

## Capítulo 7 - Estimação por intervalos

### 7.1 Intervalos de confiança

Para além duma estimativa pontual de um parâmetro é, em muitas situações, importante dispôr de alguma forma de intervalo que indique a confiança que se pode depositar na estimativa pontual.

Um intervalo de confiança (I.C.) para um parâmetro desconhecido  $\theta$  é do tipo

$$l \leq \theta \leq u,$$

onde  $l$  e  $u$  dependem do valor observado  $\hat{\theta}$  (estimativa pontual), e da distribuição por amostragem da estatística  $\hat{\Theta}$ , usada para estimar  $\theta$ .

Nota: o I. C.  $l \leq \theta \leq u$  é chamado bilateral. Em alguns casos pode ser mais importante um intervalo unilateral (superior ou inferior) do tipo

$$\begin{aligned} l \leq \theta, & \quad \text{intervalo inferior} \\ \theta \leq u & \quad \text{intervalo superior} \end{aligned}$$

### 7.2 Intervalo de confiança para a média, variância conhecida

$X$  população tal que:

$$E(X) = \mu \quad (\text{desconhecido})$$

$$V(X) = \sigma^2 \quad (\text{conhecido})$$

$(X_1, \dots, X_n)$  a. a. de dimensão  $n$

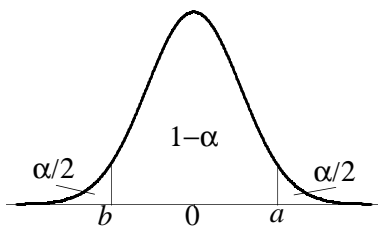
$$\hat{\mu} = \bar{X} \quad (\text{estimador pontual de } \mu)$$

Sabemos já que

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \begin{cases} \sim N(0,1) \text{ se } X \sim N(\mu, \sigma^2) \\ \sim_a N(0,1) \text{ se } X \text{ qq e } n \text{ grande} \end{cases}$$

onde o caso  $X$  qq é justificado pelo T. L. C.

Distribuição de  $Z$ :



Podemos determinar pontos,  $a$  e  $b$ , tais que

$$P\left(b \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq a\right) = 1 - \alpha = \gamma$$

onde  $\alpha \approx 0$  e  $\gamma \approx 1$ .

Para minimizar  $a - b$  deve ter-se  $b = -a$ .

$$\begin{aligned} P\left(-a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq a\right) &= \\ = P\left(-a \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq a \frac{\sigma}{\sqrt{n}}\right) &= \\ = P\left(-\bar{X} - a \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + a \frac{\sigma}{\sqrt{n}}\right) &= \\ = P\left(\bar{X} - a \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + a \frac{\sigma}{\sqrt{n}}\right) &= \\ = P\left(\mu \in \left[\bar{X} - a \frac{\sigma}{\sqrt{n}}; \bar{X} + a \frac{\sigma}{\sqrt{n}}\right]\right) &= 1 - \alpha \end{aligned}$$

com  $a: P(Z > a) = \frac{\alpha}{2}$ .

Nota: a  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  chama-se variável aleatória fulcral (deve ter distribuição conhecida e depender apenas do parâmetro desconhecido).

$\left[ \bar{X} - a \frac{\sigma}{\sqrt{n}}; \bar{X} + a \frac{\sigma}{\sqrt{n}} \right]$  é um intervalo aleatório, quando substituímos  $\bar{X}$  por  $\bar{x}$  (valor observado da média duma amostra aleatória) passamos a ter um intervalo concreto chamado intervalo de confiança.

**Definição** (Intervalo de confiança para a média com variância conhecida): Se  $\bar{x}$  for a média observada duma amostra aleatória de dimensão  $n$  duma população normal (ou duma população qualquer desde que  $n$  grande, mas nesse caso o intervalo é apenas aproximado) com variância conhecida  $\sigma^2$ , um intervalo de confiança a  $100 \times (1 - \alpha)\%$  para  $\mu$  é dado por

$$\bar{x} - a \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + a \frac{\sigma}{\sqrt{n}} \quad \text{com } a: P(Z > a) = \frac{\alpha}{2}$$

(ou  $a = z_{\alpha/2}$ , na notação do Montgomery).

