Instituto Superior de
Engenharia do Porto

Estatística

Sérgio Manuel Salazar dos Santos, Nº: 1020881

23 de fevereiro de 2021

Conteúdo

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 numero de encomendas entregues num determinado unidade de tempo [u.t.]. Nestas situações foram retiradas 120 e 90 amostras nas duas regiões respetivamente.
A primeira é uma distribuição continua, o tempo, e a segunda uma distribuição discreta.
As materias abordadas vai ser **Amostragem**, **Estimação de parâmetros** e **Testes de Hipóteses**

1 Introdução

As variáveis consideradas são:

- Regiao (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 $c = int(1+3.3log(n))$, 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

Tratamento dos dados da Variável Aleatória

X_{i_A}- "Variável aleatória que representa o tempo de demora na Região **A** da entrega de uma encomenda Expresso em u.t." i=1,2,3, ...,120
X_{i_B}- "Variável aleatória que representa o tempo de demora na Região **B** da entrega de uma encomenda Expresso em u.t." i=1,2,3, ...,120
Abaixo o resultado da tabela TEE:
Recorrendo ao excell obteve-se os seguintes resultados.

h_i	CLASSE	MARCA	n_{i_A}	n_{i_B}	$\frac{n_{i_A}}{h_i}$	$\frac{n_{i_B}}{h_i}$	f_{i_A}	f_{i_B}	F_{i_A}	F_{i_B}	e_{i_A}
$-\infty$	< 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

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

Estatística	X_A	X_B
Mínimo	7,5	7,5
Q_1:1^o Quartil	17,5	17,5
m_d: mediana	17,5	22,5
Q_3:3^o 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

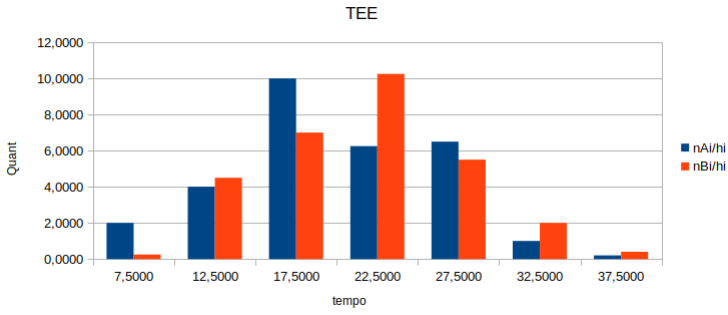

Figura 1: TEE

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

Linearização mediana **TEE**

Região **A**:
0.2 $\implies$ 12.5
0.5333 $\implies$ 17.5
$\therefore$
Mediana A =
12.5 + 0.9 x (17.5-12.5) = 17
com:
skew = -0,1051 e kurt = -0,4016

Região **B**:
0.3917 $\implies$ 17.5
0.7333 $\implies$ 22.5
$\therefore$
Mediana B =
17.5 + 0.317 x (22.5-17.5) = 19.085
com :
skew = 0,1119 e kurt = -0,1835

Na prática, considera-se que a qualidade da aproximação é suficientemente boa quando $n \geqslant 30$.
Pode-se tomar que $\delta \cong s$.

$\bar{x}_{A_0} = 20,0417 \qquad \bar{x}_{B_0} = 21,5417$
$\delta_A = 6,4494 \qquad \delta_B = 6,0909$

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

Tratamento dos dados da Segunda Variável Aleatória

Y_{i_A}- "Variavel aleatória que representa o numero de encomendas entregues pela Expresso na Região **A** por u.t." i=1,2,3, ...,90
Y_{i_B}- "Variável aleatória que representa a numero de encomendas entregues pela Expresso na Regiao **B** por u.t." i=1,2,3, ...,90
Abaixo o resultado da tabela NEE:

Y_i	n_{i_A}	n_{i_B}	f_{i_A}	f_{i_B}	F_{i_A}	F_{i_B}	e_{i_B}
< 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_1:1^o Quartil	5	6
m_d: mediana	6	7
Q_3:3^o 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

