Exercícios de Inferência Estatística

Apresenta-se a resolução de alguns exercícios enunciados nas folhas de prática - Capítulo 3. Tem-se ainda propostas de novos exercícios e respectivas resoluções. Os últimos são aplicação do teste de ajustamento do qui-quadrado

1. **(Ex. 3.27) Resolução**

   Trata-se de amostras emparelhadas (as medições são feitas nos mesmos pontos do rio).

   Considere-se as variáveis X - medições antes da campanha e Y - medições depois da campanha

   Sendo amostras emparelhadas o estudo é feito sobre a amostra aleatória das diferenças $D_i = X_i - Y_i$. Como se pretende saber se a campanha reduziu a poluição deve fazer-se um **teste à média das diferenças,** μ_D , isto é pretende-se averiguar se pode considerar-se $\mu_D > 0$.

   Como $n = 10$ é pequeno e a variância σ_D^2 é desconhecida é necessária a verificação do pressuposto da normalidade da v.a. D .

   Comecemos então pelo <u>estudo da normalidade</u> - vamos aplicar o **Teste de Shapiro-Wilk**. As hipóteses a formular são

   $H_0 : X \frown$ Normal $\qquad H_1 : X \not\frown$ Normal

   Veja-se de seguida a resolução do exercício no R

```
> x<-c(68,88,101,82,96,74,65,74,52,99)
> y<-c(67,87,90,76,98,69,68,65,59,70)
> d<-x-y;d
 [1]  1  1 11  6 -2  5 -3  9 -7 29
> shapiro.test(d)

        Shapiro-Wilk normality test

data:  d
W = 0.8781, p-value = 0.1240
```

   Tem-se $p - value > 0.05$, portanto não se rejeita H_0, isto é não podemos rejeitar a hipótese da normalidade da v.a. D.

   Vejamos agora os passos na realização do teste para verificarmos se a campanha reduziu de facto a poluição:

   1. Hipóteses: $H_0 : \mu_D \leq 0 \qquad H_1 : \mu_D > 0$
   $\alpha = 0.05$

2. **Estatística de teste:** $T = \frac{\bar{D}-\mu_D}{S_D/\sqrt{n}} \frown t_{(n-1)}$, se a hipóteses H_0 verdadeira.
3. **Região Crítica:** Rejeita-se H_0 se $T_{calc} > t_{\alpha,(n-1)} = t_{0.05,(9)}$
   $R.C.: \quad T_{calc} > 1.8333$
4. **Cálculo da Estatística de teste:**
   Tem-se $\quad \bar{d} = 5$ e $s_d = 10.10$
   $T_{calc} = \frac{\bar{d}-0}{s_d/\sqrt{n}} = 1.5655,$
5. **Conclusão:** Como $T_{calc} \notin R.C$ não rejeitamos H_0 a um nível de significância de 5%, isto é, não há motivos para acreditar que a campanha tenha reduzido a poluição.

Também podemos resolver este exercício no R, para as hipóteses formuladas em **1.**, bastando usar o comando:

```
> t.test(x,y,paired=T,alternative="greater")

        Paired t-test

data:  x and y
t = 1.5656, df = 9, p-value = 0.07595
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.8544932        Inf
sample estimates:
mean of the differences
                      5
```

As etapas **2, 3 e 4** apresentadas atrás são substituídas pelo cálculo de $p-value$. Vemos que $p-value = 0.07595$, portanto superior ao valor habitual de $\alpha = 0.05$, logo não rejeitamos H_0, então não há motivos para acreditar que a campanha tenha reduzido a poluição.

2. **(Ex. 3.30)** A resolução deste exercício vai ser apoiada no *output* do R apresentado. Vamos ver como o podemos usar.

   a) Seja X_1 - v.a. que representa a acidez total do vinho, determinada pelo método clássico; pretende-se averiguar se em média a acidez total é superior a 3.5 g/l, i.e., sendo $\mu_1 = E[X_1]$, devemos realizar o teste de hipóteses:
   H$_0$: $\mu_1 \leq 3.5$ vs. H$_1$: $\mu_1 > 3.5$.
   Observe-se que estamos em presença de uma amostra de dimensão pequena com variância populacional desconhecida, por isso (validados certos pressupostos) devemos usar o teste apresentado no comando do R
   `t.test(classico,mu=3.5,alternative="greater")`.
   Os pressupostos para a realização deste teste são:
   - garrafas seleccionadas aleatoriamente,
   - X_1 ter distribuição normal.

