Estatística[1]

[1] Última alteração a 2 de Maio de 2016

Definição

- **População:** *conjunto de elementos sobre o qual incide o estudo estatístico;*
- **Característica Estatística ou Atributo:** *característica que se observa nos elementos da população;*
- **Parâmetro:** *característica numérica da população;*
- **Amostra:** *subconjunto da população;*

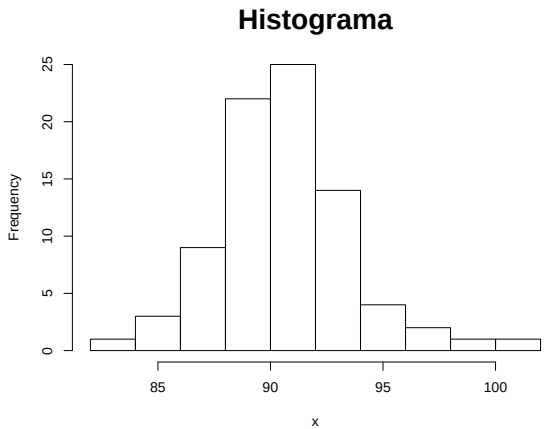
Observações:

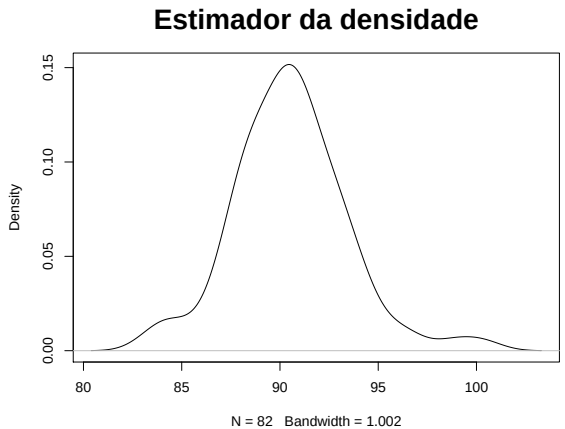
- Numa **população** podemos ter mais que uma **característica estatística**;
- Quando estamos interessados apenas numa **característica**, podemos dizer que a **população** consiste na totalidade das observações.
- Muitas vezes é impossível ou impraticável observar toda a população!

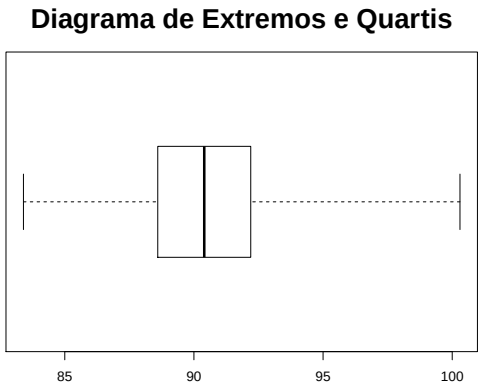
Alguns métodos gráficos que nos permitem analisar a forma da distribuição da população

Exemplo: os seguintes dados representam o nível de octanas de diversas misturas de gasolina (Snee, 1977)[2]:

88.5	94.7	84.3	90.1	89.0	89.8	91.6	90.3	90.0	91.5	89.9	98.8	88.3	90.4	91.2	90.6	92.2
87.7	91.1	86.7	93.4	96.1	89.6	90.4	91.6	90.7	88.6	88.3	94.2	85.3	90.1	89.3	91.1	92.2
83.4	91.0	88.2	88.5	93.3	87.4	91.1	90.5	100.3	87.6	92.7	87.9	93.0	94.4	90.4	91.2	86.7
94.2	90.8	90.1	91.8	88.4	92.6	93.7	96.5	84.3	93.2	88.6	88.7	92.7	89.3	91.0	87.5	87.8
88.3	89.2	92.3	88.9	89.8	92.7	93.3	86.7	91.0	90.9	89.9	91.8	89.7	92.2			







Nota: estes gráficos podem ser feitos numa folha de cálculo ou na linguagem de programação R.

[2] R. D. Snee (1977). Validation of Regression Models: Methods and Examples, *Technometrics*, Vol. **19**, No. 4, 415–428.

Notação: em geral, **letras maiúsculas** ou **letras gregas** representam características da população. **Letras minúsculas** representam as respectivas quantidades amostrais.

Por exemplo,

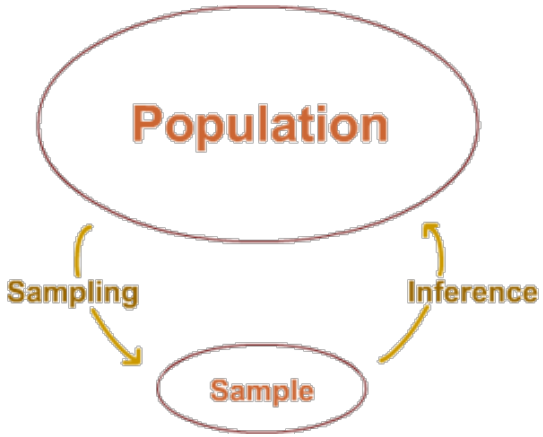
- X: elemento da população / x: elemento da amostra;
- P: proporção populacional / p: proporção amostral;
- N: dimensão da população / n: dimensão da amostra;
- μ: média populacional / $\bar{x}$: média amostral;
- σ: desvio padrão populacional / s: desvio padrão amostral;
- σ^2: variância populacional / s^2: variância amostral;

Definição (Amostra aleatória)

Vamos admitir que cada valor observado x_i *é a realização da variável aleatória* X_i*, com função de distribuição* F*. O vector* $(X_1, X_2, \ldots, X_n)$ *constitui uma* **amostra aleatória** *se e só se*

- *as* n *variáveis aleatórias* **são independentes**;
- *as* n *variáveis aleatórias* **têm todas a mesma distribuição**.

Nota: Os valores que se obtêm por concretização da **amostra aleatória** são representados por $(x_1, x_2, \ldots, x_n)$.



Definição (Estatística)

Uma **estatística** *é uma função da amostra aleatória,* $(X_1, X_2, \ldots, X_n)$*, que não depende de qualquer parâmetro desconhecido.*

Definição (Estimador)

Seja θ *um parâmetro. Um* **estimador** *de* θ *é uma estatística* $T = T(X_1, X_2, \ldots, X_n)$ *usada para estimar o parâmetro* θ*.*

Iremos estudar **estimadores pontuais** e **estimadores intervalares** de um parâmetro populacional θ.

Definição (Estimativa pontual)

Uma **estimativa pontual** *do parâmetro* θ *de uma população é o valor numérico do estimador,* $\hat{\theta} = T(x_1, x_2, \ldots, x_n)$*, calculado para uma determinada amostra.*

Nota: Não confundir a notação.

