Introdução à Probabilidade e à Estatística II

Estimação III

Lígia Henriques-Rodrigues

MAE0212 – 2º semestre 2017

Distribuição Qui-Quadrado

A v.a. contínua X tem **distribuição Qui-Quadrado com ν graus de liberdade**, $\nu > 0$ inteiro, e escrevemos $X \sim \chi^2(\nu)$, se sua f.d.p. é dada por

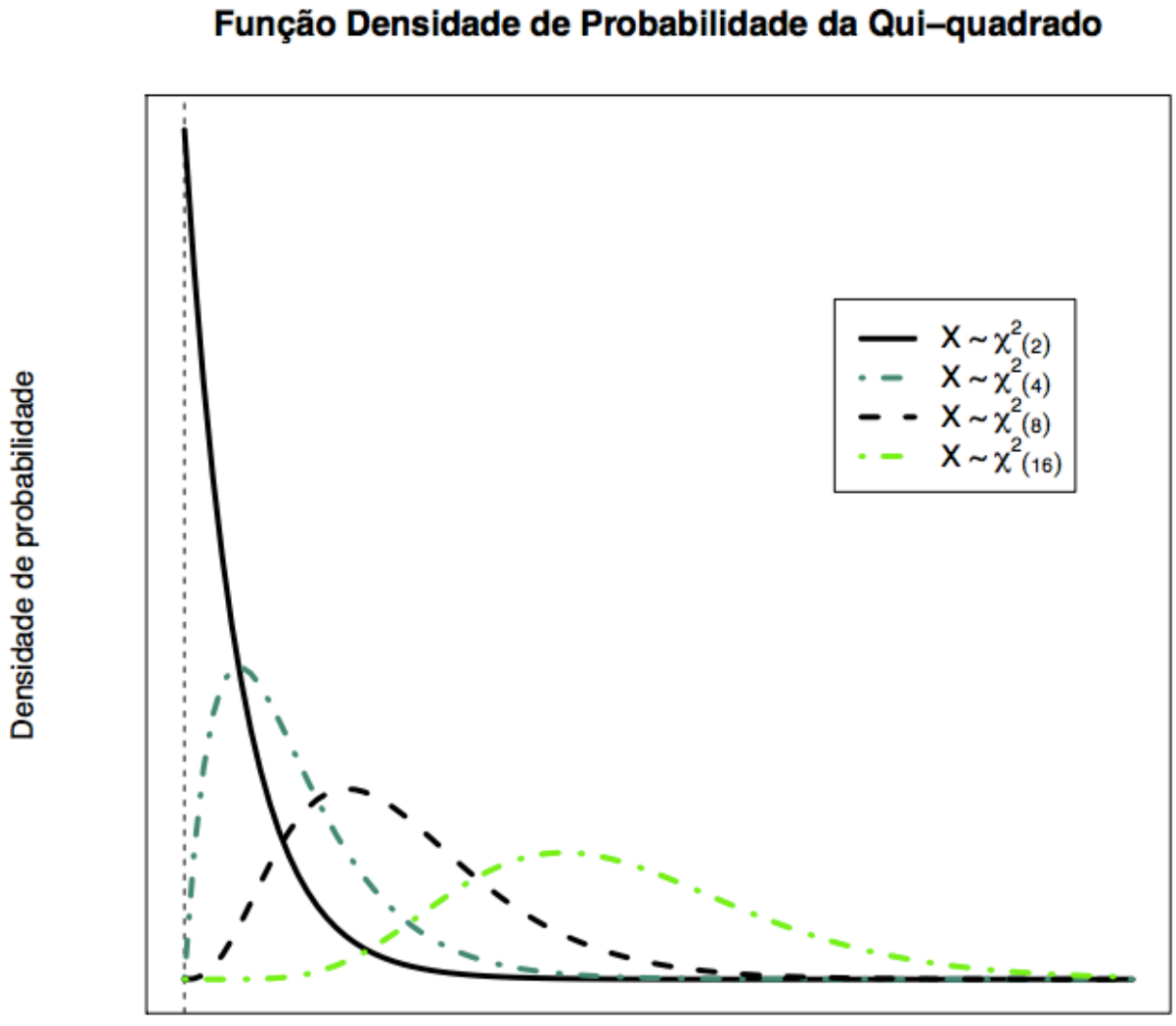
$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\nu/2)2^{\nu/2}} x^{\nu/2-1} e^{-x/2}, & x > 0 \\ 0, & x < 0, \end{cases}$$

em que $\Gamma(.)$ é a função gama,

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx,$$

satisfazendo as propriedades:

- $\Gamma(k+1) = k\Gamma(k)$;
- $\Gamma(k+1) = k!$, para k inteiro;
- $\Gamma(1/2) = \sqrt{\pi}$.


