Estatística [1]

[1] Última alteração a 26 de Abril de 2018

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;*

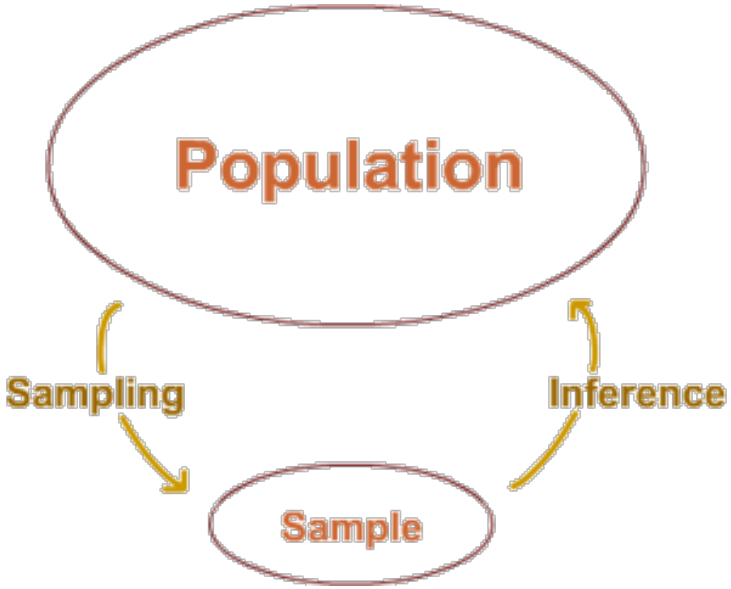
Observações:

- 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!

Estatística

- **Estatística Descritiva**: Sumaria e descreve os aspetos relevantes de um conjunto de dados. Os dados são tratados com recurso a tabelas, gráficos e indicadores numéricos.

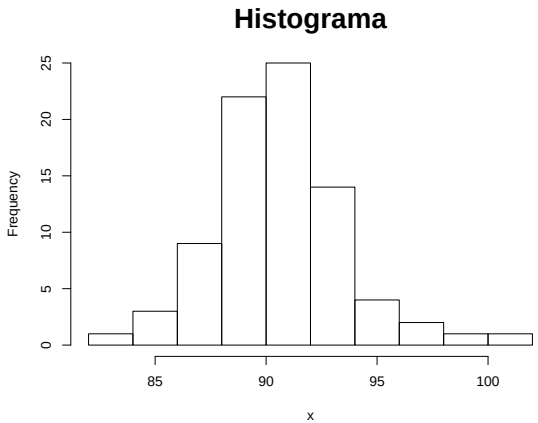
- **Inferência Estatística (ou Estatística Indutiva)**: permite-nos obter conclusões para a população, a partir do estudo de uma amostra.

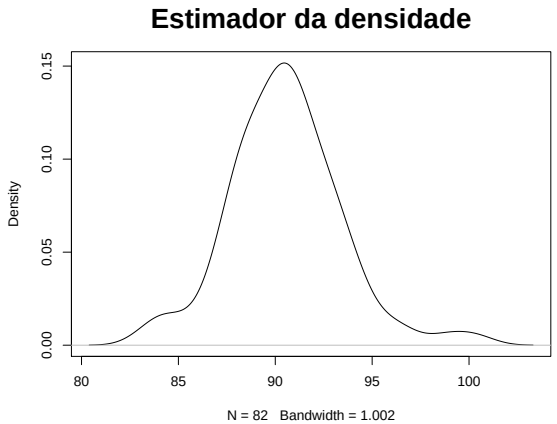


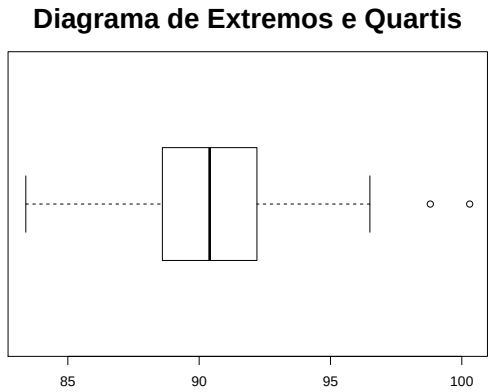
Alguns métodos gráficos que nos permitem analisar a forma da distribuição da população

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.

Definição (Amostra aleatória)

O vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *constitui uma* **amostra aleatória** *se e só se*

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

Alternativamente podemos dizer que as variáveis $X_1, X_2, \ldots, X_n$ *são independentes e idênticamente distribuídas (i.i.d.).*

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

- Na inferência Estatística Paramétrica assumimos que conhecemos uma família de distribuições de X

$$\boxed{\mathcal{P} = \{f(x;\theta);\, x \in \mathcal{X},\, \theta \in \Theta\}}$$

onde $\mathcal{X}$ é o suporte de X e Θ é o espaço de parâmetros (completamente conhecido). Caso X tenha distribuição discreta, $f(x;\theta)$ pode ser substituída por uma função de probabilidade.

- Um dos objectivos em Inferência Estatística é o de fazer inferência acerca do parâmetro desconhecido θ, usando a informação contida na amostra aleatória.

- Quando a dimensão da amostra, n, é elevada, temos uma longa lista de números, de difícil interpretação. Qual a solução?

  Resposta: sumariar a informação da amostra em "*estatísticas*".

Definição (Estatística)

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

Observação: A distribuição de probabilidade de T é designada por distribuição de amostragem de T.

Definição (Estimador)

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

Estimação Pontual

Tabela com alguns parâmetros importantes, respetivo 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 $\hat{P} = \frac{X}{n}$	$\hat{p} = \frac{x}{n}$
Coeficiente de variação $\frac{\sigma}{\mu}$	$\frac{S}{\bar{X}}$	$\frac{s}{\bar{x}}$

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

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

Parâmetro Populacional	**Estimador pontual**	**Estimativa**
Mediana	$M = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{se } n \text{ é ímpar} \\ \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} & \text{se } n \text{ é par} \end{cases}$	$m = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{se } n \text{ é ímpar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{se } n \text{ é par} \end{cases}$
Quantil	$Q_p = \begin{cases} X_{([np]+1)} & \text{se } np \notin \mathbb{N} \\ \frac{X_{(np)} + X_{(np+1)}}{2} & \text{se } np \in \mathbb{N} \end{cases}$	$q_p = \begin{cases} x_{([np]+1)} & \text{se } np \notin \mathbb{N} \\ \frac{x_{(np)} + x_{(np+1)}}{2} & \text{se } np \in \mathbb{N} \end{cases}$
Amplitude	$X_{(n)} - X_{(1)}$	$x_{(n)} - x_{(1)}$
Amplitude Interquartis	$Q_{0.75} - Q_{0.25}$	$q_{0.75} - q_{0.25}$

Nota: $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ representam as **estatísticas de ordem** ou **estatísticas ordinais** da amostra.

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;

Exemplos

1. Seja $X \sim P(\lambda)$. Então $\hat{\lambda} = \overline{X}$.
2. Seja $X \sim U(a, b)$. Então

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

e

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

3. Seja $X \sim N(\mu, \sigma^2)$. Então,

$$\hat{\mu} = \overline{X} \quad \text{e} \quad \hat{\sigma}^2 = M_2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n};$$

Definição (Estimador de máxima verosimilhança (MV))

Seja $\mathbf{X} = (X_1, \ldots, X_n)$ *uma amostra aleatória duma população cuja distribuição depende de* k *parâmetros:* $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. *Denotemos por* $\mathcal{L}(\underline{\theta}; \mathbf{x})$ *a função de verosimilhança, definida por*

$$\mathcal{L}(\underline{\theta}; \mathbf{X}) = f(\mathbf{x}; \underline{\theta}) \underset{\text{v.a.'s i.i.d.}}{=} \prod_{i=1}^{n} f(x_i; \underline{\theta})$$

ou

$$\mathcal{L}(\underline{\theta}; \mathbf{X}) = P(\mathbf{X} = \mathbf{x}; \underline{\theta}) \underset{\text{v.a.'s i.i.d.}}{=} \prod_{i=1}^{n} P(X_i = x_i; \underline{\theta}),$$

consoante a população é contínua ou discreta.
O estimador de máxima verosimilhança de $\underline{\theta}$ *é dado por*

$$\underline{\hat{\theta}} = \max_{\underline{\theta}} \mathcal{L}(\underline{\theta}; \mathbf{x})$$

Se $\mathcal{L}(\underline{\theta}; \mathbf{X})$ for diferenciável relativamente a todos os θ_i, as estimativas de MV são a solução do sistema de equações $\frac{\partial \mathcal{L}(\theta; \mathbf{X})}{\partial \theta_i} = 0, \ i = 1, \ldots, k.$

Observação:

- Os zeros da primeira derivada apenas localizam os extremos no interior do domínio da função de verosimilhança. E esses extremos podem ser extremos locais ou pontos de inflexão.

