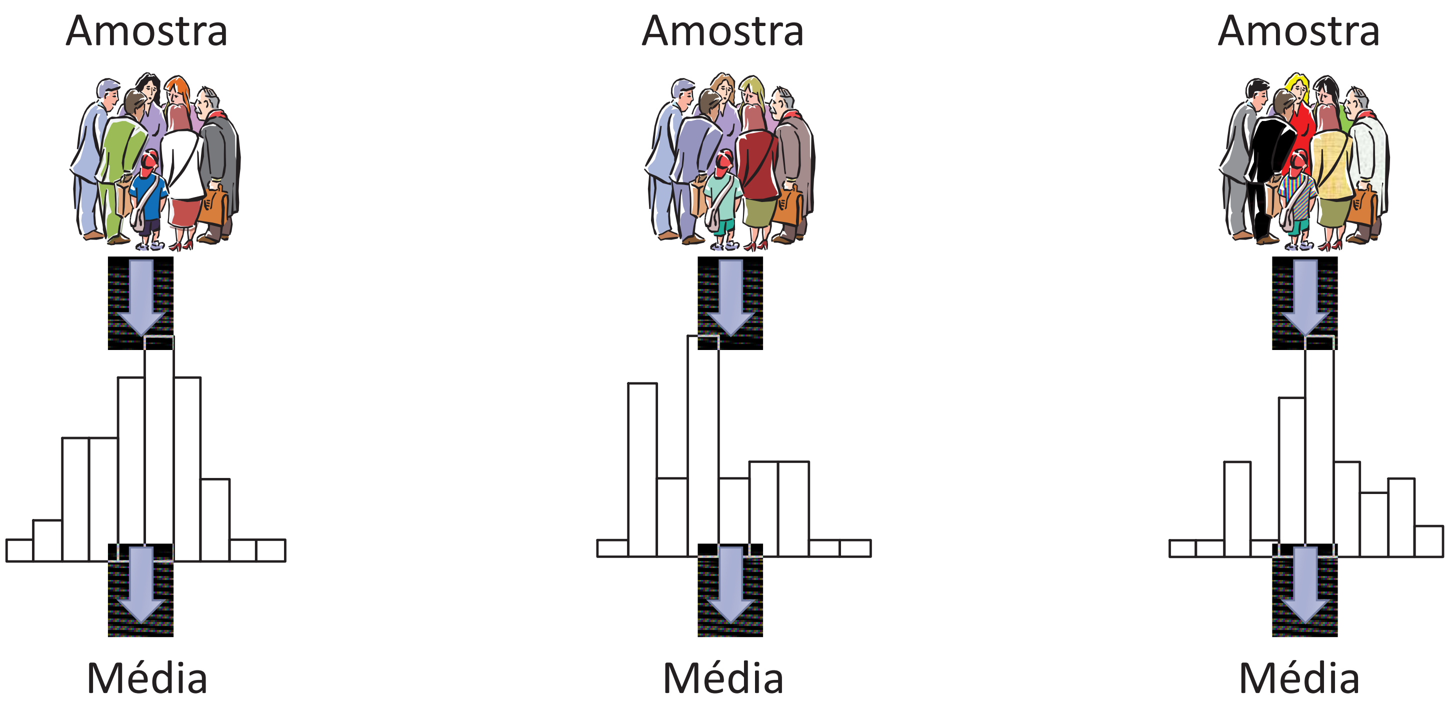
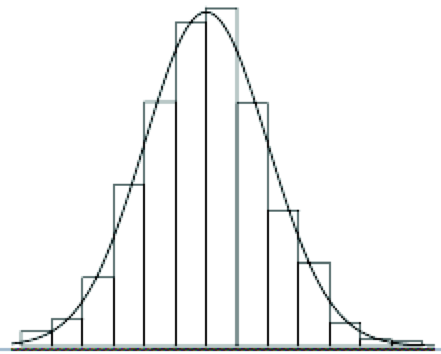
Distribuições amostrais

Distribuição amostral da média e da diferença de médias

Distribuição amostral



Distribuição amostral…



Adaptado de: http://www.webpages.uidaho.edu/css310/DATA/SAMPLE.pdf

Distribuições amostrais (1 população)

Estatística	σ^2 conhecido?	Tipo de população	Distribuição amostral
$\overline{X}$	Sim	Normal (ou qualquer se n grande*)	$Z = \dfrac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \cap N(0;1)$
	Não	Normal (ou qualquer se n grande*)	$T = \dfrac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \cap t_{(n-1)}$

*Nesse caso a distribuição amostral é aproximada em vez de exata.

