

## Estatística

*Sérgio Manuel Salazar dos Santos, N<sup>o</sup>: 1020881*

23 de Dezembro de 2019

# Conteúdo

.1	Introdução . . . . .	1
.2	O conjunto de dados . . . . .	1
.3	Metodologia Estatística . . . . .	4
.3.1	Índice de Confiança tempo médio TEE . . . . .	4
.3.2	Verificar diferença de valores num intervalo . . . . .	4
.3.3	Verificar diferenças entre as regiões . . . . .	5
.3.4	Ajuste distribuição teórica à Empírica . . . . .	6
.4	Resultados e interpretação . . . . .	7
.5	Conclusões . . . . .	7

## **Resumo**

Este trabalho consiste no estudo de Estatística das Entregas Expresso em duas regiões **A** e **B**, as variáveis em estudo é o tempo de demora das entregas e a variável de número de encomendas entregues num determinado unidade de tempo [u.t.]. Nestas situações foram retiradas 120 e 90 amostras nas duas regiões respectivamente. A primeira é uma distribuição contínua, o tempo, e a segunda uma distribuição discreta.

As matérias abordadas vão ser **Amostragem**, **Estimação de parâmetros** e **Testes de Hipóteses**

# .1 Introdução

As variáveis consideradas são:

- Região (REG): variável nominal com dois níveis  
Região A  
Região B
- Tempo de entrega (TEE), por encomenda: Variável expressa em u.t.
- Número de encomendas entregues (NEE) por u.t.

Admitindo que a amostra disponível é uma amostra aleatória representativa das populações.

Neste relatório esta-se a trabalhar com duas grandezas precisamente o tempo (TEE) e quantidade por u.t (NEE), temos recolhidos 120 registos **TEE** na qual pela regra de sturges  $c = \text{int}(1 + 3.3\log(n))$ , determina-se que é necessário sete [7] classes.

Podemos obter a amplitude de cada classe  $h = b - a$  e sua marca  $x_i = \frac{a+b}{2}$ .

## .2 O conjunto de dados

$X_{iA}$  - "Variável aleatória que representa o tempo de demora na Região **A** da entrega de uma encomenda Expresso em u.t."  $i=1,2,3, \dots, 120$

$X_{iB}$  - "Variável aleatória que representa o tempo de demora na Região **B** da entrega de uma encomenda Expresso em u.t."  $i=1,2,3, \dots, 120$

Abaixo o resultado da tabela TEE:

$h_i$	CLASSE	MARCA	$n_{iA}$	$n_{iB}$	$\frac{n_{iA}}{h_i}$	$\frac{n_{iB}}{h_i}$	$f_{iA}$	$f_{iB}$	$F_{iA}$	$F_{iB}$	$e_{iA}$
$-\infty$	$< 5$		0	0							1,1812
4	$[5,10[$	7,5	8	1	2	0,25	0,0667	0,0083	0,0667	0,0083	5,9871
4	$[10,15[$	12,5	16	18	4	4,5	0,1333	0,15	0,2	0,1583	18,8942
4	$[15,20[$	17,5	40	28	10	7	0,3333	0,2333	0,5333	0,3917	33,6282
4	$[20,25[$	22,5	25	41	6,25	10,25	0,2083	0,3417	0,7417	0,7333	33,7887
4	$[25,30[$	27,5	26	22	6,5	5,5	0,2167	0,1833	0,9583	0,9167	19,1663
4	$[30,35[$	32,5	4	8	1	2	0,0333	0,0667	0,9917	0,9833	6,1316
5	$[35,40]$	37,5	1	2	0,2	0,4	0,0083	0,0167	1	1	1,1044
$+\infty$	$>40$		0	0							0,1183
			n=120	n=120							

$n_i$  - frequência absoluta     $f_i$  - frequência relativa     $F_i$  - frequência acumulada

Recorrendo ao excel obteve-se os seguintes resultados:

Média aritmética dados classificados	Variância de uma amostra dados classificados
$\bar{x} = \frac{1}{n} \sum_{i=1}^c x_i n_i = \sum_{i=1}^c x_i f_i$	$s^2 = \frac{1}{n-1} \sum_{i=1}^c (x_i - \bar{x})^2 n_i$