Função Densidade de Probabilidade da Qui–quadrado
Densidade de probabilidade
$X \sim \chi^2_{(2)}$
$X \sim \chi^2_{(4)}$
$X \sim \chi^2_{(8)}$
$X \sim \chi^2_{(16)}$
valores de X

Nota: A distribuição Qui-quadrado é tabelada para vários valores de ν, o número de graus de liberdade.

Na tabela da distribuição qui-quadrado encontram-se os valores y_c tais que, sendo $X \sim \chi^2(\nu)$,

$$P(X > y_c) = p.$$

Momentos

- Valor esperado: $E(X) = \nu$;
- Variância: $Var(X) = 2\nu$.

Resultados importantes

1. Se $Z \sim N(0,1) \Longrightarrow X = Z^2 \sim \chi^2(1)$.
2. Se $X_1, \ldots, X_k$ são v.a. independentes com $X_i \sim \chi^2(n_i)$, $i = 1, \ldots, k$, então

$$X_1 + \ldots + X_k \sim \chi^2(n_1 + \ldots + n_k).$$

3. Se $X_1, \ldots, X_n$ são v.a. i.i.d com $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, então

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

4. Se $X \sim \chi^2(\nu)$ então

$$\frac{X - \nu}{\sqrt{2\nu}} \approx N(0,1).$$

Se $\nu \geq 30$ a aproximação pela normal será boa.

Distribuição t-Student

Teorema: Seja Z uma v.a. $N(0,1)$ e Y uma v.a. $\chi^2(\nu)$, com Z e Y independentes. Então a v.a.

$$T = \frac{Z}{\sqrt{Y/\nu}}$$

tem distribuição t-Student com ν graus de liberdade, e escrevemos, $T \sim t(\nu)$, com f.d.p. dada por

$$f(t) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi\nu}}(1 + t^2/\nu)^{-(\nu+1)/2},\ t \in \mathbb{R}.$$

