Estatística aplicada
Licenciatura em Engenharia Informática

Lino Costa

Departamento de Produção e Sistemas
Escola de Engenharia
lac@dps.uminho.pt

Revisto em 2022/2023

Sumário

Estimação intervalar

Intervalos de confiança

Estabelecer um intervalo de confiança (IC) para o parâmetro θ da população, i.e., determinar um dos seguintes tipos de intervalo:

- **intervalo bilateral** - limites inferior e superior ($\hat{\theta}_I < \theta < \hat{\theta}_S$) para θ
- **intervalos unilaterais** - limite inferior ou limite superior ($\hat{\theta}_I < \theta$ ou $\theta < \hat{\theta}_S$) para θ

Em geral, os intervalos de confiança bilaterais são mais utilizados. Os limites θ_I e θ_S dependem da distribuição amostral do estimador de θ.

Nível de confiança

Estabelece-se um nível de confiança de $100(1-\alpha)\%$ que indica que, considerando um grande número de amostras aleatórias (infinito), há $100(1-\alpha)\%$ de probabilidade do IC conter o verdadeiro valor de θ, i.e.

$$P(\hat{\theta}_L < \theta < \hat{\theta}_S) = 100(1-\alpha)\%$$

Amplitude e precisão de estimação

Amplitude de um IC (bilateral)

A amplitude de um IC é a distância entre o limite inferior e o limite superior

$$A_{IC} = \hat{\theta}_S - \hat{\theta}_I$$

Quanto maior a amplitude, maior a confiança que o intervalo contenha o verdadeiro valor de θ, mas menos informação dá sobre o verdadeiro valor de θ.

Precisão um IC

A precisão da estimação de θ pelo IC está inversamente relacionada com a amplitude do IC, i.e., quanto maior a amplitude do CI, menor é a precisão.

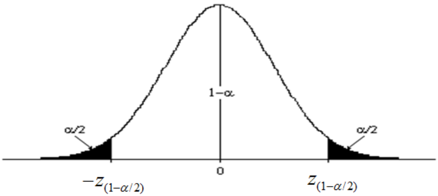
ICs para a média

Se a distribuição da média uma amostra de tamanho n é normal, i.e., $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$, com σ^2 conhecido, então

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

logo

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



ICs para a média

A partir de $P\left(-z_{(1-\frac{\alpha}{2})} < Z < z_{(1-\frac{\alpha}{2})}\right) = 1-\alpha$, como $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$, tem-se que

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

Resolvendo em ordem a μ, obtemos

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

Logo

$$\bar{X} - z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}$$

Nível de confiança

Exemplo 1

Suponha que a altura dos rapazes com 20 anos é normalmente distribuída com média $\mu = 170cm$ e desvio padrão $\sigma = 10cm$. Considere que foram recolhidas 5 amostras de 25 rapazes, tendo sido observadas as seguintes médias amostrais.

Amostra	1	2	3	4	5
$\bar{X}$	172	168	171	165	172

Com base em cada uma das amostras construa o respetivo intervalo de confiança para a média com 95%.

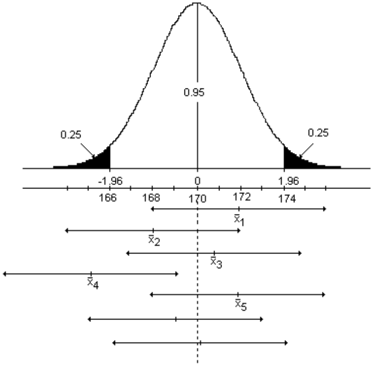
Comente os resultados obtidos em termos do significado do nível de confiança de uma intervalo de confiança.

- $\bar{X} - z_{(1-\frac{\alpha}{2})} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{(1-\frac{\alpha}{2})} \frac{\sigma}{\sqrt{n}}$
- Como $1 - \alpha = 0.95$ então $1 - \frac{\alpha}{2} = 0.975$ e $z_{(1-\frac{\alpha}{2})} = z_{0.975} = 1.96$ (tabela 5)
- $\bar{X} - 1.96 \frac{10}{\sqrt{25}} < \mu < \bar{X} + 1.96 \frac{10}{\sqrt{25}}$
- $\bar{X} \pm 4cm$

Nível de confiança

Exemplo 1

Amostra	1	2	3	4	5
IC para μ	$172 \pm 4cm$	$168 \pm 4cm$	$171 \pm 4cm$	$165 \pm 4cm$	$172 \pm 4cm$



