

- A Probabilidade e a Inferência Estatística.
- A amostragem. Principais Distribuições por Amostragem
- Tópicos de **Estimação** e de **Inferência**
 - estimação pontual
 - estimação por intervalos
 - testes de hipóteses

A Probabilidade e a Inferência Estatística

Começamos esta unidade curricular pelo estudo descritivo de uma **amostra**. Passamos depois pela **Introdução à Teoria da Probabilidade** com o estudo de alguns dos modelos de probabilidade (discretos e contínuos).

Veremos agora como utilizar todo esse conhecimento na **Inferência Estatística**. A Inferência Estatística é ... *aquilo que permite o salto que vai da amostra para a população* (Pestana e Velosa, 2008).

A Probabilidade e a Inferência Estatística

Mas ... sem conhecimento da probabilidade não é possível avançar na inferência estatística. Apenas para breve “orientação” podemos dizer que:

- **na probabilidade temos o modelo da população** que nos permite calcular a probabilidade de certos acontecimentos;
- **na inferência estatística parte-se da amostra** e pretende-se obter informação sobre o modelo ou sobre características da população.

Conceitos em Inferência Estatística

A **Inferência Estatística** pretende responder a dois grandes problemas:

- calcular valores aproximados (**estimativas**) e obter **intervalos de confiança** para parâmetros desconhecidos da população.
- formular hipóteses e verificar se há concordância entre o que se supõe e os factos – **testes de hipóteses**

Conceitos básicos:

População → conjunto completo de todas os objectos (elementos) com uma (ou mais) característica(s) comuns;

Unidade Estatística → cada elemento da população;

Amostra → conjunto dos valores efectivamente observados;

Parâmetro de uma população → **constante** desconhecida, cujo verdadeiro valor se pretende “estimar” ou “validar”.

Introdução à Teoria da Amostragem

Amostragem → procedimento de recolha de elementos da população para obter uma amostra.

De uma dada **população** podemos retirar muitas amostras:

- Amostra 1
- Amostra 2
- ...
- Amostra k
- ...

Mas ... quase sempre recolhemos só **uma amostra** para estudarmos uma **característica X** da população.

Seja **(x_1, x_2, \dots, x_n)** uma amostra de n observações da característica, obtidas após um processo de amostragem.

Introdução à Teoria da Amostragem

Vamos considerar a amostra (x_1, x_2, \dots, x_n) como uma realização de n variáveis (X_1, X_2, \dots, X_n) que são “cópias” da variável X

Recapitulando:

- **Antes** da amostragem ser realizada temos n variáveis aleatórias (X_1, X_2, \dots, X_n)
- **Depois** de efectuada a amostragem temos um conjunto de dados que constituem a amostra observada (x_1, \dots, x_n) .

Definição de amostra aleatória

Chama-se **amostra aleatória** de dimensão n a um conjunto de n variáveis aleatórias (X_1, X_2, \dots, X_n) **independentes** e **semelhantes**, i.e., tendo todas a mesma distribuição que é a da característica X em estudo na população.

Tópicos de Estimação

Quais os **parâmetros desconhecidos** da população que nos vão interessar e que **procedimentos** usar para **os estimar**?

Os **procedimentos** são métodos adequados (queremos “os melhores”) para estimar os parâmetros desconhecidos.

Em Estatística, esses métodos consistem na obtenção de variáveis aleatórias “especiais”, chamadas **estimadores**.

Definição de estimador

Chama-se **estimador de um parâmetro** a uma função da amostra aleatória (X_1, X_2, \dots, X_n) , que “permite” obter um valor para o parâmetro desconhecido (esta função não possui quantidades desconhecidas).

NOTA: Não iremos referir procedimentos para obter estimadores

Tópicos de Estimação

Um **estimador** é então **uma variável aleatória** que terá portanto uma dada distribuição (pelo menos aproximadamente).

O valor que o estimador (variável aleatória) toma quando se observa a amostra chama-se **estimativa**.

Definição de estimativa

Estimativa de um parâmetro é **um valor concreto** assumido pelo estimador.

Dada a amostra aleatória, (X_1, X_2, \dots, X_n) , i.e., v.a. **i.i.d. a X** , vamos ver os **parâmetros** que nos vão interessar, os seus **estimadores** e **estimativas**

Tópicos de Estimação

Parâmetro(s) a estimar

Estimador(es)

Estimativa(s)

$$\mu$$

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

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

$$\sigma^2$$

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

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

$$p$$

$$\hat{P} = \frac{X}{n} \text{ (a)}$$

$$\hat{p} = \frac{x}{n} \text{ (b)}$$

$$\mu_1 \text{ ?? } \mu_2$$

$$\bar{X}_1 \text{ ?? } \bar{X}_2$$

$$\bar{x}_1 \text{ ?? } \bar{x}_2$$

$$\sigma_1^2 \text{ ?? } \sigma_2^2$$

$$S_1^2 \text{ ?? } S_2^2$$

$$s_1^2 \text{ ?? } s_2^2$$

$$p_1 \text{ ?? } p_2$$

$$\hat{P}_1 \text{ ?? } \hat{P}_2$$

$$\hat{p}_1 \text{ ?? } \hat{p}_2$$

^(a) X - v.a. que conta o número de sucessos na amostra de dimensão n

^(b) x - número observado de sucessos na amostra de dimensão n .

