Probabilidades e Estatística

Prof. Caldeira Duarte e Prof^a Anabela Pereira
Departamento de Matemática
da Escola Superior de Tecnologia de Setúbal
do Instituto Politécnico de Setúbal

Fevereiro de 2010

1 VARIÁVEIS ALEATÓRIAS

Em muitas experiências aleatórias os elementos do espaço de resultados, Ω, são números reais ou conjuntos ordenados de números reais. Assim acontece com o registo de temperaturas, da pluviosidade, no lançamento de dois dados, etc. Mas já o resultado do lançamento de uma moeda ao ar não é um resultado numérico.

Quando Ω não é um conjunto numérico atribui-se muitas vezes a cada elemento ω do espaço de resultados, um número real, atribuição essa que pode ser meramente convencional.

Exemplo 1 No lançamento de uma moeda ao ar, o espaço de resultados é o conjunto $\Omega =$ {Cara,Coroa}; é usual neste caso fazer a correspondência

ω	X(ω)
Cara	1
Coroa	0

.

Para o mesmo espaço de resultados Ω, é possível estabelecer diferentes correspondências, consoante os objectivos em estudo. ■

Exemplo 2 Considere-se uma população de empresas das quais se escolhe uma ao acaso. $\Omega = \{\omega_1,\omega_2,...,\omega_m\}$ será o espaço de resultados e m o número total de empresas.
Podemos definir várias correspondências:
$\omega \to X(\omega)$,sendo $X(\omega)$ o número de empregados da empresa ω,
$\omega \to Y(\omega)$,sendo $Y(\omega)$ o volume de vendas da empresa ω,
ou quaisquer outras, conforme o objectivo em estudo. ■

Supondo agora que só estamos interessados no estudo de uma característica dos elementos de Ω, associemos a cada elemento $\omega \in \Omega$ um número real $X(\omega)$. Estamos assim a definir uma função $X : \Omega \to \mathbb{R}$. Sendo A um acontecimento, chama-se imagem de A por X, e representa-se por $X(A)$, ao conjunto dos valores que X assume para os elementos ω de A, isto é,

$$X(A) = \{X(\omega) : \omega \in A\}.$$

Por outro lado, a cada subconjunto $E \subset \mathbb{R}$, pode fazer-se corresponder o subconjunto $X^{-1}(E)$ formado por todos os elementos $\omega \in \Omega$ tais que $X(\omega) \in E$,

$$X^{-1}(E) = \{\omega : X(\omega) \in E\}.$$

A este conjunto $X^{-1}(E)$ chama-se a imagem inversa de E por X.

Exemplo 3 No lançamento de dois dados interessa somente, num dado jogo, a soma dos pontos obtida. Neste caso, o espaço de resultados é o conjunto $\Omega = \{(i,j) : i,j = 1,2,3,4,5,6\}$; defina-se a aplicação $X(i,j) = i + j$. Sendo $A = \{(1,1),(1,2),(2,1)\}$, a imagem de A por X é $X(A) = \{2,3\}$; para o acontecimento $B = \{(4,5),(5,4),(5,5),(6,6)\}$, a imagem de B por X é $X(B) = \{9,10,12\}$. Para o subconjunto real $E_1 = \{2,3\}$, a imagem inversa de E_1 por X é o acontecimento $X^{-1}(E_1) = \{(1,1),(1,2),(2,1)\}$; se $E_2 = [2,+\infty[$, $X^{-1}(E_2) = \Omega$ e se $E_3 = \left]-\infty,\frac{1}{2}\right]$, $X^{-1}(E_3) = \oslash$. ■

Estamos agora em condições de perceber a definição de variável aleatória.

Definição 1 *Uma função real $X(\omega)$ definida no conjunto Ω dos acontecimentos elementares, chama-se uma variável aleatória se a imagem inversa de qualquer intervalo I do eixo real da forma $]-\infty, x]$, é um acontecimento aleatório.*

Nota 1 Uma variável aleatória é uma função e não uma variável no sentido em que é habitualmente empregue em Análise Matemática!

1.1 Funções de Distribuição

Considere-se agora uma variável aleatória X, um intervalo real $E_x =]-\infty, x]$ e a respectiva imagem inversa $X^{-1}(E_x)$. Pela definição de variável aleatória existe sempre $P(X \leq x) = P[X^{-1}(E_x)]$. Como $P(X \leq x)$ depende de x, a igualdade $F_X(x) = P(X \leq x)$ define uma função real de variável real.

Definição 2 *A função $F_X(x)$ definida por $F_X(x) = P(X \leq x)$ chama-se a* ***Função de Distribuição*** *da variável aletória X.*

Exemplo 4 Considerem-se sucessivos lançamentos de um dado. A cada acontecimento elementar, isto é, a cada resultado possível de um lançamento, podemos associar um dos números $1, 2, 3, 4, 5, 6$, o número de pontos que aparecem na face resultante. Aqui a variável aleatória X pode tomar um de seis valores $x_i = i$ $(i = 1, 2, 3, 4, 5, 6)$ com a mesma probabilidade $P(X = x_i) = \frac{1}{6}$. A probabilidade que X seja menor que 1, é evidentemente igual a zero. $P(X < 1) = 0$.
Se x é um número satisfazendo as condições $1 \leq x < 2$,

$$P(X \leq x) = P(X = 1) = \frac{1}{6}.$$

Se $2 \leq x < 3$,

$$P(X \leq x) = P(X = 1) + P(X = 2) = \frac{1}{3}.$$

Se $3 \leq x < 4$,

$$P(X \leq x) = P(X = 1) + P(X = 2) + P(X = 3) = \frac{1}{2}$$

Se $4 \leq x < 5$,

$$P(X \leq x) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = \frac{2}{3}$$

Se $5 \leq x < 6$,

$$\begin{aligned} P(X \leq x) &= P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) \\ &= \frac{5}{6} \end{aligned}$$

Finalmente se $x \geq 6$, tem-se que

$$P(X \leq x) = 1.$$

Resumindo,

$$F(x) = \begin{cases} 0 & , \quad x < 1 \\ 1/6 & , \quad 1 \leq x < 2 \\ 1/3 & , \quad 2 \leq x < 3 \\ 1/2 & , \quad 3 \leq x < 4 \\ 2/3 & , \quad 4 \leq x < 5 \\ 5/6 & , \quad 5 \leq x < 6 \\ 1 & , \quad 6 \leq x \end{cases} .$$

Desenhando o gráfico da função $F(x) = P\left(X \leq x\right)$ deste exemplo, como uma função real da variável real x, obtem-se a função em escada da Figura 1. ■

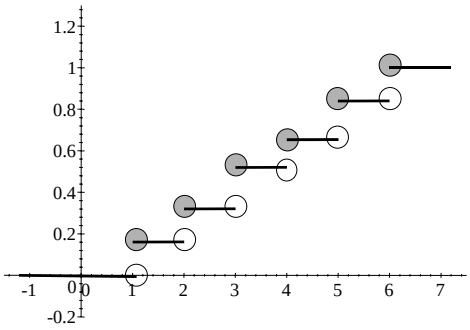


Figura 1: Função de distribuição.

Vamos agora enunciar algumas das propriedades elementares mais importantes das funções de distribuição que o exemplo anterior mostra de uma forma clara. As demonstrações destas propriedades podem ser vistas em [3]. Tem-se então que, se $F(x)$ é uma função de distribuição,

Proposição 2 $0 \leq F(x) \leq 1$.

Proposição 3 *$F(x)$ é uma função não decrescente.*

Proposição 4 $F\left(-\infty\right) = \lim_{x \to -\infty} F(x) = 0$ *e* $F\left(+\infty\right) = \lim_{x \to +\infty} F(x) = 1$.

Proposição 5 $P(a < X \leq b) = F(b) - F(a)$.

Proposição 6 *$F(x)$ é contínua à direita.*

Proposição 7 $P(X = a) = F(a) - \lim_{x \to a^-} F(x)$.

Desta proposição pode concluir-se que, se a função de distribuição é contínua para todo o $x \in \mathbb{R}$, tem-se $P(X = x) = 0$, isto é, todos os pontos reais têm probabilidade zero. Mas atenção: isto não significa forçosamente que os acontecimentos $(X = x)$ sejam impossíveis.

Proposição 8 *O conjunto de pontos de descontinuidade de qualquer função de distribuição, se não fôr vazio, é finito ou infinito numerável.*

Iremos agora tratar fundamentalmente de dois tipos de variáveis aleatórias, as do tipo discreto e as do tipo contínuo.

1.2 Variáveis Aleatórias Discretas

Definição 3 *Seja X uma variável aleatória e D o conjunto*

$$\{a : P(X = a) > 0\}$$

(conjunto dos pontos de descontinuidade da função de distribuição). A variável aleatória X diz-se do tipo ***discreto*** *quando $P(X \in D) = 1$.*

Quando a variável aleatória é discreta existe um conjunto finito ou numerável, $D = \{a_1, a_2, ..., a_n, ...\}$, tal que,

$$P(X \in D) = \sum P(X = a_1) = 1, \text{ e}$$

$$P(X = a_i) > 0, i = 1, 2, ...$$

Definição 4 *Seja D o conjunto definido anteriormente. A função,*

$$f(x) = \begin{cases} > 0 & se \quad x \in D \\ = 0 & se \quad x \in D^C \end{cases}$$

chama-se ***função de probabilidade*** *da v. a. X.*

A função de distribuição de uma variável aleatória discreta pode exprimir-se facilmente em termos da respectiva função de probabilidade:

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i).$$

Exemplo 5 Seja X a variável aleatória que representa o número de caras saídas no lançamento de 3 moedas equilibradas. O quadro seguinte apresenta a função de probabilidade desta variável:

x_i	0	1	2	3
$f(x_i)$	1/8	3/8	3/8	1/8

.

A respectiva função de distribuição (ver Figura 2) será :

$x < 0 \Rightarrow F(X) = P(X \leq x) = P(\emptyset) = 0$
$0 \leq x < 1 \Rightarrow F(X) = P(X = 0) = 1/8 = 0.125$
$1 \leq x < 2 \Rightarrow F(X) = P(X = 0) + P(X = 1) = 1/8 + 3/8 = 0.5$
$2 \leq x < 3 \Rightarrow F(X) = P(X = 0) + P(X = 1) + P(X = 2) = 1/8 + 3/8 + 3/8 = 0.875$
$3 \leq x \Rightarrow F(X) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = P(\Omega) = 1$

1.3 Variáveis Aleatórias Contínuas

Definição 5 *Seja X uma variável aleatória e $F(x)$ a respectiva função de distribuição; se,*

$$D = \{a : P(X = a) > 0\} = \emptyset,$$

resulta da proposição 2.6 que $F(x)$ não apresenta descontinuidades. Se, além disso, existe uma função não negativa, $f(x) \geq 0$, tal que para todo o número real x se verifica a relação,

$$F_X(x) = \int_{-\infty}^{x} f(u)du,$$

então a v. a. X diz-se ***contínua****.*

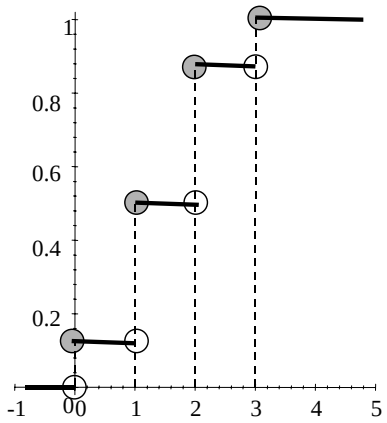


Figura 2: Função de distribuição.

Definição 6 *A função não negativa, $f(x)$, introduzida na definição anterior, chama-se* ***função de densidade de probabilidade*** *ou simplesmente função de densidade.*

Da definição de função de distribuição e da sua relação com a função de densidade, têm-se as seguintes propriedades:

$$f(x) \geq 0;$$

$$\int_{-\infty}^{+\infty} f(x)dx = 1;$$

$$\int_{a}^{b} f(x)dx = F(b) - F(a) = P(a < X < b).$$

Nota 9 Repare-se que $P(a < X < b)$ pode ser interpretada geometricamente como uma área, visto que é calculada através de um integral definido de uma função não negativa (ver Figura 3).

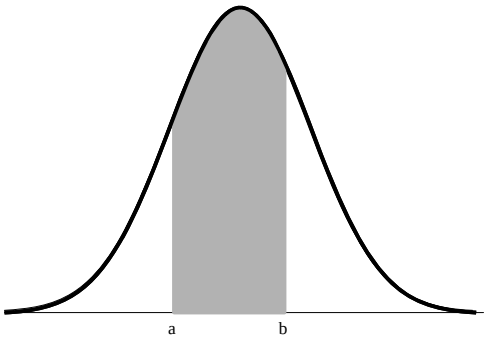


Figura 3: $P\,(a < X < b)$

Exemplo 6 Uma variável aleatória contínua, X, diz-se que tem uma **distribuição uniforme** F no intervalo $[a, b]$ (ver Figura 4), se a sua função de distribuição F, fôr dada pela seguinte expressão:

$$F(t) = \begin{cases} 0 & se \quad t \leq a \\ \frac{t-a}{b-a} & se \quad a < t < b \\ 1 & se \quad t \geq b \end{cases}$$

A derivada $F'(t)$ existe em todos os pontos da recta real excepto em $t = a$ e $t = b$. Então

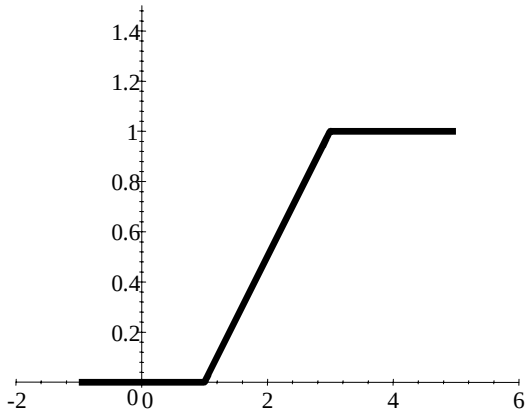


Figura 4: Função de distribuição uniforme no intervalo $[1, 3]$.

$f(t)$, a função de densidade é igual à derivada da função de distribuição em todos os pontos onde exista (ver Figura 5), e convenciona-se que é nula nos restantes; então

$$f(t) = \begin{cases} 1/(b-a) & se \quad a < t < b \\ 0 & se \quad t \leq a \vee t \geq b \end{cases}.$$

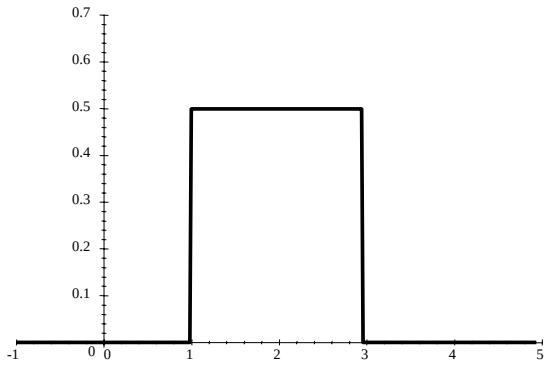


Figura 5: Função de densidade uniforme no intervalo $[1, 3]$.

1.4 Valores Esperados e Parâmetros

De acordo com [3], o conceito de valor esperado teve a sua origem nos jogos de acaso e foi, segundo se diz, introduzido por Huygens.

Exemplo 7 Considere-se um sorteio em que foram vendidos 10000 bilhetes e cujos prémios são:

1) um 1^o prémio de 500000$00
2) um 2^o prémio de 60000$00
3) três 3^o prémios de 30000$00
4) dez 4^a prémios de 10000$00.

Se fizermos a soma dos produtos dos valores dos prémios que se podem ganhar pelas respectivas probabilidades,

$$500000\$\left(\frac{1}{10000}\right)+60000\$\left(\frac{1}{10000}\right)+30000\$\left(\frac{3}{10000}\right)+10000\$\left(\frac{10}{10000}\right)=75\$$$

obtemos aquilo que se chama valor esperado ou esperança matemática do comprador de um bilhete.
Suponhamos que uma pessoa compra sistematicamente um bilhete de uma lotaria deste tipo. Ao fim n repetições, o ganho total é dado por

$$500000\$.s_n(1)+60000\$.s_n(2)+30000\$.s_n(3)+10000\$.s_n(4),$$

sendo $s_n(i)$ o número de vezes que saiu o i-ésimo prémio, $i=1,2,3,4$.
O ganho médio foi

$$500000\$.\left[s_n(1)/n\right]+60000\$.\left[s_n(2)/n\right]+30000\$.\left[s_n(3)/n\right]+10000\$.\left[s_n(4)/n\right].$$

A interpretação frequencista do conceito de probabilidade sugere que, para n grande, as frequências relativas são praticamente iguais à probabilidade. Isto significa que ao fim de um grande número de jogadas o ganho médio será aproximadamente igual ao valor da chamada esperança matemática. É por isso que se diz que um jogo é equitativo quando o que se paga para nele participar é igual à esperança matemática do jogador: após um grande número de partidas, o ganho médio por partida não se afastará muito do preço pago para participar em cada partida. ■

Com qualquer distribuição de uma variável aleatória estão sempre associados certos números chamados os **parâmetros da distribuição,** que desempenham um relevante papel na Estatística Matemática. Duas importantes famílias de parâmetros de uma distribuição são: os **momentos** e os **parâmetros de ordem**; iremos, no entanto, apenas abordar os primeiros.

Definição 7 *O **momento de ordem** k **em relação à origem** ou momento ordinário de ordem k - k inteiro positivo - de uma variável aleatória é o valor esperado da função $G(X)=X^k$,isto é,*

$$\mu_k=E\left[X^k\right].$$

Se a variável aleatória fôr discreta,

$$E\left[X^k\right]=\sum_i x_i^k p_i;$$

no caso de ser contínua,

$$E\left[X^k\right]=\int_{-\infty}^{+\infty}x^k f(x)dx.$$

Definição 8 *O momento de ordem* 1 *em relação à origem de uma variável aleatória chama-se* ***valor esperado*** *e representa-se por* μ *ou* $E[X]$**.**

Definição 9 *O* ***momento de ordem*** k ***em relação à média*** *ou* ***momento central de ordem*** k *- k inteiro positivo - de uma variável aleatória é o valor esperado da função* $G(X) = (X - \mu)^k$ *,isto é,*

$$E\left[(X-\mu)^k\right].$$

Se a variável aleatória fôr discreta,

$$E\left[(X-\mu)^k\right] = \sum_i (x_i - \mu)^k p_i;$$

no caso de ser contínua,

$$E\left[(X-\mu)^k\right] = \int_{-\infty}^{+\infty} (x-\mu)^k f(x)dx.$$

Definição 10 *O momento central de* 2^a *ordem de uma v. a. X, é chamado* ***variância*** *de X, e representa-se habitualmente por* $V[X]$ *ou* σ^2.

Definição 11 *A raiz quadrada positiva da variância de uma v. a. X,* σ*, chama-se o* ***desvio padrão****.*

O parâmetro σ^2 é uma medida de dispersão da variável aleatória em torno do seu valor esperado. Quanto mais concentrada fôr a distribuição, tanto menor será o valor de σ^2.

O papel do desvio padrão como um parâmetro que mede a dispersão de uma variável aleatória é particularmente claro quando se observa a famosa **desigualdade de Chebyshev**; esta desigualdade obtem-se a partir do teorema seguinte.

Teorema 10 *Se uma variável aleatória X toma apenas valores não negativos e tem valor esperado* $E[X]$*, então para qualquer número positivo K, tem-se*

$$P(X \geq K) \leq \frac{E[X]}{K}. \tag{2}$$

Demonstração.A demonstração será feita apenas para o caso da v. a. ser contínua (o caso discreto é análogo). Tem-se

$$E[X] = \int_0^{+\infty} xf(x)dx \geq \int_K^{+\infty} xf(x)dx \geq K\int_K^{+\infty} f(x)dx = KP(X \geq K),$$

donde, a desigualdade 2.2 sai imediatamente. ■

Teorema 11 *(Desigualdade de Chebyshev). Se X é uma variável aleatória com média* μ *e variância* σ^2*, finita, então, para um qualquer número real* $K > 0$,

$$P(|X-\mu| \geq K\sigma) \leq \frac{1}{K^2}.$$

A importância desta desigualdade advém de ser válida para toda e qualquer variável aleatória que tenha uma variância finita podendo empregar-se mesmo quando não se conhece a distribuição da v.a. (variável aleatória).

Exemplo 8 Supondo que X é uma v.a. não negativa cuja distribuição é desconhecida mas se sabe ser $E[X] = 120$ e $\sigma^2 = 100$, tem-se, por exemplo,

$$P(|X - 120| < 40) = P(80 < X < 160) \geq \frac{15}{16}.$$ ■

A seguir apresentam-se algumas propriedades da Esperança Matemática e da Variância cuja demonstração se deixa como exercício ao aluno.

Proposição 12 *Se X é uma v.a. e a e b são constantes reais*

$$E[aX + b] = aE[X] + b.$$

Proposição 13 *Seja X uma v.a. e $G(X)$ e $H(X)$ funções de X; então*

$$E[G(X) + H(X)] = E[G(X)] + E[H(X)].$$

Proposição 14 *Se X é uma v.a.,*

$$V[X] = E\left[X^2\right] - E^2[X].$$

Proposição 15 *Se X é uma v.a.,*

$$V[X] \geq 0.$$

Proposição 16 *Se X é uma v.a. constante, isto é, se $X \equiv a$, então $V[X] = 0$.*

1.5 Adenda

Se pretendermos analisar e relacionar duas variáveis aleatórias, X e Y, há parâmetros que caracterizam as ligações existentes entre ambas.

1.6 Covariância

A covariância é uma medida da distribuição conjunta dos valores dos desvios de X e Y em relação às respectivas médias; este parâmetro permite descrever o tipo de relação linear (positiva ou negativa) que existe (ou não) entre as variáveis unidimensionais X e Y.

Definição 12 *A **covariância** entre X e Y [$cov(X, Y)$ ou $\sigma_{X,Y}$] define-se como:*

$$cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = \sigma_{X,Y}.$$

Algumas das propriedades da covariância são dadas em seguida.

Proposição 17 *Sejam X e Y duas variáveis aleatórias, então a sua covariância pode calcular-se do seguinte modo,*

$$cov\,(X,Y) = E\,[XY] - E\,[X]\,E\,[Y]$$

Demonstração.

$$\begin{aligned} cov\,(X,Y) &= E\,[(X-\mu_X)\,(Y-\mu_Y)] = \\ &= E\,[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y] = \\ &= E\,[XY] - \mu_X \underbrace{E\,[Y]}_{\mu_Y} - \mu_Y \underbrace{E\,[X]}_{\mu_X} + \mu_X\mu_Y = \\ &= E\,[XY] - E\,[X]\,E\,[Y] \end{aligned}$$

Proposição 18 *Se X e Y são v.a. independentes então,*

$$E\,[XY] = E\,[X]\,E\,[Y] \quad e \quad cov\,(X,Y) = 0.$$

O recíproco deste último resultado pode não ser verdadeiro, isto é sendo $cov\,(X,Y) = 0$ não se pode inferir que X e Y sejam independentes, somente que não existe uma relação linear entre as variáveis. Pode contudo existir uma ligação não linear entre X e Y.

Proposição 19 *Sejam X e Y duas variáveis aleatórias,*

1. $E\,[X \pm Y] = E\,[X] \pm E\,[Y]$;
2. $V\,[X \pm Y] = V\,[X] + V\,[Y] \pm 2cov\,(X,Y)$;
3. *Se X e Y são independentes, então* $V\,[X \pm Y] = V\,[X] + V\,[Y]$.

2 DISTRIBUIÇÕES TEÓRICAS DISCRETAS

2.1 Distribuição Binomial

É frequente uma experiência aleatória consistir na repetição de uma série de provas, cada uma das quais apenas com dois resultados possíveis, que geralmente são designados por sucesso e insucesso; é o que acontece, por exemplo, quando se testam peças que saem de uma linha de montagem, onde cada verificação indica se a peça é defeituosa ou não; quando se lança várias vezes uma moeda regular, etc.

Exemplo 9 Considere-se a experiência aleatória que consiste no lançamento de uma moeda regular ao acaso e em que $\Omega = \{F, C\}$. A realização de 3 lançamentos é uma experiência aleatória que se pode identificar com a combinação de 3 experiências aleatórias idênticas. O espaço de resultados desta outra experiência é portanto o conjunto

$$\Pi = \{FFF, FFC, FCF, CFF, FCC, CFC, CCF, CCC\}.$$

Atendendo a que o resultado de cada lançamento é independente dos restantes, é imediato que $P(\text{saída de 3 caras}) = \left(\frac{1}{2}\right)^3$, $P(\text{saída de 2 caras e 1 coroa}) = 3 \times \left(\frac{1}{2}\right)^3$, etc. ■

Este exemplo é um caso particular de uma sucessão de **provas de Bernoulli**, isto é, de uma sucessão de experiências aleatórias independentes, em cada uma das quais se obtem o acontecimento A, que designamos por sucesso, com probabilidade p, (constante de experiência para experiência), ou o seu complementar, A^C, que designamos por insucesso, com probabilidade $q = 1 - p$.

A uma sucessão de provas de Bernoulli também se chama uma **experiência Binomial**.

Exemplo 10 A probabilidade de que um certo tipo de componente sobreviva a um teste é 3/4. Qual a probabilidade de exactamente duas dessas componentes sobrevivam ao teste, de entre as próximas 5 a serem testadas. Designemos por sucesso o acontecimento "a componente sobrevive ao teste" (simbolicamente S) e por insucesso o acontecimento "a componente não sobrevive ao teste". É imediato que neste caso estamos perante uma experiência binomial. Consideremos primeiramente a probabilidade de obter os 2 sucessos e os 5-2 insucessos por uma determinada ordem, por exemplo, 2 sucessos seguidos de 3 insucessos. Trata-se de calcular a probabilidade da sequência,

$$\underbrace{SS}\,\underbrace{S^C S^C S^C}$$

que é $\left(\frac{3}{4}\right)^2 \left(1 - \frac{3}{4}\right)^3 = \left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right)^3$. Se a ordem fôr outra, a probabilidade mantem-se, desde que o número de sucessos e de insucessos se mantenha. Ora, existem $\binom{5}{2}$ sequências diferentes em que podem ocorrer os 2 sucessos e os 3 insucessos, pelo que a probabilidade pretendida será

$$\binom{5}{2}\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)^3 = 0.0879.$$ ■

Definição 13 *Se uma prova de Bernoulli pode resultar num sucesso, com probabilidade p, ou num insucesso, com probabilidade $1 - p$, então a função de probabilidade da v. a. X, que representa o número de sucessos em n provas independentes, e se designa por* ***variável aleatória binomial****, é dada pela expressão*

$$P(X = x) = b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, ..., n.$$

Proposição 20 *Se X é uma v. a. com uma distribuição binomial, $X \sim b(n,p)$,*

$$E[X] = np$$

e

$$V[X] = np(1-p).$$

A terminar, uma referência a um importante resultado relativo à soma de v. a. com distribuição binomial.

