DEPARTAMENTO DE MATEMÁTICA
MÉTODOS ESTATÍSTICOS
2.º SEMESTRE - 2020/2021
Exame Época de Recurso

Data: 16 de julho de 2021 Duração: 2 horas e 30 minutos

Resolução

1. Variáveis: X- durabilidade (em anos) e Y- preço de venda (em euros). Dimensão da amostra: $n = 10$.

(a) Tabela de Frequências:

i	Durabilidade (em anos) x_i	Freq. Absoluta n_i	Freq. Relativa f_i	Freq. Abs. Acumulada N_i	Freq. Rel. Acumulada F_i
1	1	1	$\frac{1}{10} = 0.1$	1	$\frac{1}{10} = 0.1$
2	3	4	$\frac{4}{10} = 0.4$	5	$\frac{5}{10} = 0.5$
3	4	2	$\frac{2}{10} = 0.2$	7	$\frac{7}{10} = 0.7$
4	5	2	$\frac{2}{10} = 0.2$	9	$\frac{91}{10} = 0.9$
5	6	1	$\frac{1}{10} = 0.1$	10	$\frac{10}{10} = 1$
		$n = 10$	1		

(b) Medidas de localização central:

- MÉDIA

$$\overline{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{37}{10} = 3.7 \text{ anos}$$

- MEDIANA

$$\tilde{x} = Q_{0.50} = \frac{x_2 + x_3}{2} = \frac{3+4}{2} = 3.5 \text{ anos}$$

pois $F_2 = 0.50$

- MODA

$$moda = x_2 = 3 \text{ anos}$$

por ser o valor que apresenta maior frequência absoluta (ou relativa).

Atendendo que

$$\text{moda} < \text{mediana} < \text{média}$$

podemos dizer que a distribuição dos dados é assimétrica positiva (ou enviesada para a direita).

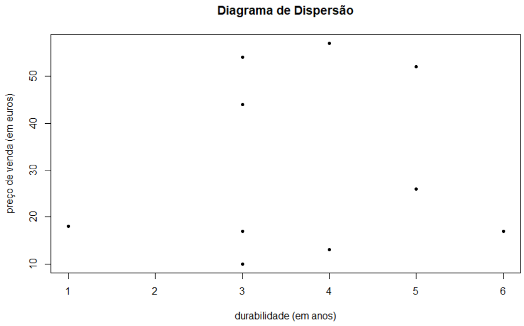
(c) Para verificarmos qual o conjunto de dados que apresenta maior dispersão relativa temos de calcular o coeficiente de variação:

$$CV_{\text{durabilidade}} = \frac{s_x}{\overline{x}} \times 100\% = \frac{\sqrt{\frac{155 - 10 \times 3.7^2}{10-1}}}{3.7} \times 100\% = 38.327\%$$

$$CV_{\text{preço de venda}} = \frac{s_y}{\overline{y}} \times 100\% = \frac{\sqrt{\frac{12645.841 - 10 \times 30.79^2}{10-1}}}{\frac{307.9}{10}} \times 100\% = 60.9113\%$$

Como $CV_{\text{durabilidade}} < CV_{\text{preço de venda}}$, conclui-se que neste conjunto de dados, o preço de venda tem uma maior dispersão do que durabilidade das sapatilhas.

(d) Para averiguarmos a existência de uma relação do tipo linear entre a durabilidade das sapatilhas e o seu preço de venda, vamos apresentar o diagrama de dispersão e calcular o coeficiente de correlação linear. Assim



$$r_{xy} = \frac{1164.63 - 10 \times \frac{37}{10} \times \frac{307.9}{10}}{\sqrt{\left(155 - 10 \times \left(\frac{37}{10}\right)^2\right) \times \left(12645.841 - 10 \times \left(\frac{307.9}{10}\right)^2\right)}} = $$
$$= 0.1061$$

Pelo diagrama de dispersão, que não apresenta uma tendência linear, e uma vez que $r_{xy} = 0.1061$ (próximo de 0), concluímos que não existe uma relação do tipo linear entre as variáveis e portanto o modelo de regressão linear simples não é adequado.

2. Seja X a variável aleatória discreta que representa o número de produtos vendidos diariamente por uma determinada marca.