ICs para a média

σ^2 conhecido

Para uma amostra de qualquer dimensão n retirada de uma população normal,

$$\bar{X} - z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}$$

σ^2 desconhecido, com $n \geq 30$

Para uma amostra de grande dimensão ($n \geq 30$) retirada de uma população com uma qualquer distribuição, a aproximação $\sigma \approx s$ é válida, logo

$$\bar{X} - z_{(1-\frac{\alpha}{2})}\frac{s}{\sqrt{n}} < \mu < \bar{X} + z_{(1-\frac{\alpha}{2})}\frac{s}{\sqrt{n}}$$

onde s é o desvio padrão amostral.

ICs para a média

Exemplo 2

O peso ao nascer é uma das variáveis mais importantes na avaliação do bem-estar de um recém-nascido. Suponha que o valor do desvio padrão para os bebés de sexo masculino é conhecido e é de 562 gramas. Num determinado centro de saúde, uma amostra de 19 recém-nascidos apresentou uma média 3222 gramas. Construa um intervalo de confiança de 95% para a média do peso dos bebés.

- IC para μ com $\sigma = 562$ conhecido
- amostra: $n = 19$, $\bar{X} = 3222$
- $\bar{X} - z_{(1-\frac{\alpha}{2})} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{(1-\frac{\alpha}{2})} \frac{\sigma}{\sqrt{n}}$
- Como $1 - \alpha = 0.95$ então $1 - \frac{\alpha}{2} = 0.975$ e $z_{(1-\frac{\alpha}{2})} = z_{0.975} = 1.96$ (tabela 5)
- $3222 - 1.96 \frac{562}{\sqrt{19}} < \mu < 3222 + 1.96 \frac{562}{\sqrt{19}}$
- $3222 - 253 < \mu < 3222 + 253$, i.e., $2969 < \mu < 3475$
- estimativa pontual de μ: $3222g$
- erro da estimativa de μ: $253g$

ICs para a média

σ^2 desconhecido, $n < 30$

Nestas condições, a distribuição de $\bar{X}$ não é normal, passando-se a utilizar a distribuição t-Student.

Distribuição t-Student

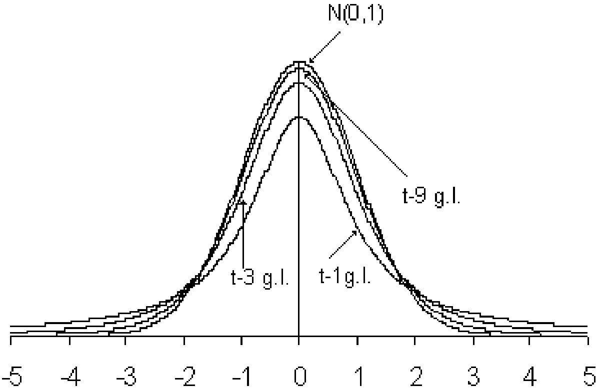
A função densidade de probabilidade da distribuição t-Student com v graus de liberdade é

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

com $E(X) = 0$ e $V(X) = \frac{v}{v-2}$ para $v > 2$. Quando $v \to \infty$, a distribuição t-Student aproxima-se da normal.

Distribuição t-Student

A distribuição t-Student é simétrica em relação a zero, unimodal, em forma de sino, mas com caudas mais pesadas do que as da distribuição normal. Aproxima-se da normal quando $v \to \infty$.



ICs para a média

σ^2 desconhecido, $n < 30$

Se $\bar{X}$ e s são a média e o desvio padrão de uma amostra aleatória de tamanho n de uma população normal com média μ e desvio padrão σ, então $T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ segue uma distribuição t-Student com $n-1$ graus de liberdade, i.e.,

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

A partir de $P\left(-t_{\frac{\alpha}{2},n-1} < T < t_{\frac{\alpha}{2},n-1}\right) = 1 - \alpha$, como $T = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$, tem-se que

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

ICs para a média

Resolvendo em ordem a μ, obtemos

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

Logo

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

onde s é o desvio padrão amostral.

σ^2 desconhecido, $n < 30$

Para uma amostra de pequena dimensão ($n < 30$) retirada de uma população normal,

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

onde s é o desvio padrão amostral.