Proposição 21 *Se as v. a. X_i, $i = 1, 2, ..., k$ são independentes*[1] *e além disso, $X_i \sim b(n_i,p)$, então a v. a. $Y_k = (X_1 + X_2 + ... + X_k) \sim b\left(\sum n_i, p\right)$.*

2.2 Distribuição de Poisson

Definição 14 *Uma variável aleatória que assume valores da sucessão infinita $0, 1, 2, 3, ...$, com probabilidades,*

$$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, 3, ..., \lambda > 0,$$

*diz-se que tem **distribuição de Poisson** com parâmetro λ, escrevendo-se simbolicamente $X \sim p(\lambda)$.*

Tendo em conta que a soma da série $\sum_{x=0}^{\infty} \frac{\lambda^x}{x!}$ é a função e^{λ}, é imediato que

$$\sum_{x=0}^{\infty} f(x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1.$$

Proposição 22 *O valor esperado e a variância de uma variável aleatória com distribuição de Poisson são iguais ao valor do parâmetro λ.*

Demonstração.

i) $E[X] = \sum_{x=0}^{\infty} x\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} x\frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=1}^{\infty} \lambda\frac{\lambda^{x-1}}{(x-1)!} = \lambda.$

ii)O segundo momento em relação à origem,

$$\begin{aligned} E\left[X^2\right] &= \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} x \frac{\lambda^{x-1}}{(x-1)!} = \\ &= \lambda e^{-\lambda} \sum_{x=1}^{\infty} (x-1) \frac{\lambda^{x-1}}{(x-1)!} + \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda^2 + \lambda \end{aligned}$$

[1] Sem preocupações de grande rigor, introduzimos aqui o conceito de variáveis aleatórias independentes. Sejam X e Y duas variáveis aleatórias discretas; X e Y dizem-se independentes se e só se, para quaisquer valores x e y, os acontecimentos $\begin{cases} X = x \\ Y = y \end{cases}$ forem independentes. Isto significa que

$$\begin{array}{l} P(X = x \,|Y = y) = P(X = x) \\ P(Y = y \,|X = x) = P(Y = y) \end{array}.$$

permite-nos concluir que

$$V[X] = E\left[X^2\right] - E^2[X] = \lambda^2 + \lambda - \lambda^2 = \lambda. \qquad \blacksquare$$

O teorema que se apresenta a seguir estabelece que a função de probabilidade da distribuição de Poisson tambem pode ser obtida como o limite de uma série de funções de probabilidade da distribuição Binomial

Teorema 23 *Seja X_n uma variável aleatória com distribuição Binomial dada pela fórmula $P(X_n = r) = \binom{n}{r} p^r (1-p)^{n-r}$, onde r toma os valores $0, 1, 2, ..., n$. Se para $n = 1, 2, 3, ...$ a relação $p = \frac{\lambda}{n}$ se mantém, onde $\lambda > 0$ é uma constante, então*

$$\lim_{n\to\infty} P(X_n = r) = \frac{\lambda^r}{r!} e^{-\lambda}.$$

Demonstração.Fazendo $p = \frac{\lambda}{n}$, vem,

$$\begin{aligned}
\binom{n}{r} p^r (1-p)^{n-r} &= \frac{n!}{r!\,(n-r)!} \left(\frac{\lambda}{n}\right)^r \left(1 - \frac{\lambda}{n}\right)^{n-r} = \\
&= \frac{n(n-1)...(n-r+1)}{n^r} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-r} \frac{\lambda^r}{r!} = \\
&= \frac{\lambda^r}{r!} . \left(1 - \frac{\lambda}{n}\right)^n . \frac{1\left(1 - \frac{1}{n}\right)...\left(1 - \frac{r+1}{n}\right)}{\left(1 - \frac{\lambda}{n}\right)^r}.
\end{aligned}$$

Como

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$$

e

$$\lim_{n\to\infty} \frac{1\left(1 - \frac{1}{n}\right)...\left(1 - \frac{r+1}{n}\right)}{\left(1 - \frac{\lambda}{n}\right)^r} = 1,$$

obtem-se o resultado pretendido. $\blacksquare$

Este resultado tem grandes aplicações práticas pois, como as figuras seguintes sugerem, a distribuiçao de Poisson pode ser considerada em certas circunstâncias, uma boa aproximação da distribuição Binomial.

Na Figura 6 são apresentados dois gráficos, um da distribuição Binomial com $n = 5$ e $p = 0.3$, donde $\lambda = np = 1.5$, e um da distribuição de Poisson com o mesmo valor esperado $\lambda = 1.5$.

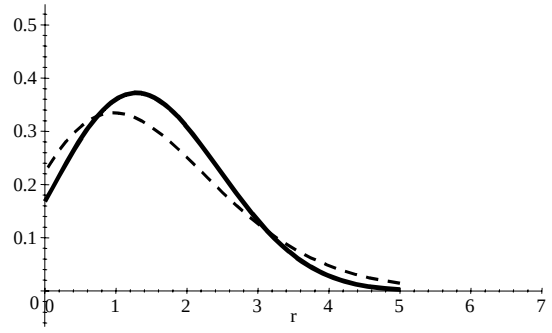


Figura 6: Distribuição DBinomial e Poisson ($n = 5$).

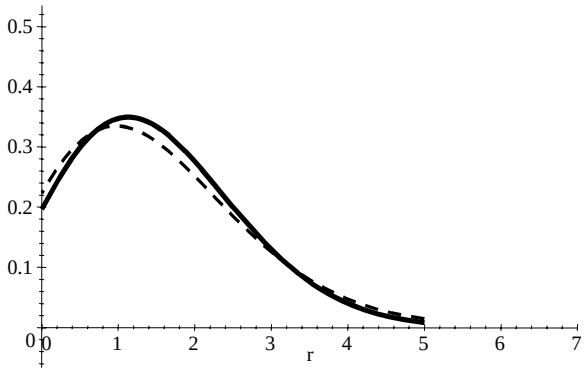


Figura 7: Distribuição Binomial e Poisson ($n = 10$).

A Figura 7 apresenta dois gráficos idênticos mas com $n = 10$ e $p = 0.15$, donde se mantem $\lambda = np = 1.5$.

Para maiores valores de n, por exemplo $n = 100$, os gráficos das distribuições Binomial e Poisson quase coincidem.

Exemplo 11 Numa comunidade com 10000 pessoas, a probabilidade de uma pessoa, num determinado dia, procurar uma cama no hospital, supõe-se igual a 1/2000. Havendo independência na procura de camas em cada dia (inexistência de epidemias, de doenças contagiosas, etc) a v.a. X que representa o número de camas procuradas em cada dia tem uma distribuição binomial com $n = 10000$ e $p = 1/2000$. Neste caso, o cálculo de qualquer probabilidade, $P(X = x)$, deixa de ser imediato. Como n é grande e p muito pequeno podemos calcular valores aproximados dessas probabilidades utilizando a distribuição de Poisson; $b(x; n, p) \approx p(x; np)$. Como exercício o aluno pode calcular alguns valores e comparar os resultados. ■

Nota 24 Na prática, se na distribuição binomial $n \geq 30$ e $np \leq 5$, pode fazer-se a aproximação pela distribuição de Poisson com parâmetro np.

3 DISTRIBUIÇÕES TEÓRICAS CONTÍNUAS

3.1 Distribuição Exponencial

Definição 15 *Uma variável aleatória X tem uma distribuição Exponencial, com parâmetro θ, se a sua função de densidade é dada pela fórmula*

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & , \theta > 0 \wedge x \geq 0 \\ 0 & , x < 0 \end{cases} .$$

Quando uma variável aleatória X tem uma distribuição Exponencial, de parâmetro θ, escreve-se simbolicamente $X \backsim E(\theta)$.

A correspondente função de distribuição tem a seguinte forma

$$F(x) = \begin{cases} 0 & , x < 0 \\ 1 - e^{-\frac{x}{\theta}} & , x \geq 0 \end{cases} .$$

Nas Figuras 8 e 9 estão representados os gráficos das funções de densidade e distribuição de uma distribuição Exponencial para $-1 \leq x \leq 5$ e com parâmetros $\theta = 1$ (tracejado) e $\theta = 3$ (contínuo):

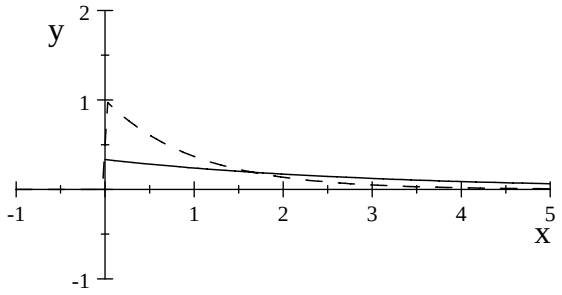


Figura 8: Função de densidade.

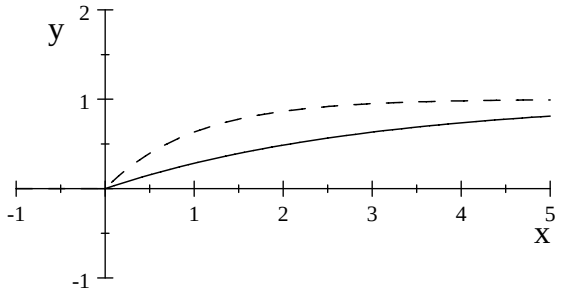


Figura 9: Função de distribuição.

Proposição 25 Se $X \backsim E(\theta)$ tem média e variância dadas, respectivamente, por

$$E[X] = \theta \qquad e \qquad V[X] = \theta^2 \ .$$

O modelo exponencial aplica-se frequentemente quando se pretende estudar tempos até à ocorrência de falhas, por exemplo em componentes electrónicos, em que se admite que o tempo que a componente vai durar é independente do tempo que esta já durou. Isto sigifica que um componente com tempo de vida exponencial tem a mesma qualidade ao longo do tempo, ou seja verifica-se a propriedade

$$P(X \geq a + b / X \geq a) = P(X \geq b)$$

Exemplo 12 Considere a variável aleatória X que representa o tempo de vida, em dias, de um dado tipo de componentes electrónicas. Esta varíavel tem a seguinte função densidade de probabilidade

$$f(x) = \begin{cases} \frac{1}{365} e^{-\frac{x}{365}} & , x \geq 0 \\ 0 & , x < 0 \end{cases} .$$

Suponha que um aparelho é constituído por três destas componentes, com comportamentos independentes entre si, e o aparelho só funciona se pelo menos duas das componentes não falham. Qual a probabilidade de que o aparelho funcione, sem falhas, pelo menos durante dois anos?

Definindo F - aparelho funciona e C - componente funciona, temos:

$$P(C) = 1 - P\left(\bar{C}\right) = 1 - \left(1 - e^{-\frac{730}{365}}\right) = e^{-2} = 0.13534$$

e

$$P(C) = 1 - P\left(\bar{C}\right) = 1 - \left(1 - e^{-\frac{730}{365}}\right) = e^{-2} = 0.13534$$

logo

$$\begin{aligned} P(F) &= P(C \cap C \cap C) + 3P\left(C \cap C \cap \bar{C}\right) = [P(C)]^3 + 3[P(C)]^2 P\left(\bar{C}\right)) = \\ &= (0.13534)^3 + 3(0.13534)^2 (0.86466) = 0.049993. \end{aligned}$$

■

3.2 Distribuição Normal

A distribuição Normal é de grande importância na teoria das probabilidades e na estatística. Na natureza e na tecnologia são inúmeros os fenómenos que apresentam características idênticas às de uma distribuição normal. Exemplos disso são, a medição da altura das pessoas de uma grande população, os erros encontrados quando se fazem muitas medições, etc. Na física, a lei das velocidades de Maxwell implica que a função de distribuição da velocidade numa dada direcção de uma molécula de massa M num gás à temperatura absoluta T, é normal com média 0 e variância $M/(kT)$, onde k é uma constante. Alem disso, sob hipóteses bastantes gerais, a distribuição normal é a distribuição limite para somas de variáveis aleatórias independentes quando o número de termos tende para infinito. Esta distribuição também é conhecida por distribuição de Gauss em homenagem ao matemático alemão Carl Gauss (1777-1855) que deduziu a sua equação.

Definição 16 *Uma variável aleatória X tem uma* ***distribuição Normal*** *se a sua função de densidade é dada pela fórmula*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \ \textit{onde } \sigma > 0 \ e \ -\infty < \mu < +\infty.$$

A distribuição Normal é definida a partir de dois parâmetros: μ e σ; demonstra-se que μ representa o valor esperado de X, e σ, o seu desvio padrão.

Quando uma variável aleatória X tem uma distribuição Normal escreve-se simbolicamente $X \sim \mathcal{N}(\mu; \sigma)$.

Nas Figuras 10 e 11 estão representados os gráficos de várias funções de densidade da distribuição Normal.

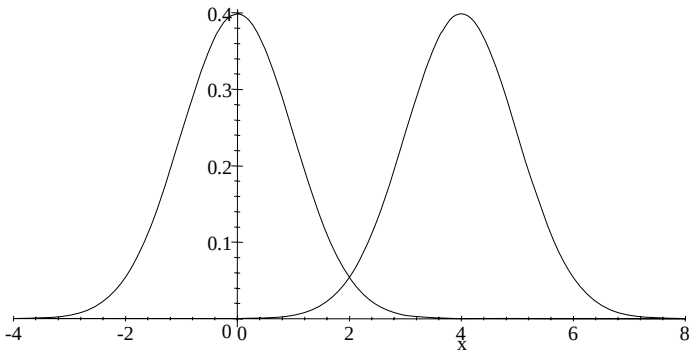


Figura 10: Função densidade da Normal $\mathcal{N}(0; 1)$ e $\mathcal{N}(4; 1)$.

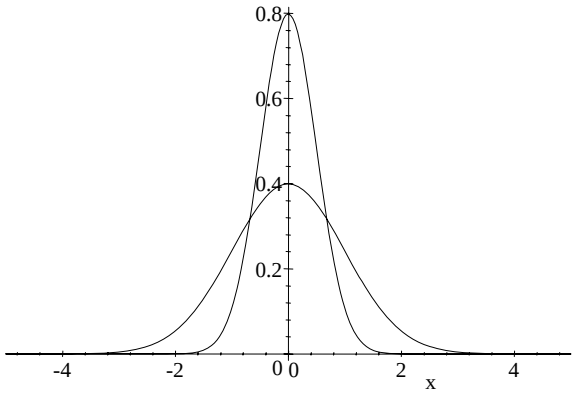


Figura 11: Função de densidade da Normal $\mathcal{N}(0; 1)$ e $\mathcal{N}(0; 0.5)$.

O estudo da função $f(x)$ permite concluir que é simétrica relativamente à recta $x = \mu$, atinge um máximo absoluto no ponto $x = \mu$, tem dois pontos de inflexão em $x = \mu \pm \sigma$ e que o eixo OX é uma assímptota horizontal ao seu gráfico.

Pode demonstrar-se que se X é uma variável aleatória com uma distribuição Normal, $X \sim \mathcal{N}(\mu; \sigma)$, a variável transformada

$$Z = \frac{X - \mu}{\sigma}$$

tem também uma distribuição Normal de média 0 e desvio padrão 1, $Z \sim \mathcal{N}(0; 1)$.

Este resultado é particularmente importante pois a função de distribuição Normal no caso especial $\mu = 0$ e $\sigma = 1$, encontra-se largamente tabelada; é a chamada distribuição Normal estandartizada ou padronizada. Neste caso, a função de distribuição (ver Figura 12)é habitualmente

representada pela letra Φ.

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx.$$

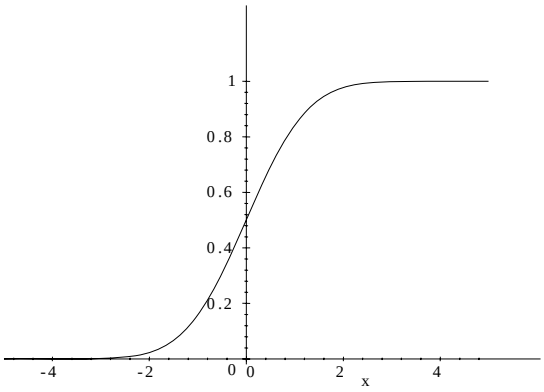


Figura 12: Função de distribuição $\mathcal{N}(0;1)$.

A variável Z designa-se por variável normal padronizada ou reduzida.
Para obter $P(a < X < b)$, sendo $X \sim \mathcal{N}(\mu;\sigma)$, basta notar,

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right),$$

e, portanto,

$$P(a < X < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Exemplo 13 A variável aleatória X tem uma distribuição $\mathcal{N}(1;2)$. Determine a probabilidade de X ser maior que 3 em valor absoluto. O que se pretende é calcular $P(|X| > 3)$. Vamos primeiro centrar e reduzir a variável aleatória X, isto é, vamos transformá-la numa outra de média $\mu = 0$ e desvio padrão $\sigma = 1$. A transformação a utilizar será definida por

$$Z = \frac{X-\mu}{\sigma}. \Leftrightarrow X = \sigma Z + \mu.$$

Neste caso $X = 2Z + 1$. Tem-se então que

$$\begin{aligned} P(|X| > 3) &= P(|2Z+1| > 3) = \\ &= P(2Z+1 > 3) + P(2Z+1 < -3) = \\ &= P(Z > 1) + P(Z < -2). \end{aligned}$$

Como

$$P(Z > 1) = 1 - P(Z \leq 1),$$

tem-se

$$P(|X| > 3) = 1 - P(Z \leq 1) + P(Z < -2).$$

Pela consulta da tabela

$$P(|X| > 3) = 1 - \Phi(1) + \Phi(-2) = 1 - 0.8413 + 0.0228 = 0.1815.$$ ■

Vamos agora referir um importante teorema sobre a distribuição Normal.

Teorema 26 *Se as variáveis aleatórias X_i, $i = 1, ..., n$, são independentes, $X_i \sim \mathcal{N}(\mu_i; \sigma_i)$, então a v.a.*

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i; \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right).$$

Teorema 27 ***(Limite Central)*** *Seja $\bar{X}$ a média de uma amostra aleatória de dimensão n, de uma população de média μ e variância σ^2, então a distribuição da soma,*

$$S_n = X_1 + X_2 + \ldots + X_n$$

ou da média

$$\bar{X} = \frac{S_n}{n}$$

tende para a distribuição Normal quando $n \to \infty$, isto é

$$S_n \sim \mathcal{N}\left(n\mu; \sqrt{n\sigma^2}\right) \to Z = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \sim \mathcal{N}(0; 1) \quad e$$
$$\bar{X} \sim \mathcal{N}\left(\mu; \frac{\sigma}{\sqrt{n}}\right) \to Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0; 1).$$

Nota 28 Na prática a convergência do Teorema do Limite Central é considerada razoável quando $n \geq 30$; quando $n < 30$ a convergência só é razoável se a distribuição for idêntica à Normal.

Exemplo 14 Uma fábrica produz e comercializa rolos de tecido cujo comprimento, em metros, é uma v.a. com valor médio 100 m e variância 156.25 m^2. abendo que o fornecimento é feito em contentores de 200 rolos, calcule a probabilidade de um contentor conter mais de 20.35 km de tecido.

Considerando X - comprimento, em metros, de um rolo de tecido, sabe-se que

$$\mu_X = 100 \text{ metros}, \sigma_X^2 = 156.25 \text{ e } n = 200.$$

Considerando que cada contentor tem 200 rolos, o comprimento de tecido, em metros, de um contentor é dado por $S_{200} = X_1 + X_2 + \ldots + X_{200}$. Aplicando o Teorema 27 temos,

$$S_{200} \sim \mathcal{N}\left(200 \times 100; \sqrt{200 \times 156.25}\right) \to Z = \frac{S_{200} - 20000}{\sqrt{31250}} \sim \mathcal{N}(0; 1).$$

Logo, a probabilidade pretendida é dada por:

$$P(S_{200} > 20350) = P\left(\frac{S_{200} - 20000}{\sqrt{31250}} > \frac{20350 - 20000}{\sqrt{31250}}\right) = P(Z > 1.98) =$$
$$= 1 - P(Z \leq 1.98) = 1 - 0.9761 = 0.0339. \quad \blacksquare$$

3.3 Aproximação da Binomial à Normal

Teorema 29 *Se X é uma v.a. binomial com média $\mu = np$ e variância $\sigma^2 = npq$, então a distribuição da v.a. $X \underset{n\to\infty}{\to} \mathcal{N}(np, \sqrt{npq})$.*

Exemplo 15 Seja X uma v.a. com uma distribuição binomial de parâmetros $n = 15$ e $p = 0.4$, $X \sim b(15, 0.4)$ e $Z = \frac{X-np}{\sqrt{npq}} = \frac{X-6}{1.9} \sim \mathcal{N}(0, 1)$. A probabilidade de a v.a. X ser igual a 4, $P(X = 4) = 0.1268$. Como a distribuição de X é discreta e se pretende obter um valor aproximado desta probabilidade à custa de uma v.a. contínua, onde as probabilidades pontuais são nulas, há que que utilizar o seguinte procedimento (ver Figura 13):

$$P_{binomial}(X = x) \approx P_{normal}(x - 0.5 < X < x + 0.5).$$

Neste caso,

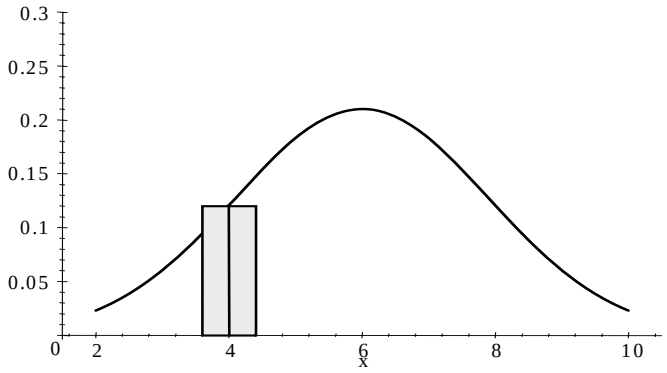


Figura 13: Aproximação da Binomial pela Normal.

$$P_{binomial}(X = 4) \approx P_{normal}(4 - 0.5 < X < 4 + 0.5) = P(3.5 < X < 4.5) =$$
$$= P(\frac{3.5 - 6}{1.9} < Z < \frac{4.5 - 6}{1.9}) = 0.1210,$$

o que fornece já uma boa aproximação. ■

Nota 30 Geralmente a distribuição normal fornece uma boa aproximação da distribuição binomial desde que $n \geq 30$ e p um valor perto de $1/2$. Como regra prática pode utilizar-se o seguinte critério: se tanto np como nq forem maiores que 5, a aproximação será aceitável.

A distribuição Normal poderá ainda ser utilizada para aproximar as distribuições Hipergeométrica e de Poisson sempre que estas, por sua vez, sejam aproximáveis por distribuições Binomiais.

Exemplo 16 Numa empresa multinacional trabalham 5000 pessoas. Seja X a variável aleatória que representa o salário dos funcionários daquela empresa e suponha-se que $X \sim \mathcal{N}(\mu; \sigma)$. Sabendo que metade deles ganham menos de 200 contos e 5% ultrapassam os 250 contos, determine:

1. μ e σ;

2. o melhor salário no grupo dos 2000 empregados pior pagos;

Para determinar μ e σ há que ter em conta que

$$P(X < 200) = 0.50$$

e que

$$P(X > 250) = 0.05.$$

Destas relações conclui-se que

$$P(Z < \frac{200 - \mu}{\sigma}) = 0.50 \tag{3}$$

e

$$P(Z > \frac{250 - \mu}{\sigma}) = 0.05, \tag{4}$$

sendo $Z \sim \mathcal{N}(0; 1)$. De 5.1 tira-se que

$$\frac{200 - \mu}{\sigma} = 0;$$

de 5.2

$$\frac{250 - \mu}{\sigma} = 1.645.$$

Resolvendo o sistema tem-se então,

$$\mu = 200 \text{ e } \sigma = 30.395.$$

Seja agora M o melhor salário no grupo dos 2000 empregados pior pagos. Isto significa que

$$P(X < M) = \frac{2000}{5000}, \tag{5}$$

isto é, a probabilidade de o salário de um indivíduo escolhido ao acaso ser inferior ao melhor salário do grupo dos 2000 empregados pior pagos é 2000/5000. = 0.4. Então, de 5.3 pode concluir-se que

$$\frac{M - 200}{30.395} = -0.7257$$

e, portanto,

$$M = 177.94 \text{ contos.}$$ ■

3.4 Distribuição do Qui-Quadrado - χ^2

Definição 17 *Uma variável aletória X tem uma **distribuição do Qui-Quadrado** com n graus de liberdade, simbolicamente $X \sim \chi^2(n)$, quando a sua função de densidade tem a forma*

$$f(x) = \frac{e^{-\frac{x}{2}} x^{\left(\frac{n}{2}-1\right)}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}, n > 0 \ e \ x > 0.$$

Definição 18 *A função Γ (ver Figura 24) é definida pela expressão*

$$\Gamma(u) = \int_0^\infty e^{-x} x^{u-1} dx, \ com \ u > 0.$$

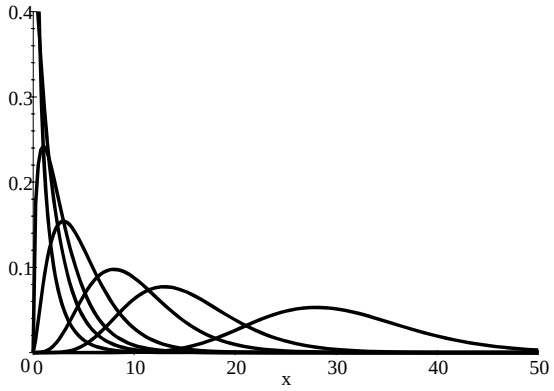


Figura 14: Distribuição do Qui-Quadrado para diferentes graus de liberdade.

A designação *graus de liberdade* dada ao parâmetro n deve-se ao facto de, em certas condições, a distribuição $\chi^2(n)$ descrever o comportamento probabilístico de uma v.a. que pode ser obtida como a soma de $m+n$ v.a., mas em que a existência de m relações lineares entre estas restringe a liberdade desse comportamento.

As distribuições do Qui-Quadrado são caracterizadas por uma dissimetria esquerda.

Proposição 31 *Se* $X \sim \chi^2(n)$,

$$E[X] = n$$

e

$$V[X] = 2n.$$

A distribuição do Qui-Quadrado encontra-se largamente tabelada para valores de $n \leq 30$, e as tabelas são geralmente apresentadas na seguinte forma:

Se $X \sim \chi^2(n)$, a pares(n,ε), para valores de n e ε em domínios convenientes, fazem corresponder o valor χ^2_ε tal que $P\left(X > \chi^2_\varepsilon\right)$=$\varepsilon$.

Por exemplo, para $n = 6$ e $\varepsilon = 0.05$ as tabelas dão $\chi^2_{0.05}(6) = 12.5916$. A probabilidade de um valor observado de χ^2 exceder 12.5916 é portanto 0.05.

Para valores de n maiores que 30, pode usar-se o resultado,

$$\sqrt{2\chi^2(n)} - \sqrt{2n} \dot{\sim} \mathcal{N}(0;1),$$

que significa que a variável aleatória do 1^o membro tem uma distribuição que, quando n tende para infinito, tende para a distribuição $\mathcal{N}(0;1)$.

Um resultado de grande importância na teoria da amostragem é o que apresentamos a seguir

Proposição 32 *Sejam* $X_1, X_2, ..., X_n$, *n variáveis aleatórias independentes com a mesma distribuição;* $X_i \sim \mathcal{N}(0;1)$, $i = 1, 2, ..., n$.
Então $\sum X_i^2 \sim \chi^2(n)$.
Por outras palavras, uma variável aleatória que resulta da soma dos quadrados de n variáveis aleatórias independentes e identicamente distribuidas $(\mathcal{N}(0;1))$, *tem uma distribuição do Qui-Quadrado com n graus de liberdade.*

3.5 Distribuição t de "Student"

Definição 19 *Uma variável aleatória X tem uma distribuição t* ***de "Student"*** *com n graus de liberdade, simbolicamente $X \sim t(n)$, quando a sua função de densidade tem a forma*

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}, -\infty < x < +\infty.$$

Na Figura 15 estão representadas as funções de densidade de três v.a., duas, com a distribuição $t(4)$ e $t(10)$ e a outra, a tracejado, com a distribuição $X \sim \mathcal{N}(0;1)$. Como se vê claramente, quanto maior é o número de graus de liberdade da distribuição t, mais o gráfico da função de densidade de t se aproxima do gráfico da densidade da Normal.