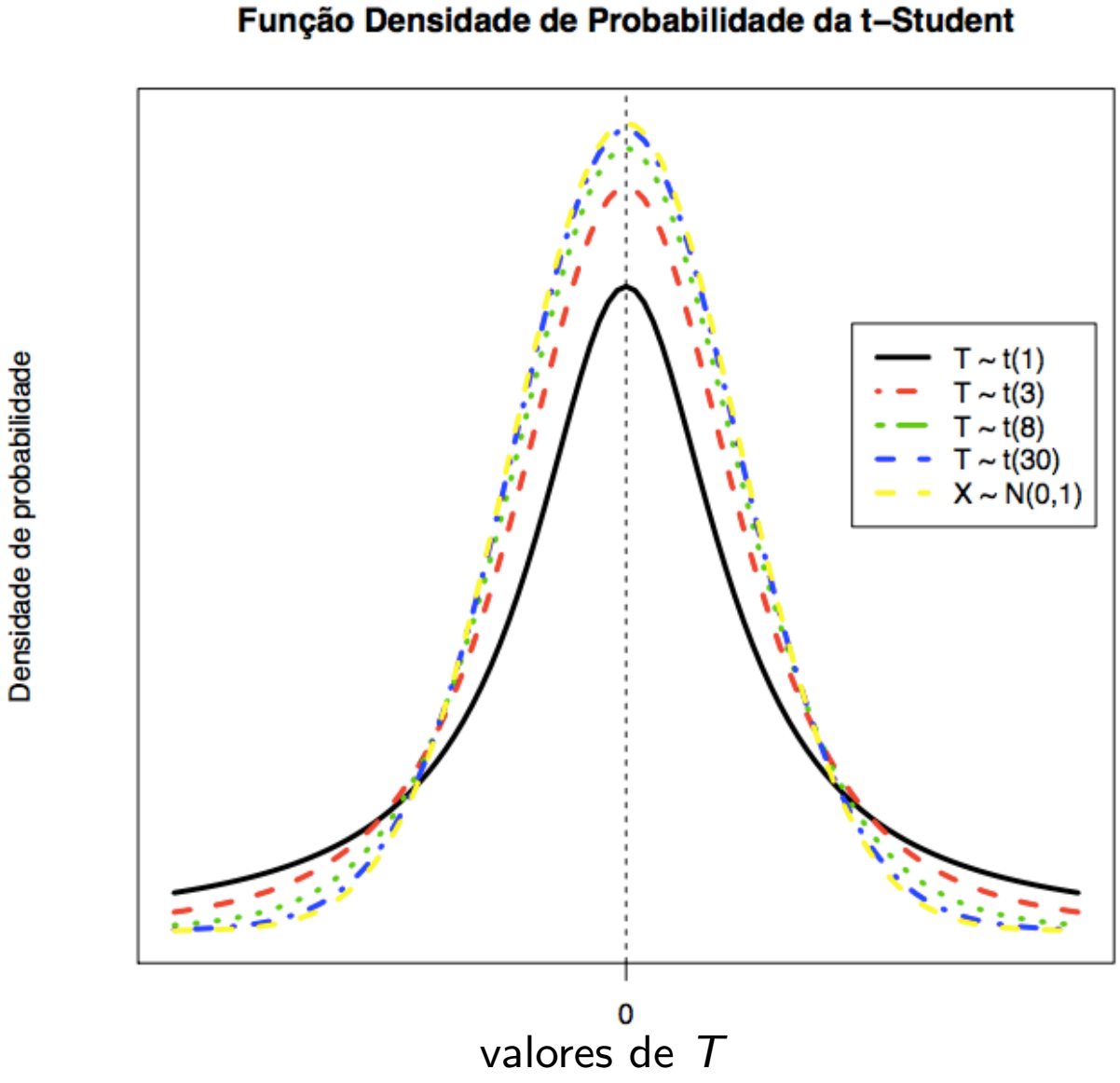
Momentos

- Valor esperado: $E(T) = 0$, $\nu > 1$;
- Variância: $Var(T) = \dfrac{\nu}{\nu - 2}$, $\nu > 2$.

Nota: A distribuição t-Student é tabelada para diversos valores de ν.

Na tabela da distribuição t-Student encontram-se os valores t_c tais que, sendo $T \sim t(\nu)$,

$$P(-t_c < T < t_c) = 1 - p.$$


Função Densidade de Probabilidade da t–Student
Densidade de probabilidade
T ~ t(1)
T ~ t(3)
T ~ t(8)
T ~ t(30)
X ~ N(0,1)
0
valores de T

2.3 Distribuições por amostragem

- **Distribuições por amostragem de $\overline{X}$**

Suponhamos que foi seleccionada uma amostra aleatória de dimensão n, $(X_1, X_2, \ldots, X_n)$, de uma população de média μ e variância σ^2. A distribuição por amostragem de $\overline{X}$ pode ser obtida sob diversas condições:

1. Suponhamos a população tem distribuição Normal e que o valor da variância da população é conhecido. Consequentemente, tendo em conta as propriedades da distribuição normal, $\overline{X} \sim N(\mu, \sigma^2/n)$, ou seja,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

2. Suponhamos a população tem distribuição Normal e que o valor da variância da população é desconhecido. Vamos usar S^2 para estimar σ^2. Nestas condições,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \text{e} \qquad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Como a população tem distribuição Normal, podemos assegurar que Z e S^2 são v.a. independentes, logo

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

3. Suponhamos que a população tem distribuição não-Normal e que o valor da variância da população é conhecida, mas a dimensão da amostra, n, é superior ou igual a 30. Neste caso, a distribuição por amostragem da média amostral pode ser aproximada pela distribuição Normal reduzida, justificado através do Teorema Limite Central:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1).$$

4. Finalmente, consideremos que seleccionámos uma amostra aleatória de uma população com distribuição não-Normal, com variância da população desconhecida e que temos um tamanho de amostra n superior ou igual a 30. Tal como no caso anterior,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1).$$

Como σ^2 é desconhecido, mas a dimensão da amostra é grande então $S \simeq \sigma$, e podemos substituir, na expressão anterior, σ por S (desvio padrão), isto é,

