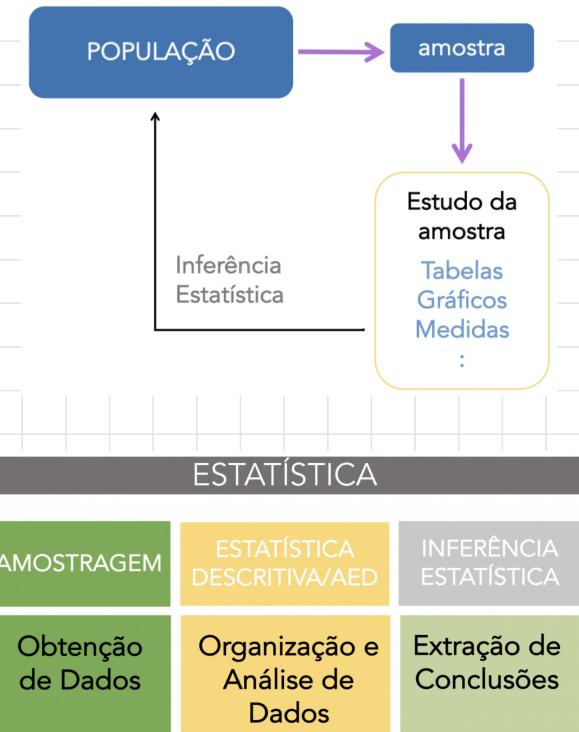


## Amostragem



População: conjunto de elementos com uma ou mais características em comum que se pretendem estudar

Unidade Estatística: elemento da população

Amostra: parte da população

Amostra Aleatória Simples: se os critérios utilizados garantirem que todas as amostras da mesma dimensão têm igual probabilidade de serem selecionadas

Amostragem  
aleatória simples

com reposição

sem reposição

- As diferenças entre amostragem com e sem reposição não são significativas quando a dimensão da população é muito maior do que a dimensão da amostra

Amostra Aleatória Não Simples: quando é impraticável ou desaconselhável a recolha de uma amostra aleatória simples

1. Agrupamento: organização das unidades estatísticas em grupos com características semelhantes e tão homogêneas quanto possível no que diz respeito a fatores que se suspeita que possam ter influência no resultado - em estudos experimentais

2. Emparelhamento: caso particular de agrupamento em que a comparação é feita em pares de unidades experimentais com características semelhantes entre si

"Agrupar o que se pode e distribuir aleatoriamente o que não for possível"  
para garantir comparações corretas no que respeita a fatores que se sabe serem importantes | para tentar tornar comparável o que diz respeito a fatores desconhecidos

3. Estatificação: organização da população em conjuntos homogêneos de indivíduos (estratos) e depois retirar amostras aleatórias de cada um dos estratos, combinando-as de modo a constituir a amostra final - em estudos observacionais

Bem...  
Bem...

{ Agrupamento: amostragem em estudos experimentais

Estatificação: amostragem em estudos observacionais

## Estudos Estatísticos

Estudo Observacional: dados obtidos apenas por observação, nem qualquer influência do observador - não é possível estabelecer associação, não causalidade

- O investigador recolhe dados apenas como observado e não como alguma que manipula condições

Estudo Retrospectivo: é usada informação já disponível acerca dos indivíduos

Estudo Prospetivo: é recolhida informação num futuro próximo

- Os indivíduos é que se auto-associam a diferentes grupos, os investigadores apenas observam o que acontece

Estudo Experimental: experiência desenhada pelo investigador para responder a alguma questão - é possível estabelecer causalidade

- O investigador controla o estudo, decidindo grupos de controlo e tratamento
- Os investigadores decidem quem estará no grupo de tratamento e quem estará no grupo de controlo

**"Controlo"** {

- Indivíduo que pertence ao grupo de controlo (não se sujeita ao tratamento)
- Experiência controlada, estudo no qual os investigadores decidem quem estará e quem não estará no grupo de tratamento

Variáveis de Perturbação: estão relacionadas com as variáveis em estudo mas podem não ser incluídas na análise dos dados, o que leva a que não se possa estabelecer uma relação de causa-efeito em estudos observacionais

Como controlar em estudos experimentais?

1. Seleção aleatória
2. Dispor de grupos homogêneos (agrupamento)
3. Administração de um placebo
4. Ensaio durante todo

- Depois da compreensão/análise dos dados têm de se tomar decisões tendo em conta a variabilidade e incerteza

## População

$X$

$Y$

$$\mu_X = E(X)$$

$$\mu_Y = E(Y)$$

$$\sigma_X = \sqrt{V(X)}$$

$$\sigma_Y = \sqrt{V(Y)}$$

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

## Amostra

$(x_1, \dots, x_n)$

$(y_1, \dots, y_n)$

$\bar{x}$  - média de  $(x_1, \dots, x_n)$

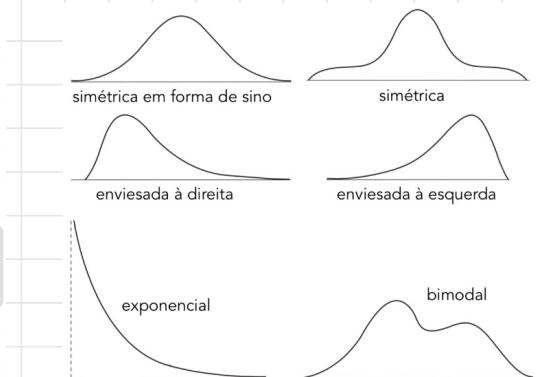
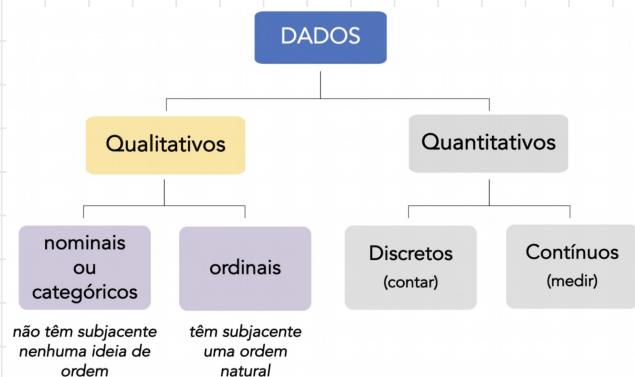
$\bar{y}$  - média de  $(y_1, \dots, y_n)$

$s_x$  - desvio padrão de  $(x_1, \dots, x_n)$

$s_y$  - desvio padrão de  $(y_1, \dots, y_n)$

$$r = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

# Estatística Descritiva



Frequência Relativa:  $f_r = \frac{f_a}{n_T}$

Classe:  $[ \dots, \dots ]$

Fórmula de Sturges: número de classes  $k = 1 + \log_2 n$

Amplitude de Classe:  $a = \frac{M - m}{k}$

# Medidas de Localização

Média:  $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{1}{n} \sum_{i=1}^n x_i$  — influenciada pelos extremos

$$\begin{aligned}\bar{y} &= a\bar{x} + b, \\ y &= ax + b\end{aligned}$$

Mediana: observação central, depois de ordenados os dados

- A mediana é mais robusta do que a média, porque a presença de valores muito diferentes da maioria afeta muito mais o valor da média do que o valor da mediana (o valor é pouco afetado por alterações — mesmo que dramáticas — num pequeno grupo de dados)

Quartil: o quantil de ordem  $q_p$  ( $0 < q_p < 1$ ) é o valor  $Q_p$  que divide as observações da amostra em duas partes, sendo que a percentagem dos elementos da amostra que não inferiores ou iguais a  $Q_p$  é pelo menos  $p$  e a percentagem dos elementos da amostra que não superiores ou iguais a  $Q_p$  é pelo menos  $1-p$

$$Q_p = \begin{cases} \frac{n_p + n_{(p+1)}}{2}, & n_p \in \mathbb{Z} \\ (k+1), & n_p \notin \mathbb{Z} \end{cases}$$

Percentil: o  $k$ -ésimo percentil ( $0 < k < 100$ ) é o valor  $P_k$  que separa as  $k\%$  menores observações da amostra das  $(100-k)\%$  maiores

Quintil: o primeiro quartil,  $Q_1$ , é o percentil 25, o segundo quartil,  $Q_2$ , é o percentil 50 (mediana) e o terceiro quartil,  $Q_3$ , é o percentil 75

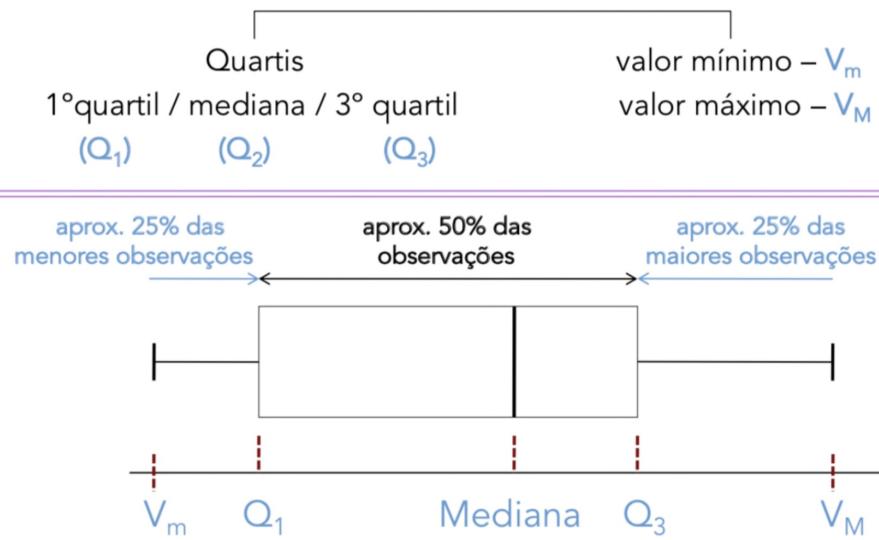
Moda: valor(s) de frequência máxima

Dado Modal: dado com frequência máxima

# Diagramas

Extremos e Quartis: mínimo;  $Q_1$ ;  $Q_2$  (mediana);  $Q_3$ ; máximo

5 números



Caixa e Bigodes:

1º bigode: menor observação  $\geq BI = Q_1 - 1,5AIQ$

2º bigode: maior observação  $\leq BS = Q_3 + 1,5AIQ$

Outliers: observações não compreendidas entre os bigodes

# Medidas de Dispersion

Amplitude Amstral:  $A = M - m$

Amplitude Interquartil:  $AID = Q_3 - Q_1$

Derro Padrão:  $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$   $\Delta_y = a s_n, y = ax + b$

Variância:  $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2)}{n-1}$

$$= \frac{\left(\sum_{i=1}^n x_i^2\right) - 2\bar{x}\left(\sum_{i=1}^n x_i\right) + n\bar{x}^2}{n-1} = \frac{\left(\sum_{i=1}^n x_i^2\right) - 2n\bar{x}^2 + n\bar{x}^2}{n-1}$$

$$\boxed{= \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}}$$

Coeficiente de Variação:  $CV = \frac{s}{\bar{x}}$

## Dados Bivariados

Covariância:  $f_{x,y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{n-1}$

Correlação:  $r = \frac{1}{n-1} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$

# Folha 1

1. Num estudo estatístico sobre uma doença neurológica foram obtidos dados relativos a sexo, peso, tipo de tratamento, número de convulsões e classificação da doença (leve, moderada e severa) de diversos doentes na mesma faixa-etária.

Classificar os diversos tipos de dados.

1. Sexo - qualitativo categórico

Peso - quantitativo contínuo

Tipo de tratamento - qualitativo categórico

Número de convulsões - quantitativo discreto

Classificação da doença - qualitativo ordinal



2. [Adaptado de (\*)] Relativamente a cada um dos cenários seguintes, identificar a(s) variável(eis) em estudo, a dimensão da amostra e o tipo de dados:

- (a) Um paleontólogo mediu a largura do molar superior em 36 espécimes do extinto *Acropithecus rigidus*.  
(b) Foram registados, o peso à nascença, o dia de nascimento e a nacionalidade da mãe de 65 bebés.  
(c) Foram registados o tipo de sangue e o nível de colesterol em 125 adultos.  
(d) Um biólogo registou o número de folhas em cada uma de 25 plantas.

2.

a) variável em estudo: largura do mola  
dimensão da amostra: 36



tipo de dados: quantitativo contínuo

b) peso - 65 - quantitativo contínuo

dia de nascimento - 65 - qualitativo ordinal

nacionalidade - 65 - qualitativo categórico



c) tipo de sangue - 125 - qualitativo categórico

nível de colesterol - 125 - quantitativo contínuo



d) número de folhas - 25 - quantitativo discreto



3. [Adaptado de (\*\*\*)]

Os dados seguintes dizem respeito a observações da magnitude (na escala de Richter) de 30 sismos na Califórnia [DadosFicha1]:

1.0	8.3	3.1	1.1	5.1
1.2	1.0	4.1	1.1	4.0
2.0	1.9	6.3	1.4	1.3
3.3	2.2	2.3	2.1	2.1
1.4	2.7	2.4	3.0	4.1
5.0	2.2	1.2	7.7	1.5

$$\left[ \text{Dados auxiliares: } \sum_{i=1}^{30} x_i = 86.1 \quad \sum_{i=1}^{30} x_i^2 = 357.61 \right]$$

(a) Representar os dados através de um diagrama de pontos e interpretar.

(b) Calcular a média, desvio padrão e mediana da amostra.

$$3.b) \text{ Média } \bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{86.1}{30} = 2,87$$



$$\text{Mediana} = 2,2$$

$$\text{Desvio Padrão: } \lambda = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = 1,95 \quad \lambda = \sqrt{\frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1}} = 1,95$$

33.1	33.4	34.8	33.8	34.7	34.3	35.6
34.5	34.6	34.1	33.9	33.6	34.6	35.2
33.7	35.8	34.2	34.0	34.7	35.2	34.3
33.4	36.0	34.5	36.1	35.1	35.1	34.6
33.7	34.9	34.2	34.2	34.2	35.3	34.2

$$\left[ \text{Dados auxiliares: } \sum_{i=1}^{35} x_i = 1207.6 \quad \sum_{i=1}^{35} x_i^2 = 41684.32 \right]$$

(a) Representar os dados através de um diagrama de pontos e um histograma adequado.

(b) Calcular a média e desvio padrão da amostra.

$$4.b) \bar{x} = \frac{1207.6}{35} = 34.5$$

$$\lambda = \sqrt{\frac{41684 - 35 \times 34.5^2}{34}} = 0,74$$



## 6. [Adaptado de (\*\*)]

Num estudo sobre a esquizofrenia, foi medida a atividade de uma determinada enzima nas plaquetas sanguíneas de 18 pacientes. Os resultados (em determinadas unidades), foram os seguintes[DadosFicha1]:

6.8	8.4	8.7	11.9	14.2	18.8
9.9	4.1	9.7	12.7	5.2	7.8
7.8	7.4	7.3	10.6	14.5	10.7

- (a) Construir um histograma para esta amostra, considerando 5 classes.
- (b) Calcular a mediana e os quartis.
- (c) Calcular o intervalo inter-quartis.
- (d) Construir o diagrama de caixa e bigodes dos dados.

6. b)  $Q_1: n_{1/4} = 18 \times 0,25 = 4,5$

$n_{3/4} = 7,4$

$Q_2: n_1 = 18 \times 0,5 = 9$

$\frac{n_9 + n_{10}}{2} = \frac{8,7 + 9,7}{2} = 9,2$



$Q_3: n_{3/4} = 18 \times 0,75 = 13,5$

$n_{11/4} = 11,9$

$18,8$

o

c)  $A1Q = Q_3 - Q_1 = 11,9 - 7,4 = 4,5$  ✓

$14,5$

d)  $BS = Q_3 + 1,5 A1Q = 11,9 + 1,5 \times 4,5 = 18,65$

$Q_3 11,9$

$BI = Q_1 - 1,5 A1Q = 7,4 - 1,5 \times 4,5 = 9,65$ 

$Q_2 1,2$

$Q_1 7,4$

$4,1$

7.

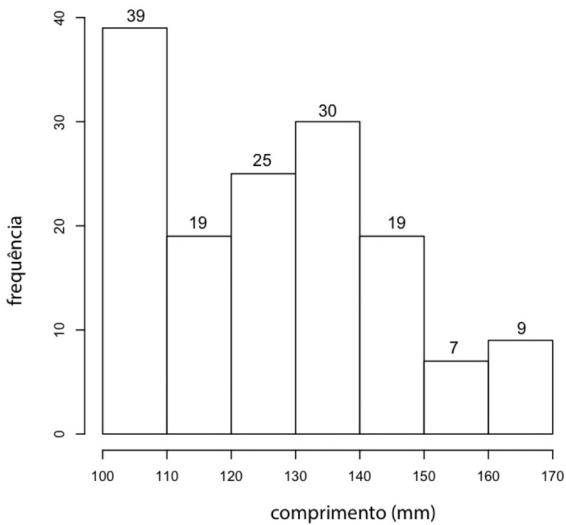
Comprimento ( $\mu m$ )	Frequência (n de indivíduos)	Comprimento ( $\mu m$ )	Frequência (n de indivíduos)
15	1	27	36
16	3	28	41
17	21	29	48
18	27	30	28
19	23	31	43
20	15	32	27
21	10	33	23
22	15	34	10
23	19	35	4
24	21	36	5
25	34	37	1
26	44	38	1

(b) Que característica do histograma sugere a interpretação de que os 500 indivíduos são um misto de dois tipos distintos?

(c) Construir um histograma dos dados usando apenas 6 classes. Comentar o facto deste histograma permitir uma interpretação qualitativamente diferente da inferida na primeira alínea.

b) distribuição assimétrica dos dados (esquerda-direita) ✓

8. Os resultados obtidos num estudo sobre o comprimento de uma certa espécie de peixes estão apresentados no seguinte histograma:



- (a) Qual a dimensão da amostra utilizada nesse estudo?
- (b) Quantos peixes apresentaram um comprimento superior ou igual a 15 cm?
- (c) Estimar a percentagem de peixes na amostra com comprimento inferior a 13.5 cm?
- (d) Alterar a escala das ordenadas para frequências relativas.
- (e) Alterar a escala das ordenadas para densidades.

8.

a)  $39 + 19 + 25 + 30 + 19 + 7 + 9 = 148$  ✓

b)  $7 + 9 = 16$  ✓

c)  $\frac{39 + 19 + 25 + 30}{148} = \frac{89}{148} = 60\%$  ✓

1) Dividi tudo por 148 ✓

2) Dividi tudo por 148 e por 10 ✓

## 9. [Adaptado de (\*\*)]

Os dados seguintes dizem respeito à idade (em anos) no momento em que foi feito o primeiro diagnóstico de diabetes tipo 2, de 20 diabéticos selecionados aleatoriamente [DadosFicha1]:

35.5	40.1	47.3	48.9	52.4
39.8	39.3	55.6	40.3	60.9
30.5	59.8	44.5	36.8	36.6
42.1	26.2	33.3	65.4	45.1

$$\left[ \sum_{i=1}^{20} x_i = 880.4 \quad \sum_{i=1}^{20} x_i^2 = 40854.56 \right]$$

(a) Representar os dados graficamente.

(b) Calcular a média e o desvio padrão dos dados.

9. b)  $\bar{x} = \frac{880,4}{20} = 44,02$



$$\Delta = \sqrt{\frac{40854,56 - 20 \times 40,2^2}{19}} = 10,5$$



## 10. [Adaptado de (\*)]

Uma bióloga mediou um certo pH em cada um de 24 sapos, obtendo valores típicos:

7.43 7.16 7.51 ...

Calculou uma média de 7.373 e um desvio padrão de 0.129 para estas medidas originais de pH. A seguir, transformou os dados, subtraindo 7 a cada observação e depois multiplicando por 100. Por exemplo 7.43 foi transformado em 43. Quais são a média e o desvio padrão dos dados transformados?

10.  $\bar{x} = 7,373 \quad \Delta_x = 0,129$

$$y = (x_i - 7) \times 100 = 100x_i - 700$$

$$\bar{y} = 100\bar{x} - 700 = 37,3$$



$$\Delta_y = 100 \Delta_x = 12,9$$

## 11. [Adaptado de (\*)]

A tabela seguinte mostra o tamanho da ninhada (número de leitões que sobrevivem 21 dias), para cada uma de 36 porcas [DadosFicha1].

Tamanho da ninhada	Frequência (no. de porcas)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2
Total	36

- (a) Representar os dados usando um gráfico adequado.
- (b) Calcular as medidas de localização da amostra.
- (c) Representar a caixa de bigodes e decidir acerca do enviesamento da distribuição dos dados.

11. b)  
 média:  $\bar{x} = 10,42$   
 mediana:  $10,5$   
 moda:  $10$

$$\begin{aligned} Q_{0,25}: \quad 36 \times 0,25 = 9 & \quad (w_9 + w_{10})/2 = 9,5 \\ Q_{0,5}: \quad 36 \times 0,5 = 18 & \quad (w_{18} + w_{19})/2 = 10,5 \\ Q_{0,75}: \quad 36 \times 0,75 = 27 & \quad (w_{27} + w_{28})/2 = 12 \end{aligned}$$

c)  $A1Q = 12 - 9 = 3$

$$\begin{aligned} BI = Q_1 - 1,5A1Q &= 5 & 18 - 7 \\ BS = Q_3 + 1,5A1Q &= 15,5 & 2B = 14 \end{aligned}$$

## 12. [Adaptado de (\*)]

Um botânico plantou 15 plantas de pimenta numa estufa. Vinte e um dias depois mediou a altura total (em cm) do caule das plantas e obteve os valores seguintes [DadosFicha1]:

12.4	12.2	13.4	12.1	12.2
11.8	13.5	12.0	10.9	13.2
12.6	11.9	13.1	14.1	12.7

12.  $\bar{x} = 188,1/15 = 12,54$  ✓

$$\sigma^2 = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n-1} = 0,6611429$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = 0,8131069$$

## 13. [Adaptado de (\*)]

Dez pacientes hipertensos participaram num estudo para avaliar a eficácia de um medicamento para reduzir a tensão arterial. A tabela abaixo mostra a tensão sistólica medida antes e depois de duas semanas de tratamento.

(a) Comparar as duas distribuições usando um diagrama de caixa e bigodes.

(b) Representar a amostra num diagrama de dispersão complementando-o com o valor do coeficiente de correlação amostral.

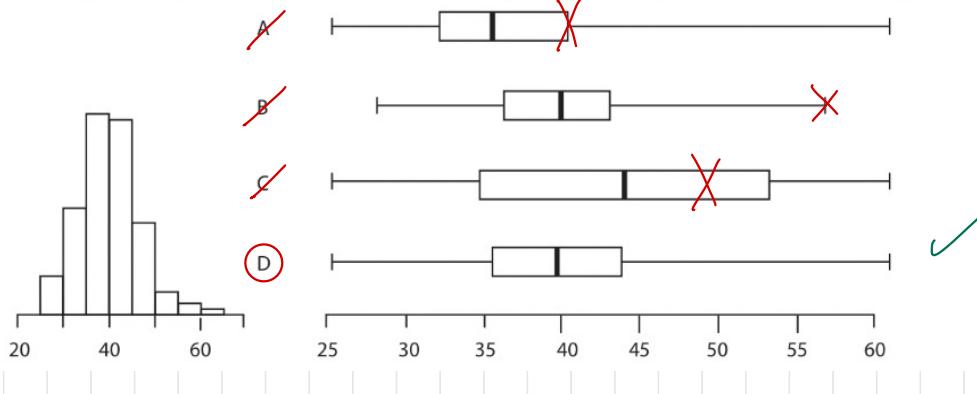
Paciente	1	2	3	4	5	6	7	8	9	10
Antes	172	186	170	205	174	184	178	156	190	168
Depois	159	157	163	207	164	141	182	171	177	138

13.

$$b) r = \frac{1}{n-1} \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = 0,51 \quad \checkmark$$

## 14. [Adaptado de (\*)]

O histograma seguinte representa os mesmos dados do que um dos diagramas de extremos e quartis. Qual?



## 15. Considerar a amostra bivariada:

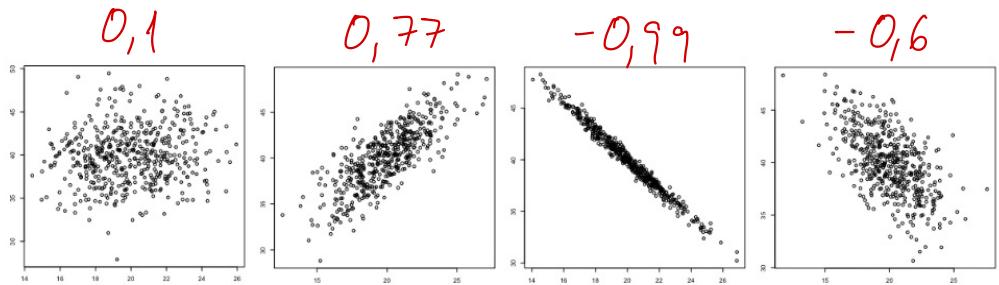
x	4.6	5.2	3.6	5.0	3.5	4.2	3.9	3.8	4.2	3.9	5.1
y	3.8	4.4	3.2	4.0	4.5	3.5	3.5	3.4	3.9	3.3	4.5

(a) Representar os dados através de um diagrama de dispersão

(b) Calcular o coeficiente de correlação de Pearson da amostra.

$$b) r = \frac{1}{n-1} \times \frac{1}{s_x s_y} \times \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = 0,5565 \quad \checkmark$$

16. Os valores 0.1, 0.77, -0.6 e -0.99 são os coeficientes de correlação de Pearson de 4 amostras bivariadas representadas abaixo por diagramas de dispersão. Associar os coeficientes aos respetivos diagramas.

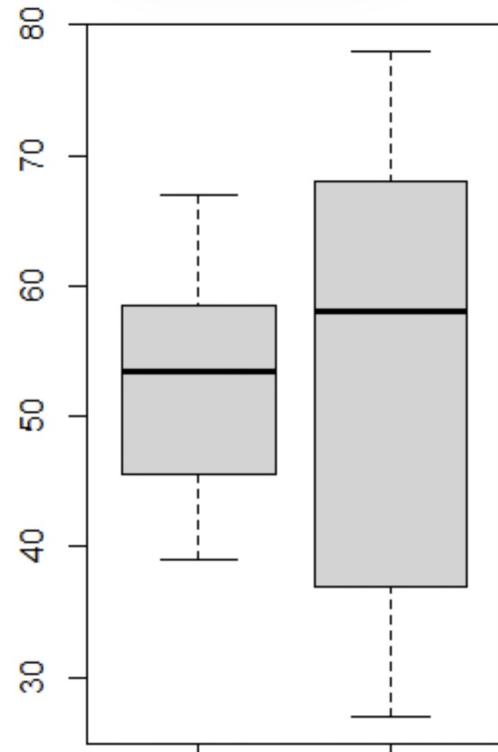


✓

17. O aumento de peso, em gramas, em dois conjuntos de animais submetidos a diferentes dietas foi:

	Grupo A	56	67	42	48	55	61	52	39	47	58	50	40	59	62	44	57
	Grupo B	78	34	37	72	58	68	27	55	65	40	75	33	66			

Utilizar os diagramas de caixa e bigodes no estudo da diferença entre os dois tipos de dieta.