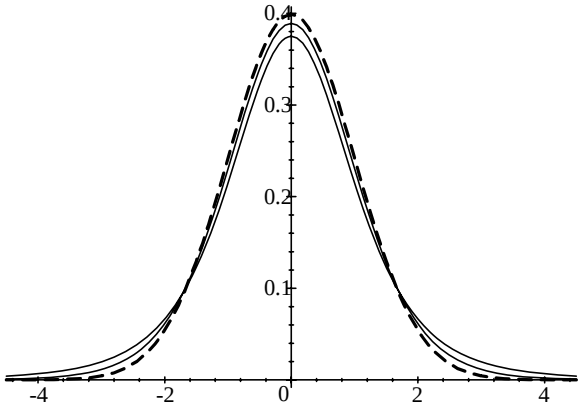


Figura 15: Distribuições t e Normal (a tracejado).

De facto, se $X \sim t(n)$ quando $n \to \infty$, pode-se demonstrar que $X \sim \mathcal{N}(0;1)$.

Proposição 33 *Se $X \sim t(n)$,*

$$E[X] = 0$$

e

$$V[X] = \frac{n}{n-2}, n > 2.$$

As principais aplicações da distribuição de "Student", resultam do teorema seguinte:

Teorema 34 *Se X e Y são variáveis aleatórias independentes, $X \sim \mathcal{N}(0;1)$ e $Y \sim \chi^2(n)$, então,*

$$T = \frac{X}{\sqrt{\frac{Y}{n}}} \sim t(n).$$

3.6 Distribuição F de "Snedcor"

Definição 20 *Uma variável aleatória X tem uma distribuição F* ***de "Snedcor"*** *com m e n graus de liberdade, simbolicamente $X \sim F(m,n)$, quando a sua função de densidade tem a forma*

$$f(x) = \frac{1}{B\left(\frac{m}{2},\frac{n}{2}\right)}.\frac{\left(\frac{m}{n}x\right)^{\frac{m}{2}-1}}{\left(1+\frac{m}{n}x\right)^{\frac{m+n}{2}}}.\frac{m}{n}, m > 0, n > 0, x > 0,$$

sendo a função $B(m,n)$, *(a função Beta), definida por*

$$B(m,n) = \int_0^{+\infty} \frac{\xi^{m-1}}{(1+\xi)^{m+n}} d\xi.$$

O gráfico da função de densidade varia, naturalmente, com os valores de m e n, tamos como podemos ver na Figura 16.

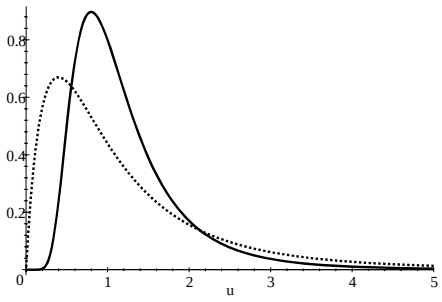


Figura 16: Funções de densidade $F(10,50)$ e $F(8,4)$ (a tracejado).

Proposição 35 *Se a v.a.* $X \sim F(m,n)$,

$$E\{X\} = \frac{n}{n-2}, n > 2$$

e

$$V\{X\} = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, n > 4.$$

A terminar este capítulo relativo às distribuições teóricas três resultados de extrema importância nas aplicações da distribuição F.

Teorema 36 *Se a v.a.* $X \sim F(m,n)$, *então* $\frac{1}{X} \sim F(n,m)$.

Teorema 37 *Se as v.a.* X *e* Y *são independentes,* $X \sim \chi^2(m)$ *e* $Y \sim \chi^2(n)$, *então, se*

$$F = \frac{(X/m)}{(Y/n)},$$

$$F \sim F(m,n).$$

Como consequência imediata deste teorema tem-se o seguinte corolário.

Corolário 38 *Se a v.a.* $X \sim t(n)$, *então* $X^2 \sim F(1,n)$.

As distribuições contínuas a que fizémos referência, as distribuições Exponencial, Normal, do Qui-Quadrado, t de "Student" e F de Snedcor, constituem o suporte teórico de mais larga utilização em questões de inferência estatística.

4 TEORIA DA AMOSTRAGEM

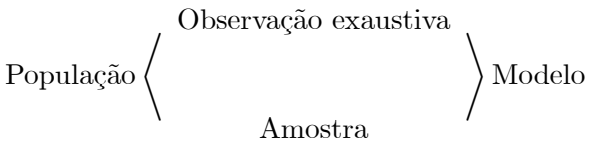
"Adivinhar é barato; adivinhar erradamente sai caro."
Antigo provérbio chinês.

Neste capítulo abordaremos métodos que permitem o cálculo de estimativas de parâmetros de distribuições de variáveis aleatórias que, com alguma credibilidade, se aproximam do verdadeiro valor que se pretende analisar.

Uma estimativa pode calcular-se segundo uma grande variedade de métodos. Pode acrescentar-se que estes métodos não fornecem valores exactos, sendo o erro um factor constante na estimação, podendo no entanto medir-se e ser objectivamente controlado.

4.1 Generalidades

Para estudarmos uma população podemos optar pela sua observação exaustiva ou por seleccionar uma amostra tal como está ilustrado na Tabela seguinte, pretendendo-se ajustar modelos da Teoria das Probabilidades a observações decorrentes de processos aleatórios.



A **Teoria da Amostragem** tem por objectivo retirar conclusões sobre uma dada **população**, quando apenas parte dela foi observada, isto é, a partir de uma **amostra**. Para tal é necessário definir um **Plano Amostral** tal como podemos observar na Figura 17.

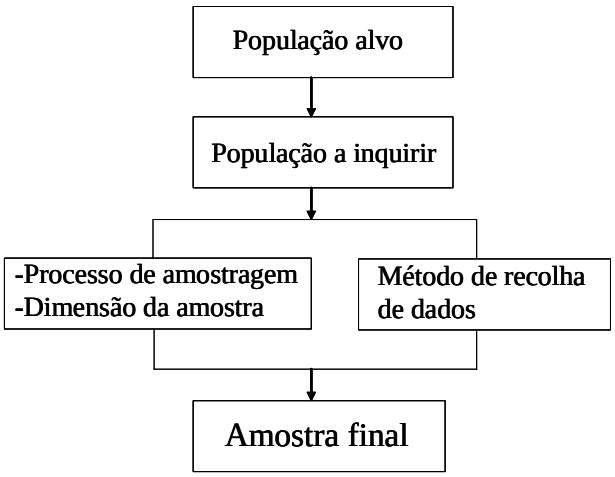


Figura 17: Plano Amostral.

Os processos de amostragem utilizados são da máxima importância, dado que a amostra a constituir tem que ser necessáriamente significativa e representativa da população. Existem vários processos de amostragem e, toda uma teoria sobre o assunto; limitar-nos-emos a descrever sumáriamente cada um dos tipos de amostragem e a respectiva importância estatística.

4.2 Processos de Amostragem

Existem, genericamente, três tipos de amostragem, isto é, três formas de seleccionar uma amostra a partir de uma população:

- **Amostragem Casual ou Aleatória** - em que se deixa completamente ao acaso a escolha dos elementos da população a incluir na amostra, isto é, a probabilidade de um elemento ser incluído na amostra é igual para todos. Existem fundamentalmente dois métodos que permitem a obtenção de amostras casuais:

  - **Método da Lotaria,** em que se atribui a cada elemento da população um número ou simbolo que depois é sorteado;
  - **Método dos Números Aleatórios**, em que se atribui a cada elemento da população um número; em seguida geram-se números aleatórios (por exemplo por computador), seleccionando-se na população os elementos correspondentes;

  Destes dois métodos, o primeiro apresenta mais desvantagens na medida em que é necessário reconstituir toda a população através de números ou outros quaisquer simbolos, assim como se impõe utilizar um processo de sorteio que não esteja sujeito a qualquer vício ou manipulação.

- **Amostragem Dirigida** - a indicação dos elementos componentes da amostra é baseada, essencialmente, no critério ou juízo do investigador. Este tipo de amostragem não tem interesse para a estatística na medida em que:

  - as amostras dirigidas são geralmente enviesadas devido às preferências pessoais do investigador;
  - não podemos medir probabilisticamente a incerteza das inferências a realizar para as populações, dado que não existe qualquer factor de casualidade inerente à constituição da amostra.

- **Amostragem Mista** - neste caso são combinados os dois tipos de amostragem anteriores. A amostragem mista tem vantagens a nível prático, quando se conhecem algumas informações da população; assim sendo define-se uma característica dos elementos a incluir na amostra, deixando-se os restantes factores ao acaso. Neste tipo de amostragem salientam-se os seguintes métodos:

  - **Amostragem Estratificada**, em que se divide a população por estratos e, dentro de cada estrato se retiram elementos ao acaso para a amostra;
  - **Amostragem por Etapas Múltiplas**, quando se analisam conjuntos da população e, em etapas sucessivas, se estudam subconjuntos desses conjuntos. Quando as unidades finais se agrupam de acordo com a sua proximidade geográfica, temos uma amostragem por áreas; quando o agrupamento se faz segundo qualquer outro critério, temos uma amostragem por conglomerado;
  - **Amostragem por Fases Múltiplas**, que consiste em seleccionar, ao acaso, um certo número de elementos de uma população (1^a fase) e, a partir dessa amostra obter-se uma subamostra (2^a fase) e assim sucessivamente. Este processo difere do anterior devido a não existir hierarquia nas unidades de amostragem, isto é, as subamostras seleccionadas na 2^a fase ou fases posteriores são da mesma categoria das encontradas na 1^a fase.

4.3 Estatísticas

Ao seleccionarmos n elementos de uma população, cujos valores observados são os de uma variável aleatória X com função densidade de probabilidade $f(x)$, vamos definir as variáveis aleatórias $X_1, X_2, ..., X_n$ em que X_i representa a i-ésima observação realizada.

As variáveis aleatórias $X_1, X_2, ..., X_n$ constituem uma **amostra aleatória** da população X com valores numéricos respectivamente $x_1, x_2, ..., x_n$.

Exemplo 17 Numa empresa com 50 empregados existem 10 novas tarefas a atribuir. Para seleccionar aleatoriamente que empregados vão desempenhar essa novas tarefas é necessário um mecanismo que permite escolher os empregados e que pode, por exemplo, ser constituído por uma urna onde se inserem 50 papeis, cada um com o nome de um dos empregados; a extracção da urna é realizada de forma perfeitamente casual, podendo seguir-se duas metodologias:

- sem reposição, em que cada empregado seleccionado só pode executar, no máximo, uma nova tarefa; neste caso e antes de qualquer extracção da urna cada um dos 50 empregados tem exactamente a mesma probabilidade de ser seleccionado, sendo esta de $\frac{1}{50}$; após a primeira extracção (em que se retira da urna o papel com o nome do primeiro empregado seleccionado) atribui-se a probabilidade de $\frac{1}{49}$ aos restantes e selecciona-se um novo empregado; este processo continua até à décima extracção em que cada empregado já tem a probabilidade $\frac{1}{41}$ de ser seleccionado e escolhe-se o último empregado;

- com reposição, em que cada empregado pode executar uma ou mais das 10 novas tarefas; neste caso em cada uma das dez extracções consideram-se sempre os 50 empregados, sendo a probabilidade de qualquer um ser escolhido para qualquer uma das novas tarefas igual a $\frac{1}{50}$; para tal basta repor na urna o papel seleccionado em cada extracção. ■

Exemplo 18 Considere o processo de fabrico de um determinado componente electrónico de um carro da marca W. O Departamento de Qualidade está interessado em conhecer a vida útil deste componente. Como a população, constituída pelos componentes que se pretendem analisar, é infinita (considerando que o processo de fabrico dos componentes opera por tempo indeterminado e em circunstâncias idênticas) é necessário proceder a uma amostragem aleatória. Nestas condições, uma forma de determinar uma amostra consiste considerar um subconjunto da população que pode ser obtido, por exemplo, através dos componentes electrónicos produzidos numa qualquer semana. ■

Se tiverem sido realizadas n observações, *independentemente umas das outras e sob as mesmas condições*, as n variáveis aleatórias $X_1, X_2, ..., X_n$ são **independentes e idênticamente distribuídas**.

Apresentam-se as seguintes definições:

Definição 21 *Seja $X_1, X_2, ..., X_n$, uma amostra aleatória de dimensão n da população X com função de densidade $f(x)$, a sua **função densidade de probabilidade conjunta** é dada por:*

$$f(x_1, x_2, ..., x_n) = f(x_1)f(x_2)...f(x_n).$$

Definição 22 *Chama-se **Estatística** a uma variável aleatória que seja apenas função de uma amostra aleatória, que não contenha parâmetros desconhecidos*

$$W = W(X_1, X_2, ..., X_n).$$

Nota 39 A estatística de uma variável aleatória representa-se por $W = W(X_1, X_2, ..., X_n)$ e o seu valor, para uma dada amostra concreta $(x_1, x_2, ..., x_n)$, por $w = w(x_1, x_2, ..., x_n)$. ■

Exemplo 19 Na tabela Seguinte apresentam-se alguns parâmetros e as estatísticas correspondentes.

Parâmetro da População X	Estatística correspondente
$\mu = E[X]$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
$\sigma^2 = E\left[(X-\mu)^2\right]$	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
$\sigma = +\sqrt{\sigma^2}$	$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

■

4.4 Estimadores

O objectivo de um problema estatístico de estimação consiste na avaliação do valor (desconhecido) de um parâmetro. Passamos agora a definir alguns conceitos: **estimador** é qualquer estatística usada para estimar o valor de um parâmetro; **estimativa** de um parâmetro de uma população é qualquer valor específico de uma estatística desse parâmetro; **estimação** é todo o processo que se baseia em utilizar um estimador para produzir uma estimativa do parâmetro.

Exemplo 20 Considere-se a amostra de uma população constante na tabela seguinte

1	1.5	3.2	4	5.1	6	7.3	8.4	9.5	10

Um estimador da média de qualquer amostra de dimensão 10 é dado por

$$\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i.$$

Concretizando para a amostra dada obtemos a estimativa $\bar{x} = 5.6$. ■

A estatística ou estimador representa-se, geralmente, por uma letra maiúscula e a estimativa pela correspondente minúscula.

Para encontrarmos estimativas dum parâmetro θ desconhecido de uma população, a partir de uma amostra, podemos utilizar dois tipos de estimação:

- **Estimação Pontual**, que consiste em encontrar um valor simples ou ponto θ^* (estimador) para θ;
- **Estimação por Intervalos**, que consiste em construir um intervalo de estimação (ou intervalo de confiança) a que θ pertence com uma certa probabilidade conhecida.

O primeiro tipo de estimação fornece-nos um valor simples que, para além de ser muito falível, também não permite uma avaliação da precisão do estimador, isto é, não permite o cálculo da diferença provável entre a estatística e o parâmetro.

No segundo tipo de estimação a *qualidade* de uma estimativa é definida associando-lhe um intervalo (de confiança) tendo uma probabilidade conhecida de conter o verdadeiro valor de θ. Como é óbvio um intervalo de confiança pode não conter o verdadeiro valor de θ, assim como qualquer outra estimativa, porém em contraste com a estimação pontual, a probabilidade de erro para o intervalo de confiança pode ser objectivamente determinada.

Em geral um estimador (ponto) ou região de estimação (intervalo de confiança) devem possuir qualidades óptimas assintóticas, isto é, válidas quando se trabalha com grandes amostras. Vamos passar a enunciar algumas dessas propriedades.

Considerando θ^* um estimador do parâmetro desconhecido θ, é desejável que o valor θ^*, observado a partir de uma amostra seja, com grande probabilidade, um valor vizinho de θ, e como tal, uma boa *estimativa* do mesmo. Conclui-se então que θ^* é um bom estimador de θ se a sua dispersão em torno deste valor for pequena. Assim sendo, pode considerar-se um estimador de um parâmetro como uma sucessão de estatísticas θ_1^*, θ_2^*,...,θ_n^* que convergem em probabilidade para θ à medida que a dimensão da amostra aumenta. A esta propriedade de uma estatística, que permite encará-la como estimador de um parâmetro, dá-se o nome de **convergência** (ou **consistência**). Formalmente:

Definição 23 *Um estimador θ^* diz-se **convergente** ou **consistente** se e só se $\lim_{n\to\infty} E[\theta^*] = \theta$ e $\lim_{n\to\infty} V[\theta^*] = 0$.*

Exemplo 21 Podemos observar no gráfico da Figura 18 o comportamento de um estimador θ^* de θ, convergente ou consistente, o qual, à medida que a dimensão da amostra aumenta tende, em valor médio, para o parâmetro θ, simultâneamente a sua dispersão tende para zero. ■

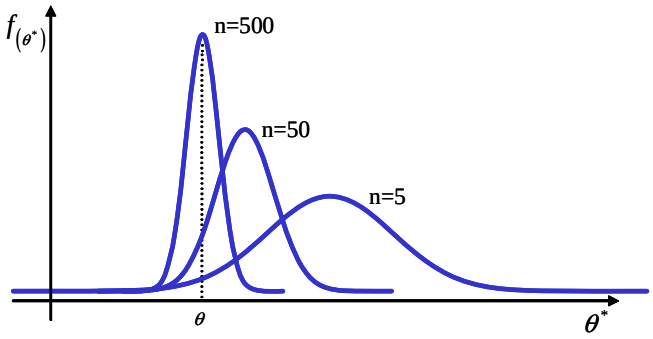


Figura 18: Estimador consistente ou convergente de θ.

Definição 24 *O **desvio** de um estimador é a quantidade $(E[\theta^*] - \theta)$.*

Definição 25 *Um estimador θ^* diz-se **centrado** ou **não enviesado** quando o seu desvio é nulo, isto é, quando $(E[\theta^*] - \theta) = 0 \Leftrightarrow E[\theta^*] = \theta$.*

Exemplo 22 Podemos observar através da Figura 19 que o estimador θ_1^* é não enviesado ou centrado e que o estimador θ_2^* é enviesado ou não centrado, sendo a diferença $(E[\theta_2^*] - \theta)$ correspondente ao enviesamento ou desvio. ■

Definição 26 *Um estimador diz-se **assintoticamente centrado** quando o desvio $(E[\theta^*] - \theta)$ tende para zero à medida que a dimensão da amostra tende para o da população (ou quando $n \to +\infty$).*

É necessário ter em atenção que o facto de um estimador estar concentrado em torno do valor real de um parâmetro pode ser mais importante do que ser centrado, desde que o desvio seja pequeno (para valores grandes de n). Através da Figura 20 seguinte, verificamos empiricamente que é preferível um estimador com pequena dispersão embora não centrado (θ_1^*) a um estimador centrado com grande dispersão (θ_2^*).

Torna-se então necessário encontrar uma forma de medir a dispersão de um estimador face a um ponto dado (geralmente o valor real do parâmetro a estimar). Uma forma possível de medir a dispersão de θ^* em torno de θ é dada por $E\left[(\theta^* - \theta)^2\right]$, logo,

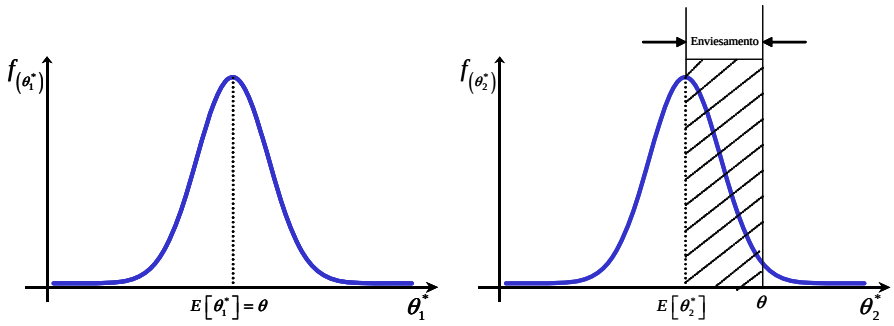


Figura 19: θ_1^* é não enviesado e θ_2^* é enviesado.

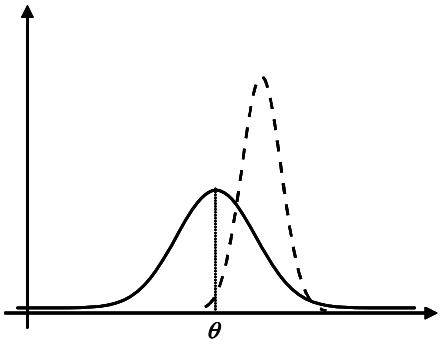


Figura 20: $f_{\left(\theta_1^*\right)}$ a tracejado e $f_{\left(\theta_2^*\right)}$ a contínuo.

Definição 27 *Um estimador θ^* diz-se* ***eficiente*** *se tem $E\left[(\theta^* - \theta)^2\right]$ mínimo.*

Para o caso dos estimadores centrados, o melhor estimador obtem-se muitas vezes pela condição da variância mínima, pois, se θ^* é centrado,

$$E\left[(\theta^* - \theta)^2\right] = V\left[\theta^*\right]$$

isto é, procura-se um estimador θ^* cuja variância seja inferior à de qualquer outro estimador centrado.

Considerando agora dois estimadores em que o primeiro é centrado mas tem uma dispersão considerável e um segundo que embora ligeiramente enviesado tem uma dispersão pequena, é necessário utilizar uma ferramenta que indique qual dos dois é melhor estimador. Para comparar e decidir qual dos dois se deve utilizar é necessário analisar a sua eficiência relativa. Formalmente:

Definição 28 *Dados dois estimadores de θ, θ_1^* e θ_2^*, define-se eficiência relativa de θ_1^* em relação a θ_2^*, pelo quociente*

$$\frac{E\left[(\theta_2^* - \theta)^2\right]}{E\left[(\theta_1^* - \theta)^2\right]}$$

Se este quociente for maior do que a unidade então θ_1^* é mais eficiente do que θ_2^*.

Exemplo 23 Considere uma população de média μ desconhecida e variância igual a σ^2 (conhecida). Suponha que o estimador da média é dado por

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Vamos estudar as qualidades deste estimador e em seguida comparar a sua eficiência com um outro estimador da média.

- **Enviesamento**:

  $\bar{X}$ é um estimador centrado ou não enviesado de μ se $E\left[\bar{X}\right] = \mu$.
  $E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n}n\mu = \mu.$
  Conclui-se que $\bar{X}$ é um estimador centrado ou não enviesado[2].

- **Convergência**:

  $\bar{X}$ é um estimador convergente se se verificar,

$$\lim_{n\to\infty} E\left[\bar{X}\right] = \mu \quad \text{e} \quad \lim_{n\to\infty} V\left[\bar{X}\right] = 0.$$

O primeiro limite, como $E\left[\bar{X}\right] = \mu$, verifica-se imediatamente, pois:

$$\lim_{n\to\infty} E\left[\bar{X}\right] = \lim_{n\to\infty} \mu = \mu.$$

Relativamente ao segundo limite, começamos por calcular a variância de $\bar{X}$:
$V\left[\bar{X}\right] = V\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}V\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V\left[X_i\right] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n},$
sendo,

$$\lim_{n\to\infty} V\left[\bar{X}\right] = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0.$$

Conclui-se que o estimador $\bar{X}$ é convergente.

- **Eficiência relativa**:

  Se considerarmos uma amostra de dimensão n dessa população e

$$\mu^* = \frac{X_1 + 2X_2 + 3X_n}{6}$$

for considerado um estimador da média, podemos comparar a eficiência entre os dois estimadores. Para o efeito vejamos se μ^* é um estimador centrado ou não enviesado, isto é, se $E\left[\mu^*\right] = \mu$:

$$E\left[\mu^*\right] = E\left[\frac{X_1 + 2X_2 + 3X_n}{6}\right] = \frac{1}{6}\left(E\left[X_1\right] + 2E\left[X_2\right] + 3E\left[X_n\right]\right) =$$
$$= \frac{1}{6}\left(\mu + 2\mu + 3\mu\right) = \mu.$$

Conclui-se que μ^* é um estimador centrado ou não enviesado. Como tal, a eficiência relativa de $\bar{X}$ relativamente a μ^* é dada por:

$$\frac{E\left[(\mu^* - \mu)^2\right]}{E\left[(\bar{X} - \mu)^2\right]} = \frac{V\left[\mu^*\right]}{V\left[\bar{X}\right]}$$

[2]Note-se que para X_i variáveis aleatórias independentes temos:
$E\left[\sum_{i=1}^{n} X_i\right] = E\left[X_1\right] + E\left[X_2\right] + \ldots + E\left[X_n\right] = \sum_{i=1}^{n} E\left[X_i\right]$
e
$V\left[\sum_{i=1}^{n} X_i\right] = V\left[X_1\right] + V\left[X_2\right] + \ldots + V\left[X_n\right] = \sum_{i=1}^{n} V\left[X_i\right].$

$$V[\mu^*] = V\left[\frac{X_1+2X_2+3X_n}{6}\right] = \frac{1}{36}\left(V[X_1]+4V[X_2]+9V[X_n]\right) =$$
$$= \frac{1}{36}\left(\sigma^2+4\sigma^2+9\sigma^2\right) = \frac{14\sigma^2}{36}.$$

Logo,

$$\frac{E\left[(\mu^*-\mu)^2\right]}{E\left[(\bar{X}-\mu)^2\right]} = \frac{V[\mu^*]}{V\left[\bar{X}\right]} = \frac{\frac{14\sigma^2}{36}}{\frac{\sigma^2}{n}} = \frac{14n}{36}.$$

Como se verifica, a eficiência relativa depende da dimensão da amostra. Se pretendermos ser mais específicos procedemos do seguinte modo:

. se $\frac{E\left[(\mu^*-\mu)^2\right]}{E\left[(\bar{X}-\mu)^2\right]} < 1 \Leftrightarrow \frac{14n}{36} < 1 \Rightarrow n \leq 2$, μ^* é mais eficiente;

. se $\frac{E\left[(\mu^*-\mu)^2\right]}{E\left[(\bar{X}-\mu)^2\right]} > 1 \Leftrightarrow \frac{14n}{36} > 1 \Rightarrow n \geq 3$, $\bar{X}$ é mais eficiente. ■

Exemplo 24 São propostos os seguintes estimadores para a variância de uma população normal de média μ conhecida,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \text{ e } S'^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

Vamos proceder ao estudo das seguintes propriedades: enviesamento e convergência. Começando por analisar S^2 :

Enviesamento: $E[S^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \frac{n}{n-1}E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] =$

$= \frac{n}{n-1}E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i)^2-\bar{X}^2\right] = \frac{n}{n-1}\left(E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i)^2\right]-E\left[\bar{X}^2\right]\right) =$

$= \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}E\left[(X_i)^2\right]-E\left[\bar{X}^2\right]\right) = \frac{n}{n-1}\left(\frac{n}{n}E\left[X^2\right]-E\left[\bar{X}^2\right]\right)$

Dado que:

$V[X] = \sigma^2 = E[X^2]-E[X]^2 \Leftrightarrow \sigma^2 = E[X^2]-\mu^2 \Leftrightarrow E[X^2] = \sigma^2+\mu^2$

$V[\bar{X}] = V\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right] = \frac{1}{n^2}V\left[\sum_{i=1}^{n}X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}V[X_i] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$

$V[\bar{X}] = \frac{\sigma^2}{n} = E[\bar{X}^2]-E[\bar{X}]^2 \Leftrightarrow \frac{\sigma^2}{n} = E[\bar{X}^2]-\mu^2 \Leftrightarrow E[\bar{X}^2] = \frac{\sigma^2}{n}+\mu^2$

então,

$\frac{n}{n-1}\left(\frac{n}{n}E\left[X^2\right]-E\left[\bar{X}^2\right]\right) = \frac{n}{n-1}\left(\sigma^2+\mu^2-\frac{\sigma^2}{n}-\mu^2\right) = \frac{n}{n-1}\left(\sigma^2-\frac{\sigma^2}{n}\right) =$

$= \frac{n}{n-1}\left(\frac{n\sigma^2-\sigma^2}{n}\right) = \frac{n}{n-1}\left(\frac{(n-1)\sigma^2}{n}\right) = \sigma^2$

Como $E[S^2] = \sigma^2$ conclui-se que S^2 é um estimador centrado ou não enviesado.