?? significa que teremos que escolher a melhor operação para fazer a comparação

Propriedades dos estimadores - só uma nota

Seja X a v.a. (população) em estudo e θ um parâmetro desconhecido.

1. Estimador centrado

O estimador $T = T(X_1, X_2, \dots, X_n)$ do parâmetro θ diz-se centrado ou não enviesado se $E[T] = \theta$.

De entre os **estimadores centrados**, a situação desejável é o estimador ter **variância mínima**.

Diz-se então que temos **o estimador mais eficiente**.

Propriedades dos estimadores - só uma nota

Na verdade a propriedade de um estimador ser centrado é uma propriedade que muitos estimadores não possuem e daí ter mais interesse definir **uma medida da diferença entre o estimador e o parâmetro** que ele pretende estimar.

É o chamado **viés** (ou em inglês '**bias**') que representaremos por **$b_{\theta}(T)$** e se define como

$$b_{\theta}(T) = E[T] - \theta$$

Tópicos de Estimação

Suponhamos que temos uma amostra

Como inferir valores de interesse para a população?

- Estimação dos parâmetros
 - Estimação pontual
 - Estimação intervalar → intervalos de confiança
- Testes de hipóteses estatísticas.

Estimação pontual → indicar um único valor como estimativa.

Exemplos de estimativas: \bar{x} , s^2 , \hat{p} , \dots .

Estimação por intervalos → neste caso calcula-se um intervalo (aleatório) que com uma probabilidade elevada contenha o verdadeiro valor do parâmetro.

Estimação por intervalos

Para a construção do intervalo é necessário conhecer a distribuição - exacta ou aproximada - do **estimador** (ou qualquer expressão dele)

Definição de intervalo de confiança

Chama-se **intervalo de confiança** ao intervalo que resulta da concretização do intervalo (aleatório) e é portanto um intervalo **(a, b)** , onde a e b são números reais e $a < b$.

Relembre-se que para obter o intervalo de confiança é então necessário conhecer **a lei do estimador** do parâmetro desconhecido.

Intervalos de confiança

Começemos então por ver como se construiria **um intervalo de confiança para μ** , (valor médio de uma característica X .)

Precisamos de procurar a distribuição do estimador de $\mu \rightarrow \bar{X}$

- Seja X v.a. c/ dist. Normal, i.e., $X \sim \mathcal{N}(\mu, \sigma)$
Para uma amostra de dimensão n tem-se $\bar{X}_n \sim \mathcal{N}(\mu, \sigma/\sqrt{n})$

Então . . .

Intervalos de confiança

- Seja X v.a. c/ dist. Normal , i.e., $X \sim \mathcal{N}(\mu, \sigma)$

Se σ conhecido \rightarrow considera-se $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$

Intervalo de confiança a $(1 - \alpha) \times 100\%$ para μ

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

($z_{\alpha/2} \rightarrow$ valor da v.a. Z tal que $P(Z > z_{\alpha/2}) = \alpha/2$)

Observações: Chama-se **precisão da estimativa** à semi-amplitude do intervalo de confiança e **confiança** ou **grau de confiança** a $(1 - \alpha) \times 100\%$

Quanto maior for o intervalo, maior é o grau de confiança, mas menor a precisão da estimativa.

Exercício 1

Um método utilizado na determinação do pH de uma dada solução fornece medições que se admite terem distribuição normal de valor médio igual ao verdadeiro valor do pH da solução e desvio padrão de 0.02.

Para avaliar o pH de uma solução, efectuaram-se 10 medições independentes tendo-se obtido os seguintes valores:

8.18 8.16 8.17 8.22 8.19 8.17 8.15 8.21 8.16 8.18

- Indique uma estimativa do valor médio do pH da solução.
- Com base nestas 10 medições, determine um intervalo a 95% de confiança para o valor médio do pH da solução.
- Para um certo processo químico é muito importante que uma dada solução tenha um pH de exactamente 8.20. Com base no resultado da alínea anterior, o que pode concluir relativamente à utilização desta solução no referido processo químico?

Distribuições por amostragem

A determinação de intervalos de confiança para os parâmetros μ , σ^2 , p , $\mu_1 - \mu_2$, σ_1^2/σ_2^2 e $p_1 - p_2$, necessita do conhecimento da distribuição dos estimadores envolvidos \longleftrightarrow **distribuições por amostragem**



São distribuições de funções da amostra aleatória (X_1, X_2, \dots, X_n) , que vamos usar para obter **Intervalos de Confiança**

Já vimos

Se $X \sim \mathcal{N}(\mu, \sigma)$ e σ conhecido usamos $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$ para construir um I.C. para o valor médio, μ .

Mas na maior parte das situações não conhecemos a variância.

Então neste caso já não é possível obter um I.C. para μ , pois os limites do intervalo dependem de σ , desconhecido.