5

### Interpretação:

- Não podemos dizer que  $\mu$  pertence ao intervalo de confiança com probabilidade  $1 - \alpha$ ;
- o que podemos dizer é que se fizermos um grande número de intervalos nestas condições, aproximadamente  $100 \times (1 - \alpha)\%$  desses intervalos conterão de facto o verdadeiro valor de  $\mu$  (que permanece desconhecido), é esta ideia que é traduzida por "confiança".

### Notas:

- 1) O comprimento do intervalo de confiança está associado à precisão, quanto menor for o comprimento mais precisa é a média.

6

2) Se diminuimos  $\alpha$ , isto é, aumentarmos  $1 - \alpha$  (grau de confiança), mantendo  $n$  fixo,  $a$  vai aumentar e consequentemente o comprimento do intervalo. Não é possível fazer  $\alpha = 0$  pois nesse caso  $a = +\infty$ .

3) Quando aumentamos  $n$ , mantendo  $\alpha$  fixo, diminui o comprimento do intervalo.

4) Qual deve ser a dimensão da amostra para que se tenha um dado erro máximo (com confiança fixa)?

Para termos uma confiança de  $100 \times (1 - \alpha)\%$  em que  $|\bar{x} - \mu| \leq E$  deve ter-se

$$n = \left( \frac{a\sigma}{E} \right)^2, \quad (\text{visto que } \frac{a\sigma}{\sqrt{n}} = E).$$

(quando  $n$  não der inteiro arredonda-se por excesso).

7

**Exemplo:**  $X \sim N(\mu, \sigma^2)$  com  $\sigma = 1$ .

a) uma amostra aleatória de dimensão 10 conduziu a  $\bar{x} = 10.1$ . Calcular um intervalo de confiança a 95% para  $\mu$ :

$$\begin{aligned} \alpha = 0.05 &\Leftrightarrow P(Z > a) = 0.025 \Leftrightarrow \\ &\Leftrightarrow P(Z < a) = 0.975 \Leftrightarrow a = \Phi^{-1}(0.975) = 1.96 \end{aligned}$$

Logo o intervalo pedido é

$$\left[ 10.1 - 1.96 \times \frac{1}{\sqrt{10}}; 10.1 + 1.96 \times \frac{1}{\sqrt{10}} \right] = [9.48; 10.72]$$

b) Qual deve ser  $n$  para que se possa garantir com 95% de confiança que  $|\bar{x} - \mu| \leq 0.25$ ?

$$n = \left( \frac{1.96 \times 1}{0.25} \right)^2 = 61.47 \quad \text{Resposta: } n = 62.$$

8

### 7.3 Intervalo de confiança para a diferença de duas médias, variâncias conhecidas

$X_1$ , população 1, com  $E(X_1) = \mu_1$  e  $V(X_1) = \sigma_1^2$  (conhecida)

$X_2$ , população 2, com  $E(X_2) = \mu_2$  e  $V(X_2) = \sigma_2^2$  (conhecida)

( $X_1$  e  $X_2$  independentes)

a. a. da população 1 ( $X_{11}, X_{12}, \dots, X_{1n_1}$ ) com média

$$\bar{X}_1 = \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1}$$

a. a. da população 2 ( $X_{21}, X_{22}, \dots, X_{2n_2}$ ) com média

$$\bar{X}_2 = \frac{\sum_{i=1}^{n_2} X_{2i}}{n_2}$$

(e a a.a.  $(X_{11}, X_{12}, \dots, X_{1n_1})$  é independente da a.a.  $(X_{21}, X_{22}, \dots, X_{2n_2})$ )

Como construir um I.C. a  $100 \times (1 - \alpha)\%$  para  $\mu_1 - \mu_2$  ?