Convergência: $\lim_{n\to\infty} E\left[S^2\right] = \lim_{n\to\infty}\sigma^2 = \sigma^2$. $\lim_{n\to\infty} V\left[S^2\right] = ?$

Como mais adiante veremos, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)} \Rightarrow V\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1)$. Então

$V\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1) \Leftrightarrow \frac{(n-1)}{\sigma^4}V\left[S^2\right] = 2(n-1) \Leftrightarrow V\left[S^2\right] = \frac{2\sigma^4(n-1)}{(n-1)^2} = \frac{2\sigma^4}{n-1}$.

Logo, $\lim_{n\to\infty} V\left[S^2\right] = \lim_{n\to\infty}\frac{2\sigma^4}{n-1} = 0$. Como $\lim_{n\to\infty} E\left[S^2\right] = \sigma^2$, e $\lim_{n\to\infty} V\left[S^2\right] = 0$, conclui-se que S^2 é um estimador convergente.

Analisando agora S'^2 :

Enviesamento: $E[S'^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i)^2-\bar{X}^2\right] =$

$= \frac{1}{n}\sum_{i=1}^{n} E\left[(X_i)^2\right] - E\left[\bar{X}^2\right] = \frac{n}{n}E\left[X^2\right] - E\left[\bar{X}^2\right] = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \sigma^2 - \frac{\sigma^2}{n}$ Como $E[S'^2] \neq \sigma^2$ conclui-se que S'^2 é um estimador não centrado ou enviesado.

Convergência: $\lim_{n\to\infty} E\left[S'^2\right] = \lim_{n\to\infty} \sigma^2 - \frac{\sigma^2}{n} = \sigma^2$. $\lim_{n\to\infty} V\left[S'^2\right] = ?$ Como,

$V\left[S'^2\right] = V\left[\frac{(n-1)S^2}{n}\right] = \frac{(n-1)^2}{n^2}V\left[S^2\right] = \frac{(n-1)^2}{n^2} \times \frac{2\sigma^4}{(n-1)} = \frac{(n-1)2\sigma^4}{n^2}$.

Logo, $\lim_{n\to\infty} V\left[S'^2\right] = \lim_{n\to\infty} \frac{(n-1)2\sigma^4}{n^2} = 0$. Como $\lim_{n\to\infty} E\left[S'^2\right] = \sigma^2$ e $\lim_{n\to\infty} V\left[S'^2\right] = 0$, conclui-se que S'^2 é um estimador convergente. ■

4.5 Distribuições Amostrais

Para uma variável aleatória definida sobre uma população os parâmetros da distribuição correspondente são fixos (média, variância, etc). No entanto, nas estatísticas correspondentes (média amostral, variância amostral, etc) as respectivas estimativas variam de amostra para amostra. Devido a esta variabilidade é necessário definir o seu comportamento, a que damos o nome de Distribuição Amostral (ou Distribuição de Amostragem):

Definição 29 *A distribuição de probabilidade de uma Estatística diz-se uma* ***Distribuição Amostral****.*

Exemplo 25 Considere-se uma população com 4 elementos, aos quais se associam os seguintes valores da variável aleatória discreta X:

$$2,\ 4,\ 6,\ 6.$$

A partir destes dados calculamos a correspondente função de probabilidade de X:

$$f(x) = \begin{cases} 1/4 & , x = 2 \\ 1/4 & , x = 4 \\ 2/4 & , x = 6 \\ 0 & , \text{caso contrário} \end{cases} .$$

Concluindo-se que:

$$\mu_X = E[X] = \sum_{i=1}^{3} x_i f(x_i) = 2 \times \frac{1}{4} + 4 \times \frac{1}{4} + 6 \times \frac{2}{4} = 4.5$$

$$\sigma_X^2 = V[X] = \sum_{i=1}^{3} (x_i - \bar{x})^2 f(x_i) =$$

$$= (2 - 4.5)^2 \times \frac{1}{4} + (4 - 4.5)^2 \times \frac{1}{4} + (6 - 4.5)^2 \times \frac{2}{4} = 2.75.$$

Vamos agora definir a distribuição amostral de X, com base em amostras de dimensão 2 obtidas aleatoriamente e com reposição. Considerando todas as amostras de dimensão 2, construímos a seguinte tabela:

Amostras	$\bar{X}$	Prob. de ocorrência
$(2,2)$	2	$1/4 \times 1/4 = 1/16$
$(2,4)$	3	$1/4 \times 1/4 = 1/16$
$(2,6)$	4	$1/4 \times 2/4 = 2/16$
$(4,2)$	3	$1/4 \times 1/4 = 1/16$
$(4,4)$	4	$1/4 \times 1/4 = 1/16$
$(4,6)$	5	$1/4 \times 2/4 = 2/16$
$(6,2)$	4	$2/4 \times 1/4 = 2/16$
$(6,4)$	5	$2/4 \times 1/4 = 2/16$
$(6,6)$	6	$2/4 \times 2/4 = 4/16$

Partindo desta informação definimos a função de probabilidade da média amostral ($\bar{X}$):

$$f(\bar{x}) = \begin{cases} 1/16 & , \bar{x} = 2 \\ 2/16 & , \bar{x} = 3 \\ 5/16 & , \bar{x} = 4 \\ 4/16 & , \bar{x} = 5 \\ 4/16 & , \bar{x} = 6 \\ 0 & , \text{caso contrário} \end{cases} .$$

e obtemos,

$$E\left[\bar{X}\right] = 2 \times \frac{1}{16} + 3 \times \frac{2}{16} + \ldots + 6 \times \frac{4}{16} = 4.5$$
$$V\left[\bar{X}\right] = (2-4.5)^2 \times \frac{1}{16} + \ldots + (6-4.5)^2 \times \frac{4}{16} = 1.375.$$

Concluindo-se que nesta distribuição amostral

$$E\left[\bar{X}\right] = E\left[X\right] = \mu = 4.5$$

e

$$V\left[\bar{X}\right] = \frac{V\left[X\right]}{n} = \frac{\sigma_X^2}{n} = \frac{2.75}{2} = 1.375.$$ ■

A dedução da distribuição amostral do exemplo anterior só foi possível dado o diminuto número de elementos da população, o que é pouco realista; deve, como tal, ser encarada como um mero exemplo académico.

Vamos então apresentar algumas distribuições amostrais, considerando que as populações em estudo são normais (ou assintoticamente normais, pelo Teorema do Limite Central ??referência cruzada).

4.5.1 Distribuição da Média Amostral

Com σ conhecido Com base no Teorema 26, sabendo que $X \sim \mathcal{N}(\mu, \sigma)$ (com μ desconhecido e σ conhecido) e tendo uma amostra independente $X_1, X_2 \ldots, X_n$, considera-se a estatística $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ um "bom" estimador de μ. Como já anteriormente verificámos,

$$E\left[\bar{X}\right] = \mu \text{ e } V\left[\bar{X}\right] = \frac{V\left[X\right]}{n} = \frac{\sigma^2}{n},$$

logo, concluímos que

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Podemos assim utilizar como estatística e correspondente distribuição amostral a normal reduzida

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1).$$

Exemplo 26 Numa praça de Lisboa estão habitualmente estacionados automóveis em transgressão. Todos os dias a polícia autua alguns deles. A receita dessas multas é uma variável aleatória de média igual a 4000 euros e variância igual a 1600000. Qual a probabilidade da receita diária média

durante um ano (225 dias), com multas deste tipo, ultrapassar os 4222 euros?
Embora não se conheça a distribuição da v.a. X, dispõe-se da seguinte informação,

$$\mu_X = 4000,\ n = 225 \text{ dias e}$$
$$\sigma_X^2 = 1600000 \Rightarrow \sigma_X = 1264.91$$

e pretende-se o cálculo da probabilidade $P(\bar{X} > 4222)$.
Podemos então aplicar a este exemplo a anterior distribuição amostral fazendo

$$Z = \frac{\bar{X} - 4000}{\frac{1264.91}{\sqrt{225}}} \sim \mathcal{N}(0,1).$$

A partir daqui o cálculo da probabilidade pretendida é simples,

$$P\left(\bar{X} > 4222\right) = P\left(\frac{\bar{X} - 4000}{\frac{1264.91}{\sqrt{225}}} > \frac{4222 - 4000}{\frac{1264.91}{\sqrt{225}}}\right) = P(Z > 2.63) =$$
$$= 1 - P(Z \leq 2.63) = 1 - 0.9957 = 0.0043. \qquad \blacksquare$$

Com σ desconhecido Neste caso colocam-se duas situações distintas:

1. se a dimensão da amostra é grande (na prática, $n \geq 30$), podemos substituir na estatística do caso anterior σ por S (calculado a partir da amostra), sem que o erro cometido com esta substituição seja grande. Como tal, a estatística e a correspondente distribuição amostral a utilizar vai ser:
$$Z = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim \mathcal{N}(0,1); \tag{6}$$

2. se a dimensão da amostra é pequena (na prática, $n < 30$) utiliza-se a estatística da variância
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
e aplica-se o teorema:

Teorema 40 *Se $X \sim \mathcal{N}(\mu, \sigma)$, a média $\bar{X}$ e a variância empírica S^2 são independentes, então $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ e a variável aleatória $Y = \frac{(n-1)S^2}{\sigma^2}$ tem uma distribuição χ^2 com $(n-1)$ graus de liberdade.*

Considerando então $Y = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{(n-1)}$ e a distribuição amostral $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$, como Y e Z são independentes temos, através do Teorema 34:

$$T = \frac{Z}{\sqrt{\frac{Y}{n-1}}} \sim t_{(n-1)}.$$

Substituíndo Z pela expressão (6):

$$T = \frac{Z}{\sqrt{\frac{Y}{n-1}}} = \frac{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1)\frac{S^2}{\sigma^2}}{n-1}}} = \frac{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \times \frac{\sigma}{S} = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}.$$

Conclui-se que, nestes casos, a estatística e correspondente distribuição amostral a utilizar é:

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}.$$

4.5.2 Distribuição para a Diferença de Duas Médias Amostrais

Com σ_1 e σ_2 conhecidos

Teorema 41 *Se duas amostras aleatórias independentes de dimensões n_1 e n_2, provenientes de duas populações (discretas ou contínuas) de médias μ_1 e μ_2 e variâncias σ_1^2 e σ_2^2 respectivamente, então a distribuição amostral da diferença de médias $\bar{X}_1 - \bar{X}_2$ é assintóticamente normal com média*

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

e variância

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Logo, a estatística $\left(\bar{X}_1 - \bar{X}_2\right)$ é uma normal com parâmetros,

$$\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

Centrando e reduzindo esta v.a. obtemos a estatística e a distribuição amostral:

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1). \tag{7}$$

Exemplo 27 Um fabricante A de máquinas de lavar roupa afirma que os seus termostatos têm uma durabilidade de 6.5 anos e uma variância de 0.81. Um outro fabricante B afirma que os seus termostatos têm uma durabilidade média de 6 anos e um desvio padrão de 0.8 anos. Qual a probabilidade de que numa amostra aleatória de 36 termostatos o fabricante A tenha uma durabilidade média de pelo menos mais um ano que a durabilidade média de uma amostra de 49 termostatos do fabricante B?

Definindo as variáveis aleatórias:
X_1 : Durabilidade (em anos) dos termostatos do fabricante A
X_2 : Durabilidade (em anos) dos termostatos do fabricante B
Neste exemplo pretende-se calcular a probabilidade,

$$P\left(\bar{X}_1 - \bar{X}_2 \geq 1\right).$$

Dispõe-se da seguinte informação,

$$\mu_{X_1} = 6.5; \mu_{X_2} = 6; \sigma^2_{X_1} = 0.81; \sigma^2_{X_2} = 0.64; n_{X_1} = 36; n_{X_2} = 49.$$

Como tal vai utilizar-se a estatística e a distribuição amostral:

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (6.5 - 6)}{\sqrt{\frac{0.81}{36} + \frac{0.64}{49}}} \sim \mathcal{N}(0,1).$$

Logo,

$$P\left(\bar{X}_1 - \bar{X}_2 \geq 1\right) = P\left(\frac{\bar{X}_1 - \bar{X}_2 - (6.5-6)}{\sqrt{\frac{0.81}{36} + \frac{0.64}{49}}} \geq \frac{1 - (6.5-6)}{\sqrt{\frac{0.81}{36} + \frac{0.64}{49}}}\right) =$$

$$= P\left(Z \geq 2.645\right) = 1 - P\left(Z < 2.645\right) = 1 - 0.996 = 0.004 \qquad \blacksquare$$

Com σ_1 e σ_2 desconhecidos À semelhança da 2^a distribuição amostral enunciada, mais uma vez encontramo-nos perante dois casos distintos:

1. se as dimensões das amostras são grandes (na prática, $n_1 \geq 30$ e $n_2 \geq 30$), podemos substituir na estatística do caso anterior σ_1 e σ_2 por S_1 e S_2 (calculados a partir das amostras correspondentes), sem que o erro cometido com esta substituição seja grande. Como tal, a estatística e a distribuição amostral a utilizar nestes casos vai ser:

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \mathcal{N}(0,1)\,;$$

2. se as dimensões das amostras são pequenas (na prática, $n_1 < 30$ ou $n_2 < 30$ e considerando ainda que $\sigma_1^2 = \sigma_2^2$), utilizamos as variáveis aleatórias (6.2) e

$$Y = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} + \frac{(n_2 - 1)S_2^2}{\sigma_2^2}, \tag{8}$$

que tem distribuição χ^2 com $(n_1 - 1 + n_2 - 1) = (n_1 + n_2 - 2)$ graus de liberdade (através da aplicação do Proposição 32).

Então, aplicando o Teorema 34 obtemos

$$T = \frac{Z}{\sqrt{\frac{Y}{n_1+n_2-2}}} \sim t_{(n_1+n_2-2)}. \tag{9}$$

Substituindo em (9) Z e Y pelas expressões definidas em (7) e (8) respectivamente e considerando que $\sigma_1^2 = \sigma_2^2 = \sigma$:

$$T = \frac{Z}{\sqrt{\frac{Y}{n_1+n_2-2}}} = \frac{\frac{\left(\bar{X}_1-\bar{X}_2\right)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}}{\sqrt{\frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}+\frac{(n_2-1)S_2^2}{\sigma_2^2}}{n_1+n_2-2}}} = \frac{\frac{\left(\bar{X}_1-\bar{X}_2\right)-(\mu_1-\mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2(n_1+n_2-2)}}} =$$

$$= \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}}.$$

Conclui-se que, nestes casos, a estatística e correspondente distribuição amostral a utilizar é[3]:

$$T = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}} \sim t_{(n_1+n_2-2)}.$$

[3]Caso as variâncias não sejam iguais as inferências sobre $\mu_1 - \mu_2$ tornam-se bastante mais complexas; esta questão é conhecida como o problema de Behrens-Fisher e tem referências, poe exemplo, nas obras de Kendall e Stuart (1967) e Cox e Hinkley (1974).

4.5.3 Distribuição para a Proporção Amostral

Vamos agora estudar a distribuição que indica a proporção de sucessos (elementos com uma característica pretendida) de X_1, X_2,..., X_n, amostra aleatória de n variáveis aleatórias de Bernoulli independentes e em que $X_i = 1$ representa sucesso e $X_i = 0$ representa insucesso (consoante o elemento observado tenha ou não a característica pretendida). Definindo

$$X = \sum_{i=1}^{n} X_i$$

então a variável aleatória X representa o n^o de sucessos e tem naturalmente uma distribuição binomial, isto é,

$$X \sim B(n, p).$$

Sendo p a proporção desconhecida de uma população e p^* a proporção calculada com base numa amostra independente de dimensão n, tem-se que,

$$p^* = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{m}{n}, \text{ com } m \leq n,$$

sendo m a totalidade dos elementos da amostra que gozam de certa caraterística. A proporção, p, pode então entender-se como um caso particular de um valor médio. Uma proporção empírica (calculada sobre uma amostra) é então o caso particular da média amostral.

A média e a variância de p^* são dados respectivamente por

$$E[p^*] = p \text{ e } V[p^*] = \frac{pq}{n}.$$

No entanto, como desconhecemos os valores de p e q, em muitos casos resolve-se o problema substituindo-os pela sua estimativa p^* e q^*.

Como na distribuição amostral da média (com σ desconhecido e $n \geq 30$) utilizamos a variável aleatória

$$Z = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim \mathcal{N}(0, 1),$$

então, para a proporção p, vamos utilizar a estatística e a distribuição amostral idêntica à anterior (com $n \geq 30$), mas ajustada ao caso em questão (de acordo o Teorema 27):

$$Z = \frac{p^* - p}{\sqrt{\frac{p^* q^*}{n}}} \dot{\sim} \mathcal{N}(0, 1)$$

4.5.4 Distribuição para a Diferença de Duas Proporções Amostrais

Dadas duas populações X_1 e X_2, considere-se a diferença de proporções $(p_1 - p_2)$ entre os seus elementos que gozam de determinada característica. Considerando duas amostras independentes de X_1 e X_2 de dimensões n_1 e n_2, cujas proporções empíricas são p_1^* e p_2^* respectivamente, então, demonstra-se que $(p_1^* - p_2^*)$ tem valor esperado

$$E[p_1^* - p_2^*] = p_1 - p_2$$

e variância

$$V[p_1^* - p_2^*] = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \text{ que se estima através de } \frac{p_1^* q_1^*}{n_1} + \frac{p_2^* q_2^*}{n_2}.$$

Desta forma, e à semelhança do que se fez para a diferença de duas médias, para as diferenças de proporções $(p_1 - p_2)$ utiliza-se a estatística e a distribuição amostral (para $n_1 \geq 30$ e $n_2 \geq 30$):

$$Z = \frac{(p_1^* - p_2^*) - (p_1 - p_2)}{\sqrt{\frac{p_1^* q_1^*}{n_1} + \frac{p_2^* q_2^*}{n_2}}} \dot{\sim} \mathcal{N}(0,1).$$

4.5.5 Distribuição para a Variância Amostral

Com μ conhecido Seja $X \sim \mathcal{N}(\mu, \sigma)$ em que μ é conhecido, e considerando uma amostra independente $X_1, ..., X_n$, então a estatística e a distribuição amostral a utilizar neste caso é dada por:

$$X^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2_{(n)}.$$

Com μ desconhecido No caso de pretendermos estudar a variância de populações normais, em que μ é desconhecido usa-se a seguinte estatística e a distribuição amostral:

$$X^2 = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{(n-1)}.$$

Exemplo 28 Recolheu-se uma amostra aleatória de dimensão 5 de uma população normal. Determine a probabilidade do desvio padrão da amostra ser inferior ao desvio padrão da população. Sabe-se que $n = 5$, e pretende-se calcular a probabilidade do desvio padrão amostral ser inferior ao da população, $P(S < \sigma)$. Vai então utilizar-se a distribuição amostral

$$X^2 = 4\frac{S^2}{\sigma^2} \sim \chi^2_{(4)}.$$

Logo,

$$P(S < \sigma) = P\left(\frac{S^2}{\sigma^2} < 1\right) = P\left(4\frac{S^2}{\sigma^2} < 4\right) = P\left(X^2 < 4\right) = 0.584. \qquad \blacksquare$$

4.5.6 Distribuição para a Razão de Duas Variâncias Amostrais

Em alguns problemas há todo o interesse em verificar se duas amostras de tamanhos n_1 e n_2, cujas variâncias são S_1^2 e S_2^2, provêm ou não da mesma população normal ou de duas populações normais com variâncias iguais. Em tais casos, e à semelhança do que já foi feito anteriormente (em que indicamos como podemos obter distribuições amostrais de diferenças, especificamente entre médias e proporções) poderiamos chegar à distribuição amostral da diferença de variâncias $\left(S_1^2 - S_2^2\right)$, no entanto, tal distribuição é bastante complicada de deduzir. No seu lugar podemos então considerar a estatística $\frac{S_1^2}{S_2^2}$, pois uma razão muito grande ou muito pequena indica uma grande diferença entre variâncias, assim como uma razão muito próxima de um indica uma pequena diferença entre variâncias.

Então, para duas amostras aleatórias independentes de dimensões n_1 e n_2 (extraídas de duas populações normais com variâncias desconhecidas σ_1^2 e σ_2^2 respectivamente), as correspondentes variâncias amostrais são dadas por S_1^2 e S_2^2, sendo,

$$Y_1 = (n_1 - 1)\frac{S_1^2}{\sigma_1^2} \sim \chi^2_{(n_1-1)} \text{ e } Y_2 = (n_2 - 1)\frac{S_2^2}{\sigma_2^2} \sim \chi^2_{(n_2-1)}.$$

Fazendo

$$F = \frac{\frac{Y_1}{(n_1-1)}}{\frac{Y_2}{(n_2-1)}} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \tag{10}$$

e aplicando o Teorema 37 obtemos

$$F = \frac{\frac{Y_1}{(n_1-1)}}{\frac{Y_2}{(n_2-1)}} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} \sim F_{(n_1-1,n_2-1)},$$

pois a variável (10) é quociente de Qui-Quadrados divididos pelos respectivos graus de liberdade. Então, neste caso, vamos utilizar a estatística e a distribuição amostral:

$$F = \frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} \sim F_{(n_1-1,n_2-1)}.$$

4.6 Intervalos de Confiança

Nos intervalos de confiança que mais adiante apresentaremos parte-se do princípio de que todas as populações em análise têm comportamento normal (ou aproximadamente normal), sendo as estatísticas e respectivas distribuições amostrais a utilizar as apresentadas no subcapítulo anterior.

Na teoria da estimação pontual temos uma avaliação (θ^*) do verdadeiro valor de um parâmetro (θ), no entanto não dispomos de informação acerca da confiança a atribuir a essa estimativa. Na estimação por intervalos o estimador θ^* de um parâmetro θ é apresentado sob a forma de um intervalo genérico $]\theta^* - d, \theta^* + d[$ (em que d representa o erro associado ao intervalo), existindo uma probabilidade conhecida desse intervalo conter o parâmetro θ.

Em resumo, se considerarmos X uma variável aleatória com função densidade de probabilidade $f(X_1, X_2, ..., X_n/\theta)$ em que θ é o parâmetro desconhecido a estimar, $X_1, X_2, ..., X_n$ uma amostra aleatória e

$$L_1(X_1, X_2, ..., X_n) \quad \text{e} \quad L_2(X_1, X_2, ..., X_n)$$

duas estatísticas tais que

$$L_1 < L_2 \ \wedge \ P(L_1 < \theta < L_2) = 1 - \alpha.$$

Nestas condições, para uma realização da amostra $x_1, x_2, ..., x_n$, calculamos l_1 e l_2 e:

- ao intervalo $]l_1, l_2[$ denominamos **intervalo de confiança** a $(1-\alpha)100\%$ para o parâmetro θ;
- à probabilidade $(1-\alpha)$ dá-se o nome de **grau de confiança** do intervalo;
- à probabilidade complementar, α, dá-se o nome de **nível de significância**;
- aos extremos do intervalo, l_1 e l_2, chamamos **limites de confiança inferior e superior**, respectivamente.

Como é óbvio pretende-se que uma estimativa possua o máximo de confiança possível, no entanto, se uma maior confiança é pretendida na estimação, esta conduz a possibilidades de erros menores (dado que um baixo nível de significância produz um intervalo de estimação maior) e, como tal, a precisão da estimação diminui.

Exemplo 29 Consideremos as seguintes afirmações proferidas por três alunos de uma escola que esperam ansiosamente a saída de uma pauta de exame de Estatística onde constam as respectivas notas:
1^o Estudante: "Tenho a sensação de que o professor de Estatística afixa a pauta na parte da manhã, como usualmente faz."
2^o Estudante: "Tenho quase a certeza de que o professor de Estatística afixa a pauta entre as 10 e as 11 horas."
3^o Estudante: "Tenho a certeza absoluta de que o professor de Estatística ou afixa a pauta às 10.30 ou já não a afixa hoje."

Estas três afirmações permitem constatar facilmente que se se pretende uma maior confiança na estimativa, se tem que permitir que a possibilidade de erro aumente. Por outro lado, se se permitir que o erro diminua, a amplitude do intervalo aumenta, perdendo a estimativa alguma precisão. No entanto há que ter em atenção que, se um intervalo de confiança tem uma amplitude demasiado grande, a estimativa não tem utilidade. ■

Resumindo, um intervalo de confiança tem uma amplitude inversamente proporcional à dimensão da amostra pois, no limite, para n a tender para a dimensão da população, o intervalo reduz-se a um único ponto, isto é, o valor do parâmetro é conhecido com exactidão. Da mesma forma, se considerarmos n fixo, a amplitude do intervalo também é inversamente proporcional ao risco ou erro a ele associado, isto é, à probabilidade do intervalo não conter o verdadeiro valor do parâmetro.

A interpretação de um intervalo de confiança é geralmente realizada de uma forma relativamente banal, mas incorrecta do ponto de vista teórico. Se for recolhido um grande número de amostras de n observações independentes da variável aleatória X, a proporção de amostras às quais correspondem particulares intervalos $]l_1, l_2[$, compreendendo o verdadeiro valor do parâmetro θ, tende a aproximar-se de $(1-\alpha)$. Assim, $(1-\alpha)$, traduz o grau de de confiança que se tem em que uma particular amostra de dimensão n de X dê origem a um intervalo que compreenda o verdadeiro valor do parâmetro θ. Isto é, a partir da igualdade

$$P(L_1 < \theta < L_2) = 1 - \alpha$$

conclui-se que a probabilidade do intervalo aleatório genérico

$$]L_1, L_2[$$

conter o verdadeiro valor do parâmetro θ é $(1-\alpha)$. Tem-se pois considerável confiança que, para uma amostra concreta de dimensão n, o particular intervalo correspondente $]l_1, l_2[$, contenha o valor de θ. Repare-se que cada intervalo particular $]l_1, l_2[$ ou contem ou não contem θ, e $(1-\alpha)$ não traduz a primeira dessa alternativas; com efeito, como l_1 e l_2 são números, a dupla desigualdade $l_1 < \theta < l_2$ ou é válida ou não é e portanto

$$P(l_1 < \theta < l_2) = 1 \quad \text{ou} \quad P(l_1 < \theta < l_2) = 0,$$

embora por desconhecimento de θ, não se saiba o que se passa. A cada particularização do intervalo $]L_1, L_2[$, associa-se pois, como grau de confiança quanto a conter θ, o número $(1-\alpha)$; de um modo sintético, qualquer particularização do referido intervalo aleatório diz-se que constitui um intervalo de confiança a $(1-\alpha)$ para θ.