ICs para a média

Exemplo 3

Numa fábrica, uma amostra de 12 eixos foi retirada da linha de produção. O diâmetro médio encontrado foi de $19.92cm$ com um desvio padrão de $0.17cm$. Construa um intervalo de confiança de 95% para o diâmetro médio dos eixos. Assuma a normalidade do diâmetro dos eixos produzidos na linha de produção.

- IC para μ com σ desconhecido
- amostra: $n = 12 < 30$, $\bar{X} = 19.92$ e $s = 0.17$
- $\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$
- Como $1 - \alpha = 0.95$ então $\frac{\alpha}{2} = 0.025$ e $t_{0.025,11} = 2.201$ (tabela 6)
- $19.92 - 2.201\frac{0.17}{\sqrt{12}} < \mu < 19.92 + 2.201\frac{0.17}{\sqrt{12}}$
- $19.92 - 0.108 < \mu < 19.92 + 0.108$, i.e., $19.812 < \mu < 20.028$
- estimativa pontual de μ: $19.92cm$
- erro da estimativa de μ: $0.108cm$

ICs para a diferença de médias

Se $\bar{X}_1$ e $\bar{X}_2$ são as médias de duas amostras aleatórias independentes de dimensões n_1 e n_2 retiradas de duas populações normais com médias e variâncias (μ_1, σ_1^2) e (μ_2, σ_2^2), respetivamente, então

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

Nota: a distribuição amostral de $\bar{X}_1 - \bar{X}_2$ é normal com média $\mu_1 - \mu_2$ e variância $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$, i.e.,

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

ICs para a diferença de médias

Duas amostras independentes, σ_1^2 e σ_2^2 conhecidos

Para duas amostras independentes de quaisquer dimensões n_1 e n_2 retirada de duas populações normais,

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

Duas amostras independentes, σ_1^2 e σ_2^2 desconhecidos, com $n_1 \geq 30$ e $n_2 \geq 30$

Para duas amostras independentes de grande dimensão ($n_1 \geq 30$ e $n_2 \geq 30$) retiradas de duas populações com quaisquer distribuições, as aproximações $\sigma_1 \approx s_1$ e $\sigma_2 \approx s_2$ são válidas, logo

$$(\bar{X}_1 - \bar{X}_2) - z_{(1-\frac{\alpha}{2})}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{(1-\frac{\alpha}{2})}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

onde s_1 e s_2 são os desvios padrão amostrais.

ICs para a diferença de médias

Exemplo 4

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

Curso	n	$\bar{X}$	s
Engenharia	40	1253	126
Relações Internacionais	30	1047	145

Construa um intervalo de confiança de 90% para a diferença entre os salários médios.

- IC para $\mu_1 - \mu_2$ com σ_1 e σ_2 desconhecidos
- amostras independentes: $n_1 = 40 \geq 30$ e $n_2 = 30 \geq 30$, $\bar{X}_1 = 1253$, $s_1 = 126$, $\bar{X}_2 = 1047$ e $s_2 = 145$
- $(\bar{X}_1 - \bar{X}_2) - z_{(1-\frac{\alpha}{2})}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{(1-\frac{\alpha}{2})}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Como $1 - \alpha = 0.90$ então $\frac{\alpha}{2} = 0.05$ e $z_{0.95} = 1.65$ (tabela 5)
- $(1253 - 1047) - 1.65\sqrt{\frac{126^2}{40} + \frac{145^2}{30}} < \mu_1 - \mu_2 < (1253 - 1047) + 1.65\sqrt{\frac{126^2}{40} + \frac{145^2}{30}}$
- $206 - 54.7 < \mu_1 - \mu_2 < 206 + 54.7$, i.e., $151.3 < \mu_1 - \mu_2 < 260.7$
- Como o IC não contém o valor 0, pode-se concluir que existem diferenças significativas entre os salários médios

ICs para a diferença de médias

Duas amostras independentes, $\sigma_1^2 = \sigma_2^2 = \sigma^2$ desconhecida, $n_1 < 30$ e/ou $n_2 < 30$

Se $\bar{X}_1$ e $\bar{X}_2$ são as médias de duas amostras aleatórias independentes de dimensões n_1 e n_2 retiradas de duas populações normais com médias μ_1 eμ_2, respetivamente, e variância comum $\sigma_1^2 = \sigma_2^2 = \sigma^2$, então

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

sendo s_p^2 uma estimativa da variância comum dada por

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

ICs para a diferença de médias

Duas amostras independentes, $\sigma_1^2 = \sigma_2^2 = \sigma^2$ desconhecida, $n_1 < 30$ e/ou $n_2 < 30$

Para duas amostras independentes de dimensões n_1 e n_2 retiradas de duas populações normais com variância comum,

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

com

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

onde s_1 e s_2 são os desvios padrão amostrais.