**Nota:** um intervalo deste tipo é útil para comparar duas experiências ou dois métodos.

O estimador pontual de  $\mu_1 - \mu_2$  é  $\bar{X}_1 - \bar{X}_2$ .

Já sabemos que se  $X_1 \sim N(\mu_1, \sigma_1^2)$  e se  $X_2 \sim N(\mu_2, \sigma_2^2)$  então

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \text{ e } \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right) \text{ pelo que}$$

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

logo a v.a. fulcral é:

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

**Notas:** a distribuição anterior também é válida, aproximadamente, para  $X_1$  e  $X_2$  com outra qualquer distribuição, não necessariamente normal, desde que  $n_1$  e  $n_2$  sejam elevados (pelo T.L.C.)

Fazendo, do mesmo modo que na secção anterior,

$$P(-a \leq Z \leq a) = 1 - \alpha, \text{ com } P(Z > a) = \frac{\alpha}{2}$$

obtem-se um I.C. a  $100 \times (1 - \alpha)\%$  para  $\mu_1 - \mu_2$ :

$$\left[ \bar{x}_1 - \bar{x}_2 - a \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}; \bar{x}_1 - \bar{x}_2 + a \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

**Notas:** Mantêm-se as notas 1) 2) e 3) da secção 7.2, relativas à interpretação e variação do comprimento do intervalo.

4) Dimensão da amostra tal que

$$\left| (\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2) \right| \leq E \text{ com } 100 \times (1 - \alpha)\% \text{ de confiança ?}$$

Não tem solução única para  $n_1$  e  $n_2$  gerais, mas se quisermos  $n_1 = n_2 = n$  obtém-se

$$n = \left( \frac{a}{E} \right)^2 (\sigma_1^2 + \sigma_2^2)$$

### 7.4 Intervalo de confiança para a média de uma população normal, variância desconhecida

X população tal que:

$$E(X) = \mu \quad (\text{desconhecido})$$

$$V(X) = \sigma^2 \quad (\text{desconhecida})$$

$(X_1, \dots, X_n)$  a. a. de dimensão  $n$

$$\hat{\mu} = \bar{X} \quad (\text{estimador pontual de } \mu)$$

Não se pode usar  $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$  porque  $\sigma$  é desconhecido. Um procedimento lógico consiste em substituir  $\sigma$  por  $S$  (desvio padrão amostral), ou seja, em considerar a v. a. fulcral  $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ .

Mas qual será o efeito de fazer isto? Não é a mesma v. a.! Qual a sua distribuição?

Se  $n$  for grande ( $n > 30$ , em geral) pode mostrar-se que o efeito é pequeno e tem-se

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

quer para  $X \sim N(\mu, \sigma^2)$ , quer para  $X$  qualquer (com  $E(X) = \mu$  e  $V(X) = \sigma^2$ ). Ou seja, o I.C. calcula-se exactamente como na Secção 7.2 substituindo  $\sigma$  por  $s$  (desvio padrão amostral observado).

Se  $n \leq 30$  o problema não é solúvel no caso geral (isto é, desconhecendo o tipo de distribuição da população). Se  $X \sim N(\mu, \sigma^2)$  o teorema seguinte fornece o resultado que se pretende.

**Teorema:** Seja  $(X_1, \dots, X_n)$  uma a.a. duma população  $X \sim N(\mu, \sigma^2)$ . A variável aleatória

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

tem distribuição  $t$  com  $n - 1$  graus de liberdade ( $T \sim t_{n-1}$ ).

**Notas:**

1) Uma v.a. com distribuição  $t$  tem função de densidade de probabilidade

$$f(x) = \frac{\Gamma[(k+1)/2]}{\sqrt{\pi k} \Gamma(k/2)} \cdot \frac{1}{\left(1 + \frac{x^2}{k}\right)^{\frac{k+1}{2}}}$$

onde  $k > 0$  é o número de graus de liberdade. Pode mostrar-se que se  $T \sim t_k$  então

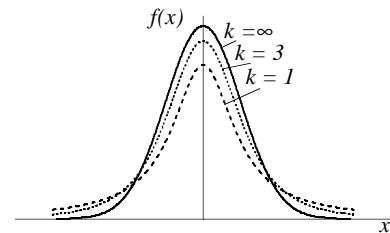
$$E(T) = 0 \quad (k > 1) \text{ e } V(T) = k/(k-2) \quad (k > 2)$$

### 2) Função Gama

$$\Gamma(r) = \int_0^{+\infty} x^{r-1} e^{-x} dx, \quad r > 0$$

se  $r$  inteiro  $\Gamma(r) = (r-1)!$ . Esta função pode ser vista como uma generalização do factorial (definido só para inteiros) aos reais positivos.