4.6.1 Intervalos de Confiança para a Média

Com σ conhecido Estatística e distribuição amostral a utilizar:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1).$$

Para determinar um intervalo de confiança para μ, vamos utilizar a estatística Z. Fixando o valor α começamos por calcular um intervalo

$$\left]-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}\right[$$

onde Z se situa. Para o cálculo dos extremos deste intervalo consulta-se o valor de $z_{1-\frac{\alpha}{2}}$ na tabela da Normal, correspondente à probabilidade

$$P\left(Z < z_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$$

(note-se que $-z_{1-\frac{\alpha}{2}}$ e $z_{1-\frac{\alpha}{2}}$ são simétricos dado que as respectivas probabilidades são complementares tal com se visualiza na Figura 21). Então, o Intervalo de Confiança para a média a

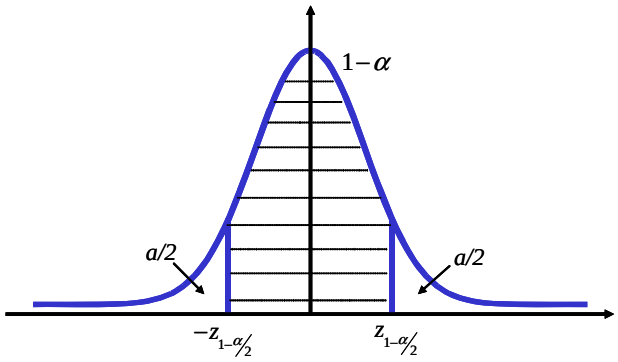


Figura 21: Intervalo de Confiança para a v.a. Z.

$(1-\alpha)100\%$ deduz-se do seguinte modo,

$$\begin{aligned}
&P\left(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow \\
\Leftrightarrow\ &P\left(-z_{1-\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow \\
\Leftrightarrow\ &P\left(-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \Leftrightarrow \\
\Leftrightarrow\ &P\left(z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} > \mu-\bar{X} > -z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \Leftrightarrow \\
\Leftrightarrow\ &P\left(z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} + \bar{X} > \mu > -z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} + \bar{X}\right) = 1-\alpha \Leftrightarrow \\
\Leftrightarrow\ &P\left(\bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.
\end{aligned}$$

Sendo $\bar{x}$ calculado a partir dos valores da amostra, da anterior igualdade resulta o intervalo de confiança para μ a $(1-\alpha)100\%$:

$$\left]\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}; \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right[.$$

Exemplo 30 A característica X em certo artigo produzido em série segue uma distribuição com variância igual a 9. Com base numa amostra de dimensão 100, que forneceu um valor médio igual a 5, determine um intervalo de confiança a 95% para o valor médio da distribuição.

Como não conhecemos a distribuição da população em causa, através do Teorema 27 vamos obter

$$Z = \frac{\bar{X} - \mu}{\frac{3}{\sqrt{100}}} \dot{\sim} \mathcal{N}(0,1).$$

Deduzindo o IC:

$$P\left(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha \Leftrightarrow$$

$$\vdots$$

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

Como

$$(1-\alpha) = 0.95 \Leftrightarrow \alpha = 0.05 \Leftrightarrow 1 - \frac{\alpha}{2} = 0.975,$$

retiramos da tabela da normal,

$$P(Z < z_{0.975}) = 0.975 \Leftrightarrow z_{0.975} = 1.96.$$

Como $\bar{x} = 5$, obtemos o intervalo para μ a 95% de confiança (ou com 5% de risco de erro):

$$\left]5 - 1.96 \times \frac{3}{10}, 5 + 1.96 \times \frac{3}{10}\right[=]4.412, 5.588[\,.$$

Em termos de interpretação do intervalo de confiança anterior, e se quisermos ser precisos, concluímos que, se observarmos um grande n^o de amostras de dimensão 100, a proporção das amostras onde podemos encontrar a média da v.a. X situada no intervalo de confiança acima definido é igual a 0.95; de uma forma mais sintética podemos afirmar que, o anterior intervalo aleatório $]4.412, 5.588[$, é um intervalo de confiança a 95% para a média de X; por último, de uma forma mais corrente, embora menos correcta em termos teóricos, é usual afirmar que, com 95% de confiança a média de X se situa entre os valores 4.412 e 5.588 . ■

Exemplo 31 Se para o exemplo 30 pretendessemos saber a dimensão da amostra para obtermos um intervalo de confiança para μ, nas condições anteriormente apresentadas, mas cuja amplitude (A) não fosse superior a 0.5, o procedimento a seguir seria:

$$A = \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} - \left(\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 2 \times z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}.$$

Logo,

$$A = 2 \times 1.96 \times \frac{1}{\sqrt{n}} \leq 0.5 \Leftrightarrow \sqrt{n} \geq 23.52 \Rightarrow n \geq 553.1904,$$

isto é, a dimensão da amostra deveria ser igual ou superior a 554. ■

Com σ desconhecido Neste tipo de intervalos de confiança, em que ambos os parâmetros são desconhecidos, podemos encontrar-nos perante duas situações distintas:

1. se a dimensão da amostra é grande (na prática, $n \geq 30$), utiliza-se a estatística e a distribuição amostral:
$$Z = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim \mathcal{N}(0,1).$$
Obtendo-se, de forma análoga ao caso anterior, o intervalos de confiança para μ a $(1-\alpha)100\%$:
$$\left]\bar{x} - z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}; \bar{x} + z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right[;$$

2. se a dimensão da amostra é pequena (na prática, $n < 30$) utiliza-se a estatística e a distribuição amostral:
$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}.$$

Fixando o valor de α começamos por calcular um intervalo

$$\left]-t_{(n-1);1-\frac{\alpha}{2}}, t_{(n-1);1-\frac{\alpha}{2}}\right[$$

onde T se situa como observamos na Figura 22. Para o cálculo dos extremos deste intervalo consulta-se o valor de $t_{(n-1);1-\frac{\alpha}{2}}$ na tabela da t-Student (note-se que, à semelhança do que acontecia na distribuição Normal, $-t_{(n-1);1-\frac{\alpha}{2}}$ e $t_{(n-1);1-\frac{\alpha}{2}}$ são simétricos), correspondente à probabilidade

$$P\left(T < t_{(n-1);1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}.$$

Então o Intervalo de Confiança para a média a $(1-\alpha)100\%$ deduz-se do seguinte modo,

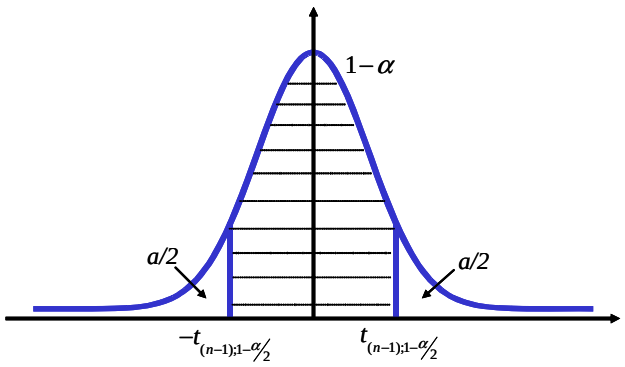


Figura 22: Intervalo de Confiança para a v.a. T.

$$P\left(-t_{(n-1);1-\frac{\alpha}{2}} < T < t_{(n-1);1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$
$$\Leftrightarrow P\left(-t_{(n-1);1-\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < t_{(n-1);1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$
$$\Leftrightarrow P\left(-t_{(n-1);1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < \bar{X}-\mu < t_{(n-1);1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1-\alpha \Leftrightarrow$$
$$\Leftrightarrow P\left(t_{(n-1);1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} > \mu-\bar{X} > -t_{(n-1);1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1-\alpha \Leftrightarrow$$
$$\Leftrightarrow P\left(\bar{X}+t_{(n-1);1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} > \mu > \bar{X}-t_{(n-1);1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1-\alpha \Leftrightarrow$$
$$\Leftrightarrow P\left(\bar{X}-t_{(n-1);1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < \mu < \bar{X}+t_{(n-1);1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1-\alpha.$$

Sendo $\bar{x}$ e s calculados a partir dos valores da amostra, resulta da anterior igualdade o intervalo de confiança para μ a $(1-\alpha)100\%$:

$$\left]\bar{x}-t_{(n-1);1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}};\bar{x}+t_{(n-1);1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right[.$$

Exemplo 32 O tempo que uma máquina leva a executar determinada operação numa peça está sujeito a variações, tendo no entanto um comportamento normal. Para verificar se as condições de funcionamento da máquina estão dentro das normas, registou-se 12 vezes o referido tempo. Os resultados (em segundos) foram os seguintes:

$$29, \quad 33, \quad 36, \quad 35, \quad 36, \quad 40, \quad 32, \quad 37, \quad 31, \quad 35, \quad 30, \quad 36.$$

Construa um intervalo de confiança a 95% para o tempo médio de execução da tarefa pela máquina em análise, sabendo que esta segue uma distribuição normal.

Para este exemplo podemos definir a nossa variável X como o "tempo, em segundos, que uma máquina leva a executar uma tarefa". Sabemos que

$$X \sim \mathcal{N}(\mu,\sigma), n = 12 \text{ e } (1-\alpha) = 0.95.$$

Como desconhecemos os parâmetros da distribuição e $n < 30$, vamos utilizar:

$$T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}.$$

Deduzindo o Intervalo de Confiança:

$$P\left(-t_{(n-1);1-\frac{\alpha}{2}} < T < t_{(n-1);1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$
$$\vdots$$
$$\Leftrightarrow P\left(\bar{X}-t_{(n-1);1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} < \mu < \bar{X}+t_{(n-1);1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1-\alpha.$$

Impõe-se então calcular $\bar{x}$ e s :

$$\bar{x} = \frac{1}{12}\sum_{i=1}^{12} x_i = 34.17$$
$$s^2 = \frac{1}{11}\sum_{i=1}^{12}(x_i-34.17)^2 = 10.08 \Rightarrow s = 3.18$$

Como

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{12}}} \sim t_{(11)} \quad \text{e} \quad 1 - \frac{\alpha}{2} = 0.975,$$

tem-se

$$P\left(T < t_{(11);0.975}\right) = 0.975 \Rightarrow t_{(11);0.975} = 2.201.$$

Para $\bar{x} = 34.17$ e $s = 3.18$, obtem-se o intervalo de confiança para μ a 95% de confiança (ou com 5% de risco de erro):

$$\left]34.17 - 2.201 \times \frac{3.18}{\sqrt{12}}, 34.17 + 2.201 \times \frac{3.18}{\sqrt{12}}\right[=]32.15, 36.19[\,. \qquad \blacksquare$$

4.6.2 Intervalos de Confiança para a Diferença de Duas Médias

Com σ_1 e σ_2 conhecidos Considerem-se duas variáveis aleatórias independentes X_1 e X_2 normais com médias μ_1 e μ_2 e desvios padrões σ_1 e σ_2 (conhecidos) respectivamente. Seleccionando duas amostras aleatórias independentes de dimensões n_1 e n_2, para determinar um intervalo de confiança para $(\mu_1 - \mu_2)$, vamos utilizar a estatística e a distribuição amostral:

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1)\,.$$

Se as populações não forem normais ou não se conhecer a sua distribuição, mas n_1 e n_2 forem grandes, Z é assintóticamente uma $\mathcal{N}(0,1)$, pelo Teorema 27.

Então o Intervalo de Confiança para a diferença de duas médias a $(1-\alpha)100\%$ deduz-se do seguinte modo,

$$P\left(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$\Updownarrow$$

$$P\left(-z_{1-\frac{\alpha}{2}} < \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha \Leftrightarrow$$

$$\Updownarrow$$

$$P\left(-z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2) < z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

$$\Updownarrow$$

$$P\left(\bar{X}_1 - \bar{X}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \bar{X}_1 - \bar{X}_2 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

Logo, o intervalo de confiança para $(\mu_1 - \mu_2)$ a $(1-\alpha)100\%$ vai ser:

$$\left]\bar{x}_1 - \bar{x}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right[.$$

Exemplo 33 Duas variáveis aleatórias X_1 e X_2 seguem distribuições normais com variâncias $\sigma_1^2 = 3.64$ e $\sigma_2^2 = 4.03$ respectivamente. Construa um intervalo de confiança a 95% para a diferença entre as suas médias, sabendo que em duas amostras recolhidas se obtiveram os seguintes resultados:

Amostra 1:	$n_1 = 32$	$\bar{x}_1 = 16.20$
Amostra 2:	$n_2 = 40$	$\bar{x}_2 = 14.85$

Vamos utilizar

$$(1-\alpha) = 0.95 \Leftrightarrow 1 - \frac{\alpha}{2} = 0.975$$

e

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{3.64}{32} + \frac{4.03}{40}}} \sim \mathcal{N}(0,1).$$

Deduzindo o Intervalo de Confiança:

$$P\left(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\vdots$$

$$P\left(\bar{X}_1 - \bar{X}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \bar{X}_1 - \bar{X}_2 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1-\alpha$$

Como $z_{0.975} = 1.96$, $\bar{x}_1 = 16.20$ e $\bar{x}_2 = 14.85$ o intervalo de confiança para $(\mu_1 - \mu_2)$ a 95% de confiança (ou com 5% de risco de erro) vai ser:

$$\left]\left(16.20 - 14.85\right) - 1.96\sqrt{0.2145}, \left(16.20 - 14.85\right) + 1.96\sqrt{0.2145}\right[=$$
$$=]0.44, 2.26[\,. \qquad \blacksquare$$

Com σ_1 e σ_2 desconhecidos Neste tipo de intervalos de confiança, em que ambos os parâmetros são desconhecidos, podemos, mais uma vez, encontrar-nos perante duas situações distintas:

1. se as dimensões das amostras são grandes (na prática, $n_1 \geq 30$ e $n_2 \geq 30$), a estatística e a correspondente distribuição amostral a utilizar é:

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \mathcal{N}(0,1).$$

   O intervalo de confiança para $(\mu_1 - \mu_2)$ a $(1-\alpha)100\%$, deduzido de forma idêntica ao anterior, é:

$$\left]\bar{x}_1 - \bar{x}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right[;$$

2. se as dimensões das amostras são pequenas (na prática, $n_1 < 30$ ou $n_2 < 30$ e considerando ainda que $\sigma_1^2 = \sigma_2^2$), a estatística e a correspondente distribuição amostral a utilizar é:

$$T = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}} \sim t_{(n_1+n_2-2)}.$$

Então o Intervalo de Confiança para a diferença das duas médias a $(1-\alpha)100\%$ deduz-se do seguinte modo $\left(\text{fazendo } a = \sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\right)$,

$$P\left(-t_{(n_1+n_2-2);1-\frac{\alpha}{2}} < T < t_{(n_1+n_2-2);1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$\Updownarrow$$
$$P\left(-t_{(n_1+n_2-2);1-\frac{\alpha}{2}} < \frac{\left(\bar{X}_1-\bar{X}_2\right)-(\mu_1-\mu_2)}{a} < t_{(n_1+n_2-2);1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$\Updownarrow$$
$$P\left(-t_{(n_1+n_2-2);1-\frac{\alpha}{2}} < \frac{\left(\bar{X}_1-\bar{X}_2\right)-(\mu_1-\mu_2)}{a} < t_{(n_1+n_2-2);1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$\Updownarrow$$
$$P\left(-t_{(n_1+n_2-2);1-\frac{\alpha}{2}}a < \left(\bar{X}_1-\bar{X}_2\right)-(\mu_1-\mu_2) < t_{(n_1+n_2-2);1-\frac{\alpha}{2}}a\right) = 1-\alpha$$
$$\Updownarrow$$
$$P\left(\bar{X}_1-\bar{X}_2-t_{(n_1+n_2-2);1-\frac{\alpha}{2}}a < \mu_1-\mu_2 < \bar{X}_1-\bar{X}_2+t_{(n_1+n_2-2);1-\frac{\alpha}{2}}a\right) = 1-\alpha$$

Logo, o intervalo de confiança para $(\mu_1-\mu_2)$ a $(1-\alpha)100\%$ vai ser:

$$\left]\left(\bar{x}_1-\bar{x}_2\right)-t_{(n_1+n_2-2);1-\frac{\alpha}{2}}a, \left(\bar{x}_1-\bar{x}_2\right)+t_{(n_1+n_2-2);1-\frac{\alpha}{2}}a\right[=$$
$$= \left]\left(\bar{x}_1-\bar{x}_2\right) \mp t_{(n_1+n_2-2);1-\frac{\alpha}{2}}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\right[.$$

Exemplo 34 Foi realizado um estudo para determinar se um certo tratamento tinha efeito corrosivo sobre um metal. Uma amostra de 100 peças foi imersa num banho durante 24 horas com o tratamento, tendo sido removido uma média de 12.2 mm de metal com um desvio padrão de 1.1 mm. Uma segunda amostra de 200 peças foi também imersa durante 24 horas mas sem tratamento, sendo a média de metal removido de 9.1 mm, com um desvio padrão de 0.9 mm. Determine um intervalo de confiança a 98% para a diferença entre as médias das populações, retirando conclusões quanto ao efeito do tratamento.

Como $n_1 \geq 30$, $n_2 \geq 30$, σ_1^2 e σ_2^2 são desconhecidos vamos utilizar:

$$Z = \frac{\left(\bar{X}_1-\bar{X}_2\right)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{100}+\frac{S_2^2}{200}}} \sim \mathcal{N}(0,1).$$

Deduzindo o Intervalo de Confiança:

$$P\left(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$
$$\vdots$$
$$P\left(\bar{X}_1-\bar{X}_2-z_{1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{100}+\frac{S_2^2}{200}} < \mu_1-\mu_2 < \bar{X}_1-\bar{X}_2+z_{1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{100}+\frac{S_2^2}{200}}\right) = 1-\alpha$$

Como $(1-\alpha) = 0.98 \Leftrightarrow 1 - \frac{\alpha}{2} = 0.99$, $z_{0.99} = 2.33$, $\bar{x}_1 = 12.2$, $\bar{x}_2 = 9.1$, $s_1^2 = 1.1^2$ e $s_1^2 = 0.9^2$ o intervalo de confiança para $(\mu_1 - \mu_2)$ a 98% de confiança (ou com 2% de risco de erro) vai ser:

$$\left](12.2-9.1) - 2.33\sqrt{\frac{1.1^2}{100}+\frac{0.9^2}{200}}, (12.2-9.1) + 2.33\sqrt{\frac{1.1^2}{100}+\frac{0.9^2}{200}}\right[=$$

$$=]2.804, 3.396[\,.$$

Como

$$(\mu_1 - \mu_2) > 0 \Leftrightarrow \mu_1 > \mu_2,$$

a média do metal removido com o tratamento é superior à média do metal removido sem este, conclui-se que o tratamento tem efeito corrosivo no metal. ■

Exemplo 35 Duas marcas de comprimidos, um deles contendo ácido acetilsalicílico (a.a.s.), são anunciados como fazendo desaparecer a dor de cabeça em tempo record. Foram feitas experiências com cada um deles, tendo os resultados (em minutos) sido os seguintes:

Comprimido 1:	9.6	9.4	9.3	11.2	11.4	12.1
(com a.a.s.)	10.4	9.6	10.2	8.8	13.0	10.2

Comprimido 2:	10.6	13.2	11.7	9.6	8.5	9.7
(sem a.a.s.)	12.3	12.4	10.8	10.8		

Assume-se por hipótese que os tempos acima referidos seguem distribuições normais (com variâncias iguais). Pretende-se saber se um dos comprimidos pode ser considerado mais eficaz do que o outro através de uma estimativa pontual e de uma estimativa por intervalos (a 95% de confiança).

Primeiro vamos obter a estimativa pontual para a diferença entre os tempos médios que cada comprimido leva a tirar a dor de cabeça. Considerando que,

X_1 representa o tempo em minutos que o comprimido com a.a.s. leva a tirar a dor de cabeça (com $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$)

X_2 representa o tempo em minutos que o comprimido sem a.a.s. leva a tirar a dor de cabeça (com $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$)

então,

$\bar{x}_1 = \frac{1}{12}\sum_{i=1}^{12} x_i = 10.4(3)$ minutos e $\bar{x}_2 = \frac{1}{10}\sum_{i=1}^{10} x_i = 10.96$ minutos.

A estimativa pontual de $(\mu_1 - \mu_2)$ é igual a

$$(\bar{x}_1 - \bar{x}_2) = -0.53 \text{ minutos},$$

concluindo-se que, em média, o comprimido sem a.a.s. leva mais meio minuto que o com a.a.s. para fazer desaparecer a dor de cabeça.

Pretende-se agora obter um intervalo de confiança para a diferença de tempos médios e retirar conclusões para o modelo. Primeiro temos de calcular:

$$s_1^2 = \frac{1}{11}\sum_{i=1}^{12}(x_i - 10.43)^2 = 1.58 \quad \text{e} \quad s_2^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - 10.96)^2 = 2.12.$$

Como $n_1 < 30$, $n_2 < 30$, σ_1 e σ_2 desconhecidos e iguais, vamos utilizar,

$$T = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{12}+\frac{1}{10}\right)\frac{11S_1^2+9S_2^2}{20}}} \sim t_{(20)}.$$

Deduzindo o Intervalo de Confiança:

$$P\left(-t_{(n_1+n_2-2);1-\frac{\alpha}{2}} < T < t_{(n_1+n_2-2);1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\vdots$$

$$P\left(\bar{X}_1 - \bar{X}_2 - t_{(n_1+n_2-2);1-\frac{\alpha}{2}} a < \mu_1 - \mu_2 < \bar{X}_1 - \bar{X}_2 + t_{(n_1+n_2-2);1-\frac{\alpha}{2}} a\right) = 1-\alpha.$$

Considerando

$$(1-\alpha) = 0.95 \Leftrightarrow 1 - \frac{\alpha}{2} = 0.975 \;\text{ e }\; t_{(20);0.975} = 2.086,$$

então o intervalo de confiança para $\mu_1 - \mu_2$ a 95% de confiança é:

$$\left]-0.53 \mp 2.086\sqrt{\frac{11}{60} \times \frac{11\times 1.58 + 9 \times 2.12}{20}}\right[=]-1.74, 0.68[\,.$$

Como o intervalo a 95% de confiança contém o valor zero (isto é, a diferença entre os dois tempos é nula) não é muito seguro afirmar que um dos comprimidos seja superior ao outro, embora exista uma ligeira tendência para o comprimido com a.a.s. ser, em média, mais rápido (pois a parte negativa do intervalo é maior do que a parte positiva). ■

4.6.3 Intervalo de Confiança para uma Proporção

Para o cálculo do intervalo de confiança para uma proporção p, vamos utilizar a estatística e a distribuição amostral:

$$Z = \frac{p^* - p}{\sqrt{\frac{p^*q^*}{n}}} \dot{\sim} \mathcal{N}(0,1)\,.$$

O Intervalo de Confiança correspondente, a $(1-\alpha)100\%$ de confiança, é dado por:

$$\left]p^* - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p^*q^*}{n}}; p^* + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p^*q^*}{n}}\right[.$$

Exemplo 36 Um banco pretende estimar a percentagem de clientes que passam cheques sem cobertura. Numa amostra de 150 clientes conclui-se que 15 deles já tinham passado cheques sem cobertura. Estime, a 95% de confiança, a verdadeira percentagem (ou proporção) de clientes do banco que passam cheques sem cobertura.

Como estamos perante uma proporção (com $n \geq 30$) vamos utilizar

$$Z = \frac{p^* - p}{\sqrt{\frac{p^*q^*}{150}}} \dot{\sim} \mathcal{N}(0,1)\,.$$

Deduzindo o IC,

$$P\left(z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$

$$\vdots$$

$$\Leftrightarrow P\left(p^* - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p^*q^*}{n}} < p < p^* + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p^*q^*}{n}}\right) = 1-\alpha.$$

Como

$$p^* = \frac{15}{150} = 0.1 \Rightarrow q^* = 1 - p^* = 0.9,$$
$$\text{e } (1-\alpha) = 0.95 \Leftrightarrow 1 - \frac{\alpha}{2} = 0.975 \text{ e } z_{0.975} = 1.96, \text{ então,}$$

o intervalo de confiança para p a 95% de confiança é:

$$\left]0.1 - 1.96\sqrt{\frac{0.1 \times 0.9}{150}}, 0.1 + 1.96\sqrt{\frac{0.1 \times 0.9}{150}}\right[=]5.20\%, 14.80\%[\,.$$

Conclui-se que, com 95% de confiança, a percentagem de clientes de um banco que passam cheques sem cobertura situa-se entre 5,2 e 14,8. ■

4.6.4 Intervalo de Confiança para a Diferença de Duas Proporções

Para o cálculo do intervalo de confiança para a diferença de duas proporções $(p_1 - p_2)$ utiliza-se a estatística e a distribuição amostral:

$$Z = \frac{(p_1^* - p_2^*) - (p_1 - p_2)}{\sqrt{\frac{p_1^* q_1^*}{n_1} + \frac{p_2^* q_2^*}{n_2}}} \dot{\sim} \mathcal{N}(0,1)$$

obtendo-se o seguinte intervalo de confiança,

$$\left](p_1^* - p_2^*) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_1^* q_1^*}{n_1} + \frac{p_2^* q_2^*}{n_2}}, (p_1^* - p_2^*) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_1^* q_1^*}{n_1} + \frac{p_2^* q_2^*}{n_2}}\right[.$$

Exemplo 37 Um comerciante de brinquedos verificou que 38 de 100 bonecos fabricados pela empresa A não satisfaziam determinada norma de segurança, enquanto que 52 dos 500 fabricados pela empresa B não obedeciam à mesma norma. Verifique, através de um intervalo de confiança a 95%, se é razoável supor que as percentagens observadas traduzem comportamentos idênticos para os dois fabricantes, no que toca ao não cumprimento da norma de segurança.

Sejam,

p_1 - proporção de bonecos da empresa A que não cumpre a norma de segurança ($p_1^* = 0.38$) e

p_2 - proporção de bonecos da empresa A que não cumpre a norma de segurança ($p_2^* = 0.104$).