- Na maior parte das situações é mais fácil trabalhar com a função de **log-verosimilhança**:

$$l(\theta; \mathbf{X}) = \ln \mathcal{L}(\theta; \mathbf{X}),$$

do que trabalhar directamente com a função de verosimilhança.

Exemplos: Seja $\mathbf{X}$ uma amostra aleatória duma população

1. $P(\lambda)$. Então,

$$l(\lambda;\mathbf{x}) = \ln L(\lambda;\mathbf{x}) = -n\lambda + \left(\sum_{i=1}^{n} x_i\right) \ln\lambda - \sum_{i=1}^{n} \ln x_i! \quad \text{se } x_i \in \mathbb{N}_0.$$

Como $\frac{\partial l(\lambda)}{\partial \lambda} = 0 \Leftrightarrow \lambda = \overline{X}$, o estimador de máxima verosimilhança de λ é

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

2. $N(\theta, 1)$, com $\theta \in \mathbb{R}$. Então,

$$l(\theta;\mathbf{x}) = -n\ln\sqrt{2\pi} - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2$$

Resulta que $\boxed{\hat{\theta} = \overline{X}}$ é o estimador de máxima verosimilhança de θ;

Algumas propriedades que permitem avaliar a qualidade de um estimador pontual:

- Enviesamento (exactidão);
- Variância (precisão);
- Eficiência (= Enviesamento & Variância)
- Consistência;

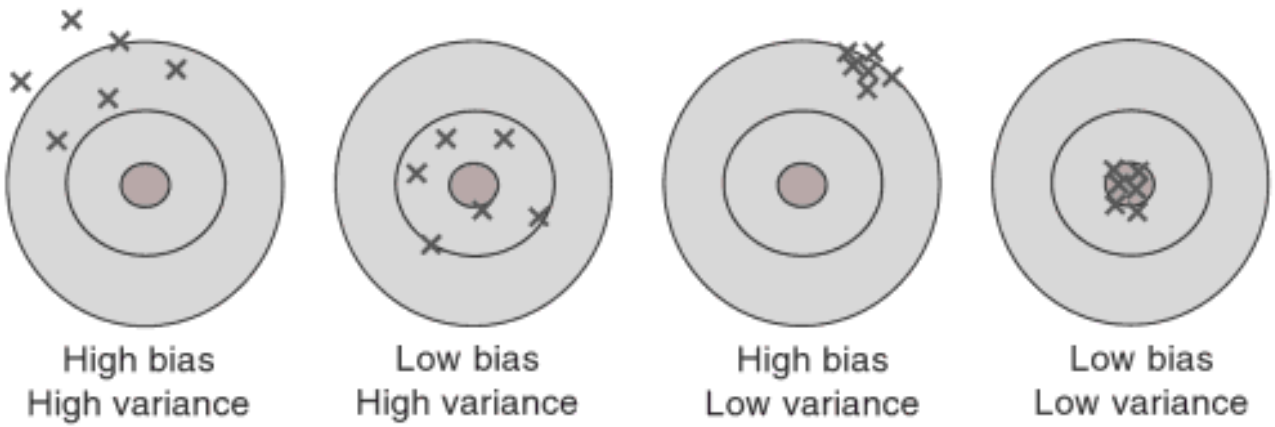


Figura: Ilustração do efeito do enviesamento (**bias**) e da variância (**variance**).

Definição (Estimador centrado)

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

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

Notas:

- Se $E(T) \neq \theta$, o estimador é **enviesado**.
- A diferença $Vies(T) = E(T) - \theta$ corresponde ao valor do **enviesamento** ou **viés** de T.
- Se $E(T) \neq \theta$, e $\lim_{n\to\infty} E(T) = \theta$, diz-se que o estimador é **assintoticamente centrado**.

Exemplo: Seja $\mathbf{X}$ uma amostra aleatória duma população com valor médio μ e variância σ^2. Então os estimadores $\bar{X}$ e S^2 são centrados para μ e σ^2, já que

$$\boxed{E(\bar{X}) = \mu}$$

e

$$\boxed{E(S^2) = \sigma^2.}$$

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 do parâmetro θ é*

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

Teorema

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

Observação: Se o estimador T 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

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

Exemplo:

Seja $\mathbf{X}$ uma amostra aleatória duma população com valor médio μ e variância σ^2. Então os estimadores $\bar{X}$ e S^2 são centrados e

$$\boxed{EQM(\bar{X}) = V(\bar{X}) = \frac{\sigma^2}{n}.}$$

Para populações normais temos $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ e

$$\boxed{EQM(S^2) = V(S^2) = \frac{2\sigma^4}{n-1}.}$$

Se considerarmos outro estimador de σ^2, $\boxed{\hat{\sigma}^2 = \frac{n-1}{n}S^2}$, obtemos

$$EQM(\hat{\sigma}^2) = \frac{(2n-1)\sigma^4}{n^2} < EQM(S^2) = \frac{2\sigma^4}{n-1}$$

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».

Definição (Intervalo Aleatório)

Seja θ um parâmetro desconhecido e $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória. 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

$$\underbrace{P(\theta \in [T_1, T_2])}_{\text{Probabilidade de cobertura}} = P(T_1 \leq \theta \leq T_2) = 1 - \alpha, \qquad 0 < \alpha < 1.$$

Então:

- *O intervalo (de confiança) aleatório $[T_1, T_2]$ é chamado* **estimador intervalar** *de 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*.*

Notas:

- *Ao intervalo* $[t_1, t_2]$ *chamamos* **intervalo de confiança** $(1-\alpha) \times 100\%$ *para* θ*.*
- *Costumamos usar níveis de confiança iguais ou superiores a 90%. Os valores mais usuais são* $90\%, 95\%$ *e* 99%*.*
- *O intervalo é que é aleatório, não o parâmetro.*
- *Por vezes, a probabilidade de cobertura não depende de* θ*.*

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. que depende apenas do parâmetro θ e cuja distribuição não depende de θ.

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:

$$P(a_1 \leq T \leq a_2) = 1 - \alpha \Leftrightarrow P\left(-t_{n-1:\alpha/2} \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t_{n-1:\alpha/2}\right) = 1 - \alpha$$

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

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 inequações é mais fácil se estimarmos o denominador da fracção anterior por $\sqrt{\hat{P}(1-\hat{P})/n}$.

Intervalo de Confiança para proporção populacional, p

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:

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

Intervalo de Confiança para proporção populacional, p

Exercício

Num inquérito destinado a estimar a proporção p de elementos da população que tem TV por cabo, foram inquiridas 200 pessoas, das quais 78 afirmaram ter este serviço. Determine a estimativa pontual de p e o respectivo intervalo de 95% de confiança.

Solução: $\hat{p} = 0.39$. Como $n = 200 > 30$ e $z_{0.05} = 1.96$, o intervalo de 95% de confiança para a proporção p é:

$$IC(p) = \left[0.39 - 1.96\sqrt{0.39(1-0.39)/200}\ ;\ 0.39 + 1.96\sqrt{0.39(1-0.39)/200}\right] = [0.322\ ,\ 0.458].$$

O intervalo de confiança de Wilson para a proporção é $IC_{95\%}(p) = [0.325\ ,\ 0.459]$

Definição (Hipótese Estatística)

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

- O objectivo de um teste de hipóteses é decidir, com base numa amostra da população em estudo qual de entre duas hipóteses é a "verdadeira".
- As duas hipóteses complementares num teste de hipóteses são designadas de **hipótese nula** (H_0) e de **hipótese alternativa** (H_1).
- Se a hipótese estatística especifica completamente a distribuição é chamada de **hipótese simples**. Caso contrário é chamada de **hipótese composta**.