### 3) Representação gráfica



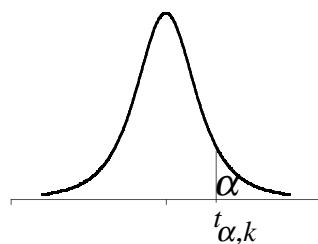
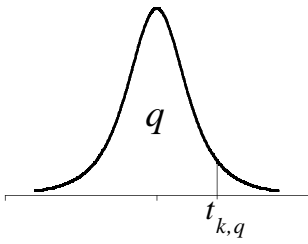
4) É fácil ver que quando  $k \rightarrow +\infty$

$$f(x) \rightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (\text{f.d.p. de } N(0,1))$$

5) Os percentis da distribuição  $t_k$  encontram-se tabelados

Tabelas da disciplina:

No livro:



$$P(T < t_{k,q}) = q$$

$$P(T > t_{\alpha,k}) = \alpha$$

$$t_{k,1-q} = -t_{k,q}$$

$$t_{1-\alpha,k} = -t_{\alpha,k}$$

Voltando à construção do I.C. e procedendo da forma habitual:

$$P\left(-a \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq a\right) = 1 - \alpha \Leftrightarrow$$

$$\Leftrightarrow P\left(\mu \in \left[\bar{X} - a \frac{S}{\sqrt{n}}; \bar{X} + a \frac{S}{\sqrt{n}}\right]\right) = 1 - \alpha$$

com  $a: P(T > a) = \frac{\alpha}{2}$ .

Obtém-se então o I.C. para a média de uma população normal com variância desconhecida

$$\bar{x} - a \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + a \frac{s}{\sqrt{n}}$$

**Observações:**

- 1) A interpretação é semelhante à que foi feita em 7.2.
- 2) Se  $\alpha$  diminuir, com  $n$  fixo, aumenta o comprimento do intervalo.

3) Se  $n$  aumentar, com  $\alpha$  fixo, espera-se que diminua o comprimento do intervalo, mas não há certeza, pois  $s$  varia de amostra para amostra.

4) Determinação de  $n$  para um dado erro (com  $\alpha$  fixo):

$$|\bar{x} - \mu| \leq E \Leftrightarrow n = \left(\frac{as}{E}\right)^2$$

Dificuldades:

i)  $a$  também depende de  $n$ ; Solução: resolução por tentativa-erro.

ii)  $s$  é desconhecido antes de se ter a amostra; Solução: obter uma amostra preliminar para ter uma ideia do valor que  $s$  pode vir a ter.

**Exemplo:** Determinação da constante de acidez do ácido orto-hidroxibenzóico.

a) (Relatório 1) Temos 4 observações da constante de acidez:

$$x_1 = 1.47566 \quad x_2 = 1.48504 \quad x_3 = 1.42970$$

$$x_4 = 1.33962 \quad (\times 10^{-3})$$

$$\bar{x} = 1.432505 \times 10^{-3} \quad n = 4$$

$$s_x = 0.0664775 \times 10^{-3}$$

Admitindo que se trata de observações independentes e identicamente distribuídas de uma mesma variável ( $X$ ) com distribuição  $N(\mu_X, \sigma_X^2)$  tem-se

$$I.C._{.95\%}(\mu_X) = \left[\bar{x} - a \frac{s}{\sqrt{n}}; \bar{x} + a \frac{s}{\sqrt{n}}\right]$$