Vamos então referir brevemente alguns resultados importantes, que nos permitem definir distribuições que precisamos de utilizar agora

Resultados (distribuições) importantes

Teorema

Sejam X_i ($i = 1, \dots, n$) v.a. independentes com distribuição $\mathcal{N}(\mu, \sigma)$.
Então:

- $\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_{(1)}^2$,

que se designa **distribuição qui-quadrado com 1 grau de liberdade**.

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

Aplicação – distribuição da variância numa população normal

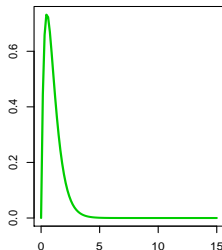
Sejam X_i ($i = 1, \dots, n$) v.a. independentes com distribuição $\mathcal{N}(\mu, \sigma)$.

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

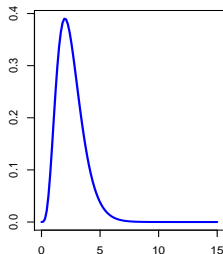
$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

Gráficos da função densidade de uma v.a. com distribuição χ^2

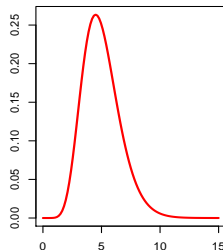
f. densidade da Qui-quadrado(n=4)



f. densidade da Qui-quadrado(n=10)



f. densidade da Qui-quadrado(n=20)



Gráficos da função densidade de uma v.a. com distribuição $\chi^2_{(4)}$, $\chi^2_{(10)}$ e $\chi^2_{(20)}$, da esquerda para a direita.

Resultados importantes (cont.)

- Sejam $Z \sim \mathcal{N}(0, 1)$ e $Y \sim \chi^2_{(n)}$ v.a. independentes, então

$$\frac{Z}{\sqrt{Y/n}} \sim t_{(n)},$$

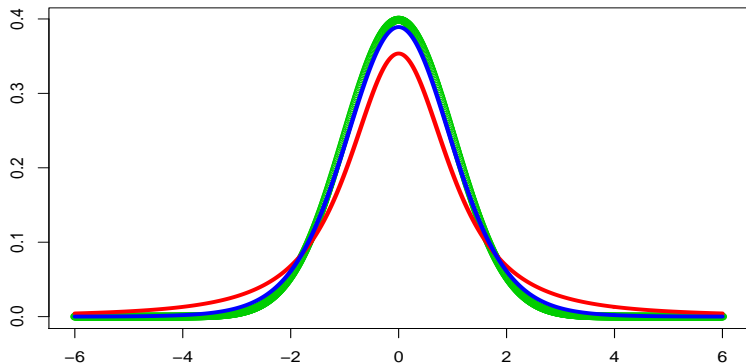
que se designa **distribuição t – Student** com n g.l.

Para mais informações ver quadro das distribuições contínuas.

Aplicação – distribuição da média numa população normal com σ desconhecido

Sejam X_i ($i = 1, \dots, n$) v.a. independentes com distribuição $\mathcal{N}(\mu, \sigma)$.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}, \quad \text{com } S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$



Gráficos da função densidade de uma v.a. com distribuição $\mathcal{N}(0, 1)$, $t_{(2)}$ e $t_{(10)}$.

Intervalos de confiança (cont.)

O Intervalo a $(1 - \alpha) \times 100\%$ de confiança para σ^2 numa população normal, constrói-se usando a v.a. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2, (n-1)}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2, (n-1)}}$$

O Intervalo a $(1 - \alpha) \times 100\%$ de confiança para μ com σ desconhecido, numa população normal, constrói-se usando a v.a.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

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

Intervalos de confiança (cont.)

Intervalo de confiança a $(1 - \alpha) \times 100\%$ para μ

Se X tem **distribuição qualquer**

É necessário dispor de uma **amostra de dimensão elevada**, i.e., **n grande** \rightarrow **aplicação do Teorema Limite Central**

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \quad \text{se } \underline{\sigma \text{ conhecido}}$$

Ou, que é o caso mais frequente,

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim \mathcal{N}(0, 1) \quad \text{se } \underline{\sigma \text{ desconhecido}}$$

$$\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Intervalo de confiança para proporções

Dada X com **distribuição binomial** de parâmetros (n, p) (p parâmetro desconhecido).

Um **estimador** de p é $\hat{P} = \frac{X}{n}$ (X é o número de sucessos em n provas)

Sendo a **estimativa** $\hat{p} = \frac{x}{n}$ (x é o número observado de sucessos em n provas)

Se $X \sim B(n, p)$ e n grande $\frac{X - np}{\sqrt{npq}} \sim \mathcal{N}(0, 1)$

Intervalo de confiança a $(1 - \alpha) \times 100\%$ para p

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} < p < \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Intervalos de confiança - duas populações

Intervalos de confiança para $\mu_1 - \mu_2$ (amostras independentes)

Caso de duas populações normais i.e.,

$X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ e $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ com variâncias conhecidas

Retiram-se duas amostras aleatórias independentes com dimensões n_1 e n_2 , respectivamente.

