51	118
Concordo parcialmente	66	6	44	116
Concordo totalmente	31	6	10	47
Total	150	150	150	450

5. A tabela apresenta o número de grávidas em trabalho de parto que chegam a um serviço de urgência de um grande hospital num determinado dia, tendo por base os registos do último ano. Verifique se as chegadas seguem uma distribuição de Poisson.

Nº de Partos	Frequência
0	2
1	13
2	37
3	46
4	61
5	55
6	62
7	41
8	22
9	16
10	7
11	2
12	1

6. A tabela apresenta as frequências observadas de parafusos produzidos por uma máquina em função do comprimento. A máquina está ajustada para produzir parafusos de 50 mm de comprimento. Verifique se os dados do comprimento dos para fusos se ajustam a uma distribuição normal.

Intervalo	Frequência Observada
< 46	3
$[46 - 47[$	6
$[47 - 48[$	12
$[48 - 49[$	12
$[49 - 50[$	24
$[50 - 51[$	16
$[51 - 52[$	11
$[52 - 53[$	10
$[53 - 54[$	5
≥ 54	1

7. A figura mostra as primeiras linhas da informação disponibilizada na Bolsa de Londres na semana de 9 a 13 de abril de 2018.

Exchange Traded Funds

Weekly Statistics - 9 Apr 2018 to 13 Apr 2018

Issuer Name	TIDM	ISIN	Traded Currency	Number of Trades	Trades %	GBP Turnover	Turnover %	Average Trade Size	Time Weighted Spread (bps)
LYX ETF FTSE 100 MH USD ACC TH	100H	LU1650492504	USD	2	0.00%	5 930	0.00%	2 965	18
AMUNDI S&P 500	500U	LU1681049018	USD	3	0.00%	11 518 972	0.33%	3 839 657	12
AMUNDI MSCI EM ASIA	AASU	LU1681044563	USD	5	0.01%	152 939	0.00%	30 588	14
SPDR MSCI ACWI UCITS ETF $	ACWD	IE00B44Z5B48	USD	139	0.19%	3 441 105	0.10%	24 756	10
SPDR MSCI ACWI UCITS ETF	ACWI	IE00B44Z5B48	GBP	9	0.01%	362 825	0.01%	40 314	12
LYXOR ETF MSCI ACWI $	ACWU	FR0011093418	USD	10	0.01%	3 764 306	0.11%	376 431	14
LYXOR MSCI AC ASIA PACIFIC EX JAPAN- ACC	AEJ	FR0010312124	USD	2	0.00%	182 172	0.01%	91 086	16
AMUNDI MSCI EUROPE EX UK	AEXK	LU1681043326	GBX	2	0.00%	1 959	0.00%	979	18

Sabe-se que o valor esperado das frequências de dígitos numa lista de números segue uma distribuição logarítmica, em primeiro lugar referida por Newcomb e, mais tarde, redescoberta por Frank Benford e, por essa razão, ficou conhecida como Lei de Benford da Distribuição dos Dígitos. De acordo com essa lei, a distribuição dos primeiros dígitos é dada por:

$$P(d) = log(1 + \frac{1}{d})$$

sendo d o dígito em causa, com $d = 1, 2, \ldots, 9$.

Primeiro Dígito	Freq.
1	375
2	225
3	147
4	107
5	103
6	67
7	67
8	53
9	61
Total	1205

Verifique se os dados da tabela seguem a Lei de Benford.

Hill, T. P. The First Digit Phenomenon. Amer. Sci. 86, 358-363, 1998.

Raimi, R.A. the peculiar distribution of first digits, 221, 109-120, 1969.

8. A tabela apresenta a listagem dos primeiros dígito dos 278 municípios por população e densidade (Hab./Km2).

Dígito	Pop.	Área
1	79	107
2	45	39
3	28	29
4	22	21
5	34	26
6	19	14
7	29	12
8	9	13
9	13	17
Total	278	278

Verifique que se os dados seguem a Lei de Benford.

Soluções

1.
```
Pearson's Chi-squared test with Yates' continuity correction
data:  data.table1
X-squared = 85.075, df = 1, p-value < 2.2e-16

x1= c(61,27)
y1 = c(75,312)
data.table1 = rbind(x1,y1)
data.table1
dimnames(data.table1)=list("HIV"=c("Pos","Neg"),"Cons"=c("Sim","Nao"))
addmargins(data.table1)
prob1=chisq.test(data.table1)
prob1
round(prop.table(data.table1)*100,2)
round(prop.table(data.table1, 1)*100,2)
round(prop.table(data.table1, 2)*100,2)
prob1$observed
round(prob1$expected,2)
round(prob1$residual,2)
```

2.
```
Pearson's Chi-squared test
data:  data.table2
X-squared = 14.728, df = 2, p-value = 0.0006335
```

```
x2= c(21,36,10)
y2 = c(26,142,52)
data.table2 = rbind(x2,y2)
data.table2
dimnames(data.table2)=list("A_Gen"=c("Sim","Nao"),"Relig"=c("Heb","Prot","Cat"))
addmargins(data.table2)
prob2=chisq.test(data.table2)
prob2
round(prop.table(data.table2)*100,2)
round(prop.table(data.table2, 1)*100,2)
round(prop.table(data.table2, 2)*100,2)
prob2$observed
round(prob2$expected,2)
round(prob2$residual,2)
```

3. Pearson's Chi-squared test

```
data:  data.table3
X-squared = 82.965, df = 3, p-value < 2.2e-16

x3= c(69,290,285,747)
y3 = c(368,760,706,1308)
data.table3 = rbind(x3,y3)
data.table3
dimnames(data.table3)=list("Grupo"=c("Cont","Exp"),"Peso"=c("<1000","1000-2000","1200-1400",
addmargins(data.table3)
prob3=chisq.test(data.table3)
prob3
round(prop.table(data.table3)*100,2)
round(prop.table(data.table3, 1)*100,2)
round(prop.table(data.table3, 2)*100,2)
prob3$observed
round(prob3$expected,2)
round(prob3$residual,2)
```

4. Pearson's Chi-squared test

```
data:  data.table4
X-squared = 185.24, df = 8, p-value < 2.2e-16
```

```
x4= c(1,59,14)
y4 = c(10,54,31)
s4 = c(42,25,51)
r4 = c(66,6,44)
t4 = c(31,6,10)
data.table4 = rbind(x4,y4,s4,r4,t4)
data.table4
dimnames(data.table4)=list("Satisf"=c("DT","DP","NC,ND","CP","CT"),"Ser"=c("Pneumo","Cardio"
addmargins(data.table4)
prob4=chisq.test(data.table4)
prob4
round(prop.table(data.table4)*100,2)
round(prop.table(data.table4, 1)*100,2)
round(prop.table(data.table4, 2)*100,2)
prob4$observed
round(prob4$expected,2)
round(prob4$residual,2)
```

5. Chi-squared test for given probabilities

```
data:  partos$dias
X-squared = 6.5308, df = 12, p-value = 0.887

partos = read.csv("cat_5.csv", header=T)
partos
names(partos)[1]="part"
partos
sum(partos$dias)
media=sum(partos$part*partos$dias)/sum(partos$dias)
media
prob1=dpois(0:11,media)
t=1-sum(dpois(0:11,media))
prob=c(prob1,t)
sum(prob)
esp=sum(partos$dias)*prob
qui=sum((partos$dias-esp)^2/esp)
qui
pchisq(qui,length(partos$dias)-1,lower.tail=F)
```

```
chisq.test(partos$dias,p=prob)
```

6. Chi-squared test for given probabilities

```
data:  paraf$obs
X-squared = 5.9424, df = 9, p-value = 0.7457
```

```
paraf = read.csv("cat_6.csv", header=T)
paraf
names(paraf)[1]="int"
paraf
sum(paraf$obs)
media=50
dp=2
ls=c(47:54)
li=c(46:53)
l1=pnorm(min(ls)-1,media,dp)
s1=pnorm(max(ls),media,dp, lower.tail = F)
prob1=pnorm(ls,media,dp)-pnorm(li,media,dp)
prob=c(l1,prob1,s1)
sum(prob)
esp=sum(paraf$obs)*prob
qui=sum((paraf$obs-esp)^2/esp)
qui
pchisq(qui,length(paraf$int)-1,lower.tail=F)
chisq.test(paraf$obs,p=prob)
```

7. Chi-squared test for given probabilities

```
data:  bolsa$obs
X-squared = 6.9626, df = 8, p-value = 0.5407
```

```
bolsa = read.csv("cat_7.csv", header=T)
bolsa
names(bolsa)[1]="dig"
bolsa
sum(bolsa$obs)
prob=log10(1+1/bolsa$dig)
sum(prob)
```

```
esp=sum(bolsa$obs)*prob
qui=sum((bolsa$obs-esp)^2/esp)
qui
pchisq(qui,length(bolsa$dig)-1,lower.tail=F)
chisq.test(bolsa$obs,p=prob)
```

8.
```
Chi-squared test for given probabilities
data:  municip$pop
X-squared = 21.539, df = 8, p-value = 0.005845

Chi-squared test for given probabilities
data:  municip$area
X-squared = 15.238, df = 8, p-value = 0.05468

municip = read.csv("cat_8.csv", header=T)
municip
names(municip)[1]="dig"
municip
sum(municip$area)
sum(municip$pop)
prob=log10(1+1/municip$dig)
sum(prob)

esp=sum(municip$pop)*prob
qui=sum((municip$pop-esp)^2/esp)
qui
pchisq(qui,length(municip$dig)-1,lower.tail=F)
chisq.test(municip$pop,p=prob)

esp=sum(municip$area)*prob
qui=sum((municip$area-esp)^2/esp)
qui
pchisq(qui,length(municip$dig)-1,lower.tail=F)
chisq.test(municip$area,p=prob)
```

Capítulo 7

Regressão e Correlação

7.1 Introdução

A Regressão Linear aborda o estudo da relação entre uma variável dependente e uma ou mais variáveis independentes. Situações como a relação entre o peso ao nascer e o número de semanas de gestação, o rendimento per capita e a população, consumo de combustível de aquecimento e temperatura ambiente podem ser estudadas através de uma análise de regressão. O modelo de regressão define uma equação que expressa a variável dependente Y como função de uma ou mais variáveis independentes $X_1, X_2, \ldots, X_k$. No caso em que só existe uma só variável independente, o objetivo da análise de regressão é determinar os parâmetros β_0, β_1 que definem a equação

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

designados por coeficientes de regressão. Assim, dado um conjunto de pares de pontos (X_i, Y_i) o ajuste do modelo aos dados será determinado através do método dos mínimos quadrados.

A Figura 7.1 pretende ilustrar um ajuste a um dado conjunto de pares (X_i, Y_i). β_0 representa a intersecção na origem, isto é, o ponto onde a reta intersecta o eixo dos Y, β_1 o declive da reta, isto é, quanto Y variará (acréscimo ou decréscimo) por cada unidade de aumento de X, ε o termo de erro, isto é, para cada valor de X_i, ε_i representa a diferença entre o valor observado Y_i e o valor ajustado $\hat{Y}_i$.

A questão que se coloca é como determinar que uma dada reta constitui um "bom" ajuste a um conjunto de pontos. Uma primeira abordagem será considerar a minimização das distâncias dos pontos observados à reta. Como se está em presença de um conjunto de pontos, a minimização deverá ser estendida à soma dos desvios.

A Figura 7.2 pretende ilustrar que esta não é uma abordagem razoável já que as duas retas da figura apresentam uma soma dos desvios nula e só a reta a cheio passa pelos dois pontos. Assim,

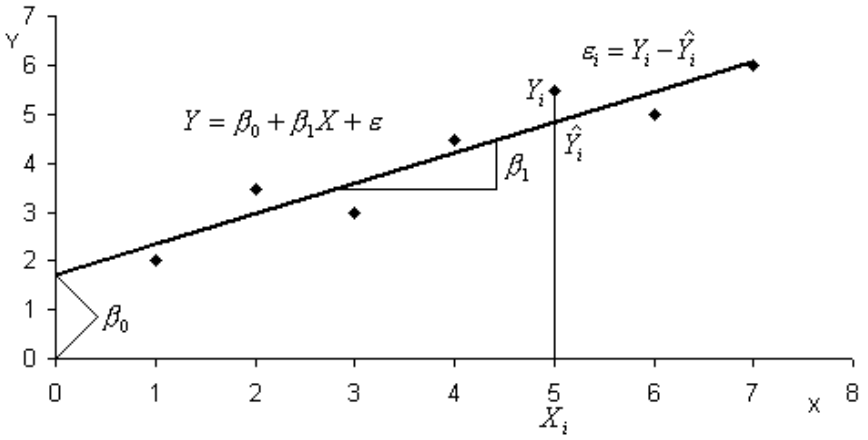


Figura 7.1: Ajuste de uma reta a um conjunto de pontos.

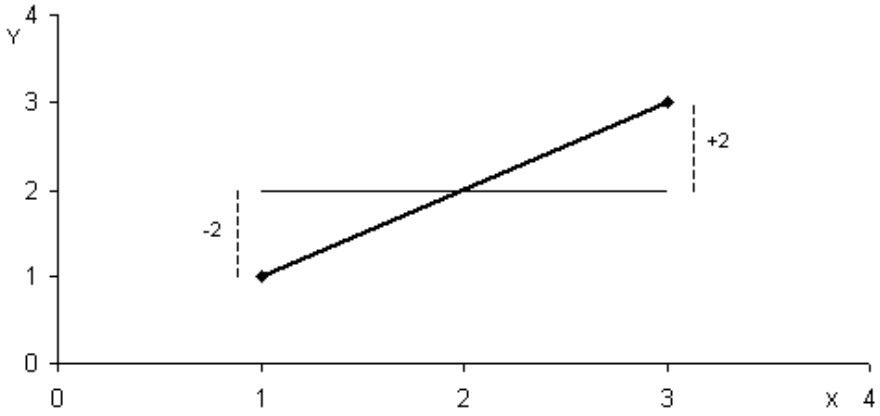


Figura 7.2: Ajuste segundo o critério de minimização de $\sum \left(Y_i - \hat{Y}_i\right)$.

parece claro que esta abordagem não satisfaz o objetivo pelo facto de as distâncias positivas cancelarem as distâncias negativas. Uma alternativa seria considerar o módulo do desvio das distâncias já que esta abordagem ultrapassaria o problema do cancelamento dos desvios.

A Figura 7.3 ilustra que, também, esta abordagem não responde ao objetivo. Nesta figura, a reta a cheio, embora apresente uma soma do módulo dos desvios inferior ao da reta a traço fino, esta última ajusta-se melhor aos três pontos considerados. Por último, a consideração do quadrado dos desvios mostra que a reta que melhor se ajusta aos dados é aquela que apresenta um menor soma dos quadrados dos desvios (para a reta a traço fino, 3, e para a reta a traço grosso, 4). A abordagem para a determinação da reta que melhor se ajusta será, assim, determinada pela minimização dos quadrados dos desvios, o que justifica o nome do método dos mínimos quadrados.

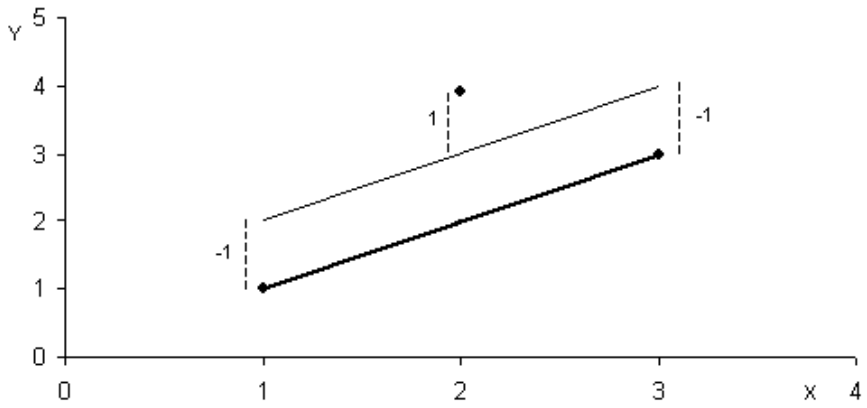


Figura 7.3: Ajuste segundo o critério de minimização de $\sum \left| Y_i - \hat{Y}_i \right|$.

Exemplo 7.1.1: Ajuste de uma reta

O conjunto de observações da tabela representa a tensão de rotura (10 × psi - libras por polegada quadrada) de provetes de cimento em função da quantidade de um aditivo (g). O diagrama de dispersão das observações é apresentado na seguinte figura.

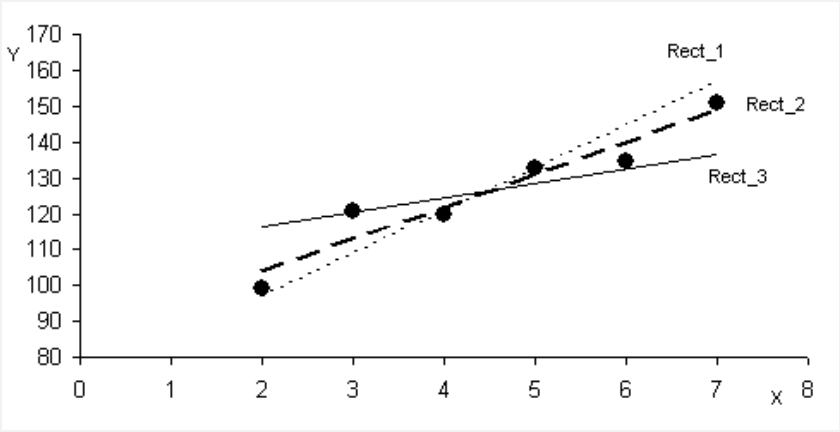


Diagrama de dispersão com três retas de ajuste.

Assumindo uma relação linear entre tensão e a quantidade de aditivo, qual das três relações melhor se ajusta aos pontos observados?

Aditivo	Tensão
X	Y
2	99
3	121
4	120
5	133
6	135
7	151

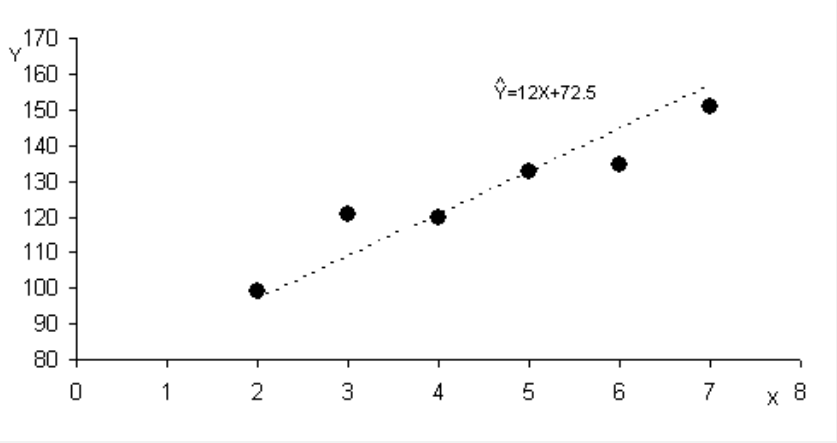
Solução

As figuras apresentam as três retas ajustadas com a respetiva equação. Para as três retas sugeridas, a tabela apresenta os valores estimados de acordo com as seguintes equações,

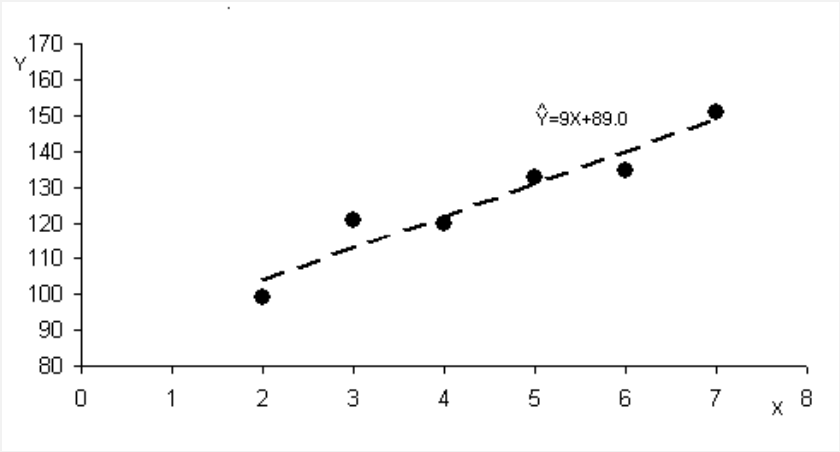
$$\hat{Y}_1 = 72.5 + 12.0X$$

$$\hat{Y}_2 = 86.0 + 9.0X$$

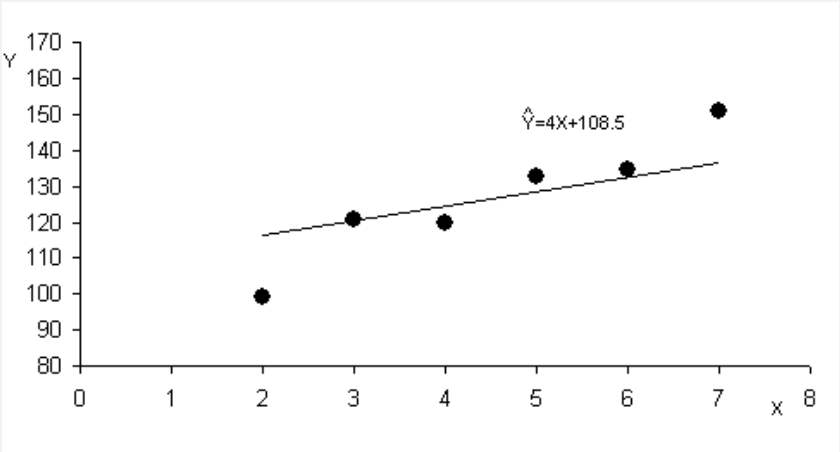
$$\hat{Y}_3 = 108.5 + 4.0X$$



Ajuste segundo a reta $\hat{Y}_1 = 72.5 + 12.0X$.



Ajuste segundo a reta $\hat{Y}_2 = 86.0 + 9.0X$.



Ajuste segundo a reta $\hat{Y}_3 = 108.5 + 4.0X$.

X	Y	$Reta_1$	$Reta_2$	$Reta_3$
2	99	96.5	104	116.5
3	121	108.5	113	120.5
4	120	120.5	122	124.5
5	133	132.5	131	128.5
6	135	144.5	140	132.5
7	151	156.5	149	136.5

A partir dos valores estimados é possível calcular os resíduos para cada um dos modelos.

$Y_i - \hat{Y}_{1,i}$	$Y_i - \hat{Y}_{2,i}$	$Y_i - \hat{Y}_{3,i}$
-2.5	-5	-17.5
-12.5	8	0.5
0.5	-2	-4.5
-0.5	2	4.5
9.5	-5	2.5
5.5	2	14.5
0	0	0

Assim, verifica-se que a soma dos desvios para as três retas é igual a zero, e a soma dos desvios absolutos é diferente de zero como seria de esperar.

$\left\lvert Y_i - \hat{Y}_{1,i} \right\rvert$	$\left\lvert Y_i - \hat{Y}_{2,i} \right\rvert$	$\left\lvert Y_i - \hat{Y}_{3,i} \right\rvert$
2.5	5	17.5
12.5	8	0.5
0.5	2	4.5
0.5	2	4.5
9.5	5	2.5
5.5	2	14.5
31	24	44

Por outro lado, verifica-se que a reta que apresenta a menor soma dos quadrados dos resíduos, a reta de mínimos quadrados, é a que visualmente melhor se ajusta aos dados.

$\left(Y_i - \hat{Y}_{1,i}\right)^2$	$\left(Y_i - \hat{Y}_{2,i}\right)^2$	$\left(Y_i - \hat{Y}_{3,i}\right)^2$
6.25	25	306.25
156.25	64	0.25
0.25	4	20.25
0.25	4	20.25
90.25	25	6.25
30.25	4	210.25

Convém referir que é possível obter retas com uma menor soma absoluta dos resíduos mas que não deixam de exibir uma maior soma dos quadrados dos resíduos. A reta

$$\hat{Y} = 81.5 + 10.0X$$

apresenta uma soma dos resíduos absolutos de 22 e uma soma dos quadrado dos resíduos de 143.5.

Por último, sendo possível determinar várias retas com soma nula dos resíduos, ou com a mesma soma absoluta de desvios, só existe uma reta cuja soma dos quadrados dos resíduos é mínima, isto é, a reta de mínimos quadrados é única.

7.2 O Método dos Mínimos Quadrados

O modelo de regressão linear simples pode ser descrito matematicamente pela seguinte equação

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

onde β_0 e β_1 são os parâmetros desconhecidos que definem a reta. Como se pode ver pela Figura 7.1, β_0 corresponde à intersecção na origem e β_1 ao declive. O termo de erro ε representa a contribuição dos erros aleatórios, umas vezes positivos, outras, negativos. Assumindo que o termo de erro tem média zero, logo $E[\varepsilon] = 0$. Assim, o valor médio de Y é dado pelo seu valor esperado

$$E[Y] = E[\beta_0 + \beta_1 X + \varepsilon] = \beta_0 + \beta_1 X + E[\varepsilon] = \beta_0 + \beta_1 X.$$

Assim, o valor médio de Y para um dado valor de X, assumido como fixo, denotado por $E[Y|X]$, representa a reta com intersecção β_0 e declive β_1.

Para uma dada observação i, o modelo representado pela equação

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

pode ser reescrito como

$$Y_i = \hat{Y}_i + \varepsilon_i$$

onde $\hat{Y}$ representa o modelo ajustado. Por isso, o termo de erro pode ser definido como a diferença entre o valor observado e o valor ajustado,

$$\varepsilon_i = Y_i - \hat{Y}_i$$

$\hat{Y}$ representa o valor médio de Y , sendo $\hat{\beta}_0$ e $\hat{\beta}_1$ as estimativas dos parâmetros β_0 e β_1. Assim, os erros podem ser definidos como

$$\left(Y_i - \hat{Y}_i\right) = \left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right],$$

sendo a Soma dos Quadrados dos Resíduos, para os pares de pontos, dada por

$$SQR = \sum_{i=1}^{n} \left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right]^2.$$

O Método dos Mínimos Quadrados determina os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizam a Soma dos Quadrados dos Resíduos. Para tal, é necessário determinar as derivadas parciais da SQR

relativamente a $\hat{\beta}_0$ e $\hat{\beta}_1$, respetivamente.

$$\frac{\partial SQR}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2\left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right](-1)$$
$$\frac{\partial SQR}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2\left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right](-X_i)$$

Igualando a zero as derivadas parciais e rearranjando os termos das equações, obtêm-se as chamadas equações normais,

$$\begin{cases} \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 X_i = 0 \\ \sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} \hat{\beta}_0 X_i - \sum_{i=1}^{n} \hat{\beta}_1 X_i^2 = 0 \end{cases} \qquad \begin{cases} \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i X_i \end{cases}$$

A resolução do sistema de equações conduz às seguintes estimativas para os parâmetros

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{\beta}_1 X_i}{n} = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{n \sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

Para simplificar as fórmulas, introduzem-se as seguintes somas,

$$S_{xx} = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2$$

$$S_{yy} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2$$

$$S_{xy} = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = \sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)$$

Assim, as estimativas dos coeficientes podem ser reexpressas como,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Exemplo 7.2.1: Estimativas dos coeficientes da reta

A tabela apresenta a resistência (em psi) em função da idade (dias) para provetes de material cerâmico.

Dias	Resistência
3	1.4
4	1.5
5	2.1
6	2.4
8	3.1
9	3.2
10	3.3

Determine a reta de mínimos quadrados. Qual a resistência para uma idade de 7 dias?

Solução

A tabela apresenta os cálculos necessários à determinação dos coeficientes.