com  $a = t_{3;0.975} = 3.182$ , logo

$$I.C._{95\%}(\mu_X) = [1.327 \times 10^{-3}; 1.538 \times 10^{-3}]$$

**b) (Relatório 2)** Temos agora 5 observações da constante de acidez:

$$\begin{aligned} x_1 = 8.0623 & \quad x_2 = 8.1398 & \quad x_3 = 7.7421 \\ x_4 = 7.7393 & \quad x_5 = 7.4780 & \quad (\times 10^{-4}) \end{aligned}$$

$$\bar{x} = 7.8323 \times 10^{-4} \quad n = 5$$

$$s_x = 0.269151 \times 10^{-4}$$

Admitindo a mesma hipótese de trabalho e com  $a = t_{4;0.975} = 2.776$ , obtém-se

$$I.C._{95\%}(\mu_X) = [7.498 \times 10^{-4}; 8.166 \times 10^{-4}].$$

**c)** Em vez de se trabalhar com os valores de  $K_a$  também se poderia ter trabalhado com os valores de  $pK_a = -\log_{10} K_a$ , obtendo-se então:

**Relatório 1:** Temos as seguintes 4 observações de  $pK_a$ :

$$\begin{aligned} y_1 = 2.8310 & \quad y_2 = 2.8283 & \quad y_3 = 2.8448 \\ y_4 = 2.8730 & & \end{aligned}$$

$$\bar{y} = 2.844275 \quad n = 4$$

$$s_y = 0.020468$$

donde, admitindo que se trata de observações independentes e identicamente distribuídas de uma mesma variável ( $Y$ ) com distribuição  $N(\mu_Y, \sigma_Y^2)$  se tem

$$I.C._{95\%}(\mu_Y) = [2.8117; 2.8768]$$

**Relatório 2:** Temos agora as seguintes 5 observações de  $pK_a$ :

$$\begin{aligned} y_1 = 3.0935 & \quad y_2 = 3.0894 & \quad y_3 = 3.1111 \\ y_4 = 3.1113 & \quad y_5 = 3.1262 & \end{aligned}$$

$$\bar{y} = 3.1063 \quad n = 5$$

$$s_y = 0.014946$$

donde, admitindo que se trata de observações independentes e identicamente distribuídas de uma mesma variável ( $Y$ ) com distribuição  $N(\mu_Y, \sigma_Y^2)$  se tem

$$I.C._{95\%}(\mu_Y) = [3.0877; 3.1249]$$

**d)** Dado que  $Y = -\log_{10} X \Leftrightarrow X = 10^{-Y}$  e sabemos que  $\mu_X \approx 10^{-\mu_Y}$  (ver Secção 5.7), dos intervalos obtidos na alínea c) também podemos obter intervalos para  $\mu_X$ , aplicando a transformação  $10^{-y}$  a ambos os extremos dos intervalos, vem então

**Relatório 1:**

$$I.C.'_{95\%}(\mu_X) = [1.328 \times 10^{-3}; 1.543 \times 10^{-3}]$$

**Relatório 2:**

$$I.C.'_{95\%}(\mu_Y) = [7.501 \times 10^{-4}; 8.171 \times 10^{-4}]$$

Estes intervalos são parecidos (mas não iguais) aos obtidos nas alíneas a) e b).

Qual é a resolução correcta? A resposta prende-se com uma das hipóteses de trabalho colocadas, isto é, que determinada variável tem distribuição  $N(\mu, \sigma^2)$ . Devemos basear o I.C. na variável ( $K_a$  ou  $pK_a$ , neste caso) em relação à qual melhor podemos garantir a verificação dessa hipótese de trabalho.