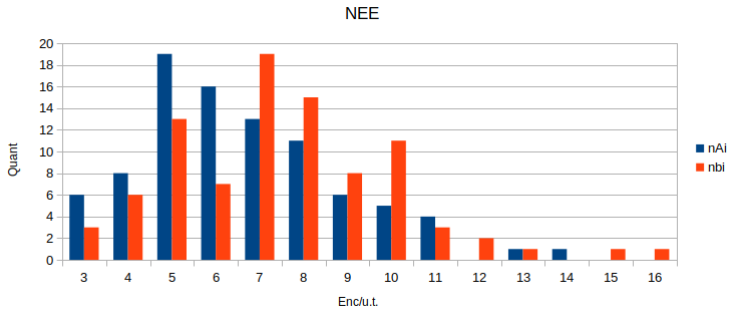

Figura 2: NEE

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

$\bar{y}_{A_0} = 6{,}6111 \qquad \bar{y}_{B_0} = 7{,}5111$
$\delta_A = 2{,}3112 \qquad \delta_B = 2{,}5140$

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

3 Metodologia Estatística

3.1 Índice de Confiança tempo médio TEE

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

$IC_{1-\alpha} = [A, B]$; para $1 - \alpha = 0.95$, $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$
Zona critica $Z_c = Z_{1-\frac{\alpha}{2}} = \Phi^{-1}(0.975) \cong 1.96$
$P\left(A \leqslant \mu \leqslant B\right) = 1 - \alpha$
$\triangle = Z_c \times \frac{\delta}{\sqrt{n}}$
$A = \bar{x} - \triangle \quad and \quad B = \bar{x} + \triangle$
$\therefore$
$IC_{A_{0.95}} = [\ 18.8877\ ,\ 21.1956\ ]$ and $IC_{B_{0.95}} = [\ 20.4519\ ,\ 22.6314\ ]$

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:

$\bar{y}_{A_0} = 6{,}6111 \qquad \bar{y}_{B_0} = 7{,}5111 \qquad n = 90$
$\delta_A = 2{,}3112 \qquad \delta_B = 2{,}5140$

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

$\hat{P_A} - \hat{P_B} \sim N\left(p_A - p_B; \frac{p_A\, q_A}{n_A} + \frac{p_B\, q_B}{n_B}\right) \qquad \triangle = z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\hat{p_A}\,\hat{q_A}}{n_A} + \frac{\hat{p_B}\,\hat{q_B}}{n_B}} \qquad q = (1-p)$

$IC_{97\%}(\hat{P_A} - \hat{P_B}) = [(\hat{p_A} - \hat{p_B}) - \triangle\ ;\ (\hat{p_A} - \hat{p_B}) + \triangle]$

$\hat{P_A} - \hat{P_B} \sim N\left(0,1233\ ;\ 0,02788\right) \qquad z_{(1-\frac{\alpha}{2})} = \phi^{-1}(0,985) = 2,1701$

Recorrendo a calculadaora casio $fx - 9860GII$:

$\triangle = InvNorm(0.985)\sqrt{\frac{0.3677(1-0.3677)}{90} + \frac{0.2444(1-0.2444)}{90}} \ \cong\ 0.3677$
$\therefore$
$IC_{97\%}(\hat{P_A} - \hat{P_B}) = [\ (\hat{p_A} - \hat{p_B})\ -\ 0,3624\ ;\ (\hat{p_A} - \hat{p_B})\ +\ 0,3624\ ]$