X	Y	X^2	XY
3	1.4	9	4.2
4	1.5	16	6
5	2.1	25	10.5
6	2.4	36	14.4
8	3.1	64	24.8
9	3.2	81	28.8
10	3.3	100	33
45	17	331	121.7

Assim, com $n = 7$, $\sum X_i = 45$, $\sum Y_i = 17$, $\sum X_i^2 = 331$ e $\sum X_i Y_i = 121.7$, e calculando

$$S_{xx} = 331 - \frac{1}{7}(45)^2 = 41.714$$

$$S_{xy} = 121.7 - \frac{1}{7}(45)(17) = 12.414$$

as estimativas são

$$\hat{\beta}_1 = \tfrac{12.414}{41.714} = 0.298$$
$$\hat{\beta}_0 = \tfrac{17}{7} - (0.298)\tfrac{45}{7} = 0.515$$

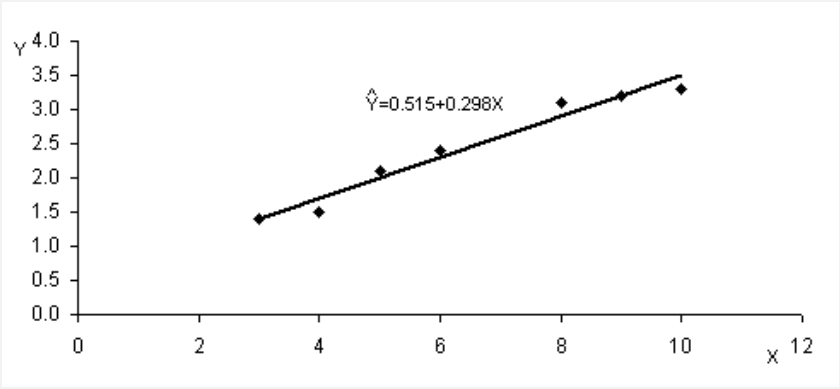
A reta de mínimos quadrados é

$$\hat{Y} = 0.515 + 0.298X$$

Para $X = 7$, obtém-se o seguinte valor para a resistência,

$$\hat{Y} = 0.515 + 0.298(7) = 2.599$$

A figura apresenta a reta ajustada.



Ajuste segundo a reta $\hat{Y} = 0.515 + 0.298X$.

O método dos mínimos quadrados constitui, como se depreende dos exemplos anteriores, uma forma de produzir estimativas dos parâmetros desconhecidos β_0 e β_1. A distribuição dos estimadores depende do termo de erro aleatório ε. Por esta razão, a validade do método depende de algumas assunções relativas aos erros.

1. A média da distribuição de probabilidade do erro é 0. Isto é, para cada valor da variável independente X, a média dos erros é 0.

$$E[\varepsilon] = 0$$

$$E[Y] = \beta_0 + \beta_1 X + E[\varepsilon] = \beta_0 + \beta_1 X$$

2. A variância da distribuição de probabilidade do erro é constante para todos os valores da variável independente X.

$$Var[\varepsilon] = \sigma^2$$

3. A distribuição de probabilidade do erro é normal.

$$\varepsilon \sim N\left(0, \sigma^2\right)$$

4. Os erros associados com quaisquer duas observações são independentes.

$$\varepsilon \sim iid$$

Graficamente, as assunções impostas ao modelo traduzem-se, esquematicamente, na Figura 7.4, que mostra que existe uma relação linear entre as variáveis dependente e independente e que a distribuição da variável dependente e do termo de erro têm a mesma forma e variância constante.

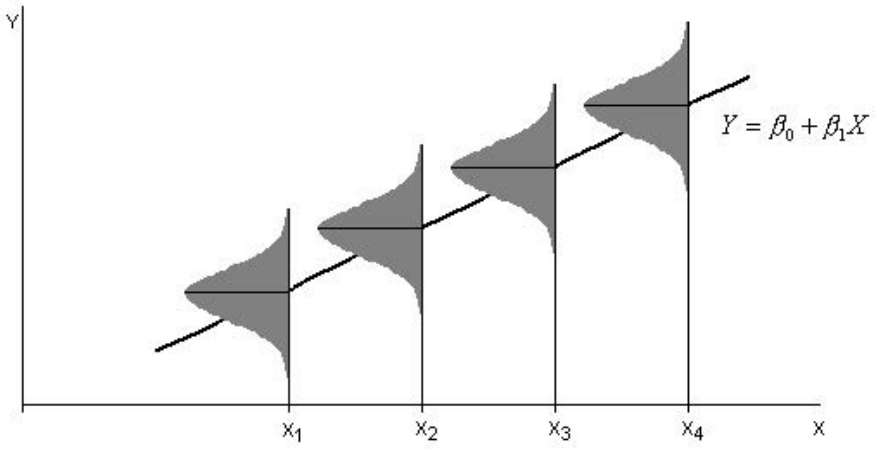


Figura 7.4: Modelo de Regressão Linear e distribuição dos erros.

7.3 Propriedades dos Estimadores de Mínimos Quadrados

7.3.1 Média e Variância de β_1

O estimador de β_1 depende dos valores conhecidos da variável independente, assumida como não aleatória.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Portanto, o termo S_{xx} pode ser tratado como uma constante. No Anexo mostra-se que a esperança de $\hat{\beta}_1$ é dada por

$$E\left[\hat{\beta}_1\right] = \beta_1$$

e, portanto, $\hat{\beta}_1$ é um estimador não enviesado de β_1, com a seguinte variância,

$$Var\left[\hat{\beta}_1\right] = \frac{\sigma^2}{S_{xx}}.$$

7.3.2 Média e Variância de β_0

O estimador de β_0 é dado pela seguinte expressão

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

com o seguinte valor esperado

$$E\left[\hat{\beta}_0\right] = \beta_0$$

e, portanto, $\hat{\beta}_0$ é também um estimador não enviesado de β_0, com a seguinte variância,

$$Var\left[\hat{\beta}_0\right] = \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)\sigma^2$$

Os coeficientes $\hat{\beta}_0$ e $\hat{\beta}_1$ podem ser formulados como combinações lineares das variáveis aleatórias Y_i. Resulta, portanto, que as estimativas de $\hat{\beta}_0$ e de $\hat{\beta}_1$ não são independentes. Assim, a sua covariância será dada por (ver Anexo).

$$Cov\left[\hat{\beta}_0, \hat{\beta}_1\right] = \sigma^2 \left(-\frac{\bar{X}}{S_{xx}}\right).$$

7.3.3 Estimação de σ^2

As condições impostas ao modelo de regressão implicam que o termo de erro

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

seja uma variável aleatória com média zero e variância constante σ^2. Como os valores de X são assumidos como fixos, Y é uma variável aleatória com média

$$Y = \beta_0 + \beta_1 X$$

e variância σ^2. Portanto, resulta que os valores estimados de $\hat{\beta}_0$ e $\hat{\beta}_1$ dependem, como foi visto, dos valores observados de Y.

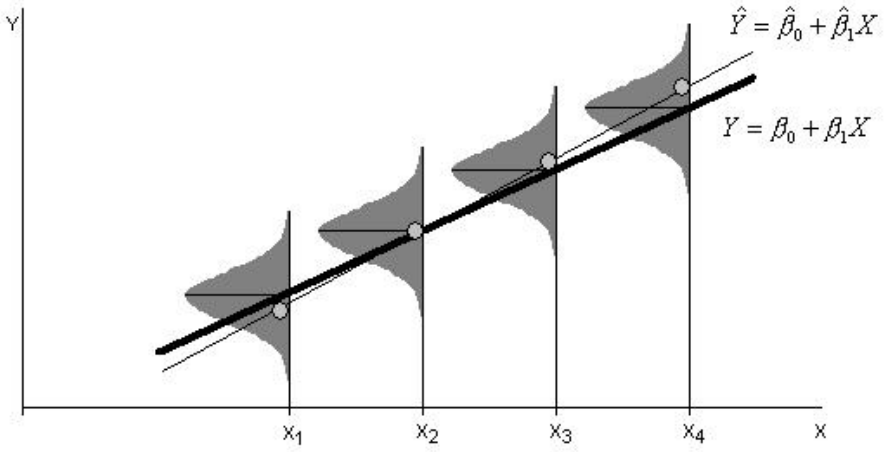


Figura 7.5: Verdadeira linha de regressão e linha estimada.

No entanto, para estabelecer inferências acerca dos coeficientes de regressão é necessário estimar a variância σ^2. Como se depreende da Figura 7.5 o termo de erro representa a flutuação aleatória à volta da verdadeira linha de regressão.

A soma dos quadrados dos resíduos é dada por

$$SQR = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

Tendo em atenção que para uma amostra aleatória de n observações $Y_1, Y_2, \ldots, Y_n$, provenientes de uma distribuição normal com média μ e variância σ^2, a distribuição amostral de

$$\frac{(n-1)\,S^2}{\sigma^2}$$

segue uma distribuição de χ^2 com $n-1$ graus de liberdade, é possível demonstrar que o valor esperado da Soma dos Quadrados dos Resíduos é

$$E\left[SQR\right] = (n-2)\,\sigma^2.$$

Portanto, para as condições especificadas relativamente aos termos de erro,

$$E\left[S^2\right] = E\left[\frac{SQR}{n-2}\right] = \frac{\sigma^2}{n-2}\,(n-2) = \sigma^2$$

é um estimador não tendencioso de σ^2 e

$$\frac{SQR}{\sigma^2} = \frac{(n-2)\,S^2}{\sigma^2}$$

possui uma distribuição de χ^2 com $n-2$ graus de liberdade, dado que dois graus de liberdade são requeridos para estimar β_0 e β_1.

Em resumo, para determinar a estimativa para σ^2, basta usar a relação

$$S^2 = \frac{SQR}{n-2}$$

com

$$SQR = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = S_{yy} - \hat{\beta}_1 S_{xy}$$

e

$$S_{yy} = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}.$$

Exemplo 7.3.1: Estimação da variância

Determine a estimativa de σ^2 para o Exemplo 7.2.

Solução

A tabela apresenta os valores observados, os valores estimados, os resíduos, o quadrado dos resíduos e o produto dos resíduos pelos valores da variável independente.

X_i	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	$\left(Y_i - \hat{Y}_i\right)^2$	$X_i\left(Y_i - \hat{Y}_i\right)$
3	1.4	1.408219	-0.008219	0.000068	-0.024658
4	1.5	1.705822	-0.205822	0.042363	-0.823288
5	2.1	2.003425	0.096575	0.009327	0.482877
6	2.4	2.301027	0.098973	0.009796	0.593836
8	3.1	2.896233	0.203767	0.041521	1.630137
9	3.2	3.193836	0.006164	0.000038	0.055479
10	3.3	3.491438	-0.191438	0.036649	-1.914384
45	17	17	0	0.13976	0

Em primeiro lugar pode verificar-se que a soma dos valores estimados $\hat{Y}_i$ é igual à soma dos valores observados Y_i. Também a soma dos resíduos $\sum\left(Y_i - \hat{Y}_i\right)$ é igual a zero. Por último, a soma do produto dos resíduos pela variável independente $X_i\left(Y_i - \hat{Y}_i\right)$ é também igual a zero, o que decorre da dedução das equações normais.

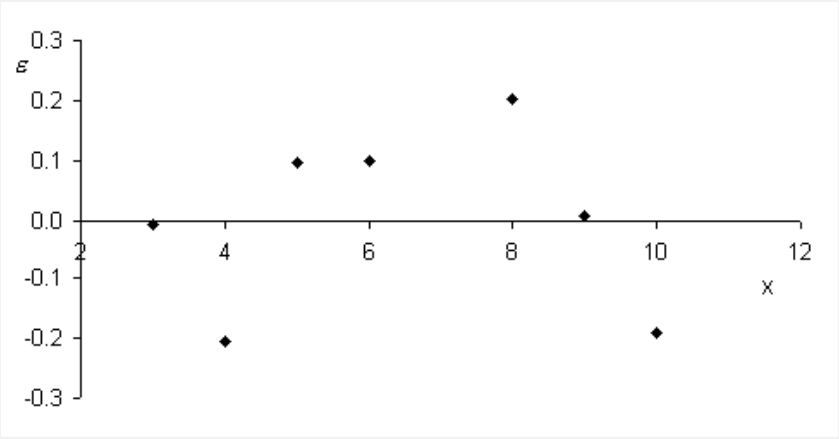
A estimativa para a variância é

$$S^2 = \frac{0.139670}{5} = 0.029752$$

a que corresponde um desvio padrão estimado

$$S = \sqrt{0.027952} = 0.167189$$

Decorre das propriedades da distribuição que a maioria das observações devem estar a menos de $2S$ do valor estimado pela reta de regressão. A próxima figura apresenta os resíduos observados em função da variável independente e, como se pode observar, os resíduos distribuem-se aleatoriamente à volta do zero.



Resíduos.

7.4 Inferências sobre os coeficientes de regressão

Em última análise, o objetivo do desenvolvimento de um modelo de regressão linear é responder à questão se a variável dependente Y está linearmente relacionada com a variável independente X, isto é, se Y varia linearmente à medida que X varia. Caso assim não seja, o coeficiente β_1 no modelo será igual a zero. Debaixo das assunções relativas à distribuição dos erros, a distribuição amostral $\hat{\beta}_1$ terá uma distribuição normal com média e variância dadas por

$$E\left[\hat{\beta}_1\right] = \beta_1 \qquad Var\left[\hat{\beta}_1\right] = \frac{\sigma^2}{S_{xx}}$$

tal como mostra a Figura 7.6.

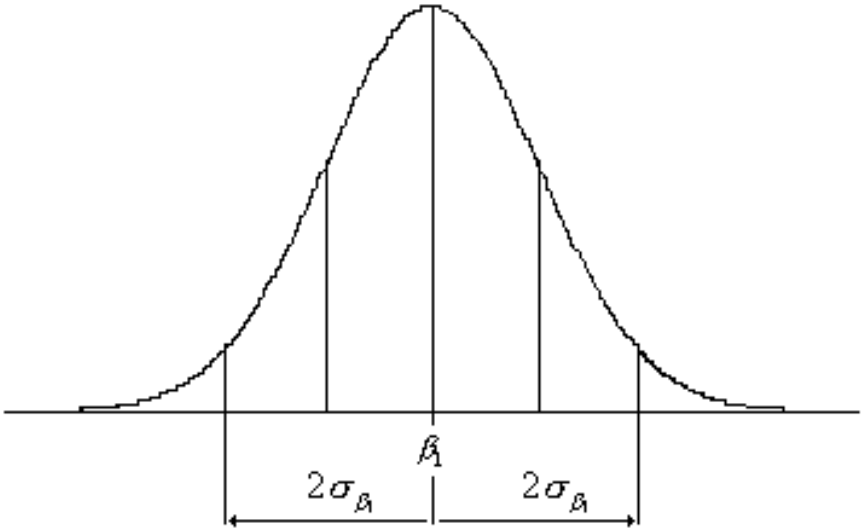


Figura 7.6: Distribuição amostral de $\hat{\beta}_1$.

Assim, é possível realizar inferências acerca do declive β_1, através de um intervalo de confiança ou de um teste de hipóteses. Deve ser referido que, em geral, não tem interesse a realização de um teste de hipóteses para a intersecção na origem β_0. Contudo, a estatística de teste neste último caso é dada por

$$t = \frac{\hat{\beta}_0 - \text{Valor especificado para } \beta_0}{S\sqrt{(1/n) + \left(\bar{X}\right)^2 \Big/ S_x}}$$

que segue uma distribuição de t-Student com $n-2$ graus de liberdade.

Definição 7.4.1: Intervalo de Confiança para o Declive da Reta de Regressão, β_1

$$\hat{\beta}_1 - t_{\alpha/2,n-2} S_{\hat{\beta}_1} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2,n-2} S_{\hat{\beta}_1}$$

onde

$$S_{\hat{\beta}_1} = \frac{S}{\sqrt{S_{xx}}}$$

Definição 7.4.2: Teste de Hipóteses acerca do Declive da Reta de Regressão, β_1 (Grandes Amostras, $n \geq 30$)

Teste Unilateral	Teste Bilateral
$H_0 : \beta_1 = 0$	$H_0 : \beta_1 = 0$
$H_1 : \beta_1 > 0$	$H_1 : \beta_1 \neq 0$
$(H_1 : \beta_1 < 0)$	

Estatística

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{S/\sqrt{S_{xx}}}$$

Região de Rejeição	Região de Rejeição
$t > t_\alpha$	$\lvert t \rvert > t_{\alpha/2}$
$(t < -t_\alpha)$	

onde t_α e $t_{\alpha/2}$ são, respetivamente, os valores da distribuição t-Student com graus de liberdade, tal que $P(t > t_\alpha) = \alpha$ e $P\left(t > t_{\alpha/2}\right) = \alpha/2$.

Exemplo 7.4.1: Teste de hipótese ao declive

Para os dados do Exemplo 7.2 teste a hipótese de que $\beta_1 = 0$.

Solução

1. Formulação das hipóteses

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

2. Região crítica

$$|t| > t_{0.025} = 2.571$$

3. Teste estatístico

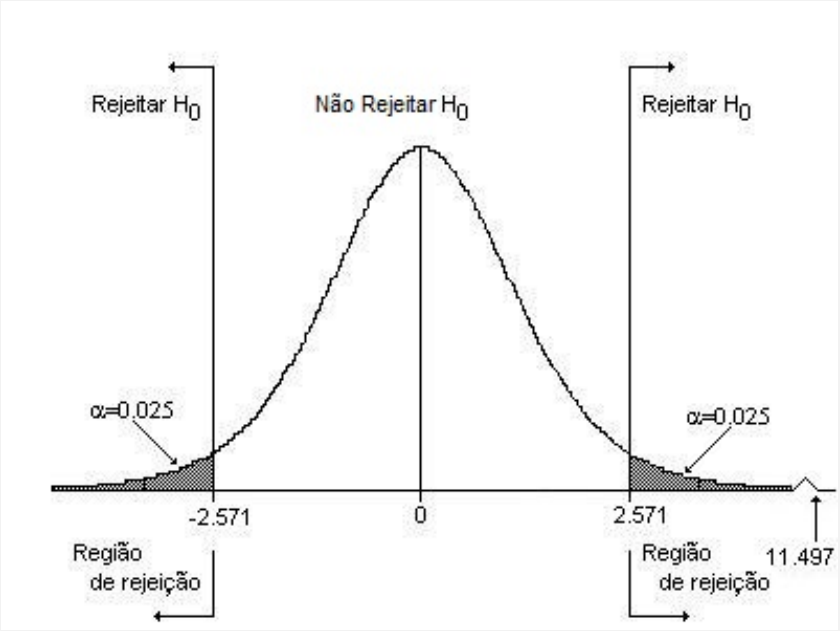
$$t = \frac{\hat{\beta}_1}{S/\sqrt{S_{xx}}} = \frac{0.298}{0.167/\sqrt{41.714}} = 11.497$$

4. Decisão

   Rejeitar a Hipótese Nula, ou seja, o declive da reta β_1, é diferente de zero e, portanto, o modelo explica a resistência em função da idade através de uma relação linear.

   O intervalo de confiança de 95% seria

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} S_{\hat{\beta}_1} = 0.298 \pm 2.571 \frac{0.161}{\sqrt{41.714}} = 0.298 \pm 0.026$$



Região de rejeição.

7.5 Estimativas Intervalares

O modelo de regressão permite responder à questão: para um dado valor de X, qual o valor previsto para Y? No Exemplo 7.2, para uma idade $X = 7$ dias, o valor previsto da resistência foi $\hat{Y} = 2.599$ psi. Contudo, o interesse pode residir no valor médio de Y, para um dado valor de X, ou no valor individual de Y para um dado valor de X. No primeiro caso, o objetivo seria estimar o valor médio da resistência para um grande número de provetes com idade igual a 7 dias; no segundo caso, o interesse residiria na previsão da resistência para um só resultado experimental, isto é, para uma dada idade. Em ambos os casos, o valor previsto é o mesmo, dado que é baseado no mesmo

modelo. Contudo, o intervalo de variação no primeiro caso será menor do que no segundo caso.

7.5.1 Inferência para um valor médio de Y_0

Assim, de uma forma geral, para um dado valor X_0, o valor esperado da resposta será

$$E\left[Y\left|X_0\right.\right] = \mu_{Y|X_0} = \beta_0 + \beta_1 X_0$$

a que corresponde o seguinte valor estimado

$$\hat{\mu}_{Y|X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

que constitui um estimador não enviesado. A variância é dada por

$$Var\left[\hat{\mu}_{Y|X_0}\right] = Var\left[\hat{\beta}_0\right] + X_0^2 Var\left[\hat{\beta}_1\right] + 2X_0 Cov\left[\hat{\beta}_0, \hat{\beta}_1\right]$$

que, substituindo pelas expressões deduzidas, conduz à seguinte expressão

$$Var\left[\hat{\mu}_{Y|X_0}\right] = \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)\sigma^2 + X_0^2\frac{\sigma^2}{S_{xx}} + 2X_0\sigma^2\left(-\frac{\bar{X}}{S_{xx}}\right)$$

que pode ser simplificada, obtendo-se

$$Var\left[\hat{\mu}_{Y|X_0}\right] = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} + \frac{X_0^2}{S_{xx}} - 2X_0\frac{\bar{X}}{S_{xx}}\right) = \sigma^2\left[\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{S_{xx}}\right]$$

Assim. o intervalo de confiança para a resposta média é dado por

$$\hat{\mu}_{Y|X_0} \pm t_{\alpha/2,n-2}\sqrt{S^2\left[\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{S_{xx}}\right]}$$

7.5.2 Inferência para um valor particular de Y_0

Na previsão de um valor individual Y_0, dado valor X_0, o erro da estimativa será dado pela diferença $Y_0 - \hat{Y}_0$.

Contudo, assumindo que Y_0 e $\hat{Y}_0$ são normalmente distribuídos, o seu erro também será normalmente distribuído, de acordo com as assunções impostas no desenvolvimento do modelo. O valor esperado do erro será nulo,

$$E\left[Y_0 - \hat{Y}_0\right] = 0.$$

A variância será dada por

$$Var\left[Y_0 - \hat{Y}_0\right] = Var\left[Y_0\right] + Var\left[\hat{Y}_0\right] - 2Cov\left[Y_0, \hat{Y}_0\right].$$

No entanto, como o valor previsto Y_0 não foi usado na dedução do modelo de regressão, Y_0 e $\hat{Y}_0$ são independentes e a sua covariância é nula. Assim sendo, a variância é dada por,

$$Var\left[Y_0 - \hat{Y}_0\right] = \sigma^2 + Var\left[\hat{\beta}_0 + \hat{\beta}_1 X_0\right] = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{S_{xx}}\right]$$

isto é,

$$Var\left[Y_0 - \hat{Y}_0\right] = \sigma^2\left[1 + \frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{S_{xx}}\right].$$

Portanto, o intervalo para a reposta individual é dado por

$$\hat{Y}_0 \pm t_{\alpha/2,n-2}\sqrt{S^2\left[1 + \frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{S_{xx}}\right]}$$

A Figura 7.7 apresenta as curvas correspondentes aos intervalos de confiança de 95% para a média e para os valores individuais, sendo as bandas para o valor individual mais largas que as bandas do valor médio.

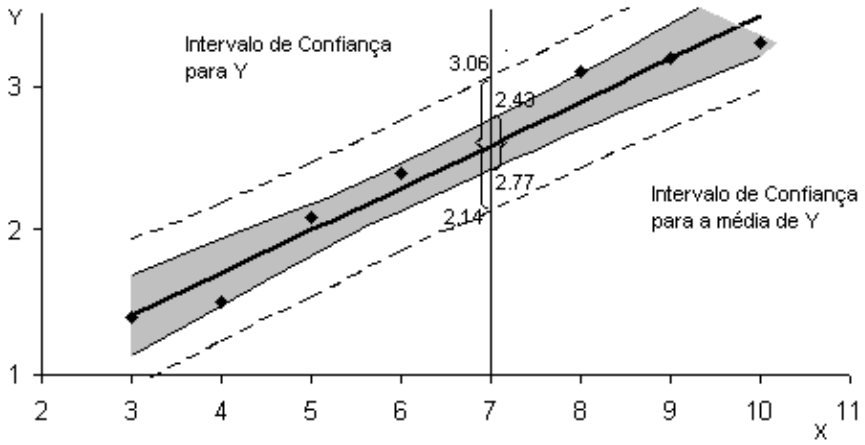


Figura 7.7: Limites de confiança para as estimativas.

7.6 Coeficiente de Correlação

O coeficiente de correlação é uma medida de como duas variáveis variam em conjunto. Dado um conjunto de pares de valores (X_i, Y_i) de uma população bivariada normal com a seguinte função densidade

$$f(x,y) = \frac{e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{(x-\mu_x)}{\sigma_x}\right)^2 - 2\rho\left(\frac{(x-\mu_x)}{\sigma_x}\right)\left(\frac{(y-\mu_y)}{\sigma_y}\right) + \left(\frac{(y-\mu_y)}{\sigma_y}\right)^2\right]}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}$$

o parâmetro ρ é chamado o coeficiente de correlação com $Cov[X,Y] = \rho\sigma_x\sigma_y$. Um coeficiente de correlação nulo implica, portanto, que as duas variáveis são independentes. Contudo, tendo em atenção a definição de covariância de duas variáveis,

$$Cov[X,Y] = E[(X-\mu_x)(Y-\mu_y)] = E[XY] - \mu_x\mu_y$$

o coeficiente de correlação pode ser expresso como função dos valores padronizados de X e de Y,

$$\rho = E\left[\left(\frac{X-\mu_x}{\sigma_x}\right)\left(\frac{Y-\mu_y}{\sigma_y}\right)\right]$$

sendo coeficiente de correlação amostral (correlação de Pearson) dado por

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{S_{\bar{X}}}\right)\left(\frac{Y_i-\bar{Y}}{S_{\bar{Y}}}\right).$$

Em primeiro lugar, deve ser realçado que o coeficiente de correlação é adimensional e o seu valor situa-se entre -1 e 1.

Exemplo 7.6.1: Coeficiente de correlação

O Índice de Desenvolvimento de Griffiths é uma medida agregadora destinada a avaliar o desenvolvimento psico-motor de crianças. Este índice é calculado através da avaliação de determinadas tarefas motoras e intelectuais. Os dados representam as avaliações motora e intelectual para 9 crianças com a idade de 4 anos. Calcule o coeficiente de correlação.

Motora	Intelectual
84	77
73	85
101	105
74	86
88	108
100	116
86	96
95	100
82	100

Solução

A tabela apresenta os cálculos necessários à determinação do coeficiente de correlação. As médias das classificações são, respetivamente, 87 e 97.

Criança	X_i	Y_i	$\underbrace{\left(X_i-\bar{X}\right)}_{x_i}$	$\underbrace{\left(Y_i-\bar{Y}\right)}_{y_i}$	x_iy_i	$\left(X_i-\bar{X}\right)^2$	$\left(Y_i-\bar{Y}\right)^2$
1	84	77	-3	-20	60	9	400
2	73	85	-14	-12	168	196	144
3	101	105	14	8	112	196	64
4	74	86	-13	-11	143	169	121
5	88	108	1	11	11	1	121
6	100	116	13	19	247	169	361
7	86	96	-1	-1	1	1	1
8	95	100	8	3	24	64	9
9	82	100	-5	3	-15	25	9
Total	783	873	0	0	751	830	1230
	$\bar{X}=87$	$\bar{Y}=97$			$S_x=10.2$	$S_y=12.4$	

O Coeficiente de Correlação pode ser calculado por esta fórmula equivalente

$$R=\frac{\sum\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\sum\left(X_i-\bar{X}\right)^2\sum\left(Y_i-\bar{Y}\right)^2}}.$$

Assim, o valor do Coeficiente de Correlação é

$$R=\frac{751}{\sqrt{(830)\,(1230)}}=0.743$$

Para uma melhor compreensão do Coeficiente de Correlação, os gráficos abaixo apresentam as diversas transformações que são operadas nos dados e como tal se traduz graficamente. Assim, a seguinte figura apresenta a dispersão dos valores originais.

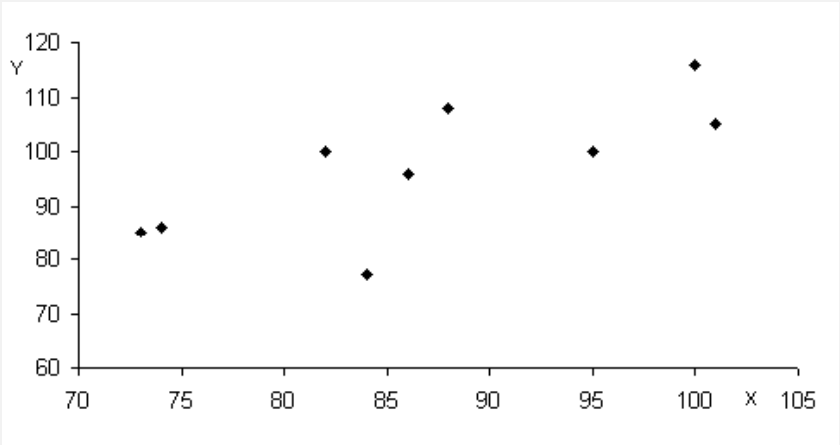
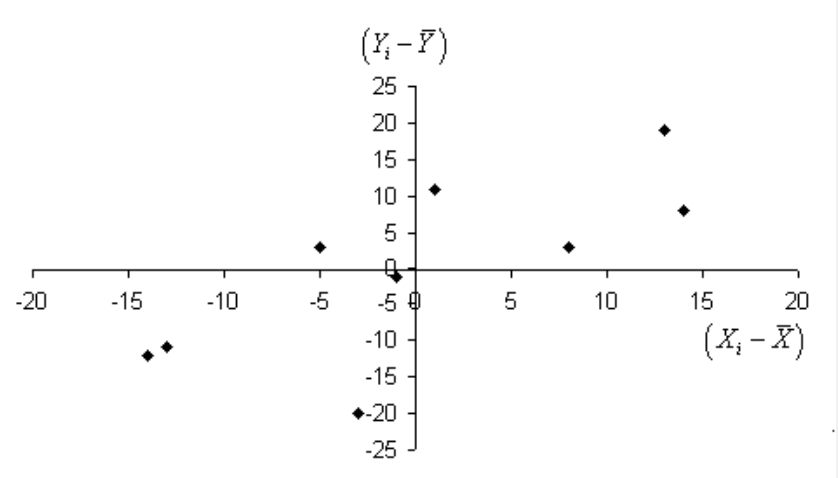


Gráfico de dispersão.