(a) Função de probabilidade:

x	0	1	2	3
$f(x)$	a	b	c	d

Os dados:

$$\begin{cases} P(X=1) = P(X=2) \\ P(X=0) = 0.2 \\ P(X>2) = 0.15 \end{cases} \Leftrightarrow \begin{cases} f(1) = f(2) \\ f(0) = 0.2 \\ P(X=3) = 0.15 \end{cases} \Leftrightarrow \begin{cases} b = c \\ a = 0.2 \\ f(3) = 0.15 \end{cases} \Leftrightarrow \begin{cases} b = c \\ a = 0.2 \\ d = 0.15 \end{cases}$$

Como $\sum_x f(x) = 1$ para termos uma função de probabilidades, então:

$$\sum_x f(x) = 1 \Leftrightarrow f(0)+f(1)+f(2)+f(3) = 1 \Leftrightarrow a+b+c+d = 1 \Leftrightarrow 0.2+c+c+0.15 = 1 \Leftrightarrow c = 0.325$$

Então a função de probabilidade:

x	0	1	2	3
$f(x)$	0.2	0.325	0.325	0.15

A Função de Distribuição:

$$F(x) = \begin{cases} 0 & , \quad x < 0 \\ 0,2 & , \quad 0 \leq x < 1 \\ 0.525 & , \quad 1 \leq x < 2 \\ 0,85 & , \quad 2 \leq x < 3 \\ 1 & , \quad x \geq 3 \end{cases}$$

pois

- Se $x < 0$, então $F(x) = 0$
- Se $0 \leq x < 1$, então $F(x) = f(0) = 0.2$
- Se $1 \leq x < 2$, então $F(x) = 0.2 + f(1) = 0.2 + 0.325 = 0.525$

- Se $2 \leq x < 3$, então $F(x) = 0.525 + f(2) = 0.525 + 0.325 = 0.85$
- Se $x \geq 3$, então $F(x) = 0.85 + f(3) = 0.85 + 0.15 = 1$

(b) Considerando que a variável aleatória X segue uma Distribuição Uniforme Discreta com domínio $D = \{0, 1, 2, 3\}$, ou seja, $X \sim U_{(4)}$.

i. Como $X \sim U_{(4)}$ então a função de probabilidade é:

x	0	1	2	3
$f(x)$	0.25	0.25	0.25	0.25

Pretende-se calcular

$$P(X < 2 | X \geq 1) = \frac{P(X < 2 \wedge X \geq 1)}{P(X \geq 1)} = \frac{P(X = 1)}{f(1) + f(2) + f(3)} = \frac{f(1)}{0.25 + 0.25 + 0.25} = \frac{0.25}{0.75} = 0.3333$$

ii. Pretende-se calcular:

$$V[1 - 3X] = (-3)^2 \times V[X] = 9 \times \frac{15}{12} = 11.25$$

pois, como o domínio da variável aleatória X é constituído por inteiros consecutivos, tem-se:

$$V[X] = \frac{(3 - 0 + 1)^2 - 1}{12} = \frac{15}{12}$$

3. Seja X a variável aleatória contínua que representa o tempo, em horas, que um atleta de alta competição em período de estágio dorme por noite, $X \sim N(8, 0.6)$ pois $E[X] = \mu = 8$ horas e $\sqrt{V[X]} = \sigma = 0.6$ horas. Portanto tem-se

$$X \sim N(8, 0.6) \Leftrightarrow Z = \frac{X - 8}{0.6} \sim N(0, 1)$$

(a) Pretende-se determinar m tal que

$$P(X \geq m) = 0.33 \Leftrightarrow P\left(Z \geq \frac{m-8}{0.6}\right) = 0.33 \Leftrightarrow 1 - P\left(Z < \frac{m-8}{0.6}\right) = 0.33 \underset{\text{v.a.contínua}}{\Leftrightarrow}$$
$$\Leftrightarrow 1 - \Phi\left(\frac{m-8}{0.6}\right) = 0.33 \Leftrightarrow \Phi\left(\frac{m-8}{0.6}\right) = 0.67 \Leftrightarrow \frac{m-8}{0.6} = z_{0.67} \Leftrightarrow \frac{m-8}{0.6} = 0.44 \Leftrightarrow$$
$$\Leftrightarrow m = 8.264 \text{ horas}$$

(b) Seja Y a variável aleatória discreta que representa o número de atletas que dormem no máximo 7 horas por noite, num grupo de 15 atletas que estão em estágio, $Y \sim B(15, 0.05)$ pois

$$n = 15 \text{ atletas}$$
$$p = P(X \leq 7) = P\left(Z \leq \frac{7-8}{0.6}\right) = P(Z \leq -1.67) = \Phi(-1.67) = 1 - \Phi(1.67) =$$
$$= 1 - 0.9525 = 0.0475 \approx 0.05$$

Pretende-se

$$P(Y \geq 4) = 1 - P(Y < 4) \underset{\text{v.a. discreta}}{=} 1 - P(Y \leq 3) = 1 - F(3) = 1 - 0.9945 = 0.0055$$

(c) Amostra: dimensão $n = 15$

Seja $\overline{X}$ - Tempo médio que dorme por noite um atleta num estágio de 15 dias. Pretende-se $P\left(\overline{X} > 8.5\right)$.