A Diferença de proporçôes 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. Considerando uma significância de 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:

$$\begin{cases} \mu = 0 \\ \delta = s \end{cases} \implies \bar{X} = \bar{X}_A - \bar{X}_B \backsim N\left(0\,,\, \frac{\delta_A^2}{n_A} + \frac{\delta_B^2}{n_B}\right) \quad ; \quad \frac{\delta_A^2}{n_A} + \frac{\delta_B^2}{n_B} \cong 0.6558$$

$$P(\bar{X}_{H_0} \leqslant C) = 0.05 \implies RC_X\,]-\infty\,,\, -1.332] \qquad \bar{x}_A - \bar{x}_B = -1.5 \in RC_X$$

$$z_0 = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{\delta_A^2}{n_A} + \frac{\delta_B^2}{n_B}}} \cong -1.8523 \qquad RC_z =]-\infty\,,\, -1.6448] \qquad pvalue = P(Z < z_0) = 0.032$$

Condição NEE:

$$\begin{cases} \mu = 0 \\ \delta = s \end{cases} \implies \bar{Y} = \bar{Y_A} - \bar{Y_B} \backsim N\left(0\,,\, \frac{\delta_A^2}{n_A} + \frac{\delta_B^2}{n_B}\right) \quad ; \quad \frac{\delta_A^2}{n_A} + \frac{\delta_B^2}{n_B} \cong 0.1296$$

$$P(\bar{Y}_{H_0} \leqslant C) = 0.05 \implies RC_Y\,]-\infty\,,\, -0.5921] \qquad \bar{y}_A - \bar{y}_B = -0.9 \in RC_Y$$

$$z_0 = \frac{\bar{y}_A - \bar{y}_B}{\sqrt{\frac{\delta_A^2}{n_A} + \frac{\delta_B^2}{n_B}}} \cong -2.5 \qquad RC_z =]-\infty\,,\, -1.6448] \qquad pvalue = P(Z < z_0) = 0.0062$$

A Hipótese de proximidade entre as regiões é falsa, ambos os critérios estão dentro da região de rejeição logo a hipótese imposta é falsa. O valor de prova também reforça a ideia pois a percentagem de favorecimento é quase nulo.

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%.

k-numero de classes ; m-numero de parâmetros

TEE Região A:

k=6 , m=2 e α=0.05

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

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

$RC_{\chi^2} = [\ InvChiCD(0.05, 3)\ ,\ +\infty\ ] \quad \rightarrow \quad RC_{\chi^2} = [\ 7.8147\ ,\ +\infty\ ]$

$q_0 = 7.2234 < 7.8147$

NEE Região B:

k=8 , m=2 e α=0.05

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

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

$RC_{\chi^2} = [\ InvChiCD(0.05, 5)\ ,\ +\infty\ ] \quad \rightarrow \quad RC_{\chi_2} = [\ 11.0705\ ,\ +\infty\ ]$

$q_0 = 8.5532 < 11.0705$

Ambas as condições propostas são aceitáveis como distribuições com um grau de confiança de 95%, pois estão fora da região de rejeição.

3.5 Relação Erro Tipo 1 e 2 da alínea 3.3

Apresente um gráfico expressando a relação entre o erro tipo I (α) e a potência do teste (1-β), para valores hipotéticos das verdadeiras diferenças de médias calculadas anteriormente no ponto 3.3.

Distribuição normal Diferença.