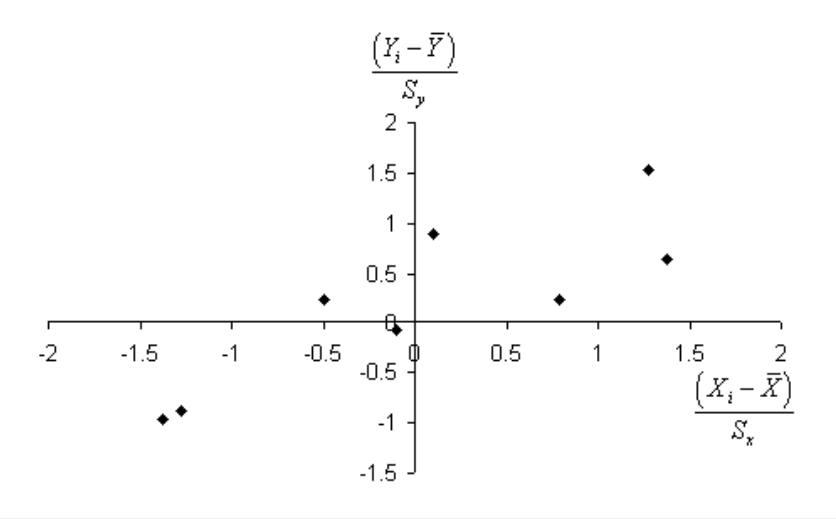
A figura apresenta as observações subtraídas das respetivas médias, o que conduz a uma

mudança dos eixos para as médias de X e de Y. Ambas as representações dependem das unidades das observações.



Representação dos desvios relativos à média das observações.

A última figura apresenta os valores padronizados, o que, relativamente à figura anterior, corresponde à divisão dos desvios das observações relativamente à sua média, pelo respetivo desvio padrão, representação que é adimensional.

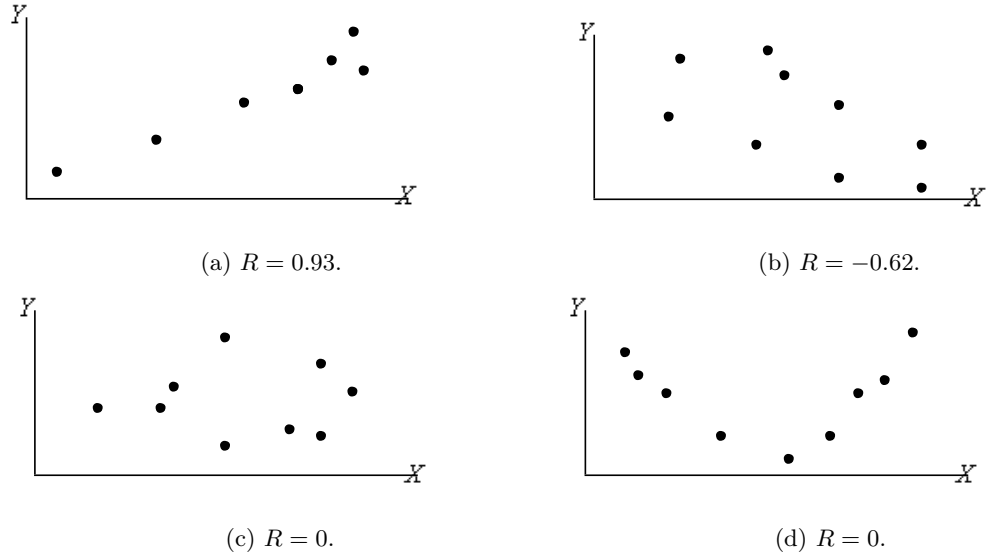


Representação das observações padronizadas.

Observações no primeiro e terceiro quadrantes apresentam uma covariação positiva, enquanto que observações no segundo e quarto quadrantes exibem um produto

$$(X_i - \bar{X})(Y_i - \bar{Y})$$

negativo.



(a) $R = 0.93$. (b) $R = -0.62$.

(c) $R = 0$. (d) $R = 0$.

Figura 7.8: Exemplos de diagramas de dispersão para diferentes valores de R.

Assim, quando a maioria das observações se situam no primeiro e terceiro quadrantes o coeficiente de correlação é positivo, indicando que X e Y variam no mesmo sentido.

Por outro lado, quando a maioria das observações se situam no segundo e quanto quadrantes, implicando que X e Y variam em sentido inverso, o coeficiente de correlação é negativo. Se as observações estiverem distribuídas pelos quatro quadrantes o coeficiente de correlação é nulo.

A Figura 7.8 mostra os diagramas de dispersão para diferentes valores do coeficiente de correlação. Uma correlação com $R = 1$ implica uma distribuição dos pontos sobre uma linha reta com declive positivo; inversamente, numa correlação com $R = -1$, os pontos distribuem-se segundo uma reta com declive negativo. Uma correlação nula implica que não existe uma relação linear entre as variáveis.

Uma última nota deve ser feita sobre o coeficiente de correlação. A correlação não implica causalidade mas tão só a possível existência de uma associação linear A causalidade entre duas variáveis pode ser estabelecida através, por exemplo, de modelos de regressão. Por isso, face a um elevado coeficiente de correlação é incorreto assumir que uma mudança em X implica uma variação em Y.

Inferências acerca do coeficiente de correlação da população ρ podem ser estabelecidas a partir do coeficiente de correlação amostral R.

Definição 7.6.1: Teste de Hipóteses para o Coeficiente de Correlação, ρ

Teste Unilateral	Teste Bilateral
$H_0 : \rho = 0$	$H_0 : \rho = 0$
$H_1 : \rho > 0$	$H_1 : \rho \neq 0$
$(H_1 : \rho < 0)$	

Estatística

$$t = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

Região de Rejeição	Região de Rejeição
$t > t_\alpha$	$\lvert t\rvert > t_{\alpha/2}$
$(t < -t_\alpha)$	

onde t_α e $t_{\alpha/2}$ são, respetivamente, os valores da distribuição t-Student com $n-2$ graus de liberdade, tal que $P\left(t > t_\alpha\right) = \alpha$ e $P\left(t > t_{\alpha/2}\right) = \alpha/2$. Os valores (X, Y) são aleatoriamente selecionados de uma população normal bivariada.

Exemplo 7.6.2: Teste de hipótese acerca da correlação

Teste a hipótese de que não existe correlação entre as avaliações motora e intelectual para os dados do exemplo anterior.

Solução

1. Formulação das hipóteses

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

2. Região crítica

$$|t| \leq t_{0.025} = 2.365$$

3. Teste estatístico

$$t = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{0.743\sqrt{9-2}}{\sqrt{1-(0.743)^2}} = 2.940$$

4. Decisão

   Rejeitar a Hipótese Nula, ou seja, o coeficiente de correlação é diferente de zero e, portanto, existe uma associação linear entre as duas variáveis.

Por último, se as duas variáveis são independentes, a sua covariância é nula e também o respetivo coeficiente de correlação. Portanto, a independência implica correlação nula. Contudo,

o inverso não é verdade, já que pode existir uma associação não linear entre as variáveis e o respetivo coeficiente de correlação ser nulo (ver Figura 7.8d).

7.7 Correlação e Regressão - Coeficiente de Determinação, R^2

Num modelo de regressão, a contribuição da variável independente X para a previsão da variável dependente Y, pode ser estudada através de uma medida derivada do coeficiente de correlação. O declive da reta de regressão e o coeficiente de correlação usam informação semelhante para o seu cálculo. Recordando as respetivas expressões,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_i - \bar{X}\right)^2} \qquad R = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum \left(X_i - \bar{X}\right)^2 \sum \left(Y_i - \bar{Y}\right)^2}},$$

e considerando a sua razão,

$$\frac{\beta_1}{R} = \frac{\sqrt{\sum \left(X_i - \bar{X}\right)^2 \sum \left(Y_i - \bar{Y}\right)^2}}{\sum \left(X_i - \bar{X}\right)^2} = \frac{\sqrt{\sum \left(Y_i - \bar{Y}\right)^2}}{\sqrt{\sum \left(X_i - \bar{X}\right)^2}}.$$

Se ambos os termos forem divididos por $n-1$, a relação entre ambos fica

$$\hat{\beta}_1 = R\frac{\sqrt{\sum \left(Y_i - \bar{Y}\right)^2}}{\sqrt{\sum \left(X_i - \bar{X}\right)^2}} = R\frac{S_y}{S_x}$$

o que explica a sua relação próxima.

Exemplo 7.7.1: Coeficiente de determinação

Para os dados do exemplo 7.6.1 determine o declive da reta de regressão.

Solução

Tendo em atenção os cálculos do exemplo anterior, com

$$S_{xy} = \sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = 751 \qquad S_{xx} = \sum \left(X_i - \bar{X}\right)^2 = 830$$

o declive é,

$$\hat{\beta}_1 = \frac{751}{830} = 0.905$$

Como o coeficiente de correlação determinado foi R=0.743, o declive da reta pode ser determinado pela expressão deduzida

$$\hat{\beta}_1 = R\frac{\sqrt{\sum \left(Y_i - \bar{Y}\right)^2}}{\sqrt{\sum \left(X_i - \bar{X}\right)^2}} = 0.743\frac{\sqrt{1230}}{\sqrt{830}} = 0.905.$$

A questão determinante é saber se a variável independente contribui para a previsão da variável dependente. Para qualquer observação é possível estabelecer a seguinte igualdade,

$$\left(Y_i - \bar{Y}\right) = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$

que basicamente traduz o facto de que o desvio de cada observação relativamente à sua média

$$\left(Y_i - \bar{Y}\right)$$

pode ser decomposto na soma do desvio da reta de regressão relativamente à média

$$\left(\hat{Y}_i - \bar{Y}\right)$$

mais o desvio da observação relativamente à reta de regressão

$$\left(Y_i - \hat{Y}_i\right),$$

tal como se pode ver na Figura 7.9.

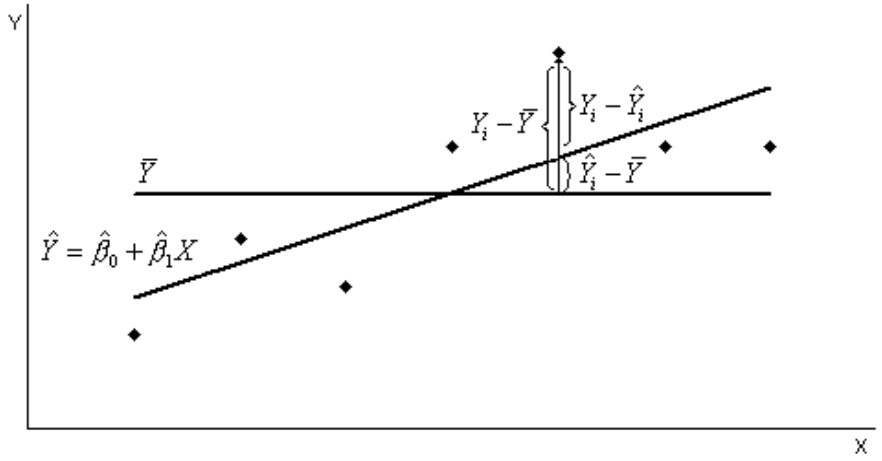


Figura 7.9: Relação entre o valor observado, o valor estimado e a média.

A Soma dos Quadrados dos Desvios de cada observação Y_i relativamente à sua média pode ser expressa como

$$\sum \left(Y_i - \bar{Y}\right)^2 = \sum \left[\left(\hat{Y}_i - \bar{Y}\right)^2 + 2\left(\hat{Y}_i - \bar{Y}\right)\left(Y_i - \hat{Y}_i\right) + \left(Y_i - \hat{Y}_i\right)^2\right]$$

e, tendo em atenção que

$$\begin{aligned}
&\sum \left(\hat{Y}_i - \bar{Y}\right)\left(Y_i - \hat{Y}_i\right) = \sum \left(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y}\right)\varepsilon_i \\
&= \sum \left(\bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i + \bar{Y}\right)\varepsilon_i = \sum \hat{\beta}_1 \left(X_i - \bar{X}\right)\varepsilon_i \\
&= \sum \hat{\beta}_1 X_i \varepsilon_i - \hat{\beta}_1 \bar{X} \sum \varepsilon_i = 0
\end{aligned}$$

como se deduz do sistema de equações normais, conclui-se que

$$\sum \left(Y_i - \bar{Y}\right)^2 = \sum \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum \left(Y_i - \hat{Y}_i\right)^2.$$

Assim, a variação total é dividida em duas parcelas, a variação explicada pela regressão

$$\sum \left(\hat{Y}_i - \bar{Y}\right)^2$$

e a variação não explicada

$$\sum \left(Y_i - \hat{Y}_i\right)^2.$$

Contudo, dado que a variação explicada pode ser expressa como

$$\sum \left(\hat{Y}_i - \bar{Y}\right)^2 = \sum \left(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y}\right)^2 = \sum \left(\bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i - \bar{Y}\right)^2$$

$$\sum \left(\hat{Y}_i - \bar{Y}\right)^2 = \sum \hat{\beta}_1^2 \left(X_i - \bar{X}\right)^2$$

e como

$$\hat{\beta}_1 = R \frac{\sqrt{\sum \left(Y_i - \bar{Y}\right)^2}}{\sqrt{\sum \left(X_i - \bar{X}\right)^2}}$$

logo

$$\sum \left(\hat{Y}_i - \bar{Y}\right)^2 = R^2 \sum \left(Y_i - \bar{Y}\right)^2$$

Assim, o quadrado do coeficiente de correlação, designado como Coeficiente de Determinação é uma função de

$$R^2 = \frac{\sum \left(Y_i - \bar{Y}\right)^2 - \sum \left(Y_i - \hat{Y}_i\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2} = 1 - \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2} = \frac{\sum \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2}$$

significando a proporção da soma dos quadrados dos desvios que é explicada pela relação linear entre X e Y.

Exemplo 7.7.2: Coeficientes de correlação e de determinação

A tabela do exemplo 7.2.1 apresenta os cálculos auxiliares para a determinação do coeficiente de correlação e de determinação.

X_i	Y_i	$\underbrace{(X_i - \bar{X})}_{x_i}$	$(X_i - \bar{X})^2$	$x_i Y_i$	$\underbrace{(Y_i - \bar{Y})}_{y_i}$	$(Y_i - \bar{Y})^2$	$x_i y_i$
3	1.4	-3.429	11.755	-4.8	-1.029	1.058	3.527
4	1.5	-2.429	5.898	-3.643	-0.929	0.862	2.255
5	2.1	-1.429	2.041	-3	-0.329	0.108	0.469
6	2.4	-0.429	0.184	-1.029	-0.029	0.001	0.012
8	3.1	1.571	2.469	4.871	0.671	0.451	1.055
9	3.2	2.571	6.612	8.229	0.771	0.595	1.984
10	3.3	3.571	12.755	11.786	0.871	0.759	3.112
45	17	0	41.714	12.414	0	3.834	12.414
$\bar{X} = 6.429$	$\bar{Y} = 2.429$						

Assim, os coeficientes podem ser calculados

$$R = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} = \frac{12.414}{\sqrt{(41.714)(3.834)}} = 0.982$$

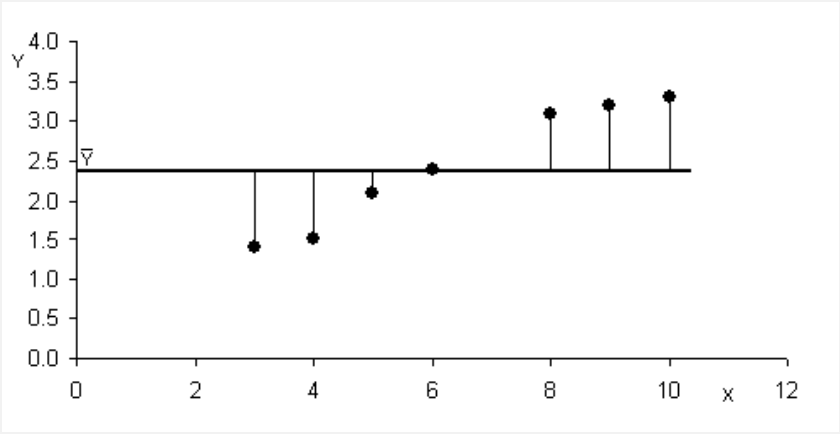
$$\hat{\beta}_1 = R \frac{\sqrt{\sum (Y_i - \bar{Y})^2}}{\sqrt{\sum (X_i - \bar{X})^2}} = 0.982 \frac{\sqrt{3.834}}{\sqrt{41.714}} = 0.298$$

com o coeficiente de determinação dado por

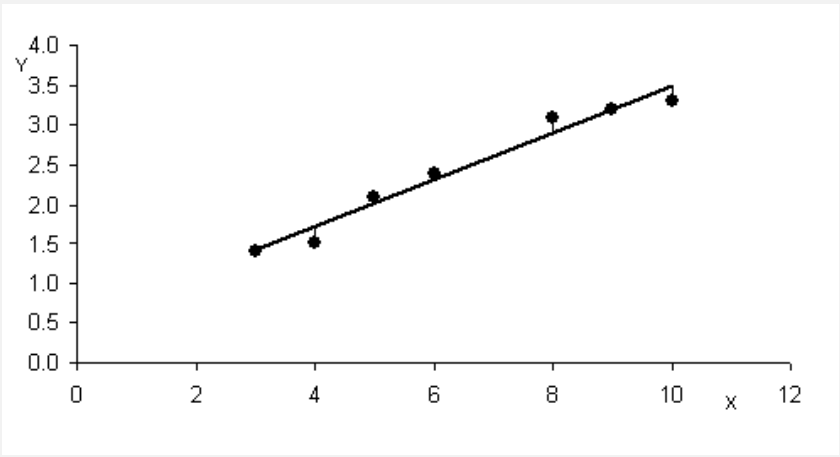
$$R^2 = 1 - \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{0.140}{3.834} = 0.964$$

que corresponde ao quadrado do coeficiente de correlação.

As figuras pretendem demonstrar que se X não contribui em nada para a previsão dos valores de Y, então os desvios em relação a $\bar{Y}$ deveriam ser aproximadamente iguais aos desvios em relação à reta de regressão. Como se pode ver, os desvios relativos à reta de regressão são muito mais pequenos que relativamente aos desvios a $\bar{Y}$, o que permite concluir que o modelo de regressão contribui para a previsão de Y. Na verdade, pode-se afirmar que aproximadamente 98% da variação observada é explicada pelo modelo de regressão linear.



A variável independente não contribui para a explicação dos valores observados.



A variável independente contribui para a explicação dos valores observados.

7.8 Análise da Variância e Regressão

Como foi visto no ponto anterior, a Soma dos Quadrados dos Desvios de cada observação Y_i relativamente à sua média pode ser expressa como

$$\sum \left(Y_i - \bar{Y}\right)^2 = \sum \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum \left(Y_i - \hat{Y}_i\right)^2,$$

isto é,

$$\text{Variação Total} = \text{Variação Explicada} + \text{Variação Não Explicada}.$$

Esta partição da Soma Total dos Quadrados tem semelhanças com a Análise da Variância. Assim, a correspondente tabela ANOVA para a regressão linear pode ser construída como

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	F
Explicada (pela regressão)	$\sum \left(\hat{Y}_i - \bar{Y}\right)^2$	1	$\sum \hat{\beta}_1^2 \left(X_i - \bar{X}\right)^2$	$\frac{\sum \hat{\beta}_1^2 \left(X_i - \bar{X}\right)^2}{S^2}$
Não Explicada (resíduos)	$\sum \left(Y_i - \hat{Y}_i\right)^2$	$n-2$	$S^2 = \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{n-2}$	
Total	$\sum \left(Y_i - \bar{Y}\right)^2$	$n-1$		

Assim, o teste relativo à Hipótese Nula $\beta_1 = 0$ pode ser respondido pela razão entre a variância explicada sobre a não explicada, isto é,

$$F = \frac{\sum \hat{\beta}_1^2 \left(X_i - \bar{X}\right)^2}{S^2}$$

que, para valores suficientemente elevados, conduz à sua rejeição.

Exemplo 7.8.1: Tabela ANOVA

Construa a tabela ANOVA para o exemplo 7.2.1.

Solução

A tabela ANOVA pode ser construída a partir dos valores previamente calculados, isto é,

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média dos Quadrados	F
Explicada (pela regressão)	3.695	1	3.695	132.174
Não Explicada (resíduos)	0.140	5	0.028	
Total	3.834	6		

O teste F da tabela ANOVA é equivalente ao teste t já que as duas estatísticas estão relacionadas, com $F = t^2$ ($132.174 = 11.947^2$).

7.9 Análise de Resíduos

A validação de um modelo de regressão linear exige a realização de inferências acerca do declive β_1, bem como o cálculo do coeficiente de determinação R^2. Contudo, a validação requer, também, uma análise dos resíduos. Deve ser referido que os erros são assumidos como independentes, normalmente distribuídos, com média zero e variância constante. Em geral, identificam-se três grandes fontes de problemas relativas ao modelo de regressão:

- Especificação errada do modelo;

- As assunções relativas aos erros não são válidas;

- Os dados podem conter uma ou mais observações discordantes.

Uma incorreta especificação do modelo significa que a parte determinística do modelo não contém variáveis importantes ou que a relação entre a variável dependente e as variáveis independentes não está corretamente definida. A normalidade dos erros pode ser verificada através do gráfico de probabilidade normal. A assunção de variância constante implica que a distribuição dos resíduos não exiba nenhum tipo de padrão. Por último, valores discordantes, isto é, valores que de alguma forma se destacam do conjunto, podem ou não influenciar as estimativas dos parâmetros. Em geral, observações que se distanciam da média mais do que três desvios padrão devem ser consideradas discordantes. Contudo, em modelos de regressão múltipla, a deteção da discordância pode ser bastante complexa. Modelos baseados em séries temporais podem exibir correlação entre os erros, que, eventualmente, pode ser detetada pela representação gráfica dos resíduos. Assim, como resulta destas considerações, a representação dos diagramas de dispersão dos resíduos em função das variáveis independentes pode permitir detetar alguns dos problemas referidos. Um modelo corretamente especificado, com erros normalmente distribuídos, com média nula e variância constante deve apresentar um diagrama de dispersão dos resíduos em que estes se distribuem aleatoriamente à volta do zero, como especificado na Figura 7.10.

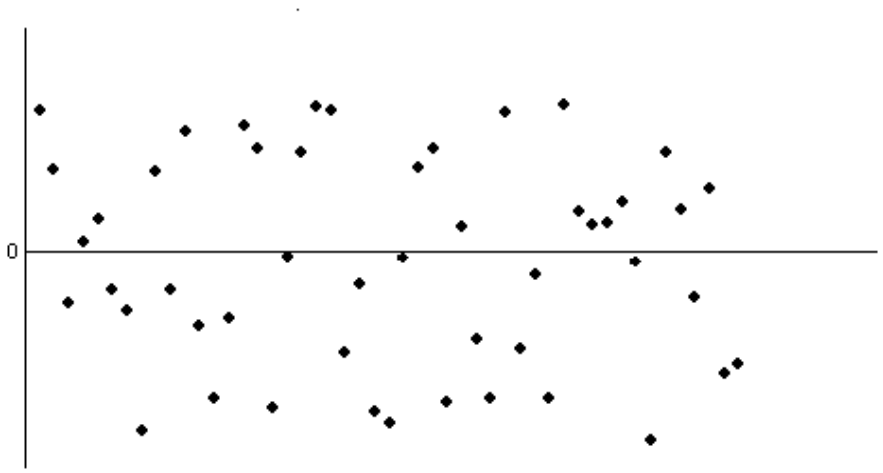


Figura 7.10: Resíduos aleatoriamente distribuídos à volta do zero.

As Figuras 7.11 a 7.13 mostram padrões de resíduos estruturados e, portanto, não conformes às assunções relativas aos erros. Nos dois primeiros casos revelam que a variância não é constante e no terceiro caso, a eventual necessidade de inclusão de um termo quadrático, com a consequente reformulação do modelo.

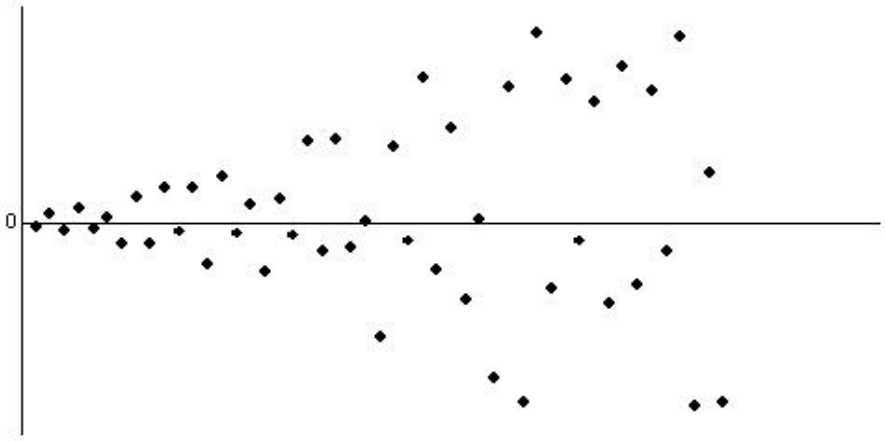


Figura 7.11: Resíduos sem variância constante, em forma de V.

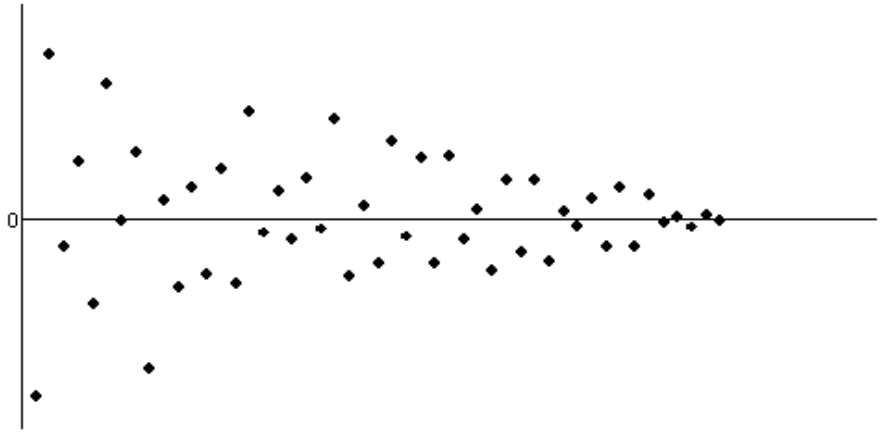


Figura 7.12: Resíduos sem variância constante, em forma de cunha.

7.10 Regressão Não Linear

A definição da relação entre a variável independente e a variável dependente deve ser estabelecida tendo em atenção o diagrama de dispersão de Y em função de X. A Figura 7.14 apresenta algumas funções matemáticas que podem ajudar a determinação da relação entre as variáveis.