ICs para a diferença de médias

Exemplo 4

Pretende-se testar duas rações no crescimento de frangos de aviário. Os frangos, distribuídos por dois pavilhões, foram alimentados durante 5 semanas com a respetiva ração. No final do período, selecionou-se uma amostra aleatória de cada um dos pavilhões, obtendo-se:

	n	$\bar{X}$	s
Ração 1	16	$1623.75g$	$192.71g$
Ração 2	10	$1588.00g$	$167.12g$

Construa um IC de 95% para a diferença entre os pesos médios dos frangos.

- IC para $\mu_1 - \mu_2$ com σ_1 e σ_2 desconhecidos, assumindo normalidade e variância comum
- amostras independentes: $n_1 = 16 < 30$ e $n_2 = 10 < 30$, $\bar{X}_1 = 1623.75$, $s_1 = 192.71$, $\bar{X}_2 = 1588.00$ e $s_2 = 167.12$
- $(\bar{X}_1 - \bar{X}_2) - t_{\frac{\alpha}{2},gl} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{\frac{\alpha}{2},gl} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ onde $gl = n_1 + n_2 - 2$
- Como $1 - \alpha = 0.95$ então $\frac{\alpha}{2} = 0.025$ e $t_{0.025,24} = 2.06$ (tabela 6)
- $(1623.75 - 1588.00) - 2.06 s_p \sqrt{\frac{1}{16} + \frac{1}{10}} < \mu_1 - \mu_2 < (1623.75 - 1588.00) + 2.06 s_p \sqrt{\frac{1}{16} + \frac{1}{10}}$
- $s_p = \sqrt{\frac{(16-1)192.71^2 + (10-1)167.12^2}{16+10-2}} = 183.53$
- $35.75 - 152.41 < \mu_1 - \mu_2 < 35.75 + 152.41$, i.e., $-116.66 < \mu_1 - \mu_2 < 188.16$
- Como o IC contém o valor 0, não se pode concluir que existam diferenças significativas entre os pesos médios dos frangos dos dois pavilhões

ICs para a diferença de médias

$\sigma_1^2 \neq \sigma_2^2$ desconhecidas, $n_1 < 30$ e/ou $n_2 < 30$

Para duas amostras independentes de dimensões n_1 e n_2 retiradas de duas populações normais com variâncias diferentes,

$$(\bar{X}_1 - \bar{X}_2) - t_{\frac{\alpha}{2},gl}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{\frac{\alpha}{2},gl}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

com

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

onde s_1 e s_2 são os desvios padrão amostrais.

ICs para a diferença de médias

Duas amostras emparelhadas, σ_1^2 e σ_2^2 desconhecidas, $n_1 = n_2 = n < 30$

Para duas amostras emparelhadas (relacionadas) de dimensões $n_1 = n_2 = n$ retiradas de duas populações normais com variâncias σ_1^2 e σ_2^2,

$$\bar{D} - t_{\frac{\alpha}{2},n-1}\frac{s_D}{\sqrt{n}} < \mu_1 - \mu_2 < \bar{D} + t_{\frac{\alpha}{2},n-1}\frac{s_D}{\sqrt{n}}$$

sendo $D_i = X_{1i} - X_{2i}$ para $i = 1, \ldots, n$,

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i \quad \text{e} \quad s_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(D_i - \bar{D}\right)^2}$$

ICs para a diferença de médias

Exemplo 5

Uma amostra de 10 trabalhadores de uma fábrica onde existe a manipulação de dioxinas foi selecionada aleatoriamente. Em cada um destes trabalhadores foi determinada a concentração (em ppm) de dioxinas no plasma e no tecido gordo.

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

Construa um intervalo de confiança de 95% para a diferença entre as concentrações médias de dioxina no plasma e no tecido gordo.