Se θ é o parâmetro populacional e Θ é o espaço–parâmetro, o formato geral das duas hipóteses é

$$H_0 : \theta \in \Theta_0 \quad versus \quad H_1 : \theta \in \Theta_1 \quad (\Theta_0 \cup \Theta_1 = \Theta).$$

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 duma população 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:

$$P(\textbf{erro do tipo I}) = P(\text{rejeitar } H_0 \,|\, H_0 \text{ é verdadeira}),$$
$$1-\beta = P(\textbf{erro do tipo II}) = P(\text{não rejeitar } H_0 \,|\, H_0 \text{ é falsa}).$$

O **nível de significância** é $\boxed{\alpha = \max_\theta P(\textbf{erro do tipo I}).}$

A (função) potência é $\boxed{\beta = 1 - P(\textbf{erro do tipo II}).}$

Algumas observações:

- O ideal é conseguirmos que ambas as probabilidades α e $1-\beta$ tomem o seu valor mínimo (**zero**).
- Contudo, é **impossível** minimizar α e $1-\beta$ simultaneamente pois, quando α diminui, $1-\beta$ aumenta e vice-versa.
- Depois de tomada uma decisão, apenas se pode cometer um dos erros. É mais fácil controlar α do que controlar $1-\beta$ (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 $\mu = 5.6$?

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

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

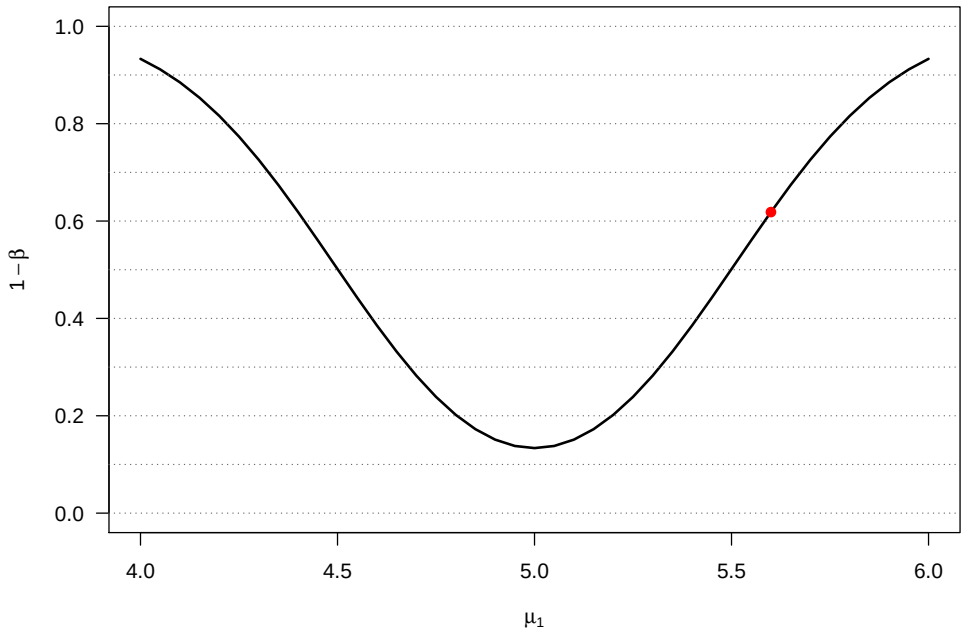


Figura: Função potência do teste de hipóteses.

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

Teste de hipóteses: 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.*

Teste de hipóteses: p-value

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. O valor-p corresponde à probabilidade de se observar um valor igual ou mais "extremo" do que o observado, w_{obs}, se H_0 é verdadeira. O cálculo desta probabilidade depende do tipo de região de rejeição da hipótese H_0, conforme indicado na seguinte tabela:

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

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. $H_0 : \sigma^2 = \sigma_0^2 \quad vs. \quad H_1 : \sigma^2 \neq \sigma_0^2$ (teste bilateral);

ou

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

ou

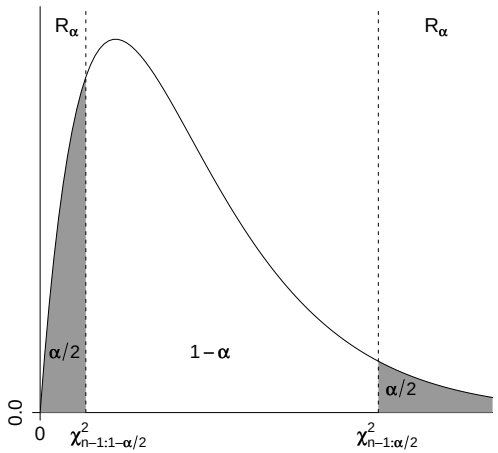
3. $H_0 : \sigma^2 \geq \sigma_0^2 \quad vs. \quad H_1 : \sigma^2 < \sigma_0^2$ (teste unilateral esquerdo);

Estatística de teste:

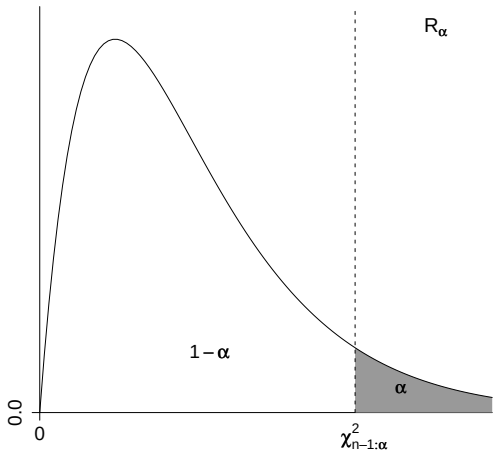
$$X^2 = \frac{(n-1)S^2}{\sigma_0^2} \overset{\text{sob } H_0,\ \sigma = \sigma_0}{\sim} \chi^2_{n-1}$$

Teste de hipóteses para a variância σ^2 populacional

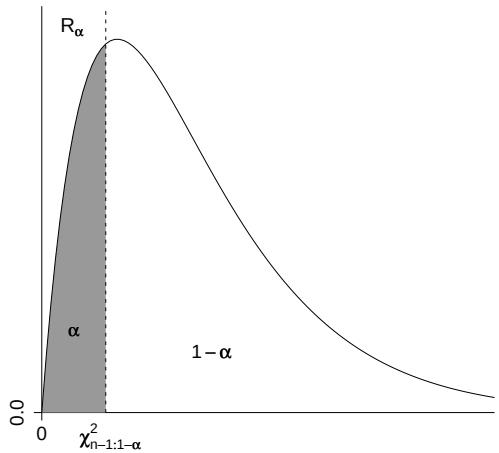
Região de rejeição do teste, para um nível de significância α:



(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:

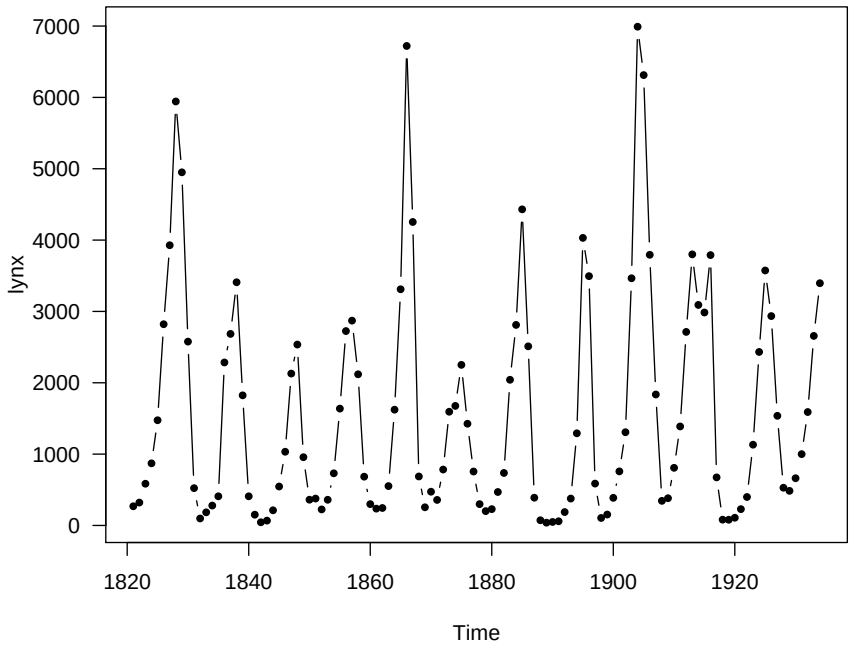
$$Z = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}} \underset{\text{sob } H_0, \text{ com } p = p_0}{\overset{\text{a}}{\sim}} N(0,1)$$