✓

# Probabilidades

Probabilidade  $P(A)$ : valor numérico que quantifica a possibilidade de ocorrência de um acontecimento A

Estatística Indutiva: 1. Hipótese 2. Observação 3. Conclusão

A probabilidade é utilizada para quantificar o grau de confiança que se pode atribuir à conclusão de um estudo estatístico - instrumento de apoio à decisão

Experiência: procedimento executado sob condições controladas, que produz resultados observáveis e mensuráveis

Experiência Aleatória: (1) conhecem-se todos os resultados possíveis, (2) cada vez que é efectuada não se conhece antecipadamente qual dos resultados possíveis vai ocorrer e (3) pode ser repetida em condições análogas - (4) o acontecimento em estudo ou ocorre ou não ocorre

Espaço Amostral  $\Omega/S$ : conjunto não vazio formado por todos os resultados possíveis de uma experiência aleatória; contínuo ou discreto

Acontecimento: qualquer subconjunto do espaço amostral/de resultados,  $\Omega$

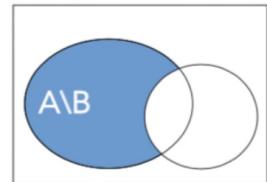
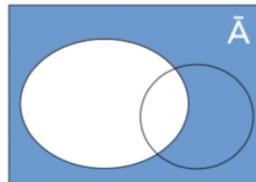
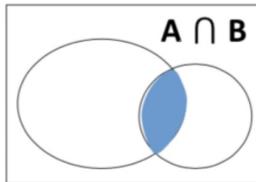
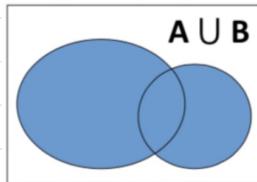
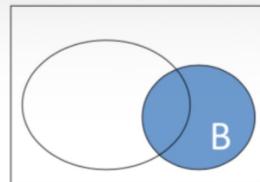
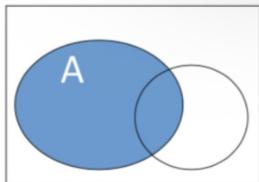
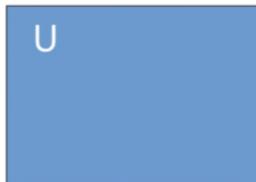
$\emptyset$  - acontecimento impossível

{a<sub>i</sub>} - acontecimento elementar

$\Omega$  - acontecimento certo

Acontecimentos Incompatíveis: a realização de um implica a não realização do outro -  $A \cap B = \emptyset$  - a intersecção é vazia; não conjuntos disjuntos, não têm elementos comuns

# Teoria de Conjuntos



Reunião:  $A \cup B = \{x : x \in A \vee x \in B\}$

Intersetção:  $A \cap B = \{x : x \in A \wedge x \in B\}$

Diferença:  $A \setminus B = \{x : x \in A \wedge x \notin B\}$

Complementar:  $\bar{A} = \{x : x \notin A\}$       -  $A \cup \bar{A} = \Omega$

Associatividade:  $A \cup (B \cup C) = (A \cup B) \cup C$  |  $A \cap (B \cap C) = (A \cap B) \cap C$

Comutatividade:  $A \cup B = B \cup A$  |  $A \cap B = B \cap A$

Distributividade:  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$  |  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Leis de De Morgan:  $\overline{A \cap B} = \bar{A} \cup \bar{B}$  |  $\overline{A \cup B} = \bar{A} \cap \bar{B}$

Idempotência:  $A \cup A = A$  |  $A \cap A = A$

Fluxo:  $A \subseteq B \Rightarrow \{A \cup B = B \wedge A \cap B = A\}$

Modular:  $A \cup \Omega = \Omega$  |  $A \cap \Omega = A$

Outras:

- $A \cap \emptyset = \emptyset$
- $A \cup \emptyset = A$
- $A \cap \bar{A} = \emptyset$
- $A \cup \bar{A} = \Omega$
- $\bar{\Omega} = \emptyset$
- $\bar{\emptyset} = \Omega$

"Yomente se realiza A":  $A \cap \bar{B} \cap \bar{C}$

"A e B realizam-se, mas não C":  $A \cap B \cap \bar{C}$

"Os três acontecimentos realizam-se simultaneamente":  $A \cap B \cap C$

"Realizam-se pelo menos um dos acontecimentos":  $A \cup B \cup C$

"Realizam-se pelo menos dois dos acontecimentos":  $(A \cap B) \cup (A \cap C) \cup (B \cap C)$

**Partição de um Conjunto:** subdivisão de S em subconjuntos  $A_1, \dots, A_n$  de S, tal que:

$$1. A_i \cap A_j = \emptyset, \forall i \neq j$$

$$2. \bigcup_{i=1}^n A_i = S$$

**Probabilidade Clássica (Laplace):** supõe-se que numa experiência aleatória se podem obter m resultados, igualmente favoráveis (equiprováveis); se k desses resultados conduzem à realização de um determinado acontecimento A, então:

$$P(A) = \frac{k}{m}$$

$k$ : número de casos favoráveis

$m$ : número de casos possíveis

Limitações:

- quando os resultados não são equiprováveis
- quando o espaço amostral é infinito

**Probabilidade Freqüentista:** supõe-se que uma experiência aleatória é repetida n vezes, e que um determinado acontecimento A se realizou  $m_n$  vezes durante as n experiências; então, a freqüência relativa de ocorrência do elemento A é  $f_n = m_n/n$ , e, quando n é grande,  $P(A) \approx m_n/n$

Limitações:

- não se aplica se a experiência for repetível
- é apenas uma interpretação - não fornece regra de cálculo

Probabilidade Axiomática de Kolmogorov: uma probabilidade é uma função função  $P(\cdot)$ , definida no espaço de acontecimentos  $\Omega$  de uma ocorrência aleatória, que toma valores reais e que satisfaz os seguintes axiomas:

$$(1) \quad P(A) \geq 0, \quad \forall A \in \mathcal{E}$$

$$(2) \quad P(\Omega) = 1$$

(3) Se  $A_1, \dots, A_n$  são acontecimentos em número finito, ou infinito numerável, tais que  $A_i \cap A_j = \emptyset$  para  $i \neq j$ , então:

$$P(A_1 \cup \dots \cup A_n) = P(A_1) + P(\dots) + P(A_n)$$

Propriedades:

$$1. \quad P(\bar{A}) = 1 - P(A)$$

$$2. \quad P(\emptyset) = 0$$

$$3. \quad A \subseteq B \Rightarrow P(A) \leq P(B)$$

$$4. \quad P(A) \leq 1$$

$$5. \quad P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$6. \quad P(A \cap \bar{B}) = P(A) - P(A \cap B)$$

Probabilidade Condicionada:  $P(A|B) = \frac{P(A \cap B)}{P(B)}$ , probabilidade de A dado B

Propriedades:

$$1. \quad P(\bar{A}|B) = 1 - P(A|B)$$

$$2. \quad P((A \cup C)|B) = P(A|B) + P(C|B) - P((A \cap C)|B)$$

$$3. \quad P(A \cap B) = P(B) \cdot P(A|B) \quad \Leftrightarrow \quad P(A \cap B) = P(A) \cdot P(B|A)$$

$> 0$   $> 0$

Acontecimentos Independentes: se o conhecimento da ocorrência de um não influencia a probabilidade de ocorrência do outro

$$P(A \cap B) = P(A) \times P(B)$$

$$\cdot \quad P(B) > 0 \Rightarrow P(A|B) = P(A)$$

$$\cdot \quad P(A) > 0 \Rightarrow P(B|A) = P(B)$$

• todo o acontecimento é independente de  $\emptyset$  e de  $\Omega$

• A e B são independentes  $\Rightarrow A \cap \bar{B}, \bar{A} \cap B, \bar{A} \cap \bar{B}$  também o são

# Folha 2

1. Numa população a presença de três caracteres genéticos,  $A$ ,  $B$  e  $C$ , está aproximadamente distribuída do seguinte modo:  $A$  está presente em 13% dos indivíduos,  $B$  em 9% e  $C$  em 22%; 5% dos indivíduos apresentam simultaneamente os caracteres  $A$  e  $B$ , 6%  $A$  e  $C$  e 5%  $B$  e  $C$ , ao passo que somente 3% dos indivíduos apresentam os três caracteres genéticos.

Escolhendo-se ao acaso um indivíduo desta população, qual é a probabilidade de ele ser portador de:

- (a) pelo menos um dos caracteres  $B$  ou  $C$ ;
- (b) somente  $C$ ;
- (c)  $B$ , no caso de se saber que o indivíduo apresenta o caractere  $C$ .

$$1. \quad P(A) = 0,13 \quad P(B) = 0,09 \quad P(C) = 0,22 \quad P(A \cap B) = 0,05 \\ P(A \cap C) = 0,06 \quad P(B \cap C) = 0,05 \quad P(A \cap B \cap C) = 0,03$$

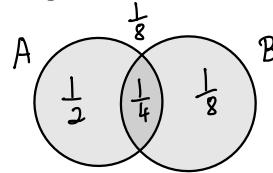
$$a) \quad P(B \cup C) = P(B) + P(C) - P(B \cap C) = 0,09 + 0,22 - 0,05 = 0,26 \quad \checkmark$$

$$b) \quad P(C \cap \bar{A} \cap \bar{B}) = P(C) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) = 0,22 - 0,06 - 0,05 + 0,03 = 0,14 \quad \checkmark$$

$$c) \quad P(B|C) = \frac{P(B \cap C)}{P(C)} = \frac{0,05}{0,22} = 0,227 \quad \checkmark$$

2. Sejam  $A$  e  $B$  dois acontecimentos definidos no mesmo espaço de probabilidade e tais que  $P(A \cup B) = 7/8$ ,  $P(A \cap B) = 1/4$  e  $P(B) = 3/8$ . Calcular:

- (a)  $P(A)$
- (b)  $P(A \cap \bar{B})$
- (c)  $P(B \cap \bar{A})$



$$2. \quad P(A \cup B) = 7/8 \quad P(A \cap B) = 1/4 \quad P(B) = 3/8$$

$$a) \quad P(A) = P(A \cup B) - P(B) + P(A \cap B) = 7/8 - 3/8 + 1/4 = 3/4 \quad \checkmark$$

$$b) \quad P(A \cap \bar{B}) = 1/2 \quad \checkmark$$

$$c) \quad P(B \cap \bar{A}) = 1/8 \quad \checkmark$$

3. Certas culturas podem ser infetadas por bactérias e por cogumelos. A probabilidade de uma cultura estar infetada por cogumelos é de 0.42. A probabilidade de uma cultura não ter bactérias é de 0.85. Sabe-se ainda que a probabilidade de ter uma cultura infetada por bactérias e cogumelos é de 0.05.

- (a) Qual é a probabilidade de uma cultura escolhida ao acaso estar infetada?
- (b) Qual é a probabilidade de uma cultura escolhida ao acaso não ter cogumelos nem bactérias?
- (c) Há independência entre a infecção por bactérias e por cogumelos? Justificar.

$$3. \quad P(C) = 0,42 \quad P(\bar{B}) = 0,85 \quad P(B \cap C) = 0,05$$

a)  $P(B \cup C) = P(B) + P(C) - P(B \cap C) = 1 - P(\bar{B}) + 0,42 - 0,05 = 0,15 + 0,37 = 0,52$  ✓

b)  $P(\bar{B} \cup \bar{C}) = P(\overline{B \cup C}) = 1 - P(B \cup C) = 1 - 0,52 = 0,48$  ✓

c)  $B \text{ e } C$  são independentes  $\Leftrightarrow P(B \cap C) = P(B) \times P(C)$

$$P(B) \times P(C) = 0,15 \times 0,42 = 0,063$$
 ✓

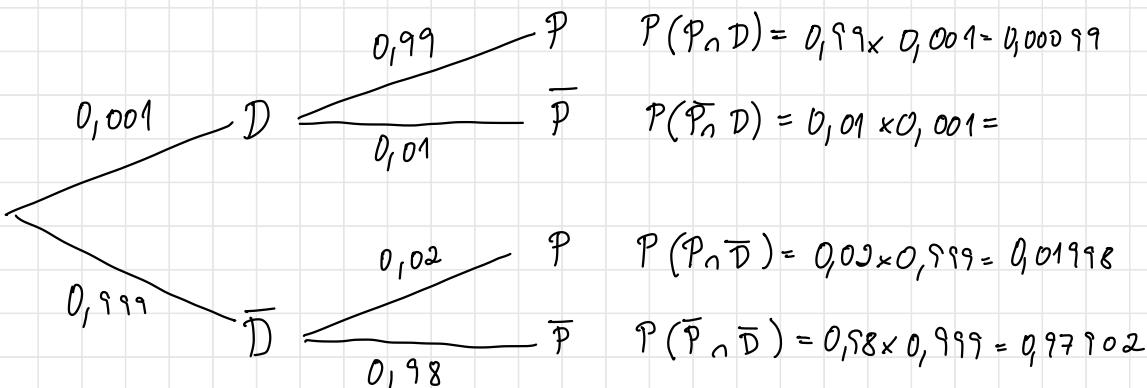
$$P(B \cap C) = 0,05 \neq 0,063 \rightarrow B \text{ e } C \text{ não são independentes}$$

4. Num teste de deteção de uma certa doença, a probabilidade do resultado ser positivo quando aplicado a um indivíduo com essa doença é 0.99, ao passo que a probabilidade de se observar um resultado positivo num indivíduo sem essa doença é 0.02.

Na população de uma região a probabilidade de uma pessoa arbitrária ter a doença é 0.001 (*prevalência*).

Para um indivíduo escolhido ao acaso nesta população, calcular:

- (a) A probabilidade do resultado do teste ser positivo.
- (b) A probabilidade do indivíduo ter a doença, dado que o resultado do teste foi positivo.
- (c) A probabilidade do indivíduo não ter a doença, dado que o resultado do teste foi negativo.
- (d) As probabilidades anteriores, mas no caso de uma população em que a prevalência da doença é 0.2.



$$4. P(P|D) = 0,99$$

$$P(P|\bar{D}) = 0,02$$

$$P(D) = 0,001$$

$$a) P(P) = P(P \cap D) + P(P \cap \bar{D}) = 0,00099 + 0,01998 = 0,02097 \quad \checkmark$$

$$b) P(D|P) = \frac{P(D \cap P)}{P(P)} = \frac{0,00099}{0,02097} = 0,04721 \quad \checkmark$$

$$c) P(\bar{D}|\bar{P}) = \frac{P(\bar{D} \cap \bar{P})}{P(\bar{P})} = \frac{0,97902}{1-0,02097} = 0,9999 \dots \quad \checkmark$$

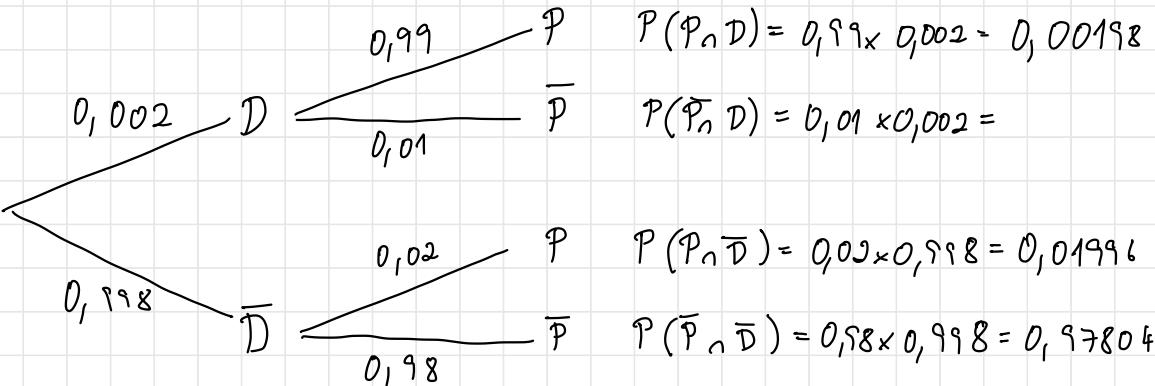
$$d) P(D) = 0,002 \quad //$$

$$P(\bar{D}) = 0,998$$

$$P(P) = P(P \cap D) + P(P \cap \bar{D}) = \\ = 0,002 \times 0,99 + 0,998 \times 0,002 = 0,02194$$

$$P(D|P) = \frac{P(D \cap P)}{P(P)} = \frac{0,002 \times 0,99}{0,02194} = 0,09024$$

$$P(\bar{D}|\bar{P}) = \frac{P(\bar{D} \cap \bar{P})}{P(\bar{P})} = \frac{0,999 \times 0,998}{1-0,02194} = 0,9999 \dots$$



5. Uma gaiola contém 12 cobaias machos e 6 cobaias fêmeas. Retiram-se 4 cobaias da gaiola.

Determinar a probabilidade de serem duas de cada género supondo que,

- (a) a extração é sem reposição;
- (b) a extração é com reposição.

5.  $12 \quad \sigma^{\rightarrow} \quad 6: \varphi \quad T = 28$       (4)

a)  $P = \frac{C_2^{12} \times C_2^6}{C_{18}^4} = \frac{12 \times 11 \times 6 \times 5}{18 \times 17 \times 16 \times 15} \times C_9^4 = \frac{11}{34}$  ✓

b)  $P = \frac{C_F}{CT} = \frac{12^2 \times 6^2}{18^4} \times C_2^4 = \frac{8}{27}$  ✓

Num estudo acerca da relação entre o rendimento familiar e riscos de saúde, um grande grupo de pessoas de uma certa população respondeu a um questionário. Alguns dos resultados encontram-se na tabela abaixo:

	Rendimento Familiar			T
	baixo	médio	alto	
sofre de stress	526	274	216	1016
não sofre de stress	1964	1680	1899	5543
T	2490	1954	2115	6557

Selecionada ao acaso uma pessoa que fez parte deste estudo, qual a probabilidade

- (a) dela sofrer de *stress*?
- (b) dela sofrer de *stress* sabendo que o seu rendimento familiar é alto?
- (c) dela sofrer de *stress* e o seu rendimento familiar ser alto?

6. a)  $P(S) = \frac{1016}{6557}$  ✓

b)  $P(S|A) = \frac{P(S \cap A)}{P(A)} = \frac{216}{2115} = \frac{24}{235}$  ✓

c)  $P(S \cap A) = \frac{216}{6557}$  ✓

7. Numa certa população de lagartos, 5% dos machos e 2% das fêmeas apresentam uma determinada mutação genética. Nessa população, 55% dos indivíduos são fêmeas.

Escolhido ao acaso um indivíduo dessa população verifica-se que tem a mutação. Qual a probabilidade de que seja um macho?

$$7. \quad P(M|\bar{F}) = 0,05$$

$$P(M|F) = 0,02$$

$$P(F) = 0,55$$

$$P(\bar{F}) = 0,45$$

$$P(\bar{F}|M) = \frac{P(\bar{F} \cap M)}{P(M)} = \frac{0,0225}{0,335} = 0,0664 \quad \checkmark$$

$$\begin{aligned} P(M \cap F) &= \\ &= P(M|F) \times P(F) = \\ &= 0,02 \times 0,55 = \\ &= 0,011 \end{aligned}$$

	F	$\bar{F}$	
M	0,011	0,0225	0,335
$\bar{M}$	0,539	0,4275	0,965
	0,55	0,45	1

$$\begin{aligned} P(M \cap \bar{F}) &= \\ &= P(M|\bar{F}) \times P(\bar{F}) = \\ &= 0,05 \times 0,45 = \\ &= 0,0225 \end{aligned}$$

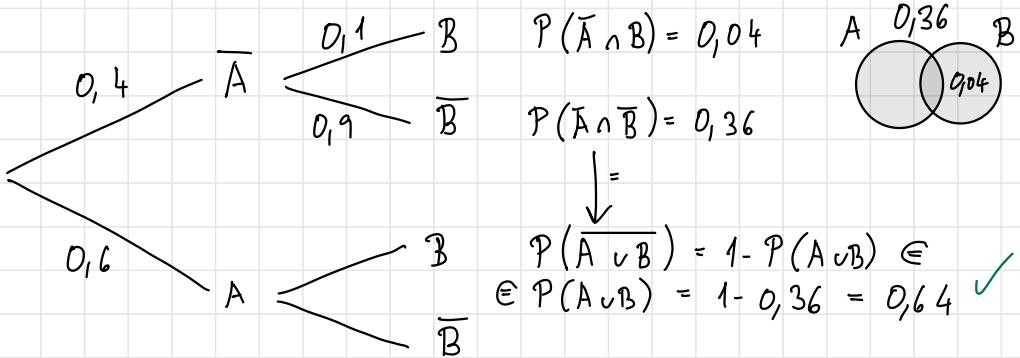
8. A probabilidade do acontecimento A ocorrer é 60%. Se A não ocorre, há uma probabilidade de 10% de ocorrer o acontecimento B.

Nestas condições, qual é a probabilidade de ocorrer pelo menos um dos acontecimentos A ou B?

$$8. \quad P(A) = 0,6$$

$$P(B|\bar{A}) = 0,10$$

$$P(A \cup B) = ?$$



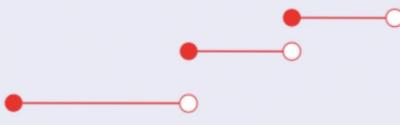
Teorema da Probabilidade Total: Se  $\{A_1, \dots, A_n\}$  é uma partição de  $\Omega$  e se  $P(A_i) > 0$ ,  $\forall i \in \{1, \dots, n\}$ , então  $\forall B: P(B) = P(A_1) \cdot P(B|A_1) + \dots + P(A_n) \cdot P(B|A_n)$

$$n=2: P(B) = P(A) \cdot P(B|A) + P(\bar{A}) \cdot P(B|\bar{A})$$

Teorema de Bayes: Se  $\{A_1, \dots, A_n\}$  é uma partição de  $\Omega$  e se  $P(A_i) > 0$ ,  $\forall i \in \{1, \dots, n\}$ , então  $\forall B, P(B) > 0: P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{P(A_1) \cdot P(B|A_1) + \dots + P(A_n) \cdot P(B|A_n)}$

## Variáveis Aleatórias e Distribuições

Variável Aleatória: função que associa um número real a cada elemento do espaço de resultados de uma experiência aleatória  $\longleftrightarrow$  População

$X$ v.a. discreta	$X$ v.a. contínua
$F_X$ constante 'aos bocados'	$F_X$ contínua
	
$\sum_i f(x_i) = 1$	$\int_{-\infty}^{+\infty} f(x) dx = 1$
$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$	$F_X(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$

Variável Aleatória Discreta: o conjunto dos seus valores possíveis é finito ou infinito numerável - é possível contar

Função de Probabilidade:  $f_X(x) = P(X=x) = \begin{cases} 1_{x_i}, & x = x_i \quad (i=1, 2, \dots) \\ 0, & \text{outros valores de } x \end{cases}$

Propriedades: 1.  $0 \leq f_X(x) \leq 1$     2.  $\sum_i f_X(x_i) = \sum 1_{x_i} = 1$     3.  $P(X \in E) = \sum_{x_i \in E} f_X(x_i)$

Função de Distribuição:  $F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$

Propriedades: 1.  $0 \leq F_X(x) \leq 1$     2.  $\lim_{x \rightarrow -\infty} = 0$ ,  $\lim_{x \rightarrow +\infty} = 1$     3.  $P(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$

$$f_x(x) = P(X=x) = P(x < X \leq x) = F_x(x) - F_x(x)$$

contínua à direita

Média:  $\mu(x) = E(x) = \sum_i x_i P(X=x_i)$

Variância:  $\sigma_x^2 = V(x) = \sum_i (x_i - \mu_x)^2 \times P(X=x_i) = \sum_i x_i^2 \times P(X=x_i) - \mu_x^2$

Experiência de Bernoulli: experiência aleatória apenas com dois resultados possíveis: sucesso (1) e insucesso (0), sendo  $P(X=1) = p$  e  $P(X=0) = 1-p$

$X \sim \text{Ber}(1)$ :  $X$  tem uma distribuição de Bernoulli de parâmetro  $p$

Considerando  $n$  experiências de Bernoulli idênticas e independentes, se render  $X$  a variável aleatória que representa o número de sucessos nas  $n$  experiências, a função de probabilidade de  $X$  é

$$f(x) = P(X=x) = C_n^x p^x (1-p)^{n-x}$$

$X$  tem distribuição binomial com parâmetros  $n$  e  $p$ :  $X \sim Bi(n, p)$

Média:  $E(X) = np$

Variância:  $V(x) = np(1-p)$

Yendo  $Y_1, \dots, Y_k$  variáveis independentes e  $Y_j \sim Bi(n_j, p)$ :  $Y_1 + \dots + Y_k \sim Bi(n_1 + \dots + n_k, p)$

Variável Aleatória Contínua: o conjunto de valores possíveis é infinito: intervalos ou reuniões de intervalos em  $\mathbb{R}$  - associadas a medidas

Função Densidade de Probabilidade:  $P(a < X < b) = \int_a^b f(x) dx$

Propriedades:

1.  $P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$
2.  $P(X \in \mathbb{R}) = \int_{-\infty}^{\infty} f(x) dx = 1$
3.  $P(X = a) = 0$

Função de Distribuição:  $F_X(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$

Propriedades:

1.  $F_X(x)$  é contínua e  $F'_X(x) = f(x)$
2.  $0 \leq F_X \leq 1$
3.  $\lim_{x \rightarrow -\infty} F_X(x) = 0$ ,  $\lim_{x \rightarrow +\infty} F_X(x) = 1$
4.  $P(h_1 < X < h_2) = F_X(h_2) - F_X(h_1)$

Média:  $\mu_X = E(X) = \int_{-\infty}^{+\infty} x f(x) dx$

Variância:  $\sigma_X^2 = V(X) = \int_{-\infty}^{+\infty} (x - \mu_X)^2 f(x) dx$

Propriedades:

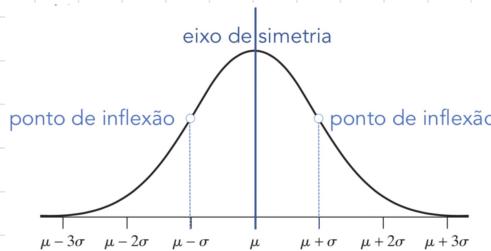
1.  $E(c) = c$
2.  $E(cX) = cE(X)$
3.  $E(aX + b) = aE(X) + b$

Propriedades:

1.  $V(c) = 0$
2.  $V(cX) = c^2 V(X)$
3.  $V(aX + b) = a^2 V(X)$

Distribuição Normal:  $f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2}$

$X$  tem distribuição normal com média  $\mu$  e variância  $\sigma^2$ :  $X \sim N(\mu, \sigma^2)$



Distribuição Normal Padrão: tem média 0 e desvio padrão 1

- $X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$
- $F(z) \equiv \phi(z) = P(Z \leq z)$

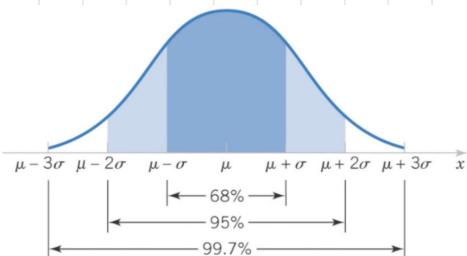
## Avaliação da normalidade

**Regra 68/95/99:** se  $X$  é uma variável aleatória com distribuição normal com média  $\mu$  e desvio padrão  $\sigma$ , então:

$$1. P(\mu - \sigma < X < \mu + \sigma) = 0,6827$$

$$2. P(\mu - 2\sigma < X < \mu + 2\sigma) = 0,9545$$

$$3. P(\mu - 3\sigma < X < \mu + 3\sigma) = 0,9973$$



**Método Gráfico:** compara as observações (valores da amostra) com os valores obtidos admitindo uma determinada distribuição normal

**Gráfico Q-Q:** gráfico constituído por pontos  $(x_i, b_i)$ , onde  $x_i$  é o percentil/quantil  $p_i$  do modelo teórico (distribuição normal) e  $b_i$  é o percentil/quantil  $p_i$  da amostra - permite compara duas distribuições basando os mesmos percentis/quantis.