Distribuições amostrais (2 populações)

Estatística	σ_1^2 e σ_2^2 conhecidos?	Tipo de populações	Distribuição amostral
$\bar{X}_1 - \bar{X}_2$	Sim	Normais (ou quaisquer se n_1 e n_2 grandes*)	$Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \cap N(0;1)$
	Não ($\sigma_1^2 = \sigma_2^2$)	Normais (ou quaisquer se n_1 e n_2 grandes*)	$T = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_1 - 1)S_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \cap t_{(n_1 + n_2 - 2)}$
	Não ($\sigma_1^2 \neq \sigma_2^2$)	Normais (ou quaisquer se n_1 e n_2 grandes*)	$Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \overset{\circ}{\cap} t_{(v)}$

v é obtido pela aproximação de Welch

Estimação

Processos de estimação

Estimador e Estimativa

Propriedades dos estimadores

Método da máxima verosimilhança

Intervalos de confiança

Processos de estimação

- Estimação **pontual**:
  - Produção de um valor, que se pretende que seja o melhor, para um determinado parâmetro da população, com base na informação amostral.
- Estimação **intervalar** (Intervalos de confiança):
  - Construção de um intervalo que, com certo grau de certeza previamente estipulado, contenha o verdadeiro valor do parâmetro da população.

Estimador e estimativa

- **Estimador** dum parâmetro da população:
  - É uma variável aleatória (v. a.) que depende da informação amostral e cujas realizações fornecem aproximações para o parâmetro desconhecido;
  - É uma função dos dados.

- **Estimativa**:
  - É um valor específico assumido pelo estimador para uma amostra em concreto;
  - É um número.
  - Amostras diferentes podem produzir estimativas diferentes, devido à aleatoriedade amostral, logo temos incerteza associada à estimativa obtida

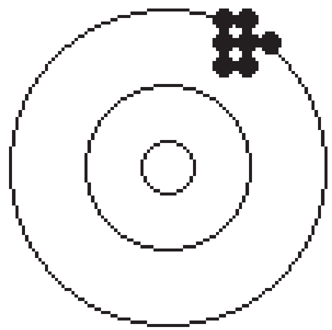
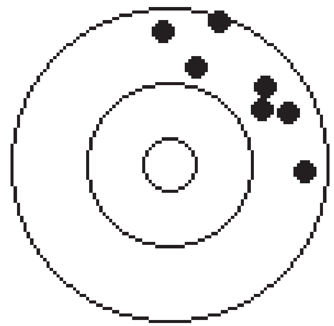
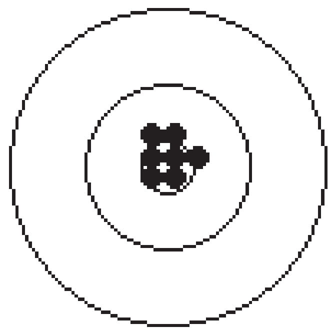
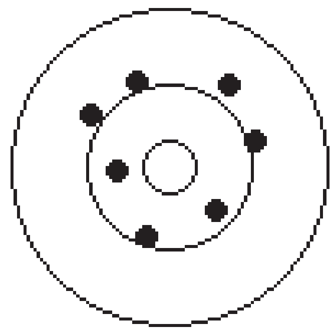
Estimador e estimativa

- Tanto o estimador como a estimativa representam-se habitualmente com um acento circunflexo sobre a letra do parâmetro $(\hat{\mu}, \hat{p}, \hat{\sigma}, \ldots)$.
- Exemplos:
  - Estimador da média populacional: $\hat{\mu} = \overline{X}$
  - Estimativa da média populacional: $\hat{\mu} = \overline{x} = 3$

Propriedades desejáveis nos estimadores

- **Não enviesamento** ou centragem:
  - em termos médios, o estimador atinge o valor real do parâmetro.
- **Eficiência:**
  - o estimador é mais eficiente quanto menor for a sua variância.
- **Suficiência:**
  - propriedade de retirar da amostra toda a informação relevante sobre o parâmetro.
- **Consistência:**
  - para *n* grande, o estimador deve ser aproximadamente igual ao parâmetro.