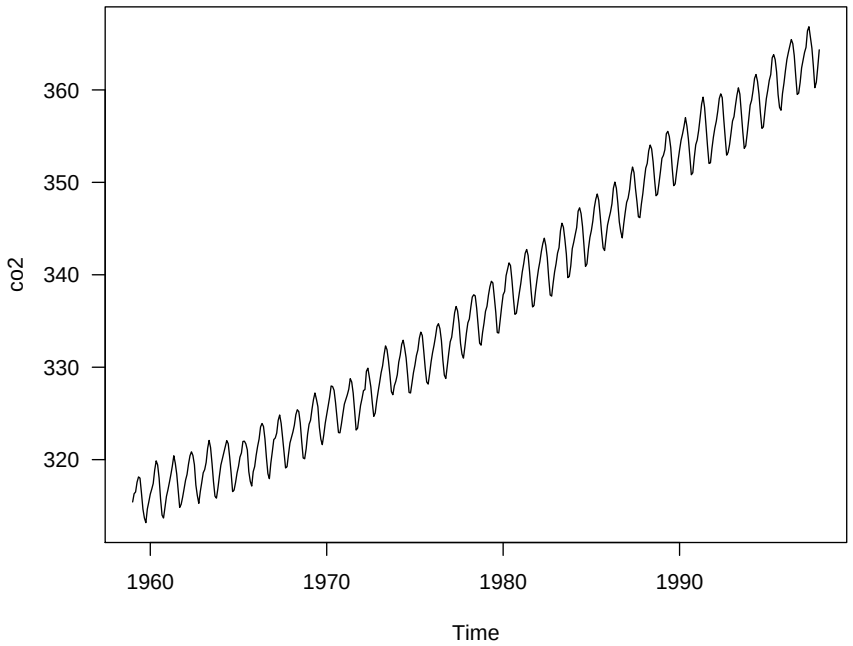
Região de rejeição do teste, para um nível de significância α pré-especificado:

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);

Teste das sequências ascendentes e descendentes

Alguns exemplos de sequências em que não há necessidade de usar o teste:





Esquerda: Annual numbers of lynx trappings for 1821–1934 in Canada. Taken from Brockwell and Davis (1991).
Direita: Mauna Loa (Hawaii) Atmospheric CO2 Concentration in parts per million (ppm): 468 monthly observations from 1959 to 1997.

Teste das sequências ascendentes e descendentes

Considere as hipóteses:

$$H_0 : \text{A amostra é aleatória} \quad vs. \quad H_1 : \text{A amostra não é aleatória}$$

- Vamos substituir pelo símbolo "$+$" cada observação precedida por uma de valor inferior, e pelo símbolo "$-$" cada observação precedida por outra de valor superior. Observações precedidas por outras de valor igual são desprezadas (corrige-se a dimensão da amostra, n).
- A estatística de teste é:

$$Z = \frac{V - \frac{2n-1}{3}}{\sqrt{\frac{16n-29}{90}}} \underset{sob\, H_0}{\overset{a}{\sim}} N(0,1).$$

  com V = número de sequências de sinais "$+$" e "$-$". Na prática considera-se que a distribuição de Z é razoável se $n \geq 26$.
- Rejeitamos H_0, ao nível de significância α, se $z_{obs} \in R_\alpha$, com

$$R_\alpha =]-\infty, -z_{\alpha/2}[\cup]z_{\alpha/2}, +\infty[$$

Teste das sequências ascendentes e descendentes

Exemplo

Consideremos uma amostra de medições de acidez (pH) de amostras de chuva, registadas em 30 locais de uma região industrial,

5.1	5.0	4.6	3.8	4.8	3.6	4.7	3.8	4.1	4.2	4.4	4.5	3.6	3.9	4.2
4.3	4.4	4.5	4.9	4.7	4.8	4.1	4.2	4.5	4.5	4.6	4.4	4.1	4.6	4.5

e testemos se estas observações constituem uma amostra aleatória.

Sinais:

.	−	−	−	+	−	+	−	+	+	+	+	−	+	+
+	+	+	+	−	+	−	+	+	0	+	−	−	+	−

Temos então $v_{obs} = 15$.

Teste de ajustamento do Qui Quadrado

Em muitas situações a distribuição da população é desconhecida, e podemos estar interessados em testar se a população tem distribuição F.

Hipóteses:

$$H_0 : X \sim F \quad vs. \quad H_1 : X \nsim F$$

Os dados (amostra) são divididos em k classes. A estatística de teste é:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{\text{a}}{\underset{\text{sob } H_0}{\sim}} \chi^2_{k-p-1},$$

onde

- O_i: nº de observações na classe i;
- E_i: número de observações que esperamos encontrar na classe i sob H_0;
- k: nº de classes;
- p: nº de parâmetros estimados, do modelo considerado em H_0.

Teste de ajustamento do Qui Quadrado

Nota:

$$\boxed{E_i = n \times p_i}$$

com

$$\boxed{p_i = P(X \in classe\ i \,|\, H_0\ verdadeira)}$$

Observação: Se houver valores E_i inferiores a 5, tipicamente correspondendo às classes dos extremos, essas classes devem ser agrupadas até o correspondente novo número esperado E_i (dado pelas somas dos correspondentes antigos $E_i's$) ultrapassar 5. Os correspondentes O_i's devem nesse caso ser também somados, diminuindo naturalmente o valor do número de classes k.

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

$$R_\alpha =]\chi^2_{k-p-1,\alpha};\ +\infty[$$

Exemplo: Geneticistas pensam que, em determinada população, a distribuição de probabilidade dos grupos sanguíneos é a seguinte:

$$\begin{cases} MM & MN & NN \\ 0.3 & 0.5 & 0.2 \end{cases}$$

Uma amostra de 200 indivíduos desta população, classificados de acordo com estes grupos sanguíneos, revelou 64 indivíduos do grupo *MM*, 96 do grupo *MN* e os restantes do grupo *NN*.
Estes dados fornecem evidência estatística para pôr em causa o pressuposto dos geneticistas? Considere $\alpha = 0.05$.

Nota: $\chi^2_{obs} = 0.427$

Exemplo: Pensa-se que o número de defeitos encontrados em circuitos eléctricos tem distribuição Poisson. Recolheu-se uma amostra aleatória de $n = 60$ circuitos e observaram-se os seguintes números de defeitos:

nº de defeitos	0	1	2	3
nº de circuitos	32	15	9	4

Pretendemos testar

$$H_0 : X \sim P(\lambda) \quad vs. \quad H_1 : X \nsim P(\lambda).$$

Como λ é desconhecido, terá de ser estimado. Assim, $\hat{\lambda} = \bar{x} = 0.75$ é a estimativa de λ.

nº de defeitos	p_i	E_i
0	0.472	28.32
1	0.354	21.24
2	0.133	7.98
3 (ou mais)	0.041	**2.46**

nº de defeitos	O_i	p_i	E_i
0	32	0.472	28.32
1	15	0.354	21.24
2 (ou mais)	13	0.174	10.44

$$\chi^2_{obs} = \frac{(32-28.32)^2}{28.32} + \frac{(15-21.24)^2}{21.24} + \frac{(13-10.44)^2}{10.44} = 2.94$$

$$\chi^2_{obs} = 2.94 \notin]\chi^2_{1,0.05}, \infty[=]3.84, \infty[$$

Teste de ajustamento do Qui Quadrado

Exemplo (distribuição normal): Os artigos produzidos em determinada fábrica são sujeitos a um controle de qualidade, resultando num índice de qualidade, X. De forma a avaliar essa qualidade recolheu-se uma amostra aleatória de 46 artigos da produção, tendo-se medido os valores seguintes do referido índice:

100, 110, 122, 132, 99, 96, 88, 75, **45**, 154, 153, **161**, 142, 99, 111, 105, 133, 142, 150, 153, 121, 126, 117, 97, 105, 117, 125, 105, 94, 90, 80, 50, 55, 102, 122, 136, 75, 104, 109, 108, 134, 135, 111, 78, 89, 154