- IC para $\mu_1 - \mu_2$ com σ_1 e σ_2 desconhecidos, assumindo normalidade
- amostras emparelhadas (dependentes): $n_1 = n_2 = n = 10 < 30$
-

Trabalhador	1	2	3	4	5	6	7	8	9	10
D_i	-2.4	-3.4	-2.4	0.3	-0.5	1.6	-2.5	0.4	-2.8	0.2

- $\bar{D} = -1.15$, $s_D = 1.734$
- $\bar{D} - t_{\frac{\alpha}{2},n-1} \frac{s_D}{\sqrt{n}} < \mu_1 - \mu_2 < \bar{D} + t_{\frac{\alpha}{2},n-1} \frac{s_D}{\sqrt{n}}$
- Como $1 - \alpha = 0.95$ então $\frac{\alpha}{2} = 0.025$ e $t_{0.025,9} = 2.262$ (tabela 6)
- $-1.15 - 2.262\frac{1.734}{\sqrt{10}} < \mu_1 - \mu_2 < -1.15 + 2.262\frac{1.734}{\sqrt{10}}$
- $-1.15 - 1.240 < \mu_1 - \mu_2 < -1.15 + 1.240$, i.e., $-2.39 < \mu_1 - \mu_2 < 0.89$
- Como o IC contém o valor 0, não se pode concluir que existam diferenças significativas entre as concentrações médias de dioxina no plasma e no tecido gordo

ICs para a proporção

Para amostra de grande dimensão $n \geq 30$, a distribuição da proporção binomial é

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Resolvendo em ordem a p, obtemos

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

Logo,

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

ICs para a proporção

Exemplo 6

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

- IC para p
- amostra de grande dimensão: $n = 500 \geq 30$, $\hat{p} = \frac{37}{500} = 0.074$
- $\hat{p} - z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Como $1 - \alpha = 0.99$ então $\frac{\alpha}{2} = 0.005$ e $z_{0.995} = 2.58$ (tabela 5)
- $0.074 - 2.58\sqrt{\frac{0.074(1-0.074)}{500}} < p < 0.074 + 2.58\sqrt{\frac{0.074(1-0.074)}{500}}$
- $0.074 - 0.030 < p < 0.074 + 0.030$, i.e., $0.044 < p < 0.104$

ICs para a diferença de proporções

Para duas amostras de grande dimensão $n_1 \geq 30$ e $n_2 \geq 30$, a distribuição da diferença de proporções binomiais é

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$$

Logo, podemos construir o intervalo de confiança para $p_1 - p_2$

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

ICs para a diferença de proporções

Exemplo 7

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

- IC para $p_1 - p_2$
- duas amostras de grande dimensão: $n_1 = 500 \geq 30$ e $n_2 = 1000 \geq 30$, $\hat{p}_1 = \frac{377}{500} = 0.754$ e $\hat{p}_2 = \frac{223}{1000} = 0.223$
- $(\hat{p}_1 - \hat{p}_2) \pm z_{(1-\frac{\alpha}{2})} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
- Como $1 - \alpha = 0.95$ então $\frac{\alpha}{2} = 0.025$ e $z_{0.975} = 1.96$ (tabela 5)
- $(0.754 - 0.223) \pm 1.96\sqrt{\frac{0.754(1-0.754)}{500} + \frac{0.223(1-0.223)}{1000}}$
- $0.531 - 0.046 < p_1 - p_2 < 0.531 + 0.046$, i.e., $0.485 < p_1 - p_2 < 0.577$
- Como o IC não contém o valor 0, pode-se concluir que existam diferenças significativas entre as proporções

IC para a variância

Distribuição Qui-quadrado (χ^2)

A função densidade de probabilidade da distribuição Qui-quadrado (χ^2) com ν graus de liberdade é

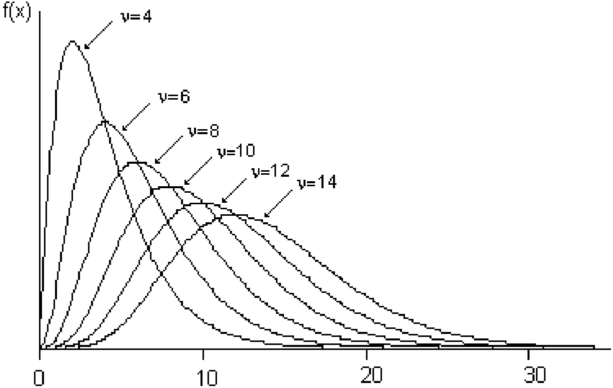
$$f(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\frac{\nu-2}{2}} e^{-x/2} & x > 0 \\ 0 & \text{outros valores} \end{cases}$$

com $E(X) = \nu$ e $V(X) = 2\nu$. Quando $\nu \to \infty$, a distribuição Qui-quadrado aproxima-se da normal.