- θ (constante desconhecida) representa um **parâmetro** populacional;
- $T = T(X_1, \ldots, X_n)$,é a **estatística** usada para estimar θ;
- $\hat{\theta} = T(x_1, \ldots, x_n)$ é a **estimativa pontual** de θ.

Repare que $T(X_1, \ldots, X_n)$ é uma **variável aleatória**, porque é função de variáveis aleatórias. A distribuição de um estimador pontual é designada por **distribuição de amostragem**.

Estimação Pontual

Tabela com alguns parâmetros importantes, respectivo estimador pontual e estimativa.

Parâmetro Populacional θ	**Estimador pontual** T	**Estimativa** $\hat{\theta}$
Média populacional μ	Média amostral $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1+\ldots+X_n}{n}$	$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Variância populacional σ^2	Variância amostral $S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}$	$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$
Desvio padrão pop. σ	Desvio padrão amostral $S = \sqrt{S^2}$	$\hat{\sigma} = s = \sqrt{s^2}$
Proporção populacional p	Proporção amostral $\hat{P} = \frac{X}{n}$	$\hat{p} = \frac{x}{n}$

Nota: No estimator $\hat{P}$, X representa o n° de vezes que ocorreu o acontecimento em estudo, na amostra aleatória.

Algumas propriedades que permitem avaliar se um estimador pontual é um bom estimador:

- Enviesamento (deve ser nulo);
- Precisão (deve apresentar pequena variabilidade);
- Eficiência (= Enviesamento & Precisão)
- Consistência;

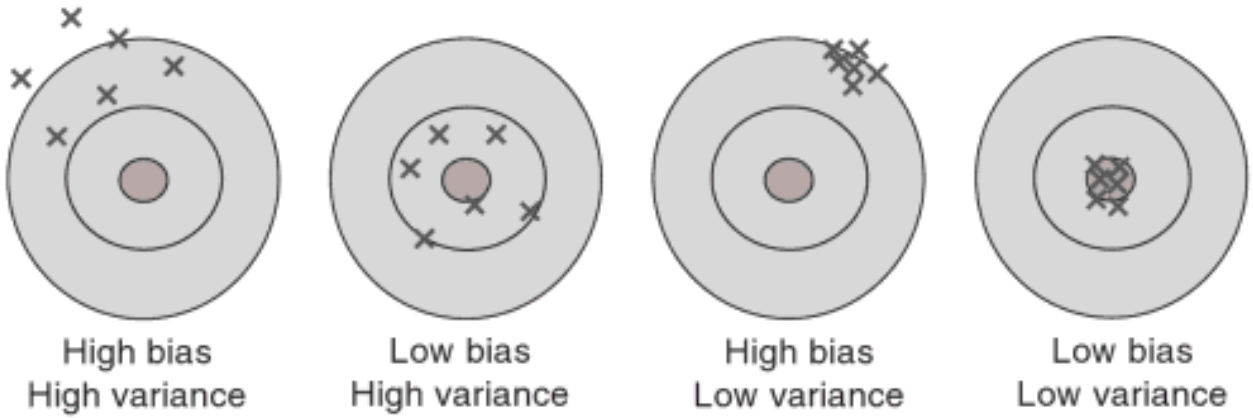


Figura: Ilustração do efeito do enviesamento (**bias**) e da precisão (**variance**).

Definição (Estimador centrado)

O estimador pontual T diz-se **centrado** *(não enviesado) para o parâmetro θ se*

$$E(T) = \theta.$$

Notas:

- Se $E(T) \neq \theta$, o estimador é **enviesado**.
- A diferença $b(T) = E(T) - \theta$ corresponde ao valor do **enviesamento** ou **viés** de T.

A variabilidade de um estimador (medida de precisão) deve ser expressa, na mesma escala de medição que a associada ao estimador, através do desvio padrão desse estimador, a que se dá o nome de **erro padrão** do estimador e representa-se por **SE**.

Definição (Erro Padrão de um estimador)

Dado um estimador pontual T, centrado, define-se o seu **erro padrão**, *que se designa SE_T, como a raiz quadrada da sua variância, caso exista:*

$$SE(T) = \sqrt{V(T)}$$

Caso **SE** envolva parâmetros desconhecidos, mas que possam ser estimados, podemos obter o **erro padrão estimado**, denotado $\widehat{SE}(T)$.

O estimador T para o parâmetro θ será tanto "melhor", quanto menor for a sua dispersão em torno do verdadeiro valor de θ.

Definição (Erro Quadrático Médio)

O **erro quadrático médio** *de um estimador pontual T de θ é*

$$EQM(T) = E[(T-\theta)^2]$$

Teorema

$$EQM(T) = V(T) + b^2(T).$$

Observação: Se o estimador for centrado, então $EQM(T) = V(T)$.

Definição (Eficiência)

Sejam T_1 e T_2 dois estimador pontuais de um parâmetro θ. Diz-se que T_1 é mais eficiente que T_2, se e só se,

$$EQM(T_1) < EQM(T_2).$$

Observação: Se ambos os estimadores forem centrados, T_1 é mais eficiente que T_2, se e só se,

$$V(T_1) < V(T_2).$$

Definição (Estimador consistente)

Um estimador T de um parâmetro θ é um estimador consistente de θ se e só se, qualquer que seja o valor real $\delta > 0$,

$$\lim_{n\to\infty} P(|T-\theta| < \delta) = 1$$

Observação: Uma condição suficiente para assegurar a consistência é

$$\lim_{n\to\infty} EQM(T) = 0.$$

Numa população cuja distribuição depende de k parâmetros, $\theta_1, \theta_2, \ldots, \theta_k$, os estimadores de momentos $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$, respectivamente, são os que resultam da resolução do sistema de k equações a k incógnitas,

$$\begin{cases} E(X) = \overline{X} \\ E((X-\mu)^2) = M_2 \\ E((X-\mu)^3) = M_3 \\ \vdots \\ E((X-\mu)^k) = M_k \end{cases} \qquad \text{onde} \qquad M_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r, \quad r > 1.$$

Inconvenientes:

1. Por vezes não existe uma única solução;
2. Por vezes a solução é inadmissível;

Exemplo 1: Seja $X \sim P(\lambda)$. Então

$$\hat{\lambda} = \overline{X}$$

Exemplo 2: Seja $X \sim U(a, b)$. Então

$$\hat{a} = \overline{X} - \sqrt{3M_2}$$

e

$$\hat{b} = \overline{X} + \sqrt{3M_2}$$

Exemplo:

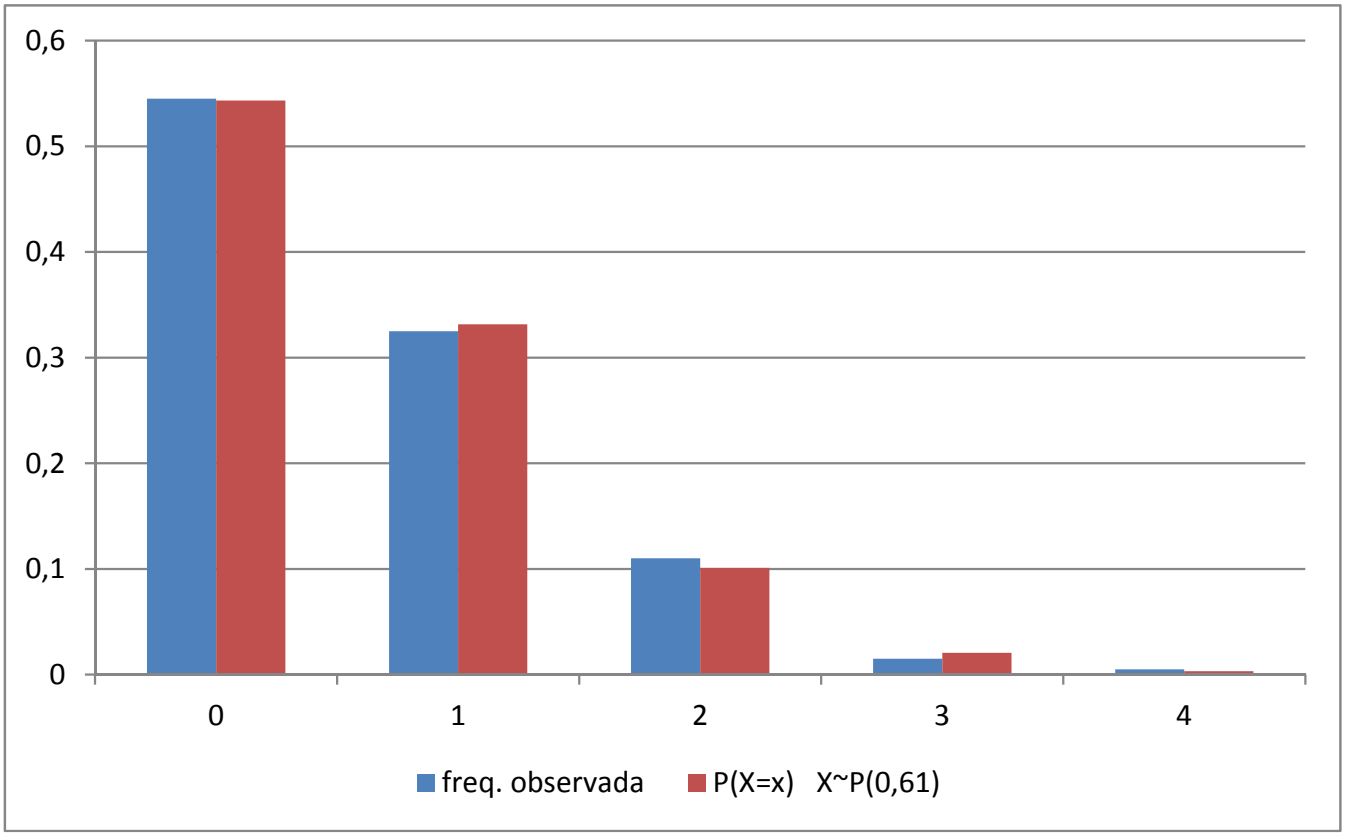
- Um dos trabalhos mais conhecido do economista Ladislaus von Bortkiewicz é o estudo do número de cavaleiros do exército prussiano mortos, por um coice de cavalo (von Bortkiewicz, L. (1898). *Das Gesetz der kleinen Zahlen* [tradução: The law of small numbers], Leipzig, Germany).

- A seguinte tabela apresenta a frequência do número de cavaleiros mortos por coice durante um ano, em 10 unidades de cavalaria, num período de 20 anos.

nº de vítimas	0	1	2	3	4	≥ 5
frequência	109	65	22	3	1	0

Considere que o modelo de Poisson é adequado e estime o parâmetro λ.

Comparação entre os valores da frequência relativa e as probabilidades dadas pelo modelo $P(0.61)$.



Distribuição por amostragem de alguns estimadores

A distribuição de um estimador é designada **distribuição por amostragem**.

Estimador	População		Distribuição
$\overline{X}$	Normal de média μ	σ^2 conhecida	$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
		σ^2 desconhecida	$T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$
	Qualquer população de média μ e $n \geq 30$	σ^2 conhecida	$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \overset{a}{\sim} N(0,1)$
		σ^2 desconhecida	$Z = \frac{\overline{X}-\mu}{S/\sqrt{n}} \overset{a}{\sim} N(0,1)$
$\hat{P}$	Qualquer população e n grande (≥ 30)		$Z = \frac{\hat{P}-p}{\sqrt{p(1-p)/n}} \overset{a}{\sim} N(0,1)$
S^2	Normal de média μ desconhecida		$X^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

Nota Importante: As distribuições por amostragem dos estimadores servirão de base à *estimação intervalar* e à realização de *testes de hipóteses* sobre os parâmetros (μ, σ ou p) da população.

Estimação por intervalo de confiança

Considere uma população de distribuição $N(170, 10^2)$. Vamos admitir que o valor médio $\mu = 170$ é desconhecido e que o pretendemos estimar a partir da amostra de dimensão $n = 20$,

171.66	171.30	161.32	172.11	141.16	168.69	165.44	180.33	184.79	157.98
171.74	163.53	167.28	169.06	162.49	161.79	173.19	170.56	163.01	164.20

A estimativa da média da população é

$$\hat{\mu} = \bar{x} = 167.08$$

Conclusão: É importante dispôr de alguma forma de intervalo que indique a confiança que podemos ter no valor da estimativa pontual. A amplitude desse intervalo dá-nos informação acerca da precisão da estimativa.

Um exemplo de aplicação:

Sondagem TVI: regressou o empate técnico

PSD perde vantagem ganha na segunda-feira e deixa de ter maioria absoluta com o CDS. Toda a esquerda sobe

27 de Maio de 2011 às 20:00 Redação / FC

PSD e PS estão de novo em empate técnico. A sondagem desta sexta-feira da INTERCAMPUS para a **TVI** e jornal «Público» indica que os sociais-democratas perderam a vantagem que tinham registado na última segunda-feira.

O PSD atinge agora os 35,8% das intenções de voto, o PS chega aos 34,1% e o CDS fica nos 11,3%. Um dado novo nesta sondagem: deixa de haver, com estes números a garantia de uma maioria absoluta entre o PSD e o CDS. A direita desceu e a esquerda subiu, dado que a CDU está agora nos 7,7% e o Bloco de Esquerda nos 6,5%.