Propriedades dos estimadores

	Eficiente	Não eficiente
Enviesado		
Não enviesado		

Exercício 1

- Às 20:00 de 13 de Junho de 2004 era possível ler a seguinte notícia na SIConline:

  "O PS é o vencedor das eleições europeias em Portugal, segundo a previsão SIC/Eurosondagem. Com 44,1 por cento a 47,9 por cento dos votos, os socialistas conseguem eleger 12 a 13 eurodeputados. A abstenção atingiu os 64 por cento. A coligação "Força Portugal" obteve 29,7 por cento a 33,5 por cento dos votos, valores que correspondem a 8 a 9 lugares no Parlamento Europeu. A CDU terá conseguido entre 10,1 por cento e 11,9 por cento e 2 a 3 deputados. Por sua vez, o Bloco de Esquerda teve 5,1 por cento a 6,9 por cento dos votos, que valem 1 eurodeputado. Os votos noutros partidos estão entre 2,8 por cento a 4,2 por cento. Os votos brancos/nulos são 1,5 por cento a 2,3 por cento, de acordo com a projecção...."

  Dê um exemplo de uma estimativa:

  - pontual
  - intervalar

  apresentadas nesta notícia.

Exercício 2

- Qual das seguintes expressões está correta? Justifique.
  i. $E(\mu) = \hat{\mu}$
  ii. $E(\hat{\mu}) = \mu$

- Qual das seguintes expressões está correta? Justifique.
  i. $P(\mu \leq x) = P(\hat{\mu} \leq x)$
  ii. $P(\bar{X} \leq x) = P(\hat{\mu} \leq x)$

- Distinga estimador de estimativa.

Métodos de estimação

- Método dos momentos:
  - Método simples e muitas vezes intuitivo;
  - Igualam-se os momentos amostrais aos populacionais
- Método dos mínimos quadrados:
  - Usado nos modelos de regressão
- Método da máxima verosimilhança:
  - Método de estimação paramétrico (assumimos uma distribuição de probabilidade) para obter estimadores.

Método da Máxima Verosimilhança (M. V.)

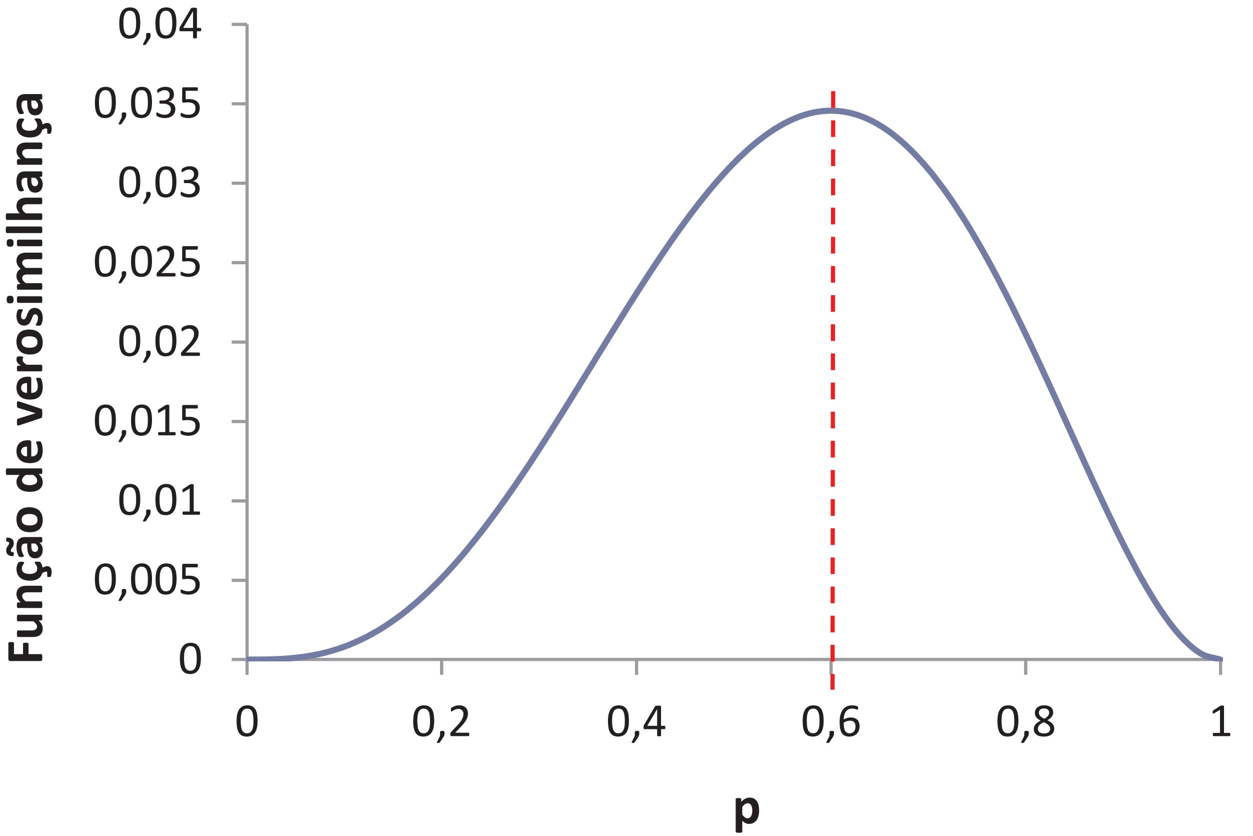
- $X_1, X_2, \ldots, X_n$ é uma a.a. de uma população com f. p. $f(x; \theta)$
- Função de verosimilhança:

$$L(.) = L(\theta) = L(\text{parâmetros}|\text{dados})$$
$$= f(x_1; \theta)\, f(x_2; \theta) \ldots f(x_n; \theta)$$