### 7.5 Intervalo de confiança para a diferença entre as médias de duas populações normais, variâncias desconhecidas

$X_1$ , população 1, com  $E(X_1) = \mu_1$  e  $V(X_1) = \sigma_1^2$  (desconhecida)

$X_2$ , população 2, com  $E(X_2) = \mu_2$  e  $V(X_2) = \sigma_2^2$  (desconhecida)

( $X_1$  e  $X_2$  independentes)

a. a. da população 1 ( $X_{11}, X_{12}, \dots, X_{1n_1}$ ) com média

$$\bar{X}_1 = \frac{\sum_{i=1}^{n_1} X_{1i}}{n_1} \text{ e variância } S_1^2 = \frac{\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2}{n_1 - 1}$$

a. a. da população 2 ( $X_{21}, X_{22}, \dots, X_{2n_2}$ ) com média

$$\bar{X}_2 = \frac{\sum_{i=1}^{n_2} X_{2i}}{n_2} \text{ e variância } S_2^2 = \frac{\sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2}{n_2 - 1}$$

25

(e a a.a.  $(X_{11}, X_{12}, \dots, X_{1n_1})$  é independente da a.a.  $(X_{21}, X_{22}, \dots, X_{2n_2})$ )

Como construir um I.C. a  $100 \times (1 - \alpha)\%$  para  $\mu_1 - \mu_2$  ?

Quando  $\sigma_1^2$  e  $\sigma_2^2$  eram conhecidas usava-se a v.a. fulcral

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

Se  $n_1 > 30$  e  $n_2 > 30$  pode-se substituir  $\sigma_1^2$  por  $S_1^2$  e  $\sigma_2^2$  por  $S_2^2$  obtendo-se

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

26

quer para  $X_1 \sim N(\mu_1, \sigma_1^2)$  e  $X_2 \sim N(\mu_2, \sigma_2^2)$  quer para  $X_1$  e  $X_2$  com outra qualquer distribuição.

Então um I.C. calcula-se exactamente como na Secção 7.4, apenas substituindo  $\sigma_1^2$  por  $s_1^2$  e  $\sigma_2^2$  por  $s_2^2$ .

Quando  $n_1 \leq 30$  ou  $n_2 \leq 30$  o problema só tem solução no caso em que  $X_1 \sim N(\mu_1, \sigma_1^2)$  e  $X_2 \sim N(\mu_2, \sigma_2^2)$  e mesmo assim para se obter uma v.a. fulcral com distribuição exacta é necessário supor que (embora desconhecidas)  $\sigma_1^2 = \sigma_2^2 = \sigma^2$  (esta suposição é razoável em muitas situações reais, e além disso pode ser testada).

$\bar{X}_1 - \bar{X}_2$  continua a ser o estimador pontual de  $\mu_1 - \mu_2$

27

e tem-se ainda que

$$V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$$

Vai ser necessário estimar  $\sigma^2$ . Um estimador natural obtém-se combinando as duas variâncias amostrais

$$\hat{\sigma}^2 = S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Note-se que quando  $n_1 = n_2$  resulta  $\hat{\sigma}^2 = \frac{S_1^2 + S_2^2}{2}$ .

Sabemos que

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$$

28

substituindo  $\sigma$  por  $S_p$  obtém-se

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

(sob as suposições, ou hipóteses de trabalho, consideradas). Então de

$$P(-a \leq T \leq a) = 1 - \alpha \quad \text{com } a: P(T > a) = \frac{\alpha}{2}$$

obtém-se um I.C. a  $100 \times (1 - \alpha)\%$  para  $\mu_1 - \mu_2$ :

$$\left[ \bar{x}_1 - \bar{x}_2 - a s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}; \bar{x}_1 - \bar{x}_2 + a s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

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

### Observações:

- 1) A determinação da dimensão da amostra é mais complicada (ver 7.3 e 7.4).
- 2) E se não for razoável admitir que  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ ? Este problema (conhecido por problema de Behrens-Fisher) não tem solução exacta. Há soluções aproximadas, o Montgomery (pág. 397) apresenta uma (não faz parte do programa).

**Exemplo:** Comparação entre os resultados obtidos pelos dois grupos na determinação da constante de acidez do ácido orto-hidroxibenzoico (vamos usar os dados relativos a pKa).