Comparando com os resultados de segunda-feira, verifica-se que o PSD perdeu 3,8%, o PS subiu nove décimas, o CDS perdeu oito décimas, a CDU sobe 1,1% e o Bloco tem mais nove décimas. O voto nos outros partidos é de 4,5%.

CONFIRA A FICHA TÉCNICA COMPLETA

Universo constituído pela população com mais de 18 anos, residente em Portugal Continental. Recolha através de entrevista telefónica num total de 1015 entrevistas: 51.8% dos entrevistados do sexo Feminino, 48.2% do sexo Masculino, com a distribuição etária e por regiões presente no quadro: 32.1% dos entrevistados com idades entre os 18 e os 34 anos, 33.7% entre os 35 e os 54 anos e 34.2% dos indivíduos com 55 e mais anos. Por regiões 17.5% dos entrevistados residem no Norte Litoral, 13.4% no Grande Porto, 19.8% no Interior, 18.2% no Centro Litoral, 21% na Grande Lisboa e 10% no Sul. O erro de amostragem, para um intervalo de confiança de 95%, é de mais ou menos 3.08%. A taxa de resposta foi de 43.8%. Nesta sondagem 25.7% dos entrevistados não revelaram a sua opção e 16.8% não indicou um partido ou indicou que não votaria. Que quando aplicável, é feita uma distribuição proporcional de registo de não respondentes, sem opinião e abstenção, passando a usar-se a expressão «Projecção».

Fonte: http://www.tvi24.iol.pt/politica/legislativas/sondagem-tvi-regressou-o-empate-tecnico

Definição (Intervalo Aleatório)

Seja θ um parâmetro desconhecido e $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória de uma população com função de distribuição F. Considere as estatísticas

$$T_1(X_1, X_2, \ldots, X_n) \qquad e \qquad T_2(X_1, X_2, \ldots, X_n),$$

que não dependem do valor de θ e que satisfazem

$$P(T_1 \leq \theta \leq T_2) = 1 - \alpha, \qquad 0 < \alpha < 1.$$

Então:

- $[T_1, T_2]$ *é o* **intervalo de confiança (aleatório)** *para θ;*
- T_1 *e* T_2 *são denominados* **limites de confiança**;
- $1 - \alpha$ *é o* **coeficiente (ou nível) de confiança** *do intervalo.*

Definição (Intervalo de Confiança)

Seja $(x_1, x_2, \ldots, x_n)$ *uma realização da amostra aleatória e sejam*

$$t_1 = T_1(x_1, x_2, \ldots, x_n) \qquad e \qquad t_2 = T_2(x_1, x_2, \ldots, x_n),$$

os valores das estatísticas T_1 *e* T_2.

- *Ao intervalo* $[t_1, t_2]$ *chamamos* **intervalo de confiança** $(1-\alpha) \times 100\%$ *para* θ.
- *O valor* $(1-\alpha)$ *representa o* **coeficiente (ou nível) de confiança** *do intervalo e* α *o* **nível de significância**.
- *Costumamos usar níveis de confiança iguais ou superiores a 90%. Os valores mais usuais são* $90\%, 95\%$ *e* 99%.

Definição (Método Pivotal)

Método para determinação de um intervalo de confiança $1-\alpha$ *para* θ,

1. *Conhecer (ou encontrar) uma* **variável pivot**[3] $T = T(X_1, X_2, \ldots, X_n, \theta)$.
2. *A partir da distribuição de* T, *determinar* a_1 *e* a_2, *tais que*

$$a_1 < a_2 \qquad e \qquad P(a_1 \leq T \leq a_2) = 1-\alpha;$$

3. *Resolver as desigualdades* $a_1 \leq T(X_1, X_2, \ldots, X_n, \theta) \leq a_2$ *em ordem a* θ,

$$a_1 \leq T \leq a_2 \;\Leftrightarrow\; T_1(X_1, X_2, \ldots, X_n) \leq \theta \leq T_2(X_1, X_2, \ldots, X_n),$$

   sendo $T_1(X_1, X_2, \ldots, X_n)$ *e* $T_2(X_1, X_2, \ldots, X_n)$ *estatísticas não dependentes de* θ;

4. $IC_{(1-\alpha)\times 100\%}(\theta) = [T_1(X_1, X_2, \ldots, X_n),\, T_2(X_1, X_2, \ldots, X_n)]$ *é um intervalo de confiança* $1-\alpha$ *para* θ.

[3] v.a. com distribuição conhecida e que depende apenas do parâmetro desconhecido.

Intervalo de Confiança para a média populacional, μ

Vamos considerar uma população com valor médio μ desconhecido. A seguinte tabela apresenta a variável pivot que se deve usar para deduzir o intervalo de confiaça para o valor médio populacional.

População	Variância (σ^2)	Variável pivot
$X \sim N(\mu, \sigma^2)$	conhecida	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
	desconhecida	$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$
Qualquer População e $n \geq 30$	conhecida	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \overset{a}{\sim} N(0,1)$
	desconhecida	$\frac{\overline{X}-\mu}{S/\sqrt{n}} \overset{a}{\sim} N(0,1)$

Intervalo de Confiança para a média populacional, μ

População $N(\mu, \sigma^2)$, σ^2 conhecida ou qualquer população e $n \geq 30$

Suponha que pretendemos obter um intervalo de confiança para μ.

1. Variável Pivot:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \begin{cases} \sim N(0,1) & \text{(População } N(\mu, \sigma^2)\text{)} \\ \overset{a}{\sim} N(0,1) & \text{(Qualquer População e } n \geq 30\text{)} \end{cases}$$

2. Determinação das constantes: $a_1 = -z_{\alpha/2}$ e $a_2 = z_{\alpha/2}$.

3. Resolução das desigualdades:

$$P(a_1 \leq Z \leq a_2) = 1 - \alpha \Leftrightarrow P\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

$$\Leftrightarrow \ldots \Leftrightarrow \quad P\left(\overline{X} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

4. $$IC_{(1-\alpha)\times 100\%}(\mu) = \left[\overline{X} - z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\ ;\ \overline{X} + z_{\alpha/2}\,\frac{\sigma}{\sqrt{n}}\right]$$

Observações

- Ao substituírmos $\boxed{\bar{X}}$ por $\boxed{\bar{x}}$ (valor observado ou estimativa da média populacional) passamos a ter um intervalo concreto chamado **intervalo de confiança**;
- Não podemos garantir que μ pertença ao intervalo de confiança com probabilidade $1-\alpha$. Mas podemos dizer que se fizermos um grande número de intervalos nestas condições, aproximadamente $100 \times (1-\alpha)\%$ desses intervalos contêm o verdadeiro valor de μ (que permanece desconhecido).
- A amplitude do intervalo de confiança está associado à precisão. Quanto menor for a amplitude, mais precisa deve ser a estimativa pontual.