Estatística	$X_A$	$X_B$
Mínimo	7,5	7,5
$Q_1$ :1º Quartil	17,5	17,5
$m_d$ : mediana	17,5	22,5
$Q_3$ :3º Quartil	27,5	27,5
Máximo	37,5	37,5
$\bar{X}$ : Média	20,0417	21,5417
$s$ : desvio-padrão	6,4494	6,0909
$m_o$ : moda	17,5	22,5
Tamanho amostral $[n]$	120	120

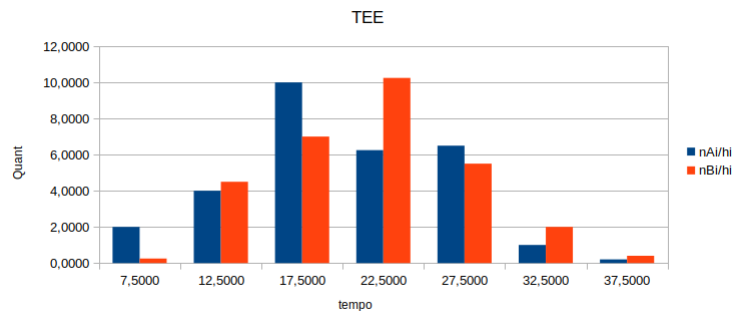


Figura 1: TEE

A mediana pode ser obtida pela frequência acumulativa quando esta é igual a 50%, ou seja,  $F_i(\text{Mediana}) = 0,5$

#### Linearização mediana **TEE**

Regiao **A**:

$$0.2 \Rightarrow 12.5$$

$$0.5333 \Rightarrow 17.5$$

$\therefore$

Midiana A =

$$12.5 + 0.9 \times (17.5 - 12.5) = 17$$

com:

$$\text{skew} = -0,1051 \text{ e kurt} = -0,4016$$

Regiao **B**:

$$0.3917 \Rightarrow 17.5$$

$$0.7333 \Rightarrow 22.5$$

$\therefore$

Midiana B =

$$17.5 + 0.317 \times (22.5 - 17.5) = 19.085$$

com :

$$\text{skew} = 0,1119 \text{ e kurt} = -0,1835$$

Na prática, considera-se que a qualidade da aproximação é suficientemente boa quando  $n \geq 30$ .

Pode-se tomar que  $\delta \cong s$ .

$$\begin{cases} \mu \\ \delta \end{cases} \Rightarrow \bar{X} = \frac{\sum_{i=1}^n X_i}{n} \sim N\left(\mu; \frac{\delta^2}{n}\right)$$

$$\bar{x}_{A_0} = 20,0417 \quad \bar{x}_{B_0} = 21,5417$$

$$\delta_A = 6,4494 \quad \delta_B = 6,0909$$

Tratamento dos dados da Segunda Variável Aleatória

$Y_{iA}$ - "Variável aleatória que representa o número de encomendas entregues pela Expresso na Região **A** por u.t."

$i=1,2,3, \dots, 90$

$Y_{iB}$ - "Variável aleatória que representa a número de encomendas entregues pela Expresso na Região **B** por u.t."

$i=1,2,3, \dots, 90$

Abaixo o resultado da tabela NEE:

$Y_i$	$n_{iA}$	$n_{iB}$	$f_{iA}$	$f_{iB}$	$F_{iA}$	$F_{iB}$	$e_{iB}$
< 3	0	0					1,2765
3	6	3	0,0667	0,0333	0,0667	0,0333	2,8549
4	8	6	0,0889	0,0667	0,1556	0,1	5,3855
5	19	13	0,2111	0,1444	0,3677	0,2444	8,6724
6	15	7	0,1667	0,0778	0,5333	0,3222	11,9216
7	13	19	0,1444	0,2111	0,6778	0,5333	13,9899
8	11	15	0,1222	0,1667	0,8	0,7	14,0145
9	6	8	0,0667	0,0889	0,8667	0,7889	11,9847
10	5	11	0,0556	0,1222	0,9222	0,9111	8,7490
11	4	3	0,0444	0,0333	0,9667	0,9444	5,4522
12	0	2	0	0,0222	0,9667	0,9667	2,9005
13	2	1	0,0222	0,0111	0,9889	0,9778	1,3172
14	1	0	0,0111	0	1	0,9778	0,5106
15	0	1	0	0,0111	1	0,9889	0,1690
16	0	1	0	0,0111	1	1	0,0477
>16	0	0					0,0330