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \approx N(0,1).$$

- **Distribuições por amostragem de S^2**

Suponhamos que foi seleccionada uma amostra aleatória de dimensão n, $(X_1, X_2, \ldots, X_n)$, de uma população Normal de média μ desconhecida e variância σ^2. Neste contexto, a distribuição por amostragem de $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ é dada por:

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

- **Distribuição por amostragem de $\hat{p}$**

Admita que os elementos de determinada população possuem uma dada característica, com uma certa probabilidade p desconhecida, independentemente uns dos outros. Suponhamos que se selecciona uma amostra aleatória de n elementos desta população. Se X denotar o número de elementos da amostra aleatória que possuem a referida característica, sabemos que $X \sim Bin(n, p)$. Se o tamanho da amostra for suficientemente grande, o Teorema Limite Central justifica que:

$$\frac{X - np}{\sqrt{np(1-p)}} \approx N(0,1).$$

Como p pode ser estimado pontualmente pela proporção de elementos da amostra possuem a referida característica, $\hat{p} = \frac{X}{n}$, a distribuição por amostragem aproximada de $\hat{p}$ é

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \approx N(0,1).$$

Tabela: Distribuições por amostragem

Estimador	População		Distribuição
$\overline{X}$	Normal de média μ	σ^2 conhecida	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
		σ^2 desconhecida	$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$
	Pop. não-Normal de média μ e $n \geq 30$	σ^2 conhecida	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$
		σ^2 desconhecida	$\frac{\overline{X}-\mu}{S/\sqrt{n}} \approx N(0,1)$
S^2	Normal de média μ desconhecida		$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$
$\hat{p}$	Qualquer população e n grande		$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \approx N(0,1)$

3. Estimação Intervalar ou Intervalo de Confiança

Para uma amostra observada, os estimadores pontuais fornecem como estimativa um único valor numérico para o parâmetro. Os estimadores pontuais são variáveis aleatórias e, portanto, possuem uma distribuição de probabilidade, em geral, denominada **distribuição amostral**.

Objetivo: Estamos interessados em atribuir um conjunto de valores ao parâmetro desconhecido θ e se possível a precisão desse conjunto.

Os intervalos de confiança são obtidos por meio da distribuição amostral do estimador pontual.

A estimativa intervalar para θ corresponde a um intervalo do tipo:

$$]\hat{\theta} - \varepsilon; \hat{\theta} + \varepsilon[,$$

onde ε é o erro amostral ou margem de erro.

Como determinar ε? Defina-se

$$P(\varepsilon) = P(|\hat{\theta} - \theta| < \varepsilon).$$

A probabilidade $P(\varepsilon)$ é designada de **coeficiente de confiança** e denotada por γ, com $0 < \gamma < 1$.

Os coeficientes (ou níveis) de confiança utilizados são superiores a 90%. O grau de confiança γ pode ser escrito como $\gamma = 1 - \alpha$, onde α é o nível de significância.

Intervalo Aleatório de Confiança

É um intervalo caracterizado por:

- Possui pelo menos um dos extremos aleatórios que dependem exclusivamente da AAS $(X_1, \ldots, X_n)$;

$$IAC(\theta; \gamma) =]T_1(X_1, \ldots, X_n), T_n(X_1, \ldots, X_n)[;$$

- A probabilidade deste intervalo conter o verdadeiro valor do parâmetro desconhecido é γ, $P(T_1 < \theta < T_2) = \gamma$.

Intervalo de Confiança

Trata-se de uma concretização do IAC quando se obtém a amostra $(x_1, \ldots, x_n)$. Sendo $t_1 = T_1(x_1, \ldots, x_n)$ e $t_2 = T_2(x_1, \ldots, x_n)$, então

$$IC(\theta; \gamma) =]t_1, t_2[.$$

- O IC ou contém ou não contém o verdadeiro valor do parâmetro desconhecido;
- **Interpretação Frequencista:** Caso escolhêssemos um número suficientemente grande de grupos de n observações e calculássemos os respectivos $IC(\theta; \gamma)$, aproximadamente $\gamma\%$ destes IC conteriam o verdadeiro valor do parâmetro desconhecido.