Como a População é Normal e $\sigma = 0.6$ conhecido, tem-se

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

então

$$P\left(\overline{X} > 8.5\right) = P\left(Z > \frac{8.5 - 8}{\frac{0.6}{\sqrt{15}}}\right) = 1 - P\left(Z \leq 3.23\right) = 1 - \Phi(3.23) = 1 - 0.9994 = 0.0006$$

(d) Populações:

População 1: X_1- resultados do treino com o método 1, $X_1 \sim N(\mu_1, \sigma_1)$

População 2: X_2- resultados do treino com o método 2, $X_2 \sim N(\mu_2, \sigma_2)$

Amostras Aleatórias Independentes:

Amostra Aleatória da População 1: $n_1 = 13$, $\bar{x}_1 = 74.5$ e $s_1^2 = 82.6$

Amostra Aleatória da População 2: $n_2 = 11$, $\bar{x}_2 = 71.8$ e $s_2^2 = 112.6$

i. Como tem-se: Populações Normais e amostras independentes, o Intervalo de confiança a $(1-\alpha) \times 100\%$ para $\frac{\sigma_1}{\sigma_2}$ é:

$$\left]\sqrt{\frac{1}{f_{(n_1-1,n_2-1);1-\frac{\alpha}{2}}} \times \frac{s_1^2}{s_2^2}}, \sqrt{\frac{1}{f_{(n_1-1,n_2-1);\frac{\alpha}{2}}} \times \frac{s_1^2}{s_2^2}}\right[$$

Pelo enunciado sabe-se que o Intervalo de confiança a $(1-\alpha)\times 100\%$ para $\frac{\sigma_1}{\sigma_2}$ é $]0.5021, 1.4203[$, então igualando, por exemplo, o extremo inferior tem-se:

$$\sqrt{\frac{1}{f_{(n_1-1,n_2-1);1-\frac{\alpha}{2}}} \times \frac{s_1^2}{s_2^2}} = 0.5021 \Leftrightarrow \sqrt{\frac{1}{f_{(13-1,12-1);1-\frac{\alpha}{2}}} \times \frac{82.6}{112.6}} = 0.5021 \Leftrightarrow$$

$$\Leftrightarrow f_{(12,11);1-\frac{\alpha}{2}} = 2.91 \Leftrightarrow 1 - \frac{\alpha}{2} = 0.95 \Leftrightarrow \alpha = 0.1 \Leftrightarrow 1 - \alpha = 0.9$$

O grau de confiança utilizado no intervalo foi de 90%.

Com 90% de confiança, os desvios padrão dos resultados dos treinos podem ser considerados iguais pois $1 \in]0.5021, 1.4203[$.

ii. Como tem-se: Populações Normais, σ_1 e σ_2 desconhecidos mas podem ser considerados iguais, $\sigma_1 = \sigma_2$ (pela alínea anterior), amostras independentes com $n_1 = 13 < 30$ e $n_2 = 11 < 30$, o Intervalo de confiança a $(1-\alpha) \times 100\%$ para $\mu_1 - \mu_2$ é:

$$\left](\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2;1-\frac{\alpha}{2}}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1+n_2-2}}, (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2;1-\frac{\alpha}{2}}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1+n_2-2}}\right[$$

grau de confiança $= 1 - \alpha = 0.90$, logo nível de significância $= \alpha = 0.10$, portanto

$$t_{n_1+n_2-2;1-\frac{\alpha}{2}} = t_{13+11-2;1-\frac{0.10}{2}} = t_{22;0.95} = 1.72$$

O Intervalo de confiança a 90% para $\mu_1 - \mu_2$ é:

$$\left](74.5 - 71.8) - 1.72 \times \sqrt{\left(\frac{1}{13} + \frac{1}{11}\right) \times \frac{(13-1)\times 82.6 + (11-1)\times 112.6}{13+11-2}},\right.$$

$$\left.(74.5 - 71.8) + 1.72 \times \sqrt{\left(\frac{1}{13} + \frac{1}{11}\right) \times \frac{(13-1)\times 82.6 + (11-1)\times 112.6}{13+11-2}}\right[=$$

$$=]-4.2125, 9.6125[$$

Com 90% de confiança, em média, os métodos de treino podem ser considerados iguais pois $0 \in]-4.2125, 9.6125[$.