Vamos verificar a hipótese da normalidade
H_0: X_1 tem distribuição normal vs. H_1: X_1 não tem distribuição normal
usando o teste de Shapiro-Wilk.

No R o teste é efectuado com o comando `shapiro.test(classico)`.

Vemos que *p-value*$= 0.3827$, que é um valor elevado, pelo que não se rejeita a hipótese da normalidade.

Passemos agora à análise do resultado da aplicação do comando
`t.test(classico,mu=3.5,alternative="greater")`.

Tem-se *p-value*$= 0.1255$, que é um valor superior ao nível de significância α=0.05 (habitualmente considerado). Logo não se rejeita a hipótese nula e portanto não se pode concluir que o vinho cumpra os requisitos ($\mu_1 > 3.5$).

b) X_2 - v.a. que representa a acidez total do vinho, determinada pelo método automático, $\mu_2 = E[X_2]$.

Trata-se de um delineamento em amostras emparelhadas por garrafa.

A resposta à questão pode ser obtida com recurso a um I.C. para $\mu_D = \mu_1 - \mu_2$ ou ao teste de hipóteses: H_0: $\mu_D = \mu_1 - \mu_2 = 0$ vs. H_1: $\mu_D = \mu_1 - \mu_2 \neq 0$, mas note-se que, sendo amostras emparelhadas, das opções apresentadas no *output* para comparar dois valores médios **só** interessa a que contém a informação `paired=TRUE`.

Os pressupostos são:
- garrafas seleccionadas aleatoriamente,
- a população das diferenças $D = X_1 - X_2$ ter distribuição normal.

A hipótese da normalidade de D é validada com recurso ao teste de Shapiro-Wilk
`shapiro.test(classico-automatico)`.

Como *p-value*$=0.08413$ e é um valor superior a 0.05(habitualmente considerado) não se rejeita a hipótese da normalidade.

Estamos ainda na situação de a variância de D não ser conhecida, então o teste a usar é `t.test(classico,automatico,paired=TRUE)`.

Tem *p-value*$=0.03493$ pelo que, aos níveis de significância usuais de 0.05 ou 0.1, se rejeita a hipótese nula concluindo-se que os dois métodos de análise conduzem a resultados diferentes.

Note-se que o I.C. que também é dado no *output* $-1.12405128 < \mu_1 - \mu_2 < -0.04594872$, nos permite concluir que com uma confiança de 95% a média das diferenças é negativa, portanto os dois métodos conduzem a resultados diferentes.

3. Mediram-se 5 espigas de trigo da variedade A e 7 espigas da variedade B. Os resultados obtidos para os comprimentos das espigas em cm, foram os seguintes:

A	12.6	13.4	11.9	12.8	13.0		
B	13.1	13.4	12.8	13.5	13.3	12.7	12.4

Pretende-se verificar se os dados obtidos permitem estabelecer qualquer diferença significativa de comprimento das espigas entre as duas variedades de trigo.

Resolução

Como é evidente, não há qualquer emparelhamento dos dados; as medições de A são independentes das medições de B.

Como as amostras são de dimensão pequena, $n_1 = 5, n_2 = 7$, teremos de verificar a validade dos seguintes pressupostos:

i) normalidade de cada uma das variáveis:
X- comprimento das espigas da variedade A e
Y- comprimento das espigas da variedade B

ii) igualdade das variâncias: $\sigma_1^2 = \sigma_2^2$?