Exemplo (Ex. 3 - 2º teste de Estatística - 27/11/2013)

3. *Medições do comprimento de 25 peças produzidas por uma máquina conduziram a uma média $\bar{x} = 140mm$. Admita que cada peça tem comprimento aleatório com distribuição normal de valor esperado μ e desvio padrão $\sigma = 10mm$, e que o comprimento de cada peça é independente das restantes.*

(a) *Construa um intervalo de confiança a 95% para o valor esperado da população.* ($IC_{95\%}(\mu) = [136.08, 143.92]$)

(b) *Qual deverá ser o tamanho da amostra de forma a que a amplitude do correspondente intervalo de confiança a 95% para a média não exceda $2mm$?* ($n \geq 385$)

Intervalo de Confiança para a média populacional, μ

População $N(\mu, \sigma^2)$, σ^2 desconhecida

Suponha que pretendemos obter um intervalo de confiança para μ.

1. Variável Pivot:
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \quad \sim t_n - 1$$
2. Determinação das constantes: $a_1 = -t_{n-1:\frac{\alpha}{2}}$ e $a_2 = t_{n-1:\frac{\alpha}{2}}$.
3. Resolução das desigualdades:
$$P(a_1 \leq T \leq a_2) = 1 - \alpha \Leftrightarrow P\left(-t_{n-1:\alpha/2} \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t_{n-1:\alpha/2}\right) =$$
$$\Leftrightarrow \ldots \Leftrightarrow \quad P\left(\overline{X} - t_{n-1:\alpha/2}\,\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1:\alpha/2}\,\frac{S}{\sqrt{n}}\right) =$$
4.
$$IC_{(1-\alpha)\times 100\%}(\mu) = \left[\overline{X} - t_{n-1:\alpha/2}\,\frac{S}{\sqrt{n}}\ ;\ \overline{X} + t_{n-1:\alpha/2}\,\frac{S}{\sqrt{n}}\right]$$

Intervalo de Confiança para a média populacional, μ

População	Variância (σ^2)	$IC_{(1-\alpha)\times 100\%}(\mu)$
$X \sim N(\mu, \sigma^2)$	conhecida	$\left[\bar{X} - z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}}\ ;\ \bar{X} + z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}}\right]$
	desconhecida	$\left[\bar{X} - t_{n-1:\frac{\alpha}{2}}\ \frac{S}{\sqrt{n}}\ ;\ \bar{X} + t_{n-1:\frac{\alpha}{2}}\ \frac{S}{\sqrt{n}}\right]$
"Qualquer" pop.[4]	conhecida	$\left[\bar{X} - z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}}\ ;\ \bar{X} + z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}}\right]$
e $n \geq 30$	desconhecida	$\left[\bar{X} - z_{\alpha/2}\ \frac{S}{\sqrt{n}}\ ;\ \bar{X} + z_{\alpha/2}\ \frac{S}{\sqrt{n}}\right]$

[4] Qualquer população com valor médio μ e variância σ^2.

Intervalo de confiança para a variância populacional

Considere a situação em que temos uma amostra aleatória $(X_1, \ldots, X_n)$ de uma população $N(\mu, \sigma^2)$, com μ desconhecido.
Pretendemos obter um intervalo de confiança $(1-\alpha) \times 100\%$ para a variância populacional, σ^2:

1. Escolha da variável pivot usada para construir o IC para σ^2:

$$X^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{(Qui-Quadrado com } n-1 \text{ graus de liberdade)},$$

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

Nota: A distribuição por amostragem de X^2 é válida para populações com distribuição normal.

Intervalo de confiança para a variância populacional

2. Depois, é necessário determinar a_1 e a_2, tais que $a_1 < a_2$ e

$$P(a_1 \leq X^2 \leq a_2) = 1 - \alpha.$$

Vamos escolher a_1 e a_2, tais que

$$P(X^2 < a_1) = \alpha/2 \qquad \text{e} \qquad P(X^2 > a_2) = 1 - \alpha/2,$$

ou seja $a_1 = \chi^2_{n-1:1-\alpha/2}$ e $a_2 = \chi^2_{n-1:\alpha/2}$.

3.

$$P(a_1 \leq X^2 \leq a_2) = 1 - \alpha \Leftrightarrow P\left(a_1 \leq \frac{(n-1)S^2}{\sigma^2} \leq a_2\right) = 1 - \alpha$$

$$\Leftrightarrow \ldots \Leftrightarrow \quad P\left(\frac{(n-1)S^2}{\chi^2_{n-1:\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1:1-\alpha/2}}\right) = 1 - \alpha$$

4. Assim,

$$\boxed{IC_{(1-\alpha)\times 100\%}(\sigma^2) = \left[\frac{(n-1)S^2}{\chi^2_{n-1:\alpha/2}} \; ; \; \frac{(n-1)S^2}{\chi^2_{n-1:1-\alpha/2}}\right]}$$

Intervalo de confiança para o desvio padrão populacional

Do resultado atrás obtido, muito simplesmente se constrói o intervalo de confiança para o **desvio padrão populacional** σ. Como

$$P(a_1 \leq X^2 \leq a_2) = 1-\alpha \Leftrightarrow P\left(a_1 \leq \frac{(n-1)S^2}{\sigma^2} \leq a_2\right) = 1-\alpha$$

$$\Leftrightarrow \ldots \Leftrightarrow \quad P\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1:\alpha/2}}} \leq \sigma \leq \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1:1-\alpha/2}}}\right) = 1-\alpha,$$

temos o Intervalo de Confiança

$$IC_{(1-\alpha)\times 100\%}(\sigma) = \left[\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1:\alpha/2}}} \; ; \; \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1:1-\alpha/2}}}\right]$$

Intervalo de confiança para a proporção populacional p

- Vamos assumir que os elementos de determinada população podem possuir uma dada **característica**, com uma certa probabilidade p desconhecida, independentemente uns dos outros.
- Suponha que se selecciona uma amostra aleatória de n elementos desta população.
- Se X denotar o número desses elementos que possuem a referida **característica**, sabemos que $\boxed{X \sim B(n,p)}$ (amostragem **com** reposição) ou $\boxed{X \sim H(N,M,n)}$ (amostragem **sem** reposição - mas se $n < 0.1N$ a distribuição Hipergeométrica pode ser aproximada pela distribuição Binomial).
- Vamos assim considerar que $\boxed{X \sim B(n,p)}$.

Intervalo de confiança para a proporção populacional p

Se a tamanho da amostra, n, for suficientemente grande, o **Teorema Limite Central** justifica que:

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \overset{a}{\sim} N(0,1)$$

Notamos ainda que o estimador de p é $\hat{P} = \frac{X}{n}$, a proporção amostral. Resulta então a seguinte variável pivot

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

Intervalo de Confiança para proporção populacional, p

Vamos agora determinar um Intervalo de Confiança para p.