- Objectivo:
  - Determinar o valor dos parâmetros (θ) que maximizam $L(.)$

Maximizar L(.)

Método da Máxima Verosimilhança (M. V.)



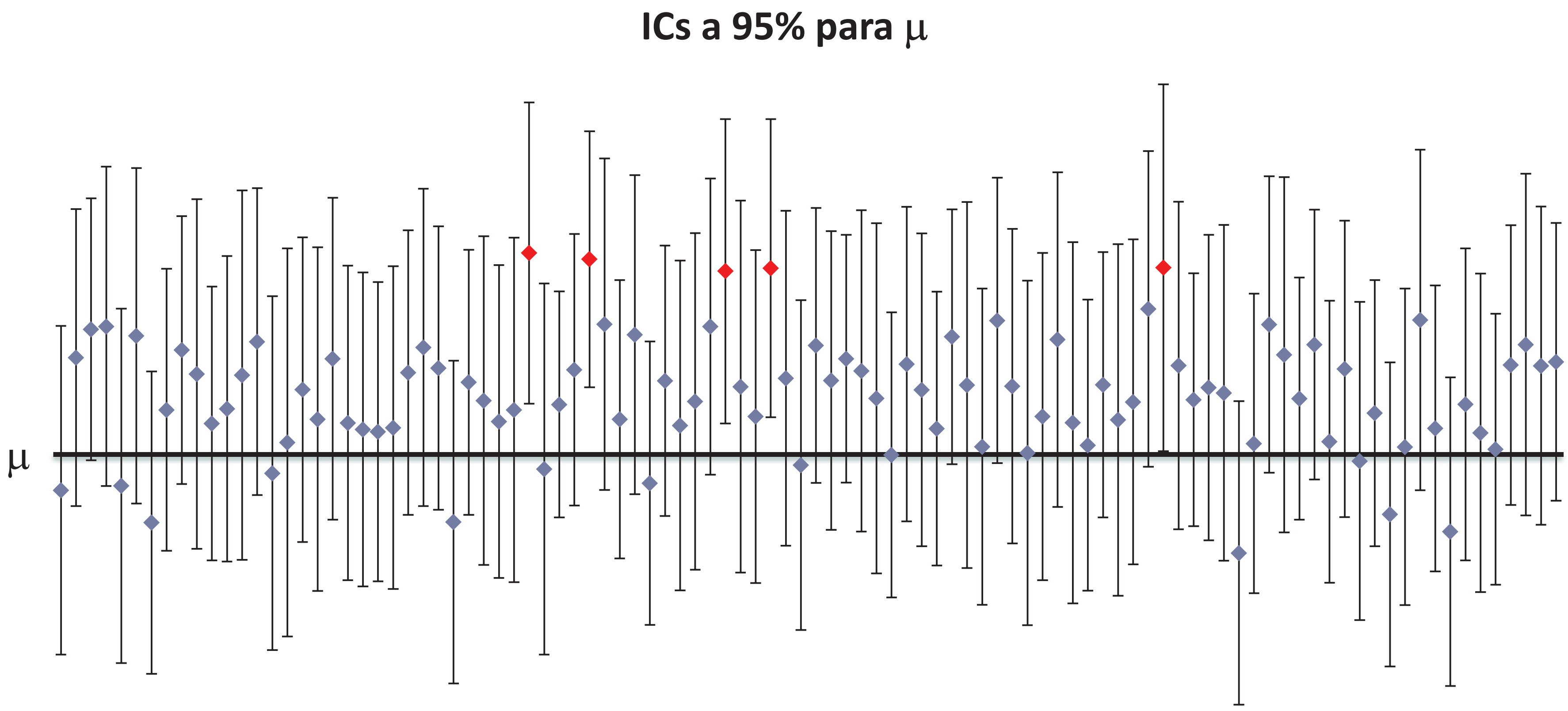
Intervalos de confiança

- Um intervalo de confiança (IC), com $100(1 - \alpha)\%$ de confiança, para um parâmetro θ desconhecido, é dado por

$$\boxed{P(Inf < \theta < Sup) = 1 - \alpha}$$

- Onde os limites deste intervalo]Inf; Sup[são aleatórios.
  - Dependem da amostra observada.