Por vezes, a relação entre a variável independente e a variável dependente pode ser não linear. A determinação dos coeficientes do modelo pode ser feita pelo método dos mínimos quadrados, desde que a relação entre a variável independente e os coeficientes do modelo seja linear, como é o caso dos exemplos apresentados na Figura 7.14. Caso tal não se verifique, o sistema de equações resultante é não linear e a sua resolução impõe métodos de otimização não linear. Como exemplo

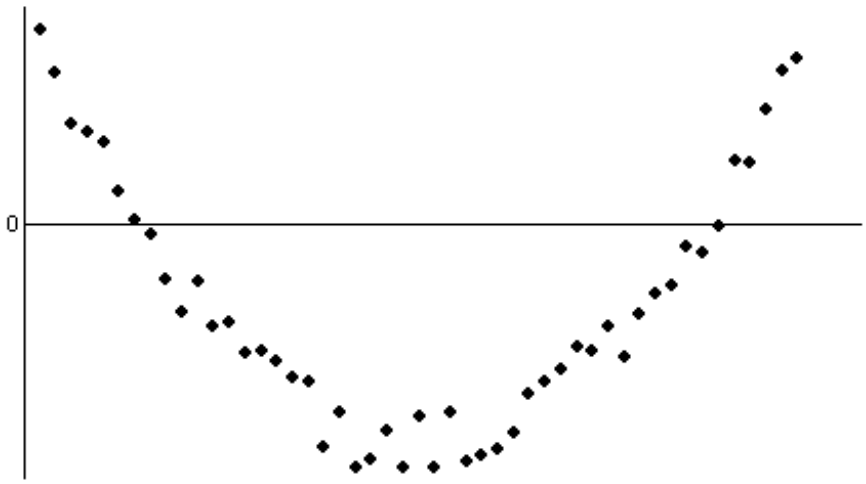


Figura 7.13: Resíduos estruturados.

considere-se o seguinte modelo não linear,

$$Y_i = ae^{-bX_i}u_i,$$

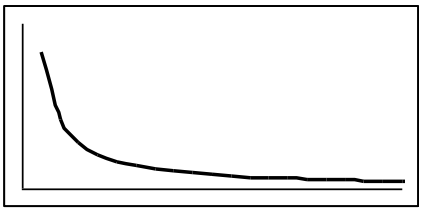
onde u_i representa o termo de erro. A aplicação de logaritmos conduz ao seguinte modelo linear

$$\ln Y_i = \ln a - bX_i + \ln u_i.$$

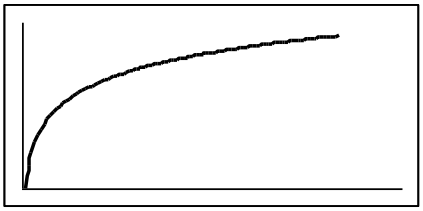
A aplicação do método dos mínimos quadrados exige que o termo de erro no modelo não linear seja multiplicativo para que no modelo transformado, o termo de erro seja aditivo, com média nula e variância constante. Em economia, é muito comum o pressuposto da homogeneidade das variâncias não ser verificado. Por isso, especialmente em situações onde a variância parece crescer com a variável independente, é usual aplicar o método dos mínimos quadrados a transformações da variável independente, como sejam o logaritmo decimal e a raiz quadrada. Tal transformação exige, de toda a maneira, a verificação dos pressupostos impostos à distribuição dos erros no modelo transformado. Deve ser referido, contudo, que a interpretação dos coeficientes nos modelos transformados é muito mais complexa.

7.11 Regressão Múltipla

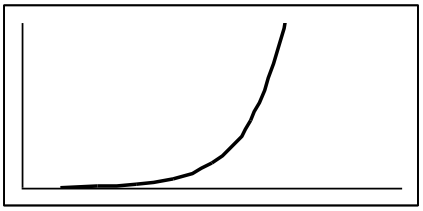
Os modelos apresentados até esta secção consideravam uma única variável independente. Quando o objetivo é investigar a influência de várias variáveis em Y, impõe-se o desenvolvimento de modelos de regressão múltipla. Modelos de regressão múltipla com muitas variáveis independentes só são possíveis de analisar por recurso a pacotes estatísticos, dado que os cálculos são incomportáveis.



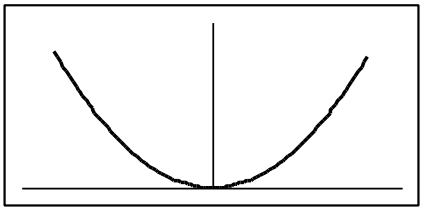
(a) $\hat{Y} = \beta_0 + \beta_1 X$.



(b) $\hat{Y} = \beta_0 + \beta_1 \ln X$.



(c) $\hat{Y} = \beta_0 + \beta_1 e^X$.



(d) $\hat{Y} = \beta_0 + \beta_1 X^2$.

Figura 7.14: Gráficos de algumas relações entre $\hat{Y}$ e X.

A título de exemplo, a dedução das equações para um modelo com duas variáveis independentes é apresentado.

O modelo com duas variáveis independentes é descrito pela seguinte equação

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

onde β_0, β_1 e β_2 são os parâmetros desconhecidos que é necessário determinar. A Soma dos Quadrados dos Resíduos, para os n pontos, é dada por

$$SQR = \sum_{i=1}^{n} \left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i \right) \right]^2.$$

As derivadas parciais da SQR relativamente a β_0, β_1 e β_2, respetivamente, são,

$$\begin{aligned}
\frac{\partial SQR}{\partial \hat{\beta}_0} &= \sum_{i=1}^{n} 2 \left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i \right) \right] (-1) \\
\frac{\partial SQR}{\partial \hat{\beta}_1} &= \sum_{i=1}^{n} 2 \left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i \right) \right] (-X_i) \\
\frac{\partial SQR}{\partial \hat{\beta}_2} &= \sum_{i=1}^{n} 2 \left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 Z_i \right) \right] (-Z_i).
\end{aligned}$$

Igualando a zero as derivadas parciais, e rearranjando os termos das equações, obtém-se um sistema de três equações,

$$\begin{cases}
\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 X_i - \sum_{i=1}^{n} \hat{\beta}_2 Z_i = 0 \\
\sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} \hat{\beta}_0 X_i - \sum_{i=1}^{n} \hat{\beta}_1 X_i^2 - \sum_{i=1}^{n} \hat{\beta}_2 Z_i X_i = 0 \\
\sum_{i=1}^{n} Y_i Z_i - \sum_{i=1}^{n} \hat{\beta}_0 X_i - \sum_{i=1}^{n} \hat{\beta}_1 X_i Z_i - \sum_{i=1}^{n} \hat{\beta}_2 Z_i^2 = 0
\end{cases}$$

$$\begin{cases} \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} Y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} \hat{\beta}_2 Z_i X_i = \sum_{i=1}^{n} Y_i X_i \\ \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i Z_i + \hat{\beta}_2 \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} Y_i Z_i \end{cases}$$

Como já foi referido, a determinação dos coeficientes exige o recurso a suporte informático. Por outro lado, como resulta do sistema de equações obtido, somente uma representação matricial permite tratar modelos com mais variáveis independentes, dado que o número de equações aumenta em uma unidade por cada variável independente adicionada. Assim, para um modelo com k variáveis independentes,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

com n observações, a representação matricial conduz a,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \\ \varepsilon_n \end{bmatrix}$$

ou seja,

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

o que conduz à seguinte representação do sistema de equações,

$$(\mathbf{X}'\mathbf{X})\,\hat{\beta} = \mathbf{X}'\mathbf{Y}$$

com a solução dada por

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

ANEXO 1

Propriedades dos Estimadores de Mínimos Quadrados

Média e Variância de O estimador de β_1 depende dos valores conhecidos da variável independente, assumida como não aleatória.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Portanto, o termo S_{xx} pode ser tratado como uma constante. S_{xy} pode ser reescrito como

$$\begin{aligned} S_{xy} &= \sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = \sum \left[\left(X_i - \bar{X}\right) Y_i - \left(X_i - \bar{X}\right)\bar{Y}\right] \\ &= \sum \left(X_i - \bar{X}\right) Y_i - \bar{Y} \sum \left(X_i - \bar{X}\right) = \sum \left(X_i - \bar{X}\right) Y_i \end{aligned}$$

dado que $|sum\left(X_i - \bar{X}\right) = 0$.

Assim, a expressão para o estimador de β_1 toma agora a forma,

$$\hat{\beta}_1 = \frac{1}{S_{xx}} \sum \left(X_i - \bar{X}\right) Y_i = \frac{\left(X_1 - \bar{X}\right)}{S_{xx}} Y_1 + \frac{\left(X_2 - \bar{X}\right)}{S_{xx}} Y_2 + \ldots + \frac{\left(X_n - \bar{X}\right)}{S_{xx}} Y_n,$$

ou seja, β_1 pode ser escrito como uma combinação linear das variáveis aleatórias Y_i,

$$\hat{\beta}_1 = \sum \omega_{1,i} Y_i \qquad\qquad \omega_{1,i} = \frac{\left(X_i - \bar{X}\right)}{S_{xx}} \ .$$

Assim sendo, o valor esperado é dado por

$$E\left[\hat{\beta}_1\right] = E\left[\frac{\left(X_1 - \bar{X}\right)}{S_{xx}} Y_1 + \frac{\left(X_2 - \bar{X}\right)}{S_{xx}} Y_2 + \ldots + \frac{\left(X_n - \bar{X}\right)}{S_{xx}} Y_n\right],$$

isto é,

$$E\left[\hat{\beta}_1\right] = \frac{\left(X_1 - \bar{X}\right)}{S_{xx}} E\left[Y_1\right] + \frac{\left(X_2 - \bar{X}\right)}{S_{xx}} E\left[Y_2\right] + \ldots + \frac{\left(X_n - \bar{X}\right)}{S_{xx}} E\left[Y_n\right].$$

Atendendo a que

$$E\left[Y\right] = \beta_0 + \beta_1 X + E\left[\varepsilon\right] = \beta_0 + \beta_1 X$$

$$E\left[\hat{\beta}_1\right] = \frac{\left(X_1 - \bar{X}\right)}{S_{xx}} \left(\beta_0 + \beta_1 X_1\right) + \frac{\left(X_2 - \bar{X}\right)}{S_{xx}} \left(\beta_0 + \beta_1 X_2\right) + \ldots + \frac{\left(X_n - \bar{X}\right)}{S_{xx}} \left(\beta_0 + \beta_1 X_n\right),$$

ou seja,

$$E\left[\hat{\beta}_1\right] = \frac{\beta_0}{S_{xx}} \sum \left(X_i - \bar{X}\right) + \frac{\beta_1}{S_{xx}} \sum \left(X_i - \bar{X}\right) X_i.$$

Como S_{xx} pode ser expresso da seguinte forma

$$\begin{aligned} S_{xx} &= \sum \left(X_i - \bar{X}\right)^2 = \sum \left[\left(X_i - \bar{X}\right) X_i - \left(X_i - \bar{X}\right)\bar{X}\right] \\ &= \sum \left[\left(X_i - \bar{X}\right) X_i\right] - \bar{X} \sum \left(X_i - \bar{X}\right) = \sum \left(X_i - \bar{X}\right) X_i \end{aligned}$$

a esperança de $\hat{\beta}_1$ é dada por

$$E\left[\hat{\beta}_1\right] = 0 + \frac{\beta_1}{S_{xx}} S_{xx} = \beta_1$$

e, portanto, $\hat{\beta}_1$ é um estimador não enviesado de β_1.

A variância $\hat{\beta}_1$, tendo em atenção que as variáveis aleatórias Y_i são independentes, o que implica que a covariância entre quaisquer dois pares é nula, é dada por

$$\begin{aligned} &Var\left[\hat{\beta}_1\right] = \frac{\left(X_1 - \bar{X}\right)^2}{\left(S_{xx}\right)^2} Var\left[Y_1\right] + \frac{\left(X_2 - \bar{X}\right)^2}{\left(S_{xx}\right)^2} Var\left[Y_2\right] + \ldots + \frac{\left(X_n - \bar{X}\right)^2}{\left(S_{xx}\right)^2} Var\left[Y_n\right] \\ &= \sigma^2 \frac{\sum \left(X_i - \bar{X}\right)^2}{\left(S_{xx}\right)^2} \end{aligned}$$

isto é,

$$Var\left[\hat{\beta}_1\right] = \frac{\sigma^2}{S_{xx}}.$$

Média e Variância de β_0

O estimador de β_0 é dado pela seguinte expressão

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

que, tomando em consideração que β_1 pode ser expresso como

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{S_{xx}} = \frac{\sum \left(X_i - \bar{X}\right) Y_i - \bar{Y} \sum \left(X_i - \bar{X}\right)}{S_{xx}} = \frac{\sum \left(X_i - \bar{X}\right) Y_i}{S_{xx}}$$

logo β_0 fica como

$$\hat{\beta}_0 = \frac{\sum Y_i}{n} - \frac{\sum \left(X_i - \bar{X}\right) Y_i}{S_{xx}} \bar{X} = \sum \left[\frac{1}{n} - \frac{\bar{X}\left(X_i - \bar{X}\right)}{S_{xx}} \right] Y_i = \sum \omega_{0,i} Y_i.$$

O valor esperado de $\hat{\beta}_0$ pode ser expresso como

$$\begin{aligned} E\left[\hat{\beta}_0\right] &= \sum \left[\tfrac{1}{n} - \tfrac{\bar{X}\left(X_i - \bar{X}\right)}{S_{xx}} \right] E\left[Y_i\right] = \sum \left[\tfrac{1}{n} - \tfrac{\bar{X}\left(X_i - \bar{X}\right)}{S_{xx}} \right] \left(\beta_0 + \beta_1 X_i\right) \\ &= \sum \tfrac{1}{n} \beta_0 - \beta_0 \tfrac{\bar{X} \sum \left(X_i - \bar{X}\right)}{S_{xx}} - \tfrac{1}{n} \beta_1 \sum X_i + \beta_1 \tfrac{\bar{X} \sum \left(X_i - \bar{X}\right) X_i}{S_{xx}} \end{aligned}$$

donde se conclui que

$$E\left[\hat{\beta}_0\right] = \beta_0.$$

Como $\hat{\beta}_0$ pode ser expresso como uma combinação linear de variáveis independentes, a sua variância é dada por

$$Var\left[\hat{\beta}_0\right] = Var\left\{ \sum \left[\frac{1}{n} Y_i - \frac{\bar{X}\left(X_i - \bar{X}\right)}{S_{xx}} Y_i \right] \right\} = \sum \left[\frac{1}{n^2} Var\left[Y_i\right] + \left[\frac{\bar{X}\left(X_i - \bar{X}\right)}{S_{xx}} \right]^2 Var\left[Y_i\right] \right]$$

e, portanto,

$$Var\left[\hat{\beta}_0\right] = \frac{1}{n^2} \sum Var\left[Y_i\right] + \frac{\bar{X}^2}{S_{xx}^2} \sum \left(X_i - \bar{X}\right)^2 Var\left[Y_i\right]$$

e, por fim

$$Var\left[\hat{\beta}_0\right] = \frac{1}{n^2} n\sigma^2 + \frac{\bar{X}^2}{S_{xx}} \sigma^2 = \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} \right) \sigma^2.$$

Os coeficientes $\hat{\beta}_0$ e $\hat{\beta}_1$ podem ser formulados, como foi visto, como combinações lineares das variáveis aleatórias Y_i. Assim,

$$\hat{\beta}_0 = \sum \left[\frac{1}{n} - \frac{\bar{X}\left(X_i - \bar{X}\right)}{S_{xx}} \right] Y_i = \sum \omega_{0,i} Y_i$$

$$\hat{\beta}_1 = \frac{1}{S_{xx}} \sum \left(X_i - \bar{X}\right) Y_i = \sum \omega_{1,i} Y_i$$

Resulta, portanto, que as estimativas de $\hat{\beta}_0$ e de $\hat{\beta}_1$ não são independentes. Atendendo a que a variância destes estimadores pode ser expressa como

$$Var\left[\hat{\beta}_0\right] = \sigma^2 \sum \omega_{0,i}^2$$

$$Var\left[\hat{\beta}_1\right] = \sigma^2 \sum \omega_{1,i}^2$$

a covariância será dada por

$$Cov\left[\hat{\beta}_0, \hat{\beta}_1\right] = \sigma^2 \sum \omega_{0,i}\omega_{1,i},$$

o que se traduz em

$$\begin{aligned}\sum \omega_{0,i}\omega_{1,i} &= \sum \left(\tfrac{1}{n} - \tfrac{\bar{X}\left(X_i - \bar{X}\right)}{S_{xx}}\right)\left(\tfrac{1}{S_{xx}} \sum \left(X_i - \bar{X}\right)\right) \\ &= \tfrac{1}{nS_{xx}} \sum \left(X_i - \bar{X}\right) - \tfrac{\bar{X}}{S_{xx}} \sum \left(X_i - \bar{X}\right)^2 = -\tfrac{\bar{X}}{S_{xx}}\end{aligned}$$

resultando na seguinte expressão para a covariância

$$Cov\left[\hat{\beta}_0, \hat{\beta}_1\right] = \frac{-\sigma^2 \sum_{i=1}^{n} X_i}{n \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \sigma^2 \left(-\frac{\bar{X}}{S_{xx}}\right).$$

ANEXO 2

Dedução das equações com variáveis desvio

O modelo de regressão linear também pode ser determinado usando como variável independente o desvio de relativamente à sua média $\left(X_i - \bar{X}\right)$. Definindo este desvio como x_i, e considerando o modelo

$$Y_i = \alpha + \beta_1 \left(X_i - \bar{X}\right) = \alpha + \beta_1 x_i,$$

as derivadas parciais em ordem aos parâmetros podem ser expressas como

$$\begin{aligned}\tfrac{\partial SQR}{\partial \hat{\alpha}} &= \sum_{i=1}^{n} 2\left[Y_i - \left(\hat{\alpha} + \hat{\beta}_1 x_i\right)\right](-1) \\ \tfrac{\partial SQR}{\partial \hat{\beta}_1} &= \sum_{i=1}^{n} 2\left[Y_i - \left(\hat{\alpha} + \hat{\beta}_1 x_i\right)\right](-x_i)\end{aligned}$$

obtendo-se as seguintes equações

$$\begin{aligned}\hat{\alpha} n + \sum_{i=1}^{n} \hat{\beta}_1 x_i &= \sum_{i=1}^{n} Y_i \\ \hat{\alpha} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} Y_i x_i\end{aligned}.$$

Contudo, atendendo a que

$$\sum_{i=1}^{n} \left(X_i - \bar{X}\right) = 0$$

as equações tomam a forma

$$\begin{aligned}\hat{\alpha} &= \frac{\sum_{i=1}^{n} Y_i}{n} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}\end{aligned}$$

com a reta de mínimos quadrados dada por

$$\hat{Y} = \alpha + \beta_1 \left(X - \bar{X}\right).$$

Graficamente, este último modelo corresponde a uma translação do eixos dos Y.

Exemplo 7.11.1: Reta de mínimos quadrados com variáveis desvio

Para os dados do exemplo 7.2.2, determine a reta de mínimos quadrados com base nas equações deduzidas para variáveis desvio.

Solução

A tabela apresenta os cálculos auxiliares.

X_i	Y_i	$\left(X_i - \bar{X}\right)$	x_i^2	Y_i
3	1.4	-3.429	11.755	-4.8
4	1.5	-2.429	5.898	-3.643
5	2.1	-1.429	2.041	-3
6	2.4	-0.429	0.184	-1.029
8	3.1	1.571	2.469	4.871
9	3.2	2.571	6.612	8.229
10	3.3	3.571	12.755	11.786
45	17	0	41.714	12.414

As estimativas dos coeficientes são

$$\alpha = \tfrac{17}{7} = 2.429$$
$$\beta_1 = \tfrac{12.414}{41.714} = 0.298$$

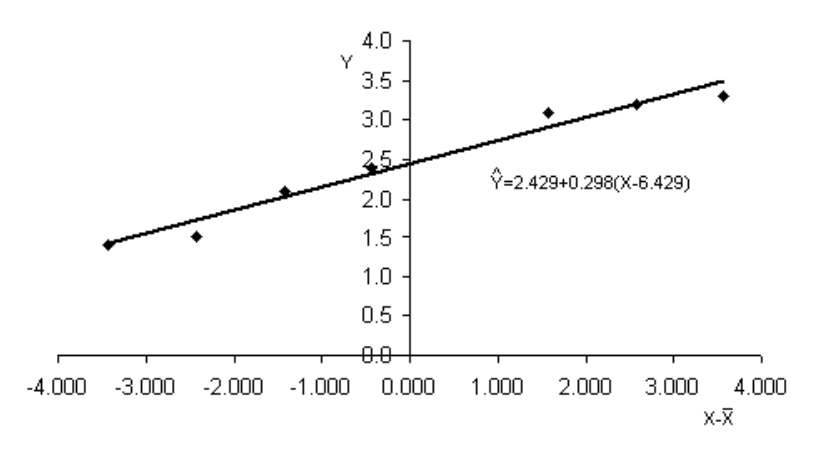
A reta de mínimos quadrados é

$$\hat{Y} = 2.429 + 0.298\,(X - 6.429)$$

Para $X = 7$, obtém-se o seguinte valor para a resistência,

$$\hat{Y} = 2.429 + 0.298\,(7 - 6.429) = 2.599$$

A seguinte figura apresenta a reta ajustada.



Ajuste segundo a reta $\hat{Y} = 2.429 + 0.298\,(X - 6.429)$**.**

Exercícios

1. A anestesia intravenosa com propofol deve conduzir à estabilidade do sistema cardiovascular. Um estudo foi conduzido em que cães foram anestesiados com quantidades crescentes de propofol, sendo registados a pressão arterial média e a frequência cardíaca.

   A tabela apresenta os resultados de um dos animais. Ajuste um modelo linear, pressão média em função da dose. De seguida, repita o exercício usando como variável independente o logaritmo da dose. Compare os dois modelos com base no coeficiente de determinação e na análise de resíduos.

Propofol	Pressão média
3.458	99
4.036	88
4.861	85
5.932	76
6.920	75
8.403	75
8.778	80
11.406	64

   Silva, A., Ribeiro, L.M., Bressan, N., Oliveira, P., Ferreira, D.A., Antunes, L.M. (2011). Dogs mean arterial pressure and heart rate responses during high Propofol plasma concentrations estimated by a pharmacokinetic model, Res Vet Sci., 91(2):278-80.

2. A tabela apresenta a Taxa de Mortalidade Infantil (TMI) em Portugal (número de óbitos de crianças com menos de um ano de vida, num ano, tendo por referência o número de nados vivos nesse mesmo ano) desde 1960 (dados da PORDATA).

   Portugal apresentou uma das maiores reduções, ocupando hoje um lugar de topo no conjunto de países com mais baixas taxas de mortalidade.

Ano	TMI
1960	77.5
1965	64.9
1970	55.5
1975	38.9
1980	24.3
1985	17.8
1990	10.9
1995	7.4
2000	5.5
2005	3.5
2010	2.5
2015	2.9

   Ajuste um modelo linear TMI em função do ano.

   De seguida repita o exercício considerando um modelo exponencial

$$\hat{Y} = ae^{bX}.$$

   Compare os dois modelos com base no coeficiente de determinação e na análise dos resíduos.

3. A tabela apresenta a distância percorrida em 6 minutos para 158 sujeitos. Sabe-se que esta distância percorrida, uma forma de avaliar a capacidade respiratória, depende da idade, do Índice de Massa Corporal (kg/m^2) e do sexo (codificado como variável binária).

   Ajuste um modelo de regressão múltipla e interprete os coeficientes. Comente os resultados em termos do coeficiente de determinação e distribuição dos resíduos.

Idade	Índice de Massa Corporal	Sexo	Distância percorrida
33	21.83	Female	720.0
31	26.3	Female	619.0
53	29.44	Female	600.0
28	23.23	Female	630.0
27	23.99	Female	646.5
27	24.5	Female	642.0
51	21.85	Female	540.0
41	25.19	Female	615.0
35	19.45	Female	624.0
29	18.9	Female	645.0
30	21.8	Female	754.5
43	20.16	Female	702.0
22	18.87	Female	693.0
29	19.85	Female	680.5
50	21.97	Female	717.0
45	22.48	Female	683.5
27	20.77	Female	717.0
30	24.28	Female	498.0
41	22.5	Female	649.0
45	22.58	Female	651.0
62	25.87	Female	661.0
32	20.34	Female	708.0
26	23.64	Female	591.0
40	22.82	Female	562.0
40	21.71	Female	630.0
24	27.51	Female	579.0
22	23.94	Female	660.0
23	21.78	Female	720.0
24	21.83	Female	597.0
23	21.5	Female	702.0
23	21.83	Female	663.0
21	20.2	Female	684.0
31	21.11	Female	630.0
66	27.41	Female	453.0
33	22.31	Female	603.0
22	22.66	Female	750.0

Idade	Índice de Massa Corporal	Sexo	Distância percorrida
58	28.95	Female	633.0
70	24.56	Female	549.0
18	24.31	Female	645.0
61	24.98	Female	681.0
63	28.44	Female	504.0
33	29.37	Female	549.0
33	19.2	Female	603.0
46	24.28	Female	570.0
52	28.55	Female	630.0
43	21.61	Female	624.0
48	22.86	Female	534.0
54	22.66	Female	570.0
53	21.19	Female	594.0
30	26.22	Female	594.0
50	26.67	Female	623.0
31	21.77	Female	612.0
54	29.69	Female	579.0
63	29.52	Female	567.0
27	18.14	Female	489.0
36	22.5	Female	660.0
45	23.23	Female	591.0
68	21.36	Female	471.0
28	20.9	Female	570.0
48	22.58	Female	564.0
46	22.64	Female	531.0
23	19.92	Female	570.0
33	28.28	Female	558.0
43	21.23	Female	573.0
54	30.33	Female	492.0
31	23.34	Female	660.0
38	21.63	Female	684.0
41	25.48	Female	513.0
64	22.35	Female	492.0
62	29.34	Female	501.0
25	21.08	Female	651.0
25	22.06	Female	588.0

Idade	Índice de Massa Corporal	Sexo	Distância percorrida
33	21.01	Female	570.0
38	28.33	Female	558.0
53	25.59	Female	603.0
50	22.1	Female	584.0
52	25.15	Female	549.0
31	20.08	Female	564.0
38	19.07	Female	552.0
47	22.66	Female	465.0
28	22.06	Female	615.0
41	21.09	Female	543.0
58	20.31	Female	606.0
50	21.37	Female	618.0
41	21.23	Female	603.0
35	24.17	Female	606.0
52	25.06	Female	602.0
30	19	Female	595.0
43	24.65	Female	535.0
26	22.41	Male	811.5
53	23.17	Male	717.0
34	21.26	Male	827.0
21	27.66	Male	621.0
25	19.61	Male	741.0
29	19.93	Male	744.0
47	27.78	Male	642.0
34	19.88	Male	630.0
33	22.53	Male	831.0
59	24.96	Male	600.0
30	29.8	Male	645.0
40	28.41	Male	672.0
28	24.3	Male	735.0
40	27.76	Male	666.0
45	30.02	Male	600.0
40	26.3	Male	669.0
40	22.2	Male	636.0
40	23.53	Male	627.0
37	25.22	Male	654.0

Idade	Índice de Massa Corporal	Sexo	Distância percorrida
60	24.51	Male	597.0
31	25.77	Male	750.0
20	20.76	Male	729.0
27	21.55	Male	699.0
36	18.04	Male	735.0
26	23.84	Male	606.0
46	26.03	Male	567.0
33	29.99	Male	630.0
33	23.89	Male	594.0
31	28.34	Male	651.0
65	22.99	Male	540.0
41	22.53	Male	723.0
28	19.31	Male	684.0
52	30.09	Male	690.0
25	25.56	Male	632.0
36	23.6	Male	648.0
31	28.08	Male	690.0
48	23.15	Male	726.0
56	27.08	Male	666.0
45	27.68	Male	537.0
55	25.21	Male	642.0
59	25.98	Male	627.0
32	23.15	Male	660.0
40	27.68	Male	636.0
37	28.7	Male	713.0
50	22.2	Male	672.0
54	26.15	Male	680.0
70	24.98	Male	510.0
36	23.66	Male	654.0
37	27.68	Male	633.0
67	25.95	Male	582.0
22	19.6	Male	714.0
65	24.39	Male	560.0
22	22.53	Male	715.0
23	20.28	Male	548.0
24	25.61	Male	630.0

Idade	Índice de Massa Corporal	Sexo	Distância percorrida
26	22.04	Male	655.0
20	24.66	Male	705.0
21	20.99	Male	690.0
23	21.8	Male	780.0
23	21.63	Male	670.0
27	21.97	Male	710.0
33	25.14	Male	612.0
37	23.85	Male	670.0
29	29.41	Male	668.0
48	23.71	Male	600.0
62	28.73	Male	471.0
33	29.06	Male	530.0
32	21.3	Male	650.0
37	23.15	Male	640.0

4. A tabela apresenta o peso (g) ao nascer de 208 bebés do sexo masculino em função da idade gestacional. Ajuste um modelo linear e também um modelo polinomial de segundo grau, com o quadrado da idade gestacional. Compare os dois modelos. O que pode concluir?