1. Variável Pivot:

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

2. Determinação das constantes: $a_1 = -z_{\alpha/2}$ e $a_2 = z_{\alpha/2}$.

3. Resolução das desigualdades:

$$P(a_1 \leq Z \leq a_2) = 1 - \alpha \Leftrightarrow P\left(-z_{\frac{\alpha}{2}} \leq \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\frac{\alpha}{2}} \right) = 1 - \alpha$$

Sendo p um parâmetro desconhecido, a resolução das desigualdades é mais fácil se estimarmos o denominador da fracção anterior por $\sqrt{\hat{P}(1-\hat{P})/n}$.

3 Então,

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\hat{P}-p}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha \quad \Leftrightarrow \ldots \Leftrightarrow$$

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

4 Solução aproximada:

$$IC_{(1-\alpha)\times 100\%}(p) = \left[\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\,;\, \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right]$$

Definição (Hipótese Estatística)

Uma hipótese estatística é uma conjetura acerca da distribuição de uma ou mais variáveis aleatórias. Essa conjectura pode incidir sobre os parâmetros de uma ou mais populações (teste paramérico) ou acerca da distribuição da população (testes não paramétrico de ajustamento).

- Se a hipótese estatística especifica completamente a distribuição é chamada de **hipótese simples**. Caso contrário é chamada de **hipótese composta**.
- Para cada hipótese que se faça, designada por **hipótese nula** e denotada por H_0, há sempre outra hipótese, designada por **hipótese alternativa** e denotada por H_1.

Exemplo

Seja $(X_1, X_2, \ldots, X_n)$ *uma amostra aleatória de uma população com* $X \sim N(\mu, 2^2)$. *A hipótese estatística de que o valor médio desta população toma o valor 8 denota-se por:*

$$H_0 : \mu = 8 \quad versus \quad H_1 : \mu \neq 8 \qquad (\textbf{Hipótese simples})$$

A hipótese estatística de que o valor médio desta população é menor ou igual a 8 denota-se por:

$$H_0 : \mu \leq 8 \qquad vs. \qquad H_1 : \mu > 8 \qquad (\textbf{Hipótese composta})$$

Definição (Teste de uma hipótese estatística)

Um teste de uma hipótese estatística H_0 é uma regra (ou critério) para decidirmos se **rejeitamos** *ou* **não rejeitamos** H_0.
A decisão é tomada com base no valor duma estatística T e de um subconjunto $R \in \mathbb{R}$. Rejeitamos H_0 se $t = T(x_1, \ldots, x_n) \in R$.

Exemplo: Seja $(X_1, \ldots, X_n)$ uma amostra aleatória da população dos pesos de formigas Solenopsis com distribuição $N(\mu, 2^2)$. Um teste possível para testar:

$$H_0 : \mu \leq 8 \quad vs. \quad H_1 : \mu > 8,$$

é rejeitar H_0 se

$$\underbrace{\frac{\bar{X} - 8}{2/\sqrt{n}} > 1.64}_{\text{regra de decisão}} \qquad \underbrace{R =]1.64, +\infty[}_{\text{região crítica}}.$$

Definição (Erros do tipo I e do tipo II)

Decisão:	**Situação real:**	
	H_0 *é verdadeira*	H_0 *é falsa*
Não rejeitar H_0	*Decisão acertada*	**erro do tipo II**
Rejeitar H_0	**erro do tipo I**	*Decisão acertada*

Definimos ainda as probabilidades:

$$\alpha = P(\textbf{erro do tipo I}) = P(\text{rejeitar } H_0 \,|\, H_0 \text{ é verdadeira}),$$
$$\beta = P(\textbf{erro do tipo II}) = P(\text{não rejeitar } H_0 \,|\, H_0 \text{ é falsa}).$$

α: **nível de significância**

$1-\beta$: **potência do teste**

Algumas observações:

- O ideal é conseguirmos que ambas as probabilidades α e β tomem o seu valor mínimo (**zero**).
- Contudo, é **impossível** minimizar α e β simultaneamente pois, quando α diminui, β aumenta e vice-versa.
- É mais fácil controlar α do que controlar β (que depende do valor do parâmetro dado pela hipótese H_1). Então:
  - rejeitar H_0 é uma conclusão "**forte**";
  - não rejeitar H_0 é uma conclusão "**fraca**".

Exemplo de um teste de hipóteses

Suponha que temos uma amostra aleatória de dimensão $\boxed{n = 9}$ duma população com distribuição $\boxed{N(\mu, 1)}$ e que pretendemos testar

$$H_0 : \mu = 5 \qquad vs. \qquad H_1 : \mu \neq 5$$

Regra de decisão: rejeitar H_0 se $\bar{X}$ estiver "longe" de 5, isto é, se $\bar{X} < 4.5$ ou se $\bar{X} > 5.5$.

- Qual o nível de significância?

$$\begin{aligned}
\alpha &= P(\bar{X} < 4.5 \text{ ou } \bar{X} > 5.5 \,|\, \mu = 5) = \\
&= P(\bar{X} < 4.5 \,|\, \mu = 5) + P(\bar{X} > 5.5 \,|\, \mu = 5) = \\
&= \Phi(\sqrt{9}(4.5 - 5)) + 1 - \Phi(\sqrt{9}(5.5 - 5)) = \\
&= 2 \times 0.0668 = 0.1336
\end{aligned}$$

Exemplo de um teste de hipóteses

- Qual a potência do teste se $\mu = 5.6$?

$$\begin{aligned} 1-\beta &= P(\bar{X} < 4.5 \text{ ou } \bar{X} > 5.5 | \mu = 5.6) = \\ &= P(\bar{X} < 4.5 \,|\, \mu = 5.6) + P(\bar{X} > 5.5 \,|\, \mu = 5.6) = \\ &= \underbrace{\Phi(\sqrt{9}(4.5 - 5.6))}_{=0.0005} + \underbrace{1 - \Phi(\sqrt{9}(5.5 - 5.6))}_{=0.6179} = 0.6184 \end{aligned}$$

- Qual a potência do teste se $\mu = \mu_1$?