Método para determinação de IC's

Se T for um estimador do parâmetro θ, e conhecida a distribuição amostral de T podemos escrever

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

onde Z é uma variável aleatória que é função da AAS e do parâmetro θ, $Z = Z(X_1, \ldots, X_n; \theta)$, e que possui distribuição independente de θ. Invertendo a desigualdade em ordem a θ será sempre possível encontrar T_1 e T_2, tais que:

$$P(T_1 < \theta < T_2) = \gamma,$$

com γ um valor fixo , $0 < \gamma < 1$.

Para uma dada amostra, teremos dois valores fixos para $t_1 = T_1(x_1, \ldots, x_n)$ e $t_2 = T_2(x_1, \ldots, x_n)$, e o intervalo de confiança para θ, com coeficiente de confiança γ, será indicado do seguinte modo:

$$IC(\theta; \gamma) =]t_1, t_2[.$$

Intervalo de confiança para a média - variância conhecida

Sendo $(X_1, \ldots, X_n)$ uma AAS de tamanho n proveniente de uma população normal com variância σ^2 conhecida, sabemos que

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \iff Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Fixado o coeficiente de confiança $\gamma = 1 - \alpha$ tal que $0 < \gamma < 1$, podemos encontrar um valor $z(\gamma)$ ou um valor $z_{\alpha/2}$ tal que

$$P(-z(\gamma) < Z < z(\gamma)) = \gamma \iff P(Z < z(\gamma)) = \frac{1+\gamma}{2}.$$

ou

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha \iff P(Z < z_{\alpha/2}) = 1 - \frac{\alpha}{2}.$$

Para obter o intervalo basta inverter a desigualdade

$$-z(\gamma) < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z(\gamma)$$

em ordem a μ, obtendo-se,

$$\overline{X} - z(\gamma)\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z(\gamma)\frac{\sigma}{\sqrt{n}}.$$

Assim, o intervalo aleatório de confiança para μ, com coeficiente de confiança γ, é dado por,

$$IAC(\mu;\gamma) = \left]\overline{X} - z(\gamma)\frac{\sigma}{\sqrt{n}}, \overline{X} + z(\gamma)\frac{\sigma}{\sqrt{n}}\right[.$$

Coletada a amostra $(x_1, \ldots, x_n)$, o intervalo de confiança para μ, com coeficiente de confiança γ, é dado por,

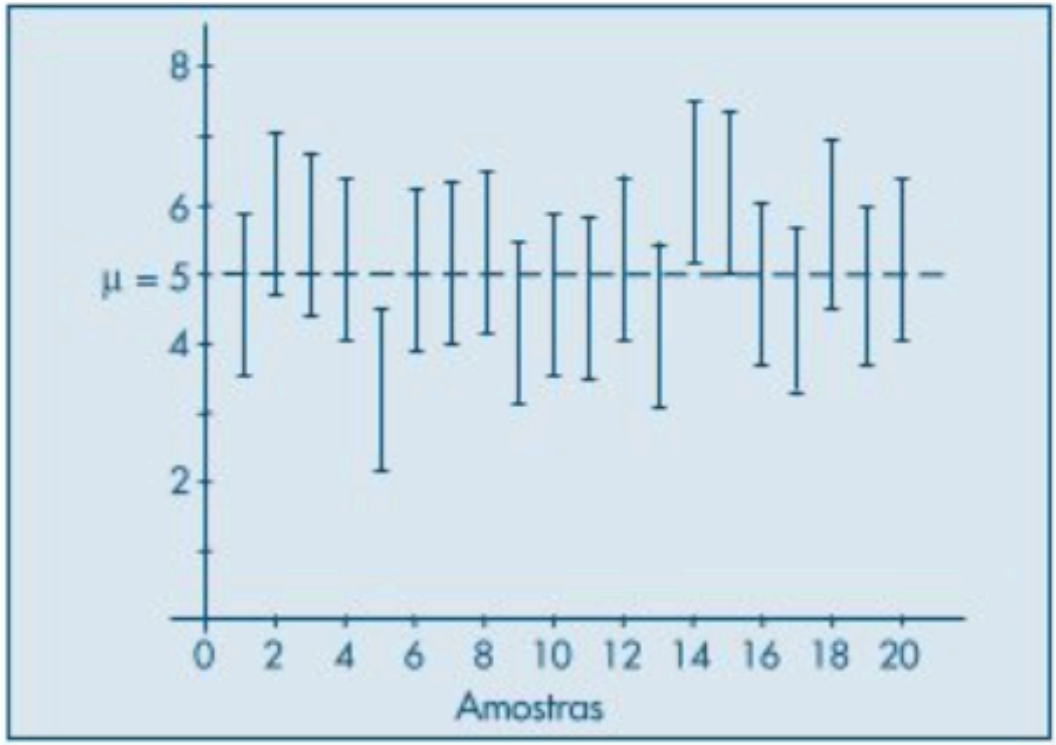
$$IC(\mu;\gamma) = \left]\overline{x} - z(\gamma)\frac{\sigma}{\sqrt{n}}, \overline{x} + z(\gamma)\frac{\sigma}{\sqrt{n}}\right[.$$