Idade	Peso	Idade	Peso	Idade	Peso	Idade	Peso
40	1907	40	1886	41	1943	40	1782
40	1949	39	2038	38	1898	40	1804
40	1830	38	1827	41	1751	40	1743
39	2039	40	2032	34	1611	40	1956
41	2070	38	1838	40	1964	39	1692
38	1802	39	1983	38	2117	36	1640
38	1995	40	1917	37	1941	38	2010
40	1733	40	1835	36	1804	40	1999
38	1954	39	1931	40	1862	42	1808
28	1154	40	1754	41	1702	41	2049
39	1988	40	1838	39	1996	30	1386
42	2089	41	1847	39	1802	41	2027
40	1970	40	2035	32	1477	40	1787
39	1819	40	2001	36	2001	41	1999
39	1940	39	1964	35	1664	39	1851

Idade	Peso	Idade	Peso	Idade	Peso	Idade	Peso
40	1721	41	1814	38	1846	37	1850
33	1662	38	1952	39	1904	41	1899
40	1859	41	1955	39	1943	40	1929
35	1801	40	2058	40	1854	39	2003
40	1753	38	1749	39	1848	40	2025
40	1847	38	1895	38	1888	38	1878
38	1871	30	1223	41	1909	40	1890
39	1896	39	1873	38	1838	39	1860
26	900	38	1802	38	1859	40	1897
39	1805	40	1868	39	1832	39	2100
39	2024	31	1534	32	1652	40	1916
40	1958	40	1881	37	1757	40	2134
38	1759	40	1983	41	1770	33	1582
41	1938	41	1918	40	1972	40	1897
27	872	42	2038	40	1906	38	1879
39	1839	38	1959	39	1983	39	1904
39	2092	40	2077	38	2078	40	1865
40	2035	40	1888	40	1871	34	1553
40	1935	40	1863	40	1882	40	1771
40	2064	39	1896	38	1971	39	1991
38	1898	38	1808	40	1843	39	2048
39	1869	39	1866	39	1913	38	1891
37	1882	39	1779	40	1572	39	1676
39	1705	40	2018	38	2049	41	1937
40	1912	41	1819	40	2006	37	1818
40	1910	32	1313	34	1563	36	1736
41	2023	39	1979	41	1989	38	1925
39	2128	40	1884	40	1748	36	1837
39	1974	27	940	39	1989	41	1891
40	1849	39	1904	39	1822	41	1887
39	1780	40	2031	40	1820	35	1710
39	1922	39	1955	40	1857	40	1900
38	1961	39	1924	38	1889	34	1596
40	2023	41	1713	29	1101	39	1984
40	1806	40	1880	39	1880	38	1735
37	1808	40	1872	41	1828	39	1770
39	1806	39	1768	35	1645	39	1933

Soluções

1.
```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    24.82263    4.07296   6.094 0.000888 ***
propofol$T_art -0.22552    0.05038  -4.477 0.004207 **
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           87.451     16.889   5.178  0.00206 **
log(propofol$T_art)  -18.440      3.857  -4.782  0.00306 **
```

```
propofol = read.csv("propofol.csv", header=T)
propofol
names(propofol)[1]="propofol"
propofol
plot(propofol$propofol~propofol$T_art)
abline(lm(propofol$propofol~propofol$T_art))
model_1=lm(propofol$propofol~propofol$T_art)
summary(model_1)
plot(model_1,1:2)

plot(propofol$propofol~log(propofol$T_art))
abline(lm(propofol$propofol~log(propofol$T_art)))
model_2=lm(propofol$propofol~log(propofol$T_art))
summary(model_2)
plot(model_2,1:2)
```

2.
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2765.1030   343.5777   8.048 1.12e-05 ***
tmi$ano       -1.3782     0.1729  -7.973 1.21e-05 ***
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 139.721053   5.618196   24.87 2.53e-10 ***
tmi$ano      -0.068975   0.002827  -24.40 3.05e-10 ***
```

```
tmi = read.csv("tmi.csv", header=T)
tmi
names(tmi)[1]="ano"
tmi

plot(tmi$tmi~tmi$ano)
abline(lm(tmi$tmi~tmi$ano))
model_1=lm(tmi$tmi~tmi$ano)
summary(model_1)
plot(model_1,1:2)

plot(log(tmi$tmi)~tmi$ano)
abline(lm(log(tmi$tmi)~tmi$ano))
model_2=lm(log(tmi$tmi)~tmi$ano)
summary(model_2)
plot(model_2,1:2)
```

3.
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 726.6280    16.4113  44.276  < 2e-16 ***
min_6$idade  -2.5338     0.3991  -6.349 2.25e-09 ***

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  761.073     44.788  16.993  < 2e-16 ***
min_6$imc     -5.597      1.866  -2.999  0.00315 **

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 769.4793    41.0495  18.745  < 2e-16 ***
min_6$idade  -2.3657     0.4252  -5.564 1.13e-07 ***
min_6$imc    -2.0754     1.8227  -1.139    0.257

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 787.1454    37.7891  20.830  < 2e-16 ***
```

```
min_6$idade  -2.0256     0.3949  -5.130 8.60e-07 ***
min_6$imc    -4.3786     1.7235  -2.540   0.0121 *
min_6$sexo   54.7302     9.9539   5.498 1.56e-07 ***

min_6 = read.csv("six_min.csv", header=T)
min_6
names(min_6)[1]="idade"
min_6

plot(min_6$dist~min_6$idade)
abline(lm(min_6$dist~min_6$idade))
plot(min_6$dist~min_6$imc)
abline(lm(min_6$dist~min_6$imc))
boxplot(min_6$dist~min_6$sexo)

model_1=lm(min_6$dist~min_6$idade)
summary(model_1)
plot(model_1,1:2)

model_2=lm(min_6$dist~min_6$imc)
summary(model_2)
plot(model_2,1:2)

model_3=lm(min_6$dist~min_6$idade+min_6$imc)
summary(model_3)
plot(model_3,1:2)

model_4=lm(min_6$dist~min_6$idade + min_6$imc + min_6$sexo)
summary(model_4)
plot(model_4,1:2)
```

4.
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -325.662    116.978  -2.784  0.00587 **
peso_rn$sem   56.401      3.025  18.644  < 2e-16 ***

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6620.274    722.509  -9.163  < 2e-16 ***
peso_rn$sem   418.377     41.230  10.147  < 2e-16 ***
peso_2         -5.128      0.583  -8.797 5.82e-16 ***
```

```
peso_rn = read.csv("peso_rn.csv", header=T)
peso_rn
names(peso_rn)[1]="sem"
peso_rn

plot(peso_rn$peso~peso_rn$sem)
abline(lm(peso_rn$peso~peso_rn$sem))
model_1=lm(peso_rn$peso~peso_rn$sem)
summary(model_1)
plot(model_1,1:2)

plot(peso_rn$peso~peso_rn$sem)
abline(lm(peso_rn$peso~peso_rn$sem))
peso_2=peso_rn$sem^2
peso_2
peso_rn$sem