Um estimador para $\mu_1 - \mu_2$ é $\bar{X}_1 - \bar{X}_2$ e

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

Intervalo de confiança a $(1 - \alpha) \times 100\%$ para $\mu_1 - \mu_2$

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Intervalos de confiança - duas populações

Caso de populações normais i.e., $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ e $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ com **variâncias desconhecidas**.

Amostras independentes:

- Se n_1, n_2 **grandes** (e neste caso não é necessário ter-se **normalidade**) pode substituir-se no intervalo anterior σ_1^2 por s_1^2 e σ_2^2 por s_2^2 .
- Se n_1 e n_2 **são pequenos**, podemos construir intervalos de confiança sob **certas hipóteses**
 - ambas as distribuições normais;
 - se for possível admitir variâncias populacionais iguais.

Então ...

Intervalos de confiança - duas populações

A variável aleatória agora usada é

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}$$

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

Intervalo de confiança a $(1 - \alpha) \times 100\%$ para $\mu_1 - \mu_2$ para duas populações normais, com variâncias desconhecidas mas supostas iguais

$$(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(t_{\alpha/2} \equiv t_{\alpha/2, (n_1+n_2-2)})$$

Quociente de duas variâncias

O que pode fazer-se para averiguar a igualdade das variâncias populacionais?

Sejam duas populações normais $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ e $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ das quais retiramos duas amostras independentes de dimensões n_1 e n_2 , respectivamente.

O estimador para $\frac{\sigma_1^2}{\sigma_2^2}$ é $\frac{S_1^2}{S_2^2}$ e para construir o intervalo de confiança

a variável aleatória agora considerar é $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$. Surge aqui uma nova distribuição associada a esta variável aleatória, que se chama **distribuição F** (de Snedecor).

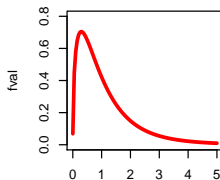
Veamos a génese (a definição) desta distribuição:

A distribuição F

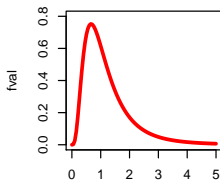
Teorema

Sejam $U \sim \chi^2_{(m)}$ e $V \sim \chi^2_{(n)}$ variáveis aleatórias independentes, a variável $X = \frac{U/m}{V/n}$ diz-se ter **distribuição F** com **(m, n) graus de liberdade** e representa-se por $F_{(m,n)}$.

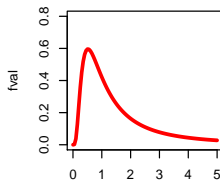
f. densidade da $F_{(3,15)}$



f. densidade da $F_{(10,10)}$



f. densidade da $F_{(15,3)}$



Gráficos da função densidade de uma v.a. com distribuição $F_{(3,15)}$, $F_{(10,10)}$ e $F_{(15,3)}$, da esquerda para a direita.

Resultados com a distribuição F

Prova-se que $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ tem distribuição F com $(n_1 - 1, n_2 - 1)$ graus de liberdade.


Intervalo de confiança a $(1 - \alpha) \times 100\%$ para $\frac{\sigma_1^2}{\sigma_2^2}$

$$\frac{s_1^2}{s_2^2 f_{\alpha/2; (n_1-1, n_2-1)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2 f_{\alpha/2; (n_2-1, n_1-1)}}{s_2^2}$$

Nota:

No caso de o intervalo de confiança para $\frac{\sigma_1^2}{\sigma_2^2}$, conduzir à conclusão que $\sigma_1^2 \neq \sigma_2^2$, Welch-Satterthwaite propõem o uso da variável aleatória

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx t_{(\nu)},$$

sendo ν calculado à custa das variâncias de cada uma da amostra — não mostraremos aqui a sua expressão. O uso deste estimador está já implementado no  - **ver a sua utilização nas aulas práticas.**

Intervalos de confiança (amostras emparelhadas)

Consideremos que numa dada experiência as observações estão relacionadas, i.e, **emparelhadas** pelo indivíduo.

Seja (X_i, Y_i) ($i = 1, \dots, n$) a **amostra emparelhada** e defina-se $D_i = X_i - Y_i$, isto é, (D_1, D_2, \dots, D_n) é a **amostra aleatória das diferenças**.

Se D_1, D_2, \dots, D_n são variáveis aleatórias provenientes de uma lei **normal** com valor médio $\mu_D = \mu_X - \mu_Y$ e variância σ_D^2 , **desconhecida** tem-se

$$\frac{\bar{D} - \mu_D}{S_D / \sqrt{n}} \sim t_{(n-1)}$$

Intervalos de confiança (cont.)