- Um gráfico Q-Q linear indica uma distribuição normal
- Um gráfico Q-Q com concavidade voltada para cima indica uma distribuição enviesada à direita.
- Um gráfico Q-Q com concavidade voltada para baixo indica uma distribuição enviesada à esquerda.

# Distribuição por Amostragem

Estimador: estimativas que são os valores possíveis de uma função dos elementos da amostra

Distribuição por Amostragem: distribuição de probabilidade de uma variável aleatória baseada numa amostra genérica e cuja variabilidade resulta da aleatoriedade da amostragem

Estatística: variável aleatória que é função de uma amostra aleatória genérica e cuja variabilidade resulta da aleatoriedade da amostragem, nem parâmetros desconhecidos  $Y = f(X_1, \dots, X_n)$

População:  $X$  é variável aleatória que representa uma característica em estudo

Amostra Aleatória: amostra genérica em que  $X_1, \dots, X_n$  são variáveis aleatórias associadas às  $n$  observações

Amostra: amostra específica em que  $x_1, \dots, x_n$  são os valores das variáveis aleatórias  $X_1, \dots, X_n$

Média Amostral: a média de  $\bar{X}$  é igual à média da população:

$$\mu_{\bar{X}} = E(\bar{X}) = \mu_X$$

Desvio Padrão Amostral: o desvio padrão de  $\bar{X}$  é igual ao desvio padrão da população dividido pela raiz quadrada do tamanho da amostra:

Distribuição Amostral: se a população tem distribuição normal, então  $\bar{X}$  também tem distribuição normal

$$\sigma_{\bar{X}} = \sqrt{V(\bar{X})} = \frac{\sigma_X}{\sqrt{n}}$$

Variância Amostral:  $V(\bar{X}) = \frac{\sigma^2}{n}$

1. Sejam  $X_1, \dots, X_n$  v.a. independentes tais que  $X_i \sim N(\mu_i, \sigma_i^2)$ , então:  
 $X_1 + \dots + X_n \sim N(\mu_1 + \dots + \mu_n, \sigma_1^2 + \dots + \sigma_n^2)$

2. Sejam  $X_1, \dots, X_n$  v.a. independentes tais que  $X_i \sim N(\mu, \sigma^2)$ , então:  
 $\bar{X} = \frac{X_1 + \dots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

A variabilidade de  $X$  diminui com o aumento do tamanho da amostra

**Teorema do Límite Central:** se o tamanho da amostra,  $n$ , for grande, então a média amostral  $\bar{X}$  tem distribuição aproximadamente normal

• Sejam  $X_1, \dots, X_n$  uma sucessão de variáveis aleatórias independentes e identicamente distribuídas com  $E(X_i) = \mu$  e  $V(X_i) = \sigma^2$

Então, sendo  $Z \sim N(0, 1)$ :  $\lim_{n \rightarrow \infty} P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) = P(Z \leq z) = \lim_{n \rightarrow \infty} P\left(\frac{\sum_i^n X_i - n\mu}{\sigma/\sqrt{n}} \leq z\right)$

Não grande tem que ser  $n$ ?  $n \geq 30$

### Proposição Amostral

• Seja uma população dicotómica (há apenas duas observações possíveis) em que o sucesso (1) tem probabilidade  $p$  e o insucesso (0) tem probabilidade  $1-p$

• Seja  $\hat{P}$  a proporção de sucessos numa amostra de tamanho  $n$  retirada aleatoriamente dessa população —  $\hat{P}$  é uma média amostral e, para  $n$  grande,  $\hat{P}$  tem distribuição normal

• Seja  $Y$  a variável aleatória representativa do número de sucessos numa amostra de tamanho  $n$

$$\hat{P} = \frac{Y}{n}$$

• Como  $Y \sim Bi(n, p)$ , então  $\mu_Y = np$  e  $\sigma_Y = \sqrt{np(1-p)}$

$$\mu_{\hat{P}} = \frac{1}{n} E(Y) = \frac{1}{n} np = p \quad \sigma_{\hat{P}} = \sqrt{V(\hat{P})} = \sqrt{V\left(\frac{Y}{n}\right)} = \sqrt{\frac{1}{n^2} V(Y)} = \sqrt{\frac{np(1-p)}{n^2}} = \sqrt{\frac{p(1-p)}{n}}$$

• Se  $n$  for grande, a distribuição por amostragem da proporção  $\hat{P}$  pode ser aproximada por uma distribuição normal de média  $p$  e desvio padrão  $\sqrt{\frac{p(1-p)}{n}}$

Se  $\hat{P}$  tem distribuição aproximadamente normal, então  $Y = n\hat{P}$  também tem distribuição aproximadamente normal

Como a distribuição de  $Y$  é binomial, conclui-se que, se  $n$  for grande, a distribuição binomial  $B(n, p)$  pode ser aproximada por uma distribuição normal de média  $\mu = np$  e desvio padrão  $\sigma = \sqrt{np(1-p)}$

**Teorema de Moivre-Laplace:** se  $X \sim B(n, p)$  e  $Y = \frac{X - np}{\sqrt{np(1-p)}}$ , então, se  $n \rightarrow \infty$   $Y \sim N(0,1)$  Y tem uma distribuição aproximadamente normal, com média 0 e variância 1

• Sendo  $X \sim B(n, p)$  e  $Z \sim N(0,1)$ ,  $n \rightarrow \infty$ , então:  $P\left(\frac{X - np}{\sqrt{np(1-p)}} < z\right) \approx P(Z < z)$

**Condições de Aproximação:**

$$\begin{aligned} 1. \quad X &\sim B(n, p) \\ 2. \quad n &> 25 \\ 3. \quad \min\{np, n(1-p)\} &> 5 \end{aligned} \quad \left. \frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1) \right\}$$

**Concessão de Continuidade:** associação de um intervalo a cada valor da variável aleatória discreta

**Concessão de Continuidade de Fisher:** somar e subtrair 0,5

$$P(X \leq k) = P(X < k + 0,5) = P\left(\frac{X - np}{\sqrt{np(1-p)}} < \frac{k + 0,5 - np}{\sqrt{np(1-p)}}\right) \approx \phi\left(\frac{k + 0,5 - np}{\sqrt{np(1-p)}}\right)$$

$$P(X \geq k) = P(X > k - 0,5) = P\left(\frac{X - np}{\sqrt{np(1-p)}} > \frac{k - 0,5 - np}{\sqrt{np(1-p)}}\right) \approx 1 - \phi\left(\frac{k - 0,5 - np}{\sqrt{np(1-p)}}\right)$$

$\phi$ : função de distribuição para a normal reduzida

$$P(X = k) \approx P(k - 0,5 < Y < k + 0,5)$$

# Folha 3

1. Considere a variável aleatória  $X$  com função de probabilidade:

x	-1	0	1	2	3	4	5
f(x)	0.35	0.46	c	0.05	0.02	0.01	0.01

- (a) Determinar  $P(X = 1)$ .
- (b) Determinar:  $P(X \leq 2)$ ,  $P(X < 2)$  e  $P(|X| \leq 1)$ .
- (c) Calcular a média, a moda e a mediana de  $X$ .
- (d) A distribuição será simétrica? Justificar.

a)  $\sum_x f(x) = 1$

$$P(X=1) = 1 - 0,35 - 0,46 - 0,05 - 0,02 - 0,01 - 0,01 = 0,1$$

b)  $P(X \leq 2) = 0,35 + 0,46 + 0,1 + 0,05 = 0,96$

$$P(X < 2) = 0,96 - 0,05 = 0,91$$

$$P(|X| \leq 1) = 0,35 + 0,46 + 0,1 = 0,91$$

c)  $E(X) = \mu_x = \sum_x x f(x) = 0 = -0,35 + 0 + 0,1 + 0,1 + 0,06 + 0,04 + 0,05$

Mediana: menor a tal que  $P(X \leq a) \geq 0,5$

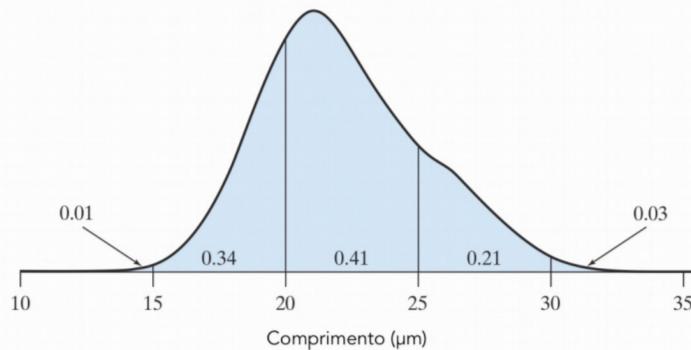
Mediana: 0

Moda: 0

d) Não é simétrico!

## 2. [Adaptado de (\*)]

Numa certa população do parasita *Trypanosoma* o comprimento dos indivíduos é distribuído de acordo com a curva de densidade da figura, onde também estão indicadas a áreas sob a curva.



Considerar  $C$  o comprimento de um indivíduo escolhido ao acaso nesta população e determinar:

- (a)  $P(20 < C < 30)$
- (b)  $P(C > 20)$
- (c)  $P(C < 20)$ .

a)  $P(20 < C < 30) = 0,41 + 0,21 = 0,62$  ✓

b)  $P(C > 20) = 0,41 + 0,21 + 0,03 = 0,65$  ✓

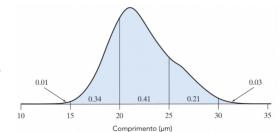
c)  $P(C < 20) = 1 - 0,65 = 0,35$  ✓

3. [Adaptado de (\*)]

Para a distribuição de comprimentos dada pela curva de densidade do exercício 2, admitir agora que se considera uma amostra aleatória com dois trypanosomas.

Qual é a probabilidade de:

- terem ambos comprimento inferior a  $20 \mu m$ ?
- o primeiro ter menos de  $20 \mu m$  de comprimento e o segundo ter mais de  $25 \mu m$ ?
- um deles ter menos de  $20 \mu m$  de comprimento e o outro ter mais de  $25 \mu m$ ?
- exactamente um deles ter menos de  $20 \mu m$ ?
- pelo menos um deles ter menos de  $20 \mu m$ ?



a)  $P(X_1 < 20 \wedge X_2 < 20) = 0,35 \times 0,35 = 0,1225$  ✓

b)  $P(X_1 < 20 \wedge X_2 > 25) = 0,35 \times 0,24 = 0,084$  ✓

c)  $P((X_1 < 20 \wedge X_2 > 25) \vee (X_1 > 25 \wedge X_2 < 20)) = 0,084 \times 2! = 0,168$  ✓

d)  $P((X_1 < 20 \wedge X_2 \geq 20) \vee (X_1 \geq 20 \wedge X_2 < 20)) = 0,35 \times 0,65 \times 2! = 0,455$  ✓

e)  $P(X_1 < 20 \vee X_2 < 20) = 0,1225 + 0,455 = 0,5775 = 1 - (0,65 \times 0,65)$

4. [Adaptado de (\*\*)]

Usando o software R, determinar,

*Probabilidade de sucesso*  
*número de ensaios* ↗

(a)  $P(X = 20)$ ;  $P(X \leq 18)$ ;  $P(X \geq 18)$ ;  $P(20 \leq X \leq 22)$ ; para  $X \sim B(35, 0.6)$ .

(b)  $P(12 \leq X \leq 18)$ ;  $P(10 < X < 18)$ ; para  $X \sim B(20, 0.45)$ .

(c) a:  $P(X \leq a) = 0.9$ , onde  $X \sim B(100, 0.6)$ .

a)  $P(X = 20) = 0,6^{20} \times 0,4^{15} \times C_{20}^{35} = 0,1275$  ✓  $\rightarrow$  binom(20, 35, 0.6)

$P(X \leq 18) = \text{binom}(18, 35, 0.6) = 0,1935$  ✓

$P(20 \leq X \leq 22) = \text{binom}(22, 35, 0.6) - \text{binom}(19, 35, 0.6) = 0,394$

b)  $P(12 \leq X \leq 18) = \text{binom}(18, 20, 0.45) - \text{binom}(11, 20, 0.45) = 0,131$

$P(10 < X \leq 18) = \text{binom}(17, 20, 0.45) - \text{binom}(10, 20, 0.45) = 0,249$  ✓

c)  $P(X \leq a) = 0.9 \Leftrightarrow a = \text{qbinom}(0.9, 100, 0.6) = 66$  ✓

## 5. [Adaptado de (\*)]

A casca do caracol *Limocolaria martensiana* pode ser listada ou lisa. Numa certa população destes caracóis 60% dos indivíduos têm a casca listada e 40% a casca lisa. É retirada uma amostra de 10 caracóis desta população.

(a) Determinar a probabilidade da percentagem de caracóis listados na amostra ser:

- i. 50%;
- ii. 60%;
- iii. 70%.

(b) Determinar o valor esperado e o desvio padrão de  $X$ , que é a v.a. que representa o nº de caracóis (em 10), que têm a casca listada.

$X$ : número de caracóis com casca listada em 10

$$X \sim \mathcal{B}(10, 0.6)$$

a)

i)  $\text{Binom}(5, 10, 0.6) = 0,201$  ✓

ii)  $\text{Binom}(6, 10, 0.6) = 0,251$  ✓

iii)  $\text{Binom}(7, 10, 0.6) = 0,215$  ✓

b)

$$E(X) = n \cdot p$$

$$E(X) = 10 \cdot 0.6 = 6$$

✓

$$V(X) = np(1-p)$$

$$\sigma_X = \sqrt{V(X)} = \sqrt{6 \cdot 0.4} = 1.55$$

✓

6. [Adaptado de (\*)]

Sabe-se que a probabilidade de um filho de dois portadores do gene do *albinismo* ser albino é  $\frac{1}{4}$  e é independente do número de filhos do casal. Se um casal de portadores deste gene tiver 8 filhos, qual é a probabilidade de:

- (a) Nenhum deles ser albino?
- (b) Exactamente dois serem albinos?
- (c) Pelo menos um ser albino?
- (d) O segundo filho ser o primeiro albino?
- (e) O sexto filho ser o segundo albino?

$X$ : número de filhos albinos em 8       $X \sim B(8, \frac{1}{4})$

a)  $P(X=0) = \text{Binom}(0, 8, \frac{1}{4}) = 0,1 = 0,1001$  ✓

b)  $P(X=2) = \text{Binom}(2, 8, \frac{1}{4}) = 0,3$  ✓

c)  $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X=0) = 1 - \text{Binom}(0, 8, \frac{1}{4}) = 0,9$  ✓

d)  $P = \frac{3}{4} \times \frac{1}{4} = \frac{3}{16}$  ?      e)  $P = 5 \times \frac{1}{4} \times \left(\frac{3}{4}\right)^4 \times \frac{1}{4} = 0,1$  ?

7. [Adaptado de (\*)]

Numa certa população uma em cada oito crianças tem um nível de chumbo no sangue superior a  $30\mu\text{g}/\text{dl}$ , classificado pelas autoridades de saúde como *nível elevado*.

Num grupo de 16 crianças escolhidas aleatoriamente nessa população qual é a probabilidade de:

- (a) Exactamente duas delas terem nível elevado de chumbo no sangue?

- (b) Nenhuma delas ter?

- (c) Três ou mais terem nível elevado?

$$P(X=k) = 1^k (1-1)^{n-k} \times C_n^k$$

$X$ : número de crianças com nível de chumbo elevado em 16

$X \sim B(16, 0.125)$

a)  $P(X=2) = \text{Binom}(2, 16, \frac{1}{8}) = 0,289 = \left(\frac{1}{8}\right)^2 \times \left(\frac{7}{8}\right)^{14} \times C_2^{16} = P(X \leq 2) - P(X \leq 1) = \frac{0,7892 + 0,5613}{2} - \frac{0,5143 + 0,0001}{2}$

b)  $P(X=0) = \text{Binom}(0, 16, \frac{1}{8}) = 0,118$

c)  $P(X \geq 3) = 1 - \text{Binom}(2, 16, \frac{1}{8}) = 0,323$

## Breve nota sobre a Distribuição de Poisson

Frequentemente os resultados de uma experiência aleatória são expressos como o número de objetos num certo espaço ou o número de ocorrências de um acontecimento num certo intervalo de tempo. Por exemplo, o número de nascimentos diárias numa maternidade, ou o número de bactérias de um certo tipo existentes num ml de água. A distribuição de Poisson é uma distribuição de probabilidade frequentemente usada para modelar estas situações. As suposições básicas para a utilização do modelo são:

- A taxa média de ocorrência do acontecimento ( $\lambda$ ) é constante ao longo do tempo (espaço).
- A informação sobre o número de ocorrências num certo período (espaço) não tem qualquer influência sobre o número de ocorrências num outro período (ou espaço) disjunto.

Diz-se que a v.a  $X$  tem distribuição de Poisson se a f.p. é da forma

$$f(x) = f(x; \lambda) = \begin{cases} e^{-\lambda} \frac{\lambda^x}{x!} & \text{para } x = 0, 1, \dots; \lambda > 0, \\ 0 & \text{caso contrário.} \end{cases}$$

- Usa-se a notação  $X \sim P(\lambda)$ .
- Mostra-se que  $E(X) = V(X) = \lambda$ .

Sejam  $X_1, X_2, \dots, X_k$  v.a. independentes com distribuição de Poisson,  $X_i \sim P(\lambda_i)$ ,  $i = 1, 2, \dots, k$ . Então a soma também segue uma distribuição de Poisson

$$X_1 + X_2 + \dots + X_k \sim P\left(\sum_{i=1}^k \lambda_i\right)$$

8. [Adaptado de (\*)]

Seja  $X$  a v.a. que representa o número de bactérias de um certo tipo existentes num  $\text{cm}^3$  de água. Supor que  $X$  tem distribuição de Poisson de parâmetro  $\lambda = 3$ .

$$X \sim P(3)$$

- Qual é a probabilidade de não haver bactérias num  $\text{cm}^3$  de água?
- Determine a probabilidade de que num  $\text{cm}^3$  de água existam pelo menos 3 bactérias.
- Qual a probabilidade de que numa amostra de dois  $\text{cm}^3$  de água existam quando muito 4 bactérias?

a)  $f(0) = f(0; 3) = e^{-3} \times \frac{3^0}{0!} = e^{-3} \times 1 = e^{-3} = 0,0499 \checkmark$   $Pois(0, 3) = f_{Pois}(0, 3)$

b)  $P(X \geq 3) = 1 - P_{Pois}(2, 3) = 0,577 \checkmark$

c)  $X$ : número de bactérias em  $1 \text{ cm}^3$  de água  
 $Y$ : número de bactérias em  $2 \text{ cm}^3$  de água

$$Y = X_1 + X_2 \quad Y \sim P(3+3) \Leftrightarrow Y \sim P(6) \quad P(Y \leq 4) = P_{Pois}(4, 6) = 0,285 \checkmark$$

9. Os nascimentos num certo hospital ocorrem aleatoriamente a uma taxa média de 0.8 por hora. Assumindo o modelo de Poisson, determinar a probabilidade de

- (a) não ocorrerem nascimentos numa certa hora nesse hospital.
- (b) ocorrerem mais do que 2 nascimentos numa certa hora nesse hospital.

$X$ : número de nascimentos numa hora  $\quad X \sim P(0.8)$

a)  $P(X=0) = \text{Pois}(0, 0.8) = e^{-0.8} \times \frac{0.8^0}{0!} = e^{-0.8} = 0.449$  ✓

b)  $P(X > 2) = 1 - P(X \leq 2) = 1 - \text{Pois}(2, 0.8) = 1 - \left( e^{-0.8} + e^{-0.8} \times \frac{0.8^1}{1!} + e^{-0.8} \times \frac{0.8^2}{2!} \right) = 0.0474$

10. Num estudo envolvendo uma certa espécie de ave, um dos interesses recai no número de ovos postos. Em 80 ninhos encontrados, o número médio de ovos por ninho foi 3.8 e a variância foi 3.1. Uma vez que a variância e a média são aproximadamente iguais, supõe-se que o número de ovos por ninho segue uma distribuição de Poisson com média 3.8. Se esta for realmente a distribuição populacional,

- (a) Qual é a probabilidade de se encontrar um ninho com 4 ovos?
- (b) Qual é a probabilidade de se encontrar um ninho com menos do que 3 ovos?
- (c) Supondo que se acabou de encontrou um ninho com 6 ovos, qual a probabilidade de se voltar a encontrar um ninho com 6 ovos?

$X$ : número de ovos por ninho  $\quad X \sim P(3.8)$

a)  $P(X=4) = \text{Pois}(4, 3.8) = e^{-3.8} \times \frac{3.8^4}{4!} = 0.194$  ✓

b)  $P(X < 3) = P(X \leq 2) = \text{Pois}(2, 3.8) = e^{-3.8} \times \left( 1 + 3.8 + \frac{3.8^2}{2!} \right) = 0.269$  ✓

c)?  $P(X=6)^2 = \text{Pois}(6, 3.8)^2 = \left( e^{-3.8} \times \frac{3.8^6}{6!} \right)^2 = 0.00875$  ✗

## 11. [Adaptado de (\*)]

Assumindo que o crescimento decorrido num período de 15 dias para uma população de girassóis segue uma distribuição normal com média 3.18 cm e desvio padrão 0.53 cm, que percentagem de plantas cresce:

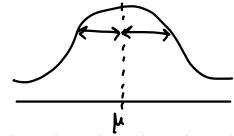
(a) 3 cm ou menos?

(b) 4 cm ou mais?

(c) entre 2.5 e 3.5 cm?

$$X \sim N(\mu, \sigma^2)$$

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



X: crescimento em 15 dias de um girassol em cm       $X \sim (3,18; 0,53^2)$

a)  $P(X < 3) = P\left(\frac{X-3,18}{0,53} < \frac{3-3,18}{0,53}\right) \approx P(Z < -0,34) = P(Z > 0,34) = 1 - P(Z < 0,34) = 1 - 0,6331 = 0,3669$  ✓

b)  $P(X \geq 4) = P\left(\frac{X-3,18}{0,53} \geq \frac{4-3,18}{0,53}\right) \approx P(Z \geq 1,55) = 1 - P(Z < 1,55) = 1 - 0,9394 = 0,0606$  ✓

c)  $P(2,5 < X < 3,5) = P\left(\frac{2,5-3,18}{0,53} < \frac{X-3,18}{0,53} < \frac{3,5-3,18}{0,53}\right) = P(Z < 0,60) - P(Z < -1,3) = 0,7257 - P(Z > 1,3) = 0,7257 - (1 - P(Z < 1,3)) = 0,7257 - 1 + P(Z < 1,3) = 0,7257 - 1 + 0,1032 = 0,6289$  ✓

12. Seja  $X \sim N(0.5, 0.04) = N(0.5, 0.2^2)$ 

(a) Calcular  $P(X < 0.3)$  e  $c$  tal que  $P(0.5 - c \leq X \leq 0.5 + c) = 0.95$ , utilizando as tabelas da distribuição normal;

(b) os mesmos cálculos que em (a), mas utilizando o software R (\*\*).

$$X \sim N(0.5; 0.04) = N(0.5; 0.2^2)$$

a)  $P(X < 0.3) = P\left(Z < \frac{0.3-0.5}{0.2}\right) = P(Z < -1) = P(Z > 1) = 1 - P(Z < 1) = 1 - 0.8413 = 0.1587 = \text{norm}(0.3, 0.5, 0.04)$

b)  $P(0.5 - c \leq X \leq 0.5 + c) = P\left(\frac{0.5-c-0.5}{0.2} \leq Z \leq \frac{0.5+c-0.5}{0.2}\right) = P(-0.2 \leq Z \leq 0.2) = P(-S_c \leq Z \leq S_c) = P(Z < S_c) - P(Z < -S_c) = P(Z < S_c) - 1 + P(Z < -S_c) = 2P(Z < S_c) - 1$

$-1 + 2P(Z < S_c) = 0.95 \Leftrightarrow P(Z < S_c) = 0.975 \Leftrightarrow S_c = 1.96 \Leftrightarrow c = 0.392$  ✓

13. Sendo  $X$  uma variável aleatória normal com média 7 e variância 9, determinar, usando a tabela da distribuição normal e o software R(\*\*):

(a)  $P(X < 8)$ ,  $P(X > 4.5)$ ,  $P(4 \leq X < 10)$ .

(b)  $a$ , tal que  $P(X > a) = 0.25$ ;  $b$ , tal que  $P(X < b) = 0.10$ . Que representam  $a$  e  $b$ ?

(c)  $P(Y > 12)$ , com  $Y = X_1 + X_2$ , onde  $X_1$  e  $X_2$  são v.a. independentes com a mesma distribuição de  $X$ .

$$X \sim N(7, 3^2)$$

a)  $P(X < 8) = \text{Pr}_{\text{norm}}(8, 7, 3) = P\left(Z < \frac{8-7}{3}\right) = P\left(Z < \frac{1}{3}\right) = 0,6293$  ✓

$$\begin{aligned} P(X > 4.5) &= 1 - P(X < 4.5) = 1 - P\left(Z < \frac{4.5-7}{3}\right) = 1 - P(Z < -0.83) = \\ &= 1 - \text{Pr}_{\text{norm}}(4.5, 7, 3) = 1 - P(Z > 0.83) = \\ &= P(Z < 0.83) = 0.7967 \end{aligned}$$

$$\begin{aligned} P(4 < X < 10) &= P(X < 10) - P(X < 4) = P(Z < 1) - P(Z < -1) = P(Z < 1) - P(Z > 1) = \\ &= \text{Pr}_{\text{norm}}(10, 7, 3) - \text{Pr}_{\text{norm}}(4, 7, 3) = P(Z < 1) - 1 + P(Z < 1) = 2P(Z < 1) - 1 = \\ &= 2 \times 0,8413 - 1 = 0,6826 \end{aligned}$$

b)  $P(X > a) = 0,25 \Leftrightarrow 1 - P(X \leq a) = 0,25 \Leftrightarrow 0,75 = P(X \leq a)$

$$\Leftrightarrow 0,75 = P\left(Z \leq \frac{a-7}{3}\right) \Leftrightarrow \frac{a-7}{3} = 0,675 \Leftrightarrow a = 9,025$$

$$P(X < b) = 0,10 \Leftrightarrow 1 - P(X > b) = 0,10 \Leftrightarrow P(X > b) = 0,9 \Leftrightarrow$$

$$\Leftrightarrow P\left(Z > \frac{b-7}{3}\right) = 0,9 \Leftrightarrow P\left(Z \leq -\frac{b-7}{3}\right) = 0,9 \Leftrightarrow \frac{7-b}{3} = 1,28 \Leftrightarrow b = 3,16$$

c)  $X_1 \sim N(\mu_1, \sigma_1^2) \wedge X_2 \sim N(\mu_2, \sigma_2^2) \wedge X_1, X_2 \text{ independentes} \Rightarrow X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

$$Y = X_1 + X_2 \sim N(7+7, 9+9) = N(14, 18) = N(14, \sqrt{18}^2)$$

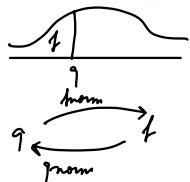
$$P(Y > 12) = 1 - P(Y \leq 12) = 1 - P\left(Z \leq \frac{12-14}{\sqrt{18}}\right) = 1 - P(Z \leq -0,47) = P(Z > 0,47) =$$

$$= P(Z < 0,47) = 0,6808 =$$

14. Assume-se que o diâmetro de uma população de oliveiras segue uma distribuição normal de média 17 cm e desvio padrão 2 cm.