As empresas terão comportamentos idênticos se $p_1 = p_2$, isto é, se $p_1 - p_2 = 0$. Como $n_1 \geq 30$ e $n_2 \geq 30$ vamos construir um intervalo de confiança para a diferença de proporções, utilizando

$$Z = \frac{(p_1^* - p_2^*) - (p_1 - p_2)}{\sqrt{\frac{p_1^* q_1^*}{100} + \frac{p_2^* q_2^*}{500}}} \dot{\sim} \mathcal{N}(0,1)\,.$$

Deduzindo o IC,

$$P\left(z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha \Leftrightarrow$$
$$\vdots$$
$$\Leftrightarrow P\left(p_1^* - p_2^* - z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_1^* q_1^*}{n_1} + \frac{p_2^* q_2^*}{n_2}} < p_1 - p_2 < p_1^* - p_2^* + z_{1-\frac{\alpha}{2}}\sqrt{\frac{p_1^* q_1^*}{n_1} + \frac{p_2^* q_2^*}{n_2}}\right) = 1 - \alpha.$$

Como $(1-\alpha) = 0.95 \Leftrightarrow 1-\frac{\alpha}{2} = 0.975$ e $z_{0.975} = 1.96$, então, o intervalo de confiança para $(p_1 - p_2)$ a 95% de confiança é:

$$\left](0.276) - 1.96\sqrt{\frac{0.2356}{100} + \frac{0.093}{500}}, (0.276) + 1.96\sqrt{\frac{0.2356}{100} + \frac{0.093}{500}}\right[=$$
$$=]17.7\%, 37.5\%[$$

Como o intervalo de confiança apenas tem valores positivos conclui-se que a verdadeira proporção (a 95% de confiança) de bonecos que não cumprem a norma de segurança é maior na empresa A do que na empresa B. Como tal, estas duas empresas não têm comportamentos idênticos quanto ao incomprimento da norma de segurança. ■

4.6.5 Intervalos de Confiança para a Variância

Com μ conhecido Para o cálculo do intervalo de confiança para σ^2 utiliza-se a estatística e a distribuição amostral:

$$X^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2_{(n)}.$$

Fixando o valor de α começamos por calcular um intervalo

$$\left]\chi^2_{(n);\frac{\alpha}{2}}, \chi^2_{(n);1-\frac{\alpha}{2}}\right[$$

onde X^2 se situa, como ilustra a Figura 23: Para o cálculo dos extremos deste intervalo consultam-se

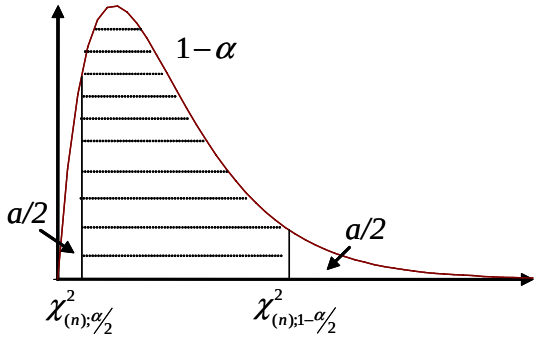


Figura 23: Intervalo de Confiança para X^2.

os valores $\chi^2_{(n);\frac{\alpha}{2}}$ e $\chi^2_{(n);1-\frac{\alpha}{2}}$ na tabela do Qui-Quadrado, correspondentes às probabilidades

$$P\left(X^2 < \chi^2_{(n);\frac{\alpha}{2}}\right) = \frac{\alpha}{2} \quad \text{e} \quad P\left(X^2 < \chi^2_{(n);1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}.$$

Então o Intervalo de Confiança para a variância a $(1-\alpha)100\%$ deduz-se do seguinte modo:

$$P\left(\chi^2_{(n);\frac{\alpha}{2}} < X^2 < \chi^2_{(n);1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$
$$\Leftrightarrow P\left(\chi^2_{(n);\frac{\alpha}{2}} < \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2} < \chi^2_{(n);1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$
$$\Leftrightarrow P\left(\frac{1}{\chi^2_{(n);\frac{\alpha}{2}}} > \frac{\sigma^2}{\sum\limits_{i=1}^{n}(X_i-\mu)^2} > \frac{1}{\chi^2_{(n);1-\frac{\alpha}{2}}}\right) = 1-\alpha \Leftrightarrow$$
$$\Leftrightarrow P\left(\frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{(n);1-\frac{\alpha}{2}}} < \sigma^2 < \frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{(n);\frac{\alpha}{2}}}\right) = 1-\alpha.$$

Então, o intervalo de confiança para σ^2 a $(1-\alpha)100\%$ é:

$$\left]\frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{(n);1-\frac{\alpha}{2}}}, \frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{(n);\frac{\alpha}{2}}}\right[.$$

Com μ desconhecido No caso de pretendermos estudar a variância de populações normais, em que μ é desconhecido usa-se a estatística e a distribuição amostral:

$$X^2 = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{(n-1)}.$$

Para deduzir o intervalo de confiança, fazemos,

$$P\left(\chi^2_{(n-1);\frac{\alpha}{2}} < X^2 < \chi^2_{(n-1);1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$
$$\Leftrightarrow P\left(\chi^2_{(n-1);\frac{\alpha}{2}} < (n-1)\frac{S^2}{\sigma^2} < \chi^2_{(n-1);1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$
$$\Leftrightarrow P\left(\frac{1}{\chi^2_{(n-1);\frac{\alpha}{2}}} > \frac{\sigma^2}{(n-1)S^2} > \frac{1}{\chi^2_{(n-1);1-\frac{\alpha}{2}}}\right) = 1-\alpha \Leftrightarrow$$
$$\Leftrightarrow P\left(\frac{(n-1)\,S^2}{\chi^2_{(n-1);1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)\,S^2}{\chi^2_{(n-1);\frac{\alpha}{2}}}\right) = 1-\alpha.$$

Então, o intervalo de confiança para σ^2 a $(1-\alpha)100\%$ é:

$$\left]\frac{(n-1)\,s^2}{\chi^2_{(n-1);1-\frac{\alpha}{2}}}, \frac{(n-1)\,s^2}{\chi^2_{(n-1);\frac{\alpha}{2}}}\right[.$$

Exemplo 38 Os dados seguintes são relativos aos pesos de 10 embalagens de adubo (em kgs) distribuídos por uma empresa,

$$46.4, \quad 46.1, \quad 45.8, \quad 47, \quad 46.1, \quad 45.9, \quad 45.8, \quad 46.9, \quad 45.2, \quad 46.$$

Determine um intervalo de confiança a 95% para a variância dos pesos, cuja distribuição se considera normal. Nas mesmas condições, determine um intervalo de confiança para o desvio padrão dos pesos e comente-o.

Como a população é Normal e μ é desconhecido vamos utilizar a distribuição

$$X^2 = 9\frac{S^2}{\sigma^2} \sim \chi^2_{(9)}.$$

Deduzindo o IC,

$$P\left(\chi^2_{(n-1);\frac{\alpha}{2}} < X^2 < \chi^2_{(n-1);1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$

$$\vdots$$

$$\Leftrightarrow P\left(\frac{(n-1)\,S^2}{\chi^2_{(n-1);1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)\,S^2}{\chi^2_{(n-1);\frac{\alpha}{2}}}\right) = 1-\alpha.$$

Temos de calcular,

$$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 46.12 \text{ kgs}$$

e

$$s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - 46.12)^2 = 0.286(2).$$

Sabendo que $(1-\alpha) = 0.95 \Leftrightarrow \frac{\alpha}{2} = 0.025 \Leftrightarrow (1-\frac{\alpha}{2}) = 0.975$, então,

$$\chi^2_{0.025}(9) = 2.7 \text{ e } \chi^2_{0.975}(9) = 19.$$

O intervalo de confiança para σ^2 a 95% de confiança é:

$$\left]\frac{9\times 0.2862}{19}, \frac{9\times 0.2862}{2.7}\right[=]0.1355, 0.9541[\,.$$

Para calcular o intervalo de confiança para σ a 95% de confiança basta fazer:

$$\left]\sqrt{\frac{9\times 0.2862}{19}}, \sqrt{\frac{9\times 0.2862}{2.7}}\right[=]0.3682, 0.9767[\,.$$

O intervalo para o desvio padrão indica que as embalagens têm uma variabilidade média no peso que pode ir de 368.2 a 976.7 gramas, com 95% de confiança. ■

4.6.6 Intervalo de Confiança para a Razão de Duas Variâncias

Neste caso vai utilizar-se a estatística e a distribuição amostral:

$$F = \frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} \sim F_{(n_1-1,n_2-1)}.$$

Fixando o valor de α começamos por calcular um intervalo

$$\left]f_{(n_1-1,n_2-1);\frac{\alpha}{2}}, f_{(n_1-1,n_2-1);1-\frac{\alpha}{2}}\right[$$

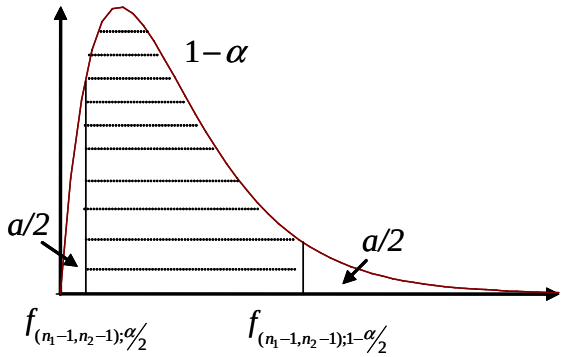


Figura 24: Intervalo de Confiança para a v.a. F.

onde F se situa como é ilustrado na Figura 24: Para o cálculo dos extremos deste intervalo consultam-se os valores $f_{(n_1-1,n_2-1);\frac{\alpha}{2}}$ e $f_{(n_1-1,n_2-1);1-\frac{\alpha}{2}}$ na tabela da F-Snedcor; o primeiro quantil não é imediato, como tal aplicamos o Teorema 36 obtendo-o do seguinte modo:

$$f_{(n_1-1,n_2-1);\frac{\alpha}{2}} = \frac{1}{f_{(n_2-1,n_1-1);1-\frac{\alpha}{2}}}.$$

O intervalo de confiança para a razão entre duas variâncias a $(1-\alpha)100\%$ deduz-se do seguinte modo:

$$\begin{aligned}
&P\left(f_{(n_1-1,n_2-1);\frac{\alpha}{2}} < F < f_{(n_1-1,n_2-1);1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow \\
\Leftrightarrow\ &P\left(f_{(n_1-1,n_2-1);\frac{\alpha}{2}} < \frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} < f_{(n_1-1,n_2-1);1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow \\
\Leftrightarrow\ &P\left(f_{(n_1-1,n_2-1);\frac{\alpha}{2}} \times \frac{S_2^2}{S_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < f_{(n_1-1,n_2-1);1-\frac{\alpha}{2}} \times \frac{S_2^2}{S_1^2}\right) = 1-\alpha \Leftrightarrow \\
\Leftrightarrow\ &P\left(\frac{1}{f_{(n_1-1,n_2-1);1-\frac{\alpha}{2}}} \times \frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{f_{(n_1-1,n_2-1);\frac{\alpha}{2}}} \times \frac{S_1^2}{S_2^2}\right) = 1-\alpha.
\end{aligned}$$

Então, o intervalo de confiança para $\frac{\sigma_1^2}{\sigma_2^2}$ a $(1-\alpha)100\%$ é:

$$\left]\frac{1}{f_{(n_1-1,n_2-1);1-\frac{\alpha}{2}}} \times \frac{s_1^2}{s_2^2}, \frac{1}{f_{(n_1-1,n_2-1);\frac{\alpha}{2}}} \times \frac{s_1^2}{s_2^2}\right[.$$

Exemplo 39 Pretende-se comparar o tempo que duas máquinas, A e B, gastam no fabrico de uma peça. A partir de 13 peças fabricadas na máquina A e de 16 peças fabricadas na máquina B, obtiveram-se os seguintes resultados para as variâncias dos tempos

$$s_1^2 = 6.32 \qquad s_2^2 = 4.80.$$

Admitindo que o tempo de fabrico das peças tem um comportamento normal, vamos determinar, a 95%, um intervalo de confiança para a razão das variâncias $\frac{\sigma_1^2}{\sigma_2^2}$.

Aplica-se

$$F = \frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} \sim F_{(12,15)}.$$

Deduzindo o IC,

$$P\left(f_{(n_1-1,n_2-1);\frac{\alpha}{2}} < F < f_{(n_1-1,n_2-1);1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$

$$\vdots$$

$$\Leftrightarrow P\left(\frac{1}{f_{(n_1-1,n_2-1);1-\frac{\alpha}{2}}} \times \frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{f_{(n_1-1,n_2-1);\frac{\alpha}{2}}} \times \frac{S_1^2}{S_2^2}\right) = 1-\alpha.$$

Como $\frac{\alpha}{2} = 0.025 \Leftrightarrow 1-\frac{\alpha}{2} = 0.975$, obtemos directamente da tabela

$$f_{(12,15);0.975} = 2.96$$

e para o outro quantil fazemos

$$f_{(12,15);0.025} = \frac{1}{f_{(15,12);0.975}} = \frac{1}{3.18} = 0.3145.$$

O intervalo de confiança para $\frac{\sigma_1^2}{\sigma_2^2}$ a 95% é:

$$\left]\frac{1}{2.96} \times \frac{6.32}{4.80}, \frac{1}{0.3145} \times \frac{6.32}{4.80}\right[=]0.4448, 4.1865[\,.$$

Conclui-se que, com 95% de confiança, a razão das variâncias se situa entre 0.4448 e 4.1865, o que significa que não deve existir grande diferença entre as variâncias dos tempos das duas máquinas (pois o valor 1, correspondente a $\sigma_1^2 = \sigma_2^2$, encontra-se no intervalo). ■

4.7 Notas sobre Distribuições Amostrais e Intervalos de Confiança

Em toda a exposição atrás realizada, considerou-se sempre o caso de amostras independentes, em que a probabilidade de escolha é a mesma para qualquer elemento da população ao longo de sucessivas tiragens. Isto implica que quando trabalhamos com populações finitas a amostragem é feita com reposição. No entanto, na prática geralmente sucede o contrário, isto é, a amostra é feita sem reposição, o que implica alterações nos parâmetros de amostragem de algumas estatísticas. Nestas condições os intervalos de confiança atrás apresentados são válidos para populações infinitas (ou populações finitas em que é utilizada a amostragem com reposição), porém, para o caso de populações finitas em que é utilizada amostragem sem reposição, é necessário corrigir os limites de confiança indicados.

Em resumo, para amostras extraídas com reposição de uma população X finita ou infinita tem-se que:

$$E\left[\bar{X}\right] = \mu \text{ e } V\left[\bar{X}\right] = \frac{\sigma^2}{n}.$$

Para populações finitas e amostras extraídas sem reposição (com N elementos de entre os quais n têm determinada característica) tem-se:

$$E\left[\bar{X}\right] = \mu \text{ e } V\left[\bar{X}\right] = \frac{\sigma^2}{n} \times \frac{N-n}{N-1}.$$

Exemplo 40 Uma companhia que transporta barris de petróleo recebe um carregamento de 100 barris, pretendendo estudar o diâmetro médio dos barris devido a problemas de carregamento dos mesmos. Uma amostra, sem reposição, de 50 barris fornece o diâmetro médio de 2.55. No passado

o desvio padrão do diâmetro da população foi de 0.07. Construa um intervalo de confiança a 99% para a média.

Como estamos perante uma população finita em que σ é conhecido e a amostra é realizada sem reposição, vamos utilizar a estatística com a distribuição amostral,

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}} \sim \mathcal{N}(0,1).$$

Como $\sigma = 0.07, N = 100, n = 50$ e $(1-\alpha) = 0.99$, temos,

$$Z = \frac{\bar{X} - \mu}{\frac{0.07}{\sqrt{50}} \times \sqrt{\frac{50}{99}}} \sim \mathcal{N}(0,1).$$

DeDuzindo o IC,

$$P\left(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \Leftrightarrow$$

$$\vdots$$

$$\Leftrightarrow P\left(\bar{X} - z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} < \mu < \bar{X} + z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}\right) = 1-\alpha.$$

Dado que $1 - \frac{\alpha}{2} = 0.995$ e $z_{0.995} = 2.58$, então, o intervalo de confiança para μ a 99% de confiança é:

$$]2.55 - 2.58 \times 0.007, 2.55 + 2.58 \times 0.007[=]2.532, 2.568[.$$

Conclui-se que o diâmetro médio dos barris se situa entre 2.532 e 2.568, com 99% de confiança. ■

5 TESTES DE HIPÓTESES

5.1 Generalidades

Todos os dias temos de tomar decisões respeitantes a determinadas populações, com base em amostras das mesmas (decisões estatísticas). Nesta tomada de decisões é util formular hipóteses sobre as populações, hipóteses essas que podem ou não ser verdadeiras. A essas hipóteses chamamos hipóteses estatísticas, as quais geralmente se baseiam em afirmações sobre as distribuições de probabilidade das populações ou alguns dos seus parâmetros. Por vezes estas hipóteses, ao serem formuladas, têm por único objectivo serem rejeitadas.

Exemplo 41 Se queremos decidir se uma dada moeda está viciada, formulamos a hipótese de que a moeda seja "*honesta*", isto é, que a probabilidade de sair por exemplo cara seja $p = 0.5$. Da mesma forma, se queremos decidir se um produto é melhor do que outro, podemos formular a hipótese de que não existe diferença entre ambos os produtos. ■

Desta forma os testes de hipóteses podem considerar-se uma segunda vertente da inferência estatística, tendo por objectivo verificar, a partir de dados observados numa ou várias amostras, a validade de certas hipóteses relativas a uma ou várias populações.

5.2 Princípios da realização dos testes de hipóteses

1. De uma forma geral emite-se uma certa hipótese a testar denominada **Hipótese Nula** e representada por H_0:

   (a) em seguida medimos o desvio observado em certas características da amostra e calculamos a probabilidade, se H_0 for verdadeira, do desvio ser "importante";

   (b) se a probabilidade anterior for "relativamente elevada" (isto é, superior a um nível de significância, α, previamente definido), consideramos plausivel H_0 e aceitamo-la, pelo menos provisóriamente; quando um teste não rejeita H_0 não se pode concluir que esta seja verdadeira, mas apenas que não está em desacordo com os factos observados, como tal utiliza-se a expressão *não rejeitar* H_0 em vez de *aceitar* H_0;

   (c) se, pelo contrário, a probabilidade for "pequena" (isto é, inferior a um nível de significância, α, previamente definido), o desvio observado mostra-se pouco compatível com H_0 e rejeitamo-la. Desta forma admitimos, implicitamente, a validade da outra hipótese, denominada por **Hipótese Alternativa** e representada por H_1.

2. O conjunto dos valores observados para os quais H_0 é admissível forma a **Região de Aceitação** (representada por RA). Os restantes valores formam a **Região de Rejeição** ou **Região Crítica** (representada por RC) como podemos ver na Figura 25.

3. Consoante o número de elementos em análise num teste, #, podemos distinguir diferentes formas de especificar H_0 (que traduz a situação estacionária, sendo usual colocar nesta hipótese a igualdade) e H_1, considerando, por exemplo, θ^* estimador de θ:

   (a) hipótese simples (ou composta) contra hipótese composta (em que $\#\{\theta_{H_0}\} = 1$ (ou $\#\{\theta_{H_0}\} > 1$) e $\#\{\theta_{H_1}\} > 1$). Podemos neste tipo de testes estar perante,

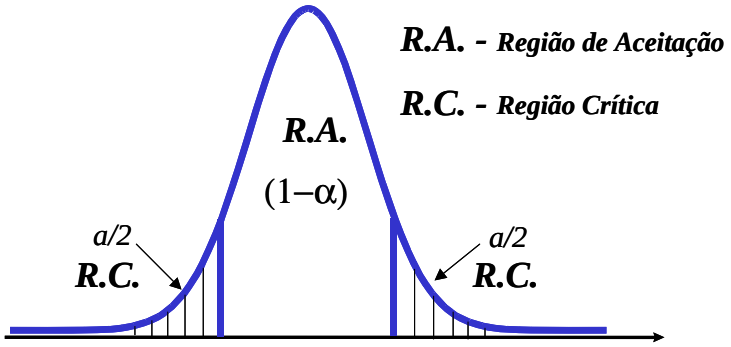


Figura 25: Região Crítica e Região de Aceitação num Teste de Hipóteses.

i. Teste Bilateral que apresenta duas regiões críticas como vemos na Figura 26.

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

Como,

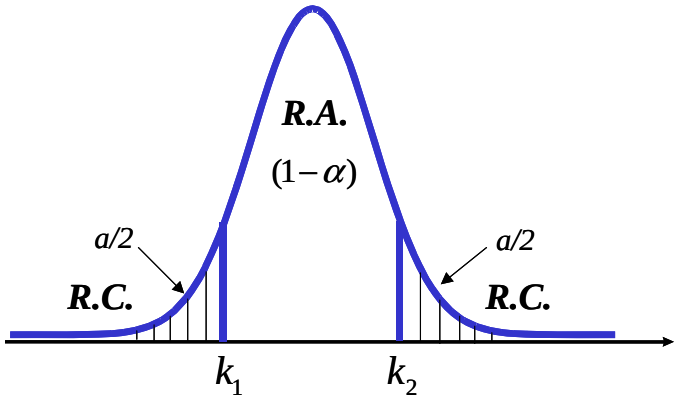


Figura 26: Teste Bilateral.

$$P\left(k_1 \leq \theta^*\right) = P\left(k_2 \geq \theta^*\right)$$

então,

$$P\left(\text{Rej. } H_0/H_0V\right) = P(\theta^* \in RC/\theta = \theta_0) = \alpha \Leftrightarrow$$
$$\Leftrightarrow \begin{cases} P\left(\theta^* \leq k_1/\theta = \theta_0\right) = \frac{\alpha}{2} \\ P\left(\theta^* \geq k_2/\theta = \theta_0\right) = \frac{\alpha}{2} \end{cases}.$$

ii. Teste Unilateral Esquerdo que apresenta a região crítica à esquerda como vemos na Figura 27.

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta < \theta_0 \end{cases} \quad \text{ou} \quad \begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$

$$P\left(\text{Rej. } H_0/H_0V\right) = P(\theta^* \in RC/\theta = \theta_0) =$$
$$= P\left(\theta^* \leq k/\theta = \theta_0\right) = \alpha.$$

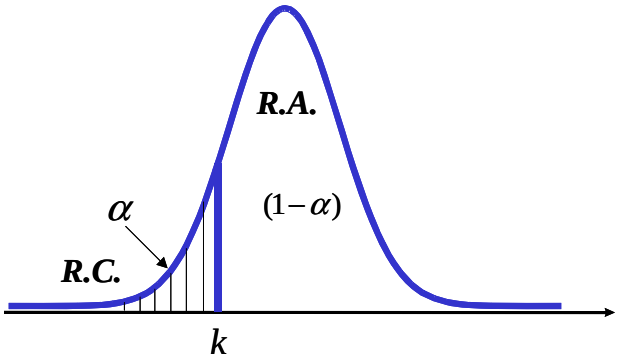


Figura 27: Teste Unilateral Esquerdo.

iii. Teste Unilateral Direito que apresenta a região crítica à direita como vemos na Figura 28.

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{cases} \quad \text{ou} \quad \begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

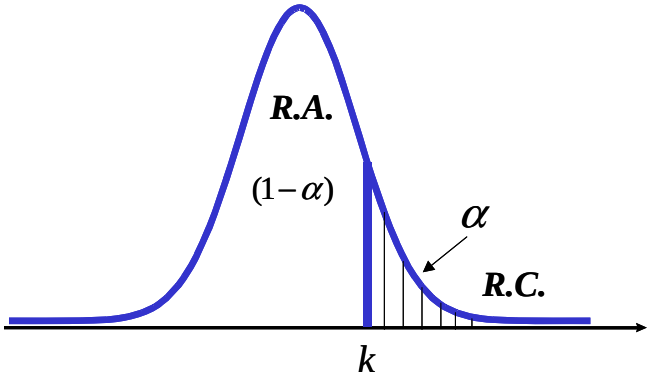


Figura 28: Teste Unilateral Direito.

$$P\left(\text{Rej. } H_0/H_0V\right) = P(\theta^* \in RC/\theta = \theta_0) = \\ = P\left(\theta^* \geq k/\theta = \theta_0\right) = \alpha.$$

(b) hipótese simples contra hipótese simples (em que $\#\{\theta_{H_0}\} = 1$ e $\#\{\theta_{H_1}\} = 1$).

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$$

Neste caso estamos sempre perante um teste unilateral considerado esquerdo, se $\theta_0 > \theta_1$, ou direito se $\theta_0 < \theta_1$.

4. Existe uma relação entre a teoria da estimação, envolvendo intervalos de confiança, e a teoria relativa aos testes de hipóteses. Quando trabalhamos com testes de hipóteses bilaterais podemos efectivamente utilizar os intervalos de confiança para testar hipóteses (pois o intervalo de confiança coincide com a região de aceitação). Resultado análogo para testes unilaterais, exigiriam intervalos de confiança unilaterais, os quais, embora de rara aplicação prática, são possíveis de definir.

5. Um teste de hipóteses nem sempre conduz a decisões correctas pois a análise de uma amostra pode, como é evidente, falsear as conclusões. Como tal podemos encontrar-nos perante quatro situações distintas apresentadas na tabela seguinte:

Decisão Tomada	Situação Real	
	H_0 Verdadeira	H_0 Falsa
Rejeita-se H_0	Erro de 1^a espécie (α)	Decisão correcta (π)
Não se rejeita H_0	Decisão correcta	Erro de 2^a espécie (β)

(a) Num erro de 1^a espécie (cuja probabilidade se representa por α, ou nível de significância do teste) rejeita-se H_0, sendo esta verdadeira, logo

$$\alpha = P(\text{Rej. } H_0/H_0V).$$

(b) Num erro de 2^a espécie (cuja probabilidade se representa por β) não se rejeita H_0, sendo esta falsa (ou H_1 verdadeira), logo

$$\beta = P(\text{Não Rej. } H_0/H_0F) = P(\text{Não Rej. } H_0/H_1V).$$

Num teste unilateral direito, os dois erros podem geometricamente representar-se como mostra a Figura 29 em que a função cujo gráfico está a tracejado representa o comportamento do verdadeiro parâmetro da população. Num teste bilateral, os dois erros

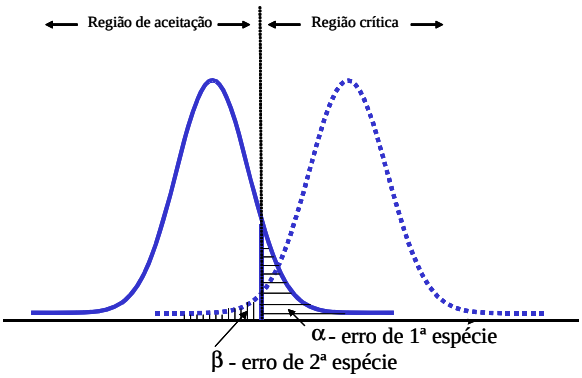


Figura 29: Erros de 1^a e 2^a espécie.

podem geometricamente representar-se como mostra a Figura 30, em que as três últimas funções representam comportamentos possíveis da função densidade de probabilidade da população, com os respectivos erros de 1^a e 2^a espécies.

(c) Chama-se **função potência** de um teste e representa-se por π à probabilidade de rejeitar H_0 quando esta é falsa (decisão correcta). Então podemos dizer que dado o erro de 2^a espécie β, a função potência é o seu complementar

$$\pi = P(\text{Rej. } H_0/H_0F) = P(\text{Rej. } H_0/H_1V) = 1 - \beta.$$

Esta probabilidade é função do grau de falsidade de H_0, logo a probabilidade de rejeição é tanto mais elevada, quanto mais falsa for H_0. Conclui-se então que a relação entre a probabilidade de rejeição de H_0 e o grau de falsidade da mesma constituem a função

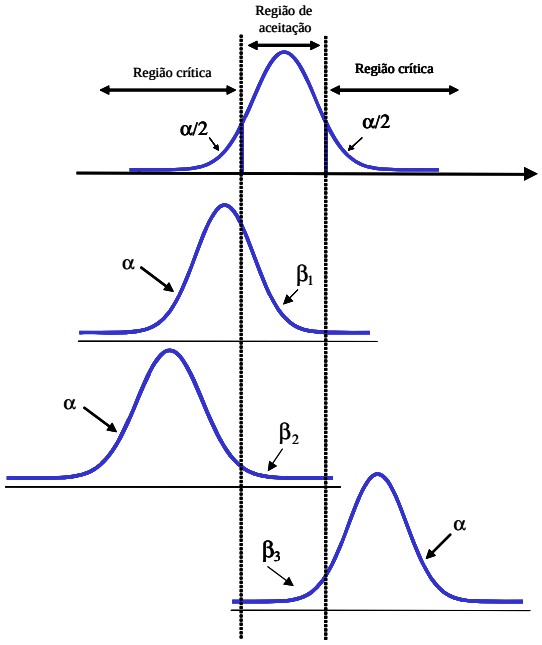


Figura 30: Erros num Teste Bilateral.

potência do teste, isto é, quanto maiores forem os valores da função potência, menor é o erro de 2^a espécie cometido, logo, melhor a qualidade do teste (teste mais potente).