Intervalo de confiança a $(1 - \alpha) \times 100\%$ para μ_D

$$\bar{d} - t_{\alpha/2, (n-1)} \frac{s_D}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2, (n-1)} \frac{s_D}{\sqrt{n}}$$

Se não for possível admitir D_i normais, mas se tenha n “grande” o intervalo de confiança $(1 - \alpha) \times 100\%$ para μ_D é

$$\bar{d} - z_{\alpha/2} \frac{s_D}{\sqrt{n}} < \mu_D < \bar{d} + z_{\alpha/2} \frac{s_D}{\sqrt{n}}$$

Intervalos de confiança para proporções

Consideremos agora o caso de duas proporções

Sejam X_1 e X_2 variáveis aleatórias tais que

$$X_1 \sim B(n_1, p_1) \quad \text{e} \quad X_2 \sim B(n_2, p_2).$$

n_1 e n_2 dimensões de amostras aleatórias independentes

Os estimadores de p_1 e p_2 são $\hat{P}_1 = X_1/n_1$ e $\hat{P}_2 = X_2/n_2$

Se n_1 e n_2 “grandes” temos

$$\hat{P}_1 - \hat{P}_2 \sim \mathcal{N} \left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \right)$$

Intervalo de confiança a $(1 - \alpha) \times 100\%$ para $p_1 - p_2$ quando as dimensões das amostras são elevadas

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Exercício 2

Um estudo pretende comparar um tipo de semente melhorada com o tipo de semente tradicional usado anteriormente. A semente melhorada passará a ser utilizada se, em média, o crescimento das plantas após 20 dias fôr superior ao verificado nas sementes tradicionais. São criadas 15 diferentes situações laboratoriais, variando temperatura e humidade. Em cada situação semeia-se uma semente de cada tipo e obtêm-se os seguintes resultados para o crescimento (em cms) das plantas após 20 dias:

Situação	1	2	3	4	5	6	7	8
sementes melhoradas	3.46	3.48	2.74	2.83	4.00	4.95	2.24	6.92
sementes tradicionais	3.18	3.67	2.92	3.10	4.10	4.86	2.21	6.91

Situação	9	10	11	12	13	14	15
sementes melhoradas	6.57	6.18	8.30	3.44	4.47	7.59	3.87
sementes tradicionais	6.83	6.19	8.05	3.46	4.18	7.43	3.85

Poderá considerar-se haver diferença significativa entre os dois tipos de sementes? Responda justificando e explicitando quaisquer hipóteses adicionais que seja necessário impôr.

Exercício 3

Pretende-se avaliar se um certo adubo A aumenta a produção de determinada cultivar do cereal T . Para tal efeito um experimentador plantou 2 talhões com a referida cultivar, tendo aplicado o adubo A só num deles. De cada talhão foram então amostradas 12 áreas de 1m^2 . Em cada uma destas áreas foram colhidas todas as plantas e pesado o grão. Os dados obtidos, expressos em gramas foram os seguintes:

Talhão c/ adubo	422	460	455	466	475	472	465	456	452	430	458	470
Talhão s/ adubo	470	437	429	447	432	457	422	425	432	474	452	442

O que decidiria quanto à utilização do adubo?

Introdução aos Testes de Hipóteses

- Hipóteses estatísticas. Hipótese nula/Hipótese alternativa.
- Passos necessários à elaboração de um teste de hipóteses.
- Testes sobre parâmetros de uma ou de duas populações.
- Teste de normalidade e “teste do qui-quadrado” de ajustamento.

Testes de Hipóteses

Até aqui estivemos a tratar de **Intervalos de Confiança**, isto é, a construir intervalos de valores que, com uma confiança elevada, contêm o verdadeiro valor do parâmetro desconhecido.

Mas ... vamos agora ver uma outra forma de decidir se uma dada suposição (relativa ao(s) parâmetro(s) desconhecido(s) ou outros aspectos da característica em estudo) **é** ou **não** suportada pela informação fornecida pelos dados de uma amostra.

Vamos então **formular Testes de Hipóteses** e indicar como, **com base numa amostra observada**, se pode tomar decisões relativamente às hipóteses formuladas.

Testes de Hipóteses

Note-se que os quadros com Intervalos de confiança e os quadros dos Testes de Hipóteses seguem a mesma estrutura, portanto as variáveis usadas na construção dos intervalos de confiança são as que iremos também aqui usar.

Uma hipótese estatística é qualquer conjectura sobre aspectos desconhecidos da população (que podem ser parâmetros ou mesmo a forma da distribuição).

Se a hipótese diz respeito a:

- **um parâmetro**, supondo conhecida a forma da distribuição, a hipótese diz-se **paramétrica**.
- **investigar a forma da distribuição**, ou **um parâmetro** sem admitir o conhecimento da forma da distribuição, a hipótese diz-se **não paramétrica**.

Testes de Hipóteses

Um teste de hipóteses é então um procedimento que permite decidir se uma dada hipótese **é** ou **não** suportada pela informação fornecida pelos dados de uma amostra.

Um teste diz-se paramétrico se as **hipóteses** envolvidas são **paramétricas**, i.e., dizem respeito ao(s) parâmetro(s), supondo conhecida, pelo menos aproximadamente, a forma da distribuição.

Metodologia a seguir num teste de hipóteses:

- Como uma dada afirmação pode ser verdadeira ou falsa, formulam-se duas hipóteses alternativas
- **Hipótese nula** - H_0 - corresponde ao *statu quo*, i.e., é a hipótese que especifica o valor “actual” do(s) parâmetro(s) (é considerada verdadeira até haver evidência estatística para a sua rejeição) e
- **Hipótese alternativa** - H_1 - onde se especificam o(s) valor(es) a “aceitar” quando se rejeita a hipóteses nula.