- Qual é a probabilidade do diâmetro de uma oliveira, escolhida ao acaso nesta população, estar compreendido entre 14 e 20 cm?
- Qual é a probabilidade do diâmetro de uma oliveira ser inferior a 18 cm?
- Para efeitos de tratamento pretende-se classificar as oliveiras de acordo com o diâmetro do seguinte modo:

- tipo A - oliveiras correspondentes aos 20% dos menores diâmetros
- tipo B - oliveiras com os 50% dos valores seguintes dos diâmetros
- tipo C - oliveiras com os 30% maiores diâmetros.



Determinar os limites do diâmetro para cada classe.

$X$ : diâmetro de uma oliveira na população  $X \sim N(17, 2^2)$

$$\text{a)} P(14 < X < 20) = P(X < 20) - P(X < 14) = P\left(Z < \frac{3}{2}\right) - P\left(Z < -\frac{3}{2}\right) = \\ = P(Z < 1,5) - P(Z > 1,5) = P(Z < 1,5) - 1 + P(Z < -1,5) = 2 \times 0,9332 - 1 = 0,8664 \quad \checkmark$$

$$\text{b)} P(X < 18) = P\left(Z < \frac{18-17}{2}\right) = P(Z < 0,5) = 0,6915 \quad \checkmark$$

$$\text{c)} \begin{aligned} A: \quad P(X < a) &= 0,10 \Leftrightarrow P\left(Z < \frac{a-17}{2}\right) = 0,10 \Leftrightarrow P\left(Z < \frac{17-a}{2}\right) = 0,10 \\ \Leftrightarrow \frac{17-a}{2} &= 0,84 \Leftrightarrow a = 15,32 \quad \checkmark \end{aligned}$$

$$A = [0; 15,32]$$

$$\begin{aligned} B: \quad P(X < b) &= 0,7 \Leftrightarrow P\left(Z < \frac{b-17}{2}\right) = 0,7 \Leftrightarrow \frac{b-17}{2} = 0,52 \\ \Leftrightarrow b &= 18,04 \end{aligned}$$

$$B = [15,32; 18,04[$$

$$C = [18,04; +\infty[$$

121	82	100	151	68	58	48	95	42
95	145	64	201	101	25	123	70	163
84	57	139	60	78	93	92	110	94
119	104	110	113	118	62	83	67	203

A média da amostra é 98.3 U/l e o desvio padrão é 40.4 U/l

- (a) Que percentagem das observações está a menos de um desvio padrão (dp) da média? E a 2 dp? E a 3 dp? Analisar os resultados obtidos.
- (b) Construir um histograma e desenhar o diagrama caixa-dos-bigodes para esta amostra
- (c) Tendo em conta as representações obtidas, será de sugerir alguma distribuição como aproximação da distribuição de probabilidade da população subjacente?

a)

$\pm 1\text{ dp}:$	26	$\rightarrow 72\%$	}
$\pm 2\text{ dp}:$	34	$\rightarrow 94\%$	
$\pm 3\text{ dp}:$	36	$\rightarrow 100\%$	

? cumprido  $68/95/99 \rightarrow ?$  é normal

Considerar a amostra do exercício 3 da Folha 1 (DadosFolha1.csv), relativa a observações da magnitude, na escala de Richter, de 30 sismos na Califórnia:

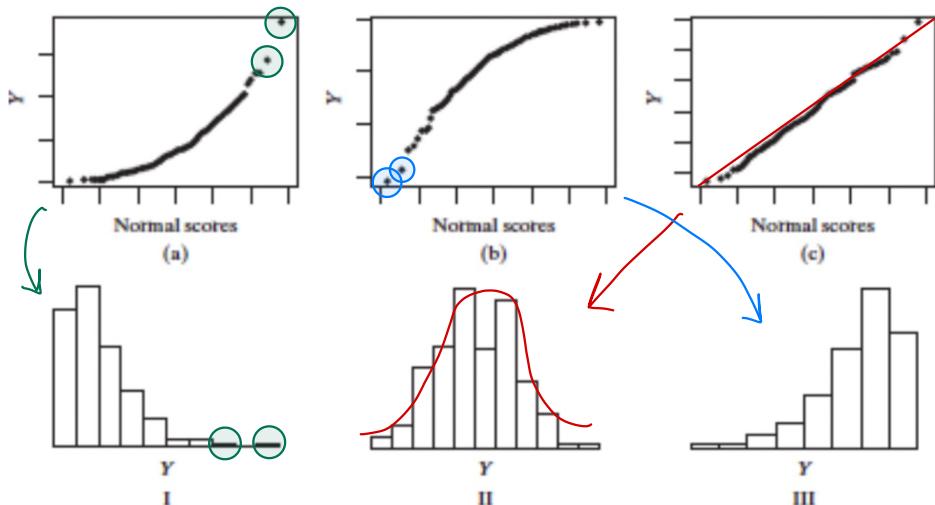
1.0	8.3	3.1	1.1	5.1
1.2	1.0	4.1	1.1	4.0
2.0	1.9	6.3	1.4	1.3
3.3	2.2	2.3	2.1	2.1
1.4	2.7	2.4	3.0	4.1
5.0	2.2	1.2	7.7	1.5

$$\left[ \sum_{i=1}^{30} x_i = 86.1 \sum_{i=1}^{30} x_i^2 = 357.61 \right]$$

investigar se é razoável admitir que esta amostra provém de uma distribuição normal, através de um gráfico QQ.



Os seguintes gráficos QQ dizem respeito a 3 amostras, representadas abaixo através de histogramas.  
Associar os gráficos QQ aos respetivos histogramas.



Considerar a amostra constituída pelos valores  $x_i$  (ficheiro Ex18F3.csv disponível no Moodle).

- Investigar se é razoável admitir que esta amostra provém de uma distribuição normal, através de um gráfico QQ complementado com um histograma.
- Repetir o estudo feito em na alínea anterior para os dados transformados  $y_i = \log(x_i)$ . A conclusão é a mesma?

a) Concordade para cima  $\rightarrow$  envergada à direita



b) NÃO, JÁ É NORMAL! (QQ linear)



19. O famoso antropólogo Sir Francis Galton, obteve no Laboratório de Antropologia na Exposição Internacional de 1884 diversas medidas relativas a adultos, algumas das quais se encontram sumariadas na tabela abaixo. Supor que todas as variáveis em estudo têm distribuições normais com médias e desvios padrão dados nessa tabela.

Variável	Homens		Mulheres	
	$\mu$	$\sigma$	$\mu$	$\sigma$
altura (cm)	172	7.2	161	6.6
envergadura (cm)	178	7.6	160	7.4
massa (kg)	64.9	7.03	55.8	6.49

- (a) i. Qual a proporção de homens que medem menos do que 160 cm?  
ii. Acima de que massa (em kg) estão os 10% dos homens mais pesados?  
iii. Abaixo de que valor medem as 20% mulheres mais baixas?
- (b) Indicar a distribuição das variáveis seguintes:  
i. Média das alturas (em cm) de 24 mulheres selecionadas aleatoriamente.  
ii. Envergadura média (em cm) de 10 homens selecionados aleatoriamente.  
iii. Média das massas (em kg) de 5 mulheres selecionadas aleatoriamente.
- (c) Determinar a probabilidade da altura média de 10 homens selecionados aleatoriamente ser superior a 165 cm.

a)

i.  $h \sim N(172, 7.2^2)$   $P(h < 160) = P\left(Z < \frac{160 - 172}{7.2}\right) = P\left(Z < -\frac{12}{7.2}\right) = P(Z < -1.66) = 1 - P(Z < 1.66) = 1 - 0.9525 = 0.0475$  ✓

ii.  $m \sim N(64.9, 7.03^2)$   $P(m > ?) = 0.1 \Leftrightarrow 1 - P(m < ?) = 0.1$   
 $\Leftrightarrow P(m < ?) = 0.9 \Leftrightarrow P\left(Z < \frac{? - 64.9}{7.03}\right) = 0.9 \Leftrightarrow \frac{? - 64.9}{7.03} = 1.285$   
 $\Leftrightarrow ? = 73.9$  ✓

iii.  $h \sim N(161, 6.6^2)$   $P(h < ?) = 0.2 \Leftrightarrow 1 - P(h < -?) = 0.2$   
 $\Leftrightarrow P(h < -?) = 0.8 \Leftrightarrow P\left(Z < \frac{-? + 161}{6.6}\right) = 0.8 \Leftrightarrow \frac{-? + 161}{6.6} = 0.845$   
 $\Leftrightarrow ? = 155$  ✓

b) i.  $\bar{X} \sim N\left(161, \frac{6.6^2}{24}\right)$  ✓ ii.  $\bar{X} \sim N\left(178, \frac{7.6^2}{10}\right)$  ✓ iii.  $\bar{X} \sim N\left(55.8, \frac{6.49^2}{5}\right)$  ✓

c)  $\bar{X} \sim N\left(172, \frac{7.2^2}{10}\right)$   $P(\bar{X} > 165) = 1 - P\left(Z < \frac{165 - 172}{7.2^2/10}\right) = 1 - P(Z < 3.07) = 0.07$  ✓

20. Registou-se o crescimento decorrido num período de 15 dias em 50 girassóis selecionados ao acaso na população de girassóis considerada no exercício 11.

- Calcular a probabilidade de que pelo menos 60% das plantas cresçam mais do que 3 cm (sem recurso ao software R).
- Qual a probabilidade de se observar um crescimento inferior a 3 cm em 30% das plantas?
- Desses girassóis quantos se espera que tenham um crescimento de pelo menos 4 cm?

$$X \sim N(3.18, 0.53^2)$$

a)  $P(X > 3) = P\left(Z > \frac{3-3.18}{0.53}\right) = P(Z > -0.34) = P(Z < 0.34) = 0,6331$  ✓

$$Y \sim B(50, 0.6331) \quad n > 25 \quad \min\{ny, n(1-y)\} > 5$$

$$P(Y \geq 30) = P(Y \geq 29,5) = P\left(Z \geq \frac{29,5 - 50 \times 0,6331}{\sqrt{50 \times 0,6331 \times (1-0,6331)}}\right) = P(Z \geq -0,63) = 1 - P(Z \leq 0,63) = 0,7357$$

b)  $P(X < 3) = P(Z < -0,34) = 1 - P(Z < 0,34) = 1 - 0,6331 = 0,3669$  ✓

$$Y \sim B(50, 0.3669) \quad n > 25 \quad \min\{ny, n(1-y)\} > 5$$

$$\begin{aligned} P(Y = .3 \times 50) &= P(Y = 15) = P(14,5 < Y < 15,5) = P(-1,13 < Z < -0,84) = \\ &= P(Z > -1,13) - P(Z < -0,84) = 1 - P(Z < 1,13) - 1 + P(Z < 0,84) = 0,8708 - 0,7915 = 0,0793 \end{aligned}$$
 ✓

c)  $P(X > 4) = P(Z > 1,55) = 1 - P(Z < 1,55) = 1 - 0,94 = 0,06$

$$0,06 \times 50 = 3$$
 ✓

21. Em Portugal cerca de 42.3% da população tem sangue do tipo O. Supor que é selecionada aleatoriamente uma amostra de 100 pessoas. Determinar a probabilidade de que 50 delas tenham sangue do tipo O, usando

- (a) a fórmula da distribuição binomial
- (b) a aproximação pela distribuição normal.

a)  $X \sim B(100, 0.423)$

$$P(X=50) = \frac{1^k \times (1-p)^{n-k} \times C_k^n}{n!} = 0,423^{50} \times (1-0,423)^{50} \times \frac{100!}{50!} = 24%$$

b)  $n > 25$  e  $\min \left\{ np = 42.3, n(1-p) = 57.7 \right\} = 42.3 > 5$

$$\frac{X - np}{\sqrt{np(1-p)}} \sim N(0, 1) \quad P(X \geq 50) = P(X > 50 - 0.5) = P(X > 49.5) =$$

$$= P\left(\frac{X - 42.3}{\sqrt{42.3 \times 0.577}} > \frac{49.5 - 42.3}{\sqrt{42.3 \times 0.577}}\right) = P\left(\frac{X - 42.3}{4.94} > 1.457\right) =$$

$$= 1 - P\left(\frac{X - 42.3}{4.94} < 1.457\right) = 0.9279$$

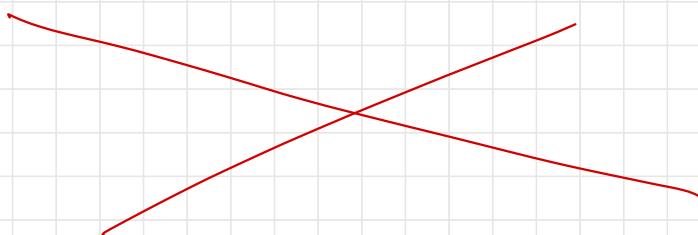
Portanto  $P(X=50, 100, 42.3) = 0.9279$  ...

22. (Para resolver utilizando o R)

Obter uma amostra aleatória de dimensão 10 de uma população com distribuição Binomial com parâmetros  $n = 20$  e  $p = 0.3$ .

Determinar a média, o desvio padrão e a mediana, e desenhar o histograma e o diagrama caixa de bigodes.

Repetir o exercício com amostras de dimensão 100, 1000. O que se observa?



23. O comprimento médio dos insetos de determinada espécie é 15 mm, e o respetivo desvio padrão é 3 mm. Vai ser selecionada uma amostra aleatória de 49 insetos dessa espécie.

- Qual a probabilidade aproximada da média dos comprimentos dos insetos da amostra ser superior a 16 mm?
- Qual a probabilidade aproximada da média dos comprimentos dos insetos da amostra se encontrar a menos de 3 mm da média populacional?

$$X \sim N(15, 3^2)$$

a)  $\bar{X} \sim N\left(15, \frac{3^2}{49}\right)$   $P(\bar{X} > 16) = 1 - P(Z < 2,33) = 1 - 0,9901 = 0,0099$

b)  $P(|\bar{X} - 15| < 3) = P(12 < \bar{X} < 18) = P(\bar{X} < 18) - P(\bar{X} < 12) = P(Z < 2) - P(Z < -2) = 1 - 0 = 1$

24. O diâmetro médio das laranjas de determinada espécie é 7.8 cm, e o respetivo desvio padrão é 1.6 cm. Vai ser selecionada uma amostra aleatória de 100 laranjas dessa espécie.

Calcular a probabilidade aproximada da média dos diâmetros das laranjas na amostra

- ser inferior ao diâmetro médio das laranjas dessa espécie;
- diferir da média populacional mais do que 5 mm.

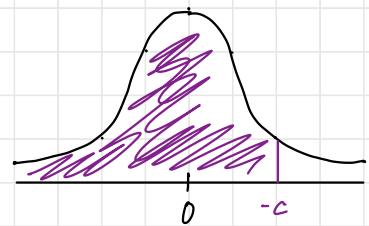
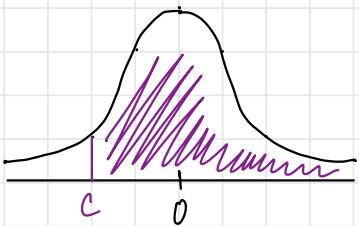
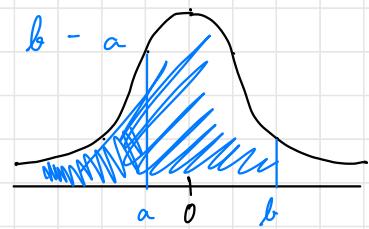
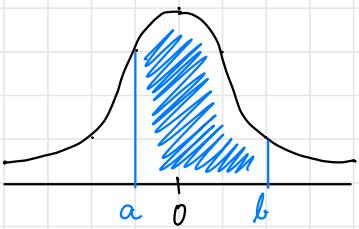
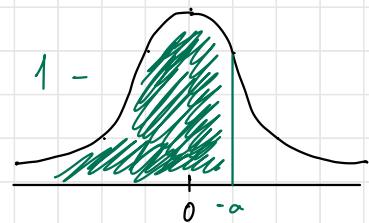
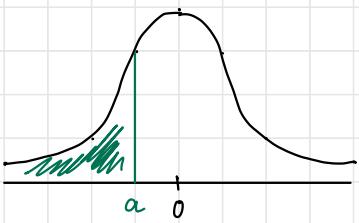
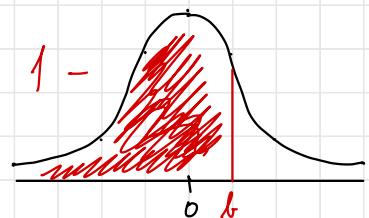
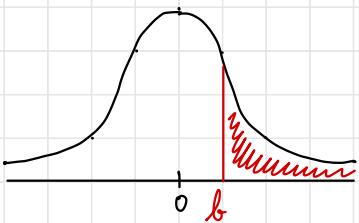
$$\bar{X} \sim N\left(7.8, \frac{1.6^2}{100}\right)$$

a)  $P(\bar{X} < 7.8) = P(Z < 0) = 0.5$

b)  $P(|\bar{X} - 7.8| > 0.5) = P(X < 7.3) + P(X > 8.3) = P(X < 7.3) + 1 - P(X < 8.3)$

$$= P(Z < -3,125) + 1 - P(Z < 3,125) =$$

$$= 1 - P(Z < 3,125) = 2 - 2P(Z < 3,125) = 2 - 2 \times 0,9991 = 0,0018$$



# RESUMO

Média:  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

$$\cdot \bar{y} = ax + b$$

Mediana: observação central, depois de ordenados os dados

Quantil  $Q_k = \begin{cases} \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \frac{n}{2} \in \mathbb{Z} \\ x_{k+1}, & \frac{n}{2} \notin \mathbb{Z} \end{cases}$

Extremos e Quantis: mínimo;  $Q_1$ ;  $Q_2$ ;  $Q_3$ ; máximo

Caixa e Bigodes:

- $B_1$ : menor observação  $\geq B_1 = Q_1 - 1,5 AIQ$
- $B_3$ : maior observação  $\leq B_3 = Q_3 + 1,5 AIQ$

Amplitude Interquantil:  $M - m$

Amplitude Interquantil:  $AIQ = Q_3 - Q_1$

Derroio Padrão:  $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$

$$\cdot s_y = a_1 x, \quad y = ax + b$$

Variância:  $s^2 = \frac{\sum_{i=1}^n x_i^2 - n \bar{x}^2}{n-1}$

Coeficiente de Variação:  $CV = \frac{s}{\bar{x}}$

Covariância:  $r_{x,y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{n-1}$

Correlação:  $r = \frac{1}{n-1} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$

Probabilidade:  $P(A) = \frac{\text{A favoráveis}}{\text{Total possíveis}}$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cap \bar{B}) = P(A) - P(A \cap B)$

Probabilidade Condicionada:  $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Acidentes Independentes:  $P(A \cap B) = P(A) \times P(B)$

Teorema da Probabilidade Total:  $P(B) = P(A_1) \times P(B|A_1) + \dots + P(A_m) \times P(B|A_m)$

Teorema de Bayes:  $P(A_i|B) = \frac{P(A_i) \times P(B|A_i)}{P(A_1) \times P(B|A_1) + \dots + P(A_m) \times P(B|A_m)}$

Distribuição Binomial:  $f(n) = P(X=n) = C_n^m p^n (1-p)^{n-m}$ ,  $X \sim B(m, p)$

Média:  $E(X) = np$       Variância:  $V(X) = np(1-p)$

$\cdot Y_1 + \dots + Y_k \sim B(m_1 + \dots + m_k, p)$

Variável Aleatória Discreta:

- $f_X(x) = P(X=x) = p_x$
- $F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$

$\cdot f_X(x) = P(X=x) = P(x \leq X \leq x) = F_X(x) - F_X(x^-)$

Média:  $\mu(X) = E(X) = \sum_i x_i p(X=x_i)$

Variância:  $\sigma_X^2 = V(X) = \sum_i x_i^2 p(X=x_i) - \mu_X^2$

Variável Aleatória Contínua:

- $P(a < X < b) = \int_a^b f(x) dx$
- $F_X(x) = P(X \leq x) = \int_{-\infty}^x (x - \mu_x)^2 f(x) dx$

Média:  $\mu(X) = E(X) = \int_{-\infty}^{+\infty} x f(x) dx$

Variância:  $\sigma_X^2 = V(X) = \int_{-\infty}^{+\infty} (x - \mu_x)^2 f(x) dx$

Distribuição Normal:  $X \sim N(\mu, \sigma^2)$

Distribuição Normal Padrão:  $X \sim N(0, 1)$

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

$$F(z) = \Phi(z) = P(Z \leq z)$$

Regra 68/95/99:

1.  $P(\mu - \sigma < X < \mu + \sigma) = 0,68$
2.  $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0,95$
3.  $P(\mu - 3\sigma < X < \mu + 3\sigma) = 0,99$

Grafico QQ U → simétrica à direita

Grafico QQ Λ → simétrica à esquerda

$$X_1 \sim N(\mu_1, \sigma_1^2) \wedge X_2 \sim N(\mu_2, \sigma_2^2) \wedge X_1, X_2 \text{ independentes} \Rightarrow X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

A distribuição normal padrão é simétrica em relação à origem e a área total é 1

Média amostral:  $\mu_x = E(\bar{X}) = \mu_{\bar{X}}$

Variância amostral:  $V(\bar{X}) = \frac{\sigma^2}{n}$

Desvio Padrão Amostral:  $\sigma_{\bar{X}} = \sqrt{V(\bar{X})} = \frac{\sigma_x}{\sqrt{n}}$

$$X_1 + \dots + X_n \sim N(\mu_1 + \dots + \mu_n, \sigma_1^2 + \dots + \sigma_n^2)$$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$1. X \sim B(n, p)$$

$$2. n > 25$$

$$3. \min\{np, n(1-p)\} > 5$$

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

$$P(X=k) \approx P(k-0,5 < Y < k+0,5)$$

1. A função de probabilidade de uma variável aleatória  $X$ , é dada por:

$x$	0	1	2	3	4
$f(x)$	a	0.1	0.1	0.1	b

Os valores de  $a$  e  $b$  são reais e pertencem ao intervalo  $[0, 1]$ .

Sabendo que a média da variável aleatória  $X$ , é 1, os valores de  $a$  e  $b$  são,

- $a = 0.1$  e  $b = 0.6$ ;  
  $a = 0.7$  e  $b = 0.0$ ;  
  $\cancel{a = 0.6}$  e  $b = 0.1$ ;  
  $a = 0.0$  e  $b = 0.7$ ;  
 nenhuma das hipóteses anteriores está correta.

$$0a + 1 \times 0,1 + 2 \times 0,1 + 3 \times 0,1 + 4b = 1 \quad \Leftrightarrow \quad 0,6 + 4b = 1$$

$$\Leftrightarrow 4b = 0,4$$

$$a + 0,1 + 0,1 + 0,1 + 0,1 = 1 \quad \Leftrightarrow \quad a = 0,6 \quad \Leftrightarrow \quad b = 0,1$$

2. Num estudo sobre o consumo de cocaína foram entrevistados 75 homens e 36 mulheres consumidores.

Os resultados podem ser visualizados na seguinte tabela:

	Homens	Mulheres	Total
A (1 a 19 vezes)	32	7	39
B (20 a 99 vezes)	18	20	38
C ( $\geq 100$ vezes)	25	9	34
<b>Total</b>	<b>75</b>	<b>36</b>	<b>111</b>

Escolhendo ao acaso um indivíduo dos 111, qual é probabilidade de que seja um homem, e que tenha usado cocaína menos de 20 vezes?

- 71.18%  
 93.69%  
  $\cancel{28.82\%}$   
 6.31  
 nenhuma das hipóteses anteriores está correta.

$$P(H \cap A) = \frac{32}{111}$$

3. O peso dos queijos artesanais produzidos numa pequena empresa, segue uma distribuição normal com média  $\mu = 500$  g e desvio padrão  $\sigma = 120$  g. Seja  $\bar{X}$  a variável aleatória representativa da média (em gramas) dos pesos de 16 desses queijos selecionados aleatoriamente, e sejam  $\mu_{\bar{X}}$  e  $\sigma_{\bar{X}}$  a média e o desvio padrão de  $\bar{X}$ .

Então a média e o desvio padrão de  $\bar{X}$ , são,

$\mu_{\bar{X}} = 31.25$  e  $\sigma_{\bar{X}} = 7.5$ ;

$\mu_{\bar{X}} = 500$  e  $\sigma_{\bar{X}} = 30$ ;

$\mu_{\bar{X}} = 31.25$  e  $\sigma_{\bar{X}} = 30$ ;

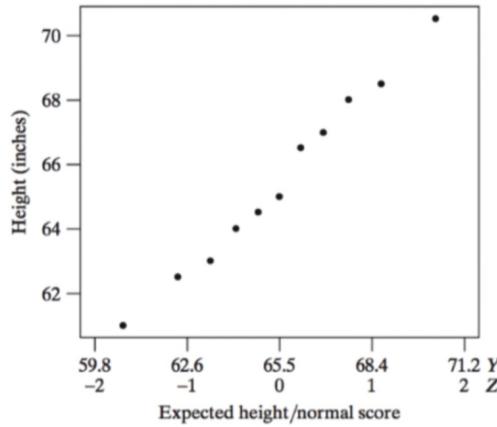
$\mu_{\bar{X}} = 500$  e  $\sigma_{\bar{X}} = 120$ ;

nenhuma das hipóteses anteriores está correta.

$$\mu_{\bar{X}} = \mu_X = 500$$

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{120}{\sqrt{16}} = 30$$

4. Um gráfico QQ é um gráfico que permite comparar duas distribuições (a distribuição dos valores da amostra e uma distribuição teórica) traçando os seus percentis/quantis. Posto isto, uma relação aproximadamente linear, como exemplificado na figura, indica que:



- Os valores observados (amostra) não são concordantes com os valores da distribuição teórica
- Os valores observados (amostra) seguem uma distribuição normal
- Os valores observados (amostra) seguem uma distribuição t-student
- Os valores observados (amostra) são concordantes com os valores da distribuição teórica
- Nenhuma das hipóteses anteriores está correta.

5. Sejam  $A$  e  $B$  acontecimentos independentes tais que  $P(A) = 0.1$  e  $P(B) = 0.2$ . Então  $P(\overline{A \cup B})$ , é:

0.72

0.30

0.98

0.70

nenhuma das hipóteses anteriores está correta.

$$\begin{aligned}
 P(\overline{A \cup B}) &= 1 - P(A \cup B) = \\
 &= 1 - [P(A) + P(B) - P(A \cap B)] = \\
 &= 1 - [0,1 + 0,2 - P(A) \times P(B)] = \\
 &= 1 - (0,1 + 0,2 - 0,1 \times 0,2) = \\
 &= 1 - (0,3 - 0,02) = \\
 &= 1 - 0,28 = \\
 &= 0,72
 \end{aligned}$$

6. Sabe-se que na terceira semana de vida, a população de juvenis de uma certa espécie de répteis têm um aumento de peso que segue uma distribuição normal, e sabe-se também que, nesse período de tempo, a probabilidade desses juvenis aumentarem o seu peso mais de 30 g, é 0.64. Selecionaram-se de forma aleatória 50 indivíduos dessa população, e registrou-se o seu crescimento no intervalo de tempo referido.

Nestas condições, a probabilidade de que pelo menos 30 juvenis tenham aumentado de peso mais do que 30 g, é:

0.7704;

0.6745;

0.4239;

0.2296;

nenhuma das hipóteses anteriores está correta.

$$\mu = 0,64$$

$$n = 50$$

$$\frac{X - \mu}{\sqrt{\mu(1-\mu)}} \sim N(0,1)$$

$$\begin{aligned}
 P(X \geq 30) &= P(X \geq 29,5) = P\left(Z \geq \frac{29,5 - 5 \times 0,64}{\sqrt{50 \times 0,64 \times (1-0,64)}}\right) = \\
 &\approx P(Z \geq -0,737) = P(Z \leq 0,737) = 0,7704
 \end{aligned}$$

7. Se  $X$  e  $Y$  são duas variáveis aleatórias, tais que  $Y = aX + b$  com  $a, b \in \mathbb{R}$ , então,

$E(Y) = aE(X) + b$  e  $V(Y) = aV(X) + b$ ;

$E(Y) = aE(X) + b$  e  $V(Y) = a^2V(X) + b$ ;

nenhuma das hipóteses anteriores está correta.

