IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia de Setúbal

DEPARTAMENTO DE MATEMÁTICA
MÉTODOS ESTATÍSTICOS
2.º SEMESTRE - 2020/2021
EXAME

Exame modelo — Duração: 2h30

Resolução

1. Dados relativos ao peso, em kg, de 24 pessoas.

   (a) número de classes $= k = \lfloor 1 + \frac{ln(24)}{ln(2)} \rfloor = 5$ classes

   Mínimo dos dados = 62 kg

   Máximo dos dados = 71 kg

   amplitude de cada classe $h = \frac{71-62}{5} = 1.8$ kg

i	classe$_i$	n_i	f_i	N_i	F_i
1	[62,63.8]	2	$\frac{2}{24} = 0.083$	2	$\frac{2}{24} = 0.083$
2	]63.8,65.6]	4	$\frac{4}{24} = 0.167$	$2+4=6$	$\frac{6}{24} = 0.25$
3	]65.6,67.4]	6	$\frac{6}{24} = 0.25$	$6+6=12$	$\frac{12}{24} = 0.5$
4	]67.4,69.2]	8	$\frac{8}{24} = 0.333$	$12+8=20$	$\frac{20}{24} = 0.833$
5	]69.2,71]	4	$\frac{4}{24} = 0.167$	$20+4=24$	$\frac{24}{24} = 1$
		$n = 24$	1		

   (b) X = peso dos pais aos vinte anos, Y = peso dos filhos aos vinte anos

   i. Como os dados estão na mesma unidade de medida e grandeza basta comparar as variâncias:

$$s_x^2 = \frac{53418 - 12 \times \left(\frac{800}{12}\right)^2}{12-1} = 7.697$$
$$s_y^2 = \frac{54849 - 12 \times \left(\frac{811}{12}\right)^2}{12-1} = 3.538$$

   Como $s_x^2 > s_y^2$, então os dados referentes ao peso dos pais aos vinte anos apresenta maior dispersão.

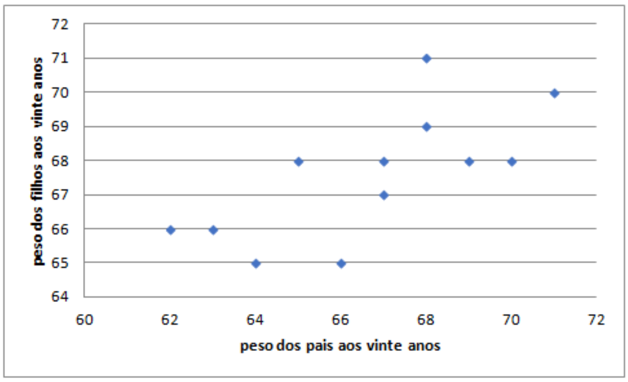
   ii. Variáveis:

   - Y, o peso dos filhos aos vinte anos, como variável dependente
   - X, o peso dos pais aos vinte anos, como variável independente.

   Amostra: $n = 12$

   Verificar se o modelo de regressão linear simples é adequado:

   - Diagrama de dispersão:



- Coeficiente de correlação linear:

$$r_{XY} = \frac{54107 - 12 \times \frac{800}{12} \times \frac{811}{12}}{\sqrt{\left(53418 - 12 \times \left(\frac{800}{12}\right)^2\right) \times \left(54849 - 12 \times \left(\frac{811}{12}\right)^2\right)}} = 0.703$$

Com base no diagram de dispersão e no coeficiente de correlação linear podemos dizer que a correlação linear existente é positiva pois é possível imaginar uma reta com declive positivo a passar pela nuvem de pontos e $r_{XY} > 0$. A correlação linear positiva existente não aparenta ser forte, mas também não pode ser considerada fraca, visto estar mais próxima de 1 do que de zero. Com base no diagrama de dispersão e no coeficiente de correlação linear parece ser adequado o modelo de regressão linear.