Verificando-se a validade destes pressupostos, o estudo da existência de diferença de comprimento entre os dois tipos de espigas pode ser feito considerando, por exemplo, um intervalo de confiança para

$$\mu_1 - \mu_2$$

com μ_1 - comprimento médio das espigas da variedade A e
μ_2 - comprimento médio das espigas da variedade B

- Vejamos se o pressuposto i) é verificado. Para isso vamos recorrer ao R

```
> x<-c(12.6,13.4,11.9,12.8,13.0)
> y<-c(13.1,13.4,12.8,13.5,13.3,12.7,12.4)
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.9706, p-value = 0.879

> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.9426, p-value = 0.6622
```

  Como qualquer dos valores de $p - value$ é superior ao valor 0.05 (habitualmente considerado), não se rejeita a hipótese da normalidade de cada uma das variáveis X e Y.

- Passemos agora à verificação do pressuposto ii). Podemos calcular o intervalo de confiança para σ_1^2/σ_2^2.
  O intervalo a $(1-\alpha) \times 100\%$ de confiança é dado por:

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

onde $s_1^2 = 0.308$ e $s_2^2 = 0.1657$
O comando no R, para obter o I.C. a 95% é:

```
> var.test(x,y)

        F test to compare two variances

data:  x and y
F = 1.8586, num df = 4, denom df = 6, p-value = 0.4739
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.29847 17.09431
sample estimates:
ratio of variances
          1.858621
```

Tem-se então o I.C. $]0.29847; 17.09431[$.
Como este intervalo contém o valor **1**, não temos motivos para duvidar que $\sigma_1^2 = \sigma_2^2$, com uma confiança de 95% .

Vamos então agora calcular o I.C. a 95% para $\mu_1 - \mu_2$, que é dado por

$$(\overline{x} - \overline{y}) - t_{0.025;(n_1+n_2-2)}\ s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\overline{x} - \overline{y}) + t_{0.025;(n_1+n_2-2)}\ s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

com $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$.
Tem-se $n_1 = 5; n_2 = 7; t_{0.025;(10)} = 2.22814; \bar{x} = 12.74; \bar{y} = 13.03$ e $s_p^2 = 0.2262$, donde o I.C. a 95% é então

$$-0.90416 < \mu_1 - \mu_2 < 0.32702$$

Portanto como este I.C. contém o valor **0** (que significa a igualdade dos dois valores médios) não podemos dizer que haja diferença no comprimento médio de cada uma das variedades, com uma confiança de 95%. A resposta a esta questão é obtida no R, executando o comando

```
> t.test(x,y,paired=F,var.equal=T)

        Two Sample t-test

data:  x and y
t = -1.0445, df = 10, p-value = 0.3208
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9041582  0.3270154
sample estimates:
```

```
mean of x mean of y
 12.74000  13.02857
```

Podemos dar a resposta usando o *p-value* ou o intervalo de confiança.

Vejamos agora dois exercícios para ilustrar a utilização do teste de ajustamento do qui-quadrado

4. Realizou-se um estudo para averiguar se o número de faltas ao trabalho dos empregados de uma empresa se podia considerar ter uma distribuição uniforme pelos dias da semana. O registo efectuado ao longo de várias semanas está resumido no seguinte quadro:

Segunda	Terça	Quarta	Quinta	Sexta
24	18	15	21	28

Resolução: Dizer que a distribuição do nº de faltas, X, é uniforme na semana, significa que $P[X=i]=1/5$, com $i=1,\ldots,5$, designando os dias da semana de Segunda a Sexta. Pretendemos então testar

$H_0: p_i = 1/5,\ i=1,\ldots 5$ *vs* H_1 : pelo menos dois $p_i \neq 1/5,\ i=1,\ldots,5$

Vamos então construir um quadro análogo ao apresentado no slide 182, das aulas teóricas

	Segunda	Terça	Quarta	Quinta	Sexta
O_i	24	18	15	21	28
p_i	1/5	1/5	1/5	1/5	1/5
$E_i = np_i$	21.2	21.2	21.2	21.2	21.2

$X^2_{cal} = \sum_{i=1}^{5} \frac{(O_i-E_i)^2}{E_i} = 0.3698+0.483+1.8132+0.0019+2.1811 = 4.849$ e $\chi^2_{0.05,4} = 9.488$.

Então $X^2_{cal} < \chi^2_{0.05,4}$, portanto, a um nível de significância de 5%, não se rejeita a hipótese de o número de faltas ser uniforme na semana.