$E(Y) = aE(X)$  e  $V(Y) = a^2V(X)$ ;

$E(Y) = aE(X) + b$  e  $V(Y) = a^2V(X)$ ;

$$E(Y) = a E(X) + b$$

$$V(Y) = \sigma^2(Y) = [a \sigma(X)]^2 = a^2 \sigma^2(X) = a^2 V(X)$$

8. O Manuel vai estar fora e pede a um amigo que lhe regue uma planta enquanto não regressa. Sabe-se que sem água, há uma probabilidade de 80% da planta morrer. Sabe-se ainda que, mesmo que seja regada, a probabilidade da planta morrer é de 15%. Contudo, o Manuel tem 90% de certeza que o amigo se vai lembrar de regar a planta.

Então, a probabilidade da planta estar viva quando o Manuel regressar é:

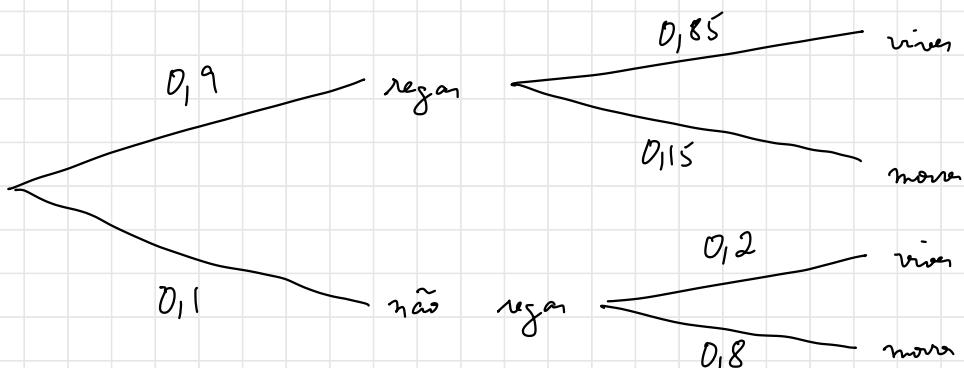
0.4750

0.7850

0.7075

0.9892

Com a informação dada não é possível fazer o cálculo.



$$P = 0.9 \times 0.85 + 0.1 \times 0.2 = 0.785$$

9. Para estudar a produção fruticula de um pomar, foi selecionada uma amostra aleatória de 13 árvores desse pomar e registado o número de frutos de cada uma. Os resultados foram os seguintes:

~~36 36 40 28 31 51 32 30 28 30 27 28 33~~  
~~27 28 28 28 30 30 31 32 33 35 36 40 51~~

Então os valores de Q1 (1º quartil) e Q3 (3º quartil), são respetivamente:

- Q1 = 28 e Q3 = 35;       Q1 = 28.5 e Q3 = 34;       Q1 = 28.5 e Q3 = 36;  
 Q1 = 28 e Q3 = 35.5;       nenhuma das hipóteses anteriores está correta.

$$Q_1: n_f = 13 \times 0,25 = 3,25 \rightarrow 3 \rightarrow 28$$

$$Q_3: n_f = 13 \times 0,75 = 9,75 \rightarrow 10 \rightarrow 35$$

10. Considerar uma variável aleatória discreta  $X$ , com a função de probabilidade que é dada na tabela seguinte:

$x$	20	21	22	23
$f(x)$	0.19	0.43	0.23	0.15

A média e a variância de  $X$ , são respetivamente:

- 21.34 e 456.30       21.34 e 0.904       21.50 e 0.904       21.50 e 456.30  
 nenhuma das hipóteses anteriores está correta.

$$E(X) = 20 \times 0,19 + 21 \times 0,43 + 22 \times 0,23 + 23 \times 0,15 = 21,34$$

$$V(X) = 20^2 \times 0,19 + 21^2 \times 0,43 + 22^2 \times 0,23 + 23^2 \times 0,15 = 0,904$$

O bico de determinada espécie de aves pode ser cinzento ou amarelo. Sabe-se que numa certa região, 60% dos indivíduos dessa espécie têm o bico cinzento.

Selecionadas aleatoriamente 6 dessas aves na referida região, a probabilidade de 5 no máximo terem o bico amarelo, é:

0.996

0.833

0.037

0.400

nenhuma das hipóteses anteriores está correta

$$P(C) = 0,6 \quad P(A) = 0,4$$

$$\begin{aligned} P &= C_0^6 \times 0,4^0 \times 0,6^6 + C_1^6 \times 0,4^1 \times 0,6^5 + C_2^6 \times 0,4^2 \times 0,6^4 + \\ &+ C_3^6 \times 0,4^3 \times 0,6^3 + C_4^6 \times 0,4^4 \times 0,6^2 + C_5^6 \times 0,4^5 \times 0,6^1 = \\ &= 0,996 = 1 - 0,4^6 \end{aligned}$$

Numa pesquisa sobre sementes de linho, estudou-se a relação do nível de ácido palmítico nas sementes (baixo, médio, alto), com a sua cor (castanha ou matizada).

As variável representativa do nível de ácido palmítico nas sementes de linho, é

discreta

categórica

qualitativa

contínua

nenhuma das hipóteses anteriores está correta

(baixo, médio, alto) → qualitativo ordinal

A altura dos alunos da FEUP inscritos em Métodos Estatísticos segue uma distribuição Normal com valor esperado igual a 1.74 m e com desvio padrão igual a 0.07 m.

Selecionado ao acaso um aluno de Métodos Estatísticos da FEUP, a probabilidade da sua altura ser inferior a 1.70 m é:

0.0108

0.0784

0.3852

0.2843

nenhuma das hipóteses anteriores está correta

$$X \sim N(1.74, 0.07^2)$$

$$\begin{aligned} P(X < 1.70) &= P\left(Z < \frac{1.70 - 1.74}{0.07}\right) = P(Z < -0.57) = \\ &= 1 - P(Z < 0.57) = \\ &= 1 - 0.7157 = \\ &= 0.2843 \end{aligned}$$

Uma unidade fabril produz parafusos de dois tamanhos; 42.3% são do tamanho A e os restantes do tamanho B. À saída da produção, antes da separação automática dos parafusos de acordo com o tamanho, foi selecionada uma amostra aleatória de 100 parafusos.

A probabilidade de ter 50 parafusos do tamanho A nesta amostra, é:

0.500

0.024

0.423

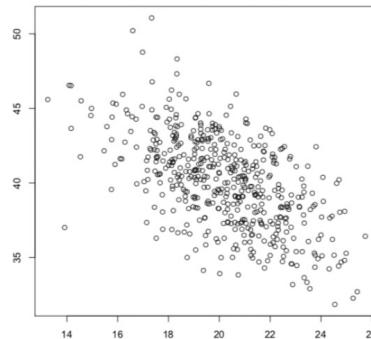
0.212

nenhuma das hipóteses anteriores está correta

$$\mu = 0.423$$

$$P(X = 50) = C_{50}^{100} \times 0.423^{50} \times (1 - 0.423)^{50} = 0.024$$

Considerando o seguinte diagrama de dispersão de uma amostra bivariada,



o coeficiente de correlação linear de Pearson para esta amostra, é:

- 0.98       0.02       -0.61       0.57
- nenhuma das hipóteses anteriores está correta

Para analisar o número de defeitos presentes num determinado tipo de máquina experimental, foi registado o número de defeitos em 120 dessas máquinas escolhidas aleatoriamente. Os dados registados estão na tabela seguinte:

número de defeitos	1	2	3	4	5	6
número de máquinas	44	41	21	9	3	2

Sendo  $\bar{x}$  a média e  $Q_2$  a mediana desta amostra, tem-se:

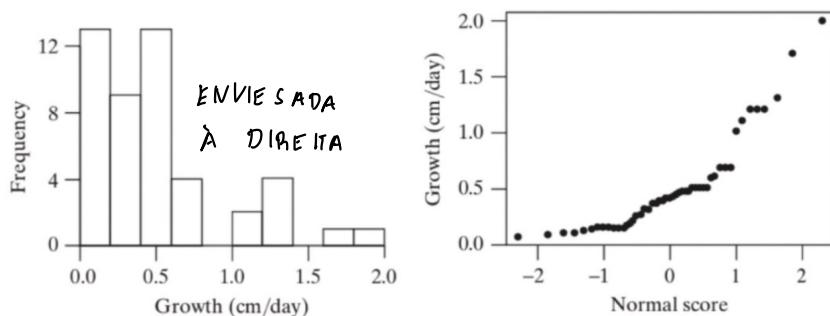
- $\bar{x} = 3.5$  e  $Q_2 = 2$         $\bar{x} = 2.1$  e  $Q_2 = 3.5$         $\bar{x} = 3.5$  e  $Q_2 = 3$         $\bar{x} = 2.1$  e  $Q_2 = 2$
- nenhuma das hipóteses anteriores está correta

$$\bar{x} = \frac{1 \times 44 + 2 \times 41 + 3 \times 21 + 4 \times 9 + 5 \times 3 + 6 \times 2}{120} = 2,1$$

$$Q_2: n_1 = 120 \times 0.5 = 60$$

$$Q_2 = \frac{n_{60} + n_{61}}{2} = \frac{2 + 2}{2} = 2$$

O histograma e o gráfico QQ de probabilidade normal representados abaixo, são relativos a taxa de crescimento (em cm/dia) de 47 carvalhos:



Trata-se de uma distribuição enviesada à direita, o que sugere que a população de carvalhos, donde provêm os dados, **não segue** uma distribuição normal

Trata-se de uma distribuição enviesada à esquerda, o que sugere que a população de carvalhos, donde provêm os dados, **segue** uma distribuição normal

Trata-se de uma distribuição enviesada à direita, o que sugere que a população de carvalhos, donde provêm os dados, **segue** uma distribuição normal

Trata-se de uma distribuição enviesada à esquerda, o que sugere que a população de carvalhos, donde provêm os dados, **não segue** uma distribuição normal

nenhuma das hipóteses anteriores está correta

Sejam  $A$  e  $B$  dois acontecimentos independentes definidos no mesmo espaço de probabilidade, e tais que  $P(B) = 2/5$  e  $P(\bar{A}) = 4/5$ . Nestas condições, o valor de  $P(A|B)$ , é

0.3

0.4

0.7

0.8

nenhuma das hipóteses anteriores está correta

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A) = 1 - P(\bar{A}) = 1 - \frac{4}{5} = \frac{1}{5}$$

Um determinado tipo de máquinas de limpeza rotativas, tem um peso que segue uma distribuição normal com média 50 Kg e desvio padrão 12 Kg. Seja  $\bar{X}$  a variável aleatória representativa da média (em quilogramas) dos pesos de uma amostra aleatória de 16 dessas máquinas.

Então a média e o desvio padrão de  $\bar{X}$ , são respectivamente,

- 3.13 Kg e 12 Kg
- 3.13 Kg e 3 Kg
- 50 Kg e 12 Kg
- 50 Kg e 3 Kg
- nenhuma das hipóteses anteriores está correta

$$\mu = 50 \quad \sigma = 12$$

$$\mu_{\bar{x}} = \mu = 50 \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{16}} = \frac{12}{4} = 3$$

Foi selecionada uma amostra aleatória de 20 classificações obtidas num teste de matemática, pelos alunos de uma determinada escola secundária de Lisboa. Os resultados encontram-se na tabela abaixo. As classificações estão na escala de 0 a 100.

35.5	40.1	47.3	48.9	52.4
39.8	39.3	55.6	40.3	60.9
30.5	59.8	44.5	36.8	36.6
42.1	26.2	33.3	65.4	45.1

$$\left[ \sum_{i=1}^{20} x_i = 880.4; \quad \sum_{i=1}^{20} x_i^2 = 40854.56 \right]$$

Então, a média e o desvio padrão dos dados da amostra, são respectivamente:

- 44.02 e 17.01
- 44.02 e 10.51
- 51.73 e 10.51
- 51.73 e 17.01
- nenhuma das hipóteses anteriores está correta

$$\mu = \frac{880.4}{20} = 44.02$$

$$\sigma^2 = \frac{1}{19} \left( 40854.56 - \frac{1}{20} 880.4^2 \right) = 110.49 \rightarrow \sigma = 10.51$$

# Inferência Estatística

Inferência Estatística: inferir sobre características e interse da população da qual foram obtidos os dados — tirar conclusões acerca de uma população a partir do estudo de uma amostra

Probabilidade: raciocínio deductivo — do geral para o particular  
≠

Inferência Estatística: raciocínio induutivo — do particular para o geral

Estimação Paramétrica: calcular, a partir da análise de uma amostra, um valor (ou intervalo de valores) que serve como aproximação do correspondente parâmetro na população

## Objetivos:

1. obter uma estimativa para um parâmetro da população
2. avaliar a qualidade da estimativa

Testes de Hipótese: responder, a partir da análise de uma amostra, a uma questão (formulada como uma hipótese) sobre a característica em estudo da população

Erro Padrão da Média: Nendo a média amostral  $\bar{X}$  uma variável aleatória baseada em  $n$  variáveis aleatórias  $X_1, \dots, X_n$  independentes e identicamente distribuídas com média  $\mu$  e desvio padrão  $\sigma$ :  $E(\bar{X}) = \mu$  e  $V(\bar{X}) = \sigma^2/n$

1.  $\sigma$  conhecido:  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

2.  $\sigma$  desconhecido:  $\sigma_{\bar{X}} = \frac{s}{\sqrt{n}}$

desvio padrão da amostra

desvio padrão da média amostral  
erro padrão da média

se

1: Descreve a variabilidade na amostra  
 → refere-se à dispersão dos dados que constituem a amostra  
 usar  $n$  se pretende descrever a variabilidade numa amostra

se: indica a variabilidade associada com a média amostral, isto é, fornece uma indicação sobre a qualidade da estimativa  
 → tendo os  $n$  dados (da amostragem) da média amostral  $\bar{X}$ , descreve a fiabilidade da média da amostra como estimativa da média da população  
 usar  $n$  o objetivo é indicar a imprecisão associada à estimativa  $\bar{x}$ , da média da população  $\mu$

**Intervalo de Confiança (IC):** Para um parâmetro  $\theta$  da população (cujo valor é desconhecido), é um intervalo construído a partir de uma amostra aleatória retirada da população e que contém  $\theta$  com uma certa "garantia"

IC para  $\mu$  de uma população normal com variância conhecida:

$$X \sim N(\mu, \sigma^2) \rightarrow \bar{X} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right) \rightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

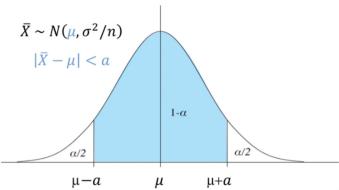
Um IC centrado na média amostral é  $[\bar{X} - a, \bar{X} + a]$ ,  $a > 0$

→  $P(\mu \in [\bar{X} - a, \bar{X} + a]) = 1 - \alpha \Leftrightarrow P(|\bar{X} - \mu| \leq a) = 1 - \alpha$  **grau de confiança**

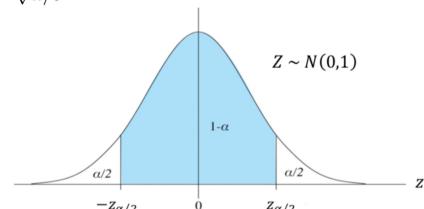
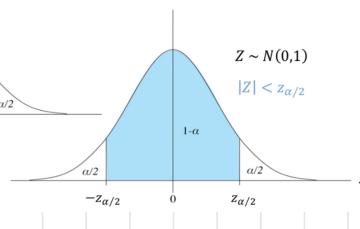
**Nível de Significância:**  $\alpha$

$$X \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Rightarrow Z \sim N(0, 1)$$

$$\begin{aligned} P(|\bar{X} - \mu| < a) &= 1 - \alpha \Leftrightarrow P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha \\ &\Leftrightarrow P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \end{aligned}$$



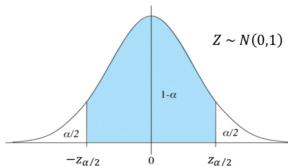
onde  $z_{\alpha/2} = \frac{a}{\sqrt{n}/\sigma}$



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

Alguns casos particulares:

$1 - \alpha$	$z_{\alpha/2}$
90%	1.65
95%	1.96
99%	2.58



$$P\left(\bar{X} - 1.65 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.65 \frac{\sigma}{\sqrt{n}}\right) = 0.90$$

$$P\left(\bar{X} - 2.58 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 2.58 \frac{\sigma}{\sqrt{n}}\right) = 0.99$$

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



$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Margem de Erro:  $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$  (aproxima sempre para cima)

Considerando todas as amostras aleatórias de tamanho  $n$  que é possível retirar da população e construindo os respectivos IC, P% desses intervalos não contém o parâmetro  $\theta$ ; não existe P% de probabilidade de um IC conter  $\theta$  😞

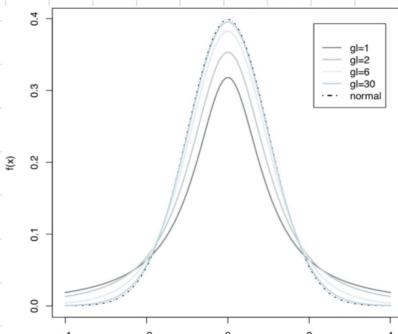
IC para  $\mu$  de uma população normal com variância desconhecida:  $X \sim N(\mu, \sigma^2)$

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

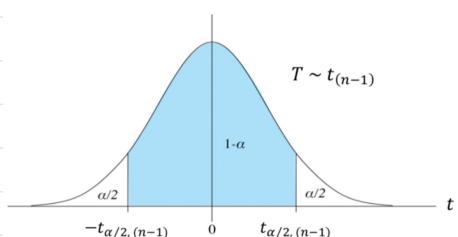
Distribuição t de Student: distribuição simétrica que depende de um só parâmetro, o número de "graus de liberdade"

T tem uma distribuição t de Student com  $n-1$  graus de liberdade:

$$T \sim t_{(n-1)}$$



$$X \sim N(\mu, \sigma^2) \Rightarrow T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$



$$T \sim t_{(n-1)} \Rightarrow P(-t_{\alpha/2, (n-1)} \leq T \leq t_{\alpha/2, (n-1)}) = 1 - \alpha$$

$$P(-t_{\alpha/2, (n-1)} \cdot S/\sqrt{n} \leq \bar{X} - \mu \leq t_{\alpha/2, (n-1)} \cdot S/\sqrt{n}) = 1 - \alpha$$

$$P(\bar{X} - t_{\alpha/2, (n-1)} \cdot S/\sqrt{n} \leq \mu \leq \bar{X} + t_{\alpha/2, (n-1)} \cdot S/\sqrt{n}) = 1 - \alpha$$

**IC:**  $\left( \bar{X} \pm t_{\alpha/2, (n-1)} \frac{S}{\sqrt{n}} \right)$

Caso  $1 - \alpha = 0.95$ , (95% de confiança)

Para o caso  $1 - \alpha = 0.95$ , (95% de confiança), temos:

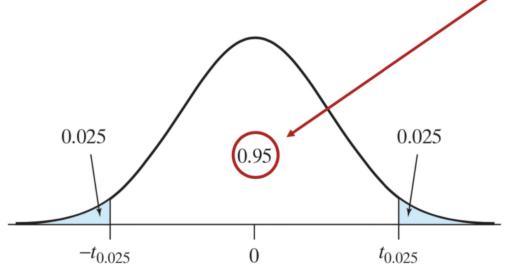
$$P(-t_{0.025, (n-1)} \cdot S/\sqrt{n} \leq \bar{X} - \mu \leq t_{0.025, (n-1)} \cdot S/\sqrt{n}) = 0.95$$

$$P(\bar{X} - t_{0.025, (n-1)} \cdot S/\sqrt{n} \leq \mu \leq \bar{X} + t_{0.025, (n-1)} \cdot S/\sqrt{n}) = 0.95$$

e o IC aleatório

$$\left( \bar{X} \pm t_{0.025, (n-1)} \cdot \frac{S}{\sqrt{n}} \right)$$

$$T \sim t_{(n-1)} \Rightarrow P(-t_{0.025, (n-1)} \leq T \leq t_{0.025, (n-1)}) = 0.95$$



Para uma amostra concreta, de dimensão  $n$ , com média  $\bar{x}$  e desvio padrão  $s$ , obtenho o intervalo

$$\left( \bar{x} - t_{\alpha/2, (n-1)} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2, (n-1)} \cdot \frac{s}{\sqrt{n}} \right)$$

intervalo de confiança, com grau de confiança  $1 - \alpha$  para a média  $\mu$  de uma população normal com desvio padrão desconhecido que é um



• Dado  $\bar{x}$  a média,  $s$  o desvio padrão e  $n$  a dimensão da amostra

1. Variância  $\sigma^2$  conhecida:  $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \rightarrow N(0, 1)$

2. Variância  $\sigma^2$  desconhecida:  $\bar{x} \pm t_{\alpha/2, n-1} s \rightarrow t_{(n-1)}$   $\left( se = \frac{s}{\sqrt{n}} \right)$

$\lambda$ : desvio padrão numa amostra — estimativa de  $\sigma$

$\lambda e$ : erro padrão (da média) — estimativa de  $\sigma_{\bar{x}}$

$\sigma$ : desvio padrão populacional

$S$ : desvio padrão amostral — estimador

} parâmetros

•  $X \sim N(\mu, \sigma^2)$

•  $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$  média amostral

•  $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

•  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$  desvio padrão da média amostral

$\bar{x}$ : média da amostra

### População Normal

- variância ( $\sigma_x^2$ ) conhecida

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \longrightarrow N(0, 1)$$

- variância ( $\sigma_x^2$ ) desconhecida

$$\bar{x} \pm t_{\alpha/2, n-1} se \longrightarrow t_{(n-1)}; \quad \left( se = \frac{s_x}{\sqrt{n}} \right)$$

### População NÃO Normal ( $n \geq 30$ )

$$\bar{x} \pm z_{\alpha/2} \frac{s_x}{\sqrt{n}} \longrightarrow N(0, 1)$$

$$P(Z > z_{\alpha/2}) = \alpha/2; \quad Z \sim N(0, 1)$$

$$P(T > t_{\alpha/2, n-1}) = \alpha/2; \quad T \sim t(n-1)$$

## Intervalos de Confiança

- Se a população não tiver distribuição normal (e variância desconhecida), então não é possível construir um intervalo de confiança para a média da população se o tamanho da amostra for grande ( $n \geq 30$ ) - TLC

$$IC: \bar{x} \pm z_{\alpha/2} s_e, \quad s_e = \frac{s}{\sqrt{n}}$$

- ICs baseados em amostras de menor dimensão tenderão a ter menor amplitude

## RESUMO

### População Normal

- variância ( $\sigma_x^2$ ) conhecida

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \rightarrow N(0, 1)$$

- variância ( $\sigma_x^2$ ) desconhecida

$$\bar{x} \pm t_{\alpha/2, n-1} s_e \rightarrow t_{(n-1)}; \quad \left( s_e = \frac{s}{\sqrt{n}} \right)$$

### População NÃO Normal ( $n \geq 30$ )

$$\bar{x} \pm z_{\alpha/2} \frac{s_x}{\sqrt{n}} \rightarrow N(0, 1)$$

$$P(Z > z_{\alpha/2}) = \alpha/2; \quad Z \sim N(0, 1)$$

$$P(T > t_{\alpha/2, n-1}) = \alpha/2; \quad T \sim t(n-1)$$

→ IC - Diferença de Médias ( $\mu_x - \mu_y$ ) - Duas Populações X e Y

$$\cdot E(\bar{X} - \bar{Y}) = \mu_x - \mu_y \quad \cdot V(\bar{X} - \bar{Y}) = V(\bar{X}) + V(\bar{Y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$$

$$S_{(\bar{X}-\bar{Y})} = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

$$se_{(\bar{X}-\bar{Y})} = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

$$\cdot se_{(\bar{X}-\bar{Y})} = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} = \boxed{\sqrt{se_x^2 + se_y^2}} \quad se_x = \frac{s_x}{\sqrt{n_x}}, \quad se_y = \frac{s_y}{\sqrt{n_y}}$$

Média Ponderada das Variâncias Amostrais:

$$S_p^2 = \frac{(n_x - 1) S_x^2 + (n_y - 1) S_y^2}{n_x + n_y - 2}$$

Erro Padrão Ponderado:  
Só SE  $s_x = s_y$ !

$$SE_p = \sqrt{S_p^2 \left( \frac{1}{n_x} + \frac{1}{n_y} \right)} = \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}$$

- Admitindo  $\sigma_x^2 \neq \sigma_y^2$

$$SE_{(\bar{X}-\bar{Y})} = \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}$$

- Admitindo  $\sigma_x^2 = \sigma_y^2 = \sigma^2$

$$SE_{(\bar{X}-\bar{Y})} \equiv SE_p = \sqrt{\frac{S_p^2}{n_x} + \frac{S_p^2}{n_y}}$$

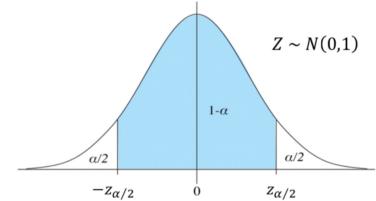
Admitindo populações normais e amostras independentes:

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$$

Nesta situação, com um grau de confiança  $1 - \alpha$ , temos:

$$P\left(|\bar{X} - \bar{Y} - (\mu_X - \mu_Y)| \leq z_{\alpha/2} \cdot \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}\right) \approx 1 - \alpha$$

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim N(0, 1)$$



variância conhecida

$$\text{IC: } \bar{X} - \bar{Y} \pm z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}$$

variância desconhecida

$$\text{IC: } \bar{X} - \bar{Y} \pm t_{\alpha/2, df} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

$$df = \frac{\left(\bar{x}_X^2 + \bar{x}_Y^2\right)^2}{\frac{\bar{x}_X^4}{n_X-1} + \frac{\bar{x}_Y^4}{n_Y-1}} \approx n_X + n_Y - 2$$

Para populações não normais mas amostras grandes:  $\bar{X} - \bar{Y} \pm z_{\alpha/2}$  se

$(\mu_X - \mu_Y)$  - Amostras Independentes - Resumo

#### Populações Normais

- variâncias conhecidas

$$\bar{X} - \bar{Y} \pm z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \leftarrow N(0, 1)$$

- variâncias desconhecidas

$$\bar{X} - \bar{Y} \pm t_{\alpha/2, n_X + n_Y - 2} \text{se} \quad \leftarrow t_{(n_X + n_Y - 2)}$$

#### Populações NÃO Normais - Amostras Grandes ( $n \geq 30$ )

$$\bar{X} - \bar{Y} \pm z_{\alpha/2} \text{se} \quad \leftarrow N(0, 1)$$

$$\text{com se} = \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

$$= \bar{X} - \bar{Y} \pm z_{\alpha/2} \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}$$

# Folha 4

Um farmacologista mediou a concentração cerebral de dopamina numa amostra de ratos. A concentração média obtida foi de 1 269 ng/g e o desvio padrão de 145 ng/g.