- Se Z segue uma distribuição normal padrão, i.e., $Z \sim N(0,1)$, então Z^2 segue a distribuição de Qui-Quadrado com 1 grau de liberdade.
- Se $X_1, X_2, \ldots, X_n$ são n variáveis aleatórias independentes que seguem uma distribuição de Qui-Quadrado com $\nu_1, \nu_2, \ldots, \nu_n$ graus de liberdade, então $Y = \sum_{i=1}^{n} X_i$ segue a distribuição de Qui-Quadrado com $\nu_1 + \nu_2 + \ldots + \nu_n$ graus de liberdade.

Distribuição Qui-quadrado

A distribuição Qui-quadrado é unimodal, assimétrica à direita e aproxima-se da normal quando $\nu \to \infty$.



IC para a variância

Se $\bar{X}$ e s são a média e o desvio padrão de uma amostra aleatória de tamanho n de uma população normal com média μ e desvio padrão σ, então $X^2 = \frac{(n-1)s^2}{\sigma^2}$ segue uma distribuição Qui-quadrado com $n-1$ graus de liberdade, i.e.,

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

A partir de $P\left(\chi^2_{1-\frac{\alpha}{2},n-1} < X^2 < \chi^2_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$, como $X^2 = \frac{(n-1)s^2}{\sigma^2}$, tem-se que

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

IC para a variância

Resolvendo em ordem a σ^2, obtemos

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

Logo

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

onde s é o desvio padrão amostral.
Para o desvio padrão σ, tem-se que

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

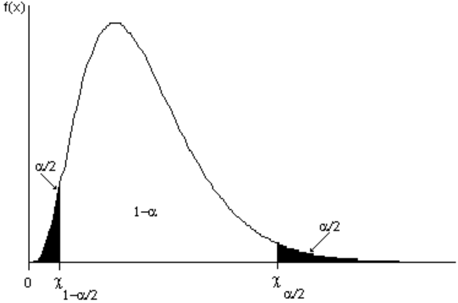
onde s é o desvio padrão amostral.

IC para a variância

Para uma amostra de qualquer dimensão retirada de uma população normal,

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

onde s é o desvio padrão amostral.



IC para a variância

Exemplo 8

Uma fábrica de cimento produz sacos de 15 kg. Para além da quantidade média de enchimento, a variação dessa quantidade é igualmente importante na medida em que alguns sacos conterão mais, e outros menos, de 15 kg. Para estimar a variação, o controlador de qualidade recolheu uma amostra de 15 sacos de cimento, com uma média de $15.02\,kg$ e um desvio padrão de $0.075\,kg$. Assumindo a normalidade do enchimento, construa um intervalo de 90% de confiança para a variância do enchimento.

- IC para σ^2, assumindo normalidade
- amostra: $n = 15$, $\bar{X} = 15.02$ e $s = 0.075$
- $\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}$
- Como $1 - \alpha = 0.90$ então $\frac{\alpha}{2} = 0.05$ e $\chi^2_{0.05,14} = 23.685$, $\chi^2_{0.95,14} = 6.571$ (tabela 7)
- $\frac{(15-1)0.075^2}{23.685} < \sigma^2 < \frac{(15-1)0.075^2}{6.571}$
- $0.003 < \sigma^2 < 0.012$

IC para a razão de variâncias

Distribuição F de Fisher

A função densidade de probabilidade da distribuição F de Fisher com ν_1 e ν_2 graus de liberdade é