Testes de Hipóteses- Exemplos

1. $H_0 : \mu = 1$ v.s. $H_1 : \mu \neq 1$
2. $H_0 : \mu = 3$ v.s. $H_1 : \mu < 3$
3. $H_0 : \mu = 1$ v.s. $H_1 : \mu > 1$
4. $H_0 : \mu \geq \mu_0$ v.s. $H_1 : \mu < \mu_0$
5. $H_0 : X \sim \text{Normal}$ v.s. $H_1 : X$ segue outra
distribuição

O último exemplo é um teste para verificar se uma dada característica segue uma distribuição (neste caso a normal).

Designa-se habitualmente por **teste de normalidade**.

Exemplo 1

Numa linha de engarrafamento de azeite a quantidade deitada em cada garrafa é uma variável aleatória que se admite ter distribuição normal. O processo de enchimento considera-se regulado se $\mu = 1$ litro, não sendo de admitir grandes desvios. Para controlar o processo de enchimento escolheram-se ao acaso 20 garrafas da produção diária. Suponha que se obteve uma média de 0.965 litros com um desvio padrão de 0.08 litros.

Poder-se-á dizer que o processo não está regulado? Justifique convenientemente a resposta.

Resolução: Para averiguar se o processo não está regulado, podemos formular um teste das hipóteses:

$$H_0 : \mu = 1$$

v.s.

$$H_1 : \mu \neq 1$$

Testes de Hipóteses

- A resposta num teste de hipóteses é dada na forma
 - **Rejeitar H_0** - significa que os dados observados testemunham fortemente **contra H_0** - neste caso será **adoptada a hipótese H_1** ou
 - **Não rejeitar H_0** - significa que **não há evidência suficiente para rejeitar H_0** .

Tomar decisões para a população com base numa amostra possui riscos (i.e. pode cometer-se erros). Vejamos o quadro:

	Decisões	
Realidade	Rejeitar H_0	Não rejeitar H_0
H_0 verdadeira	ERRO de tipo I	não há erro
H_0 falsa	não há erro	ERRO de tipo II

Testes de Hipóteses Paramétricos

$P(\text{erro de tipo I ou } 1^{\text{a}} \text{ espécie}) = P(\text{Rejeitar } H_0 | H_0 \text{ verdadeira}) = \alpha$
 α - nível de significância

$P(\text{erro de tipo II ou } 2^{\text{a}} \text{ espécie}) = P(\text{Não rejeitar } H_0 | H_0 \text{ falsa}) = \beta$

Costuma atribuir-se um valor muito baixo à probabilidade do erro de 1^{a} espécie, i.e, habitualmente considera-se

$\alpha = P(\text{Rejeitar } H_0 | H_0 \text{ verdadeira}) = 0.05$ ou 0.01

Rejeitar H_0 significa que os dados testemunham fortemente contra H_0

Testes de Hipóteses

Passos a seguir na construção de um **teste estatístico**:

1. Identificar o(s) parâmetro(s); especificar H_0 e H_1 e o nível de significância α .
2. Escolher uma variável aleatória – **estatística de teste**, que sob H_0 terá distribuição conhecida (pelo menos aproximadamente).
3. Definir a **região de rejeição** ou **região crítica** – **RC** (conjunto de valores da estatística que são menos “plausíveis” caso H_0 seja verdadeira), portanto levam a **rejeitar H_0** .
4. Calcular **o valor** da estatística de teste, para a amostra observada.
5. **Se o valor calculado \in RC \rightarrow rejeita-se H_0**
Se o valor calculado \notin RC \rightarrow não se rejeita H_0

Testes de Hipóteses

Iremos considerar testes para os parâmetros: valor médio, variância, diferença de valores médios no caso de amostras independentes e emparelhadas, quociente de variâncias e proporções.

Para todos estes testes os passos indicados no slide anterior encontram-se explicitados no quadro “Testes a Médias, Variâncias e Proporções”.

Exemplo: de formulação de hipóteses para o valor médio:

$H_0 : \mu = \mu_0$	v.s.	$H_1 : \mu \neq \mu_0$	teste bilateral
$H_0 : \mu \leq \mu_0$	v.s.	$H_1 : \mu > \mu_0$	teste unilateral (à direita)
$H_0 : \mu \geq \mu_0$	v.s.	$H_1 : \mu < \mu_0$	teste unilateral (à esquerda)

Regras de decisão em Testes de Hipóteses

A indicação do valor observado da estatística de teste, por exemplo Z_{cal} , e a indicação de um valor crítico z_{α} para decidir, por exemplo,

Rejeitar H_0 se $Z_{cal} > z_{\alpha}$ (num teste unilateral à direita) tem sido recentemente “substituído” pelo cálculo de

– a probabilidade de se observar um valor igual ou mais extremo do que o observado, se a hipótese nula é verdadeira – a que se chama **valor de prova; valor- p (p -value)**

Nota: é esta quantidade que hoje em dia qualquer *software* está preparado para calcular quando se manda realizar um teste.