Vamos usar estes dados para testar, ao nível de significância 5%,

$$H_0 : X \sim N(\mu, \sigma^2) \quad vs. \quad H_1 : X \nsim N(\mu, \sigma^2)$$

A estatística de teste é, $X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{\text{sob } H_0}{\sim} \chi^2_{k-p-1}$

Como não conhecemos os valores populacionais de μ e σ^2, vamos estimá-los:

$$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{46} x_i}{46} = \frac{5109}{46} = 111.0652;$$

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{46} x_i^2 - 46 \times \bar{x}^2}{46-1} = \frac{325117}{414} = 785.3068$$

Teste de ajustamento do Qui Quadrado

Pela regra de Sturges o número **aconselhado** de classes a considerar é:

$$k \approx 1 + log(n)/log(2) = 1 + log(46)/log(2) \approx 6.523562$$

Consideramos $k = 7$. A amplitude de cada classe é aproximadamente

$$\frac{L}{7} = \frac{161-45}{7} \approx 16.571$$

Vamos aproximar este valor a 17, ou seja, considerar as classes:
$]-\infty; 62]$ $]62; 79]$ $]79; 96]$ $]96; 113]$ $]113; 130]$ $]130; 147]$ $]147; \infty[$

Devemos contar quantas observações estão em cada um dos intervalos anteriores, para obter os valores de O_i,

classe	O_i
$]-\infty; 62]$	3
$]62;\ 79]$	3
$]79;\ 96]$	6
$]96;\ 113]$	14
$]113;\ 130]$	7
$]130;\ 147]$	7
$]147;\ \infty[$	6

Teste de ajustamento do Qui Quadrado

Vamos determinar os valores de p_i e $E_i = n \times p_i = 46 \times p_i$.

	Classe	O_i	p_i	E_i
1	$]-\infty; 62]$	3	0.0400	**1.839**
2	$]62; 79]$	3	0.0863	**3.969**
3	$]79; 96]$	6	0.1692	7.782
4	$]96; 113]$	14	0.2321	10.676
5	$]113; 130]$	7	0.2229	10.251
6	$]130; 147]$	7	0.1498	6.889
7	$]147; \infty[$	6	0.0999	**4.594**

	Classe	O_i	p_i	E_i
1	$]-\infty; 79]$	6	0.1263	5.808
2	$]79; 96]$	6	0.1692	7.782
3	$]96; 113]$	14	0.2321	10.676
4	$]113; 130]$	7	0.2229	10.251
5	$]130; \infty[$	13	0.2496	11.483

A estatística de teste é, $X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \overset{\text{sob } H_0}{\sim} \chi^2_{k-p-1} = \chi^2_2$

Regra de decisão do teste: Rejeitar H_0 ao nível de significância 5% se $x^2_{obs} \in R_{0.05} \equiv]5.99, +\infty[$. Como $x^2_{obs} = 2.68$ não rejeitamos, ao nível de significância de 5% a hipótese da população ter distribuição Normal.

Observação: O último exercício também pode ser resolvido considerando todas as classes com o mesmo valor $p_i = \frac{1}{k}$ (neste caso $p_i = \frac{1}{7}$, $i = 1, \ldots, 7$).

A 1ª classe é $]-\infty, a_1]$ onde $P(x \leq a_1) = \Phi(\frac{a_1 - 111.07}{28.02}) = \frac{1}{7} = 0.1429$.
Resulta que $\frac{111.07 - a_1}{28.02} = \Phi^{-1}(0.8571) \approx 1.07 \Leftrightarrow a_1 = 81.1$.

Nota: Os pontos extremos das restantes classes obtêm-se de modo análogo.

i	Classe	O_i	p_i	E_i
1	$]-\infty; 81.1]$	7	$1/7$	6.571
2	$]81.1; 95.2]$	4	$1/7$	6.571
3	$]95.2; 106]$	10	$1/7$	6.571
4	$]106; 116.1]$	5	$1/7$	6.571
5	$]116.1; 126.9]$	7	$1/7$	6.571
6	$]126.9; 141]$	5	$1/7$	6.571
7	$]141; +\infty[$	8	$1/7$	6.571

Decisão: Rejeitar H_0 ao nível de significância 5% se $x^2_{obs} \in R_{0.05} =]9.49, +\infty[$.
Como $x^2_{obs} = 3.91$ não rejeitamos, ao nível de significância de 5% a hipótese da população ter distribuição Normal.

Teste de ajustamento do Qui Quadrado

Exercício (2° teste de Probabilidades e Estatística - 25 de Junho de 2008) Os seguintes valores (ordenados) resultaram da implementação de um algoritmo, numa linguagem de programação. Podemos afirmar que o algoritmo gera valores da distribuição $U(0,1)$?

0.089	0.139	0.140	0.208	0.325	0.412	0.447	0.449	0.462
0.516	0.601	0.614	0.628	0.644	0.647	0.683	0.692	0.798
0.810	0.810	0.850	0.875	0.883	0.887	0.909	0.944	0.996

Nota: Resolva o exercício considerando 5 classes, com igual probabilidade $p_i = 0.2, \quad i = 1, 2, \ldots, 5.$

Nota: $\chi^2_{obs} = 6.89$

[Para alunos da Licenciatura em Bioquímica & em Química Aplicada]

- A nossa capacidade para observar o mundo real não é perfeita. Muitas observações que fazemos não representam exactamente o processo que estamos a observar. Matematicamente temos,

$$\boxed{\text{valor real} = \text{medição experimental} \pm \text{erro.}}$$

- Nas medições experimentais o erro (experimental) pode não se referir a um equívoco, mas a uma imprecisão resultante do processo de medição. Para podermos garantir conclusões adequadas devemos estudar os erros experimentais e o efeito da sua propagação.

- **Exemplo:** Considere duas substâncias, com massa molecular (relativa à unidade de massa atómica u) iguais a $18.2u$ e $18.5u$, respectivamente. Podemos concluir que a segunda substância tem massa molecular superior. Contudo se também nos disserem o erro das medições é $\pm 0.5u$, já não conseguimos determinar qual a substância com maior massa molecular.

[Para alunos da Licenciatura em Bioquímica & em Química Aplicada]

Nas medições podemos ter dois tipos de erros experimentais:

- **erros sistemáticos ou fortuitos** – afectam o resultado, mas podem ser eliminados ou corrigidos.
  Estes erros podem resultar de muitos factores, por exemplo, defeitos ou má calibração da instrumentação utilizada, aplicação errada dos modelos teóricos, ...

- **erros aleatórios** – erros que não se podem controlar.
  Se o erro é aleatório (não enviesado) e uma medição experimental puder ser obtida várias vezes (nas mesmas condições), podemos estudar o erro aleatório através de análise estatística . Por exemplo, através do cálculo da média e do erro padrão.

Exemplo de medição com erro aleatório: Medição do diâmetro duma bola de ténis, com uma régua.

[Para alunos da Licenciatura em Bioquímica & em Química Aplicada]

Apresentação dos resultados

Quando obtemos medições, é usual apresentar os resultados na forma

$$\boxed{\text{medição} = \text{“melhor estimativa”} \pm \text{incerteza.}}$$

Exemplo: a medida $6.03g \pm 0.02g$ indica que estamos confiantes que o valor real da quantidade medida se situa entre $6.01g$ e $6.05g$.

[Para alunos da Licenciatura em Bioquímica & em Química Aplicada]

Se uma medida (sujeita a erro aleatório) é obtida a partir duma observação, podemos apresentar o resultado na forma $\boxed{x \pm \triangle x}$. Caso seja possível obter várias medições da mesma quantidade, os resultados podem ser apresentados na forma:

$$\overline{x} \pm \triangle \overline{x}$$

- $\overline{x}$ é a média amostral, uma estimativa do valor médio de X;
- $\triangle \overline{x}$ pode, por exemplo, ser determinado através erro padrão de $\overline{X}$, ou seja, $\triangle \overline{x} = \sigma/\sqrt{n}$. Se σ for desconhecido podemos estimar o erro padrão através do seu estimador, $S/\sqrt{n}$.

Se a população tiver distribuição normal, esperamos ter aproximadamente 68% dos valores observados no intervalo $\overline{x} \pm \triangle x$.