4. Populações:

X- Tempos obtidos por nadadores com calções tradicionais,

Y - Tempos obtidos por nadadores com novo tipo de calções,

Amostras emparelhadas.

(a) Amostras emparelhadas - Teste de Wilcoxon

Com $D = Y - X$

Pretende-se testar:

$$\begin{cases} H_0 : M_D = 0 \rightarrow & \text{o novo tipo de calções não permite obter melhores resultados} \\ vs & \\ H_1 : M_D > 0 \rightarrow & \text{o novo tipo de calções permite obter melhores resultados} \end{cases}$$

Teste unilateral direito, com $\alpha = 0.05$.

X	Y	$D = Y - X$	Sinal	$\|D\|$	Ordem
53.97	53.94	-0.03	-	0.03	4
55.88	55.86	-0.02	-	0.02	$\frac{2+3}{2} = 2.5$
49.85	49.83	-0.02	-	0.02	$\frac{2+3}{2} = 2.5$
58.90	58.89	-0.01	-	0.01	1
55.90	55.90	0	0	0	0
56.95	57.01	0.06	+	0.06	5

Soma das posições com o sinal "-"$= T^-_{obs} = 1 + 2.5 + 2.5 + 4 = 10$

Soma das posições com o sinal "+"$= T^+_{obs} = 5$

$n = 6 - 1 = 5$

Como o teste é unilateral direito tem-se

$$T_{obs} = T^-_{obs} = 10$$

e consideramos a tabela "one-Tailed Test" tem-se

$$RC = [0, T_{n,\alpha}] = [0, T_{5,0.05}] = [0, T_{n,\alpha}] = [0, 0] = \{0\}$$

Como $T_{obs} \notin RC$, não se rejeita H_0, não havendo evidências estatísticas que o novo calção permita obter melhores resultados.

(b) Como as amostras são emparelhadas, é necessário considerar a amostra das diferenças. Seja $D = Y - X$, então tem-se a amostra das diferenças:

d_i	-0.03	-0.02	-0.02	-0.01	0	0.06

Pretende-se testar:

$$\begin{cases} H_0 : \mu_D = 0 \rightarrow & \text{o tipo de calções não influência os resultados} \\ vs & \\ H_1 : \mu_D \neq 0 \rightarrow & \text{o tipo de calções influência os resultados} \end{cases}$$

Como a população é normal e desconhece-se o σ_D, e a dimensão da amostra é $n_D = 6 < 30$ então a Variável Fulcral:

$$T = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n_D}}} \sim t_{(n_D - 1)} \Leftrightarrow T \sim t_{(6-1)} \Leftrightarrow T \sim t_{(5)}$$

Então sob a hipótese nula tem-se

$$T_{obs} = \frac{-0.0033}{\frac{0.0327}{\sqrt{6}}} = -0.25$$

pois

$$\overline{d} = \frac{(-0.03) + (-0.02) + (-0.02) + (-0.01) + 0 + 0.06}{6} = -0.0033$$

$$s_d = \sqrt{\frac{(-0.03)^2 + (-0.02)^2 + (-0.02)^2 + (-0.01)^2 + 0^2 + 0.06^2 - 6 \times (-0.0033)^2}{6-1}} = 0.0327$$

Como temos um teste bilateral e a distribuição t de Student é simétrica:

$$\begin{aligned} valor - p &= 2 \times P\left(T \geq |T_{obs}|\right) = 2 \times P\left(T \geq |-0.25|\right) = 2 \times P\left(T \geq 0.25\right) = \\ &= 2 \times \left(1 - P\left(T < 0.25\right)\right) \underset{\text{v.a.contínua}}{=} 2 \times \left(1 - F\left(0.25\right)\right) \approx 2 \times \left(1 - F\left(0.267\right)\right) = \\ &= 2 \times (1 - 0.60) = 0.80 \end{aligned}$$

Como o $valor - p = 0.80 > \alpha = 0.05$ não se rejeita a hipótese H_0, não havendo evidências estatísticas que, em média, o tipo de calções influência os resultados.