Regras de decisão em Testes de Hipóteses

Podemos interpretar o **valor de prova** ou **valor- p** ou **p -value** como a **medida do grau de concordância entre os dados e H_0**

Assim:

- **Quanto menor for o p -value, menor é a consistência entre os dados e a hipótese nula**

Habitualmente adopta-se como regra de decisão:

rejeitar H_0 se p -value $\leq \alpha$

Testes de Hipóteses ... outros

Nesta breve introdução à Teoria dos Testes de Hipóteses vimos como formular testes a parâmetros desconhecidos da população, em particular:

- ao valor médio, variância e a uma proporção;
- diferença de dois valores médios em amostras independentes e em amostras emparelhadas;
- quociente de variâncias e diferença de duas proporções.

Mas ... outros testes de hipóteses podem ser necessários nas mais variadas situações.

Note-se, todavia, que os passos a seguir serão sempre os indicados no slide 202.

Como ilustração podemos referir três testes:

- **Teste de normalidade de Shapiro-Wilk**;
- **Teste do Qui-quadrado de ajustamento**;
- **Teste à igualdade de três ou mais valores médios em amostras independentes** (este é o procedimento estatístico conhecido como **Análise de Variância** que na sua ideia mais básica corresponde à generalização do problema da comparação de duas médias populacionais.)

O último ponto não será tratado neste curso.

Um teste de Normalidade

Consideremos agora o caso de se pretender **testar a forma da distribuição**.

Os testes a considerar agora são testes não paramétricos que se designam por **testes de ajustamento**.

Começemos com um teste muito importante nas nossas aplicações - **um teste de ajustamento à distribuição normal** - para averiguar se um dado conjunto de observações se pode considerar proveniente de uma população com distribuição normal – é um **teste de normalidade**,

O Teste de Shapiro Wilk

que se tem revelado ser um dos mais potentes. Vejamos em síntese como se processa.

O teste de Shapiro-Wilk

Seja X a característica em estudo na população.

Formulam-se as hipóteses:

H_0 : X tem distribuição Normal

H_1 : X não tem distribuição Normal

Calcula-se o valor da estatística de teste

$$W_{cal} = \frac{b^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

com b constante a determinar a partir dos dados e com recurso a uma tabela (que não será aqui dada).

O teste de Shapiro-Wilk

Valores pequenos de W_{cal} indicam não normalidade, i.e.

$$\text{RC: } W_{cal} < W_{\alpha}$$

Com W_{α} – **valor crítico** a consultar numa Tabela.

Alternativamente pode usar-se o **valor- p** ,

- **Rejeita-se H_0** se **valor- $p \leq \alpha$** , significando que não se pode admitir que X tem distribuição normal;
- Se **valor- $p > \alpha$** não se rejeita **H_0** , o que significa que a distribuição normal é uma distribuição possível para X .

Exercício 2 (reescrito)

Um estudo pretende comparar um tipo de semente melhorada com o tipo de semente tradicional. A semente melhorada passará a ser utilizada se, em média, o crescimento das plantas após 20 dias for superior ao das obtidas das sementes tradicionais. São criadas 15 diferentes situações laboratoriais, variando temperatura e humidade. Em cada situação semeia-se uma semente de cada tipo e obtêm-se os seguintes resultados para o crescimento (em cm) das plantas após 20 dias :

Situação	1	2	3	4	5	6	7	8
Semente melhorada	3.46	3.48	2.74	2.83	4.00	4.95	2.24	6.92
Semente tradicional	3.18	3.67	2.92	3.10	4.10	4.86	2.21	6.91
Situação	9	10	11	12	13	14	15	
Semente melhorada	6.57	6.18	8.30	3.44	4.47	7.59	3.87	
Semente tradicional	6.83	6.19	8.05	3.46	4.18	7.43	3.85	


Deverá passar a usar-se as sementes melhoradas? Responda justificando e explicitando quaisquer hipóteses adicionais que seja necessário impor.

Resolução - Exercício 2

Trata-se de um problema de **amostras emparelhadas** e como se pretende averiguar se as sementes melhoradas apresentam melhor desempenho, o teste a realizar é o teste unilateral

$$H_0 : \mu_D = 0 \quad \text{vs.} \quad H_1 : \mu_D > 0$$

D designa a diferença entre o comprimento das plantas das sementes melhoradas e das tradicionais.

O comando do  para realizar este teste (e da execução do qual se obtém tb o intervalo de confiança unilateral, que não fez parte do programa) é o seguinte:

```
> t.test(nsem, vsem, alternative='greater'), paired=TRUE)
```

Resolução - Exercício 2 (cont.)

Mas ... para utilizar aquele teste é necessário verificar previamente se pode admitir-se que a **diferença entre alturas das plantas provenientes das sementes melhoradas e das tradicionais segue uma distribuição normal**.

Começamos pelo teste de normalidade de Shapiro-Wilk para analisar a hipótese de normalidade:

```
> shapiro.test(nsem-vsem)
Shapiro-Wilk normality test
data:  nsem-vsem
W = 0.9428, p-value = 0.4184
```

Conclusão: Dado que o *p-value* é **0.4184**, conclui-se que **não se rejeita a hipótese da normalidade**, pelo que pode realizar-se o teste *t*.