model_2=lm(peso_rn$peso~peso_rn$sem + peso_2)
summary(model_2)
plot(model_2,1:2)
```

Capítulo 8

Estatística Não Paramétrica

8.1 Introdução

De uma forma geral, os métodos estatísticos relativos a intervalos de confiança e testes de hipótese têm subjacente a assunção da normalidade, isto é, que as amostras aleatórias são retiradas de populações normais. Por esta razão, estes métodos são referidos como métodos paramétricos, dado que a maioria dos parâmetros estão especificados. Assim, por exemplo, no teste t para a comparação de médias, é assumido que as populações são aproximadamente normais com variâncias iguais. Contudo, se a forma das distribuições de onde as amostras independentes são retiradas é desconhecida, subsiste, na mesma, a questão de saber se as duas distribuições são ou não idênticas. Apesar de não haver uma definição consensual, os métodos não paramétricos não impõem nenhuma condição relativa à distribuição dos dados. Estes métodos são matematicamente simples, de execução fácil e aplicam-se não só a dados quantitativos mas, também, a dados de natureza nominal e ordinal, como por exemplo, respostas categóricas do tipo sim não, ou respostas em graduações. Os procedimentos apresentados constituem alternativas aos métodos paramétricos usuais e, por essa razão, têm sido objeto de comparação quer em populações normais quer em populações não normais. Em geral, os métodos não paramétricos são menos eficientes em presença de populações normais e, neste caso, os métodos paramétricos são preferíveis. Contudo, o seu uso é aconselhável quando se verificam grandes afastamentos da normalidade.

Alguns dos métodos não paramétricos são baseados em graduações, isto é, em vez de trabalharem com as observações, trabalham com as graduações. Em situações onde as variáveis observadas são de natureza ordinal, os valores em si mesmo não têm significado a não ser pela ordem que estabelecem. Por essa razão, a utilização de graduações não altera a informação original. Por outro lado, se as observações seguem uma qualquer distribuição não normal, a distribuição da estatística de teste pode ser desconhecida enquanto que a distribuição das estatísticas baseadas

em graduações pode ser mais facilmente determinada em muitos casos.

8.2 Teste a duas amostras independentes

Os dados representam duas amostras aleatórias em que $x_{11}, x_{12}, \ldots, x_{n_1 1}$ constituem as n_1 observações da amostra aleatória da população 1, e $x_{12}, x_{22}, \ldots, x_{n_2}$ as n_{22} observações da amostra aleatória retirada da população 2. $R(x_{i1})$ e $R(x_{j2})$ representam às graduações atribuídas a cada uma das $n_1 + n_2 = N$ observações x_{i1}, x_{j2}.

Definição 8.2.1: Teste de Mann-Whitney

O cálculo da estatística de teste baseia-se na soma das graduações das duas amostras,

$$R_1 = \sum_{i=1}^{n_1} R(x_{i1}) \qquad R_2 = \sum_{j=1}^{n_2} R(x_{j1})$$

entre as quais existe a seguinte relação, tendo em atenção que $N = n_1 + n_2$,

$$R = R_1 + R_2 = \frac{N(N+1)}{2}.$$

As estatísticas U_1 e U_2 são definidas como

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \qquad U_2 = n_2 n_1 + \frac{n_2(n_2+1)}{2} - R_2$$

$$U_1 = n_1 n_2 - U_2.$$

Sejam $F(x)$ e $G(x)$ as distribuições correspondentes às duas populações, o teste bilateral corresponde a testar as seguintes hipóteses:

$$\begin{aligned} H_0 &: F(x) = G(x) \quad \text{para todo o } x \\ H_1 &: F(x) \neq G(x) \quad \text{para pelo menos um valor de } x \end{aligned}$$

Assumindo que existe uma diferença na localização das distribuições, a diferença entre as distribuições implica que $P(X_1 < X_2)$, os testes de hipóteses bilateral e unilaterais podem ser escritos como

Teste Unilateral	Teste Bilateral	Teste Unilateral
$H_0 : P(X_1 < X_2) \leq \frac{1}{2}$	$H_0 : P(X_1 < X_2) = \frac{1}{2}$	$H_0 : P(X_1 < X_2) \geq \frac{1}{2}$
$H_1 : P(X_1 < X_2) > \frac{1}{2}$	$H_1 : P(X_1 < X_2) \neq \frac{1}{2}$	$H_1 : P(X_1 < X_2) < \frac{1}{2}$
Estatística		
$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$	$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$	$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$
Região de Rejeição		
$U_1 < w_{1-\alpha}$	$U_1 < w_{1-\alpha/2}$	$U_1 > w_{1-\alpha}$

Os valores críticos são retirados da tabela dos quantis da estatística de Mann-Whitney, para o nível de significância especificado α. O teste assume que os dados constituem amostras aleatórias independentes e que a escala de medida é pelo menos ordinal.

Se houver muitas observações iguais (empates) a estatística pode ser modificada para acomodar esta situação de acordo com

$$U_1 = \frac{\frac{R_1 - n_1(N+1)}{2}}{\sqrt{\frac{n_1 n_2}{N(N-1)} R^2 - \frac{n_1 n_2 (N+1)^2}{4(N-1)}}}.$$

em que $R^2 = \sum_{i=1}^{n_1} R\left(x_{i1}\right)^2 + \sum_{j=1}^{n_2} R\left(x_{j1}\right)^2$.

Exemplo 8.2.1: Teste de Mann-Whitney

A tabela apresenta os valores de colesterol de duas amostras aleatórias de homens, com idade compreendidas entre 50 e 59 anos, tratados com dois medicamentos diferentes. Verifique se há diferenças entre o teor de colesterol nas duas amostras.

Med. 1	Med. 2
243	205
231	203
225	198
224	195
220	185
213	
200	

Solução

1. Hipóteses

   $H_0 : F(x) = G(x)$ para todo o x

   $H_1 : F(x) \neq G(x)$ para pelo menos um valor de x

2. Região crítica

   $$U_1 < w_{1-0.025} = 34$$

3. Estatística

Med. 1	Med. 2	Grad. M1	Grad. M2
243	205	1	7
231	203	2	8
225	198	3	10
224	195	4	11
220	185	5	12
213		6	
200		9	
		$R_1 = 30,\ U_1 = 33$	$R_2 = 48,\ U_2 = 2$

$$n_1 = 7 \quad n_2 = 5 \quad N = 12$$

$$U_1 = 7 \times 5 + \frac{7(7+1)}{2} - 30 = 33 \qquad U_2 = 5 \times 7 + \frac{5(6+1)}{2} - 48 = 2$$

$$U_1 < w_{1-0.025} = w_{0.975} = 34$$

4. Decisão

   Rejeitar H_0, ou seja, os medicamentos são diferentes em termos da redução dos valores de colesterol.

8.3 Teste a duas amostras emparelhadas

Os dados representam pares de observações $(x_{11}, x_{21}), (x_{12}, x_{22}), \ldots, (x_{1n}, x_{2n})$, retiradas de uma distribuição bivariada. Para cada par é calculada a diferença,

$$d_i = x_{1i} - x_{2i}$$

sendo o teste baseado no número das diferenças positivas e negativas, assumindo que a distribuição das diferenças é simétrica: Neste caso, a média (se existir) e a mediana coincidem com a linha de simetria.

Definição 8.3.1: Teste de Wilcoxon

O teste requer o cálculo das diferenças absolutas entre cada par

$$|d_i| = |x_{1i} - x_{2i}| .$$

No caso de a diferença ser zero, o par não é considerado. A cada diferença absoluta é atribuída uma graduação $R(|d_i|)$, da menor para a maior. Em seguida, são calculadas as somas das graduações positivas T_+ e negativas T_- ,

$$T_+ = \sum_{i=1,\text{sign}(d_i)>0}^{n} \text{sign}(d_i)d_i \qquad T_- = \sum_{i=1,\text{sign}(d_i)<0}^{n} \text{sign}(d_i)d_i.$$

As relações entre estas duas somas são dadas pelas as expressões,

$$T_+ = \frac{n(n+1)}{2} - T_- \qquad T_- = \frac{n(n+1)}{2} - T_+.$$

A estatística de teste é dada por

$$T = \frac{T_+}{\sqrt{\sum_{i=1}^{n} R(|d_i|)^2}}$$

A distribuição dos valores de T pode ser calculada por enumeração, existindo tabelas para as regiões críticas. Para grandes valores de pares de pontos ($n > 5$), a estatística de teste pode ser aproximada pela distribuição normal, com a média e a variância da distribuição das somas dadas por

$$\mu_T = \frac{T_+ + T_-}{2} \text{ e } \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

$$z_T = \frac{T - \mu_T}{\sigma_T}$$

Teste Unilateral	Teste Bilateral	Teste Unilateral
$H_0 : d_{0.50} \geq 0$	$H_0 : d_{0.50} = 0$	$H_0 : d_{0.50} \leq 0$
$H_1 = d_{0.50} < 0$	$H_1 = d_{0.50} \neq 0$	$H_1 = d_{0.50} > 0$

Estatística

$$T = \frac{T_+}{\sqrt{\sum_{i=1}^{n} R(|d_i|)^2}}$$

$$z_T = \frac{T - \mu_T}{\sigma_T}$$

Região de Rejeição

$$T > w_{1-\alpha/2}$$

$$z > z_{1-\alpha/2}$$

onde $d_{0.50}$ representa a mediana dos d_i.

Exemplo 8.3.1: Teste de Wilcoxon

A capacidade respiratória em pacientes com doenças respiratórias pode ser avaliada pela prova de marcha. Nesta prova é pedido a cada sujeito que caminhe durante 6 minutos a distância máxima possível, sem correr. A tabela apresenta os resultados das provas de marcha de 14 sujeitos, que as realizaram por duas vezes com um intervalo de meia hora. Pretende-se avaliar se há uma melhoria de uma prova para outra, o chamado efeito de aprendizagem. Verifique se há diferenças entre as distâncias percorridas nas duas provas. Use $\alpha = 0.10$

Suj.	P_1	P_2
1	696	755
2	690	693
3	662	681
4	693	717
5	489	498
6	600	651
7	610	661
8	630	633
9	543	579
10	561	564
11	594	558
12	513	510
13	684	726
14	558	603

Solução

1. Hipóteses

$$H_0 : d_{0.50} \leq 0$$

$$H_1 : d_{0.50} > 0$$

2. Região crítica

$$z > 1.65$$

3. Estatística

P_1	P_2	Dif. d_i	Grad($\|d_i\|$)	Sinal Grad($\|d_i\|$)
696	755	-59	14	-14
690	694	-4	4	-4
662	681	-19	6	-6
693	717	-24	7	-7
489	498	-9	5	-5
601	651	-50	12	-12
610	661	-51	13	-13
630	633	-3	2.5	-2.5
543	579	-36	8.5	-8.5
562	564	-2	1	-1
594	558	36	8.5	8.5
513	510	3	2.5	2.5
684	726	-42	10	-10
558	603	-45	11	-11
			$\sum_{i=1}^{n} R(x_i) = 105$	$T+ = 11$ $T- = 94$

$$\sigma_T = \sqrt{\frac{14(14+1)(2\times 14+1)}{24}} = 15.93$$

$$z_T = \frac{11 - 52.5}{15.93} = -2.61$$

4. Decisão

   Rejeitar H_0, isto é os sujeitos percorrem maiores distâncias na segunda prova.

8.4 Teste a várias amostras independentes

A comparação de várias amostras independentes pretende responder à questão se as populações de onde provêm as várias amostras são idênticas. Portanto, em termos de planeamento experimental, a comparação é o equivalente não paramétrico ao planeamento completamente casual. A diferença essencial reside no facto de que o cálculo da estatística de teste ser baseado nas graduações das observações e não nas observações originais.

Definição 8.4.1: Teste de Kruskal-Wallis

Hipóteses

H_0 : Todas as k populações são idênticas (As populações têm médias idênticas).

H_1 : Pelo menos uma das populações gera observações mais elevadas do que pelo menos uma das populações.

Estatística

$$H = \frac{12}{N\left(N+1\right)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3\left(N+1\right)$$

Região de Rejeição

$$H > w_{1-\alpha}$$

Cada uma das k amostras contém n_i observações. As N observações são graduadas por ordem crescente, isto é, a cada observação x_{ij} é atribuída uma graduação $R\left(x_{ij}\right)$, com $N = \sum_{i=1}^{k} n_i$:

$$R_i = \sum_{i=1}^{k} R\left(x_{ij}\right).$$

Caso existam observações iguais, a estas serão atribuídas as graduações médias. Os valores críticos são retirados da tabela dos quantis da estatística de Kruskal-Wallis, para o nível de significância especificado α. Quando existem muitos empates (observações com a mesma graduação), ou não existem tabelas exatas para o número e dimensão das amostras, os quantis podem ser aproximados pela distribuição de Qui-Quadrado, com $k-1$ graus de liberdade. O teste assume que todas as amostras são aleatórias e independentes, retiradas das respetivas k populações, com uma escala de medida que é pelo menos ordinal.

Se existirem muitas observações repetidas, a estatística calculada deve ser ajustada do seguinte modo,

$$H' = \frac{H}{1 - \frac{\sum_{j=1}^{l} q_l\left(q_l^2-1\right)}{N\left(N^2-1\right)}}$$

onde l representa o número de conjuntos com observações repetidas e q_j é o número de elementos no conjunto j.

Exemplo 8.4.1: Teste de Kruskal-Wallis

Quatro reatores químicos diferentes foram testados para determinar o rendimento de uma dada reação ($\mathrm{kg/m^3}$). Os resultados observados foram os seguintes:

Reator 1	Reator 2	Reator 3	Reator 4
59	60	42	27
56	82	27	49
73	62	34	22
78	55	49	42
53	42		52
	32		

Verifique se os reatores têm um rendimento médio idêntico. Use $\alpha = 0.05$.

Solução

A tabela apresenta as observações, as respetivas graduações e a correspondente soma por reator.

Reator 1	$R(x_{ij})$	Reator 2	$R(x_{ij})$	Reator 3	$R(x_{ij})$	Reator 4	$R(x_{ij})$
59	15	60	16	42	7	27	2.5
56	14	82	20	27	2.5	49	9.5
73	18	62	17	34	5	22	1
78	19	55	13	49	9.5	42	7
53	12	42	7			52	11
		32	4				
	$R_1 = 78$		$R_2 = 77$		$R_3 = 24$		$R_4 = 31$

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) = \frac{12}{20(21)} \left[\frac{78^2}{5} + \frac{77^2}{6} + \frac{24^2}{4} + \frac{31^2}{5} \right] - 3(21) = 9.605$$

O valor tabelado de $\chi^2_{0.05,3} = 7.815$, implica que a Hipótese Nula será rejeitada, isto é, as populações de onde foram retiradas as amostras, não possuem médias idênticas.

8.5 Teste a várias amostras relacionadas

Experiências planeadas para detetar diferenças entre tratamentos, com as observações organizadas em blocos, são referidas como experiências com blocos aleatórios. O teste não paramétrico de Quade permite testar o efeito dos tratamentos, tendo por base não as observações, mas as suas graduações dentro de cada bloco, e da graduação dos blocos.

Definição 8.5.1: Teste de Quade

Hipóteses

H_0 : Qualquer graduação dentro de um bloco é igualmente provável
(os tratamentos são idênticos).

H_1 : Pelo menos um dos tratamentos gera observações mais elevadas do que pelo menos um dos outros tratamentos.

Estatística

$$T = \frac{(b-1)\,B_1}{A_1 - B_1}$$

Região de Rejeição

$$T > w_{1-\alpha}$$

A cada observação x_{ij}, dentro de cada bloco i, é atribuída uma graduação $R(x_{ij})$. Caso existam observações iguais, a estas serão atribuídas as graduações médias. Para cada bloco é determinada a amplitude das observações originais, sendo os blocos graduados de acordo. A graduação de cada bloco, Q_i, é feita por ordem crescente das amplitudes. Em seguida, para cada observação é calculada a seguinte estatística

$$S_{ij} = Q_i \left[R(x_{ij}) - \frac{k+1}{2} \right].$$

S_{ij} representa, dentro de cada bloco, o contributo relativo de cada observação, ajustado pela importância de cada bloco. Para cada tratamento é definida a seguinte soma

$$S_j = \sum_{i=1}^{b} S_{ij}.$$

Para o cálculo da estatística de teste são ainda necessárias as seguintes somas,

$$A_1 = \sum_{i=1}^{b}\sum_{j=1}^{k} S_{ij}^2 \qquad B_1 = \frac{1}{b}\sum_{j=1}^{k} S_j^2.$$

Os valores críticos são retirados da tabela dos quantis da distribuição F, para o nível de significância especificado α, com $k-1$ e $(b-1)(k-1)$ graus de liberdade. O teste assume que todas as amostras são aleatórias e independentes, que os resultados de um bloco não influenciam os resultados de outro qualquer, e cada observação dentro de um bloco pode ser graduada.

Exemplo 8.5.1: Teste de Quade

Uma empresa têxtil pretende estudar a influência da percentagem de algodão na composição de um tecido, em termos da sua resistência. Para tal usa algodão de quatro origens diferentes, em diferentes percentagens na composição do tecido. A tabela apresenta os resultados da experiência.

	60%	70%	80%	90%
Rússia	19.6	25.1	35.8	34.9
Índia	21.6	22.7	35.4	33.2
Bulgária	20.7	22.8	36.3	34.8
Turquia	23.2	26.3	33.3	36.8

Verifique se existem diferenças entre os tecidos. Use $\alpha = 0.05$.

Solução

A tabela apresenta os cálculos auxiliares necessários à determinação da estatística .

Blocos		1	2	3	4	Amplitude
		19.6	25.1	35.8	34.9	16.2
1	$R(x_{1j})$	1	2	4	3	$Q_1 = 4$
	S_{1j}	-6	-2	6	2	
	S_{1j}^2	36	4	36	4	$\sum_{j=1}^{k} S_{1j}^2 = 80$
		21.6	21.6	35.4	33.2	13.8
2	$R(x_{2j})$	1.5	1.5	4	3	$Q_2 = 2$
	S_{2j}	-2	-2	3	1	
	S_{2j}^2	4	4	9	1	$\sum_{j=1}^{k} S_{2j}^2 = 18$
		22.8	20.7	36.3	34.8	15.6
3	$R(x_{3j})$	2	1	4	3	$Q_3 = 3$
	S_{3j}	-1.5	-4.5	4.5	1.5	
	S_{3j}^2	2.25	20.25	20.25	2.25	$\sum_{j=1}^{k} S_{3j}^2 = 45$
		23.2	26.3	33.3	36.8	13.6
4	$R(x_{4j})$	1	2	3	4	$Q_4 = 1$
	S_{4j}	-1.5	-0.5	0.5	1.5	
	S_{4j}^2	2.25	0.25	0.25	2.25	$\sum_{j=1}^{k} S_{4j}^2 = 5$
	S_i	-11	-9	14	6	$\sum_{i=i}^{b}\sum_{j=1}^{k} S_{ij}^2 = 148$
	S_i^2	121	81	196	36	$\sum_{j=1}^{k} S_j^2 = 434$

Os valores das estatísticas são:

$$A_1 = \sum_{i=1}^{b}\sum_{j=1}^{k} S_{ij}^2 = 148 \qquad B_1 = \frac{1}{b}\sum_{j=1}^{k} S_j^2 = \frac{434}{4} = 108.5$$

$$T = \frac{(b-1)\,B_1}{A_1 - B_1} = \frac{(4-1)\,108.5}{148 - 108.5} = 8.241$$

Assim, dado que o valor de $F_{0.05,3,9} = 3.86$ é menor que o valor da estatística T, a Hipótese Nula é rejeitada, isto é, há diferenças na resistência ditadas pela percentagem de algodão.

8.6 Testes de Kolmogorov-Smirnov

Os testes de Kolmogorov-Smirnov são baseados na máxima distância vertical entre duas funções de distribuição. Para o efeito, a função de distribuição empírica é definida com base numa amostra aleatória.

Definição 8.6.1: Distribuição empírica

A Distribuição Empírica de $x_1, x_2, \ldots, x_n$, uma amostra aleatória, é uma função de x, que iguala a fração dos x_i que são menores ou iguais a x_i, para cada x, com $-\infty < x < \infty$.

Exemplo 8.6.1: Função empírica

Uma amostra aleatória de 5 lâmpadas foi retirada da produção de um dia de uma determinada fábrica. Os tempos de vida, em milhares de horas, observados foram:

1.73	2.74	4.72	4.84	8.06

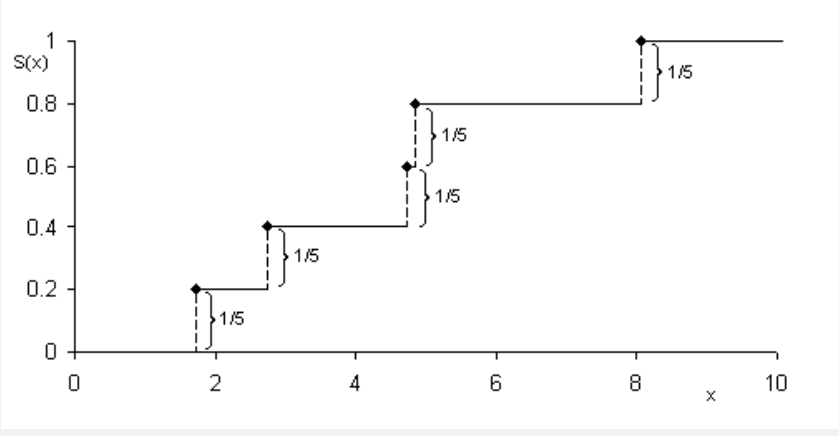
Construa a função empírica.

Solução

A variável aleatória tem a seguinte distribuição empírica de probabilidade:

x	P(X=x)
1.73	0.2
2.74	0.4
4.72	0.6
4.84	0.8
8.06	1

A seguinte figura apresenta a função empírica



Exemplo 8.6.2: Função empírica

Uma amostra aleatória de 8 alunos foi selecionada de uma turma. O número de moedas que cada aluno possuía nos bolsos foi o seguinte:

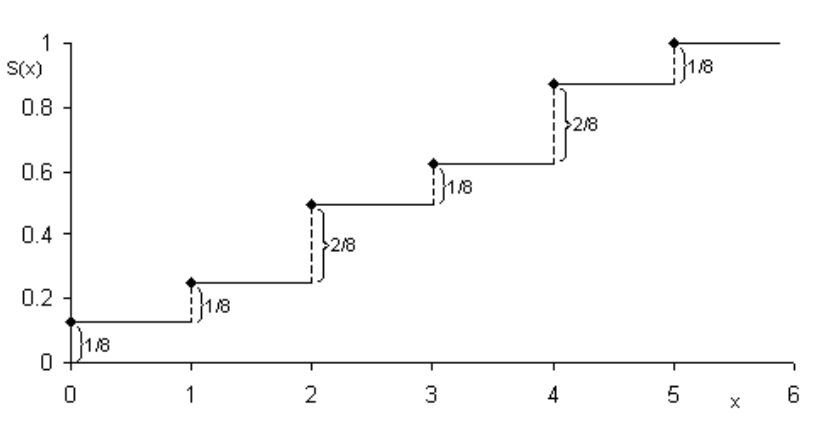
0	1	2	2	3	4	4	5

Construa a função empírica.

Solução

x	P(X=x)
0	0.125
1	0.250
2	0.500
2	0.500
3	0.625
4	0.875
4	0.875
5	1.000

A seguinte figura apresenta a função empírica



Como se pode ver pelos exemplos anteriores, a função empírica é sempre uma função em escada, em que cada degrau corresponde a uma altura de $\frac{1}{n}$, nos pontos observados. A função empírica é contínua à direita, definida pelos traços a cheio, embora, por vezes, por facilidade, os traços verticais sejam também apresentados a cheio. O valor da função empírica é zero até ao menor dos valores observados, e igual a 1 para todos os valores superiores ao maior dos valores observados. Obviamente, a forma da função empírica varia de amostra para amostra.

8.6.1 Teste de Bom Ajuste de Kolmogorov

Este teste constitui uma alternativa ao teste de bom ajuste do Qui-Quadrado, para dados ordinais. O objetivo de um teste de aderência é determinar se a amostra aleatória provém de facto de uma distribuição especificada $F^*(x)$.

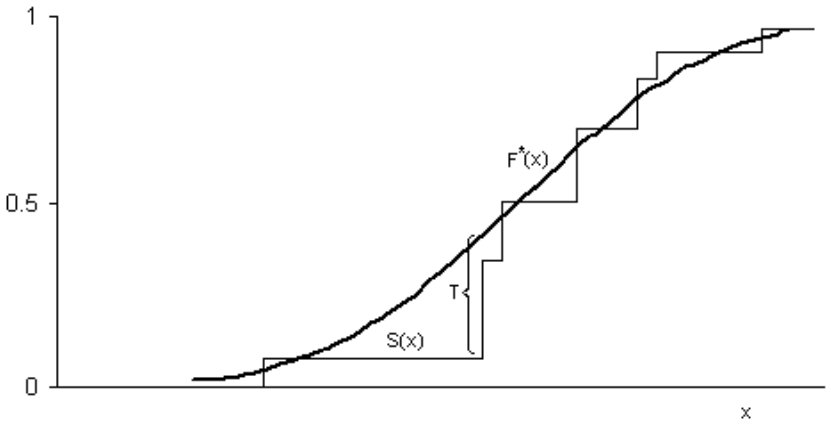


Figura 8.1: T, máxima distância vertical.

O teste de Kolmogorov avalia o ajuste comparando a função empírica $S(x)$ com a função de distribuição $F^*(x)$, através da distância vertical entre as duas curvas. A Figura 8.1 pretende retratar um exemplo. Assim, T representa a máxima distância vertical entre as duas curvas. Se a curva da distribuição empírica estiver muito próxima da curva da distribuição $F^*(x)$, o valor da maior distância vertical será pequeno e a Hipótese Nula de que os dados provêm da distribuição especificada não será rejeitada. Caso contrário, o valor de T será elevado, e a Hipótese Nula será rejeitada.

Num gráfico, a determinação da maior distância vertical é relativamente simples. Contudo, a identificação desta distância sem recurso a um gráfico, em determinados casos, nem sempre coincide com a diferença $F^*(x_i) - S(x_i)$ para um dado valor de x_i. A Figura 8.2 pretende retratar duas situações em que a maior distância vertical é determinada pela diferença $F^*(x_i) - S(x_{i-1})$. Assim, por exemplo, a maior distância vertical entre as duas curvas, no ponto x_{i+1}, $T1$, é dada pela diferença $F^*(x_{i+1}) - S(x_i)$. O mesmo se passa no ponto x_{i+3}, para a determinação da distância $T2$. Devido à forma em escada da função empírica, a maior distância vertical só coincide com a diferença $F^*(x_i) - S(x_i)$, quando a curva empírica $S(x)$ está acima da função especificada.

Existe alguma controvérsia sobre qual o melhor teste de bom ajuste, de Kolmogorov ou de Qui-Quadrado. Em geral, o teste de Kolmogorov deve ser usado se a amostra é pequena, uma vez que o teste de Qui-Quadrado pressupõe que a amostra é suficientemente grande para que a

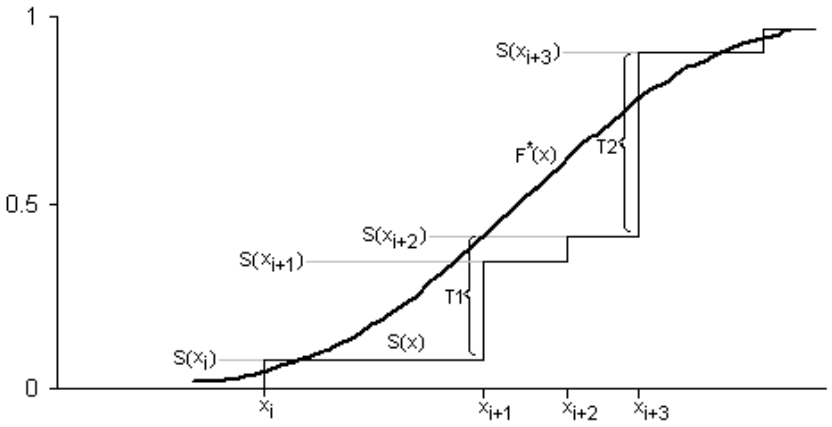


Figura 8.2: Máxima distância vertical.

aproximação à estatística de Qui-Quadrado seja válida.

Definição 8.6.2: Teste de Bom Ajuste de Kolmogorov

Teste Unilateral	Teste Bilateral	Teste Unilateral
$H_0 : F(x) \geq F^*(x)$	$H_0 : F(x) = F^*(x)$	$H_0 : F(x) \leq F^*(x)$
$-\infty < x < \infty$	$-\infty < x < \infty$	$-\infty < x < \infty$
$H_1 : F(x) < F^*(x)$	$H_1 : F(x) \neq F^*(x)$	$H_1 : F(x) > F^*(x)$
para pelo menos	para pelo menos	para pelo menos
um valor de x	um valor de x	um valor de x
	Estatística	
$T^+ = \sup_x [F^*(x) - S(x)]$	$T = \sup_x \lvert F^*(x) - S(x) \rvert$	$T^- = \sup_x [S(x) - F^*(x)]$
	Região de Rejeição	
$T^+ > w_{1-\alpha}$	$T > w_{1-\alpha/2}$	$T^- > w_{1-\alpha}$

Os valores críticos são retirados da tabela dos quantis da estatística de Kolmogorov, para o nível de significância especificado α. O teste assume que os dados constituem uma amostra aleatória de uma distribuição desconhecida $F(x)$ e que a função $F^*(x)$ é completamente especificada. O teste é exato se a distribuição se $F^*(x)$ é contínua, sendo conservador no caso discreto.

Exemplo 8.6.3: Teste de Bom Ajuste de Kolmogorov

Os dados representam uma amostra aleatória das concentrações de um poluente (em ppm) num determinado lago.

1.10	1.40	1.60	1.82	2.10	2.36	2.40	2.45	2.56	2.88

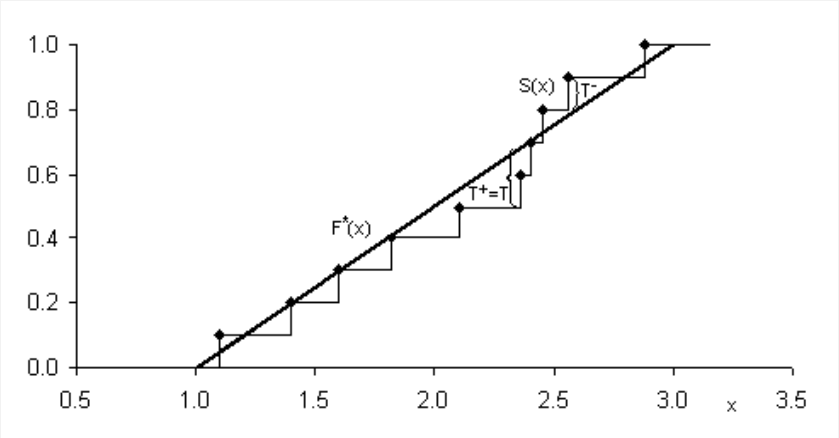
Verifique se as concentrações constituem uma amostra aleatória de uma distribuição uniforme U(1,3).

Solução

A tabela apresenta os cálculos necessários à determinação da maior distância vertical.

i	x_i	$F(x_i)$	$S(x_i)$	$F^*(x_i) - S(x_i)$	$F^*(x_i) - S(x_i)$
1	1.1 0	0.05	0.10	-0.05	0.05
2	1.40	0.20	0.20	0.00	0.10
3	1.60	0.30	0.30	0.00	0.10
4	1.82	0.41	0.40	0.01	0.11
5	2.10	0.55	0.50	0.05	0.15
6	2.36	0.68	0.60	0.08	**0.18**
7	2.40	0.70	0.70	0.00	0.10
8	2.45	0.73	0.80	-0.07	0.03
9	2.56	0.78	0.90	**-0.12**	-0.02
10	2.88	0.94	1.00	-0.06	0.04

A seguinte figura apresenta o gráfico das duas funções, sendo assinaladas as várias estatísticas do teste de Kolmogorov.



Os valores das estatísticas unilaterais são:

$$T^+ = \sup_x [F^*(x) - S(x)] = F^*(2.36) - S(2.10) = 0.68 - 0.50 = 0.18$$

$$T^{-} = \sup_{x} \left[S\left(x\right) - F^{*}\left(x\right) \right] = S\left(2.56\right) - F^{*}\left(2.56\right) = 0.90 - 0.78 = 0.12$$

A estatística bilateral T iguala o maior valor das estatísticas unilaterais, isto é, $T = 0.18$. Assim, para um nível de significância de 5%, o valor tabelado para o teste bilateral é $w_{1-0.025} = 0.409$, o que conduz à não rejeição da Hipótese Nula, isto é, os dados seguem uma distribuição uniforme $U(1,3)$.

8.7 Testes para famílias de distribuições

O teste de Kolmogorov é usado, como foi visto, para testar se uma amostra aleatória está de acordo com uma distribuição especificada, isto é, quando são conhecidos todos os parâmetros da distribuição. O teste de Kolmogorov foi adaptado, por Lilliefors, para permitir testar se os dados provêm de distribuições não completamente especificadas na Hipótese Nula, em particular para a distribuição normal e exponencial. O cálculo das estatísticas de teste é feito da mesma forma, isto é, através da máxima distância vertical, embora a tabela de valores críticos seja diferente para cada distribuição.

Definição 8.7.1: Teste para famílias de distribuições

Teste Bilateral

$H_0 : F\left(x\right) = F^{*}\left(x\right) \quad -\infty < x < \infty$

$H_1 : F\left(x\right) \neq F^{*}\left(x\right)$ para pelo menos um valor de x

Estatística

$$T_1 = \sup_{x} \left| F^{*}\left(x\right) - S\left(x\right) \right|$$

Região de Rejeição

$$T > w_{1-\alpha/2}$$

8.7.1 Teste de Lilliefors para a Normal

O teste assume que os dados constituem uma amostra aleatória de uma distribuição desconhecida $F\left(x\right)$ e que a função$F^{*}\left(x\right)$ corresponde a uma distribuição normal para a qual a média e a variância não são especificadas. Os valores críticos são retirados da tabela dos quantis da estatística de Lilliefors, para o nível de significância especificado α. Os valores da média e do desvio padrão são

estimados a partir dos dados da amostra através das seguintes fórmulas,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}},$$

com os valores padronizados calculados através de

$$z_i = \frac{x_i - \bar{x}}{s}.$$

Exemplo 8.7.1: Teste de Lilliefors para a Normal

Uma amostra de 20 parafusos de uma máquina foi retirada da produção de um determinado dia. Verifique se os comprimentos seguem uma distribuição normal.

31.4	23.1	23.4	32.2	26.5	24.8	34.2	38.7	29.2	37.6
23.9	35.3	23.6	25.1	37.3	33.4	31.2	28.9	30.1	27.2

Solução

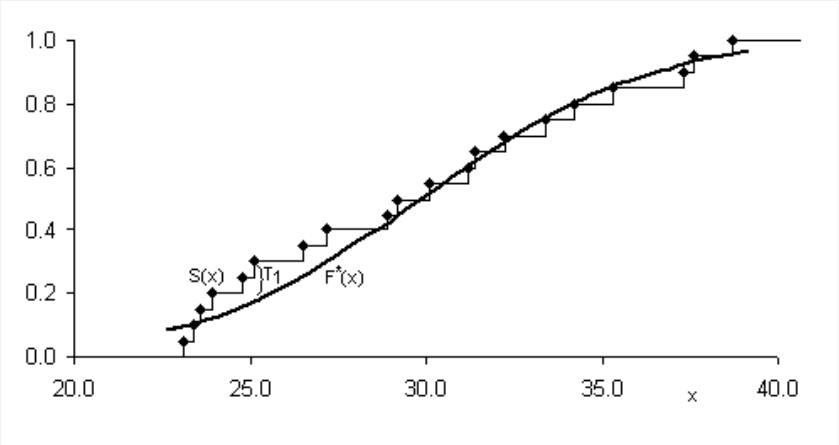
A tabela apresenta os cálculos necessários à determinação da maior distância vertical. A média e o desvio padrão são:

$$\bar{x} = 29.86 \qquad s = 5.10$$

i	x_i	z_i	$F^*(x_i)$	$S(x_i)$	$F^*(x_i) - S(x_i)$	$F^*(x_i) - S(x_i)$
1	23.1	-1.32	0.0927	0.0500	0.0427	0.0927
2	23.4	-1.27	0.1028	0.1000	0.0028	0.0528
3	23.6	-1.23	0.1100	0.1500	-0.0400	0.0100
4	23.9	-1.17	0.1215	0.2000	-0.0785	-0.0285
5	24.8	-0.99	0.1608	0.2500	-0.0892	-0.0392
6	25.1	-0.93	0.1756	0.3000	**-0.1244**	-0.0744
7	26.5	-0.66	0.2553	0.3500	-0.0947	-0.0447
8	27.2	-0.52	0.3013	0.4000	-0.0987	-0.0487
9	28.9	-0.19	0.4257	0.4500	-0.0243	0.0257
10	29.2	-0.13	0.4489	0.5000	-0.0511	-0.0011
11	30.1	0.05	0.5192	0.5500	-0.0308	0.0192
12	31.2	0.26	0.6040	0.6000	0.0040	0.0540
13	31.4	0.30	0.6190	0.6500	-0.0310	0.0190
14	32.2	0.46	0.6772	0.7000	-0.0228	0.0272

i	x_i	z_i	$F^*(x_i)$	$S(x_i)$	$F^*(x_i) - S(x_i)$	$F^*(x_i) - S(x_i)$
15	33.4	0.70	0.7565	0.7500	0.0065	0.0565
16	34.2	0.85	0.8029	0.8000	0.0029	0.0529
17	35.3	1.07	0.8572	0.8500	0.0072	0.0572
18	37.3	1.46	0.9278	0.9000	0.0278	0.0778
19	37.6	1.52	0.9356	0.9500	-0.0144	0.0356
20	38.7	1.73	0.9586	1.0000	-0.0414	0.0086

A seguinte figura apresenta o gráfico das duas funções, sendo assinalada a maior distância vertical.



O valor da estatística é:

$$T_1 = \sup_x |F^*(x) - S(x)| = |F^*(25.1) - S(25.1)| = |0.1756 - 0.30| = 0.1244$$

Assim, para um nível de significância de 5%, o valor tabelado para o teste bilateral é $w_{1-0.025} = 0.190$, o que conduz à não rejeição da Hipótese Nula, isto é, não se rejeita que os dados seguem uma distribuição normal. De notar que se o enunciado do problema especificasse uma média e um desvio padrão, o teste apropriado seria o teste de Kolmogorov.

8.7.2 Teste de Lilliefors para a Exponencial

O teste assume que os dados constituem uma amostra aleatória de uma distribuição desconhecida $F(x)$ e que a função $F^*(x)$ corresponde a uma distribuição exponencial de parâmetro λ desconhecido.

$$F^*(x) = \begin{cases} 1 - e^{-x/\lambda} & x > 0 \\ 0 & \text{outros valores} \end{cases}.$$

Os valores críticos são retirados da tabela dos quantis da estatística de Lilliefors, para o nível de significância especificado α. A média é estimada a partir dos dados da amostra através da seguinte fórmula,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

calculando-se os valores auxiliares z_i através de

$$z_i = \frac{x_i}{\bar{x}}.$$

Exemplo 8.7.2: Teste de Lilliefors para a Exponencial

Uma amostra aleatória de lâmpadas foi retirada da produção de um determinado dia. O tempo de vida (em milhares de horas) foi determinado, tendo-se obtido os seguintes resultados:

3.15	1.12	2.91	2.49	4.20	2.36
1.77	3.21	4.25	2.38	4.70	2.58

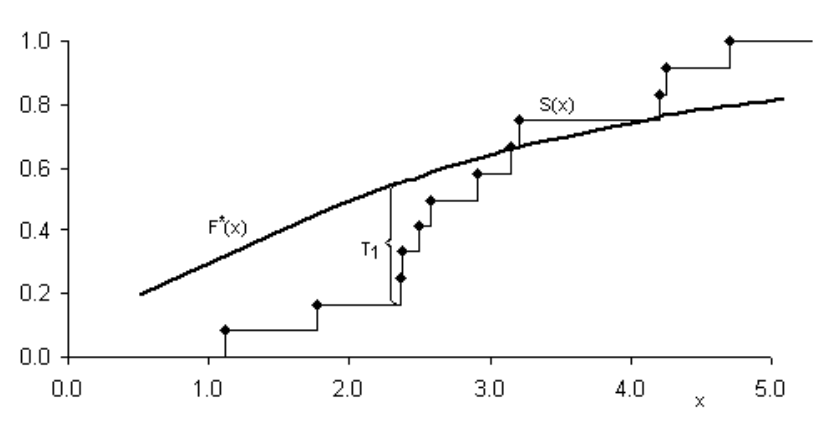
Verifique se os dados se ajustam a uma distribuição exponencial.

Solução

A tabela apresenta os cálculos necessários à determinação da maior distância vertical. A média das observações é,

i	x_i	z_i	$F^*(x_i)$	$S(x_i)$	$F^*(x_i) - S(x_i)$	$F^*(x_i) - S(x_i)$
1	1.12	0.3827	0.3180	0.0833	0.2346	0.3180
2	1.77	0.6048	0.4538	0.1667	0.2871	0.3705
3	2.36	0.8064	0.5535	0.2500	0.3035	**0.3869**
4	2.38	0.8132	0.5566	0.3333	0.2232	0.3066
5	2.49	0.8508	0.5729	0.4167	0.1563	0.2396
6	2.58	0.8815	0.5859	0.5000	0.0859	0.1692
7	2.91	0.9943	0.6300	0.5833	0.0467	0.1300
8	3.15	1.0763	0.6591	0.6667	-0.0075	0.0758
9	3.21	1.0968	0.6661	0.7500	-0.0839	-0.0006
10	4.20	1.4351	0.7619	0.8333	-0.0714	0.0119
11	4.25	1.4522	0.7659	0.9167	-0.1507	-0.0674
12	4.70	1.6059	0.7993	1.0000	-0.2007	-0.1174

A seguinte figura apresenta o gráfico das duas funções, sendo assinalada a maior distância vertical.



O valor da estatística é:

$$T_1 = \sup_x |F^*(x) - S(x)| = |F^*(2.36) - S(1.77)| = |0.5535 - 0.1667| = 0.3869.$$

Para um nível de significância de 5%, o valor tabelado para o teste bilateral é $w_{1-0.025} = 0.2981$, o que conduz à rejeição da Hipótese Nula, isto é, os dados não seguem uma distribuição exponencial. Caso o parâmetro λ tivesse sido especificado na Hipótese Nula, o problema seria resolvido através do teste de Kolmogorov.

8.8 Teste para duas amostras independentes

Em presença de duas amostras, eventualmente provenientes de duas populações, o objetivo é determinar se as duas funções de distribuição são idênticas.

Existem vários testes que permitem comparar duas amostras em termos da sua localização. Contudo, o teste de Smirnov permite comparar duas distribuições não só em termos das suas localizações mas também das suas variâncias.

Definição 8.8.1: Teste de Smirnov

Teste Unilateral	Teste Bilateral	Teste Unilateral
$H_0 : F(x) \geq G(y)$	$H_0 : F(x) = G(y)$	$H_0 : F(x) \leq G(y)$
$-\infty < x < \infty$	$-\infty < x < \infty$	$-\infty < x < \infty$
$-\infty < y < \infty$	$-\infty < y < \infty$	$-\infty < y < \infty$
$H_1 : F(x) < G(y)$	$H_1 : F(x) \neq G(y)$	$H_1 : F(x) > G(y)$
para pelo menos	para pelo menos	para pelo menos
um valor de x	um valor de x	um valor de x
	Estatística	
$T^+ = \sup_x [S_1(x) - S_2(y)]$	$T = \sup_x \lvert S_1(x) - S_2(y) \rvert$	$T^- = \sup_x [S_2(y) - S_1(x)]$
	Região de Rejeição	
$T^+ > w_{1-\alpha}$	$T > w_{1-\alpha/2}$	$T^- > w_{1-\alpha}$

Os valores críticos são retirados da tabela dos quantis da estatística de Smirnov, para o nível de significância especificado α. O teste assume que os dados constituem duas amostras aleatórias independentes de duas distribuições desconhecidas. O teste é exato se as variáveis aleatórias são contínuas.

Exemplo 8.8.1: Teste de Smirnov

Para estudar dois tipos de filtros de poluição atmosférica, duas amostras foram recolhidas durante a operação de cada um dos filtros. Os dados representam as concentrações em (g/m^3) de dióxido de enxofre após a filtragem. Verifique se as duas distribuições são idênticas.

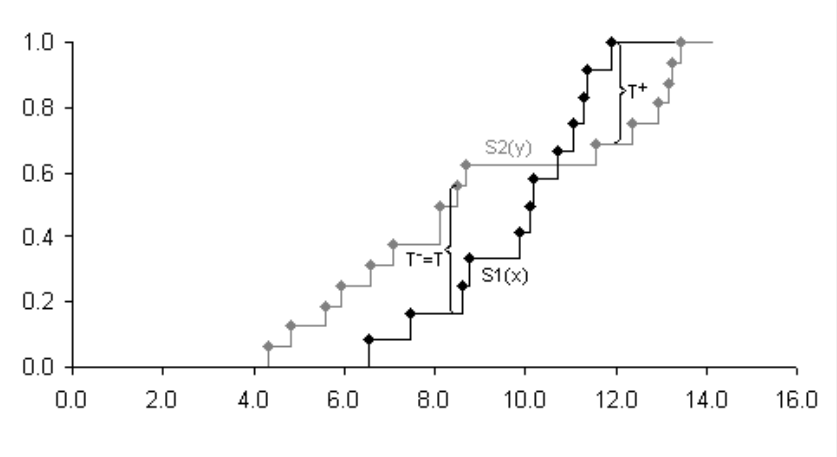
Amostra 1	8.8	11.4	11.1	8.6	11.3	10.7	6.5	10.2	9.9	10.1	11.9	7.5
Amostra 2	13.2	6.6	7.1	8.6	13.0	5.6	13.3	8.1	4.3	13.4	4.8	11.6
	12.4	5.9	8.1	8.5								

Solução

A tabela apresenta os cálculos das funções empíricas para as duas amostras, sendo assinaladas as maiores distâncias verticais.

i	x_i	j	y_j	$S_1(x)$	$S_2(y)$	$S_1(x) - S_2(y)$
		1	4.3	0.0000	0.0625	-0.0625
		2	4.8	0.0000	0.1250	-0.1250
		3	5.6	0.0000	0.1875	-0.1875
		4	5.9	0.0000	0.2500	-0.2500
1	6.5			0.0833	0.2500	-0.1667
		5	6.6	0.0833	0.3125	-0.2292
		6	7.1	0.0833	0.3750	-0.2917
2	7.5			0.1667	0.3750	-0.2083
		7	8.1	0.1667	0.4375	-0.2708
		8	8.1	0.1667	0.5000	-0.3333
		9	8.5	0.1667	0.5625	**-0.3958**
3	8.6			0.2500	0.5625	-0.3125
		10	8.7	0.2500	0.6250	-0.3750
4	8.8			0.3333	0.6250	-0.2917
5	9.9			0.4167	0.6250	-0.2083
6	10.1			0.5000	0.6250	-0.1250
7	10.2			0.5833	0.6250	-0.0417
8	10.7			0.6667	0.6250	0.0417
9	11.1			0.7500	0.6250	0.1250
10	11.3			0.8333	0.6250	0.2083
11	11.4			0.9167	0.6250	0.2917
		11	11.6	0.9167	0.6875	0.2292
12	11.9			1.0000	0.6875	**0.3125**
		12	12.4	1.0000	0.7500	0.2500
		13	13	1.0000	0.8125	0.1875
		14	13.2	1.0000	0.8750	0.1250
		15	13.3	1.0000	0.9375	0.0625
		16	13.4	1.0000	1.0000	0.0000

A seguinte figura apresenta o gráfico das duas funções, sendo assinaladas as várias estatísticas do teste de Smirnov.



Os valores das estatísticas unilaterais são:

$$T^{+} = \sup_{x} \left[S_1(x) - S_2(y) \right] = S_1(11.9) - S_2(11.6) = 1.0 - 0.6875 = 0.3125$$

$$T^{-} = \sup_{x} \left[S_2(y) - S_1(x) \right] = S_2(8.5) - S_1(7.5) = 0.5625 - 0.1667 = 0.3958.$$

A estatística bilateral T iguala o maior valor das estatísticas unilaterais, isto é, $T = 0.3958$. Para um nível de significância de 5%, o valor tabelado para o teste bilateral é $w_{1-0.025} = 0.5$, o que conduz à não rejeição da Hipótese Nula, isto é, à não rejeição de que as amostras provêm de duas distribuições idênticas.

8.9 Correlação baseada em graduações

A correlação entre duas variáveis (X_i, Y_i) é uma medida da sua associação linear, como, por exemplo, o peso e a estatura de pessoas, os resultados em diversos testes de capacidade intelectual. O Coeficiente de Correlação de Pearson,

$$R = \frac{\sum \left(X_i - \bar{X} \right) \left(Y_i - \bar{Y} \right)}{\sqrt{\sum \left(X_i - \bar{X} \right)^2 \sum \left(Y_i - \bar{Y} \right)^2}}$$

impõe que a distribuição bivariada das variáveis aleatórias X e Y seja normal. Uma forma de ultrapassar esta dependência é usar as graduações das observações. Spearman desenvolveu um coeficiente de correlação semelhante ao de Pearson, com a ressalva de que em vez dos valores observados, são usadas graduações. Contudo, uma medida de correlação não paramétrica deve satisfazer os mesmos requisitos do Coeficiente de Correlação de Pearson, isto é, deve ser uma medida que varia entre -1 e 1, positiva quando as duas variáveis variam no mesmo sentido e negativa quando aos maiores valores de uma variável correspondem os menores valores da outra.

Para além disso, se as duas variáveis forem independentes, o seu coeficiente de correlação deve ser nulo.

Definição 8.9.1: Coeficiente de Correlação de Spearman

O coeficiente de correlação de Spearman possui as características referidas anteriormente, sendo a sua expressão dada por

$$R_s = 1 - \frac{6\sum_{i=1}^{n}\left[R\left(x_i\right) - R\left(y_i\right)\right]^2}{n\left(n^2-1\right)} = 1 - \frac{6T}{n\left(n^2-1\right)}$$

onde T representa o quadrado da diferença entre as graduações. No caso de existirem muitos empates nas graduações, deve ser usada a seguinte fórmula,

$$R_S = \frac{\sum_{i=1}^{n} R\left(x_i\right) R\left(y_i\right) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\sum_{i=1}^{n} R\left(x_i\right)^2 - n\left(\frac{n+1}{2}\right)^2}\sqrt{\sum_{i=1}^{n} R\left(y_i\right)^2 - n\left(\frac{n+1}{2}\right)^2}}$$

que é justamente o coeficiente de correlação de Pearson baseado nas graduações das observações.

Definição 8.9.2: Teste de Hipóteses para o Coeficiente de Correlação

Teste Unilateral	Teste Bilateral	Teste Unilateral
	$H_0 : \rho = 0$	
$H_1 : \rho < 0$	$H_1 : \rho \neq 0$	$H_1 : H_1 : \rho > 0$
	Estatística	
	$R_s = 1 - \frac{6T}{n\left(n^2-1\right)}$	
	Região de Rejeição	
$R_s < -w_{1-\alpha}$	$\lvert R_S\rvert > w_{1-\alpha/2}$	$R_S > w_{1-\alpha}$

onde $w_{1-\alpha}$ e $w_{1-\alpha/2}$ são, respetivamente, os quantis valores da distribuição R_s. Os valores (X, Y) são aleatoriamente selecionados de uma população bivariada.

Exemplo 8.9.1: Teste de Hipóteses para o Coeficiente de Correlação

O Índice de Desenvolvimento de Griffiths é uma medida agregadora destinada a avaliar o desenvolvimento psico-motor de crianças. Este índice é calculado através da avaliação de determinadas tarefas motoras e intelectuais. Os dados representam as avaliações motora e

intelectual para 9 crianças com a idade de 4 anos. Calcule o coeficiente de correlação de Spearman e teste a sua significância ao nível de 5

Motora	Intelectual
84	77
73	85
101	105
74	86
88	108
100	116
86	96
95	100
82	100

Solução

A tabela apresenta as graduações, as diferenças d_i e quadrados respetivos.

x_i	84	73	101	74	88	100	86	95	82
$R(x_i)$	4	1	9	2	6	7	5	6	3
y_i	77	85	105	86	108	116	96	100	100
$R(y_i)$	1	2	7	3	8	9	4	5.5	5.5
d_i	3	-1	2	-1	-2	-2	1	0.5	-2.5
d_i^2	9	1	4	1	4	4	1	0.25	6.25

O valor de T é igual a 30.5. Assim, o coeficiente de correlação de Spearman toma o seguinte valor,

$$R_s = 1 - \frac{6T}{n\left(n^2-1\right)} = 1 - \frac{6\left(30.5\right)}{9\left(81-1\right)} = 0.7458$$

Para um nível de significância de 5%, $w_{1-0.025} = 0.6833$, donde se conclui pela rejeição da Hipótese Nula, isto é, há uma correlação positiva entre os índices de desenvolvimento motor e intelectual.

Exercícios

1. Um engenheiro tem a possibilidade de adquirir componentes elétricos de dois fornecedores. Para decidir qual dos dois deve comprar, pediu amostras e sujeitou-as a teste, registando o tempo até à falha. Sabe-se que o tempo até à falha deste tipo de componentes elétricos não segue uma distribuição normal. A tabela apresenta os tempos, em horas, até à falha debaixo

de uma utilização contínua. Use um teste não paramétrico apropriado para verificar se há diferenças nos tempos de falha dos aparelhos dos dois fornecedores. O que pode concluir?

F1	125	199	371	143	125	443	933	699	158	678
F2	547	472	443	526	1943	433	489	1898	1236	651

2. A tabela apresenta a Pressão Parcial de Oxigénio de doentes com doença obstrutiva crónica. Verifique se a distribuição da Pressão Parcial segue uma distribuição normal. Use um teste não paramétrico. O que pode concluir?

41.80	45.20	49.20	49.90	50.20	50.20	55.00	56.30
56.60	56.80	60.20	61.00	61.90	63.90	64.00	64.10
64.60	65.30	65.50	66.60	66.60	67.10	69.40	81.80

3. A adesão de células a superfícies sintéticas depende das propriedades das superfícies. Poucos estudos analisaram as propriedades das superfícies no comportamento das células, em particular do sangue. Para estudar a adesão, cinco superfícies foram preparadas com diferentes concentrações do composto C18 (0% a 10%). A tabela apresenta o valor médio das células aderidas em cada experiência para as diferentes concentrações de C18. Sabe-se que a adesão não segue uma distribuição normal nem que as variâncias sejam homogéneas. Por estas razões, verifique se existem diferenças na adesão, usando um teste não paramétrico adequado. O que pode concluir?

0%C18	1% C18	2.5% C18	5% C18	10% C18
11.5	3.8	6.9	2.3	7
2.8	4.4	4.8	2.9	5.9
4.7	4.1	3.8	4.1	5.3
4.1	4.4	3.8	6.1	7.7

4. A tabela apresenta os valores da Pressão Parcial de Dióxido de Carbono de dois doentes com doença obstrutriva pulmonar crónica. Verifique se há diferenças nas distribuições, nomeadamente quanto à localização e à forma.

Doente 1	67.3	62.5	66.5	60.5	60.2	66.2	71.4	61.2	57.2	72.1	61.9
	59.2	56.5	49.8	55.7	59.7	56.6	59.9	66.5			
Doente 2	55.0	49.1	44.5	45.4	56.5	67.3	54.5	53.0	62.9	61.2	55.4
	49.5	60.0	55.4	59.7							

5. A tabela apresenta o tempo que ratos de laboratório com um determinada doença aguentaram correr num aparelho ROTAROD. Os tempos foram registados nas 12 e 24 e pretendia-se verificar se a doença tinha algum efeito na capacidade física dos animais. O que pode concluir

12h	28	15	86	33	18	33	32	31	32	14	78	41	47	12	14	12	80	42	47
24h	10	18	97	88	18	75	63	52	37	19	40	24	86	12	30	8	69	49	91

6. Durante um semestre um estudante obteve notas em várias matérias como mostra a tabela que se segue. Determinar se existe diferença entre as notas para um nível de significância de 0.01.

Matemática	72	80	83	75	
Ciências	81	74	77		
Inglês	88	82	90	87	80
Economia	74	71	77	70	

7. Os artigos fabricados por uma companhia são produzidos por três operários usando três máquinas diferentes. O fabricante deseja determinar se existe diferença entre os operários e entre as máquinas. Realiza-se uma experiência para determinar o número de artigos, por dia, produzidos por cada operário em cada máquina. O que pode concluir?. Use um nível de significância de 0.05.

Operário	Máquina A	Máquina B	Máquina C
1	23	34	28
2	27	30	25
3	24	28	27

8. Os enólogos de uma grande adega cooperativa admitem que, em média, o álcool provável (expresso em graus) das uvas anualmente entregues para vinificação pelo conjunto dos seus sócios, segue uma distribuição Normal com média 10^o e desvio padrão 1^o. Pretende-se testar a validade desta conjetura a partir dos dados obtidos nos últimos 12 anos:

11.9	10.6	13.3	11.6	12.9	10.4	11.3	13.5	9.1	8.2	11.6	10.0

9. Os dados seguintes representam as percentagens de óxidos de azoto removidos numa central térmica a carvão:

91	95	90	83	91	65	55	42	55
81	89	38	20	45	58	85	78	70

Para estes dados $n = 18$, $\bar{x} = 68.39$, $s = 21.54$. Teste a hipótese de que a população é normal.

10. Um agricultor deseja determinar a existência de diferença nas produções entre duas variedades diferentes de trigo. A tabela que se segue apresenta a produção de trigo por unidade de área usando as duas variedades. Que pode o agricultor concluir quanto às distribuições da produção dos diferentes tipos de trigo. (Use $\alpha = 0.05$)

Trigo I	15.9	15.3	16.4	14.9	15.3	16.0	14.6	15.3	14.5	16.6
Trigo II	16.4	16.8	17.1	16.9	18.0	15.6	18.1	17.2	15.4	

11. Os valores listados representam as concentrações de zinco (mg/kg) em vários locais numa zona interior e numa zona exterior dum local de depósito de lixos no oceano.

Interior	13.5	23.8	20.9	23.8	20.0	24.4	16.4	18.3	17.6	25.4	23.3
Exterior	26.4	20.6	19.8	15.0	16.8	20.4	23.4	21.5			

Verifique se existem diferenças na distribuição da concentração nas duas zonas.

Soluções

1.
```
Wilcoxon rank sum test with continuity correction
data:  temp_f$tempo by temp_f$fact
W = 22.5, p-value = 0.0411
alternative hypothesis: true location shift is not equal to 0