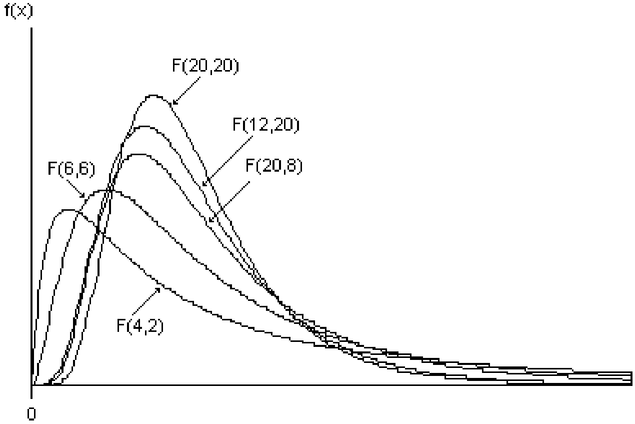
$$f(x) = \begin{cases} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right) x^{\frac{\nu_1}{2}-1} \left(1+\frac{\nu_1}{\nu_2}x\right)^{-\frac{1}{2}(\nu_1+\nu_2)} & x > 0 \\ 0 & \text{outros valores} \end{cases}$$

com $E(X) = \frac{\nu_2}{\nu_2-2}$ para $\nu_2 > 2$ e $V(X) = \frac{2\nu_2^2(\nu_1+\nu_2-2}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ para $\nu_2 > 4$.

- Se U e V são variáveis aleatórias independentes seguindo distribuições de Qui-Quadrado com ν_1 e ν_2 graus de liberdade, então $F = \frac{\frac{U}{\nu_1}}{\frac{V}{\nu_2}}$ é uma variável aleatória seguindo a distribuição F com ν_1 e ν_2 graus de liberdade.

Distribuição F de Fisher

A distribuição F é unimodal e assimétrica à direita.



IC para a razão de variâncias

Se s_1 e s_2 são os desvios padrão de duas amostras aleatórias independentes de tamanhos n_1 e n_2 retiradas de populações normais com desvios padrão σ_1 e σ_2, então $F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ segue uma distribuição F com $n_1 - 1$ e $n_2 - 1$ graus de liberdade, i.e.,

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

A partir de $P\left(F_{1-\frac{\alpha}{2},n_1-1,n_2-1} < F < F_{\frac{\alpha}{2},n_1-1,n_2-1}\right) = 1 - \alpha$, como $F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$, tem-se que

$$P\left(F_{1-\frac{\alpha}{2},n_1-1,n_2-1} < \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} < F_{\frac{\alpha}{2},n_1-1,n_2-1}\right) = 1 - \alpha$$

IC para a razão de variâncias

Resolvendo em ordem a $\frac{\sigma_1^2}{\sigma_2^2}$, obtemos

$$P\left(\frac{s_1^2}{s_2^2}F_{1-\frac{\alpha}{2},n_2-1,n_1-1} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}F_{\frac{\alpha}{2},n_2-1,n_1-1}\right) = 1-\alpha$$

Como, $F_{1-\frac{\alpha}{2},n_2-1,n_1-1} = \frac{1}{F_{\frac{\alpha}{2},n_1-1,n_2-1}}$, obtém-se

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

onde s_1 e s_2 são os desvios padrão amostrais.

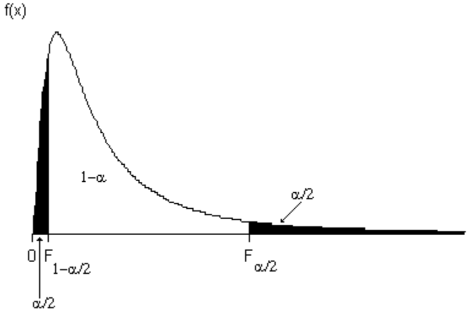
IC para a razão de variâncias

IC para a razão de variâncias

Para duas amostras independentes de quaisquer dimensões retiradas de populações normais,

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

onde s_1 e s_2 são os desvios padrão amostrais.



ICs para a variância

Exemplo 9

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

- IC para $\frac{\sigma_1^2}{\sigma_2^2}$, assumindo normalidade
- amostra: $n_1 = 13$, $n_2 = 9$, $s_1 = 0.071$ e $s_2 = 0.075$
- $\frac{s_1^2}{s_2^2}\frac{1}{F_{\frac{\alpha}{2},n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}F_{\frac{\alpha}{2},n_2-1,n_1-1}$
- Como $1-\alpha = 0.90$ então $\frac{\alpha}{2} = 0.05$ e $F_{0.05,12,8} = 3.28$, $F_{0.05,8,12} = 2.85$ (tabela 8)
- $\frac{0.071^2}{0.075^2}\frac{1}{3.28} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{0.071^2}{0.075^2}2.85$
- $0.27 < \frac{\sigma_1^2}{\sigma_2^2} < 2.55$
- como o IC contém o valor 1, não podemos concluir que existam diferenças significativas entre as variâncias