5. Realizou-se um estudo do comportamento do gorgulho Azuki do feijão. Introduziram-se larvas desse gorgulho nos feijões que as alimentaram. As crisálidas saíram através de um buraco feito no feijão e, como tal, o nº de buracos por feijão indica-nos o nº de adultos que saíram. Os resultados observados em 100 feijões encontram-se na seguinte tabela:

nº de gorgulhos saídos de 1 feijão (x_i)	0	1	2	3	≥ 4
frequência observada (n_i)	60	22	10	5	3

Será o nº. de gorgulhos por feijão uma v.a. com distribuição de Poisson?

Resolução: Seja X a v.a. que designa o nº. de gorgulhos por feijão. As hipóteses a formular são:

$H_0 : X$ tem distribuição de Poisson $\quad vs \quad H_1 : X$ não tem distribuição de Poisson.

Note-se que não conhecemos o parâmetro da distribuição, então terá que ser estimado. Como é sabido, se $X \frown P(\lambda)$, tem-se $E[X] = \lambda$, então podemos considerar como estimativa para λ, $\hat{\lambda} = \bar{x}$.

No nosso problema temos $n = 100$ e $\bar{x} = \frac{\sum n_i x_i}{100} = 0.69$.

Temos que calcular a probabilidade de a variável X tomar os valores 0, 1, 2, 3 e ≥ 4, que são fáceis de obter por consulta das tabelas da Poisson com $\lambda = 0.7$. Vejamos como obter estes valores no R

```
> p1<-dpois(0:3,0.7);p1
[1] 0.49658530 0.34760971 0.12166340 0.02838813
```

o comando `dpois` permite obter $P[X = i] = \frac{e^{-\lambda}\lambda^i}{i!}$, neste caso $i = 0, 1, 2, 3$. Para se calcular $P[X \geq 4] = 1 - P[X \leq 3]$, no R obtém-se com o comando `ppois`

```
> p2<-1-ppois(3,0.7)
```

O vector que contém as probabilidades para $x = 0, 1, 2, 3, \geq 4$, no caso de $X \frown P(0.7)$ é

```
> pi<-c(p1,p2);pi
[1] 0.496585304 0.347609713 0.121663399 0.028388127 0.005753458
```

Vamos então construir o quadro análogo ao apresentado no slide 182, das aulas teóricas

	0	1	2	3	≥ 4
O_i	60	22	10	5	3
$\hat{p}_i$	0.49659	0.34761	0.12166	0.02839	0.00575
$E_i = n\hat{p}_i$	49.659	34.761	12.166	2.839	0.575

Como a última classe apresenta uma frequência esperada (estimada) inferior a 1 (ver slide 183, teóricas) devem juntar-se as duas últimas classes, isto é o quadro deve ser

	0	1	2	≥ 3
O_i	60	22	10	8
$\hat{p}_i$	0.49659	0.34761	0.12166	0.03414
$E_i = n\hat{p}_i$	49.659	34.761	12.166	3.414

$X^2_{cal} = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = 13.384 \quad$ e $\quad \chi^2_{0.05,2} = 5.99$.

Então $X^2_{cal} > \chi^2_{0.05,2}$, logo rejeita-se a hipótese de a distribuição ser $Poisson(0.7)$ a um nível de significância de 5%.

Vamos agora executar o teste do qui-quadrado, para o que é necessário dar o vector com os valores observados.

```
> oi<-c(60,22,10,8)
> chisq.test(oi, p = pi)
Chi-squared test for given probabilities

data:  oi
X-squared = 13.3836, df = 3, p-value = 0.003876

Warning message:
In chisq.test(oi, p = pi) : Chi-squared approximation may be incorrect
```

Atennda-se ao alerta dado, informando que deve haver cuidado na leitura do teste. Efectivamente tendo havido aqui necessidade de estimar o parâmetro da distribuição de Poisson o nº de graus de liberdade para o χ^2 é 2 e não 3 como está indicado.

Como é habitual a leitura do valor de $p - value$ dado no *output* conduz, tal como atrás vimos, à rejeição da hipótese nula.