Num teste bilateral o gráfico da função potência tem a forma de um V como se visualiza na Figura 31. Um V estreito (com um declive acentuado) indica que o valor do parâmetro definido na hipótese nula e os diversos valores da hipótese alternativa estão bem discriminados; se pelo contrário, o V for largo, indica uma fraca discriminação nos valores formulados nas hipóteses. Num teste unilateral direito o gráfico da função

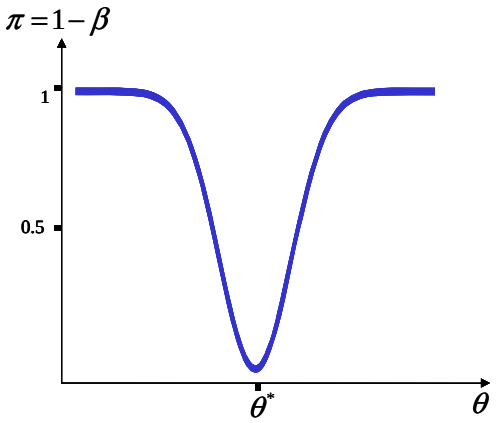


Figura 31: Função potência de um Teste Bilateral.

potência tem a forma de um S como se observa na Figura 32. Mais uma vez, o declive acentuado indica que o valor do parâmetro definido na hipótese nula e diversos valores da hipótese alternativa estão bem discriminados; se pelo contrário o declive for pouco acentuado indica uma fraca discriminação nos valores formulados nas hipóteses.

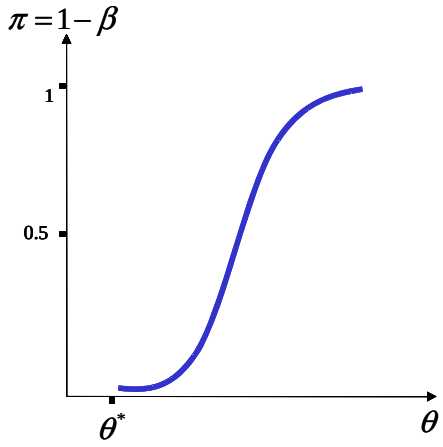


Figura 32: Função potência num Teste Unilateral Direito.

(d) É através das probabilidades α e β que se procura o melhor teste de hipóteses, sendo o teste ideal o que minimiza simultâneamente ambos os valores. No entanto, e como α e β variam em sentidos contrários, tal não é possível. O que na maior parte dos casos se faz (com base no Teorema de *Neyman-Pearson*[4]) consiste em fixar α (para amostras de dimensão n) para tentar minimizar β. Note-se ainda que é possível fixar α e β *a priori*, ficando n livre; no entanto este método que se baseia em valores pequenos de α e β conduz a valores de n grandes, o que nem sempre é conveniente.

6. Os erros anteriores não podem ser completamente evitados, no entanto, pode-se manter pequena a probabilidade de os cometer. Na prática fixa-se um limite superior de risco de erro de 1^a espécie (α), que na maior parte dos casos se situa entre 1% e os 5% ($\alpha = 0.01$ até $\alpha = 0.05$). Este limite, ou nível de significância do teste, é que permite definir a condição de rejeição de H_0.

5.3 Testes de Hipóteses Paramétricos

Nos testes de hipóteses paramétricos ou realizados a parâmetros de uma população, e ao contrário dos intervalos de confiança, em vez de procurarmos uma estimativa ou um intervalo para o parâmetro, vamos admitir um valor hipotético para o mesmo e depois utilizar a informação da amostra para rejeitar ou não esse valor. Nos casos que em seguida apresentamos vamos debruçar-nos apenas sobre populações com distribuições normais (ou aproximadamente normais).

Passemos a enunciar, de uma forma geral, a metodologia a utilizar num teste de hipóteses paramétrico:

1. formulação das hipóteses;
2. fixação do erro de 1^a espécie ou nível de significância do teste
$$\alpha = P(\text{Rej. } H_0/H_0V);$$
3. escolha da estatística (também denominada por estatística teste ou variável fulcral) e respectiva distribuição amostral adequadas;

[4]Página 307 e seguintes de Bento Murteira, *Probabilidades e Estatística*, Volume II, McGraw-Hill,1990.

4. cálculo de RC a partir do nível de significância do teste, α;

5. com base na amostra calcula-se o estimador θ^* do parâmetro θ, e aplica-se a regra de decisão:

$$\begin{cases} \text{se } \theta^* \in RC \Rightarrow \text{rejeitar } H_0 \\ \text{se } \theta^* \in RA \Rightarrow \text{não rejeitar } H_0. \end{cases}$$

5.3.1 Testes de Hipóteses para a Média

Neste caso a estatística a utilizar é

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Como vimos no capítulo anterior, consoante os restantes parâmetros sejam ou não conhecidos e a dimensão da amostra seja grande ou pequena, vamos utilizar diferentes estatísticas teste e respectivas distribuições amostrais. Para os testes de hipóteses este procedimento repete-se, logo vamos utilizar:

1. se σ é conhecido,
$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1);$$

2. se σ é desconhecido e $n \geq 30$,
$$Z = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim \mathcal{N}(0,1);$$

3. se σ é desconhecido e $n < 30$,
$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{(n-1)}.$$

Exemplo 42 De um universo normal, de média e variância desconhecidas, foi retirada uma amostra aleatória de 9 observações, cujos resultados foram:

$$\sum_{i=1}^{9} x_i = 36 \quad \text{e} \quad \sum_{i=1}^{9} x_i^2 = 162.$$

Proceda ao seguinte ensaio de hipóteses:

$$\begin{cases} H_0 : \mu = 5 \\ H_1 : \mu = 6 \end{cases}$$

para um nível de significância de 5%.

A estatística para o estudo do parâmetro μ é $\bar{X}$. Como desconhecemos a variância da populção e $n < 30$, utilizamos a estatística teste e a distribuição amostral:

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{9}}} \sim t_{(8)}.$$

Impõe-se então calcular $\bar{x}$ e s:

$$\bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = 4$$

$$s^2 = \frac{1}{8}\sum_{i=1}^{9} x_i^2 - \frac{9}{8}(4)^2 = 2.25 \Rightarrow s = 1.5.$$

Partindo de

$$T = \frac{\bar{X} - \mu}{\frac{1.5}{\sqrt{9}}} \sim t_{(8)}$$

e de $\alpha = 0.05$, vamos calcular RC de um teste unilateral direito (dado que H_1 está sempre associada a RC) como podemos ver na Figura 33. Para tal podemos seguir duas metodologias equivalentes

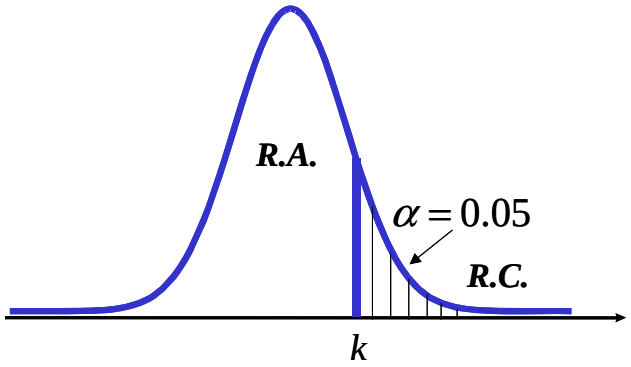


Figura 33: Região crítica associada ao teste.

mas que correspondem a escalas diferentes no cálculo da região crítica. Na primeira metodologia k (que separa RA de RC) é obtido a partir da estatística $\bar{X}$, assim como a tomada de decisão consiste em verificar se a estimativa da média amostral se situa em RC ou RA. Na segunda metodologia k é obtido a partir da estatística teste T, assim como a tomada de decisão consiste em verificar se a estimativa da estatística teste, T^*, se situa em RC ou RA. Ambos os procedimentos são equivalentes, variando apenas a escala utilizada para o cálculo de RC, assim como para a tomada de decisão.

Vamos começar por resolver este exemplo através da primeira metodologia:

$$P(\text{Rej. } H_0/H_0V) = \alpha \Leftrightarrow P\left(\bar{X} \in RC/H_0V\right) = P\left(\bar{X} \geq k/\mu = 5\right) = 0.05 \Leftrightarrow$$

$$\Leftrightarrow P\left(T \geq \frac{k-\mu}{\frac{1.5}{\sqrt{9}}}/\mu = 10\right) = 0.05 \Leftrightarrow P\left(T < \frac{k-5}{\frac{1.5}{3}}\right) = 0.95 \Leftrightarrow$$

$$\Leftrightarrow \frac{k-5}{\frac{1.5}{3}} = 1.85 \Leftrightarrow k = 5.93.$$

Então $RC = [5.93, +\infty[$, como podemos visualizar na Figura 34. Como $\bar{x} = 4 < 5.93$ se encontra na região de aceitação (RA), não se rejeita H_0.

Na segunda metodologia k é obtido a partir da estatística teste T, sendo RC calculado a partir do quantil que lhe corresponde, isto é, RC começa a partir do quantil referente a $(1-\alpha) = (1-0.05) = 0.95$. Como

$$t_{(8);0.95} = 1.86, \text{ então, } RC = [1.86, +\infty[\,.$$

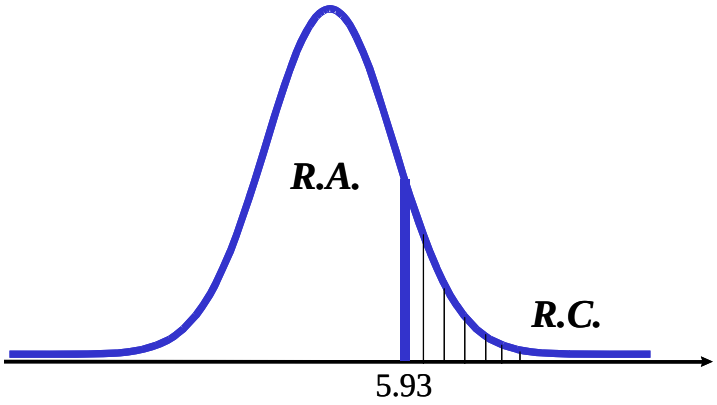


Figura 34: Região crítica calculada em função de $\bar{X}$.

Podemos desta forma visualizar na Figura 35 a mudança de escala da região crítica utilizando a 2^a metodologia. Como a estimativa de T é dada por

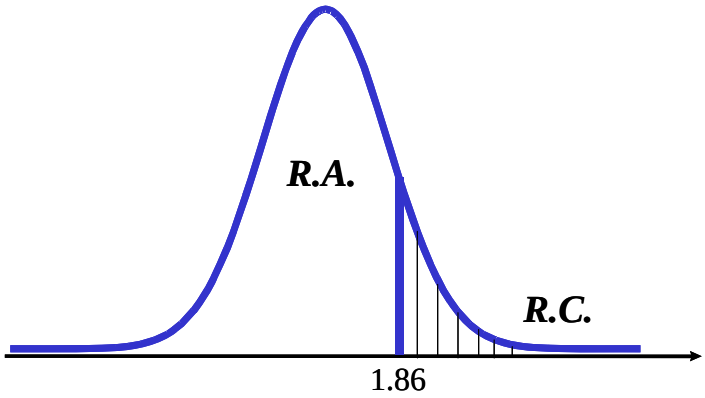


Figura 35: Região Crítica calculada em função de T.

$$T^* = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{9}}} = \frac{4-5}{\frac{1.5}{3}} = -2.0 \in RA,$$

não se rejeita H_0. ■

Exemplo 43 Para $X \sim \mathcal{N}(\mu, 100)$, $n = 25$, $\bar{x} = 980$ e $\alpha = 0.05$, vamos calcular RC, erros de 2^a espécie e a função potência para

$$\begin{cases} H_0 : \mu = 1000 \\ H_1 : \mu < 1000 \end{cases}$$

A estatística para o estudo do parâmetro μ é $\bar{X}$. Como conhecemos a variância da distribuição, utilizamos a estatística teste e a distribuição amostral

$$Z = \frac{\bar{X} - \mu}{\frac{100}{\sqrt{25}}} \sim \mathcal{N}(0, 1).$$

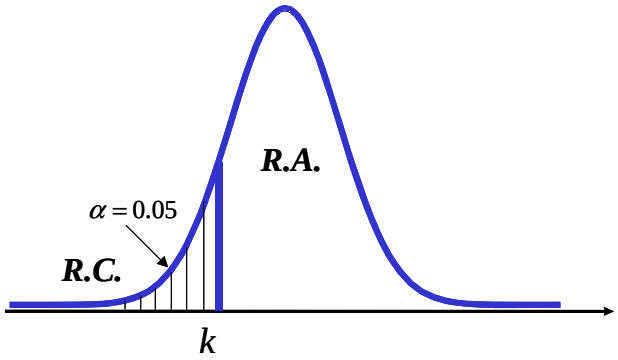


Figura 36: Região crítica associada ao teste.

Temos um teste unilateral esquerdo tal como podemos ver na Figura 36. Utilizando a primeira metodologia temos:

$$P(\text{Rej. } H_0/H_0V) = \alpha \Leftrightarrow P\left(\bar{X} \in RC/H_0V\right) = P\left(\bar{X} \leq k/\mu = 1000\right) = 0.05 \Leftrightarrow$$
$$\Leftrightarrow P\left(Z \leq \frac{\bar{X}-\mu}{\frac{100}{5}}/\mu = 1000\right) = 0.05 <=> P\left(Z \leq \frac{k-1000}{20}\right) = 0.05 \Leftrightarrow$$
$$\Leftrightarrow \frac{k-1000}{20} = -1.645 \Leftrightarrow k = 967.1$$

Então $RC =]-\infty, 967.1]$ como se visualiza na Figura 37.

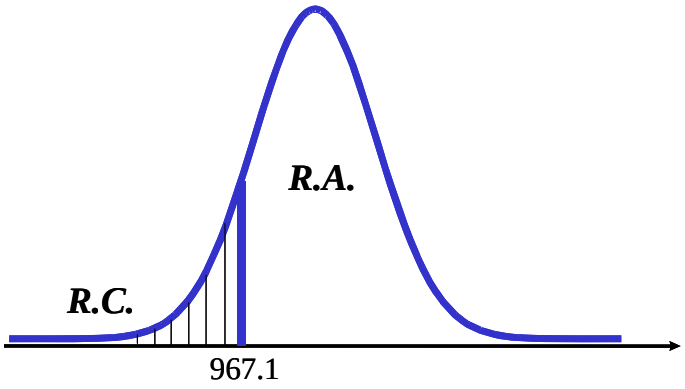


Figura 37: Região crítica calculada em função de $\bar{X}$.

Como $\bar{x} = 980 \in RA$, não se rejeita H_0.
Utilizando a segunda metodologia,

$$z_{0.05} = -z_{0.95} = -1.645, \text{ isto é, } RC =]-\infty, -1.645].$$

Como a estimativa de Z é dada por

$$Z^* = \frac{\bar{x}-\mu}{\frac{100}{\sqrt{25}}} = \frac{980-1000}{\frac{100}{\sqrt{25}}} = -1 \in RA,$$

não se rejeita H_0.

Embora a segunda metodologia seja mais rápida, a primeira é mais directa quando pretendemos calcular o erro de 2^a espécie, como em seguida veremos:

$$\begin{aligned}\beta &= P(\text{Não Rej. } H_0/H_0F) = P(\bar{X} \in RA/H_1V) = \\ &= P(\bar{X} > k/\mu < 1000) = P(\bar{X} > 967.1/\mu < 1000) = \\ &= 1 - P\left(Z \leq \frac{967.1-\mu}{20}/\mu < 1000\right).\end{aligned}$$

Atribuindo alguns valores a μ, por exemplo 999, 990, 970, 950, 930 e 910, calculamos o respectivo erro de 2^a espécie e correspondente função potência cujos valores se encontram na tabela seguinte:

μ	$\beta(\mu)$	$\pi(\mu)$
999	0.9446	0.0554
990	0.8729	0.1271
970	0.5557	0.4443
950	0.1977	0.8023
930	0.0318	0.9682
910	0.0022	0.9978

Podemos ainda expressar graficamente estas duas funções através da Figura 38.

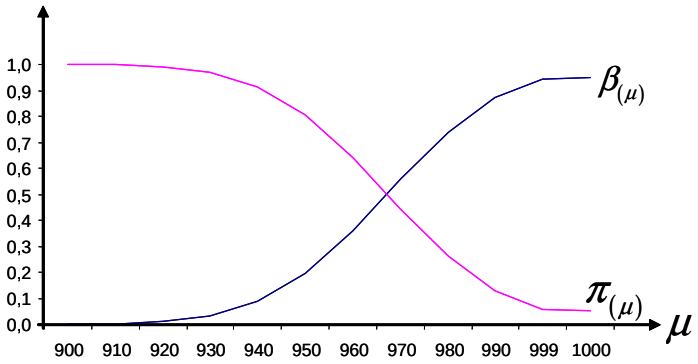


Figura 38: Erro de 2^a espécie e correspondente função potência do teste. ■

5.3.2 Testes de Hipóteses para a Diferença de Duas Médias

Nestes casos há que diferenciar mais uma vez as estatísticas teste e respectivas distribuições amostrais a utilizar:

1. se os desvios padrões são conhecidos,

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1);$$

2. se os desvios padrões são desconhecidos, $n_1 \geq 30$ e $n_2 \geq 30$,

$$T = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \mathcal{N}(0,1);$$

3. se os desvios padrões são desconhecidos (e iguais), $n_1 < 30$ ou $n_2 < 30$,

$$T = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}} \sim t_{(n_1+n_2-2)}.$$

Exemplo 44 A altura média de 50 atletas de um dado clube que tiveram bons resultados em competições desportivas, é de 68.2 polegadas, com desvio padrão de 2.5 polegadas, enquanto que um grupo de 50 atletas do mesmo clube com resultados inferiores nessas competições tem altura média de 67.5 polegadas com desvio padrão de 2.8 polegadas. Vamos testar a hipótese de que os atletas que obtiveram bons resultados nas competições são, em média, mais altos do que os restantes (com $\alpha = 0.05$).

Devemos então proceder ao teste de hipóteses:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases}$$

o que significa confrontar a inexistência de diferença entre as médias das alturas dos dois grupos de atletas, contra a altura média do 1^o grupo de atletas ser superior à do 2^o grupo.

Como os desvios padrões são desconhecidos, $n_1 \geq 30$ e $n_2 \geq 30$ utilizamos a estatística teste e a distribuição amostral

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{2.5^2}{50} + \frac{2.8^2}{50}}} \sim \mathcal{N}(0,1).$$

Vamos resolver o exemplo recorrendo à segunda metodologia.

Como estamos perante um teste unilateral direito, RC começa a partir do quantil associado a $(1-\alpha) = (1-0.05) = 0.95$, logo,

$$RC = [z_{0.95}, +\infty[= [1.645, +\infty[.$$

Sendo $(\bar{x}_1 - \bar{x}_2) = (68.2 - 67.5) = 0.7$, a estimativa de Z é dada por

$$Z^* = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{2.5^2}{50} + \frac{2.8^2}{50}}} = \frac{0.7 - 0}{\sqrt{\frac{2.5^2}{50} + \frac{2.8^2}{50}}} = 1.4 \in RA.$$

Conclui-se que não se rejeita H_0, isto é, não se prova que a altura média dos dois grupos de atletas difira para o valor de α considerado. ■

A tomada de decisão face ao resultado de um teste de hipóteses não dá garantia de que estejamos a agir de forma correcta pois basta alterar o tipo de teste (por exemplo de unilateral para bilateral), o nível de significância ou a dimensão da amostra, para que o resultado do teste possa ser completamente diferente. Esta situação está ilustrada no exemplo seguinte.

Exemplo 45 Se, para o exemplo anterior, quiséssemos que a diferença observada entre as alturas médias, de 0.7 polegadas, fosse significativa, qual deveria ser a dimensão das amostras (mantendo a igualdade entre as mesmas)?

Neste caso, pretendemos que $Z^* \in RC$, isto é,

$$\begin{aligned} Z^* \geq 1.645 &\Leftrightarrow \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{2.5^2}{n} + \frac{2.8^2}{n}}} \geq 1.645 \Leftrightarrow \frac{0.7 - 0}{\sqrt{\frac{2.5^2}{n} + \frac{2.8^2}{n}}} \geq 1.645 \Leftrightarrow \\ &\Leftrightarrow 1.645\sqrt{\frac{2.5^2}{n} + \frac{2.8^2}{n}} \leq 0.7 \Leftrightarrow 1.645\frac{3.75}{\sqrt{n}} \leq 0.7 \Leftrightarrow \\ &\Leftrightarrow \sqrt{n} \geq \frac{1.645 \times 3.75}{0.7} \Rightarrow n \geq 78. \end{aligned}$$

■

Exemplo 46 Os quocientes de inteligência (QI) de 16 estudantes de um dado bairro de uma cidade apresentaram uma média de 107 com um desvio padrão de 10; entretanto, noutro bairro da mesma cidade, analisaram-se 14 estudantes cujos QI tinham uma média de 112 e um desvio padrão de 8. Há diferenças significativas entre os QI dos dois grupos (considere $\alpha = 0.01$ e as populações normais com $\sigma_1 = \sigma_2$)?

Para resolver esta questão vamos elaborar o seguinte teste de hipóteses bilateral:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

O que significa confrontar a inexistência de diferença significativa entre as médias dos QI contra a existência dessa mesma diferença.

Como as populações são Normais, $\sigma_1 = \sigma_2$, $n_1 < 30$ e $n_2 < 30$, vamos utilizar,

$$T = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{16} + \frac{1}{14}\right) \frac{15\times 10^2 + 13\times 8^2}{28}}} \sim t_{(28)}.$$

Estamos perante um teste bilateral traduzido na Figura 39.

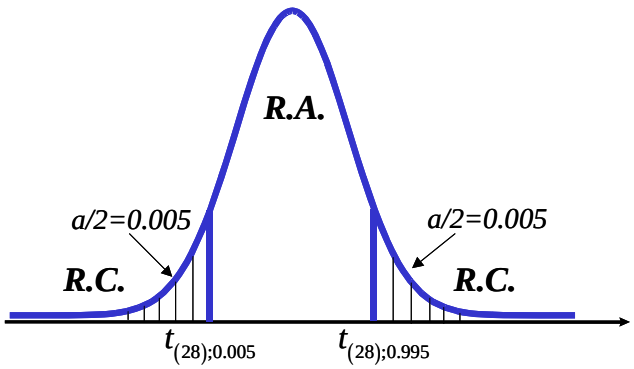


Figura 39: Teste bilateral.

A região crítica calcula-se fazendo,

$$RC = \left]-\infty, t_{(28);0.005}\right] \cup \left[t_{(28);0.995}, +\infty\right[=]-\infty, -2.763] \cup [2.763, +\infty[\,.$$

Como $(\bar{x}_1 - \bar{x}_2) = (107 - 112) = -5$, a estimativa de T é dada por

$$T^* = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{16} + \frac{1}{14}\right) \frac{15\times 10^2 + 13\times 8^2}{28}}} = \frac{-5 - 0}{\sqrt{\frac{8745}{784}}} = -1.4971 \in RA$$

não se rejeita H_0, isto é, conclui-se que não existem diferenças significativas entre as médias dos QI dos dois grupos. ■

5.3.3 Teste de Hipóteses para uma Proporção

Neste caso, e considerando p^* a proporção observada na amostra, estimativa da proporção desconhecida (p) da população, vamos utilizar a variável aleatória

$$Z = \frac{p^* - p}{\sqrt{\frac{pq}{n}}} \sim \mathcal{N}(0, 1)$$

para grandes amostras (na prática $n \geq 30$).

Chama-se a atenção para o facto da estatística utilizada nos Intervalos de Confiança ser uma aproximação desta estatística teste (no respeitante à variância de p^*).

5.3.4 Teste de Hipóteses para a Diferença de Duas Proporções

Neste caso, e considerando p_1^* e p_2^* as proporções observadas nas amostras, estimativas das proporções desconhecidas (p_1 e p_2) das populações, vamos utilizar

$$Z = \frac{(p_1^* - p_2^*) - (p_1 - p_2)}{\sqrt{\frac{p_1^* q_1^*}{n_1} + \frac{p_2^* q_2^*}{n_2}}} \sim \mathcal{N}(0,1)$$

para grandes amostras (na prática $n_1 \geq 30$ e $n_2 \geq 30$).

Exemplo 47 Sabendo que existem dois grupos de individuos X e Y (cada um com 100 indivíduos) portadores de uma doença, aplica-se um antibiótico apenas ao 1^o grupo. De resto, ambos os grupos são tratados em condições idênticas. Constata-se que, nos grupos X e Y 73% e 65% dos indivíduos, respectivamente, se curaram da doença. Teste a hipótese de que o antibiótico não é eficiente para o nível de significância de 0.01.

Vamos considerar p_1 e p_2 as proporções das populações curadas, aplicando-se o antibiótico e não se aplicando o mesmo respectivamente. Devemos então decidir entre as hipóteses:

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 > p_2 \end{cases}$$

o que significa confrontar a inexistência de diferenças entre proporções (antibiótico ineficiente) contra proporção de indivíduos curados no primeiro grupo ser superior à do segundo (antibiótico eficiente).

Utilizamos para o efeito,

$$Z = \frac{(p_1^* - p_2^*) - (p_1 - p_2)}{\sqrt{\frac{0.73 \times 0.27}{100} + \frac{0.65 \times 0.35}{100}}} \dot{\sim} \mathcal{N}(0,1)\,.$$

Uma vez que a região crítica se situa à direita e $\alpha = 0.01$, então

$$RC = [z_{0.99}, +\infty[= [2.326, +\infty[\,.$$

Como $(p_1^* - p_2^*) = (0.73 - 0.65) = 0.08$, a estimativa de Z é dada por

$$Z = \frac{(p_1^* - p_2^*) - (p_1 - p_2)}{\sqrt{\frac{0.73 \times 0.27}{100} + \frac{0.65 \times 0.35}{100}}} = \frac{0.08 - 0}{\sqrt{\frac{0.73 \times 0.27}{100} + \frac{0.65 \times 0.35}{100}}} = 1.228 \in RA$$

não se rejeita H_0, não se podendo concluir que o antibiótico seja eficiente. ■

5.3.5 Testes de Hipóteses para a Variância

Neste caso, para populações Normais, vamos utilizar:

1. se μ é conhecido,

$$X^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_{(n)};$$

2. se μ é desconhecido,

$$X^2 = (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{(n-1)}.$$

Exemplo 48 O peso dos pacotes cheios por uma máquina de empacotamento tem uma distribuição normal com desvio padrão 0.25 kg. Extraíndo uma amostra de 20 pacotes registou-se um desvio padrão de 0.32 kg. Este aumento de variabilidade é significativo ao nível de significância de 5%?

Vamos considerar o teste de hipóteses:

$$\begin{cases} H_0 : \sigma = 0.25 \\ H_1 : \sigma > 0.25 \end{cases}$$

o que significa confrontar a inexistência do aumento de variabilidade contra o aumento da mesma.