Qual o erro padrão obtido para a média, considerando:

- (a) Uma amostra de 8 ratos?
- (b) Uma amostra de 30 ratos?

a)  $\text{se} = \frac{\sigma}{\sqrt{n}} = \frac{145}{\sqrt{8}} = 51,265$  ✓

b)  $\text{se} = \frac{\sigma}{\sqrt{n}} = \frac{145}{\sqrt{30}} = 26,473$  ✓

Considerar o exercício anterior e suponha que a variável em estudo é normalmente distribuída. Determinar para cada um dos casos (amostra de 8 e 30 ratos) um intervalo de confiança a 95% para a média.

Comparar os dois intervalos e comentar o resultado.

$$\left( \bar{X} \pm t_{\alpha/2} (n-1) \times \frac{\sigma}{\sqrt{n}} \right)$$

a)  $\left( 1269 \pm t_{0,025} (7) \times \frac{145}{\sqrt{8}} \right) = \left( 1269 \pm 2,3646 \times \frac{145}{\sqrt{8}} \right) = [1147,8; 1390,2]$  ✓

b)  $\left( 1269 + t_{0,025} (29) \times \frac{145}{\sqrt{30}} \right) = \left( 1269 + 2,0452 \times \frac{145}{\sqrt{30}} \right) = [1214,9; 1323,1]$  ✓

Supor que estamos a planejar uma experiência para testar o efeito de uma dieta no aumento de peso de uma população de perus. Seja  $Y$  a variável que representa o aumento de peso em 3 semanas relativo a essa dieta. Experiências anteriores sugerem que o desvio padrão de  $Y$  é aproximadamente 80 g.

Determinar quantos perus deverão constituir a amostra da experiência se se pretender que o erro padrão da média seja não superior a 20 g.

$$\text{se } \leq 20, \quad \text{se} = \frac{\sigma}{\sqrt{n}} = \frac{80}{\sqrt{n}}$$

$$\frac{80}{\sqrt{n}} \leq 20 \Leftrightarrow \sqrt{n} \geq 4 \Leftrightarrow n \geq 16 \quad \therefore n=16 \quad \checkmark$$

Sabe-se, por experiência passada, que o desvio padrão relativo ao tempo necessário para a recuperação total das pessoas que sofrem de uma determinada doença, é de 15 dias. Depois de analisar uma amostra de 16 doentes obteve-se um tempo médio de 85 dias para recuperar um doente.

Estimar, com um nível de confiança de 95%, o tempo médio (em dias), necessário para um doente que sofra da doença recuperar totalmente.

$$\sigma = 15 \quad \bar{x} = 85 \quad n = 16 \quad \alpha = 1 - 0,95 = 0,05$$

$$\left( \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

$$\text{IC: } \left( 85 \pm z_{0,05} \times \frac{15}{\sqrt{16}} \right) = \left( 85 \pm 1,96 \times \frac{15}{4} \right) = [77,65, 92,35]$$

• A amostra não é grande ( $n < 30$ ) e como tal tem de se assumir que a amostra provém de uma população com distribuição normal.

Numa experiência integrada num estudo sobre o desenvolvimento da glândula timo, os investigadores pesaram as glândulas de cinco embriões de frangos após 14 dias de incubação. Os pesos da *glândula timo* (em mg) foram os seguintes:

29.6 21.5 28.0 34.6 44.9

Para estes dados, a média é 31.7 mg, e o desvio padrão 8.7 mg.

- Calcular o erro padrão da média.
- Construir um intervalo de confiança a 90% para a média da população.

a)  $\lambda_e = \frac{\Delta}{\sqrt{n}} = \frac{8,7}{\sqrt{5}} = 3,89$  ✓

b) IC:  $(\bar{x} \pm t_{\alpha/2}(n-1) \times \lambda_e) = \left(31,7 \pm t_{0,05}(4) \times \frac{8,7}{\sqrt{5}}\right) = (23,4; 40,0)$  ✓

6. Considerar os dados seguintes (exercício 4 da Folha 1), referentes a uma amostra das medições do perímetro cefálico (em cm) de 35 recém nascidos do sexo masculino.

33.1	33.4	34.8	33.8	34.7	34.3	35.6
34.5	34.6	34.1	33.9	33.6	34.6	35.2
33.7	35.8	34.2	34.0	34.7	35.2	34.3
33.4	36.0	34.5	36.1	35.1	35.1	34.6
33.7	34.9	34.2	34.2	34.2	35.3	34.2

Dados auxiliares:  $\sum_{i=1}^{35} x_i = 1207,6$   $\sum_{i=1}^{35} x_i^2 = 41684,32$

Determinar um intervalo de confiança a 95% para o perímetro cefálico médio da população.

• A amostra é grande ( $n \geq 30$ ) logo:  $IC = (\bar{x} \pm z_{\alpha/2} \lambda_e)$

$$\bar{x}_e = \frac{\sum x_i}{n} = \frac{1207,6}{35} \quad \lambda_e = \frac{\Delta}{\sqrt{n}}$$

$$\Delta = \sqrt{\frac{1}{n-1} \left( \sum x_i^2 - \frac{1}{n} \left( \sum x_i \right)^2 \right)} = \sqrt{\frac{1}{34} \left( 41684,32 - \frac{1}{35} \times 1207,6^2 \right)}$$

$$z_{\alpha/2} = 1,96$$

$$IC = (34,25736, 34,74836)$$

O nível de ácido úrico  $X$  (mg/dl) numa determinada população de homens adultos e saudáveis tem uma distribuição normal de valor esperado  $\mu$  mg/dl e desvio padrão  $\sigma = 1$  mg/dl. Recolheu-se uma amostra de dimensão 100, cujos níveis de ácido úrico apresentaram valor médio igual a 5.5 mg/dl.

- (a) Determinar um intervalo de confiança a 99% para  $\mu$ .
- (b) Admitir agora que se tem  $X \sim N(5.5, 1)$ . Calcular a dimensão da amostra de modo a que seja pelo menos igual a 0.9, a probabilidade da média da amostra se situar entre 5 e 6 mg/dl.

a) IC para população normal de variância conhecida

$$IC: \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 5,5 \pm z_{0,005} \times \frac{1}{\sqrt{100}} = 5,5 \pm z_{0,005} \times \frac{1}{10} = 5,5 \pm \frac{2,5758}{10} = \\ = (5,24; 5,76) \quad \checkmark$$

b)  $z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 0,5 \quad z_{0,05} \times \frac{1}{\sqrt{n}} = 0,5 \Leftrightarrow \frac{1,6449}{\sqrt{n}} = 0,5$   
 $\Leftrightarrow n = 11 \quad \checkmark$

A distribuição dos diâmetros dos caules nas plantas de determinada espécie tem distribuição normal. Uma amostra de 5 plantas apresentou os seguintes diâmetros (em mm):

25.4 25.2 25.3 25.0 25.4

Construir um intervalo de confiança a 98% para o diâmetro médio dos caules das plantas da espécie em causa.

$$\bar{x} = 25,26 \quad s_n = 0,1673$$

$$IC: 25,26 \pm t_{0,01,4} \times \frac{0,1673}{\sqrt{5}} = 25,26 \pm 3,7469 \times 0,0748 = \\ = 25,26 \pm 0,28 = \\ = (24,98; 25,54) \quad \checkmark$$

Uma amostra aleatória retirada de uma população normal, produziu os intervalos de confiança seguintes, com graus de confiança distintos para a média dessa população:

$$]28.84, 31.10[ \quad ]28.48, 31.46[$$

- (a) Qual o valor da média amostral?  
(b) Qual dos dois intervalos corresponde a um menor grau de confiança?

a)  $\frac{28,84 + 31,10}{2} = \frac{28,48 + 31,46}{2} = 29,97 \quad \checkmark$

b) O intervalo de menor amplitude corresponde a um menor grau de confiança:  $]28,84; 31,10[ \quad \checkmark$

Numa amostra de doentes com uma determinada doença, verificou-se que o tempo médio de vida, após o diagnóstico, foi de 7 anos.

Admitindo que o tempo de vida médio destes doentes segue uma distribuição normal de média  $\mu$  e desvio padrão  $\sigma = 1$  ano, indicar o tamanho da amostra para ter pelo menos 95% de confiança de que o erro de estimação seja inferior a 0.05.

$$z_{\alpha/2} \times \frac{\Delta}{\sqrt{n}} \leq 0,05 \Rightarrow 1,96 \times \frac{1}{\sqrt{n}} \leq 0,05 \Leftrightarrow n \geq 1537 \quad \checkmark$$

Dois grupos de frangos, escolhidos aleatoriamente e de modo independente, foram submetidos a duas dietas diferentes. Após 2 semanas, observou-se o aumento de peso dos frangos, obtendo-se os seguintes valores (em gramas):

$$S_p^2 = \frac{(n-1) S_1^2 + (n-1) S_2^2}{n_1 + n_2 - 2}$$

ERRO PADRÃO PONDERADO

$$\text{ERRO PADRÃO } \Delta e = \sqrt{\frac{\Delta_1^2}{n_1} + \frac{\Delta_2^2}{n_2}} = \sqrt{\Delta e_1^2 + \Delta e_2^2}$$

	Dieta 1	Dieta 2
$\bar{x}$	165	180
s	42	56

$$SE_p = \sqrt{\frac{\Delta_1^2}{n_1} + \frac{\Delta_2^2}{n_2}}$$

Determinar o erro padrão e o erro padrão ponderado de  $\bar{X}_1 - \bar{X}_2$ , considerando as dimensões das amostras tal como indicado, e comentar os resultados obtidos:

(a)  $n_1 = 10$  e  $n_2 = 15$

(b)  $n_1 = 25$  e  $n_2 = 25$ .

a)  $SE = \sqrt{\frac{42^2}{10} + \frac{56^2}{15}} = 9,6$  ✓

$$SE_p = \sqrt{\left( \frac{9 \times 42^2 + 14 \times 56^2}{10 + 15 - 2} \right)} \left( \frac{1}{10} + \frac{1}{15} \right) = 20,8$$

b)  $SE = \sqrt{\frac{42^2}{25} + \frac{56^2}{25}} = 14$  ✓

$$SE_p = \sqrt{\left( \frac{24 \times 42^2 + 24 \times 56^2}{25 + 25 - 2} \right)} \left( \frac{1}{25} + \frac{1}{25} \right) = 13,6$$

Considerar os dados do exercício anterior e supor que o aumento de peso para cada dieta pode ser considerado como tendo uma distribuição normal.

Construir um intervalo de confiança a 95% para a diferença entre os aumentos de peso médios das populações correspondentes. IC para populações normais com variâncias desconhecidas

(a)  $n_1 = 10$  e  $n_2 = 15$   $\bar{Y}_1 - \bar{Y}_2 \pm t_{\alpha/2, n_1 + n_2 - 2} \Delta e$

(b)  $n_1 = 25$  e  $n_2 = 25$ .

a)  $165 - 180 \pm 2,0687 \sqrt{\frac{42^2}{10} + \frac{56^2}{15}} = (-55,62, 25,62)$  ✓

b)  $165 - 180 \pm 2,0106 \times \sqrt{\frac{42^2}{25} + \frac{56^2}{25}} = (-43,15, 13,15)$  ✓

Foram medidos os níveis de destruição dos pulmões em 9 indivíduos não fumadores e em 12 indivíduos fumadores, tendo-se obtido os resultados seguintes:

fumadores	18.1	6.0	10.8	11.0	7.7	17.9	8.5	13.0	18.9			
não fumadores	16.6	13.9	11.3	26.5	17.4	15.3	15.8	12.3	18.6	12.0	24.1	16.5

Construir um intervalo de confiança a 95% para a diferença das médias dos níveis de destruição dos pulmões nos dois grupos. O que se pode concluir?

$$\bar{x} = 12,433$$

$$\bar{y} = 16,692$$

$$\bar{x} - \bar{y} = -4,259$$

$$IC: \bar{x} - \bar{y} \pm t_{\alpha/2, n_1 + n_2 - 2} \cdot s$$

$$\alpha = 0,05 \rightarrow \alpha/2 = 0,025$$

$$gl = n_1 + n_2 - 2 = 9 + 12 - 2 = 19$$

$$t_{0.025, 19} = 2,0930$$

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$s_1^2 = 23,51$$

$$s_2^2 = 21,47$$

$$se = \sqrt{\frac{23,51}{9} + \frac{21,47}{12}} = 2,10$$

$$IC: -4,259 \pm 2,0930 \cdot 2,10 = (-8,65; 0,136)$$

✓

Pretende-se testar se a altura média  $\mu_1$ , dos pais numa dada população difere significativamente da altura média  $\mu_2$  dos respetivos filhos. Para tal, foi selecionada uma amostra de 12 pais e respetivos filhos adultos, tendo-se registado as seguintes alturas (em cm):

altura do pai ( $x_i$ )	190	184	183	182	181	178	175	174	170	168	165	164
altura do filho ( $y_i$ )	189	186	180	179	187	182	183	171	170	178	174	165
DIFERENÇA	1	-2	3	3	-6	-4	-8	3	0	-10	-9	-1

Dados auxiliares:

$$\sum_{i=1}^{12} x_i = 2114 \quad \sum_{i=1}^{12} y_i = 2144 \quad \sum_{i=1}^{12} x_i^2 = 373160 \quad \sum_{i=1}^{12} y_i^2 = 383666 \quad \sum_{i=1}^{12} (x_i - y_i)^2 = 330$$

- (a) Construir um intervalo de confiança a 95% para  $\mu_1 - \mu_2$ . O que se pode concluir?  
 (b) Como se poderia obter a partir desta amostra um intervalo de confiança com o dobro da precisão?  
 As conclusões seriam as mesmas?

a)  $\bar{x} - \bar{y} \pm t_{\alpha/2, df} \text{ e } t_{\alpha/2, df} = t_{0.025, 11}$

$$IC: \frac{\bar{x}}{12} - \frac{\bar{y}}{12} \pm 2.2010 \times \sqrt{\frac{s_{df}^2}{12}}$$

$$s_{df}^2 = \frac{1}{11} \left( \sum_{i=1}^{12} (x_i - y_i)^2 - \frac{1}{12} \left( \sum_{i=1}^{12} (x_i - y_i) \right)^2 \right) = \\ = \frac{1}{11} \left( 330 - \frac{1}{12} \times 30^2 \right) = 23,18$$

$$IC: \frac{-30}{12} \pm 2.2010 \times \sqrt{\frac{23,18}{12}} = -2,5 \pm 3,06 = (-5,56; 0,56)$$

b) aumentar o tamanho da amostra para o quádruplo ...

# Testes de Hipóteses

**Hipótese Estatística:** afirmação acerca de aspectos desconhecidos de uma variável aleatória  $X$

**Testes de Hipóteses:** procedimentos estatísticos que, com base em amostras, permitem tomar uma decisão acerca de uma hipótese estatística: rejeitar ou não rejeitar — avaliar a validade

Decidir uma afirmação relativa à distribuição de probabilidade de uma variável aleatória!

**Testes de Hipóteses** {  
· paramétricos — acerca de um parâmetro da população  
· paramétricos  
· de independência  
· de homogeneidade / igualdade

**Testes Paramétricos:** formuladas duas hipóteses,  $H_0$  e  $H_1$ , acerca de um parâmetro da distribuição da variável  $X$ , testá-las, definindo um critério que, em face dos dados, permite:

1. Não rejeitar  $H_0$  (hipótese nula)

ou

2. Rejeitar  $H_0$ , aceitando implicitamente  $H_1$  (hipótese alternativa)

↓  
· Desenhado para rejeitar  $H_0$

· Avalia quão fortes são as evidências a favor de  $H_1$

**Erro Tipo I:** rejeição da hipótese nula quando é verdadeira

**Erro Tipo II:** não rejeição da hipótese nula, quando é falsa

Decisão	$H_0$ é verdadeira	$H_0$ é falsa
Não rejeição de $H_0$	Decisão Correta	Erro de Tipo II
Rejeição de $H_0$	Erro de Tipo I	Decisão Correta

$\alpha$ : nível de significância — probabilidade de erro tipo I

- $P(\text{Rejeitar } H_0 \mid H_0 \text{ verdadeira}) = \alpha$

- $\alpha$  redy  $\alpha$ :

- aumentar região de não rejeição
- aumentar tamanho da amostra

Testes t: testar hipóteses acerca de médias de variáveis quantitativas, em amostras grandes ou normais, baseados na distribuição t de Student

- Tipos:
- Comparar uma média desconhecida com um valor específico, a partir de uma amostra
  - Comparar 2 médias desconhecidas a partir de 2 amostras independentes
  - Comparar 2 médias desconhecidas a partir de 2 amostras emparelhadas

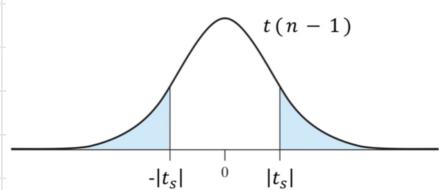
1)  $H_0: \mu = VE$        $H_1: \mu \neq VE$

$t_s$ : estatística do teste,      
$$t_s = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - VE}{\sigma/\sqrt{n}}$$

- Se  $POP \sim N$  e  $H_0: \mu = V$ ,  $T \sim t_{n-1}$

- Se  $A_{\text{ajd}} < \alpha$ , rejeita-se  $H_0$

$$P(\text{Rejeitar } H_0 \mid H_0 \text{ verdadeira}) = \alpha$$



VALOR-P

Valor - P:

- probabilidade de se obter uma estatística de teste igual ou mais extrema do que a observada, sob a condição de  $H_0$  ser verdadeira
- medida da evidência dos dados em favor de  $H_1$
- medida de compatibilidade entre os dados e  $H_0$
- menor nível de significância  $\alpha$  com que se rejeitaria  $H_0$

## Testes t – Procedimento Geral

- ① Identificar o parâmetro de interesse
- ② Definir a hipótese nula
- ③ Especificar uma hipótese alternativa apropriada
- ④ Escolher  $\alpha$  – nível de significância
- ⑤ Determinar a dimensão da amostra
- ⑥ Determinar uma estatística de teste apropriada
- ⑦ Determinar a região de rejeição/não rejeição
- ⑧ Tomar a decisão se  $H_0$  deve ser rejeitada ou não

valor -f- <  $\alpha$ : evidência de  $H_1$  a um nível de significância de  $\alpha$

Testes de Aleatorização: avaliam a variabilidade na diferença de duas médias amostrais

"valor - f": ~~X~~ possibilidades que satisfazem a hipótese / ~~X~~ Toda

## Comparação de 2 Médias

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Amostras Independentes:

$$t_s = \frac{\bar{x}_1 - \bar{x}_2}{se}, \quad se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

•  $t_s$  é o valor de uma variável aleatória  $T$ , que no caso de populações normais e sendo a hipótese  $H_0$  verdadeira, segue aproximadamente uma distribuição  $t$  de Student com  $n_1 + n_2 - 2$  graus de liberdade

Amostras Emparelhadas: considera-se a amostra constituída pelas diferenças e aplica-se o teste  $t$  para uma só amostra —  $H_0: \mu = 0$ ,  $H_1: \mu \neq 0$

Significância Estatística: permite verificar a discrepância de uma hipótese estatística em relações aos dados observados, utilizando uma medida de evidência — valor - f

• Há significância estatística quando valor - f < α

TH & IC: dado um  $IC = [a, b]$  com confiança  $1 - \alpha$ , todos os valores no intervalo são plausíveis para o parâmetro a estimar e os valores fora do intervalo não são considerados implausíveis

• Se o valor do parâmetro especificado por  $H_0$  pertence ao  $IC$   $[a, b]$ , então  $H_0$  não pode ser rejeitada a um nível  $\alpha$ , não rejeita - se  $H_0$  ao nível de  $\alpha$

## Associação e Causalidade:

Y: variável de resposta — representa a característica de interesse

X: variável explicativa — usada para explicar/preser a resposta

Num estudo experimental, é possível avaliar se existem evidências que diferenças em X causam diferenças em Y

Num estudo observacional, não é possível estabelecer relações de causalidade; apenas é possível avaliar se existe evidência de que diferenças em X estejam associadas a diferenças em Y

Teste Bilaterais: a hipótese alternativa é da forma  $\mu_1 \neq \mu_2$  (VE)

Teste Unilaterais: desigualdades ( $</>$ ) no lugar de igualdades (=)

## Condições de Aplicabilidade dos Testes t:

1. Os dados devem ser obtidos aleatoriamente das respetivas populações
2. As observações em cada amostra devem ser independentes
3. As distribuições das médias amostrais devem ser aproximadamente normais

# Folha 5

Recolheu-se uma amostra aleatória de 15 suspensões celulares, e registaram-se os valores do consumo de oxigénio (em ml), observado para cada uma das células durante o período de incubação.

Na sequência de indicações de estudos anteriores pretende-se saber se a média do consumo de oxigénio da população de células é ou não 12 ml.

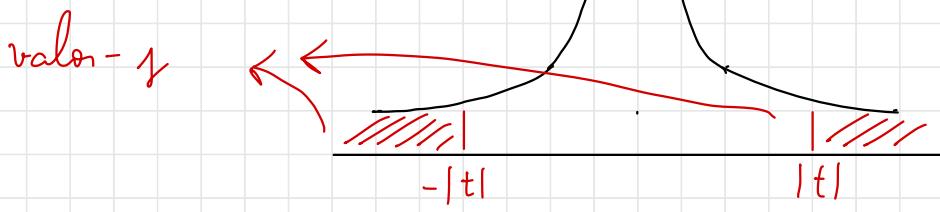
- (a) Formular as hipóteses mais adequadas para efetuar um teste de hipóteses para o problema.
- (b) Tendo em conta que valor-p = 0.0187, dizer qual a decisão que deve ser tomada a um nível de significância 0.01, 0.03 e 0.05.

$X$ : consumo de oxigénio em ml  $n = 15$   $\mu_X ?$

a)  $H_0: \mu = 12$  ✓ é

$H_1: \mu \neq 12$  ✓ não é

b)  $T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$   $T = \frac{\bar{X} - 12}{S/\sqrt{15}} \sim t_{14 = n-1}$



• Se  $\alpha = 0,01 \Rightarrow \text{valor - 1} > \alpha$ , logo, não se rejeita  $H_0$



• Se  $\alpha = 0,03 \Rightarrow \text{valor - 1} < \alpha$ , logo, rejeita-se  $H_0$



• Se  $\alpha = 0,05 \Rightarrow \text{valor - 1} < \alpha$ , logo, rejeita-se  $H_0$



Para estudar o efeito de um determinado medicamento no nível de colesterol no sangue, foi efetuado um teste t para testar a hipótese nula  $H_0 : \mu_d = 0$ , 'contra' a hipótese alternativa  $H_1 : \mu_d \neq 0$ , onde  $d$  representa a diferença no nível de colesterol (antes do tratamento – depois do tratamento).

Para isso foi utilizada uma amostra de 9 pessoas que se submeteram ao tratamento, e os valores numéricos de interesse encontram-se na tabela seguinte:

Nível de Colesterol			
	Antes	Depois	Diferença
Média	231	207	24
Desvio Padrão	0.85	1.09	34.04

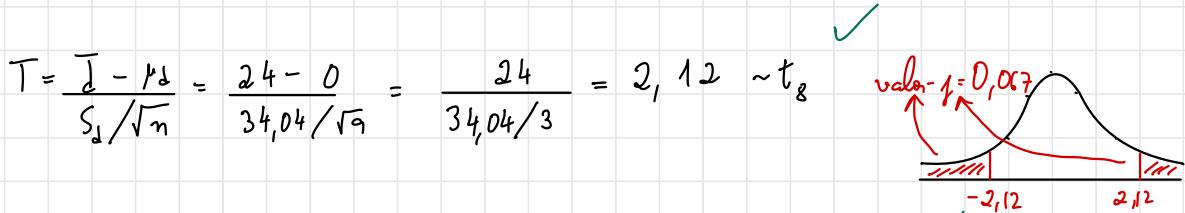
Admitindo populações normais, dizer se os dados permitem dizer se existe, ou não, evidência suficiente para concluir que, depois do tratamento, o nível de colesterol no sangue é diferente.

Considerar os níveis de significância 1%, 2%, 5% e 10%

$$t_s = \frac{24}{\sqrt{\frac{34,04^2}{9}}} = 2,12 \quad gl = n-1 = 8 \quad 2,12 \sim t_8 \rightarrow V_f = 2 \times 0,035 = 0,07 \quad \checkmark$$

$$IC_{95\%} (\mu_d) = 231 - 207 \pm t_{0,025, 8} \times 1.06 = 24 \pm 2,306 \times \frac{34,04}{\sqrt{9}} = (-2,17; 50,17) \quad \checkmark$$

∴ Como  $0 \in IC_{95\%} (\mu_d)$ , logo não se rejeita  $H_0$  para  $\alpha = 0,05$



•  $\alpha = 0,01 \Rightarrow \beta > \alpha \Rightarrow$  não se rejeita  $H_0$  ✓

•  $\alpha = 0,02 \Rightarrow \beta > \alpha \Rightarrow$  não se rejeita  $H_0$  ✓

•  $\alpha = 0,05 \Rightarrow \beta > \alpha \Rightarrow$  não se rejeita  $H_0$  ✓

•  $\alpha = 0,10 \Rightarrow \beta < \alpha \rightarrow$  rejeita-se  $H_0$  ✓

Num estudo acerca da possível influência do crómio em indivíduos diabéticos, alguns ratos foram alimentados com uma dieta de baixo teor de crómio e outros foram alimentados com uma dieta normal.

Uma variável de resposta foi a atividade da enzima hepática GITH, a qual foi medida utilizando uma molécula marcada radioativamente. Na tabela seguinte encontram-se os resultados obtidos (expressos em milhares de contagens por minuto por grama de fígado).

Existem 10 formas distintas de dividir uma amostra de 5 observações em duas partes, uma com 3 observações e outra com as restantes 2. Isto é, há 10 possíveis aleatorizações das cinco observações em dois grupos, de tamanhos 3 e 2.

baixo teor de crómio	normal
42.3	53.1
51.5	50.7
53.7	

- (a) Construir a lista destas 10 divisões (aleatorizações) e para cada uma calcule a diferença das médias relativas ao grupo de 3 observações (dieta de baixo teor de crómio) e ao grupo de 2 observações (dieta normal).
- (b) Em quantos dos 10 casos essa diferença é superior ou igual à observada?
- (c) As observações evidenciam que o crómio afeta a atividade da enzima hepática GITH? Justifique.

	$\{42.3, 50.7, 51.5, 53.1, 53.7\}$	
1.	$\{42.3, 50.7, 51.5\}$	$\{53.1, 53.7\}$
2.	$\{42.3, 50.7, 53.1\}$	$\{51.5, 53.7\}$
3.	$\{42.3, 50.7, 53.7\}$	$\{51.5, 53.1\}$
4.	$\{42.3, 51.5, 53.1\}$	$\{50.7, 53.7\}$
5.	$\{42.3, 51.5, 53.7\}$	$\{50.7, 53.1\}$
6.	$\{42.3, 53.1, 53.7\}$	$\{50.7, 51.5\}$
7.	$\{50.7, 51.5, 53.1\}$	$\{42.3, 53.7\}$
8.	$\{50.7, 51.5, 53.7\}$	$\{42.3, 53.1\}$
9.	$\{50.7, 53.1, 53.7\}$	$\{42.3, 51.5\}$
10.	$\{51.5, 53.1, 53.7\}$	$\{42.3, 50.7\}$

9 em 10 aleatorizações conduzem a valores absolutos das diferenças de médias maiores ou iguais ao observado

$$\text{valor - } \bar{x} = \frac{9}{10}$$

As observações não evidenciam diferenças significativas nas duas dietas. De facto, se os dados fossem obtidos a partir de uma única população, as diferenças das médias em duas amostras (de tamanhos 3 e 2) seriam superiores ou iguais à diferença observada em 90% dos casos. Sendo assim, a diferença observada pode ser justificada apenas pela variabilidade decorrente do acaso.