Estatística	$Y_A$	$Y_B$
Mínimo	3	3
$Q_1$ : 1º Quartil	5	6
$m_d$ : mediana	6	7
$Q_3$ : 3º Quartil	8	9
Máximo	14	16
$\bar{Y}$ : Média	6,6111	7,5111
$s$ : desvio-padrão	2,3112	2,5140
$m_o$ : moda	5	7
Tamanho amostral $[n]$	90	90

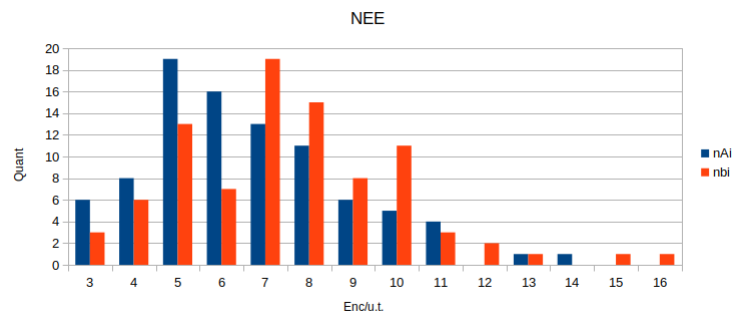


Figura 2: NEE

Na Região **A** a Média > Mediana > Moda

com skew = 0.74553 e kurt = 0.49789

Na Região **B** a Média > Mediana = Moda

com skew = 0.67659 e kurt = 1.01076

$$\begin{cases} \mu \\ \delta \end{cases} \Rightarrow \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} \sim N\left(\mu; \frac{\delta^2}{n}\right)$$

$$\bar{y}_{A_0} = 6,6111 \quad \bar{y}_{B_0} = 7,5111$$

$$\delta_A = 2,3112 \quad \delta_B = 2,5140$$

### .3 Metodologia Estatística

#### .3.1 Índice de Confiança tempo médio TEE

Estimação do tempo médio para as regiões **A** e **B** com um índice de confiança de 95%.

$$IC_{1-\alpha} = [A, B]; \text{ para } 1 - \alpha = 0.95, \alpha = 0.05, \frac{\alpha}{2} = 0.025$$

$$\text{Zona critica } Z_c = Z_{1-\frac{\alpha}{2}} = \Phi^{-1}(0.975) \cong 1.96$$

$$P(A \leq \mu \leq B) = 1 - \alpha$$

$$\Delta = Z_c \times \frac{\delta}{\sqrt{n}}$$

$$A = \bar{x} - \Delta \quad \text{and} \quad B = \bar{x} + \Delta$$

$\therefore$

$$IC_{A_{0.95}} = [18.8877, 21.1956] \quad \text{and} \quad IC_{B_{0.95}} = [20.4519, 22.6314]$$

Pode-se estimar que o tempo médio  $[\mu]$  de entrega na população esta dentro dos intervalos acima mencionados com 95% de confiança.

#### .3.2 Verificar diferença de valores num intervalo

Verificar se os dados permitem afirmar que existe diferença significativa entre a % de períodos com menos de 6 entregas por u.t. na região **A** e na região **B**. Responda com base num intervalo de confiança de 97%.

Distribuição discreta:

$$\bar{y}_{A_0} = 6,6111 \quad \bar{y}_{B_0} = 7,5111 \quad n = 90$$

$$\delta_A = 2,3112 \quad \delta_B = 2,5140$$

$$\left\{ \begin{array}{l} \mu \\ \delta \end{array} \right\} \implies \bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} \sim N\left(\mu; \frac{\delta^2}{n}\right)$$

$$P(Y_A < 6) = P(Y_A \leq 5) = F_{i_B}(5) \cong 0,3677 \quad \text{e} \quad P(Y_B < 6) = P(Y_B \leq 5) = F_{i_B}(5) \cong 0,2444$$

$$\hat{P}_A - \hat{P}_B \sim N\left(p_A - p_B; \frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}\right) \quad \Delta = z_{(1-\frac{\alpha}{2})} \sqrt{\frac{\hat{p}_A \hat{q}_A}{n_A} + \frac{\hat{p}_B \hat{q}_B}{n_B}} \quad q = (1 - p)$$

$$IC_{97\%}(\hat{P}_A - \hat{P}_B) = [(\hat{p}_A - \hat{p}_B) - \Delta; (\hat{p}_A - \hat{p}_B) + \Delta]$$

$$\hat{P}_A - \hat{P}_B \sim N(0,1233; 0,02788) \quad z_{(1-\frac{\alpha}{2})} = \phi^{-1}(0,985) = 2,1701$$