Notas:

- A probabilidade do IAC conter o verdadeiro valor do parâmetro desconhecido é γ.
- Quando se coleta a amostra $(x_1, \ldots, x_n)$, o intervalo passa a ser numérico e a interpretação conveniente é: se obtivermos várias amostras do mesmo tamanho e, para cada uma delas, calcularmos os correspondentes intervalos de confiança com coeficiente de confiança γ, esperamos que a proporção de intervalos que contenham o verdadeiro valor do parâmetro desconhecido, μ, seja igual a γ.

Interpretação gráfica

Figura 11.4: Intervalos de confiança para a média de uma N(5, 9), para 20 amostras de tamanho n = 25.



[Fonte: Bussab & Morettin]

Exercício 1: Uma máquina enche pacotes de café de acordo com uma distribuição normal com variância igual a 100 g^2. Ela estava regulada para encher pacotes com 500 g, em média. Agora, ela se desregulou, e queremos saber qual a nova média μ. Uma amostra de 25 pacotes apresentou uma média igual a 485 g. Construa um intervalo de confiança com 95% de confiança para μ.

A amplitude do intervalo de confiança é dada por

$$L = 2z(\gamma)\frac{\sigma}{\sqrt{n}},$$

e designamos de erro envolvido na estimação a semiamplitude,

$$\varepsilon = z(\gamma)\frac{\sigma}{\sqrt{n}}.$$

Exercício 2: Calcule a dimensão da mostra que será necessário recolher para obter um intervalo com amplitude $5,6$.

Intervalo de confiança para a média - variância desconhecida

1o CASO: $n < 30$

Sendo $(X_1, \ldots, X_n)$ uma AAS de tamanho n proveniente de uma população normal com variância desconhecida, sabemos que

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Fixado o coeficiente de confiança γ tal que $0 < \gamma < 1$, podemos encontrar um valor t_γ tal que

$$P(-t_\gamma < T < t_\gamma) = \gamma.$$

Para obter o intervalo basta inverter a desigualdade

$$-t_\gamma < \frac{\overline{X} - \mu}{S/\sqrt{n}} < t_\gamma$$

em ordem a μ, obtendo-se,

$$\overline{X} - t_\gamma \frac{S}{\sqrt{n}} < \mu < \overline{X} + t_\gamma \frac{S}{\sqrt{n}}.$$

Assim, o intervalo de confiança para μ, com coeficiente de confiança γ, é dado por,

$$IC(\mu;\gamma) = \left]\overline{x} - t_\gamma \frac{s}{\sqrt{n}}, \overline{x} + t_\gamma \frac{s}{\sqrt{n}}\right[.$$

2o CASO: $n \geq 30$

Neste caso, sabemos pelo TLC, que a distribuição aproximada de $\overline{X}$ é

$$\overline{X} \approx N\left(\mu, \frac{S^2}{n}\right) \iff Z = \frac{\overline{X} - \mu}{S/\sqrt{n}} \approx N(0,1),$$

obtendo-se o seguinte intervalo de confiança aproximado para μ, com coeficiente de confiança γ,

$$IC(\mu;\gamma) = \left]\overline{x} - z(\gamma) \frac{s}{\sqrt{n}}, \overline{x} + z(\gamma) \frac{s}{\sqrt{n}}\right[.$$

Exercício 3: Para estimar a vida útil média de uma válvula produzida em uma companhia foram escolhidas 10 válvulas, tendo-se obtido a vida média de 800 horas e desvio padrão de 100 horas. Sobre a hipótese de normalidade da distribuição populacional construa o intervalo de confiança a 99% para a vida média de uma válvula.