Num estudo acerca da cigarra *Magicicada Septendecim* foram medidos os comprimentos (em  $\mu\text{m}$ ) da tibia traseira em 110 indivíduos.

Os resultados relativos a machos e fêmeas encontram-se na tabela abaixo:

grupo	n	média	desvio padrão
machos	60	78.42	2.87
fêmeas	50	80.44	3.52

Pretendendo-se investigar se existe, nesta espécie, uma associação entre o comprimento da tibia traseira e o género,

- (a) Estabelecer as hipóteses nula e alternativa adequadas para conduzir um teste  $t$ ;
- (b) Poder-se-á concluir, com base nesse teste, a um nível de significância de 1% que existe uma associação entre essas duas variáveis?

a)  $H_0: \bar{m} - \bar{f} = 0$  ✓

$H_1: \bar{m} - \bar{f} \neq 0$  ✓

b)  $T = \frac{\bar{M} - \bar{F}}{SE} \sim t_{n_1+n_2-2}$  gl = 110 - 2 = 108

$$t = \frac{78,42 - 80,44}{\sqrt{\frac{2,87^2}{60} + \frac{3,52^2}{50}}} = -3,29$$

valor- $|t| < 0,002$  ✓  
2x0,001

$\therefore \alpha = 0,01$ , rejeita-se  $H_0$  ✓

Para cada uma das seguintes situações supor que se está a testar  $H_0 : \mu_1 = \mu_2$  contra  $H_1 : \mu_1 \neq \mu_2$ .

Indicar se há ou não evidência significativa de  $H_1$  nas situações seguintes:

- (a) valor  $P = 0.085$ , nível de significância=0.10.
- (b) valor  $P = 0.065$ , nível de significância=0.05.
- (c)  $t_s = 3.75$  com 19 graus de liberdade,  $\alpha = 0.01$ .

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

a)  $P = 0,085, \quad \alpha = 0,10, \quad P < \alpha \rightarrow$  afirma- $\neq$   $H_1$

b)  $P = 0,065, \quad \alpha = 0,05, \quad P > \alpha \rightarrow$  não  $\neq$  afirma  $H_1$

c)  $t_s = 3,75, \quad t_{19}, \quad \alpha = 0,01, \quad P < 0,02, \quad P < \alpha \rightarrow$  afirma- $\neq$   $H_1$

- Quando  $P > \alpha$ , aceita- $\neq$   $H_0$
- Quando  $P < \alpha$ , aceita- $\neq$   $H_1$

Pretende-se avaliar se os dados recolhidos por investigadores evidenciam uma diferença nas médias dos níveis de ácido úrico entre indivíduos normais e indivíduos com uma certa doença. O conjunto de dados consiste em medidas do ácido úrico (mg/100 ml) em 12 indivíduos doentes, e em 15 de um grupo de controlo. As médias correspondentes são de 4.5 e 3.4, e os desvios padrão de 1 e 1.5, respetivamente.

Supondo que as populações respetivas são normais, usar um teste  $t$  para fazer essa avaliação a um nível de significância de 5%.

$$n_1 = 12$$

$$\bar{x}_1 = 4,5$$

$$\sigma_1 = 1$$

$$n_2 = 15$$

$$\bar{x}_2 = 3,4$$

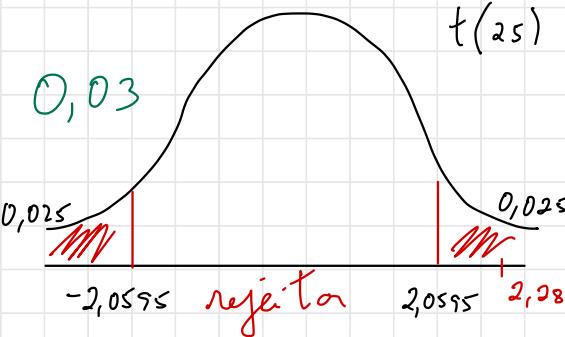
$$\sigma_2 = 1,5$$

$$H_0: \mu_1 - \mu_2 = 0 \quad \checkmark$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx t(n_1 + n_2 - 2)$$

$$t = \frac{4,5 - 3,4}{\sqrt{\frac{1^2}{12} + \frac{1,5^2}{15}}} = 2,28 \quad \checkmark$$



$\therefore |t| > 2,0595$ , logo, rejeita-se  $H_0$  para um nível de significância de 0,05 e afirma-se que as médias não são diferentes.

Num estudo acerca do metabolismo das raízes de uma certa espécie de planta, foram cultivadas diversas dessas plantas em estufa em condições distintas:  $C_1$  e  $C_2$ . Ao fim de 4 dias o conteúdo de ATP nas raízes foi registado (em nmol/mg) tendo-se obtido:

	$C_1$	$C_2$
	1.45	1.70
	1.19	2.04
	1.05	1.49
	1.07	1.91
$n$	4	4
$\bar{y}$	1.190	1.785
$s$	0.184	0.241

Usar um teste  $t$  para investigar os efeitos das condições de cultivo no metabolismo das raízes da planta em causa (nível de significância de 5%).

$$n_1 = 4$$

$$\bar{y}_1 = 1,190$$

$$s_1 = 0,184$$

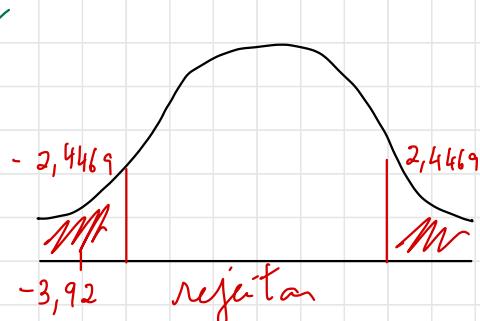
$$n_2 = 4$$

$$\bar{y}_2 = 1,785$$

$$s_2 = 0,241$$

$$t = \frac{1,19 - 1,785}{\sqrt{\frac{0,184^2}{4} + \frac{0,241^2}{4}}} = -3,92 \sim t_6 \quad \checkmark$$

$$0,01$$



∴ Como  $|t| > 2,4469$ , logo, rejeita-se  $H_0$  para um nível de significância  $\alpha = 0,05$

Supor que  $(-7.4, -2.3)$  é um intervalo de confiança a 95% para  $\mu_1 - \mu_2$ . Se testarmos  $H_0 : \mu_1 = \mu_2$  contra  $H_1 : \mu_1 \neq \mu_2$  o que se conclui para níveis de significância de 5% e 10%?

$$IC_{95\%} (\mu_1 - \mu_2) = (-7.4; -2.3)$$

•  $0 \notin IC$ , logo, rejeita-se  $H_0$  para  $\alpha = 0,05$  ✓

•  $IC_{90\%} < IC_{95\%}$ , logo,  $0 \notin IC_{90\%}$  pelo que se rejeita  $H_0$   
para  $\alpha = 0,1$  ✓

O ficheiro *Hema.csv* disponibilizado no Moodle, contém os valores do hematórito (percentagem de volume ocupada pelos glóbulos vermelhos no volume total de sangue) de 94 adultos, 51 homens e 43 mulheres.

Usando o software *R*, pretende-se fazer um teste-t para investigar se existe evidência estatística (a um nível de 5%) da associação entre o hematórito e o género, em pessoas adultas.

$$\bar{x}_h = 44,4$$

$$s_h^2 = 10,5$$

$$t = 5,55$$

$$\bar{x}_m = 40,5$$

$$s_m^2 = 12,4$$

rejeita  $H_0$

Para investigar se o exercício físico regular poderá reduzir os níveis de triglicerídeos, foi medida a concentração de triglicerídeos (em mmol/L) no soro sanguíneo de 12 voluntários do sexo masculino, antes e depois da participação num programa de exercício físico de 10 semanas.

Os resultados obtidos encontram-se no ficheiro *Tri.csv* disponibilizado no Moodle.

Usando o software *R*, fazer um teste t para decidir, com base nestas observações, se existe evidência estatística a um nível de 1%, de que o exercício físico regular reduz o nível de triglicerídeos (assume-se normalidade).

teste de hipóteses para a diferença de médias em amostras emparelhadas

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{11}$$

$$\text{valor-}p : 0,006$$

$$IC_{95\%} = (0,04; \infty)$$

∴ Para  $\alpha = 0,01$ , como o valor- $p < \alpha$ , logo, rejeita  $H_0$

# Proporções

$\hat{p}$ : proporção populacional - proporção de indivíduos com uma característica

$X$ : variável aleatória que representa o número de indivíduos, numa amostra aleatória  $X_1, \dots, X_n$  de tamanho  $n$ , retirada de uma população dicotómica, que têm numa dada característica

Estimador:  $\hat{P} = \frac{X}{n}$

Proposição Amostral:  $\hat{P} = \frac{X}{n}$

Valor Esperado:  $E(\hat{P}) = \frac{1}{n} E(X) = \frac{1}{n} \mu_X = \frac{1}{n} n\pi = \pi$

Variância:  $\sigma_{\hat{P}}^2 = \frac{1}{n^2} \sigma_X^2 = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}$

Ero Padrão:  $\text{Ero}_{\hat{P}} = \sqrt{\frac{\pi(1-\pi)}{n}}$

Para  $n$  grandes,  $n\pi > 5$  e  $n(1-\pi) > 5$ , a distribuição de  $\hat{P}$  pode ser aproximada à distribuição normal

$$\hat{P} - \pi \sim N(0,1)$$

MARGEM DE ERRO

Método de Wald:  $\hat{P} \pm \frac{\sqrt{\frac{\pi(1-\pi)}{n}}}{\text{Ero}_{\hat{P}}}$  é um IC a  $1-\alpha$  para  $\pi$

Condições:

- amostras grandes
- $\pi$  não muito próximo de 0 nem de 1

Método de Agresti-Coull:  $\tilde{P}$  - estimador de Wilson ajustado

$$\cdot \text{se}_{\tilde{P}} = \sqrt{\frac{\tilde{P}(1-\tilde{P})}{n+4}}$$

$$\cdot \tilde{P} = \frac{x+2}{n+4}$$

PARA  
 $\alpha = 95\%$

Notas:

- bom para amostras pequenas
- mais fiáveis ICs em  $\tilde{P}$  do que em  $\hat{P}$
- $IC = \tilde{P} \pm z_{\alpha/2} \times \text{se}_{\tilde{P}}$

## RESUMO

Método de Wald: estimador  $\hat{P}$

$$\cdot \hat{P} = \frac{X}{n}$$

$$\cdot \text{se}_{\hat{P}} = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

Método de Agresti-Coulli: estimador  $\tilde{P}$

$$\cdot \tilde{P} = \frac{x+2}{n+4}$$

$$\cdot \text{se}_{\tilde{P}} = \sqrt{\frac{\tilde{P}(1-\tilde{P})}{n+4}}$$

IC: estimador  $\pm z_{\alpha/2} \times \text{se}_{\text{estimador}}$

Qual deverá ser o tamanho da amostra?

$$|\hat{p} - p| \leq \epsilon$$

$\Downarrow$

$$z_{\alpha/2} \cdot \text{Ne}_{\tilde{p}} = z_{\alpha/2} \sqrt{\frac{p(1-p)}{n+4}} \leq \epsilon$$

$$n \geq \left(\frac{z_{\alpha/2}}{\epsilon}\right)^2 p(1-p) - 4$$

A desconhecido?

$$\hat{p}(1 - \hat{p}) = \hat{p}^2 \leq 0,25 \quad \Leftrightarrow \quad \sqrt{\hat{p}(1-\hat{p})} \leq 0,5$$

$\Downarrow$

$$n \geq \left(\frac{z_{\alpha/2}}{\epsilon}\right)^2 \times 0,25 - 4$$

Intervalos de Confiança Bilaterais: (estimativa  $\pm$  margem de erro)

método	estimativa	margem de erro
Wald	$\hat{p}$	$z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Agresti-Coull ( $\alpha=0.05$ )	$\tilde{p}$	$z_{\alpha/2} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$

Intervalos de Confiança Unilaterais: um dos limites é  $\infty / 0 / 1$

método	estimativa	margem de erro
Wald	$\hat{p}$	$z_{\alpha} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Agresti-Coull ( $\alpha=0.05$ )	$\tilde{p}$	$z_{\alpha} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$

# IC para Diferença de Proporções

Método de Wald:  $\hat{p}_i = \frac{x_i}{n_i}$

$$\text{se } \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \text{ se}$$

Método de Agresti-Coull:  $\tilde{p}_i = \frac{x_i + 1}{n_i + 2}$

$$\text{se } \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$$

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{\alpha/2} \text{ se}$$

A distribuição binomial pode ser usada para determinar a distribuição por amostragem da diferença  $\hat{p}_1 - \hat{p}_2$

## Testes para a Proporção de uma População

Definimos

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

Com o estimador  $\hat{P} = \frac{X}{n}$  para a proporção  $p$ , utilizamos a estatística de teste seguinte:

$$Z_0 = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

e aceitamos  $H_0$  se  $z_0$  está na região de aceitação, definida por:

$$[-z_{\alpha/2}, z_{\alpha/2}]$$

# Folha 6

Numa floresta, 25% dos pinheiros brancos estão infetados com *ferrugem da bolha*. Tendo sido escolhida uma amostra aleatória de 4 pinheiros brancos, considerar  $\tilde{P}$  a proporção amostral de árvores infetadas, ajustada pelo método de Wilson.

(a) Calcular a probabilidade de  $\tilde{P}$  ser igual a

- i. 2/8
- ii. 3/8
- iii. 4/8
- iv. 5/8
- v. 6/8

(b) Representar graficamente a distribuição amostral de  $\tilde{P}$ .

$$\tilde{P} = \frac{x+2}{n+4} \quad X \sim B(4, 0.25)$$

a)

$$i. P(\tilde{P} = 2/8) = P(X=0) = C_0^4 \times 0,25^0 \times 0,75^4 = 0,3164 \quad \checkmark$$

$$ii. P(\tilde{P} = 3/8) = P(X=1) = C_1^4 \times 0,25^1 \times 0,75^3 = 0,4219 \quad \checkmark$$

$$iii. P(\tilde{P} = 4/8) = P(X=2) = C_2^4 \times 0,25^2 \times 0,75^2 = 0,2109 \quad \checkmark$$

$$iv. P(\tilde{P} = 5/8) = P(X=3) = C_3^4 \times 0,25^3 \times 0,75^1 = 0,046835 \quad \checkmark$$

$$v. P(\tilde{P} = 6/8) = P(X=4) = C_4^4 \times 0,25^4 \times 0,75^0 = 0,0039 \quad \checkmark$$

Numa experiência sobre uma mutação numa dada espécie de planta, foram examinados  $n$  indivíduos, destes, 20% foram considerados mutantes.

Determinar o erro padrão de  $\tilde{P}$  para:

- (a)  $n = 100$  (20 mutantes).
- (b)  $n = 400$  (80 mutantes).

a)  $\text{er}_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} = 0,04$  ✓

b)  $\text{er}_{\tilde{p}} = 0,02$  ✓

Num estudo acerca do tipo sanguíneo em primatas não humanos, foram selecionados aleatoriamente 71 orangotangos. Desses, 14 tinham sangue do tipo B.

Construir um intervalo de confiança a 95% para a percentagem de portadores de sangue do tipo B na população de orangotangos.

$$\tilde{p} = \frac{y+2}{n+4} = \frac{16}{75} \quad \tilde{p} \pm 1.96 \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}; \quad \tilde{p} = \frac{y+2}{n+4}$$

I.C.:  $\tilde{p} \pm 1,96 \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} = 0,213 \pm 1,96 \times 0,0461 = [12\%; 31\%]$  ✓

Uma amostra aleatória de 100 indivíduos com uma determinada patologia experimentou um novo medicamento. Em 55% dos casos observaram-se melhorias significativas.

- (a) Determinar limites de confiança, a 95% e a 99%, para a percentagem de indivíduos da população com a patologia que apresentaram melhoria do estado de saúde depois de tomarem o medicamento.
- (b) Pode-se afirmar, com 95% de confiança, que mais de metade das pessoas melhoram com o medicamento?

a)  $\hat{p} = \frac{55 + 2}{100 + 4} = 0,548$   $\text{se} \hat{p} = \sqrt{\frac{0,548(1-0,548)}{100+4}} = 0,0488$

95% ] 45%; 65% [ ✓ 100% ] 42%; 68% [ ✓

b) não, o intervalo contém valores menores do que 50%. ✓

Num controlo de qualidade, ao analisar 200 ovos escolhidos ao acaso da produção total de um aviário, encontraram-se 22 com salmonela.

- (a) Construir um intervalo de confiança a 95% para a proporção de ovos com salmonela nesse aviário.
- (b) Qual deveria ser o tamanho da amostra para se obter uma estimativa com uma precisão dupla e a mesma confiança?
- (c) Poderá afirmar-se que, a um nível de significância de 5%, a percentagem de ovos com salmonela no aviário é diferente de 15 %?

a) IC: (0,073, 0,162) ✓

c) não, 15% ja tem ✓

b) A precisão é metade da amplitude: 0,0445 - ε

$$\varepsilon' = \frac{\varepsilon}{2} = 0,02225 \quad z_{\alpha/2} \times \text{se} \leq \varepsilon' \quad \hat{p} = 0,11765$$

$$1,96 \times \sqrt{\frac{0,11765 \times (1-0,11765)}{n+4}} \leq 0,02225$$

$$\varepsilon = \frac{0,3222}{\sqrt{n+4}} \leq 0,01135 \quad \Leftrightarrow n+4 \geq 805,86 \Rightarrow n \geq 802 \quad \checkmark$$

Vai ser conduzida uma experiência para determinar a proporção  $p$ , de moscas *Drosophila* com uma certa mutação.

- Determinar a dimensão mínima da amostra para que se possa construir um intervalo de confiança para  $p$ , com margem de erro não superior a 0.05 e grau de confiança de 95%.
- Supor agora que 78 de 400 moscas selecionadas aleatoriamente, apresentavam a mutação. Com base nestes valores construir um intervalo de confiança a 95% para  $p$ .

a)  $\hat{p}_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n+4}} \leq 0,05 \Rightarrow 1,96 \times \sqrt{\frac{0,25}{n+4}} \leq 0,05$

$$\Rightarrow n \geq 381 \quad \checkmark$$

b)  $\hat{p} = \frac{78+2}{400+4} = 0,196 \quad \Delta p = \sqrt{\frac{0,196(1-0,196)}{400+4}} = 0,0197$

IC:  $0,196 \pm 1,96 \times 0,0197 = [15,7\%; 23,5\%] \quad \checkmark$

Numa experiência clínica para tratar pacientes com *alterações de ansiedade*, foi administrada hidroxizina em 71 pacientes e 30 deles melhoraram. Num outro grupo, o grupo de controlo, 70 pacientes receberam um placebo e 20 deles melhoraram. Sejam  $p_1$  e  $p_2$  as probabilidades de melhoria utilizando hidroxizina e o placebo, respectivamente.

A partir dos dados da experiência construir um intervalo de confiança a 95% para  $p_1 - p_2$ .

$$\hat{p}_1 = 0,4267 \quad \hat{p}_2 = 0,297 \quad \hat{p}_1 - \hat{p}_2 = 0,1297$$

$$\Delta p = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1+4} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2+4}} = 0,078$$

IC:  $0,1297 \pm 1,96 \times 0,078 = [0; 28\%] \quad \checkmark$

Foi realizado um estudo comparativo entre um tratamento inovador e o tratamento convencional para uma certa doença. Assim, foram selecionadas aleatoriamente 100 pessoas com essa doença, e divididas aleatoriamente em 2 grupos. Um grupo de 49 indivíduos recebeu o tratamento convencional e os restantes 51 receberam o tratamento inovador. No grupo que seguiu o tratamento convencional, 12 pessoas apresentaram efeitos secundários associados à medicação, e no grupo sujeito ao novo tratamento esse número foi de 16 indivíduos.

- Construir um intervalo de confiança a 95% para a diferença entre as proporções de indivíduos com efeitos secundários em cada uma das populações consideradas.
- Haverá diferenças entre os dois tratamentos?

a)  $\hat{p}_1 = 0,264$

$$\hat{p}_2 = 0,327$$

$$\hat{p}_1 - \hat{p}_2 = -0,063$$

$$se = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{49+4} + \frac{\hat{p}_2(1-\hat{p}_2)}{51+4}} = 0,0876$$

$$IC_{95\%}: -0,063 \pm 1,96 \times 0,0876 = ] -23\%; 11\% [ \quad \checkmark$$

b) não ✓  
OÉ

É clinicamente aconselhado repouso no período final da gravidez, a mulheres grávidas de gémeos, para reduzir o risco de parto prematuro. Uma amostra aleatória de 212 mulheres com gestações gemelares foi dividida aleatoriamente em dois grupos. Num deles foi prescrito este procedimento, e no outro grupo (de controlo) não, i.e. sem período final de repouso.

A tabela a seguir mostra a incidência de parto prematuro (menos de 37 semanas de gestação).

	Repouso na cama	Controlo
número de partos prematutos	32	20
número de mulheres	105	107

Construir um intervalo de confiança de 95% para  $p_r - p_c$ , sendo  $p_r$  e  $p_c$  as probabilidades de parto prematuro no caso de repouso e no caso de controlo, respetivamente.

O intervalo de confiança sugere que o período final de gestação passado em repouso é benéfico?

$$\hat{p}_r = 0,312$$

$$\hat{p}_c = 0,198$$

$$\hat{p}_r - \hat{p}_c = 0,114$$

$$se = \sqrt{\frac{\hat{p}_r(1-\hat{p}_r)}{105+4} + \frac{\hat{p}_c(1-\hat{p}_c)}{107+4}} = 0,05366$$

$$IC: 0,114 \pm 1,96 \times 0,05366 = ] 0,88\%; 22\% [ \quad \text{NÃO!} \quad \checkmark$$

# Dados Qualitativos

Testes Não Paramétricos: testes do  $\chi^2$  (qui-quadrado)

1. Fixar o nível de significância
2. Definir as hipóteses
3. Definir a estatística de teste

Teste do  $\chi^2$  1. qualidade e ajustamento  
2. independência  
3. homogeneidade

Estatística de Teste:  $\chi_s^2$

Distribuição  $\chi^2$ :  $\boxed{\sum_{i=1}^n Z_i^2 + \dots + Z_n^2 \sim \chi^2_n}$ ,  $Z_i \sim N(0, 1)$ ,  $n = gl$

•  $Y_e$  W ~  $\chi^2_n$ :  $E(W) = n$        $V(W) = 2n$

Teste de (Qualidade de) Ajustamento: avaliar se os dados podem ser considerados provenientes de uma população com distribuição ...

$H_0$ : probabilidade das categorias envolvidas

$$\boxed{\chi_s^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}}$$

$O_i$ : frequência observada

$E_i$ : frequência esperada ( $H_0$ )

$E_i = n f_i$

Sob  $H_0$ , a estatística de teste  $\chi^2_S = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$  segue aproximadamente numa distribuição  $\chi^2_S$  com  $k-1$  graus de liberdade:

$$\chi^2_S \sim \chi^2_{k-1}$$

Se num teste de ajustamento a variável é dicotómica, então a hipótese nula não é composta e podem-nos considerar alternativas e conclusões unilaterais.

- O teste de ajustamento do  $\chi^2$  é baseado na **comparação das frequências absolutas**.
- A **soma das frequências** observadas é igual à soma das esperadas, ou seja, igual ao tamanho da amostra ( $n$ ).
- As **hipóteses** de um teste de ajustamento podem envolver mais do que uma afirmação, como no primeiro exemplo,

$$H_0 : p_1 = \frac{3}{16}; \dots; p_5 = \frac{2}{16}$$

Como foi dito, tais hipóteses dizem-se **hipóteses compostas**.

- A **distribuição da estatística** do teste sob  $H_0$  é **aproximada**, pelo que a **dimensão da amostra deve ser suficientemente elevada**.
- Sugere-se** que, em geral, o **valor das frequências** esperadas seja **pelo menos 5**, e pode ser conveniente agrupar categorias.
- Se num teste de ajustamento a variável é **dicotómica**, então a **hipótese nula não é composta**, e é possível considerar alternativas, e **conclusões unilaterais**.

**Teste de Independência:** a partir de uma amostra, testar a hipótese de duas variáveis aleatórias serem independentes

**Tabela de Contingência:** tabela  $2 \times 2$  — duas variáveis observadas em cada um dos individuos de uma amostra aleatória única

$H_0$ : São independentes

$H_1$ : não São independentes

$$\chi_s^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

- Se os acontecimentos são independentes,  $P(A \cap B) = P(A) \times P(B)$
- Se  $H_0$  for verdadeira, com  $\chi_s^2$  segue aproximadamente uma distribuição de qui-quadrado com  $gl$  graus de liberdade
- Se  $gl < \chi_{\alpha/2}$ , rejeita-se  $H_0$

A partir de uma amostra aleatória de tamanho  $N$  ( $T_G$ ), constrói-se a **tabela de contingência** com as frequências observadas. Determina-se o **valor da estatística do teste**  $\chi_s^2$ , considerando todas as  $r \times k$  células:

Para cada uma das  $r \times k$  células determina-se a **frequência esperada**.

$$\chi_s^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

$T_L$ : Total da Linha

$T_C$ : Total da Coluna

$T_G$ : Total Global

$$e_i = \frac{T_L \times T_C}{T_G}$$

		freq observadas ( $o_i$ )	Y			
			B <sub>1</sub>	...	B <sub>k</sub>	Total
X	A <sub>1</sub>					
	A <sub>r</sub>					
		Total			N	

$\chi_s^2$  é o valor de uma variável aleatória  $\chi_s^2$  (estatística de teste).

Sob  $H_0$ ,  $\chi_s^2$  segue **aproximadamente** uma distribuição do **qui-quadrado** com um número de graus de liberdade que é dado por:  $gl = (r - 1) \times (k - 1)$ .

**Teste de Homogeneidade:** a partir de uma amostra, testar a hipótese de várias populações terem uma mesma característica

## Tabela de Contingência

	P 1	P 2	TOTAL
A 1			
A 2			
TOTAL			

$H_0$ : não existe associação

$H_1$ : existe associação

Pretendemos testar se uma variável  $X$ , que tem  $r$  categorias ( $A_1, \dots, A_r$ ), tem a mesma distribuição em  $k$  populações.

$H_0$ : A distribuição de probabilidade de  $X$  é igual em todas as populações.

A partir de  $k$  amostras aleatórias independentes retiradas de cada uma das populações constrói-se a tabela de contingência com as frequências observadas.

$H_1$ : A distribuição de probabilidade de  $X$  **não** é igual em todas as populações.

freq observadas ( $O_i$ )		População			
		$P_1$	$\dots$	$P_k$	
X	$A_1$				N
	$\vdots$				
$A_r$					
Total				N	

Para cada uma das  $r \times k$  células determina-se a respetiva frequência esperada:

$$e_i = \frac{\text{Total linha} \times \text{Total coluna}}{\text{Total global}}$$

freq esperadas ( $e_i$ )		População			
		$P_1$	$\dots$	$P_k$	
X	$A_1$	$\frac{N_{A_1} \times n_1}{N}$			$N_{A_1}$
	$\vdots$				$\vdots$
$A_r$					$N_{Ar}$
Total		$n_1$	$\dots$	$n_k$	N

Depois de calcular as  $e_i$ , determinamos o valor da estatística do teste, considerando todas as  $r \times k$  células:

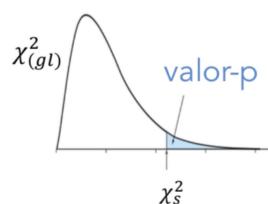
$$\chi_s^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

$\chi_s^2$  é o valor de uma v.a.  $\chi_S^2$  que, sob  $H_0$ , segue aproximadamente uma distribuição  $\chi^2$  com  $(r - 1) \times (k - 1)$  graus de liberdade.