Intervalos de confiança



Intervalos de confiança

Parâmetro	σ^2 conhecido?	Tipo de população	Intervalo de confiança
μ	Sim	Normal (ou qualquer se n grande)	$\left]\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} ; \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right[$
	Não	Normal (ou qualquer se n grande)	$\left]\bar{X} - t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} ; \bar{X} + t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right[$

Exercício 3

- Suponha que no aviário *ApenasOvos* o peso dos ovos produzidos se pode considerar Normal com média 60 gr e desvio-padrão 5 gr.

  Há 15 dias, foi efetuada uma mudança na alimentação das galinhas deste aviário que se julga aumentar o peso dos ovos. Ontem recolheu-se uma amostra aleatória de ovos neste aviário, tendo-se obtido os seguintes resultados:

Statistics

Peso dos ovos (gr)

N	Valid	25
Mean		63,65

a) Forneça uma estimativa pontual para o peso médio dos ovos depois da mudança na alimentação das galinhas.

Exercício 3 (cont.)

b) Com 95% de confiança, depois da mudança de alimentação, o peso médio dos ovos situa-se entre ____ gr e ____ gr.

c) Como procederia caso pretendesse reduzir a amplitude do intervalo de confiança anterior?

d) Qual deveria ser a dimensão da amostra para <u>reduzir o erro de estimativa</u> do I. C. da alínea b) para 1 gr.

e) Se reduzir o erro de estimativa na alínea anterior, o que espera que aconteça à dimensão da amostra? E se aumentar o nível de significância?

f) Qual deveria ser a dimensão da amostra para <u>reduzir a amplitude</u> do I. C. da alínea b) para 1,5 gr.

g) Suponha agora que se desconhece a variância populacional e que se observou s = 5 gr. Resolva novamente a alínea b) e comente as diferenças obtidas.

Intervalos de confiança

Parâmetro	σ_1^2 e σ_2^2 conhecidos?	Tipo de populações	Distribuição amostral
$\mu_1 - \mu_2$	Sim	Normais (ou quaisquer se n_1 e n_2 grandes)	$\left]\bar{X}_1 - \bar{X}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}};\bar{X}_1 - \bar{X}_2 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right[$
	Não ($\sigma_1^2 = \sigma_2^2$)	Normais (ou quaisquer se n_1 e n_2 grandes)	$\left]\bar{X}_1 - \bar{X}_2 - t_{n_1+n_2-2;1-\frac{\alpha}{2}} S^*;\bar{X}_1 - \bar{X}_2 + t_{n_1+n_2-2;1-\frac{\alpha}{2}} S^*\right[$ $S^* = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
	Não ($\sigma_1^2 \neq \sigma_2^2$)	Normais (ou quaisquer se n_1 e n_2 grandes)	$\left]\bar{X}_1 - \bar{X}_2 - t_{v;1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}};\bar{X}_1 - \bar{X}_2 + t_{v;1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right[$ v é obtido pela aproximação de Welch

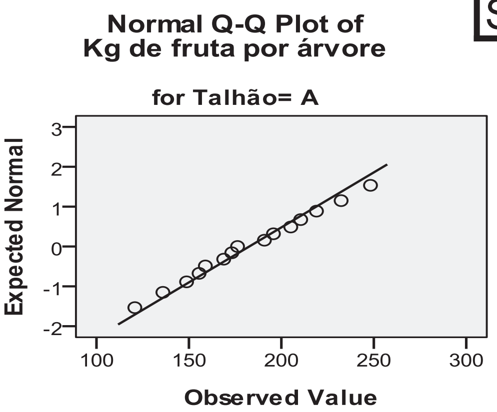
Exercício 4

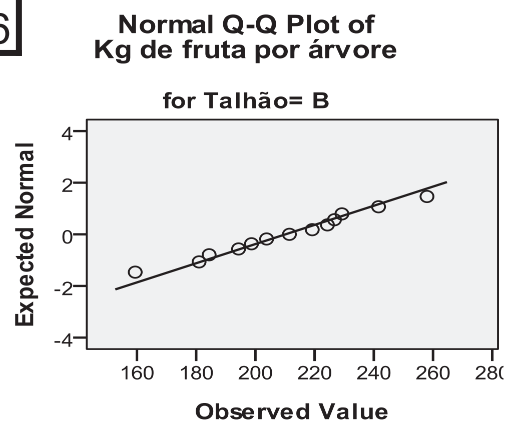
- Um pomar de laranjeiras Baía, plenamente desenvolvido, foi dividido em 2 talhões. Em cada um dos talhões utilizou-se um adubo diferente (A e B). Aquando da colheita das laranjas, selecionaram-se aleatoriamente algumas árvores em cada um dos talhões e registou-se o número de kg de fruto por árvore.