Se quisermos apresentar um resultado mais preciso, então devemos estimar a medida através dum intervalo de confiança, com um nível de confiança superior a 68% ($\triangle \overline{x} = z_{\alpha/2}\sigma/\sqrt{n}$ ou $\triangle \overline{x} = t_{n-1,\alpha/2}S/\sqrt{n}$ ou $\triangle \overline{x} = z_{\alpha/2}S/\sqrt{n}$).

[Para alunos da Licenciatura em Bioquímica & em Química Aplicada]

Exemplo

Para se determinar o pH da água duma piscina foram efectuadas 7 medições, que forneceram os seguintes resultados:

$$7.18, \quad 7.26, \quad 7.15, \quad 7.16, \quad 7.26, \quad 7.19$$

Temos $\bar{x} = 7.20$ e $\frac{s}{\sqrt{6}} = 0.02$.

Se admitirmos que a população tem distribuição normal, o intervalo de 95% de confiança do pH médio é $7.20 \pm (2.57 \times 0.02)$.

[Para alunos da Licenciatura em Bioquímica & em Química Aplicada]

Propagação dos erros

Depois de recolhermos **diferentes medições** experimentais, por vezes elas são combinadas de acordo com uma fórmula, de modo a obtermos o resultado desejado.

Para determinarmos o valor da incerteza do resultado, precisamos de saber como combinar a incerteza de cada medição.
Considere sem perda de generalidade, Z dependente de X e Y, isto é $Z = f(X, Y)$. Então, se considerarmos que X e Y são independentes, o valor de Z pode ser apresentado na forma

$$z \pm \triangle z$$

com $z = f(x, y)$ e $\triangle z = \sqrt{\left(\frac{\partial f}{\partial x}(x, y)\right)^2 (\triangle x)^2 + \left(\frac{\partial f}{\partial y}(x, y)\right)^2 (\triangle y)^2}$.

[Para alunos da Licenciatura em Bioquímica & em Química Aplicada]

Exemplo: Suponha que obteve duas medições, $5.32 \pm 0.02 cm$ e $0.103 \pm 0.001 s$, associadas às grandezas X e Y, respectivamente. Dada a grandeza G,

$$G = f(X, Y) = \frac{2X}{Y^2},$$

o seu valor poderá ser apresentado na forma $g \pm \triangle g$ com

$$g = 2 \times 5.32/0.103^2 = 1002.92\, cm/s^2$$

e

$$\triangle g = \sqrt{\left(\frac{\partial f}{\partial x}(x,y)\right)^2 (\triangle x)^2 + \left(\frac{\partial f}{\partial y}(x,y)\right)^2 (\triangle y)^2}$$

$$= \sqrt{\left(\frac{2}{y^2}\right)^2 (\triangle x)^2 + \left(-\frac{4x}{y^3}\right)^2 (\triangle y)^2} \underset{\substack{\text{fazendo } x=5.32, \triangle x=0.02, \\ y=0.103, \triangle y=0.001}}{=} 19.84\, cm/s^2$$

Regressão

É uma técnica estatística que permite estudar a relação entre uma **variável resposta** Y (também designada por **variável dependente**) e uma ou mais **variáveis explicativas**, $x_1, x_2, \ldots$ (também designadas por **variáveis independentes**).

Equação de regressão: Modelo matemático que relaciona as variáveis.

Regressão linear simples: modelo com uma variável dependente Y, uma variável independente x e a equação de regressão é linear, isto é,

$$\boxed{Y = \beta_0 + \beta_1 x + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2).}$$

Exemplo com dados reais:

- Y - comprimento da tíbia de crianças entre 7-12 anos;
- x - comprimento da perna (medido com fita métrica) de crianças entre 7-12 anos.

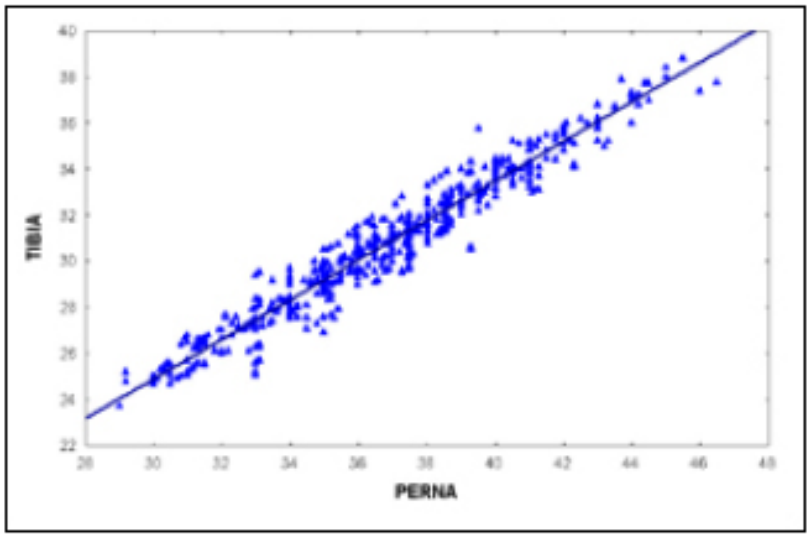


Figura: Diagrama de dispersão da medida radiográfica da tíbia, em função da medida clínica da perna (em cm) e recta de regressão linear estimada. Figura retirada de ACTA FISIATR 2007; 14(2): 95 - 99.

$$\boxed{Y = \beta_0 + \beta_1 x + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2).}$$

Observações:

1. $\beta_0 + \beta_1 x$ é a componente determinística do modelo;
2. $\varepsilon \sim N(0, \sigma^2)$ é um erro aleatório;
3. Os parâmetros β_0 e β_1 terão de ser estimados a partir dos dados.
4. $Y = \beta_0 + \beta_1 x + \varepsilon$ é também **variável aleatória**. Isto é,

$$\boxed{Y|x \sim N(\beta_0 + \beta_1\ x, \sigma^2)}$$

Cálculo do valor médio e váriância:
$E(Y|x) = E(\beta_0 + \beta_1\ x + \varepsilon|x) = \beta_0 + \beta_1\ x + E(\varepsilon) = \beta_0 + \beta_1\ x + 0 = \beta_0 + \beta_1\ x$
$V(Y|x) = V(\beta_0 + \beta_1\ x + \varepsilon|x) = V(\varepsilon) = \sigma^2$

Regressão Linear Simples: Estimadores dos parâmetros

Suponha que se observam um conjunto de n valores da variável resposta Y e da variável independente x,

$$(x_i, Y_i), \quad i = 1, 2, \ldots, n,$$

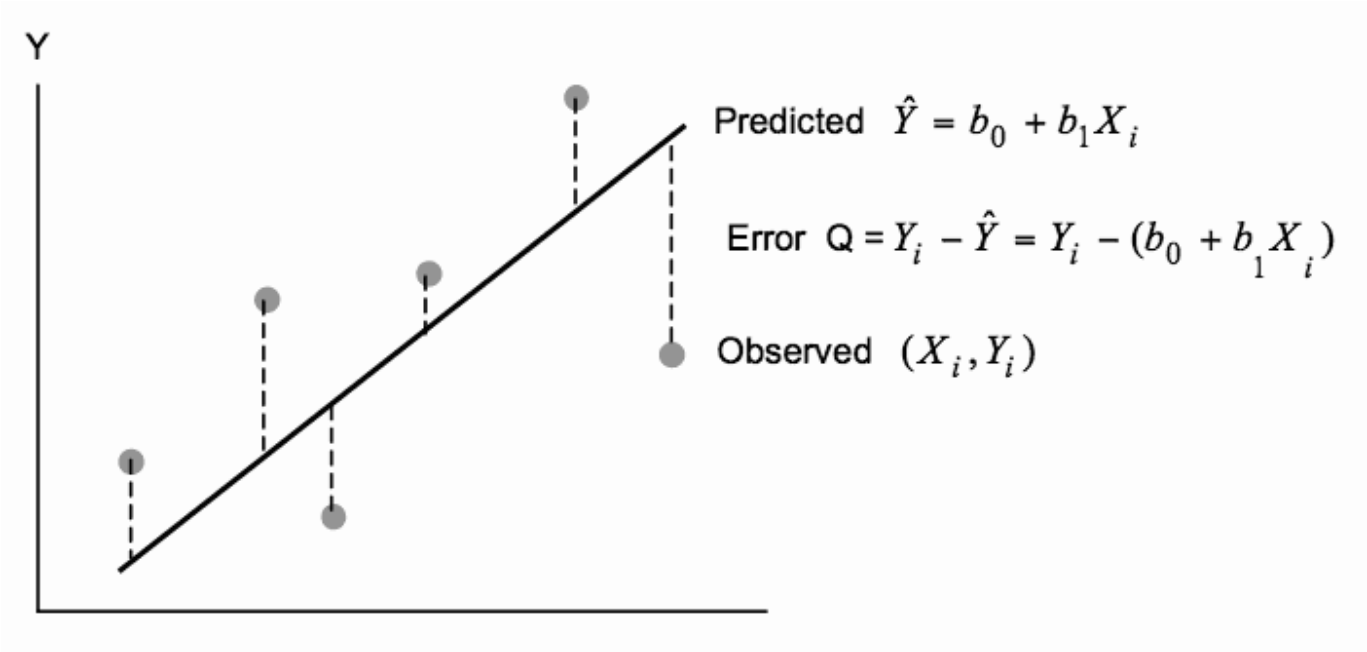
e que se pretendem usar estes valores para estimar os parâmetros β_0 e β_1 do modelo de regressão linear simples