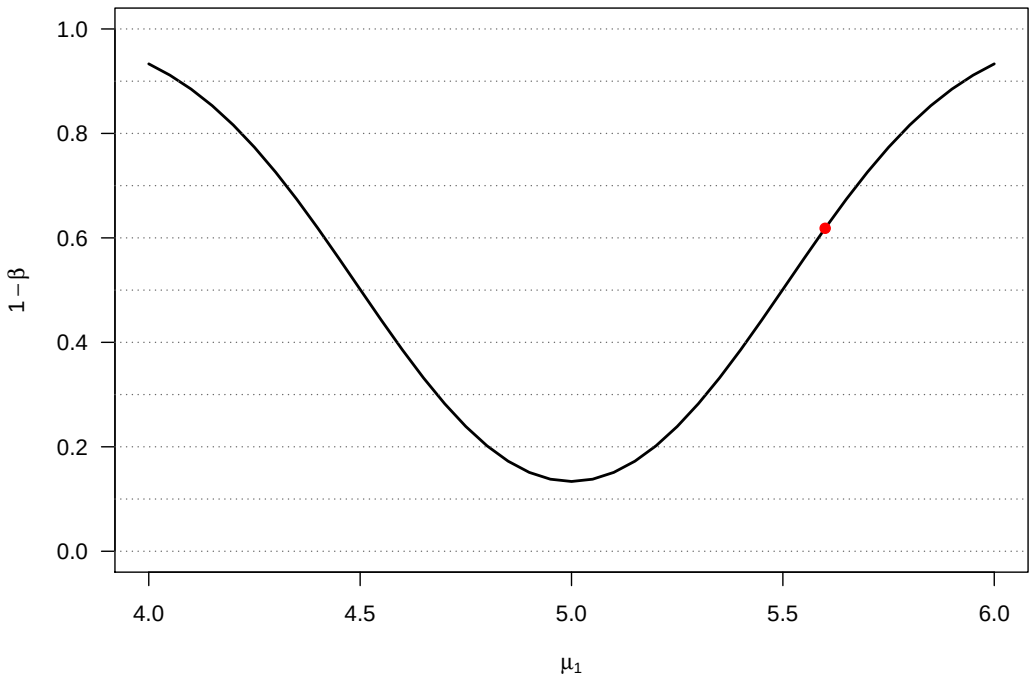


Figura: Função potência do teste de hipóteses.

Procedimento genérico de um teste de hipóteses:

1. Especificar a **hipótese** nula (H_0) e a hipótese alternativa (H_1);
2. Escolher uma **estatística de teste** adequada;
3. Escolher o **nível de significância**, α;
4. Determinar a **região crítica** do teste de hipóteses;
5. Calcular o **valor observado da estatística de teste** com base na amostra recolhida.
6. **Decidir** sobre a rejeição ou não da hipótese nula, H_0.

Teste de hipóteses para o valor médio populacional, μ

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0 \qquad \text{(teste \textbf{bilateral})}$$

População	Variância (σ^2)	Estat. de Teste	Rejeitar H_0 se
$X \sim N(\mu, \sigma^2)$	conhecida	$Z \underset{\text{Sob } H_0}{\sim} N(0,1)$	$\lvert z_{obs}\rvert > z_{\alpha/2}$
	desconhecida	$T \underset{\text{Sob } H_0}{\sim} t_{n-1}$	$\lvert t_{obs}\rvert > t_{n-1:\alpha/2}$
Qualquer população e $n \geq 30$	conhecida	$Z \underset{\text{Sob } H_0}{\overset{a}{\sim}} N(0,1)$	$\lvert z_{obs}\rvert > z_{\alpha/2}$
	desconhecida	$Z \underset{\text{Sob } H_0}{\overset{a}{\sim}} N(0,1)$	$\lvert z_{obs}\rvert > z_{\alpha/2}$

Nota: $\boxed{Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}}$ $\boxed{Z = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}}$ $\boxed{T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}}$

Teste de hipóteses para o valor médio populacional, μ

$$H_0 : \mu \leq \mu_0 \quad vs. \quad H_1 : \mu > \mu_0 \qquad \text{(teste unilateral } \textbf{direito}\text{)}$$

População	Variância	Estat. de Teste	Rejeitar H_0 se
$X \sim N(\mu, \sigma^2)$	σ^2, conhecida	$Z \underset{\text{Sob } H_0}{\sim} N(0,1)$	$z_{obs} > z_\alpha$
	σ^2, desconhecida	$T \underset{\text{Sob } H_0}{\sim} t_{n-1}$	$t_{obs} > t_{n-1:\alpha}$
Qualquer população e $n \geq 30$	σ^2, conhecida	$Z \underset{\text{Sob } H_0}{\overset{a}{\sim}} N(0,1)$	$z_{obs} > z_\alpha$
	σ^2, desconhecida	$Z \underset{\text{Sob } H_0}{\overset{a}{\sim}} N(0,1)$	$z_{obs} > z_\alpha$

Nota: $\boxed{Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}}$ $\boxed{Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}}$ $\boxed{T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}}$

Teste de hipóteses para o valor médio populacional, μ

$H_0 : \mu \geq \mu_0 \quad vs. \quad H_1 : \mu < \mu_0$ (teste unilateral **esquerdo**)

População	Variância	Estat. de Teste	Rejeitar H_0 se
$X \sim N(\mu, \sigma^2)$	σ^2, conhecida	$Z \underset{\text{Sob } H_0}{\sim} N(0,1)$	$z_{obs} < -z_\alpha$
	σ^2, desconhecida	$T \underset{\text{Sob } H_0}{\sim} t_{n-1}$	$t_{obs} < -t_{n-1:\alpha}$
Qualquer população e $n \geq 30$	σ^2, conhecida	$Z \underset{\text{Sob } H_0}{\overset{a}{\sim}} N(0,1)$	$z_{obs} < -z_\alpha$
	σ^2, desconhecida	$Z \underset{\text{Sob } H_0}{\overset{a}{\sim}} N(0,1)$	$z_{obs} < -z_\alpha$

Nota: $Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ $Z = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$ $T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}}$

Teste de hipóteses: p-value

Em vez de decidirmos em função da região crítica conter, ou não, o valor observado da estatística de teste, podemos determinar, com base no valor observado da estatística de teste, para que nível de significância a decisão muda.

Definição (Valor-p ou "p-value")

De um modo informal, podemos definir o **valor-p** *ou "***p-value***" como o mais pequeno nível de significância que leva à rejeição de H_0. Assim,*

- *um valor-p pequeno é desfavorável a H_0.*
- *um valor-p elevado indica que as observações são consistentes com H_0.*

Regra de cálculo do valor-p:
Seja $(x_1, x_2, \ldots, x_n)$ a concretização da amostra aleatória e

$$w_{obs} = W(x_1, x_2, \ldots, x_n),$$

o valor observado da estatística de teste W. Vamos assumir que W tem distribuição contínua.