Statistics

Kg de fruta por árvore

		Talhão A	Talhão B
N	Valid	15	13
Mean		182,16	195,43
Std. Deviation		26,938	34,626





Exercício 4 (cont.)

a) Forneça uma estimativa pontual para o número médio de kg de fruto por laranjeira no talhão B. Qual o erro padrão associado à média da amostra?

b) Pela análise do QQ-plot Normal pode considerar-se que os dados são provenientes de uma população com distribuição Normal?

c) Construa um intervalo de confiança (IC) a 90% para a produção média (em kg) por laranjeira no talhão B.

d) Admita que o desvio-padrão populacional no talhão B é conhecido e é igual a 34,626 kg. Resolva novamente a alínea c) e compare os resultados obtidos.

e) Com 95% de confiança a variância da produção por árvore do talhão A varia entre …. e ….?

Adaptado da 2ª frequência: 21 de Junho de 2011

Exercício 4 (cont.)

One-Sample Test

	Test Value = 0						
						95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error	Lower	Upper
Kg de fruta por árvore Talhão A	?	14	?	182,16	6,9554	?	197,079

f) Com 95% de confiança, a produção média (em kg) por laranjeira no talhão A varia entre que valores?

g) Com base num intervalo de confiança a 95% pode-se considerar que a variabilidade nas produções por árvore nos dois talhões é idêntica?

h) Se alterasse o nível de significância da alínea anterior para 1%, manteria a sua decisão? Justifique <u>sem efetuar cálculos</u>.

Adaptado da 2ª frequência: 21 de Junho de 2011

Exercício 4 (cont.)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
							90% Confidence Interval of the Difference	
		F	Sig.	t	df	Std. Error Difference	Lower	Upper
Kg de fruta por árvore	Equal variances assumed	1,590	,219	?	?	11,643	-33,133	?
	Equal variances not assumed			-1,119	22,568	11,858	-33,631	7,091

j) Com base num intervalo de confiança a 95%, podemos considerar que a produção média por árvore não é idêntica nos dois talhões?

k) Se alterasse o nível de significância da alínea anterior para 10%, manteria a sua opinião? E se $\alpha = 1\%$?

Adaptado da 2ª frequência: 21 de Junho de 2011

Testes de hipóteses

Hipóteses

Decisão

Tipos de erro, valor p e potência do teste

Etapas

Teste de hipótese para a média e diferença de médias populacionais

Referência ao teste de Kolmogorov-Smirnov e Shapiro-Wilks

Introdução

- **Objetivo:** testar afirmações sobre a população
- Classificação dos testes de hipóteses:
  - Paramétricos:
    - Têm como pressuposto uma distribuição para a população.
    - Utilizam-se para afirmações sobre os parâmetros populacionais.
  - Não paramétricos:
    - Dizem-se "distribution free".
    - São mais simples.
    - Utilizam-se tanto para afirmações sobre os parâmetros populacionais como sobre a distribuição populacional.

Introdução

- Hipóteses:
  - H_0: hipótese nula:
    - Contém sempre uma igualdade,
    - O que se aceita por defeito até prova em contário.
  - H_1: hipótese alternativa:
    - O que se pretende testar.
- Tipos de testes:
  - Bilaterais,
  - Unilaterais: esquerdo ou direito.
- Decisão:
  - Rejeitar H_0;
  - Não rejeitar H_0.