temp_f = read.csv("npar_1.csv", header=T)
temp_f
names(temp_f)[1]="tempo"
temp_f
temp_f$fact=as.factor(temp_f$fact)
wilcox.test(temp_f$tempo ~ temp_f$fact, exact=FALSE)
```

2.
```
One-sample Kolmogorov-Smirnov test
data:  press$pao2
D = 0.13776, p-value = 0.7526
alternative hypothesis: two-sided

press = read.csv("n_par_2.csv", header=T)
press
names(press)[1]="pao2"
press
mean(press$pao2)
sd(press$pao2)
ks.test(press$pao2, "pnorm", mean=mean(press$pao2),sd=sd(press$pao2))
```

3. Kruskal-Wallis rank sum test

   data: ad_cel$ades and ad_cel$fact

   Kruskal-Wallis chi-squared = 6.0409, df = 4, p-value = 0.1961

```
ad_cel = read.csv("n_par_3.csv", header=T)
ad_cel
names(ad_cel)[1]="ades"
ad_cel
ad_cel$fact=as.factor(ad_cel$fact)
kruskal.test(ad_cel$ades,ad_cel$fact)
```

4. Wilcoxon rank sum test with continuity correction

   data: press_2$pao2 by press_2$doent

   W = 223, p-value = 0.005503

   alternative hypothesis: true location shift is not equal to 0

```
press_2 = read.csv("n_par_4.csv", header=T)
press_2
names(press_2)[1]="pao2"
press_2
press_2$doent=as.factor(press_2$doent)
wilcox.test(press_2$pao2 ~ press_2$doent, exact=FALSE)
```

5. Wilcoxon signed rank test with continuity correction

   data: tempos$t_12 and tempos$t_24

   V = 40.5, p-value = 0.09277

   alternative hypothesis: true location shift is not equal to 0

```
tempos = read.csv("n_par_5.csv", header=T)
tempos
names(tempos)[1]="t_12"
tempos
wilcox.test(tempos$t_12,tempos$t_24, paired=TRUE)
wilcox.test(tempos$t_12-tempos$t_24)
```

6. Kruskal-Wallis rank sum test

   data: testes$notas and testes$fact

   Kruskal-Wallis chi-squared = 9.2657, df = 3, p-value = 0.02596

```
testes = read.csv("n_par_6.csv", header=T)
testes
names(testes)[1]="notas"
testes
teste$fact=as.factor(testes$fact)
kruskal.test(testes$notas,testes$fact)
```

7. Quade F = 4, num df = 2, denom df = 4, p-value = 0.1111

```
fabr = matrix(c(23, 34, 28,
                27, 30, 25,
                24, 28, 27),
                nrow = 3, byrow = TRUE,
                dimnames =
                  list(Oper = as.character(1:3),
                       Maq = LETTERS[1:3]))
fabr
quade.test(fabr)
```

8. One-sample Kolmogorov-Smirnov test
   data: alcool$alc
   D = 0.48653, p-value = 0.006819
   alternative hypothesis: two-sided

```
alcool = read.csv("n_par_8.csv", header=T)
alcool
names(alcool)[1]="alc"
alcool
mean(alcool$alc)
sd(alcool$alc)
ks.test(alcool$alc, "pnorm", mean=10,sd=1)
```

9. One-sample Kolmogorov-Smirnov test
   data: dioxido$co2
   D = 0.16771, p-value = 0.6919
   alternative hypothesis: two-sided

```
dioxido = read.csv("n_par_9.csv", header=T)
```

```
dioxido
names(dioxido)[1]="co2"
dioxido
mean(dioxido$co2)
sd(dioxido$co2)
ks.test(dioxido$co2, "pnorm", mean=mean(dioxido$co2),sd=sd(dioxido$co2))
```

10. Two-sample Kolmogorov-Smirnov test
    data: prod$rend_1 and prod$rend_2
    D = 0.66667, p-value = 0.02968
    alternative hypothesis: two-sided

```
prod = read.csv("n_par_10.csv", header=T)
prod
names(prod)[1]="rend_1"
prod
ks.test(prod$rend_1,prod$rend_2)
```

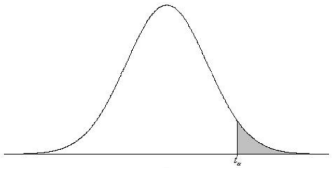
11. Two-sample Kolmogorov-Smirnov test
    data: zinc$int and zinc$ext
    D = 0.23864, p-value = 0.9546
    alternative hypothesis: two-sided