Região de rejeição	*valor-p*
$]-\infty, -c[\cup]c, +\infty[$ ou $]0, b[\cup]c, +\infty[$	$2 \times \min\left(P(W < w_{obs} \mid H_0), P(W > w_{obs} \mid H_0)\right)$
$]-\infty, c[$ ou $]0, c[$	$P(W < w_{obs} \mid H_0)$
$]c, +\infty[$	$P(W > w_{obs} \mid H_0)$

Teste de hipóteses: p-value

Aviso:

- Para os testes cuja estatística de teste tem distribuição **normal**, conseguimos calcular facilmente o valor-p.
- Para os testes cuja estatística de teste tem outra distribuição (**t de Student** ou **qui-quadrado**), o valor-p só pode ser obtido com precisão usando um software adequado. Recorrendo às tabelas, usadas nas aulas, o melhor que conseguimos é obter um valor aproximado ou um intervalo que contém o valor-p.

Testes de hipóteses para a variância σ^2 populacional

Considere a situação em que temos uma amostra aleatória $(X_1, X_2, \ldots, X_n)$ de uma população $N(\mu, \sigma^2)$, com μ **desconhecido**.

Hipóteses:

1. $H_0 : \sigma = \sigma_0 \quad vs. \quad H_1 : \sigma \neq \sigma_0$ (teste bilateral);

ou

2. $H_0 : \sigma \leq \sigma_0 \quad vs. \quad H_1 : \sigma > \sigma_0$ (teste unilateral direito);

ou

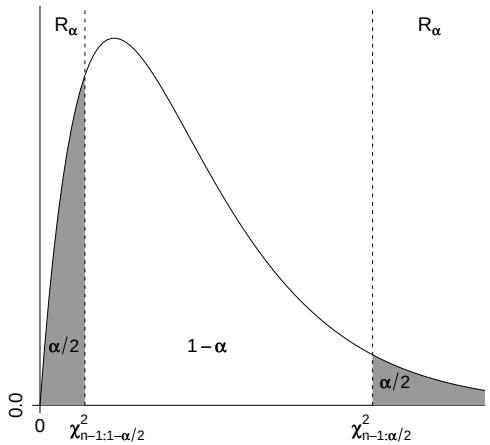
3. $H_0 : \sigma \geq \sigma_0 \quad vs. \quad H_1 : \sigma < \sigma_0$ (teste unilateral esquerdo);

Estatística de teste:

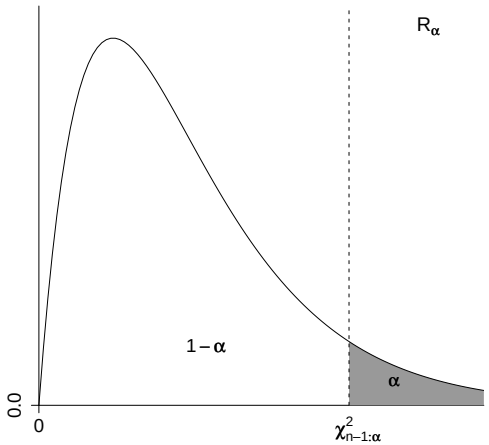
$$X^2 = \frac{(n-1)S^2}{\sigma_0^2} \overset{\text{sob } H_0,\ \sigma = \sigma_0}{\sim} \chi^2_{n-1}$$

Teste de hipóteses para a variância σ^2 populacional

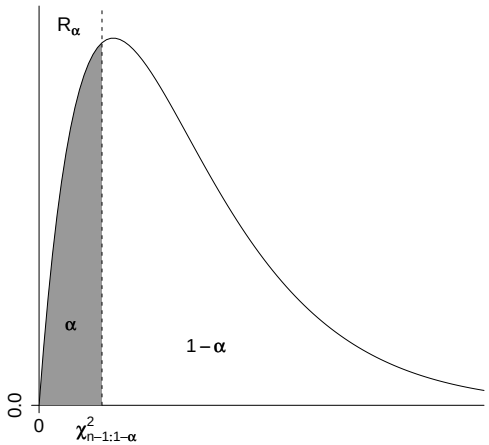
Região de rejeição do teste, para um nível de significância α:



(1)



(2)



(3)

1. $R_\alpha =]0; \chi^2_{n-1:1-\frac{\alpha}{2}}[\ \cup \ ]\chi^2_{n-1:\frac{\alpha}{2}}; +\infty[$ (teste bilateral);
2. $R_\alpha =]\chi^2_{n-1:\alpha}; +\infty[$ (teste unilateral direito);
3. $R_\alpha =]0; \chi^2_{n-1:1-\alpha}[$ (teste unilateral esquerdo);

Teste de hipóteses para a variância σ^2 populacional

Exemplo (Ex. 4 - 2º teste de Estatística - 27/11/2013)

4. *Foram efectuados estudos numa cidade com o objectivo de determinar a concentração de monóxido de carbono (CO) perto das vias rápidas. Para o efeito recolheram-se amostras de ar para as quais se determinou a respectiva concentração de CO. Os resultados (em ppm) foram os seguintes:*

$$101.1; \quad 98.2; \quad 100.4; \quad 96.8; \quad 88.2$$

Para estes dados $\sum_{i=1}^{5}(x_i - \bar{x})^2 = 107.272$. *Assuma que tais concentrações se distribuem normalmente.*

(a) *Teste, sem utilizar o valor-p, para um nível de significância* $\alpha = 0.05$, *as hipóteses*

$$H_0 : \sigma^2 \geq 28.53 \quad vs. \quad H_1 : \sigma^2 < 28.53.$$

(b) *Determine o valor-p e indique a decisão a tomar no teste* $H_0 : \sigma^2 \leq 28.53$ $vs.$ $H_1 : \sigma^2 > 28.53$, *para* $\alpha = 0.01$, *sabendo que, com base numa outra amostra de igual dimensão* ($n = 5$), *se obteve um valor observado da estatística de teste igual a 9.49.*

Teste de hipóteses para a proporção populacional, p

Suponha que observamos uma amostra aleatória de dimensão n de uma população, em que determinada proporção desconhecida p dos seus elementos possui certa característica.
Hipóteses:

1. $H_0 : p = p_0 \quad vs. \quad H_1 : p \neq p_0$ (teste bilateral);
2. $H_0 : p \leq p_0 \quad vs. \quad H_1 : p > p_0$ (teste unilateral direito);
3. $H_0 : p \geq p_0 \quad vs. \quad H_1 : p < p_0$ (teste unilateral esquerdo);

Estatística de teste:

$$Z = \frac{P - p_0}{\sqrt{p_0(1-p_0)/n}} \underset{\text{sob } H_0, \text{ com } p = p_0}{\overset{\text{a}}{\sim}} N(0,1)$$

Região de rejeição do teste, para um nível de significância α pré-especificado:

1. $R_\alpha =]-\infty; -z_{\frac{\alpha}{2}}[\cup]z_{\frac{\alpha}{2}}; +\infty[$ (teste bilateral);
2. $R_\alpha =]z_\alpha; +\infty[$ (teste unilateral direito);
3. $R_\alpha =]-\infty; -z_\alpha[$ (teste unilateral esquerdo);