Se  $\text{valor-p} < \alpha$  então rejeitamos  $H_0$ , isto é, há evidência de  $H_1$  a um nível de significância  $\alpha$ .

Note-se que estes testes são **unidirecionais**, isto é, pretendemos saber se

$$\text{valor-p} < \alpha \quad (\text{ou } \chi_s^2 > \chi_{(gl, \alpha)}^2)$$



Num teste de **independência** são observados os valores de **duas variáveis** numa **única população**.

Temos portanto uma única amostra na qual cada elemento é observado relativamente aos diversos atributos, pretendemos saber se as variáveis são ou não independentes.

As hipóteses, neste caso, formulam-se como:

$H_0$ : As variáveis são independentes.

$H_1$ : As variáveis **não** são independentes.

Num teste de **homogeneidade** são observados os valores de **uma variável** em **várias populações**.

Temos várias amostras independentes e queremos saber se as distribuições de probabilidade são iguais nas diferentes populações.

Neste caso temos:

$H_0$ : As distribuições de probabilidade da variável nas diversas populações são iguais.

$H_1$ : As distribuições de probabilidade da variável nas diversas populações **não** são iguais.

*Independência: duas variáveis, uma população*

*Homogeneidade: uma variável, duas populações*

# Folha 7

Um cruzamento entre abóboras brancas e amarelas conduziu a descendentes com as cores na tabela:

Cor	Branco	Amarelo	Verde
Número de Descendentes	155	40	10

Usar um teste de  $\chi^2$ , e um nível de significância  $\alpha = 0.05$ , para verificar se os dados são consistentes com o rácio 12:3:1, previsto por um modelo genético.

$H_0$ :	C	B	A	V
P	$12/16$	$3/16$	$1/16$	

$H_1$ :	C	B	A	V
$\hat{\theta}_i$	155	40	10	205
$\hat{\ell}_i$	153,75	38,44	12,81	

$$gl = 3 - 1 = 2$$

$$\chi^2_s = \frac{\sum_i (\hat{\theta}_i - \hat{\ell}_i)^2}{\hat{\ell}_i} = 0,691$$

$$\chi^2_{0.05, 2} = 5,9915$$

$$\chi^2_s < \chi^2_{2, 0.05}$$

↓

valor- $\chi^2_s > \alpha \rightarrow$  não se rejeita  $H_0$



Considerar o enunciado do exercício anterior, e supor que a amostra tinha a mesma composição, mas de tamanho 10 vezes superior, ou seja:

1550 descendentes brancos

400 amarelos

100 verdes

Estes dados seriam consistentes com a previsão do modelo?

$$\chi^2_S = 6,91 > \chi^2_{0.05,2} \rightarrow \text{valor-1} < \alpha$$

$\downarrow$   
rejeita- $H_0$

✓

Entre os  $n$  bebés nascidos numa determinada cidade, 51% eram rapazes.

Supor que queremos testar a hipótese de que a probabilidade de nascer um rapaz é 0.5.

Considerando uma alternativa bilateral, calcular o valor observado da estatística do  $\chi^2$  e dizer o que se pode concluir para  $\alpha = 0.05$ .

Considerar os seguintes tamanhos da amostra:  $n = 1000$ ,  $n = 5000$  e  $n = 10000$  (para cada caso calcular o valor-p usando o R).

$$H_0: p = 0,5$$

$$H_1: p \neq 0,5$$

$$gl = 1$$

	$\delta$	$\varphi$
$\theta_i$	$0,51_n$	$0,49_n$
$\ell_i$	$0,5_n$	$0,5_n$

$$n = 1000:$$

$$\chi^2_s = 0,4 < \chi^2_{0.05,1} = 3,8415$$

$V_f > \alpha \rightarrow$  não se rejeita  $H_0$  ✓

$$n = 5000:$$

$$\chi^2_s = 2 < \chi^2_{0.05,1} = 3,8415$$

$V_f < \alpha \rightarrow$  não se rejeita  $H_0$  ✓

$$n = 10000:$$

$$\chi^2_s = 4 > \chi^2_{0.05,1} = 3,8415$$

$V_f > \alpha \rightarrow$  rejeita-se  $H_0$  ✓

Pessoas que colhem cogumelos selvagens por vezes comem acidentalmente o cogumelo tóxico *Amanita phalloides*.

Ao analisar 205 casos de envenenamento com este tipo de cogumelo de 1971 a 1980, os pesquisadores verificaram que 45 das vítimas tinham morrido.

Utilizar um teste de hipóteses para  $\alpha = 0.05$ , para comparar a taxa de mortalidade no período 1971-80, com a taxa de mortalidade de 30% registada até 1970.

Os dados fornecem evidência de que a taxa baixou?

$$H_0: \mu = 0,3$$

$$H_1: \mu \neq 0,3$$

DA

	$\hat{\mu}$	$\hat{\sigma}$
$H_0$	45	160
$H_1$	61,5	143,5

$$\chi^2_s = 6,32$$

>

$$\chi^2_{0.05,1} = 3,84$$

$$V_1 < \alpha$$

rejeitar  $H_0$



Foi realizada uma experiência para investigar a eficácia de protetores externos das ancas na prevenção de fraturas da anca em idosos, com uma amostra de  $n = 1801$  pessoas.

As pessoas foram distribuídas aleatoriamente em dois grupos.

Num dos grupos foram utilizados protetores de anca ( $n_1 = 653$ ) e o outro serviu de grupo controlo ( $n_2 = 1148$ ).

Foi registado o número de fraturas em cada grupo e os dados constam da tabela abaixo.

P	C
29	51
624	1097

Resposta	Tratamento		Total
	Protetor	Controlo	
Fratura Anca	13	67	80
Não Fratura Anca	640	1081	1721
	653	1148	1801

$$H_0: \lambda_{\mu A} = \lambda_{\mu N}$$

$$H_1: \lambda_{\mu A} \neq \lambda_{\mu N}$$

$$\chi^2_s = 14,5$$

$$\chi^2_{0.01,1} = 6,63 \rightarrow \chi^2_s > \chi^2_{0.01,1} \rightarrow V_A < \alpha \rightarrow \text{rejeita } H_0$$

Num ensaio clínico, uma amostra de indivíduos que tiveram AVC foi dividida aleatoriamente em 2 grupos. A um grupo foi administrado 'Ancrod'. Ao outro um placebo.

Uma variável de resposta foi a ocorrência ou não de hemorragia intracraniana. Os resultados estão na tabela:

Hemorragia	Tratamento		
	Ancrod	Placebo	
Sim	13	5	18
Não	235	247	482
Total	248	252	500

Usar um teste  $\chi^2$  para determinar se a diferença entre as taxas de hemorragia é significativa. Considerar  $\alpha = 0.05$  e uma alternativa bilateral.

A	P
8,9	9,1
239,1	242,9

$$\chi_s^2 = 3,82 \quad < \quad \chi_{0.05, 1}^2 = 3,84$$

$$V_s > \alpha$$

não se rejeita  $H_0$  ✓

$$H_0: \mu_A = \mu_B$$

$$H_1: \mu_A \neq \mu_B$$

O tipo de habitat da mosca Drosophila foi estudado capturando moscas de 2 locais diferentes. As moscas foram marcadas com o local de captura e depois libertadas num local intermédio entre os locais originais.

Mais tarde as moscas foram recapturadas nos 2 locais. Os resultados estão resumidos na tabela seguinte:

Local Original Captura	Local Recaptura		134
	I	II	
I	78	56	134
II	33	58	91
	111	114	225

Testar a hipótese nula de **independência** entre os locais de captura e recaptura.

$$H_0: \pi_1 = \pi_n$$

$$H_1: \pi_1 \neq \pi_n$$

	I	II
66,1	63,9	
44,9	46,1	

$$\chi^2_s = 1,0,45$$

$$\chi^2_{0.01,1} = 6,63$$

$\chi^2 < \alpha$  → rejeitar  $H_0$  ✓

Num ensaio clínico, pacientes com osteoartrite dolorosa do joelho selecionados ao acaso numa população, foram aleatoriamente atribuídos a um de cinco tratamentos: Glucosamina, Condroitina, ambos, placebo, ou Celebrex (terapia usual).

Para cada tratamento registou-se o número de pacientes com melhoria na dor ou na capacidade de movimentação.

Tratamento	Tamanho amostra	Resultado sucesso	
		Número	Percentagem
Glucosamina	317	190	60.0
Condroitina	318	202	63.5
ambos	317	208	65.6
placebo	313	178	56.9
Celebrex	318	214	67.3

$$\chi^2_s = 9,696 \quad > \quad \chi^2_{0.05, 4} = 9,49$$

$\downarrow$

$$V_t < \alpha \rightarrow \text{rejeitar } H_0 \quad \checkmark$$

Numa clínica de alergologia foram selecionados 245 pacientes com idade inferior a 19 anos. Cada um foi classificado segundo a idade e a presença ou não de alergia aos ovos. De 133 pacientes com idade superior a 3 anos, 30 eram alérgicos a ovos e dos 112 pacientes com idade inferior ou igual a 3 anos, 32 exibiam essa alergia.

$\text{Idade}$	$A$	$\bar{A}$	$\text{Total}$
$< 3$	32	80	112
$> 3$	30	103	133
Total	62	183	245

$\text{Idade}$	$A$	$\bar{A}$	$\text{Total}$
$< 3$	28,3	83,7	112
$> 3$	33,7	99,3	133
Total	62	183	245

$$H_0: \pi_A = \pi_{\bar{A}}$$

$$H_1: \pi_A \neq \pi_{\bar{A}}$$

$$\chi^2_s = 1,19 < \chi^2_{0.05,1} = 3,84$$

$$V_1 > \alpha$$

não se rejeita  $H_0$  ✓

Com o objectivo de identificar factores de risco de doença coronária analisaram-se duas amostras de pessoas com essa doença. Dos 215 homens e das 1140 mulheres das amostras, observou-se que 58 dos homens e 217 das mulheres tinham diabetes. Com estes dados construiu-se o intervalo de confiança a 95%, para a diferença das proporções de diabéticos nas duas populações ( $p_{Hd} - p_{Md}$ ).

De acordo com esse intervalo, e com 95% de confiança, podemos afirmar que,

- a diabetes é um fator de risco superior para os homens
- a diabetes é um fator de risco superior para as mulheres
- a diabetes é um fator de risco igual para as duas populações
- com estes dados não é possível retirar conclusões sobre a diferença de risco nas duas populações
- nenhuma das hipóteses anteriores está correta.

$$H_0: \hat{p}_{Hd} - \hat{p}_{Md} = 0$$

$$H_1: \hat{p}_{Hd} - \hat{p}_{Md} \neq 0$$

$$\begin{array}{lll} 215 & \text{♂} & 58 \\ 1140 & \text{♀} & 217 \end{array} \quad 95\% \quad \alpha = 0,05$$

$$\tilde{p}_H = \frac{58 + 1}{215 + 2} = 0,272$$

$$\tilde{p}_M = \frac{217 + 1}{1140 + 2} = 0,191$$

$$\Delta \tilde{p} = \sqrt{\frac{\tilde{p}_H(1-\tilde{p}_H)}{215+2} + \frac{\tilde{p}_M(1-\tilde{p}_M)}{1140+2}} = 0,0324$$

$$\tilde{p}_H - \tilde{p}_M \pm \Delta \tilde{p}_{\alpha/2} = 0,081 \pm 0,0324 \times 1,96 = (0,0175; 0,144)$$

$$\hat{p}_{Hd} > \hat{p}_{Md} \quad \leftarrow \quad \hat{p}_{Hd} - \hat{p}_{Md} > 0 \quad \text{↓}$$

Foi efetuado um estudo num aviário para avaliar o consumo semanal de água por frango (em litros) num determinado período do ano. Foram selecionados aleatoriamente 15 frangos, e registaram-se os valores do consumo de água (em litros) observado para cada um dos frangos, durante uma semana. Estudos anteriores evidenciam, nesse período, uma média de consumo semanal de água da população de frangos do referido aviário, de 1.2 litros.

Para concluir o estudo é agora necessário testar a hipótese  $H_0 : \mu = 1.2$  litros, face à hipótese alternativa  $H_1 : \mu \neq 1.2$  litros. Para isso calculou-se o valor-p relativo à amostra, obtendo-se o valor 0.0254. Com este valor-p apenas uma das opções seguintes é incorreta. Indicar qual.

- Com um nível de significância de 0.02 não rejeitamos  $H_0$  ✓
- Com um nível de significância de 0.05 aceitamos  $H_1$  ✓
- Com um nível de significância de 0.03 rejeitamos  $H_0$  ✓
- Com um nível de significância de 0.01 não aceitamos  $H_1$  ✓
- Nenhuma das hipóteses anteriores está correta.

$$n = 15 \quad \bar{x} = 1,2 \quad H_0: \mu = 1,2 \quad H_1: \mu \neq 1,2$$

$$\text{valor-}p: 0,0254 \rightarrow \text{confiança em } H_0$$

Na construção de um intervalo de confiança para a média de uma população normal com variância conhecida, se diminuirmos a confiança, a amplitude do intervalo,

- Aumenta
- Diminui
- Não se altera
- aumenta se a variância e a dimensão da amostra se mantiverem constantes
- nenhuma das hipóteses anteriores está correta.

$$\begin{aligned} &\text{diminui confiança } (1-\alpha) \rightarrow \text{aumenta significância } (\alpha) \rightarrow z \text{ diminui} \\ &IC = \dots \pm z_{\alpha/2} \end{aligned}$$

Num teste de hipóteses, se a hipótese nula for “Os holandeses não diferem dos ingleses em altura”, qual deverá ser a hipótese alternativa?

- Os holandeses são mais altos que os ingleses
- Os ingleses são mais altos que os holandeses
- As afirmações anteriores podem ser consideradas como hipóteses alternativas
- A altura dos holandeses difere da altura dos ingleses
- Não há dados para formular uma hipótese alternativa.

$$H_0: h_H - h_M = 0$$

$$H_1: h_H - h_M \neq 0$$

$$H_1: h_H - h_M > 0$$

$$H_1: h_H - h_M < 0$$

Foram recolhidas quatro amostras aleatórias distintas de uma população normal. Com base em cada uma das amostras, foram construídos intervalos de confiança para a média dessa população com determinados graus de confiança.

Os intervalos de confiança obtidos foram os seguintes:

$$I_1 = ]28.84, 31.10[; \quad I_2 = ]29.63, 32.19[; \quad I_3 = ]26.19, 28.21[; \quad I_4 = ]27.07, 29.63[$$

O(s) intervalo(s) que corresponde(m) a um maior grau de confiança, é (são):

$I_2$  e  $I_4$

$I_1$

$I_3$

$I_1$  e  $I_3$

O grau de confiança é igual nos quatro intervalos.

$$A_1 = 2,26$$

$$A_2 = 2,56$$

$$A_3 = 2,02$$

$$A_4 = 2,56$$

$\rightarrow$  maior  $A \rightarrow$  maior confiança  $(1-\alpha)$

Para estimar a concentração média de açúcares num refrigerante de uma dada marca, foi medida a concentração de açúcares em 49 embalagens escolhidas aleatoriamente entre a produção dessa marca numa determinada semana. Obteve-se uma média (em unidades arbitrárias), de 10.12 e um desvio padrão de 0.65.

O intervalo de confiança a 90% para  $\mu$ , construído com base nesta amostra, é:

]8.81, 11.43[

]9.47, 10.77[

]9.96, 10.28[

]9.03, 11.21[

nenhuma das hipóteses anteriores está correta.

$$\begin{aligned} IC : \bar{x} \pm t_{\alpha/2, n-1} s_e &= 10,12 \pm t_{0,05, 48} \frac{s}{\sqrt{n}} = \\ &= 10,12 \pm t_{0,05, 48} \times \frac{0,65}{\sqrt{49}} = 10,12 + 1,6772 \times 0,093 \\ &= (9,96; 10,28) \end{aligned}$$

Numa exploração de gado, foram testados dois novos medicamentos ( $A$  e  $B$ ) em 25 animais doentes, selecionados aleatoriamente, que foram divididos em dois grupos. A um foi administrado o medicamento  $A$  e ao outro o  $B$ . O estudo pretendia saber qual dos medicamentos permitia a recuperação média mais rápida dos animais (unidades arbitrárias). Os valores de interesse para a realização de um teste-t, estão resumidos na tabela seguinte:

	$A$	$B$
$n$	10	15
$\bar{x}$	165	180
$s$	42	56

Para um nível de significância de 5%, podemos afirmar que,

- o tempo médio de recuperação com  $A$  é superior ao de  $B$ , com valor-p > 0.001
- os dados não evidenciam diferença entre  $A$  e  $B$  relativamente ao tempo médio de recuperação
- o tempo médio de recuperação com  $B$  é superior ao de  $A$ , com valor-p < 0.001
- o tempo médio de recuperação com  $B$  é superior ao de  $A$ , com  $0.01 < \text{valor-p} < 0.05$
- nenhuma das hipóteses anteriores está correta.

$$H_0: \mu_1 - \mu_2 = 0$$

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{42^2}{10} + \frac{56^2}{15}} = 19,63$$

$$t = \frac{165 - 180}{19,63} = -0,764$$

$$gl = 10 + 15 - 2 = 23$$

$$\text{valor-p} = 2 \times P(|T| > 0,764) = 2 \times 0,15 = 0,3$$

$$\alpha = 0,05$$

$$\text{valor-p} = 0,3$$

$$0,3 > 0,05 \rightarrow \text{aceitas } H_0$$

OU

0

↓

$$IC: \bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2, n_1+n_2-2} \times se (-55,61 ; 25,61)$$

$$= -15 \pm t_{0,025, 23} \times 19,63 = -15 \pm 2,0687 \times 19,63 =$$

Num teste de hipóteses, o Erro de Tipo I é o erro que se pode cometer quando,

rejectar  $H_0$  quando não devíamos

- Se aceita a hipótese nula
- A hipótese alternativa é verdadeira
- Só se conhece a distribuição da estatística de teste no caso da hipótese nula ser verdadeira
- Se rejeita a hipótese nula
- Nenhuma das hipóteses anteriores está correta.

Testou-se um novo inseticida sistémico contra os afídeos da família *Aphididae*, numa amostra aleatória de 100 pessegueiros atacados por esta praga. Em 63% dos casos observaram-se melhorias significativas.

Relativamente aos pessegueiros que sofrem desta praga, e que foram tratados como novo inseticida, podemos afirmar que (considerar o método de Wald),

- a probabilidade de melhorarem é 0.63
- a probabilidade de melhorarem é pelo menos 95%
- com 95% de confiança, podemos garantir que menos de 50% melhoraram
- com 95% de confiança, mais de metade melhoraram
- nenhuma das hipóteses anteriores está correta.

$$n = 100$$

$$n = 63$$

$$\hat{p} = 0,63$$

$$\alpha = 0,05$$

$$IC: \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 0,63 \pm z_{0,025} \times \sqrt{\frac{0,63(1 - 0,63)}{100}}$$

$$= 0,63 \pm 1,96 \times 0,05 = (0,532; 0,728)$$

Foi realizado um teste-t para testar a hipótese  $H_0 : \mu = 4$  contra a hipótese alternativa  $H_1$ . A amostra utilizada tinha dimensão 50, e os valores obtidos para a estatística do teste e valor-p, foram respectivamente:  $t_s = 2.02$  e valor-p = 0.048.

De acordo com estes valores, podemos concluir que,

- Sob  $H_0$ , a probabilidade de não rejeitar  $H_0$  é 0.048
- Há evidência estatística de  $H_1$  a um nível de significância de 1%
- Há evidência estatística de  $H_0$  a um nível de significância de 5%
- A probabilidade de  $H_1$  ser verdadeira é 0.952
- Nenhuma das hipóteses anteriores está correta.

$$H_0: \mu = 4 \quad H_1: \mu \neq 4 \quad n = 50$$
$$t_s = 2,02 \quad \text{valor-p} = 0,048$$

Das afirmações abaixo indicar a única que é verdadeira.

- O desvio padrão populacional é obtido através da recolha de apenas uma amostra
- A variância populacional é obtida através da recolha de apenas uma amostra
- O erro padrão é obtido através da recolha de apenas uma amostra
- O erro padrão é um parâmetro que permite calcular o centro de uma distribuição
- Nenhuma das afirmações anteriores é verdadeira.

Seja  $X$  a variável aleatória que representa o crescimento (num mês) de uma determinada espécie animal quando submetida a uma dieta específica. Estudos anteriores sugerem que o desvio padrão é aproximadamente 91.5 g.

Dos pares seguintes, relativos respetivamente à dimensão amostral e erro padrão da média, apenas um está correto. Assinalar qual.

- 9 e 32.3  $\sigma = 96,9$
  - 16 e 21.2  $\sigma = 84,8$
  - 9 e 30.5  $\sigma = 91,5$
  - 16 e 24.1  $\sigma = 96,4$
- Nenhuma das hipóteses anteriores está correta.

$$\sigma_e = \frac{\sigma}{\sqrt{n}} \Leftrightarrow \sigma_e = \sigma \sqrt{\frac{1}{n}}$$

Com o objetivo de comparar as médias de um parâmetro em duas populações A e B, foram recolhidas amostras de cada uma das populações, representadas respectivamente por  $x$  e  $y$ , e foi realizado um teste-t.

Alguns dos resultados obtidos foram os seguintes:

$$\mu_x = 58.458; \quad \mu_y = 62.591$$

Estatística do teste: -1.686

Número de graus de liberdade da distribuição t: 44

Valor-p: 0.099

Intervalo de confiança a 95% para a diferença de médias: ] - 9.072; 0.807[

Com estes valores podemos concluir que,

- Ao nível de 5%, as populações A e B são iguais
- As médias das populações A e B não são iguais com 95% de confiança
- As médias das populações A e B são iguais
- Ao nível de 5%, há evidências para considerar que as médias das populações A e B são iguais
- Nenhuma das hipóteses anteriores está correta.

Foi realizado um estudo relativo à temperatura mínima necessária para o desenvolvimento de determinada bactéria, que se admite ser normalmente distribuída. Recolheu-se uma amostra de dimensão 17, para a qual se obteve, em média, um valor de 5.5 graus, e um desvio padrão de 0.5 graus.

Então o valor que pertence ao intervalo de confiança a 95% para a temperatura mínima média, é,

- 5.84
- 5.25
- 5.79
- 5.21

Nenhuma das hipóteses anteriores está correta.

$$n = 17 \quad \bar{x} = 5,5 \quad s = 0,5 \quad \alpha = 0,05$$

$$IC: \bar{x} \pm t_{\alpha/2, n-1} s = 5,5 \pm t_{0,025, 17-1} \times \frac{s}{\sqrt{n}} =$$

$$= 5,5 \pm t_{0,025, 16} \times \frac{0,5}{\sqrt{17}} = 5,5 \pm 2,1199 \times 0,121$$

$$= (5,24; 5,76) \quad OU \quad 5,5 \pm$$

A amplitude de um intervalo de confiança para a média  $\mu$  de uma população normal diminui com um aumento do

- Desvio padrão da amostra       Tamanho da amostra       Grau de confiança
- Desvio padrão da população       Nenhuma das hipóteses anteriores está correta.

maior  $\frac{\sigma}{\sqrt{n}} \rightarrow$  maior amplitude

$$\frac{\sigma}{\sqrt{n}}$$

Pretende-se saber qual a percentagem  $p$  de cura de uma certa doença quando os doentes são submetidos a um determinado tratamento. Numa amostra aleatória de 40 doentes submetidos a esse tratamento, 25 ficaram curados.

Usando o método de Agresti-Coull obtém-se o seguinte intervalo de confiança a 95% para  $p$ :

- ]46.9, 75.8[       ]49.9, 75.1[       ]56.3, 66.4[       ]57.5, 67.5[

- Nenhuma das hipóteses anteriores está correta.

$$\tilde{p} = \frac{n + 2}{n + 4} = \frac{25 + 2}{40 + 4} = \frac{27}{44} = 0,614$$

$$\text{se} = \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}} = \sqrt{\frac{0,614 \times (1 - 0,614)}{44}} = 0,0734$$

$$p = \tilde{p} \pm 1,96 \text{ se} = 0,614 \pm 1,96 \times 0,0734 = (0,47; 0,76)$$

Para testar a hipótese  $H_0 : \mu = 75$  contra a hipótese  $H_1 : \mu \neq 75$  foi conduzido um teste-t a partir de uma amostra de dimensão 40 para a qual o valor da estatística foi  $t_s = -2.405$  e o valor-p foi 0.021.

Com base nestes dados, podemos afirmar que,

- Há evidência estatística de  $H_1$  a um nível de significância de 1%
- A probabilidade da hipótese  $H_1$  ser verdadeira é 0.021
- A um nível de significância de 5% devemos aceitar  $H_0$
- Sob  $H_0$ , a probabilidade de se ter uma estatística de teste igual ou mais extrema que  $-2.405$  é 0.021
- Nenhuma das hipóteses anteriores está correta.

Num teste de hipóteses, admitindo verdadeira a hipótese nula, dizemos que um resultado é estatisticamente significativo se a probabilidade de obter um resultado tão extremo quanto o obtido for menor que uma probabilidade pré-especificada (o nível de significância do teste).

Diz-se também que estamos perante um erro de Tipo I quando, *rejeita  $H_0$  quando não devíamos*

- Concluímos que não há evidências a favor de um ‘novo’ valor para um parâmetro da população quando na realidade ele existe *aceita  $H_0$*
- Concluímos que a estatística do teste é significativa quando, de facto, não o é *→ aceita  $H_1$*
- Os dados amostrais não são representativos do fenómeno que está a ser estudado
- Concluímos que há evidências a favor de um ‘novo’ valor para um parâmetro da população quando de facto não é verdade *aceita  $H_1$  & rejeita  $H_0$*
- Nenhuma das hipóteses anteriores está correta.

Para estimar a concentração média  $\mu$  de chumbo no sangue das crianças de uma certa população, foi medida a concentração de chumbo no sangue (em  $\text{ng ml}^{-1}$ ) de 49 crianças escolhidas aleatoriamente dessa população. Obteve-se uma média de 10.12 e um desvio padrão de 0.65. O intervalo de confiança a 95% para  $\mu$  construído com base nesta amostra é:

]9.93, 10.31[

]8.81, 11.43[

]9.03, 11.21[

]9.47, 10.77[

Nenhuma das hipóteses anteriores está correta.

$$\text{IC: } \bar{x} = t_{\alpha/2, n-1} \times s_e = 10,12 \pm t_{0.025, 48} \times \frac{0,65}{\sqrt{49}}$$

$$= 10,12 \pm 2,0106 \times \frac{0,65}{7} = (9,93; 10,31)$$

Usando o método de Wald, qual o tamanho mínimo que deve ter uma amostra para que se possa construir, com base nela, um intervalo de confiança para uma proporção com margem de erro não superior a 0.05, e grau de confiança de 95%?

835

485

Não há dados suficientes para responder

95

Nenhuma das hipóteses anteriores está correta.

$$\tilde{p}(1 - \tilde{p}) \leq 0,25$$

$$1,96 \times \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n+4}} \leq 0,05$$

$$\Leftrightarrow 1,96 \times \sqrt{\frac{0,5^2}{n+4}} < 0,05 \Rightarrow n \geq \left( \frac{1,96 \times 0,5}{0,05} \right)^2 - 4$$

$$\rightarrow n \geq 381 \rightarrow \text{A.C}$$

$$\underline{\text{WALD: } 381 + 4 = 385}$$

$$n \geq 385$$