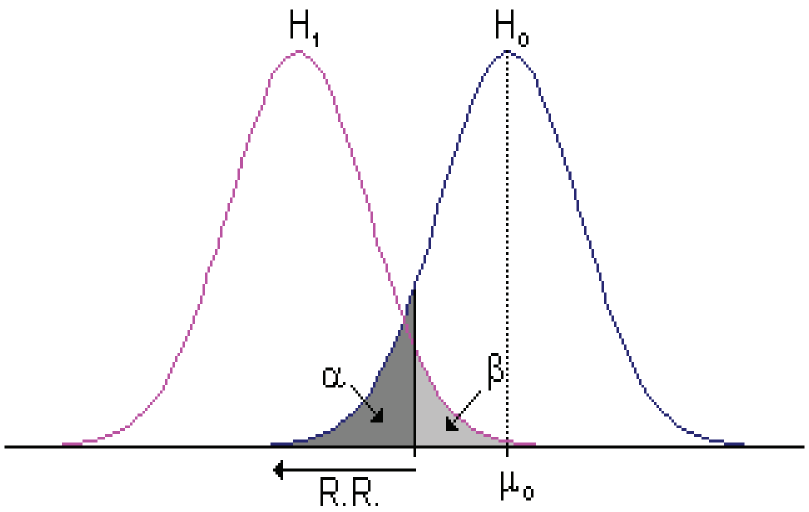
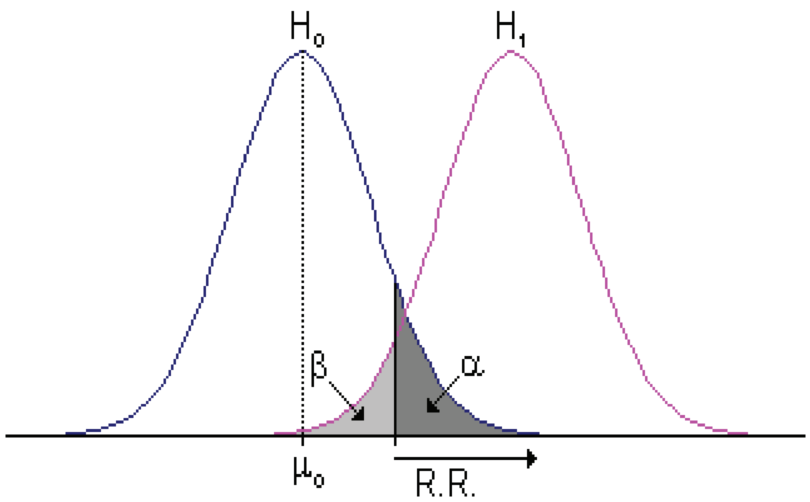
Erros nos testes de hipóteses

Decisão	Situação real	
	H_0 é verdadeira	H_0 é falsa
Rejeitar H_0	Decisão incorrecta $P(\text{rej } H_0 \mid H_0 \text{ verd.}) \leq \alpha$	Decisão correcta $P(\text{Rej. } H_0 \mid H_1 \text{ verd.}) = 1 - \beta$
Não rejeitar H_0	Decisão correcta $P(\text{Não rej. } H_0 \mid H_0 \text{ verd.}) > 1 - \alpha$	Decisão incorrecta $P(\text{Não rej. } H_0 \mid H_1 \text{ verd.}) = \beta$

Erros nos testes de hipóteses

Decisão	Situação real	
	H_0 é verdadeira	H_0 é falsa
Rejeitar H_0	Decisão incorrecta $P(\text{rej } H_0 \mid H_0 \text{ verd.}) \leq \alpha$ Erro tipo I	Decisão correcta $P(\text{Rej. } H_0 \mid H_1 \text{ verd.}) = 1 - \beta$
Não rejeitar H_0	Decisão correcta $P(\text{Não rej. } H_0 \mid H_0 \text{ verd.}) > 1 - \alpha$	Decisão incorrecta $P(\text{Não rej. } H_0 \mid H_1 \text{ verd.}) = \beta$ Erro de Tipo II

Erros de tipo I e II

Teste unilateral esquerdo ou bilateral	Teste unilateral direito ou bilateral
	

Valor p e Potência do teste

- **Valor p:**
  - Menor nível de significância a partir do qual se rejeita H_0.
  - Se $\alpha \geq$ valor p $\Rightarrow$ rejeitar H_0.
- **Potência do teste (π):**
  - Mede a capacidade do teste decidir corretamente quando H_0 é falsa.
  - $\pi(\theta_1) = P(\text{Rej. } H_0 | H_0 \text{ falsa}) = 1 - \beta(\theta_1)$, onde θ_1 é o valor de θ em H_1.
  - Quanto mais afastado estiver θ_1 de θ_0 maior é a potência do teste.

Etapas

Definir H_0 e H_1.

Escolher a estatística de teste (ET).

Escolher o nível de significância (α).

Recolher amostra e calcular ET_{obs}.

Opção a)

Calcular valor p.

Se $\alpha <$ valor p $\Rightarrow$ não rejeitar H_0;
Se $\alpha \geq$ valor p $\Rightarrow$ rejeitar H_0.

Opção b)

Definir R.A. e R.R.

Se $ET_{obs} \in$ R.A. $\Rightarrow$ não rejeitar H_0;
Se $ET_{obs} \in$ R.R. $\Rightarrow$ rejeitar H_0.