Resolução - Exercício 2 (cont.)

```
> t.test(nsem,vsem,alternative='greater', paired=TRUE)
```

Paired t-test

data: nsem and vsem

t = 0.1396, df = 14, p-value = 0.4455

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-0.07744645 Inf

sample estimates: mean of the differences

0.006666667

Como o *p-value* = 0.4455 é superior a 0.05, **não se rejeita H_0** .

Conclusão: não se pode afirmar que μ_D seja superior a 0, portanto não se pode aconselhar o uso das sementes melhoradas face às tradicionais.

Teste de ajustamento do qui-quadrado—introdução

Falámos já num teste de ajustamento - o teste Shapiro - Wilk - **teste de normalidade**.

Mas, ... a necessidade de considerar **o ajustamento a outras distribuições** (contínuas ou discretas) pode surgir nas mais variadas aplicações.

O “teste do qui-quadrado” foi desenvolvido por Pearson (1900) como **teste de ajustamento** para **dados nominais** ou **dados classificados em classes**, i.e., é um teste que usa **a contagem do número de valores que se observam em classes** ou **categorias**.

Vamos introduzir o **teste do qui-quadrado** com um **Exemplo**, lembrando as etapas necessárias à realização de um teste de hipóteses - slide 202.

Teste de ajustamento do qui-quadrado—introdução

Exemplo

A descendência originada pelo cruzamento de dois dados tipos de plantas pode ser qualquer um dos três genótipos que representaremos por c , cC e C . Um modelo teórico de sucessão genética indica que os tipos c , cC e C devem aparecer na razão 1 : 2 : 1. Efectuou-se o cruzamento daqueles dois tipos tendo-se classificado 90 plantas. A sua classificação genética foi registada na tabela:

Genótipos	c	cC	C
	18	44	28

Estão estes dados de acordo com o modelo genético?

O que se pretende aqui averiguar é se **X segue uma dada distribuição** ou **não**

... no caso do exemplo, aquela averiguação inicia-se pela **formulação das seguintes hipóteses:**

Teste de ajustamento do qui-quadrado—introdução

$H_0 : p_1 = 0.25, \quad p_2 = 0.5, \quad p_3 = 0.25$

H_1 : pelo menos uma das probabilidades é diferente do formulado.

O teste sugerido por Karl Pearson (para o caso de as observações se encontrarem classificadas em k classes) consiste em considerar a estatística de teste:

$$X^2 = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i}$$

k — nº de classes; O_i — frequência observada, é a frequência absoluta em cada classe; e_i — frequência esperada, dada por $e_i = n p_i$ com p_i a probabilidade da classe i , se a hipótese H_0 verdadeira; n é o número (total) de observações independentes.

O valor de X^2 em cada classe é uma medida da proximidade entre os dados e a Hipótese nula. Quanto menor for o valor de X^2 mais plausível é a hipótese H_0 .

Teste de ajustamento do qui-quadrado—introdução

Pearson mostrou que X^2 tem distribuição assintótica

$X^2 \sim \chi^2_{(k-1)}$ onde k é o número de classes em que as observações foram agrupadas.

A região crítica, para um nível de significância α , é definida como:

$$\text{RC: } X_{cal}^2 > \chi_{\alpha, (k-1)}^2$$

Resolução:

	c	cC	C
O_i	18	44	28
p_i	0.25	0.5	0.25
$E_i = np_i$	22.5	45	22.5

$$X_{cal}^2 = \frac{(18-22.5)^2}{22.5} + \frac{(44-45)^2}{45} + \frac{(28-22.5)^2}{22.5} = 2.27 \quad \text{e} \quad \chi_{0.05, (2)}^2 = 5.99.$$

Então $X_{cal}^2 < \chi_{0.05, (2)}^2$, logo não se rejeita H_0 a um nível de significância de 5%, portanto não há razões para duvidar que os dados estejam de acordo com o modelo genético.

Teste do qui-quadrado - Observações

- A variável X^2 tem distribuição assintótica. Qual a dimensão que a amostra deverá ter para que a aproximação à distribuição $\chi^2_{(k-1)}$ seja válida?
 - Alguns autores consideram que a aproximação só é válida se a **frequência esperada, e_i , verifica $e_i \geq 5$** ;
 - Cochran (1954) sugeriu que se podia considerar classes com $e_i = 1$, desde que **80% das classes apresente $e_i > 5$** . Sempre que as frequências esperadas de algumas classes forem inferiores a 1 essas classes devem agrupar-se com as adjacentes por forma a atingir a frequência mínima desejada.
- Sempre que para determinar a probabilidade p_i for necessário estimar parâmetros, a distribuição χ^2 virá com

($k - 1 - n^\circ$ de parâmetros estimados) graus de liberdade

**Votos de que os resultados correspondam às vossas expectativas e trabalho realizado ao longo do semestre.
BOA SORTE!!**

**BOAS FESTAS
Dez/2018**