Relatório 1	Relatório 2
$\bar{y}_1 = 2.844275$	$\bar{y}_2 = 3.1063$
$s_{y1} = 0.020468$	$s_{y2} = 0.014946$
$n_1 = 4$	$n_2 = 5$

Admitimos que (hipóteses de trabalho):

- A primeira amostra é uma concretização de uma a.a. de uma população  $Y_1 \sim N(\mu_1, \sigma_1^2)$ ;
- A segunda amostra é uma concretização de uma a.a. de uma população  $Y_2 \sim N(\mu_2, \sigma_2^2)$ ;
- $Y_1$  e  $Y_2$  são independentes;
- $\sigma_1^2 = \sigma_2^2 = \sigma^2$  (parece razoável porque  $s_{y1}$  e  $s_{y2}$  são da mesma ordem de grandeza).

$$s_p = \sqrt{\frac{3 \times 0.020468^2 + 4 \times 0.014946^2}{7}} = 0.017527$$

vamos fixar  $\alpha = 0.01$  (grau de confiança = 99%)

$$a = t_{7;0.995} = 3.499 \quad (n_1 + n_2 - 2 = 7)$$

$$\text{então } I.C._{99\%}(\mu_1 - \mu_2) = [-0.3031; -0.2209].$$

Como este intervalo não contém o ponto  $\mu_1 - \mu_2 = 0$  concluímos que os resultados obtidos nos permitem afirmar (com um grau de confiança de 99%) que  $\mu_1 \neq \mu_2$ .

Se tivéssemos usado um grau de confiança de 95% era tudo idêntico excepto  $a = t_{7;0.975} = 2.365$  e obtinha-se

$$I.C._{95\%}(\mu_1 - \mu_2) = [-0.2898; -0.2342].$$

### 7.6 Intervalo de confiança para a variância de uma população normal

$$X \sim N(\mu, \sigma^2) \quad (X_1, \dots, X_n) \text{ a.a.}$$

Queremos um I.C. a  $100 \times (1 - \alpha)\%$  para  $\sigma^2$ .

O estimador pontual de  $\sigma^2$  é  $S^2$ . A v.a. fulcral obtém-se do teorema seguinte.

**Teorema:** Seja  $(X_1, \dots, X_n)$  uma a.a. duma população  $X \sim N(\mu, \sigma^2)$ . A v.a.

$$Q = \frac{(n-1)S^2}{\sigma^2}$$

tem distribuição do chi-quadrado com  $n - 1$  graus de liberdade ( $Q \sim \chi_{n-1}^2$ ).

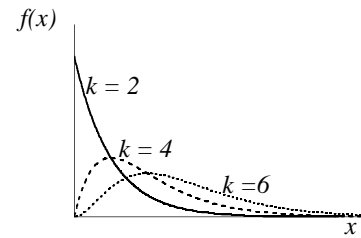
### Notas:

1) Uma v.a. com distribuição do chi-quadrado tem função de densidade de probabilidade

$$f(x) = \frac{1}{2^{k/2} \Gamma(k/2)} x^{(k/2)-1} e^{-x/2}, x > 0$$

onde  $k > 0$  é o número de graus de liberdade. Pode mostrar-se que se  $Q \sim \chi_k^2$  então  $E(Q) = k$  e  $V(Q) = 2k$ .

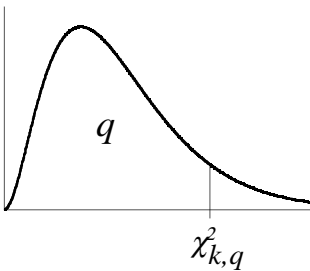
2) Representação gráfica



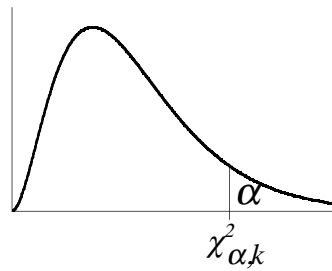
3) Os percentis da distribuição  $\chi_k^2$  encontram-se tabelados

Tabelas da disciplina:

No livro:



$$P(Q < \chi_{k,q}^2) = q$$



$$P(Q > \chi_{\alpha,k}^2) = \alpha$$

Para construir um I.C. a  $100 \times (1 - \alpha)\%$  para  $\sigma^2$  parte-se de

$$P(a \leq Q \leq b) = 1 - \alpha$$

Dado que a distribuição de  $Q$  não é simétrica é preciso determinar  $a$  e  $b$ . Faz-se então

$$a: P(Q < a) = \frac{\alpha}{2} \quad \text{e} \quad b: P(Q > b) = \frac{\alpha}{2}.$$

$$P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right) = 1 - \alpha \Leftrightarrow$$

$$\Leftrightarrow P\left(\sigma^2 \in \left[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right]\right) = 1 - \alpha$$

Um I.C. a  $100 \times (1 - \alpha)\%$  para  $\sigma^2$  é então

$$\left[\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right]$$

**Exemplo:** Vamos determinar intervalos de confiança a 95% ( $\alpha = 0.05$ ) para o desvio padrão da variável pKa considerada atrás.

**Relatório 1:**  $Y_1 \sim N(\mu_1, \sigma_1^2)$

$n_1 = 4 \quad s_{y1} = 0.020468$

$a = \chi_{3;0.025}^2 = 0.216 \quad b = \chi_{3;0.975}^2 = 9.348$

$I.C._{.95\%}(\sigma^2) = [1.344 \times 10^{-4}; 5.819 \times 10^{-3}]$

e

$I.C._{.95\%}(\sigma) = [0.012; 0.076]$

**Relatório 2:**  $Y_2 \sim N(\mu_2, \sigma_2^2)$

$n_2 = 5 \quad s_{y2} = 0.014946$

$a = \chi_{4;0.025}^2 = 0.484 \quad b = \chi_{4;0.975}^2 = 11.14$

$I.C._{.95\%}(\sigma^2) = [8.021 \times 10^{-5}; 1.846 \times 10^{-3}]$

e

$I.C._{.95\%}(\sigma) = [0.009; 0.043]$

**7.7 Intervalo de confiança para uma proporção**

$(X_1, \dots, X_n)$  amostra aleatória de uma população muito grande ou infinita.

Seja  $Y(\leq n)$  o número de observações desta amostra que pertencem a uma dada categoria de interesse.

Seja  $p$  a proporção de indivíduos na população que pertencem a essa categoria de interesse.

**Exemplos:**

População	Categoria
Peças	ser defeituosa
Eleitores	vota no partido X

O estimador pontual de  $p$  é  $\hat{P} = \frac{Y}{n}$ .

Sabemos que  $Y \sim Bin(n, p)$  e que se  $n$  for grande

$Y \underset{a}{\sim} N(np, np(1-p))$

ou ainda  $\hat{P} = \frac{Y}{n} \underset{a}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$  pois

$E(\hat{P}) = E\left(\frac{Y}{n}\right) = \frac{E(Y)}{n}$  e  $V(\hat{P}) = V\left(\frac{Y}{n}\right) = \frac{V(Y)}{n^2}$

Então tem-se  $Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \underset{a}{\sim} N(0,1)$

$P\left(-a \leq \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq a\right) \approx 1 - \alpha, a: P(Z > a) = \frac{\alpha}{2}$

Resolver em ordem a  $p$ ? Aproximação que dá resultados satisfatórios

$P\left(-a \leq \frac{\hat{P} - p}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}} \leq a\right) \approx 1 - \alpha$

O I.C. (aproximado) a  $100 \times (1 - \alpha)\%$  para  $p$  é:

$\left[\hat{p} - a\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + a\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$

**Exemplo:** População de eleitores portugueses. Sondagem (aleatória) a 1200 eleitores revelou que 683 tencionam votar no partido ABC. Determinar a e.m.v. de  $p$  (proporção de eleitores na população que tencionam votar no partido ABC) e um I.C. aproximado a 95% para  $p$ .

$\hat{p} = 683/1200 = 0.569$

$a = 1.96 \quad I.C._{.95\%}(p) = [0.541; 0.597]$