Reta de regressão:

$$\widehat{y} = 35.85 + 0.476x$$

pois

$$b = \frac{54107 - 12 \times \frac{800}{12} \times \frac{811}{12}}{53418 - 12 \times \left(\frac{800}{12}\right)^2} = 0.476 \qquad \text{e} \qquad a = \frac{811}{12} - 0.476 \times \frac{800}{12} = 35.85$$

Previsões:

$x = 98$ kg $\mapsto \widehat{y}\,(98) = 35.85 + 0.476 \times 98 = 82.498$ kg

Como o modelo de regressão linear simples foi considerado adequado, logo a previsão poderá ser considerada adequadas se o valor considerado para x se encontrar entre $[62, 71]$ ou numa sua vizinhança. A previsão efetuada para $x = 98 \notin [62, 71]$, logo a previsão efetuada, com base no modelo linear ajustado, não é de confiança pois não sabemos se a reta encontrada ainda se mantém na zona que estamos a fazer a previsão.

iii. Variável aleatória contínua Y = peso dos filhos aos vinte anos, $Y \sim N\left(\mu_y, \sigma_y\right)$.
Pretende-se determinar o grau de confiança de um intervalo para a média com uma margem de erro = 1. Como a População é Normal, σ_y é desconhecido e $n_y = 12 < 30$, então o Intervalo de Confiança a $(1-\alpha) \times 100\%$ para μ_y é

$$\left]\bar{y} - t_{(n_y-1);1-\frac{\alpha}{2}} \frac{s_y}{\sqrt{n_y}}, \bar{y} + t_{(n_y-1);1-\frac{\alpha}{2}} \frac{s_y}{\sqrt{n_y}}\right[$$

Portanto tem-se

$$\text{margem de erro} = \frac{\left(\bar{y} + t_{(n_y-1);1-\frac{\alpha}{2}} \frac{s_y}{\sqrt{n_y}}\right) - \left(\bar{y} - t_{(n_y-1);1-\frac{\alpha}{2}} \frac{s_y}{\sqrt{n_y}}\right)}{2} = t_{(n_y-1);1-\frac{\alpha}{2}} \frac{s_y}{\sqrt{n_y}} =$$

$$= t_{(12-1);1-\frac{\alpha}{2}} \frac{\sqrt{3.538}}{\sqrt{12}} = t_{(11);1-\frac{\alpha}{2}} \frac{\sqrt{3.538}}{\sqrt{12}}$$

logo pretende-se determinar o grau de confiança $= 1 - \alpha$ tal que

$$\text{margem de erro} = 1 \Leftrightarrow t_{(11);1-\frac{\alpha}{2}} \frac{\sqrt{3.538}}{\sqrt{12}} = 1 \Leftrightarrow t_{(11);1-\frac{\alpha}{2}} = 1.8 \Leftrightarrow$$

$$\Leftrightarrow 1 - \frac{\alpha}{2} = 0.95 \Leftrightarrow \alpha = 0.10 \Leftrightarrow 1 - \alpha = 0.90$$

iv. Variável aleatória contínua $Y =$ peso dos filhos aos vinte anos, $Y \sim N(\mu_y, \sigma_y)$
Pretende-se verificar se existem evidências de que o desvio padrão do peso dos filhos aos vinte anos é de 2kg, então vamos construir um intervalo de confiança de 95% para σ_y. Como a População é Normal e μ_y é desconhecido, então o Intervalo de Confiança a $(1-\alpha) \times 100\%$ para σ_y é

$$\left]\sqrt{\frac{(n_y - 1) \times s_y^2}{x^2_{(n_y-1);1-\frac{\alpha}{2}}}}, \sqrt{\frac{(n_y - 1) \times s_y^2}{x^2_{(n_y-1);\frac{\alpha}{2}}}}\right[$$

- grau de confiança $= 1 - \alpha = 0.95$
- nível de significância $= \alpha = 0.05$
- $x^2_{(n_y-1);\frac{\alpha}{2}} = x^2_{(12-1);\frac{0.05}{2}} = x^2_{(11);0.025} = 3.82$
- $x^2_{(n_y-1);1-\frac{\alpha}{2}} = x^2_{(12-1);1-\frac{0.05}{2}} = x^2_{(11);0.975} = 21.9$

Intervalo com 95% de confiança para o desvio padrão do peso dos filhos aos vinte anos:

$$\left]\sqrt{\frac{(12-1) \times 3.538}{21.9}}, \sqrt{\frac{(12-1) \times 3.538}{3.82}}\right[=]1.333, 3.192[$$

Com base na amostra recolhida e para um grau de confiança de 95%, há evidência estatística que o desvio padrão do peso dos filhos aos vinte anos poderá ser de 2 kg pois $2 \in]1.333, 3.192[$.