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \ldots, n.$$

Vamos assumir que os erros aleatórios ε_i, para cada elemento amostral Y_i, são **independentes** seguindo todos a mesma distribuição $N(0, \sigma^2)$.

Regressão Linear Simples: Estimadores dos parâmetros

Assim, queremos determinar estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$, dos coeficientes da recta de regressão β_0 e β_1, respectivamente, para obtermos a **recta estimada** $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$.



Aos desvios das n observações da amostra, $\varepsilon_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, $i = 1, 2, \ldots, n$, damos o nome de **resíduos**.

Regressão Linear Simples: Estimadores dos parâmetros

Usando o **método dos mínimos quadrados**, os estimadores $\hat{\beta}_0$ e β_1 devem ser obtidos de modo a minimizar a soma do quadrado dos resíduos,

$$\boxed{SQ = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 \, x_i)^2.}$$

A minimização é conseguida resolvendo, em ordem a β_0 e β_1, as equações,

$$\left\{ \begin{array}{l} \frac{\partial \ SQ}{\partial \beta_0} = 0 \\ \\ \frac{\partial \ SQ}{\partial \beta_1} = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} -2 \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 \ x_i) = 0 \\ \\ -2 \sum x_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 \ x_i) = 0 \end{array} \right. \Leftrightarrow$$

$$\left\{ \begin{array}{l} \sum Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i \\ \\ \sum x_i Y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \bar{x} \\ \\ \hat{\beta}_1 = \frac{\sum x_i Y_i - n\bar{x}\overline{Y}}{\sum x_i^2 - n\bar{x}^2} \end{array} \right.$$

Regressão Linear Simples: Estimadores dos parâmetros

Observação: De modo a simplificar a notação, podemos escrever:

$$\boxed{\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}}} \qquad \boxed{\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x},}$$

com

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2;$$

$$S_{xY} = \sum_{i=1}^{n} (Y_i - \overline{Y})(x_i - \overline{x}) = \sum_{i=1}^{n} Y_i (x_i - \overline{x}) = \sum_{i=1}^{n} x_i Y_i - n\overline{x}\overline{Y}.$$

Aviso:

Só devemos fazer estimação, através da recta de regressão, para valores x que estejam dentro do intervalo das observações obtidas para essa variável.

Qualidade do Ajuste

Podemos escrever,

$$SQ_R = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = S_{YY} - \frac{S_{xY}^2}{S_{xx}} = S_{YY} - \hat{\beta}_1^2 S_{xx},$$

com $S_{YY} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$.

Definição (Coeficiente de Determinação)

$$R^2 \quad = \quad 1 - \frac{SQ_R}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = \hat{\beta}_1^2 \frac{S_{xx}}{S_{YY}} = \frac{S_{xY}^2}{S_{xx} S_{YY}} \qquad (0 \leq R^2 \leq 1).$$

Esta medida compara a soma de quadrados dos resíduos (SQ_R) do modelo de regressão linear simples (RLS) com a SQ_R do modelo de RLS com $\beta_1 = 0$. Quanto mais próximo estiver de 1, maior a importância de x na determinação do valor de Y. Consideramos o ajustamento razoável se $R^2 \geq 0.8$.

Temos que

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \qquad \hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{nS_{xx}}\sum_{i=1}^{n} x_i^2\right)$$

Justificação:

- $\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} Y_i$ é uma **combinação linear** de v.a.'s Y_i independentes e com distribuição **normal**. Logo $\hat{\beta}_1$ tem distribuição **normal**.
- $\mathrm{E}(\hat{\beta}_1) = \mathrm{E}\left(\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} Y_i\right) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} \mathrm{E}(Y_i) = \ldots = \frac{\beta_1 S_{xx}}{S_{xx}} = \beta_1.$
- $\mathrm{V}(\hat{\beta}_1) = \mathrm{V}\left(\frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{S_{xx}}\right) \underset{Y_i's \ indep.}{=} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \mathrm{V}(Y_i)}{S_{xx}^2} = \frac{S_{xx}}{S_{xx}^2}\sigma^2 = \frac{\sigma^2}{S_{xx}}.$

Regressão Linear Simples: Propriedades dos estimadores

- $\text{Cov}(\overline{Y}, \hat{\beta}_1) = \text{Cov}(\overline{Y}, \frac{S_{xY}}{S_{xx}}) = \frac{1}{nS_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) \text{Cov}\,(Y_i, Y_i) = 0$
- Como $\overline{Y}$ e $\hat{\beta}_1$ são **independentes** e têm distribuição **normal**, então $\hat{\beta}_0$ também tem distribuição **normal**.
- $\text{E}(\hat{\beta}_0) = \text{E}(\overline{Y}) - \text{E}(\hat{\beta}_1)\overline{x} = \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} = \beta_0$
-

$$\text{V}(\hat{\beta}_0) = \text{V}(\overline{Y}) + \overline{x}^2 \text{V}(\hat{\beta}_1) - 2\overline{x}\,\text{Cov}(\overline{Y}, \hat{\beta}_1) = \frac{\sigma^2}{n} + \overline{x}^2 \frac{\sigma^2}{S_{xx}} - 0 =$$

$$= \frac{\sigma^2}{n} \left(1 + \frac{n\overline{x}^2}{S_{xx}}\right) = \frac{\sigma^2}{nS_{xx}} \left(S_{xx} + n\overline{x}^2\right) = \frac{\sigma^2}{nS_{xx}} \left(\sum_{i=1}^{n} x_i^2\right)$$

Regressão Linear Simples: Estimação de σ^2

- Como $E(SQ_R) = (n-2)\sigma^2$,

$$\boxed{\hat{\sigma}^2 = \frac{SQ_R}{n-2}}$$

é um **estimador** (centrado) de σ^2.