Exercício 4: Nas condições do exercício anterior, suponha que foi retirada uma amostra de 50 válvulas, tendo-se obtido a vida média de 850 horas e desvio padrão de 110 horas. Obtenha o intervalo de confiança a 95% para vida média de uma válvula.

Intervalo de confiança para a variância

Sendo $(X_1, \ldots, X_n)$ uma AAS de tamanho n proveniente de uma população normal com variância desconhecida, sabemos que

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

Fixado o coeficiente de confiança γ tal que $0 < \gamma < 1$, podemos encontrar valores χ_1^2 e χ_2^2 tais que

$$P(\chi_1^2 < T < \chi_2^2) = \gamma.$$

Para obter o intervalo basta inverter a desigualdade

$$\chi_1^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_2^2$$

em ordem a σ^2, obtendo-se,

$$\frac{(n-1)S^2}{\chi_2^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_1^2}.$$

Assim, o intervalo de confiança para σ^2, com coeficiente de confiança γ, é dado por,

$$IC(\sigma^2;\gamma) = \left]\frac{(n-1)s^2}{\chi^2_2}, \frac{(n-1)s^2}{\chi^2_1}\right[.$$

Exercício 5: Suponha que as vendas diárias, em reais, durante uma semana, de carros de uma revendedora segue distribuição $N(\mu, \sigma^2)$. Com base nos valores observados das vendas

$$253 \quad 187 \quad 96 \quad 450 \quad 320 \quad 105,$$

construa um IC$(\sigma^2; 0,90)$.

Intervalo de confiança para a proporção

Sendo $(X_1, \ldots, X_n)$ uma AAS de tamanho n proveniente de uma população Bernoulli, sabemos pelo TLC que

$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right) \iff Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \approx N(0,1)$$

Fixado o coeficiente de confiança γ tal que $0 < \gamma < 1$, podemos encontrar um valor $z(\gamma)$ tal que

$$P(-z(\gamma) < Z < z(\gamma)) = \gamma \iff P(Z < z(\gamma)) = \frac{1+\gamma}{2}.$$

Para obter o intervalo basta inverter a desigualdade

$$-z(\gamma) < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z(\gamma)$$

em ordem a p.

1o Caso: Abordagem otimista

Substitui-se $p(1-p)$ por $\hat{p}(1-\hat{p})$, sendo $\hat{p}$ um estimador consistente para p,

$$-z(\gamma) < \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} < z(\gamma)$$

$$\Longleftrightarrow$$

$$\hat{p} - z(\gamma)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z(\gamma)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

E o intervalo de confiança aproximado para p, com coeficiente de confiança γ, é dado por,

$$IC(p;\gamma) = \left]\hat{p} - z(\gamma)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z(\gamma)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right[.$$

2o Caso: Abordagem conservativa

Substitui-se $p(1-p)$ por $1/4$, um valor certamente maior do que o real,

$$-z(\gamma) < \frac{\hat{p}-p}{1/\sqrt{4n}} < z(\gamma)$$

$$\Longleftrightarrow$$

$$\hat{p} - z(\gamma)\frac{1}{\sqrt{4n}} < p < \hat{p} + z(\gamma)\frac{1}{\sqrt{4n}}.$$

E o intervalo de confiança aproximado para p, com coeficiente de confiança γ, é dado por,

$$IC(p;\gamma) = \left]\hat{p} - z(\gamma)\frac{1}{\sqrt{4n}}, \hat{p} + z(\gamma)\frac{1}{\sqrt{4n}}\right[.$$

Nota: Na abordagem conservativa aceitamos uma menor precisão para $\hat{p}$, o que se reflete numa maior amplitude do intervalo de confiança, quando comparado com o intervalo de confiança otimista.

Exercício 6: Suponha que estamos interessados em estimar a proporção p de pacientes com menos de 40 anos diagnosticados com câncer nos pulmões que sobrevivem pelo menos 5 anos. Em uma amostra aleatoriamente seleccionada de 52 pacientes, somente 6 sobreviveram mais de 5 anos. Construa um intervalo de confiança aproximado de 95% para p.