Recorrendo a calculadora casio  $fx - 9860GII$  :

$$\Delta = InvNorm(0.985) \sqrt{\frac{0.3677(1-0.3677)}{90} + \frac{0.2444(1-0.2444)}{90}} \cong 0.3677$$

$\therefore$

$$IC_{97\%}(\hat{P}_A - \hat{P}_B) = [(\hat{p}_A - \hat{p}_B) - 0,3624; (\hat{p}_A - \hat{p}_B) + 0,3624]$$

A Diferença de proporções é 36,24%.

### .3.3 Verificar diferenças entre as regiões

Testar se a região (REG) tem um efeito estatisticamente significativo sobre TEE e NEE ao nível de diferença de médias. Considere uma significância à sua escolha inferior ou igual a 5%. Use o critério do valor de prova para fundamentar a decisão.

$$\begin{cases} H_0 : & \mu_A - \mu_B = 0 \\ H_1 : & \mu_A - \mu_B < 0 \end{cases}$$

Condição TEE:

$$\begin{cases} \mu = 0 \\ \delta = s \end{cases} \implies \bar{X}_A - \bar{X}_B \sim N\left(0, \frac{\delta_A^2}{n_A} + \frac{\delta_B^2}{n_B}\right)$$

$$\bar{x}_A - \bar{x}_B = -1.5$$

$$\frac{\delta_A^2}{n_A} + \frac{\delta_B^2}{n_B}$$

$$z_0 = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{\delta_A^2}{n_A} + \frac{\delta_B^2}{n_B}}} \quad RC_z = ]-\infty, -z_{1-\alpha}] \quad pvalue = P(Z < z_0)$$

Condição REE:

$$\begin{cases} \mu = 0 \\ \delta = s \end{cases} \implies \bar{Y}_A - \bar{Y}_B \sim N\left(0, \frac{\delta_A^2}{n_A} + \frac{\delta_B^2}{n_B}\right)$$

$$\bar{y}_A - \bar{y}_B = -0.9$$

$$\frac{\delta_A^2}{n_A} + \frac{\delta_B^2}{n_B}$$

$$z_0 = \frac{\bar{y}_A - \bar{y}_B}{\sqrt{\frac{\delta_A^2}{n_A} + \frac{\delta_B^2}{n_B}}} \quad RC_z = ]-\infty, -z_{1-\alpha}] \quad pvalue = P(Z < z_0)$$



### .3.4 Ajuste distribuição teórica à Empírica

Ajuste uma distribuição teórica à distribuição empírica das variáveis TEE na região A (considerando as classes definidas) e NEE na região B. Verifique a qualidade do ajuste ao nível de 5%.

TEE Região A:

$$\begin{cases} H_0 : X \sim N(20.0417, 6.4494^2) \\ H_1 : X \sim N(20.0417, 6.4494^2) \end{cases}$$

$$q_0 = \sum_{i=1}^n \frac{(n_i - e_i)^2}{e_i} \sim \chi^2_{(k-p-1)}$$

NEE Região B:

$$\begin{cases} H_0 : X \sim N(7.5111, 2.5140^2) \\ H_1 : X \sim N(7.5111, 2.5140^2) \end{cases}$$

$$q_0 = \sum_{i=1}^n \frac{(n_i - e_i)^2}{e_i} \sim \chi^2_{(k-p-1)}$$

## **.4 Resultados e interpretação**

fazer tabela só com resultados

## **.5 Conclusões**

A Distribuição normal tem a Média = Mediana = Moda, devido a ter uma distribuição simétrica, quando estamos a analisar valores discretos isto não acontece devido a não ser simétrico podendo ter vários casos diferentes, e quanto menor o número de amostras da população maior a dificuldade de se poder inferir e estimar valores.

Falar da skew e curt.

# Lista de Figuras

1	TEE . . . . .	2
2	NEE . . . . .	3
[ ]	<sup>1</sup>	

---

<sup>1</sup>Apontamentos Estatística