-

$$\boxed{\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{SQ_R}{\sigma^2} \sim \chi^2_{n-2}}$$

Como geralmente não conhecemos o valor de σ^2, vamos ainda necessitar dos seguintes resultados

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} = \sqrt{S_{xx}}\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}} \sim t_{n-2}$$

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\hat{\sigma}^2}{nS_{xx}}\sum_{i=1}^{n} x_i^2}} = \sqrt{\frac{nS_{xx}}{\sum_{i=1}^{n} x_i^2}}\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}} \sim t_{n-2}$$

Intervalo de Confiança para β_1

Temos que

$$P\left(-t_{n-2:\alpha/2} \leq \sqrt{S_{xx}}\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}} \leq t_{n-2:\alpha/2}\right) = \ldots =$$

$$= P\left(\hat{\beta}_1 - t_{n-2:\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{n-2:\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right) = 1 - \alpha$$

Intervalo de Confiança a $(1-\alpha)\times 100\%$ para β_1

$$IC_{(1-\alpha)\times 100\%}(\beta_1) \equiv \left[\hat{\beta}_1 - t_{n-2:\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}; \hat{\beta}_1 + t_{n-2:\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right]$$

Teste de Hipóteses para β_1

1. Hipóteses:
   - $H_0 : \beta_1 = a \;\; vs. \;\; H_1 : \beta_1 \neq a$ (teste bilateral);
   - $H_0 : \beta_1 \leq a \;\; vs. \;\; H_1 : \beta_1 > a$ (teste unilateral direito);
   - $H_0 : \beta_1 \geq a \;\; vs. \;\; H_1 : \beta_1 < a$ (teste unilateral esquerdo);

2. Estatística de teste:

$$T = \sqrt{S_{xx}}\frac{\hat{\beta}_1 - a}{\hat{\sigma}} \overset{\text{sob } H_0\ (\beta_1 = a)}{\sim} t_{n-2}$$

3. Região de rejeição do teste, para um nível de significância α:
   - $R_\alpha =]-\infty; -t_{n-2:\alpha/2}[\cup]t_{n-2:\alpha/2}; +\infty[$ (teste bilateral);
   - $R_\alpha =]t_{n-2:\alpha}; +\infty[$ (teste unilateral direito);
   - $R_\alpha =]-\infty; -t_{n-2:\alpha}[$ (teste unilateral esquerdo);

4. Decisão: Rejeitar H_0 ao nível de significância α se $t_{obs} \in R_\alpha$.

Intervalo de Confiança para β_0

Com base nas seguintes igualdades

$$P\left(-t_{n-2:\alpha/2} \leq \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2 \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}}} \leq t_{n-2:\alpha/2}\right) = \ldots =$$

$$= P\left(\hat{\beta}_0 - t_{n-2:\alpha/2}\sqrt{\hat{\sigma}^2 \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}} \leq \beta_0 \leq \hat{\beta}_0 + t_{n-2:\alpha/2}\sqrt{\hat{\sigma}^2 \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}}\right)$$

$$= 1 - \alpha,$$

obtemos o seguinte intervalo de confiança

Intervalo de Confiança a $(1-\alpha)\times 100\%$ para β_0

$$IC_{(1-\alpha)\times 100\%}(\beta_0) \equiv \left[\hat{\beta}_0 - t_{n-2:\alpha/2}\sqrt{\hat{\sigma}^2 \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}}; \hat{\beta}_0 + t_{n-2:\alpha/2}\sqrt{\hat{\sigma}^2 \frac{\sum_{i=1}^n x_i^2}{nS_{xx}}}\right]$$

Intervalo de confiança para σ^2

Sabendo que

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

$$= P\left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;\alpha/2}} \leq \sigma^2 \leq \frac{(n-2)\hat{\sigma}^2}{\chi^2_{n-2;1-\alpha/2}}\right) = 1-\alpha,$$

obtemos os seguintes intervalos de confiança

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

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

Testes de hipóteses para σ^2

Hipóteses:

1. $H_0 : \sigma = \sigma_0 \;\; vs \;\; H_1 : \sigma \neq \sigma_0$ (teste bilateral);
2. $H_0 : \sigma \leq \sigma_0 \;\; vs \;\; H_1 : \sigma > \sigma_0$ (teste unilateral direito);
3. $H_0 : \sigma \geq \sigma_0 \;\; vs \;\; H_1 : \sigma < \sigma_0$ (teste unilateral esquerdo);

Estatística de teste:

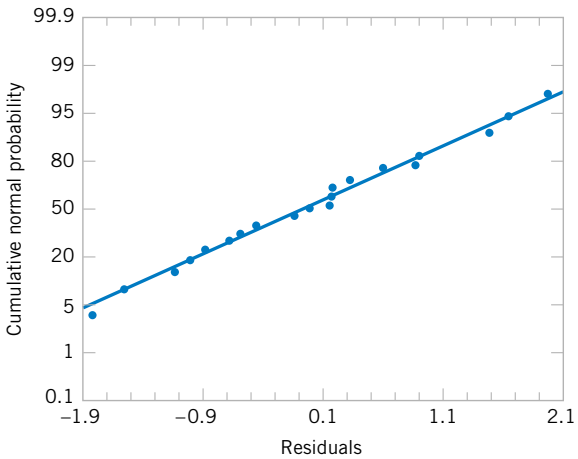
$$X^2 = \frac{SQ_R}{\sigma_0^2} = \frac{(n-2)\hat{\sigma}^2}{\sigma_0^2} \overset{\text{sob } H_0}{\sim} \chi^2_{n-2}$$

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

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

Análise (informal) dos resíduos.

É importante verificar se os resíduos da regressão, $\varepsilon_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$, verificam os pressupostos do modelo (ε_i iid, $\varepsilon_i \sim N(0, \sigma^2)$). Podemos realizar teste estatísticos para verificar os pressupostos, ou validar de modo informal através de representação gráfica.



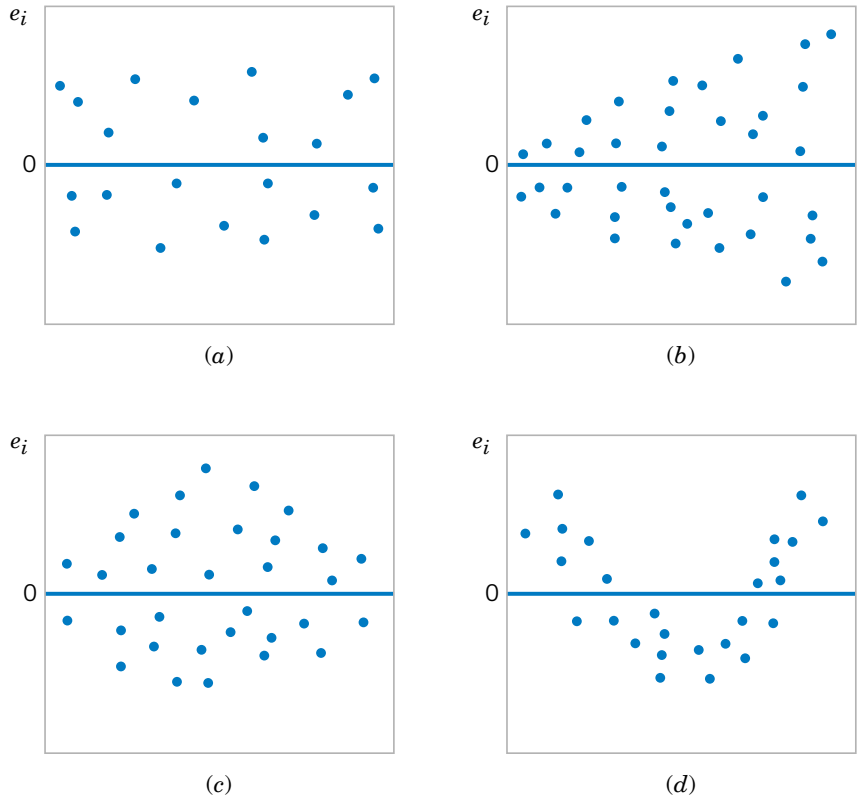


Figura: Esquerda: Validação informal do modelo normal; Direita: Gráfico dos Resíduos contra $\hat{y}$ (ou x).

Resumo dos resultados apresentados

$\hat{\sigma}^2 = \frac{SQ_R}{n-2}$	$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} Y_i$	$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$
$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{SQ_R}{\sigma^2} \sim \chi^2_{n-2}$	$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$	$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} x_i^2\right)$
	$Z = \sqrt{S_{xx}} \frac{\hat{\beta}_1 - \beta_1}{\sigma} \sim N(0,1)$	$Z = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} x_i^2}} \sim N(0,1)$
	$T = \sqrt{S_{xx}} \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}} \sim t_{n-2}$	$T = \sqrt{\frac{nS_{xx}}{\sum_{i=1}^{n} x_i^2}} \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}} \sim t_{n-2}$

$$IC_{(1-\alpha)\times 100\%}(\beta_1) = \left[\hat{\beta}_1 - t_{n-2;\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \; ; \; \hat{\beta}_1 + t_{n-2;\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right]$$

$$IC_{(1-\alpha)\times 100\%}(\beta_0) = \left[\hat{\beta}_0 - t_{n-2;\alpha/2}\sqrt{\hat{\sigma}^2 \frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}} \; ; \; \hat{\beta}_0 + t_{n-2;\alpha/2}\sqrt{\hat{\sigma}^2 \frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}}\right]$$

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

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