Como a população é Normal e μ é desconhecido, utilizamos

$$X^2 = 19\frac{S^2}{\sigma^2} \sim \chi^2_{(19)}.$$

Uma vez que a região critíca se situa à direita e $\alpha = 0.05$, então,

$$RC = \left[\chi^2_{(19);0.95}, +\infty\right[= [30.1435, +\infty[\,.$$

Como $s^2 = 0.32^2 = 0.1024$, a estimativa de X^2 é dada por

$$X^2 = 19\frac{s^2}{\sigma^2} = 19\frac{0.1024}{0.25^2} = 31.13 \in RC,$$

rejeitando-se H_0, isto é, conclui-se que há um aumento significativo na variabilidade do peso dos pacotes. ■

5.3.6 Teste de Hipóteses para a Razão de Duas Variâncias

Neste teste de hipóteses, para duas populações Normais, vamos usar:

$$F = \frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} \sim F_{(n_1-1,n_2-1)}.$$

Exemplo 49 Um professor de estatística tem duas turmas, cujas notas têm um comportamento normal. A turma A tem 16 alunos e a turma B tem 21 alunos. Num exame, embora não tenha havido diferença significativa entre as notas médias, a turma A registou um desvio padrão de 9% e a turma B de 12%. Podemos concluir que a variabilidade da turma B é superior à da turma A ($\alpha = 0.01$)?

Vamos considerar o teste de hipóteses, utilizando os indíces 1 e 2 para as turmas A e B respectivamente:

$$\begin{cases} H_0 : \frac{\sigma_1}{\sigma_2} = 1 \\ H_1 : \frac{\sigma_1}{\sigma_2} < 1 \end{cases}$$

O que significa confrontar a inexistência de diferença de variabilidade entre as notas das duas turmas, contra a variabilidade das notas da turma B ser superior à da turma A.

A estatística teste e a distribuição amostral a utilizar é,

$$F = \frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} \sim F_{(15,20)}.$$

Dado que a região critíca se situa à esquerda e $\alpha = 0.01$, temos

$$f_{(15,20);0.01} = \frac{1}{f_{(20,15);0.99}} = \frac{1}{3.09} = 0.32362.$$

Logo,

$$RC = \left]0, f_{(15,20);0.01}\right] =]0, 0.32362] .$$

Como $s_1 = 0.09$ e $s_2 = 0.12$, então

$$s_1^2 = 0.0081, \;\; s_2^2 = 0.0144 \;\; \text{e} \;\; \frac{s_1^2}{s_2^2} = 0.5625$$

sendo a estimativa de F dada por

$$F^* = \frac{s_1^2}{s_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} = 0.5625 \times 1 = 0.5625 \in RA.$$

Conclui-se pela não rejeição de H_0, isto é, segundo os dados do problema não existe diferença de variabilidade significativa entre as notas das duas turmas. ■

6 REGRESSÃO LINEAR SIMPLES

6.1 Dados Bivariados

Por vezes certos fenómenos em estudo não se descrevem apenas através de uma variável, sendo necessária a observação de duas (ou mais) variáveis para termos uma visão global do problema. Quando tal ocorre, cada unidade estatística pode contribuir com um conjunto de dois valores passando a trabalhar-se com dados bivariados. Exemplos de dados bivariados são: a altura e peso da população portuguesa, o rendimento mensal de um agregado familiar e o respectivo montante de despesas mensais, as horas de estudo de um aluno e notas obtidas nas disciplinas, etc.

6.2 Representação de Dados Bivariados

A informação da população que se pretende estudar aparece sob a forma de pares de valores da amostra, isto é, cada unidade estatística contribui com um conjunto de dois valores. Surge então o problema de como estudar a existência ou não de relações entre essas variáveis observadas.

Como ponto de partida para o estudo da existência (ou não) de relação estatística (correlação) entre duas variáveis ou características de uma amostra podemos representá-las graficamente através de um **Diagrama de Dispersão** ou **Nuvem de Pontos.** Esta representação gráfica para os dados bivariados consiste em marcarmos os valores das observações realizadas, x_i e y_i, num sistema de eixos cartesianos e obtermos os pontos correspondentes aos pares ordenados (x_i, y_i).

Exemplo 50 Considerando as idades de 16 conjuges na data dos seus casamentos representadas na tabela seguinte (em que X representa a idade do marido e Y a idade da mulher):

X	18	20	21	21	22	23	23	23	24	25	25	26	26	26	28	28
Y	17	20	20	22	22	21	22	23	23	24	25	23	24	27	26	27

Estes dados podem representar-se no Diagrama de Dispersão ou Nuvem de Pontos da Figura 40. Este diagrama, de forma intuitiva, sugere-nos a existência de uma relação linear entre as duas variáveis em estudo, isto é, uma relação que se pode traduzir geometricamente através de uma recta.

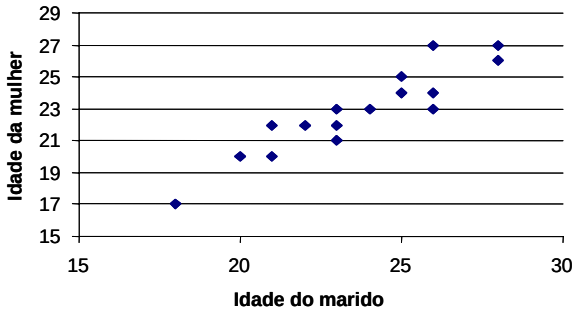


Figura 40: Diagrama de Dispersão ou Nuvem de Pontos. ■

Através da simples observação do diagrama de dispersão ou nuvem de pontos podemos concluir acerca da existência ou não de correlação linear entre duas variáveis X e Y.

Exemplo 51 Os gráficos das Figuras 41 e 42 ilustram vários tipos de correlações lineares entre duas variáveis.

Embora o Diagrama de Dispersão seja um método simples de detecção de relação linear é, no entanto, insuficiente para quantificar a correlação, assim como, quando há observações que se repetem, o diagrama não realça a sua frequência.

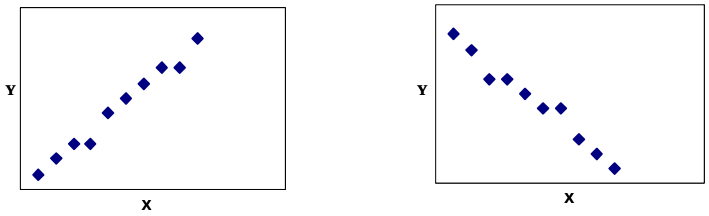


Figura 41: Correlação Linear Positiva (forte) à esquerda e Negativa (forte) à direita.

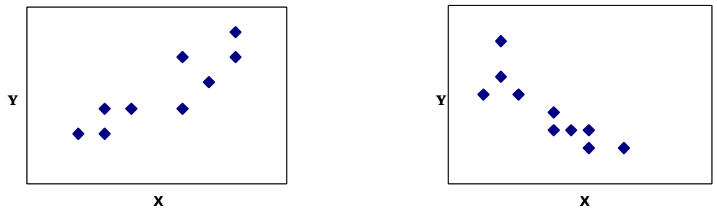


Figura 42: Correlação Linear Positiva (fraca) à esquerda e Negativa (fraca) à direita. ■

6.3 Coeficiente de Correlação Linear Empírico

O **Coeficiente de Correlação Linear Empírico** (ou Amostral), r_{XY}, mede o grau de associação linear entre dados bivaridos, sendo calculado através da expressão:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

em que

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i - \frac{n}{n-1}\bar{x}\bar{y}$$

se denomina de **covariância amostral**, sendo uma medida de variabilidade conjunta entre as variáveis X e Y; S_X e S_Y são os desvios padrões amostrais de X e Y respectivamente.

Nota 42 A covariância amostral e o coeficiente de correlação linear empírico são estatísticas respectivamente da covariância e do coeficiente de correlação linear da população. ■

Deste modo podemos reescrever o coeficiente de correlação linear empírico como:

$$r_{XY} = \frac{\text{covariância}_{XY}}{\sqrt{\text{variância}_X \times \text{variância}_Y}}$$

O coeficiente de correlação linear empírico é um número do intervalo $[-1, 1]$. O sinal do mesmo indica se uma variável aumenta à medida que a outra também aumenta ($r_{XY} > 0$) ou se uma

variável aumenta à medida que a outra diminui ($r_{XY} < 0$). A magnitude indica a proximidade dos pontos em relação a uma linha recta, isto é, quanto mais próximo r_{XY} estiver dos extremos do intervalo $[-1, 1]$, maior é o grau de associação linear; em particular se $r_{XY} = \pm 1$ existe uma correlação linear perfeita estando todos os pontos situados na recta; se $r_{XY} = 0$ a correlação linear é nula (embora possa existir uma relação não linear entre X e Y).

O valor de r_{XY} só é válido dentro da amplitude de valores x e y da amostra. Não se pode inferir que este coeficiente terá o mesmo valor quando se consideram valores de x e y mais extremos do que os constantes na amostra.

É possível trocar a variável dependente e independente sem alterar o valor de r_{XY}.

A existência de um "bom"[5] coeficiente de correlação linear empírico entre X e Y, por si só, não implica necessariamente uma relação de "causa e efeito". Como tal, este coeficiente deve ser sempre acompanhado pelo diagrama de dispersão. Na Figura 43 temos exemplos de situações em que r_{XY} tem um valor próximo dos extremos do intervalo $[-1, 1]$ e, no entanto, não são adequados os modelos lineares; conclui-se deste modo que o simples cálculo r_{XY} é, por vezes, insuficiente.

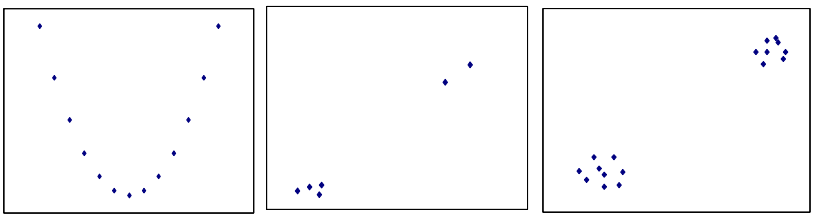


Figura 43: À esquerda temos uma relação quadrática; ao centro temos observações isoladas; à direita temos dados que compreendem subgrupos.

O exemplo seguinte ilustra uma situação deste género, com um caso concreto.

Exemplo 52 Considere o conjunto de observações da tabela

X	1	1.5	1.6	8	8.25	1.9	9.1	8.9	2	8.75	1	8.1	8.5	1.5
Y	3	3.75	3	10.5	11.5	2.6	11	11.5	3.1	10	2.5	10	10.75	2.35

Vamos verificar que o simples cálculo do coeficiente de correlação linear empírico é insuficiente para concluir se existe associação linear entre X e Y.

$$r_{XY} = \frac{\frac{1}{131}\sum_{i=1}^{13}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{13}\sum_{i=1}^{13}(x_i - \bar{x})^2 \times \frac{1}{13}\sum_{i=1}^{13}(y_i - \bar{y})^2}} = 0.989.$$

Pela simples leitura de r_{XY} seríamos levados a concluir que existiria uma boa associação linear entre X e Y. No entanto tal é falso como podemos verificar pelo diagrama de dispersão da Figura 44, onde é nitida a existência de dois subgrupos nas observações em análise.

[5] Vamos considerar como "bom" um coeficiente de correlação linear empírico que se situe no intervalo $[-1, -0.8] \cup [0.8, 1]$. Este intervalo, no entanto, depende dos objectivos e dos dados da pesquisa; como tal, deve ser entendido como um intervalo indicativo e não fixo.

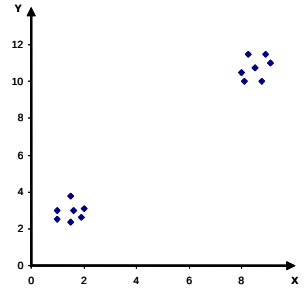


Figura 44: Observações com subgrupos. ■

6.4 Recta de Regressão

Tem-se por objectivo a construção de um modelo matemático que expresse a relação de tipo linear existente entre duas variáveis, com base nos correspondentes valores amostrais.

Considera-se, em geral, X a **variável independente** (explicativa ou explanatória) e Y a **variável dependente** (explicada ou resposta). O modelo matemático que relaciona as duas variáveis permite efectuar previsões para Y.

A recta de regressão pode calcular-se quando no

. Diagrama de Dispersão se averiguar a existência de uma relação linear entre as variáveis

e no

. Coeficiente de Correlação Linear Empírico se obtiver um valor considerado "bom".

Quando se verifica uma forte correlação linear entre as variáveis sob observação podemos descrever a relação entre X e Y, traçando na nuvem de pontos uma recta que seja (segundo algum critério) a que melhor se ajusta aos dados.

Um dos métodos mais conhecidos de ajustar uma recta a um conjunto de dados, é o Método dos Mínimos Quadrados (MMQ), que consiste em determinar a recta que minimiza a soma dos quadrados das distâncias verticais entre os valores observados e a recta (denominadas por erros ou resíduos)

$$e_i^2 = (y_i - \hat{y}_i)^2$$

tal como é ilustrado na Figura 45. O modelo matemático que expressa a relação linear de X sobre

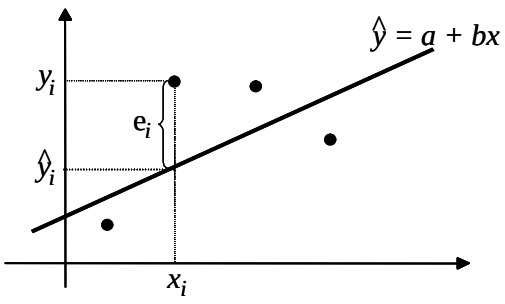


Figura 45: Ajustamento da recta de regressão.

Y é a recta de regressão

$$\hat{y} = a + bx$$

obtida de tal modo que os desvios ou resíduos quadráticos das observações em relação à recta sejam mínimos,

$$\min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} \left[y_i - (a + bx_i) \right]^2 .$$

Como tal, é necessário calcular os pontos de estacionariedade através das primeiras derivadas:

$$\left\{ \begin{array}{l} \frac{\partial}{\partial a} \sum\limits_{i=1}^{n} (y_i - a - bx_i)^2 = 0 \\ \frac{\partial}{\partial b} \sum\limits_{i=1}^{n} (y_i - a - bx_i)^2 = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} -2 \sum\limits_{i=1}^{n} (y_i - a - bx_i) = 0 \\ -2 \sum\limits_{i=1}^{n} (y_i - a - bx_i)\, x_i = 0 \end{array} \right. \Leftrightarrow$$

$$\Leftrightarrow \left\{ \begin{array}{l} \sum\limits_{i=1}^{n} y_i - na - b \sum\limits_{i=1}^{n} y_i = 0 \\ \sum\limits_{i=1}^{n} x_i y_i - a \sum\limits_{i=1}^{n} x_i - b \sum\limits_{i=1}^{n} x_i^2 = 0 \end{array} \right. \Leftrightarrow$$

$$\Leftrightarrow \left\{ \begin{array}{l} a = \frac{\sum\limits_{i=1}^{n} y_i}{n} - b \frac{\sum\limits_{i=1}^{n} x_i}{n} \\ \sum\limits_{i=1}^{n} x_i y_i - \left(\frac{\sum\limits_{i=1}^{n} y_i}{n} - \frac{b \sum\limits_{i=1}^{n} x_i}{n} \right) \sum\limits_{i=1}^{n} x_i - b \sum\limits_{i=1}^{n} x_i^2 = 0 \end{array} \right. \Leftrightarrow$$

$$\Leftrightarrow \left\{ \begin{array}{l} a = \bar{y} - b\bar{x} \\ n \sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} y_i \sum\limits_{i=1}^{n} x_i + b \left(\sum\limits_{i=1}^{n} x_i \right)^2 - nb \sum\limits_{i=1}^{n} x_i^2 = 0 \end{array} \right. \Leftrightarrow$$

$$\Leftrightarrow \left\{ \begin{array}{l} a = \bar{y} - b\bar{x} \\ b \left(n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i \right)^2 \right) = n \sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} y_i \sum\limits_{i=1}^{n} x_i \end{array} \right. \Leftrightarrow$$

$$\Leftrightarrow \left\{ \begin{array}{l} a = \bar{y} - b\bar{x} \\ b = \frac{n \sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n \sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i \right)^2} \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} a = \bar{y} - b\bar{x} \\ b = \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} a = \bar{y} - b\bar{x} \\ b = \frac{S_{XY}}{S_X^2} \end{array} \right.$$

Com base nas segundas derivadas obtem-se a matriz hessiana,

$$H = \left[\begin{array}{cc} \frac{\partial^2}{\partial a^2} \sum\limits_{i=1}^{n} (y_i - a - bx_i)^2 & \frac{\partial^2}{\partial a \partial b} \sum\limits_{i=1}^{n} (y_i - a - bx_i)^2 \\ \frac{\partial^2}{\partial b \partial a} \sum\limits_{i=1}^{n} (y_i - a - bx_i)^2 & \frac{\partial^2}{\partial b^2} \sum\limits_{i=1}^{n} (y_i - a - bx_i)^2 \end{array} \right] =$$

$$= \left[\begin{array}{cc} n & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^2 \end{array} \right]$$

que tem uma forma quadrática definida positiva[6], isto é, os pontos de estacionaridade obtidos, a (ordenada na origem) e b (declive da recta), conduzem a desvios quadráticos mínimos.

6.5 Análise Elementar de Resíduos

Uma das formas de verificar se o modelo linear ajustado é adequado, é através da análise dos resíduos.

6.5.1 Diagrama de Dispersão dos Resíduos

Uma forma simples de visualizar os resíduos (e_i) é através de um diagrama de dispersão, representando os pontos (x_i, e_i). Num modelo bem ajustado os resíduos não podem ser "muito grandes" e devem apresentar-se de forma aleatória sem nenhum padrão particular definido.

Exemplos de resíduos com padrões típicos de ajustamentos inadequados são ilustrados na Figura 46.

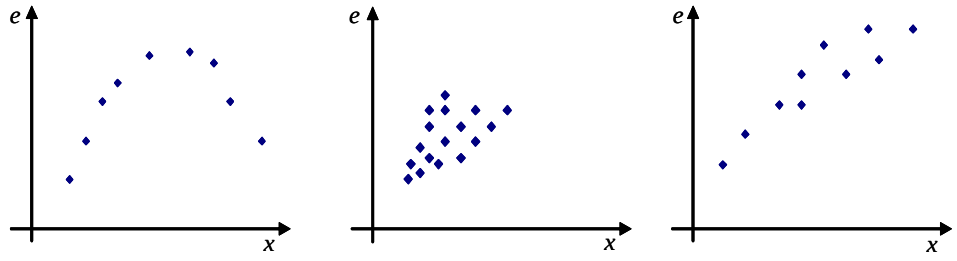


Figura 46: Diagramas de dispersão de resíduos.

Exemplo 53 Admita-se que X e Y representam, respectivamente, a altura e o peso de 12 estudantes seleccionados ao acaso entre os alunos de uma escola estando os dados representados na

[6] Esta hesseana é definida positiva pois,
$m_1 = |n| = n > 0$ e
$$m_2 = \begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix} = n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 > 0.$$

tabela seguinte,

Altura (cm)	Peso (kg)
155	70
150	63
180	72
135	60
156	66
168	70
178	74
160	65
132	62
145	67
139	67
152	68

Vamos começar por analisar estas duas variáveis através do diagrama de dispersão da Figura 47 e do coeficiente de correlação linear empírico.

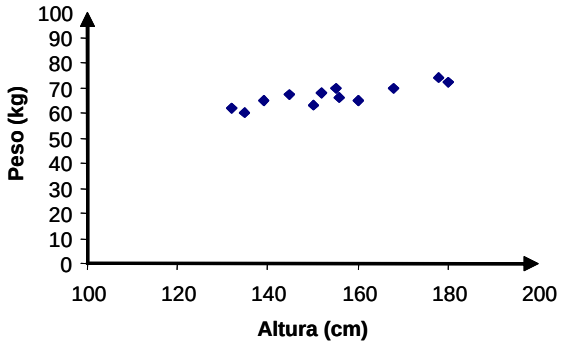


Figura 47: Diagrama de dispersão das alturas e pesos.

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{11}\sum_{i=1}^{12} x_i y_i - \frac{12}{11}\bar{x}\bar{y}}{\sqrt{\frac{1}{11}\sum_{i=1}^{12} x_i^2 - \frac{12}{11}\bar{x}^2}\sqrt{\frac{1}{11}\sum_{i=1}^{12} y_i^2 - \frac{12}{11}\bar{y}^2}} = 0.863.$$

Conclui-se que, tanto através do diagrama de dispersão como do coeficiente de correlação linear empírico, é favorável o ajustamento de uma recta de regressão linear. Vamos então proceder ao seu

cálculo com base nos valores da tabela seguinte:

	x	x^2	y	y^2	xy
	155	24025	70	4900	11550
	150	22500	63	3969	9450
	180	32400	72	5184	12960
	135	18225	60	3600	8100
	156	24336	66	4356	102960
	168	28224	70	4900	11760
	178	31684	74	5476	13172
	160	26500	65	4225	10400
	132	17424	62	3844	8184
	145	21025	67	4489	9715
	139	19321	65	4225	9035
	152	23104	68	4624	10336
Total	1850	287868	802	53792	124258

$$b = \frac{s_{xy}}{s_x^2} = \frac{12\sum_{i=1}^{12} x_i y_i - \sum_{i=1}^{12} x_i \sum_{i=1}^{12} y_i}{12\sum_{i=1}^{12} x_i^2 - \left(\sum_{i=1}^{12} x_i\right)^2} =$$

$$= \frac{12 \times 124258 - 1850 \times 802}{12 \times 287868 - (1850)^2} = 0.231733$$

$$a = \bar{y} - b\bar{x} = \frac{802}{12} - 0.231733 \times \frac{1850}{12} = 31.10778$$

Logo, a recta de regressão é

$$\hat{y} = 31.10778 + 0.231733x.$$

Graficamente, na Figura 48, está ajustada a recta de regressão à nuvem de pontos: A análise da

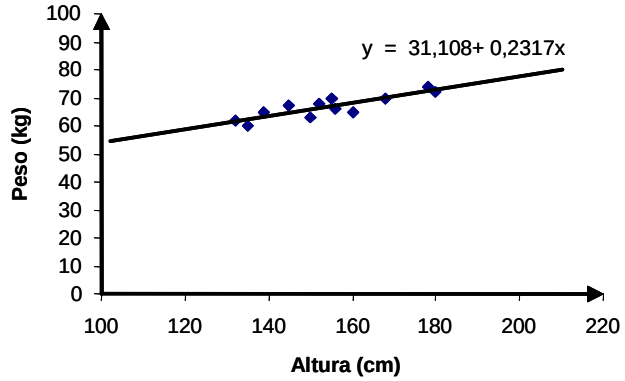


Figura 48: Ajustamento da recta de regressão à nuvem de pontos.

qualidade do ajustamento pode ainda fazer-se através da análise de resíduos. Procedendo ao cálculo

dos mesmos temos:

	x	y	$\hat{y}$	Resíduos ($e = y - \hat{y}$)
	155	70	67.03	2.97
	150	63	65.87	−2.87
	180	72	72.82	−0.82
	135	60	62.39	−2.39
	156	66	67.26	−1.26
	168	70	70.04	−0.04
	178	74	72.36	1.64
	160	65	68.19	−3.19
	132	62	61.70	0.30
	145	67	64.71	2.29
	139	65	63.32	1.68
	152	68	66.33	1.67
Total	1850	802	802.00	0.00

Podemos representar estes desvios graficamente através do diagrama de dispersão dos resíduos representado na Figura 49. Este diagrama tem desvios pequenos (inferiores a 4 kgs) e exibe um

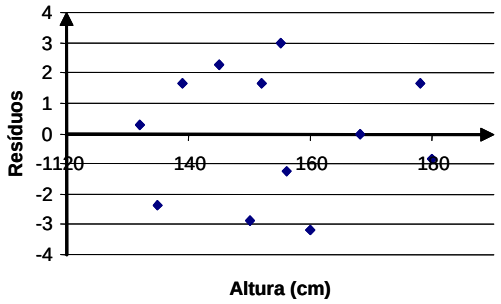


Figura 49: Diagrama de dispersão dos resíduos.

padrão aleatório, concluindo-se que o modelo é adequado aos dados. ■

6.6 Outliers

Designa-se por *outlier* uma observação que se destaca das restantes. Os *outliers* podem existir devido a erros de recolha ou registo de dados ou pelo simples facto dos dados em análise possuirem observações com comportamentos distintos em relação às restantes. Observações deste tipo podem, de uma forma sumária, dividir-se em duas classes:

- *outliers* não influentes, em que a sua existência não altera o modelo linear ajustado;
- *outliers* influentes, em que a sua existência altera o modelo linear ajustado. Este tipo de outliers deve ser examinado e omitir-se quando se conclui que decorre de um erro; caso contrário deve ser estudado cuidadosamente.

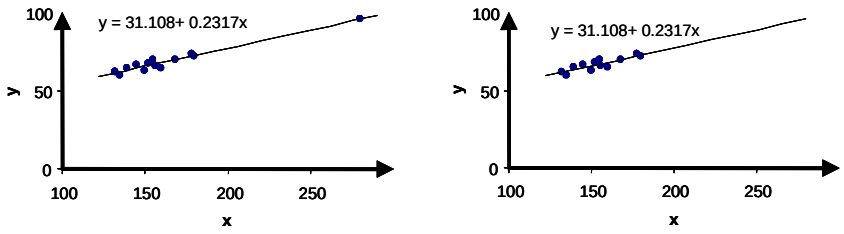


Figura 50: Diagrama de dispersão com outlier (esquerda) e sem outlier (direita).

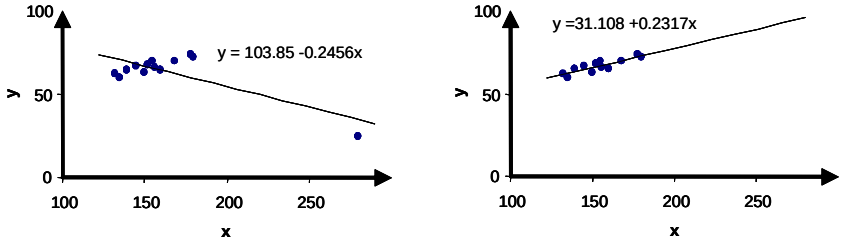


Figura 51: Diagrama de dispersão com outlier (esquerda) e sem outlier (direita).

Exemplo 54 No primeiro diagrama de dispersão da Figura **??** estamos perante um outlier não influente, pois o facto de este ser considerado, ou não, não altera o modelo linear ajustado. No primeiro diagrama de dispersão da Figura 51 estamos perante um outlier influente, pois o facto de este ser considerado, ou não, altera completamente o modelo linear ajustado. ■

Referências

[1] FISZ, M., *Probability Theory and Mathematical Statistics*, Jonh Wiley & Sons, Inc., New York, 1963.

[2] GUIMARÃES, R.C. e CABRAL, J.A.S., *Estatística*, McGraw-Hill de Portugal, Lisboa, 1997.

[3] MURTEIRA, B.J.F., *Probabilidades e Estatística*, McGraw-Hill de Portugal, Lisboa, 1979.

[4] SPIEGEL, M.R., *Probabilidade e Estatística*, Colecção Schaum, McGraw-Hill do Brasil, São Paulo, 1978

[5] OLIVEIRA, J.T., *Probabilidades e Estatística*, vol. I, Escolar Editora, Lisboa,1967.

[6] MELLO, F., *Introdução aos Métodos Estatísticos*, vol. I e II, Cadernos do Instituto de Orientação Profissional, Lisboa, 1973