2. Seja X a variável aleatória discreta que representa o número de vezes em que uma dada máquina é desligada, num dia em que trabalhe 10 horas, $X \sim P(2)$ pois $V[X] = \lambda = 2$.

(a) Seja T a variável aleatória discreta que representa número de vezes em que uma dada máquina é desligada, num dia em que trabalhe 8 horas, $T \sim P(1.6)$ pois

$$\begin{array}{lcl} 10 \text{ horas} & \mapsto & \lambda = 2 \\ 8 \text{ horas} & \mapsto & \lambda_T = 1.6 \end{array}$$

Pretende-se

$$P(T = 0) = f(0) = 0.2019$$

(b) Seja V a variável aleatória contínua que representa o tempo, em horas, que a máquina está a trabalhar sem ser desligada, pela relação entre a distribuição de Poisson e a distribuição Exponencial, tem-se $V \sim Exp\,(5)$, pois $\theta = \frac{10}{2} = 5$. A função de distribuição da variável aleatória V é dada por:

$$F\,(v) = P\,(V \leq v) = \begin{cases} 0 & , \; v < 0 \\ 1 - e^{-\frac{v}{5}} & , \; v \geq 0 \end{cases}$$

$$P(V \geq 7) = 1 - P(V < 7) \underset{\text{v.a. contínua}}{=} 1 - F(7) = 1 - \left(1 - e^{-\frac{7}{5}}\right) = 0.2466$$

(c) Seja Y a variável aleatória contínua que representa o tempo, em horas, que o operador da máquina trabalha ininterruptamente por dia, $Y \sim U_{(2,5)}$, logo $b = 5$ e $a = 2$.

$$V\,[W] = V\left[\frac{1-Y}{2}\right] = V\left[\frac{1}{2} - \frac{1}{2}Y\right] = \left(-\frac{1}{2}\right)^2 V[Y] = \frac{1}{4} \times \frac{(5-2)^2}{12} = 0.1875$$

3. Seja X a variável aleatória contínua que representa o tempo, em minutos, para o percurso em hora de ponta de uma dada artéria de uma cidade, $X \sim N\,(\mu, 3)$ pois $\sigma = \sqrt{V[X]} = 3$, portanto

$$Z = \frac{X - \mu}{3} \sim N\,(0,1)$$

(a) Pretende-se determinar $\mu = E[X]$ sabendo

$$P(X > 13) = 0.1587 \Leftrightarrow P\left(Z > \frac{13-\mu}{3}\right) = 0.1587 \Leftrightarrow 1 - P\left(Z \leq \frac{13-\mu}{3}\right) = 0.1587 \Leftrightarrow$$
$$\Leftrightarrow 1 - \Phi\left(\frac{13-\mu}{3}\right) = 0.1587 \Leftrightarrow \Phi\left(\frac{13-\mu}{3}\right) = 0.8413 \Leftrightarrow \frac{13-\mu}{3} = z_{0.8413} \Leftrightarrow \frac{13-\mu}{3} = 1 \Leftrightarrow$$
$$\Leftrightarrow \mu = 10 \text{ minutos}$$

(b) Supondo que $\mu = E[X] = 10$ minutos. Seja $T = \sum_{i=1}^{15} X_i$ a variável aleatória contínua que representa o tempo total de percurso dos 15 veículos, com $X_i \sim N\,(10, 3)$ o tempo, em minutos, para o percurso do veículo i, com $i = 1, 2, ..., 15$. Pela aditividade da distribuição Normal tem-se $T \sim N\left(150, \sqrt{135}\right)$, pois

- $\mu_T = \sum_{i=1}^{15} \mu_i = 15 \times 10 = 150$
- $\sigma_T = \sqrt{\sum_{i=1}^{15} \sigma_i^2} = \sqrt{15 \times 3^2} = \sqrt{135}$

logo

$$Z = \frac{T - 150}{\sqrt{135}} \sim N\,(0,1)$$

Como 2horas = 120 minutos, pretende-se

$$P(T > 120) = P\left(Z > \frac{120 - 150}{\sqrt{135}}\right) = P(Z > -2.58) = 1 - P(Z \leq -2.58) = 1 - \Phi(-2.58) =$$
$$= 1 - (1 - \Phi(2.58)) = \Phi(2.58) = 0.9951$$