```
zinc = read.csv("n_par_11.csv", header=T)
zinc
names(zinc)[1]="int"
zinc
ks.test(zinc$int,zinc$ext)
```

Apêndice A

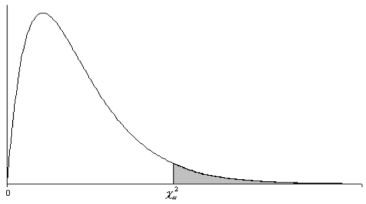
Tabelas estatísticas

A.1 Tabela da distribuição t-Student



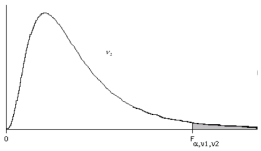
g.l.	0.25	0.15	0.10	0.05	0.025	0.010	0.005
1	1.000	1.963	3.078	6.314	12.706	31.821	63.657
2	0.816	1.386	1.886	2.920	4.303	6.965	9.925
3	0.765	1.250	1.638	2.353	3.182	4.541	5.841
4	0.741	1.190	1.533	2.132	2.776	3.747	4.604
5	0.727	1.156	1.476	2.015	2.571	3.365	4.032
6	0.718	1.134	1.440	1.943	2.447	3.143	3.707
7	0.711	1.119	1.415	1.895	2.365	2.998	3.499
8	0.706	1.108	1.397	1.860	2.306	2.896	3.355
9	0.703	1.100	1.383	1.833	2.262	2.821	3.250
10	0.700	1.093	1.372	1.812	2.228	2.764	3.169
11	0.697	1.088	1.363	1.796	2.201	2.718	3.106
12	0.695	1.083	1.356	1.782	2.179	2.681	3.055
13	0.694	1.079	1.350	1.771	2.160	2.650	3.012
14	0.692	1.076	1.345	1.761	2.145	2.624	2.977
15	0.691	1.074	1.341	1.753	2.131	2.602	2.947
16	0.690	1.071	1.337	1.746	2.120	2.583	2.921
17	0.689	1.069	1.333	1.740	2.110	2.567	2.898
18	0.688	1.067	1.330	1.734	2.101	2.552	2.878
19	0.688	1.066	1.328	1.729	2.093	2.539	2.861
20	0.687	1.064	1.325	1.725	2.086	2.528	2.845
21	0.686	1.063	1.323	1.721	2.080	2.518	2.831
22	0.686	1.061	1.321	1.717	2.074	2.508	2.819
23	0.685	1.060	1.319	1.714	2.069	2.500	2.807
24	0.685	1.059	1.318	1.711	2.064	2.492	2.797
25	0.684	1.058	1.316	1.708	2.060	2.485	2.787
26	0.684	1.058	1.315	1.706	2.056	2.479	2.779
27	0.684	1.057	1.314	1.703	2.052	2.473	2.771
28	0.683	1.056	1.313	1.701	2.048	2.467	2.763
29	0.683	1.055	1.311	1.699	2.045	2.462	2.756
30	0.683	1.055	1.310	1.697	2.042	2.457	2.750
40	0.681	1.050	1.303	1.684	2.021	2.423	2.704
60	0.679	1.045	1.296	1.671	2.000	2.390	2.660
120	0.677	1.041	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.036	1.282	1.645	1.960	2.326	2.576

A.2 Tabela da distribuição Qui-quadrado



g.l.	0.995	0.990	0.975	0.950	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169

A.3 Tabela da distribuição F



g.l.		ν_1 1	2	3	4	5	6	7	8	9	10
ν_2 1	10%	39.863	49.500	53.593	55.833	57.240	58.204	58.906	59.439	59.858	60.195
	5%	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
	2.5%	647.789	799.500	864.163	899.583	921.848	937.111	948.217	956.656	963.285	968.627
	1%	4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473	6055.847
2	10%	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392
	5%	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
	2.5%	38.506	39.000	39.165	39.248	39.298	39.331	39.355	39.373	39.387	39.398
	1%	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399
3	10%	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230
	5%	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
	2.5%	17.443	16.044	15.439	15.101	14.885	14.735	14.624	14.540	14.473	14.419
	1%	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229
4	10%	4.545	4.325	4.191	4.107	4.051	4.010	3.979	3.955	3.936	3.920
	5%	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
	2.5%	12.218	10.649	9.979	9.605	9.364	9.197	9.074	8.980	8.905	8.844
	1%	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546
5	10%	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297
	5%	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
	2.5%	10.007	8.434	7.764	7.388	7.146	6.978	6.853	6.757	6.681	6.619
	1%	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051
6	10%	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937
	5%	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
	2.5%	8.813	7.260	6.599	6.227	5.988	5.820	5.695	5.600	5.523	5.461
	1%	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874
7	10%	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703
	5%	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
	2.5%	8.073	6.542	5.890	5.523	5.285	5.119	4.995	4.899	4.823	4.761
	1%	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620
8	10%	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538
	5%	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
	2.5%	7.571	6.059	5.416	5.053	4.817	4.652	4.529	4.433	4.357	4.295
	1%	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814
9	10%	3.360	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.440	2.416
	5%	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
	2.5%	7.209	5.715	5.078	4.718	4.484	4.320	4.197	4.102	4.026	3.964
	1%	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257
10	10%	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323
	5%	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
	2.5%	6.937	5.456	4.826	4.468	4.236	4.072	3.950	3.855	3.779	3.717
	1%	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
11	10%	3.225	2.860	2.660	2.536	2.451	2.389	2.342	2.304	2.274	2.248
	5%	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
	2.5%	6.724	5.256	4.630	4.275	4.044	3.881	3.759	3.664	3.588	3.526
	1%	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539

g.l.		ν_1 12	15	20	24	30	40	60	120	∞
ν_2 1	10%	60.705	61.220	61.740	62.002	62.265	62.529	62.794	63.061	63.328
	5%	243.906	245.950	248.013	249.052	250.095	251.143	252.196	253.253	254.313
	2.5%	976.708	984.867	993.103	997.249	1001.414	1005.598	1009.800	1014.020	1018.253
	1%	6106.321	6157.285	6208.730	6234.631	6260.649	6286.782	6313.030	6339.391	6365.833
2	10%	9.408	9.425	9.441	9.450	9.458	9.466	9.475	9.483	9.491
	5%	19.413	19.429	19.446	19.454	19.462	19.471	19.479	19.487	19.496
	2.5%	39.415	39.431	39.448	39.456	39.465	39.473	39.481	39.490	39.498
	1%	99.416	99.433	99.449	99.458	99.466	99.474	99.482	99.491	99.499
3	10%	5.216	5.200	5.184	5.176	5.168	5.160	5.151	5.143	5.134
	5%	8.745	8.703	8.660	8.639	8.617	8.594	8.572	8.549	8.526
	2.5%	14.337	14.253	14.167	14.124	14.081	14.037	13.992	13.947	13.902
	1%	27.052	26.872	26.690	26.598	26.505	26.411	26.316	26.221	26.125
4	10%	3.896	3.870	3.844	3.831	3.817	3.804	3.790	3.775	3.761
	5%	5.912	5.858	5.803	5.774	5.746	5.717	5.688	5.658	5.628
	2.5%	8.751	8.657	8.560	8.511	8.461	8.411	8.360	8.309	8.257
	1%	14.374	14.198	14.020	13.929	13.838	13.745	13.652	13.558	13.463
5	10%	3.268	3.238	3.207	3.191	3.174	3.157	3.140	3.123	3.105
	5%	4.678	4.619	4.558	4.527	4.496	4.464	4.431	4.398	4.365
	2.5%	6.525	6.428	6.329	6.278	6.227	6.175	6.123	6.069	6.015
	1%	9.888	9.722	9.553	9.466	9.379	9.291	9.202	9.112	9.021
6	10%	2.905	2.871	2.836	2.818	2.800	2.781	2.762	2.742	2.722
	5%	4.000	3.938	3.874	3.841	3.808	3.774	3.740	3.705	3.669
	2.5%	5.366	5.269	5.168	5.117	5.065	5.012	4.959	4.904	4.849
	1%	7.718	7.559	7.396	7.313	7.229	7.143	7.057	6.969	6.880
7	10%	2.668	2.632	2.595	2.575	2.555	2.535	2.514	2.493	2.471
	5%	3.575	3.511	3.445	3.410	3.376	3.340	3.304	3.267	3.230
	2.5%	4.666	4.568	4.467	4.415	4.362	4.309	4.254	4.199	4.142
	1%	6.469	6.314	6.155	6.074	5.992	5.908	5.824	5.737	5.650
8	10%	2.502	2.464	2.425	2.404	2.383	2.361	2.339	2.316	2.293
	5%	3.284	3.218	3.150	3.115	3.079	3.043	3.005	2.967	2.928
	2.5%	4.200	4.101	3.999	3.947	3.894	3.840	3.784	3.728	3.670
	1%	5.667	5.515	5.359	5.279	5.198	5.116	5.032	4.946	4.859
9	10%	2.379	2.340	2.298	2.277	2.255	2.232	2.208	2.184	2.159
	5%	3.073	3.006	2.936	2.900	2.864	2.826	2.787	2.748	2.707
	2.5%	3.868	3.769	3.667	3.614	3.560	3.505	3.449	3.392	3.333
	1%	5.111	4.962	4.808	4.729	4.649	4.567	4.483	4.398	4.311
10	10%	2.284	2.244	2.201	2.178	2.155	2.132	2.107	2.082	2.055
	5%	2.913	2.845	2.774	2.737	2.700	2.661	2.621	2.580	2.538
	2.5%	3.621	3.522	3.419	3.365	3.311	3.255	3.198	3.140	3.080
	1%	4.706	4.558	4.405	4.327	4.247	4.165	4.082	3.996	3.909
11	10%	2.209	2.167	2.123	2.100	2.076	2.052	2.026	2.000	1.972
	5%	2.788	2.719	2.646	2.609	2.570	2.531	2.490	2.448	2.405
	2.5%	3.430	3.330	3.226	3.173	3.118	3.061	3.004	2.944	2.883
	1%	4.397	4.251	4.099	4.021	3.941	3.860	3.776	3.690	3.603

g.l.		ν_1 1	2	3	4	5	6	7	8	9	10
ν_2 12	10%	3.177	2.807	2.606	2.480	2.394	2.331	2.283	2.245	2.214	2.188
	5%	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
	2.5%	6.554	5.096	4.474	4.121	3.891	3.728	3.607	3.512	3.436	3.374
	1%	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296
13	10%	3.136	2.763	2.560	2.434	2.347	2.283	2.234	2.195	2.164	2.138
	5%	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
	2.5%	6.414	4.965	4.347	3.996	3.767	3.604	3.483	3.388	3.312	3.250
	1%	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100
14	10%	3.102	2.726	2.522	2.395	2.307	2.243	2.193	2.154	2.122	2.095
	5%	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
	2.5%	6.298	4.857	4.242	3.892	3.663	3.501	3.380	3.285	3.209	3.147
	1%	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939
15	10%	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059
	5%	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
	2.5%	6.200	4.765	4.153	3.804	3.576	3.415	3.293	3.199	3.123	3.060
	1%	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
16	10%	3.048	2.668	2.462	2.333	2.244	2.178	2.128	2.088	2.055	2.028
	5%	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
	2.5%	6.115	4.687	4.077	3.729	3.502	3.341	3.219	3.125	3.049	2.986
	1%	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691
17	10%	3.026	2.645	2.437	2.308	2.218	2.152	2.102	2.061	2.028	2.001
	5%	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
	2.5%	6.042	4.619	4.011	3.665	3.438	3.277	3.156	3.061	2.985	2.922
	1%	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593
18	10%	3.007	2.624	2.416	2.286	2.196	2.130	2.079	2.038	2.005	1.977
	5%	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
	2.5%	5.978	4.560	3.954	3.608	3.382	3.221	3.100	3.005	2.929	2.866
	1%	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508
19	10%	2.990	2.606	2.397	2.266	2.176	2.109	2.058	2.017	1.984	1.956
	5%	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
	2.5%	5.922	4.508	3.903	3.559	3.333	3.172	3.051	2.956	2.880	2.817
	1%	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434
20	10%	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965	1.937
	5%	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
	2.5%	5.871	4.461	3.859	3.515	3.289	3.128	3.007	2.913	2.837	2.774
	1%	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
21	10%	2.961	2.575	2.365	2.233	2.142	2.075	2.023	1.982	1.948	1.920
	5%	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
	2.5%	5.827	4.420	3.819	3.475	3.250	3.090	2.969	2.874	2.798	2.735
	1%	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310
22	10%	2.949	2.561	2.351	2.219	2.128	2.060	2.008	1.967	1.933	1.904
	5%	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
	2.5%	5.786	4.383	3.783	3.440	3.215	3.055	2.934	2.839	2.763	2.700
	1%	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258
23	10%	2.937	2.549	2.339	2.207	2.115	2.047	1.995	1.953	1.919	1.890
	5%	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
	2.5%	5.750	4.349	3.750	3.408	3.183	3.023	2.902	2.808	2.731	2.668
	1%	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211

g.l.		ν_1 12	15	20	24	30	40	60	120	∞
ν_2 12	10%	2.147	2.105	2.060	2.036	2.011	1.986	1.960	1.932	1.904
	5%	2.687	2.617	2.544	2.505	2.466	2.426	2.384	2.341	2.296
	2.5%	3.277	3.177	3.073	3.019	2.963	2.906	2.848	2.787	2.725
	1%	4.155	4.010	3.858	3.780	3.701	3.619	3.535	3.449	3.361
13	10%	2.097	2.053	2.007	1.983	1.958	1.931	1.904	1.876	1.846
	5%	2.604	2.533	2.459	2.420	2.380	2.339	2.297	2.252	2.206
	2.5%	3.153	3.053	2.948	2.893	2.837	2.780	2.720	2.659	2.596
	1%	3.960	3.815	3.665	3.587	3.507	3.425	3.341	3.255	3.166
14	10%	2.054	2.010	1.962	1.938	1.912	1.885	1.857	1.828	1.797
	5%	2.534	2.463	2.388	2.349	2.308	2.266	2.223	2.178	2.131
	2.5%	3.050	2.949	2.844	2.789	2.732	2.674	2.614	2.552	2.487
	1%	3.800	3.656	3.505	3.427	3.348	3.266	3.181	3.094	3.004
15	10%	2.017	1.972	1.924	1.899	1.873	1.845	1.817	1.787	1.755
	5%	2.475	2.403	2.328	2.288	2.247	2.204	2.160	2.114	2.066
	2.5%	2.963	2.862	2.756	2.701	2.644	2.585	2.524	2.461	2.395
	1%	3.666	3.522	3.372	3.294	3.214	3.132	3.047	2.959	2.869
16	10%	1.985	1.940	1.891	1.866	1.839	1.811	1.782	1.751	1.718
	5%	2.425	2.352	2.276	2.235	2.194	2.151	2.106	2.059	2.010
	2.5%	2.889	2.788	2.681	2.625	2.568	2.509	2.447	2.383	2.316
	1%	3.553	3.409	3.259	3.181	3.101	3.018	2.933	2.845	2.753
17	10%	1.958	1.912	1.862	1.836	1.809	1.781	1.751	1.719	1.686
	5%	2.381	2.308	2.230	2.190	2.148	2.104	2.058	2.011	1.960
	2.5%	2.825	2.723	2.616	2.560	2.502	2.442	2.380	2.315	2.248
	1%	3.455	3.312	3.162	3.084	3.003	2.920	2.835	2.746	2.653
18	10%	1.933	1.887	1.837	1.810	1.783	1.754	1.723	1.691	1.657
	5%	2.342	2.269	2.191	2.150	2.107	2.063	2.017	1.968	1.917
	2.5%	2.769	2.667	2.559	2.503	2.445	2.384	2.321	2.256	2.187
	1%	3.371	3.227	3.077	2.999	2.919	2.835	2.749	2.660	2.566
19	10%	1.912	1.865	1.814	1.787	1.759	1.730	1.699	1.666	1.631
	5%	2.308	2.234	2.155	2.114	2.071	2.026	1.980	1.930	1.878
	2.5%	2.720	2.617	2.509	2.452	2.394	2.333	2.270	2.203	2.133
	1%	3.297	3.153	3.003	2.925	2.844	2.761	2.674	2.584	2.489
20	10%	1.892	1.845	1.794	1.767	1.738	1.708	1.677	1.643	1.607
	5%	2.278	2.203	2.124	2.082	2.039	1.994	1.946	1.896	1.843
	2.5%	2.676	2.573	2.464	2.408	2.349	2.287	2.223	2.156	2.085
	1%	3.231	3.088	2.938	2.859	2.778	2.695	2.608	2.517	2.421
21	10%	1.875	1.827	1.776	1.748	1.719	1.689	1.657	1.623	1.586
	5%	2.250	2.176	2.096	2.054	2.010	1.965	1.916	1.866	1.812
	2.5%	2.637	2.534	2.425	2.368	2.308	2.246	2.182	2.114	2.042
	1%	3.173	3.030	2.880	2.801	2.720	2.636	2.548	2.457	2.360
22	10%	1.859	1.811	1.759	1.731	1.702	1.671	1.639	1.604	1.567
	5%	2.226	2.151	2.071	2.028	1.984	1.938	1.889	1.838	1.783
	2.5%	2.602	2.498	2.389	2.331	2.272	2.210	2.145	2.076	2.003
	1%	3.121	2.978	2.827	2.749	2.667	2.583	2.495	2.403	2.306
23	10%	1.845	1.796	1.744	1.716	1.686	1.655	1.622	1.587	1.549
	5%	2.204	2.128	2.048	2.005	1.961	1.914	1.865	1.813	1.757
	2.5%	2.570	2.466	2.357	2.299	2.239	2.176	2.111	2.041	1.968
	1%	3.074	2.931	2.781	2.702	2.620	2.535	2.447	2.354	2.256

g.l.		ν_1 1	2	3	4	5	6	7	8	9	10
ν_2 24	10%	2.927	2.538	2.327	2.195	2.103	2.035	1.983	1.941	1.906	1.877
	5%	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
	2.5%	5.717	4.319	3.721	3.379	3.155	2.995	2.874	2.779	2.703	2.640
	1%	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168
25	10%	2.918	2.528	2.317	2.184	2.092	2.024	1.971	1.929	1.895	1.866
	5%	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
	2.5%	5.686	4.291	3.694	3.353	3.129	2.969	2.848	2.753	2.677	2.613
	1%	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129
26	10%	2.909	2.519	2.307	2.174	2.082	2.014	1.961	1.919	1.884	1.855
	5%	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
	2.5%	5.659	4.265	3.670	3.329	3.105	2.945	2.824	2.729	2.653	2.590
	1%	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094
27	10%	2.901	2.511	2.299	2.165	2.073	2.005	1.952	1.909	1.874	1.845
	5%	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
	2.5%	5.633	4.242	3.647	3.307	3.083	2.923	2.802	2.707	2.631	2.568
	1%	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062
28	10%	2.894	2.503	2.291	2.157	2.064	1.996	1.943	1.900	1.865	1.836
	5%	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
	2.5%	5.610	4.221	3.626	3.286	3.063	2.903	2.782	2.687	2.611	2.547
	1%	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032
29	10%	2.887	2.495	2.283	2.149	2.057	1.988	1.935	1.892	1.857	1.827
	5%	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
	2.5%	5.588	4.201	3.607	3.267	3.044	2.884	2.763	2.669	2.592	2.529
	1%	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005
30	10%	2.881	2.489	2.276	2.142	2.049	1.980	1.927	1.884	1.849	1.819
	5%	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
	2.5%	5.568	4.182	3.589	3.250	3.026	2.867	2.746	2.651	2.575	2.511
	1%	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979
40	10%	2.835	2.440	2.226	2.091	1.997	1.927	1.873	1.829	1.793	1.763
	5%	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
	2.5%	5.424	4.051	3.463	3.126	2.904	2.744	2.624	2.529	2.452	2.388
	1%	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801
60	10%	2.791	2.393	2.177	2.041	1.946	1.875	1.819	1.775	1.738	1.707
	5%	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
	2.5%	5.286	3.925	3.343	3.008	2.786	2.627	2.507	2.412	2.334	2.270
	1%	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632
120	10%	2.748	2.347	2.130	1.992	1.896	1.824	1.767	1.722	1.684	1.652
	5%	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910
	2.5%	5.152	3.805	3.227	2.894	2.674	2.515	2.395	2.299	2.222	2.157
	1%	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472
∞	10%	2.706	2.303	2.084	1.945	1.847	1.774	1.717	1.670	1.632	1.599
	5%	3.842	2.996	2.605	2.372	2.214	2.099	2.010	1.939	1.880	1.831
	2.5%	5.024	3.689	3.116	2.786	2.567	2.408	2.288	2.192	2.114	2.048
	1%	6.635	4.605	3.782	3.319	3.017	2.802	2.640	2.511	2.408	2.321

g.l.		ν_1 12	15	20	24	30	40	60	120	∞
ν_2 24	10%	1.832	1.783	1.730	1.702	1.672	1.641	1.607	1.571	1.533
	5%	2.183	2.108	2.027	1.984	1.939	1.892	1.842	1.790	1.733
	2.5%	2.541	2.437	2.327	2.269	2.209	2.146	2.080	2.010	1.935
	1%	3.032	2.889	2.738	2.659	2.577	2.492	2.403	2.310	2.211
25	10%	1.820	1.771	1.718	1.689	1.659	1.627	1.593	1.557	1.518
	5%	2.165	2.089	2.007	1.964	1.919	1.872	1.822	1.768	1.711
	2.5%	2.515	2.411	2.300	2.242	2.182	2.118	2.052	1.981	1.906
	1%	2.993	2.850	2.699	2.620	2.538	2.453	2.364	2.270	2.170
26	10%	1.809	1.760	1.706	1.677	1.647	1.615	1.581	1.544	1.504
	5%	2.148	2.072	1.990	1.946	1.901	1.853	1.803	1.749	1.691
	2.5%	2.491	2.387	2.276	2.217	2.157	2.093	2.026	1.954	1.878
	1%	2.958	2.815	2.664	2.585	2.503	2.417	2.327	2.233	2.132
27	10%	1.799	1.749	1.695	1.666	1.636	1.603	1.569	1.531	1.491
	5%	2.132	2.056	1.974	1.930	1.884	1.836	1.785	1.731	1.672
	2.5%	2.469	2.364	2.253	2.195	2.133	2.069	2.002	1.930	1.853
	1%	2.926	2.783	2.632	2.552	2.470	2.384	2.294	2.198	2.097
28	10%	1.790	1.740	1.685	1.656	1.625	1.592	1.558	1.520	1.478
	5%	2.118	2.041	1.959	1.915	1.869	1.820	1.769	1.714	1.654
	2.5%	2.448	2.344	2.232	2.174	2.112	2.048	1.980	1.907	1.829
	1%	2.896	2.753	2.602	2.522	2.440	2.354	2.263	2.167	2.064
29	10%	1.781	1.731	1.676	1.647	1.616	1.583	1.547	1.509	1.467
	5%	2.104	2.027	1.945	1.901	1.854	1.806	1.754	1.698	1.638
	2.5%	2.430	2.325	2.213	2.154	2.092	2.028	1.959	1.886	1.807
	1%	2.868	2.726	2.574	2.495	2.412	2.325	2.234	2.138	2.034
30	10%	1.773	1.722	1.667	1.638	1.606	1.573	1.538	1.499	1.456
	5%	2.092	2.015	1.932	1.887	1.841	1.792	1.740	1.683	1.622
	2.5%	2.412	2.307	2.195	2.136	2.074	2.009	1.940	1.866	1.787
	1%	2.843	2.700	2.549	2.469	2.386	2.299	2.208	2.111	2.006
40	10%	1.715	1.662	1.605	1.574	1.541	1.506	1.467	1.425	1.377
	5%	2.003	1.924	1.839	1.793	1.744	1.693	1.637	1.577	1.509
	2.5%	2.288	2.182	2.068	2.007	1.943	1.875	1.803	1.724	1.637
	1%	2.665	2.522	2.369	2.288	2.203	2.114	2.019	1.917	1.805
60	10%	1.657	1.603	1.543	1.511	1.476	1.437	1.395	1.348	1.292
	5%	1.917	1.836	1.748	1.700	1.649	1.594	1.534	1.467	1.389
	2.5%	2.169	2.061	1.944	1.882	1.815	1.744	1.667	1.581	1.482
	1%	2.496	2.352	2.198	2.115	2.028	1.936	1.836	1.726	1.601
120	10%	1.601	1.545	1.482	1.447	1.409	1.368	1.320	1.265	1.193
	5%	1.834	1.750	1.659	1.608	1.554	1.495	1.429	1.352	1.254
	2.5%	2.055	1.945	1.825	1.760	1.690	1.614	1.530	1.433	1.311
	1%	2.336	2.192	2.035	1.950	1.860	1.763	1.656	1.533	1.381
∞	10%	1.546	1.487	1.421	1.383	1.342	1.295	1.240	1.169	1.008
	5%	1.752	1.666	1.571	1.517	1.459	1.394	1.318	1.222	1.010
	2.5%	1.945	1.833	1.709	1.640	1.566	1.484	1.388	1.269	1.012
	1%	2.185	2.039	1.878	1.791	1.697	1.592	1.473	1.325	1.015

A.4 Tabela de valores críticos do teste de Mann-Whitney

Notas: 1. Os pares de valores são valores críticos aproximados para testes bilaterais para $\alpha = 0.1$ e $\alpha = 0.05$.

2. Nos testes unilaterais utilizar os valores de $\alpha = 0.1$ para um teste unilateral com $\alpha = 0.05$.

			amostra maior, n_2						
		α	4	5	6	7	8	9	10
amostra menor, n_1	4	0.1	12,24	13,27	14,30	15,33	16,36	17,39	18,42
		0.05	11,25	12,28	12,32	13,35	14,38	15,41	16,44
	5	0.1		19,36	20,40	22,43	23,47	25,50	26,54
		0.05		18,37	19,41	20,45	21,49	22,53	24,56
	6	0.1			28,50	30,54	32,58	33,63	35,67
		0.05			26,52	28,56	29,61	31,65	33,69
	7	0.1				39,66	41,71	43,76	46,80
		0.05				37,68	39,73	41,78	43,83
	8	0.1					52,84	54,90	57,95
		0.05					49,87	51,93	54,98
	9	0.1						66,105	69,111
		0.05						63,108	66,114
	10	0.1							83,127
		0.05							79,131

A.5 Tabela de valores críticos do teste de Wilcoxon

Notas: 1. Os valores da tabela são os valores críticos de T para testes bilaterais para níveis de confiança α. Um valor T é significativo se for menor ou igual aos valores tabelados.

2. Para testes unilaterais, deve-se consultar para o valor de α dividido por 2, considerando $T-$ e $T+$.

n	$\alpha = 0.10$	$\alpha = 0.05$
5	2	-
6	2	0
7	3	2
8	5	3
9	8	5
10	10	8
11	14	10
12	17	13
13	21	17
14	26	21
15	30	25
16	36	29
17	41	34
18	47	40
19	53	46
20	60	52
21	67	58
22	75	65
23	83	73
24	91	81
25	100	89

A.6 Tabela de valores críticos do teste de Kruskal-Wallis (k amostras)

Notas: 1. Na tabela, os valores críticos indicam os valores de nível de significância aproximados, não excedendo o valor nominal de α. Os níveis exatos de significância estão entre parêntesis.

2. Quando a tabela não é aplicável e $k = 3$ com o tamanho dos três grupos menor do que 5 ou $k > 3$ com todos os grupos com tamanho inferior a 4, utilizar a aproximação $c = \chi^2_{\alpha,k-1}$.

Tamanho amostras			α			
			0.10	0.05	0.025	0.01
2	2	2	4.571 (.06667)	- -	- -	- -
3	2	1	4.286 (.10000)	- -	- -	- -
3	2	2	4.500 (.06667)	4.714 (.04762)	- -	- -
3	3	1	4.571 (.10000)	5.143 (.04286)	- -	- -
3	3	2	4.556 (.10000)	5.361 (.03214)	5.556 (.02500)	- -
3	3	3	4.622 (.10000)	5.600 (.05000)	5.956 (.02500)	7.200 (.00357)
4	2	1	4.500 (.07619)	- -	- -	- -
4	2	2	4.458 (.10000)	5.333 (.03333)	5.500 (.02381)	- -
4	3	1	4.056 (.09286)	5.208 (.05000)	5.833 (.02143)	- -
4	3	2	4.511 (.09841)	5.444 (.04603)	6.000 (.02381)	6.444 (.00794)
4	3	3	4.709 (.09238)	5.791 (.04571)	6.155 (.02476)	6.745 (.01000)
4	4	1	4.167 (.08254)	4.967 (.04762)	6.167 (.02222)	6.667 (.00952)
4	4	2	4.555 (.09778)	5.455 (.04571)	6.327 (.02413)	7.036 (.00571)
4	4	3	4.545 (.09905)	5.598 (.04866)	6.394 (.02476)	7.144 (.00970)
4	4	4	4.654 (.09662)	5.692 (.04866)	6.615 (.02424)	7.654 (.00762)
5	2	1	4.200 (.09524)	5.000 (.04762)	- -	- -
5	2	2	4.373 (.08995)	5.160 (.03439)	6.000 (.01852)	6.533 (.00794)
5	3	1	4.018 (.09524)	4.960 (.04762)	6.044 (.01984)	- -
5	3	2	4.651 (.09127)	5.251 (.04921)	6.004 (.02460)	6.909 (.00873)
5	3	3	4.533 (.09697)	5.648 (.04892)	6.315 (.02121)	7.079 (.00866)
5	4	1	3.987 (.09841)	4.985 (.04444)	5.858 (.02381)	6.955 (.00794)
5	4	2	4.541 (.09841)	5.273 (.04877)	6.068 (.02482)	7.205 (.00895)
5	4	3	4.549 (.09892)	5.656 (.04863)	6.410 (.02496)	7.445 (.00974)
5	4	4	4.668 (.09817)	5.657 (.04906)	6.673 (.02429)	7.760 (.00946)
5	5	1	4.109 (.08586)	5.127 (.04618)	6.000 (.02165)	7.309 (.00938)
5	5	2	4.623 (.09704)	5.338 (.04726)	6.346 (.02489)	7.338 (.00962)
5	5	3	4.545 (.09965)	5.705 (.04612)	6.549 (.02436)	7.578 (.00968)
5	5	4	4.523 (.09935)	5.666 (.04931)	6.760 (.02490)	7.823 (.00978)
5	5	5	4.560 (.09952)	5.780 (.04878)	6.740 (.02475)	8.000 (.00946)

Tamanho amostras			α			
			0.10	0.05	0.025	0.01
6	1	1	- -	- -	- -	- -
6	2	1	4.200 (.09524)	4.822 (.04762)	5.600 (.02381)	- -
6	2	2	4.545 (.08889)	5.345 (.03810)	5.745 (.02063)	6.655 (.00794)
6	3	1	3.909 (.09524)	4.855 (.05000)	5.945 (.02143)	6.873 (.00714)
6	3	2	4.682 (.08528)	5.348 (.04632)	6.136 (.02294)	6.970 (.00909)
6	3	3	4.590 (.09773)	5.615 (.04968)	6.436 (.02229)	7.410 (.00779)
6	4	1	4.038 (.09437)	4.947 (.04675)	5.856 (.02424)	7.106 (.00866)
6	4	2	4.494 (.09986)	5.340 (.04906)	6.186 (.02453)	7.340 (.00967)
6	4	3	4.604 (.09997)	5.610 (.04862)	6.538 (.02498)	7.500 (.00966)
6	4	4	4.595 (.09847)	5.681 (.04881)	6.667 (.02495)	7.795 (.00990)
6	5	1	4.128 (.09271)	4.990 (.04726)	5.951 (.02453)	7.182 (.00974)
6	5	2	4.596 (.09807)	5.338 (.04729)	6.196 (.02481)	7.376 (.00982)
6	5	3	4.535 (.09932)	5.602 (.04956)	6.667 (.02452)	7.590 (.00999)
6	5	4	4.522 (.09974)	5.661 (.04991)	6.750 (.02473)	7.936 (.00998)
6	5	5	4.547 (.09835)	5.729 (.04973)	6.788 (.02484)	8.028 (.00988)
6	6	1	4.000 (.09774)	4.945 (.04779)	5.923 (.02381)	7.121 (.00932)
6	6	2	4.438 (.09824)	5.410 (.04993)	6.210 (.02443)	7.467 (.00982)
6	6	3	4.558 (.09948)	5.625 (.04999)	6.725 (.02462)	7.725 (.00985)
6	6	4	4.548 (.09982)	5.724 (.04950)	6.812 (.02458)	8.000 (.00998)
6	6	5	4.542 (.09987)	5.765 (.04993)	6.848 (.02489)	8.124 (.00990)
6	6	6	4.643 (.09874)	5.801 (.04905)	6.889 (.02493)	8.222 (.00994)
7	7	7	4.594 (.09933)	5.819 (.04911)	6.954 (.02446)	8.378 (.00992)
8	8	8	4.595 (.09933)	5.805 (.04973)	6.995 (.02485)	8.465 (.00991)

A.7 Tabela de valores críticos do teste de Kolmogorov-Smirnov

n	0.1	0.05	0.025	0.01	0.005
1	0.9000	0.9500	0.9750	0.9900	0.9950
2	0.6838	0.7764	0.8419	0.9000	0.9293
3	0.5648	0.6360	0.7076	0.7846	0.8290
4	0.4927	0.5652	0.6239	0.6889	0.7342
5	0.4470	0.5094	0.5633	0.6272	0.6685
6	0.4104	0.4680	0.5193	0.5774	0.6166
7	0.3815	0.4361	0.4834	0.5384	0.5758
8	0.3583	0.4096	0.4543	0.5065	0.5418
9	0.3391	0.3875	0.4300	0.4796	0.5133
10	0.3226	0.3687	0.4092	0.4566	0.4889
11	0.3083	0.3524	0.3912	0.4367	0.4677
12	0.2958	0.3382	0.3754	0.4192	0.4490
13	0.2847	0.3255	0.3614	0.4036	0.4325
14	0.2748	0.3142	0.3489	0.3897	0.4176
15	0.2659	0.3040	0.3376	0.3771	0.4042
16	0.2578	0.2947	0.3273	0.3657	0.3920
17	0.2504	0.2863	0.3180	0.3553	0.3809
18	0.2436	0.2785	0.3094	0.3457	0.3706
19	0.2373	0.2714	0.3014	0.3369	0.3612
20	0.2316	0.2647	0.2941	0.3287	0.3524
21	0.2262	0.2586	0.2872	0.3210	0.3443
22	0.2212	0.2528	0.2809	0.3139	0.3367
23	0.2165	0.2475	0.2749	0.3073	0.3295
24	0.2120	0.2424	0.2693	0.3010	0.3229
25	0.2079	0.2377	0.2640	0.2952	0.3166
26	0.2040	0.2332	0.2591	0.2896	0.3106
27	0.2003	0.2290	0.2544	0.2844	0.3050
28	0.1968	0.2250	0.2499	0.2794	0.2997
29	0.1935	0.2212	0.2457	0.2747	0.2947
30	0.1903	0.2176	0.2417	0.2702	0.2899
31	0.1873	0.2141	0.2379	0.2660	0.2853
32	0.1844	0.2108	0.2342	0.2619	0.2809
33	0.1817	0.2077	0.2308	0.2580	0.2768
34	0.1791	0.2047	0.2274	0.2543	0.2728
35	0.1766	0.2018	0.2242	0.2507	0.2690
36	0.1742	0.1991	0.2212	0.2473	0.2653
37	0.1719	0.1965	0.2183	0.2440	0.2618
38	0.1697	0.1939	0.2154	0.2409	0.2584
39	0.1675	0.1915	0.2127	0.2379	0.2552
40	0.1655	0.1891	0.2101	0.2349	0.2521
> 40	$1.07/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.52/\sqrt{n}$	$1.63/\sqrt{n}$

A.8 Tabela de valores críticos do teste de Lillierfors para a Normal

Os valores indicados na tabela são os valores críticos para diferentes tamanhos de amostra.

n	.40	.30	.20	.10	.02
4	.300	.319	.352	.381	.417
5	.285	.299	.315	.337	.405
6	.265	.277	.294	.319	.364
7	.247	.258	.276	.300	.348
8	.233	.244	.261	.285	.331
9	.223	.233	.249	.271	.311
10	.215	.224	.239	.258	.294
11	.206	.217	.230	.249	.284
12	.199	.212	.223	.242	.275
13	.190	.202	.214	.234	.268
14	.183	.194	.207	.227	.261
15	.177	.187	.201	.220	.257
16	.173	.182	.195	.213	.250
17	.169	.177	.189	.206	.245
18	.166	.173	.184	.200	.239
19	.163	.169	.179	.195	.235
20	.160	.166	.174	.190	.231
25	.142	.147	.158	.173	.200
30	.131	.136	.144	.161	.187
> 30	$.736/\sqrt{n}$	$.768/\sqrt{n}$	$.805/\sqrt{n}$	$0886/\sqrt{n}$	$1.031/\sqrt{n}$

A.9 Tabela de valores críticos do teste de Lillierfors para a Exponencial

Os valores indicados na tabela são os valores críticos para diferentes tamanhos de amostra.

n	0.2	0.15	0.1	0.05	0.01
3	.451	.479	.511	.551	.600
4	.396	.422	.449	.487	.548
5	.359	.382	.406	.442	.504
6	.331	.351	.375	.408	.470
7	.309	.327	.350	.382	.442
8	.291	.308	.329	.360	.419
9	.277	.291	.311	.341	.399
10	.263	.277	.295	.325	.380
11	.251	.264	.283	.311	.365
12	.241	.254	.271	.298	.351
13	.232	.245	.261	.287	.338
14	.224	.237	.252	.277	.326
15	.217	.229	.244	.269	.315
16	.211	.222	.236	.261	.306
17	.204	.215	.229	.253	.297
18	.199	.210	.223	.246	.289
19	.193	.204	.218	.239	.283
20	.188	.199	.212	.234	.278
25	.170	.180	.191	.210	.247
30	.155	.164	.174	.192	.226
> 30	$.86/\sqrt{n}$	$.91/\sqrt{n}$	$.96/\sqrt{n}$	$1.06/\sqrt{n}$	$1.25/\sqrt{n}$

A.10 Tabela de valores críticos do teste de Smirnov para duas amostras

A tabela indica os valores críticos para $\alpha = 0.05$ (valor de cima) e $\alpha = 0.01$ (valor de baixo) para diferentes tamanhos de amostra.

* significa que não se rejeita H_0 qualquer que seja o valor da estatística observado.

	n_1									
n_2	3	4	5	6	7	8	9	10	11	12
1	*	*	*	*	*	*	*	*	*	*
	*	*	*	*	*	*	*	*	*	*
2	*	*	*	*	*	16/16	18/18	20/20	22/22	24/24
	*	*	*	*	*	*	*	*	*	*
3	*	*	15/15	18/18	21/21	21/24	24/27	27/30	30/33	30/36
	*	*	*	*	*	24/24	27/27	30/30	33/33	36/36
4		16/16	20/20	20/24	24/28	28/32	28/36	30/40	33/44	36/48
		*	*	24/24	28/28	32/32	32/36	36/40	40/44	44/48
5			*	24/30	30/35	30/40	35/45	40/50	39/55	43/60
			*	30/30	35/35	35/40	40/45	45/50	45/55	50/60
6				30/36	30/42	34/48	39/54	40/60	43/66	48/72
				36/36	36/42	40/48	45/54	48/60	54/66	60/72
7					42/49	40/56	42/63	46/70	48/77	53/84
					42/49	48/56	49/63	53/70	59/77	60/84
8						48/64	46/72	48/80	53/88	60/96
						56/64	55/72	60/80	64/88	68/96
9							54/81	53/90	59/99	63/108
							63/81	70/90	70/99	75/108
10								70/100	60/110	66/120
								80/100	77/110	80/120
11									77/121	72/132
									88/121	86/132
12										96/144
										84/14

Para amostras grandes, o valor crítico aproximado é dado por $c(\alpha)\sqrt{\frac{n_1+n_2}{n_1 n_2}}$, sendo $c(\alpha)$ dado por

α	0.1	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

A.11 Tabela de valores críticos do teste de Spearman

Nota: Na tabela, os valores críticos indicam os valores de nível de significância aproximados, não excedendo o valor nominal de α.

	α					
n	0.10	0.05	0.025	0.01	0.005	0.001
4	1.000	1.000	-	-	-	-
5	0.800	0.900	1.000	1.000	-	-
6	0.657	0.829	0.886	0.943	1.000	-
7	0.571	0.714	0.786	0.893	0.929	1.000
8	0.524	0.643	0.738	0.833	0.881	0.952
9	0.483	0.600	0.700	0.783	0.833	0.917
10	0.455	0.564	0.648	0.745	0.794	0.879
11	0.427	0.536	0.618	0.709	0.755	0.845
12	0.406	0.503	0.587	0.678	0.727	0.818
13	0.385	0.484	0.560	0.648	0.703	0.791
14	0.367	0.464	0.538	0.626	0.679	0.771
15	0.354	0.446	0.521	0.604	0.654	0.750
16	0.341	0.429	0.503	0.582	0.635	0.729
17	0.328	0.414	0.488	0.566	0.618	0.711
18	0.317	0.401	0.472	0.550	0.600	0.692
19	0.309	0.391	0.460	0.535	0.584	0.675
20	0.299	0.380	0.447	0.522	0.570	0.662
21	0.292	0.370	0.436	0.509	0.556	0.647
22	0.284	0.361	0.425	0.497	0.544	0.633
23	0.278	0.353	0.416	0.486	0.532	0.621
24	0.271	0.344	0.407	0.476	0.521	0.609
25	0.265	0.337	0.398	0.466	0.511	0.597
26	0.259	0.331	0.390	0.457	0.501	0.586
27	0.255	0.324	0.383	0.449	0.492	0.576
28	0.250	0.318	0.375	0.441	0.483	0.567
29	0.245	0.312	0.368	0.433	0.475	0.558

(Continuação)

	α					
30	0.240	0.306	0.362	0.425	0.467	0.549
31	0.236	0.301	0.356	0.419	0.459	0.540
32	0.232	0.296	0.350	0.412	0.452	0.532
33	0.229	0.291	0.345	0.405	0.446	0.525
34	0.225	0.287	0.340	0.400	0.439	0.517
35	0.222	0.283	0.335	0.394	0.433	0.510
36	0.219	0.279	0.330	0.388	0.427	0.503
37	0.215	0.275	0.325	0.383	0.421	0.497
38	0.212	0.271	0.321	0.378	0.415	0.491
39	0.210	0.267	0.317	0.373	0.410	0.485
40	0.207	0.264	0.313	0.368	0.405	0.479
41	0.204	0.261	0.309	0.364	0.400	0.473
42	0.202	0.257	0.305	0.359	0.396	0.468
43	0.199	0.254	0.301	0.355	0.391	0.462
44	0.197	0.251	0.298	0.351	0.386	0.457
45	0.194	0.248	0.294	0.347	0.382	0.452
46	0.192	0.246	0.291	0.343	0.378	0.448
47	0.190	0.243	0.288	0.340	0.374	0.443
48	0.188	0.240	0.285	0.336	0.370	0.439
49	0.186	0.238	0.282	0.333	0.366	0.434
50	0.184	0.235	0.279	0.329	0.363	0.430
51	0.182	0.233	0.276	0.326	0.359	0.426
52	0.180	0.231	0.274	0.323	0.356	0.422
53	0.179	0.228	0.271	0.320	0.352	0.418
54	0.177	0.226	0.268	0.317	0.349	0.414
55	0.175	0.224	0.266	0.314	0.346	0.411
56	0.174	0.222	0.264	0.311	0.343	0.407
57	0.172	0.220	0.261	0.308	0.340	0.404
58	0.171	0.218	0.259	0.306	0.337	0.400
59	0.169	0.216	0.257	0.303	0.334	0.397
60	0.168	0.214	0.255	0.301	0.331	0.394