Quadro resumo

Parâmetro	σ^2 conhecido?	Tipo de população	Estatística de teste (ET)
μ	Sim	Normal (ou qualquer se n grande*)	$Z=\dfrac{\bar{X}-\mu_0}{\dfrac{\sigma}{\sqrt{n}}} \cap N(0;1)$
	Não	Normal (ou qualquer se n grande*)	$T=\dfrac{\bar{X}-\mu_0}{\dfrac{S}{\sqrt{n}}} \cap t_{(n-1)}$

*Nesse caso a distribuição amostral é aproximada em vez de exata.

Quadro resumo

Parâmetros	σ_1^2 e σ_2^2 conhecidos?	Tipo de populações	Estatística de teste (ET)
$\mu_1 - \mu_2$	Sim	Normais (ou quaisquer se n_1 e n_2 grandes*)	$Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \cap N(0;1)$
	Não ($\sigma_1^2 = \sigma_2^2$)	Normais (ou quaisquer se n_1 e n_2 grandes*)	$T = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_1 - 1)S_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \cap t_{(n_1 + n_2 - 2)}$
	Não ($\sigma_1^2 \neq \sigma_2^2$)	Normais (ou quaisquer se n_1 e n_2 grandes*)	$T = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \overset{\circ}{\cap} t_{(v)}$

v é obtido pela aproximação de Welch

SPSS: TH para a média populacional (σ desconhecido)

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
X	10	3,3000	,64291	,20331

$\frac{s'}{\sqrt{n}}$

n $\quad \bar{x}$ $\quad s'$

μ_0

One-Sample Test

	Test Value = 3					
	t	df	Sig. (2-tailed)	Mean Difference	99% Confidence Interval of the Difference Lower	99% Confidence Interval of the Difference Upper
X	1,476	9	,174	,3000	-,3607	,9607

t_{obs}

graus de liberdade (g. l.)

p-value (teste bilateral)

$\bar{x} - \mu_0$

SPSS: TH para a diferença de médias populacionais (σ_1 e σ_2 desconhecidos)

Group Statistics

	GRUPO	N	Mean	Std. Deviation	Std. Error Mean
X	1,00	10	3,3000	,64291	,20331
	2,00	7	3,6857	,60945	,23035
		n_i	$\bar{x}_i$	s'_i	$\frac{s'_i}{\sqrt{n}}$

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
X	Equal variances assumed	,093	,765	-1,243	15	,233	-,3857	,31034	-1,0472	,2758
	Equal variances not assumed			-1,255	13,519	,231	-,3857	,30724	-1,0469	,2755
				t_{obs}	g. l.	p-value (teste bilateral)	$\bar{x}_1 - \bar{x}_2$	s^*		

Se $\sigma_1^2 = \sigma_2^2$

Se $\sigma_1^2 \neq \sigma_2^2$

Teste à igualdade das variâncias populacionais

SPSS: TH para a diferença de médias populacionais (amostras emparelhadas e σ desconhecido)

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	TEORICA	79,7500	24	20,24255	4,13199
	PRATICA	74,2917	24	17,87026	3,64775

Std. Error Mean ← $\frac{s_i'}{\sqrt{n}}$

Mean ↑ $\bar{x}_i$; N ↑ n_i; Std. Deviation ↑ s'_i

Paired Samples Correlations

	N	Correlation	Sig.
Pair 1 TEORICA & PRATICA	24	,972	,000

N ↑ n_i; Correlation ↑ r; Sig. ↑ p-value (teste bilateral H_0: $\rho = 0$)

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference: Lower	95% Confidence Interval of the Difference: Upper			
Pair 1 TEORICA - PRATICA	5,4583	5,0560	1,0321	3,3234	7,5933	5,289	23	,000

Mean ↑ $\bar{x}_D$; Std. Deviation ↑ s'_D; t ↑ t_{obs}; df ↑ g. l.; Sig. (2-tailed) ↑ p-value (teste bilateral)

Testes à Normalidade: K-S e S-W

- **Teste de Kolmogorov-Smirnov:**
  - Parâmetros da dist. Normal não definidos, usar correcção de Lilliefors;
  - Usar com amostras grandes.
- **Teste de Shapiro-Wilk**
  - Usar com amostras pequenas.

Hipóteses a testar: H_0: X tem distribuição Normal vs

H_1: X não tem distribuição Normal

Regra de decisão: Rejeitar H_0 quando $\alpha \geq$ valor p

SPSS: Testes à Normalidade de K-S e S-W

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
X	,217	17	,033	,814	17	,003

a. Lilliefors Significance Correction

ET_{obs} Graus de liberdade Valor p ET_{obs} Graus de liberdade Valor p