4. Seja X = comprimento do osso da perna do macho do casal e Y = comprimento do osso da perna da fêmea do casal. Amostras emparelhadas, logo $D = Y - X$

   (a) Como as amostras são amostras aleatórias emparelhadas, então é necessário construir a amostra das diferenças, $d_i = y_i - x_i$:

d= fêmea-macho	-0.6	-1.1	-2.4	0.6	-1.8	-1.3	0.3	0.3

Teste de hipóteses paramétrico

$H_0 : \mu_D \geq 0$ $\rightarrow$ em média, o tamanho do macho não pode ser considerado superior ao tamanho da fêmea
contra
$H_1 : \mu_D < 0$ $\rightarrow$ em média, o tamanho do macho pode ser considerado superior ao tamanho da fêmea

Estatística de Teste: Populações Normais com σ_D desconhecido e $n_D = 8 < 30$. Portanto

$$T = \frac{\bar{X}_D - \mu_D}{\frac{S_D}{\sqrt{n_D}}} \sim t_{(n_D - 1)}$$

Valor observado da Estatística de Teste sob a Hipótese H_0:

$$T_{obs} = \frac{-0.75 - 0}{\frac{1.089}{\sqrt{8}}} = -1.948$$

pois

- $\bar{x}_D = \frac{-0.6+(-1.1)+(-2.4)+0.6+(-1.8)+(-1.3)+0.3+0.3}{8} = -0.75$
- $s_D = \sqrt{\frac{((-0.6)^2+(-1.1)^2+(-2.4)^2+0.6^2+(-1.8)^2+(-1.3)^2+0.3^2+0.3^2)-8\times(-0.75)^2}{8-1}} = 1.089$

Região Crítica: o teste é unilateral esquerdo e $\alpha = 0.01$

$$RC = \left]-\infty, -t_{(n_D-1);1-\alpha}\right] = \left]-\infty, -t_{(8-1);1-0.01}\right] = \left]-\infty, -t_{(7);0.99}\right] =]-\infty, -3]$$

Como $T_{\text{obs}} = -1.948 \notin RC$, então Não se Rejeita H_0. Com base nas amostras e para um nível de significância de 1%, não existe evidência estatística que, em média, o tamanho do pinguim macho seja significativamente superior ao tamanho do pinguim fêmea.

OU

Valor-p: o teste é unilateral esquerdo e $T \sim t_{(7)}$ pois $n_D - 1 = 8 - 1 = 7$

valor-p $= P(T \leq T_{obs}) = P(T \leq -1.948) = F(-1.948) = 1 - F(1.948) \approx 1 - F(1.90) = 1 - 0.95 = 0.05$

Como valor-p $= 0.05 > 0.01 = \alpha$, então Não se Rejeita H_0. Com base nas amostras e para um nível de significância de 1%, não existe evidência estatística que, em média, o tamanho do pinguim macho seja significativamente superior ao tamanho do pinguim fêmea.

   (b) Teste de Wilcoxon $\rightarrow$ as amostras são emparelhadas

$$\begin{cases} H_0 : M_D = 0 \rightarrow & \text{não existem diferenças entre o tamanho do macho e da fêmea nos casais} \\ vs & \\ H_1 : M_D \neq 0 \rightarrow & \text{existem diferenças entre o tamanho do macho e da fêmea nos casais} \end{cases}$$

onde D = Fêmea - Macho

Valor observado da Estatística de Teste sob a Hipótese H_0:

$$T_{\text{obs}} = min\{T^-_{\text{obs}}, T^+_{\text{obs}}\} = min\{29.5, 6.5\} = 6.5$$

pois

- a soma das $n = 8$ posições (não há zeros) é $\frac{8\times(8+1)}{2} = 36$
- soma das posições com o sinal "+" $= T^+_{\text{obs}} = V = 6.5$
- soma das posições com o sinal "-" $= T^-_{\text{obs}} = 36 - 6.5 = 29.5$

Região Crítica: $\alpha = 0.05$, $n = 8$ e o teste é Bilateral

$$RC = [0, T_{n;\alpha}] = [0, T_{8;0.05}] = [0, 3]$$

Como $T_{\text{obs}} = 6.5 \notin RC$, então Não se Rejeita H_0. Com base nas amostras e para um nível de significância de 5%, não há evidência estatística que existam diferenças significativas entre o tamanho do macho e da fêmea nos casais destas aves.