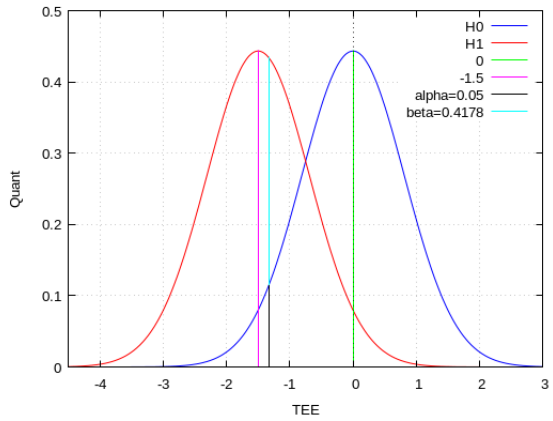

Figura 3: TEE Diferênça

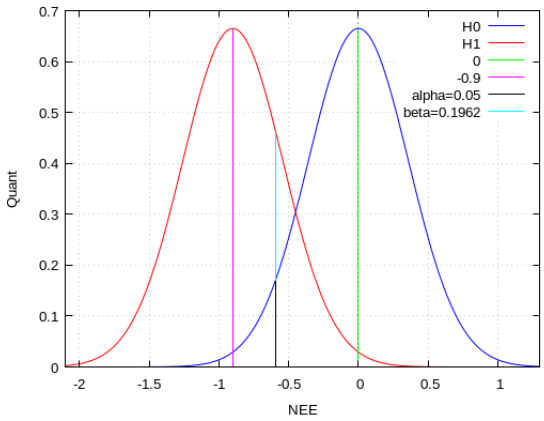

Figura 4: NEE Diferença

Continuação de 3.3

TEE Região A:

$$\begin{cases} H_0 : \bar{X}_{H_0} \backsim N(0\ ,\ 0.6558) \\ \\ H_1 : \bar{X}_{H_1} \backsim N(-1.5\ ,\ 0.6558) \end{cases}$$

$\beta = P(AceitarH_0|H_0Falsa)$
$\beta = (\bar{X}_{H_1} > -1.332)$
$\beta = NormCD(-1.332, 99999999, \sqrt{0.6558}, -1.5) = 0.4178$
Potência do teste
$1 - \beta = P(RejeitarH_0|H_0Falsa) = 0.5822$

NEE Região B:

$$\begin{cases} H_0 : \bar{Y}_{H_0} \backsim N(0\ ,\ 0.1296) \\ \\ H_1 : \bar{Y}_{H_1} \backsim N(-0.9\ ,\ 0.1296) \end{cases}$$

$\beta = P(AceitarH_0|H_0Falsa)$
$\beta = (\bar{Y}_{H_1} > -0.5921)$
$\beta = NormCD(-0.5921, 99999999, \sqrt{0.1296}, -0.9) = 0.1962$
Potência do teste
$1 - \beta = P(RejeitarH_0|H_0Falsa) = 0.8038$

Sem tempo para fazer gráfico relação α com β.

Hipóteses na qual as amostras refletem:

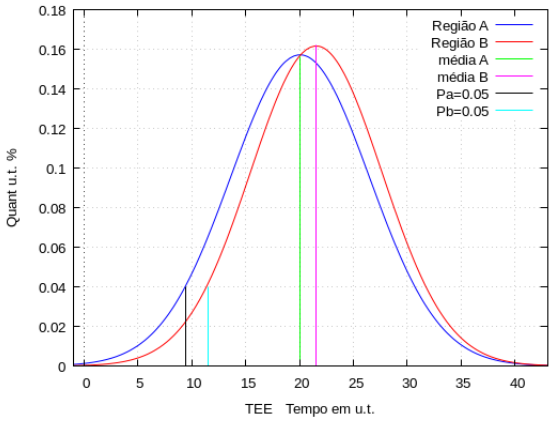

Figura 5: TEE Normal

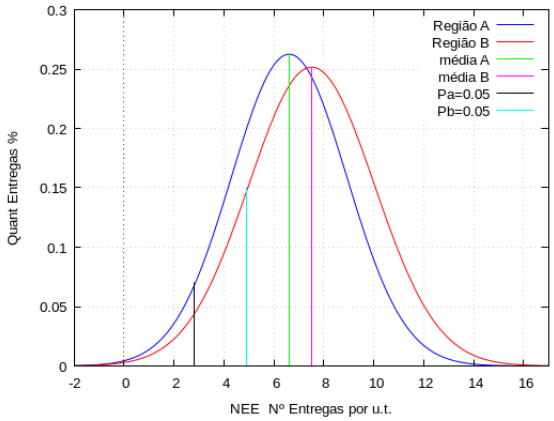

Figura 6: NEE Normal

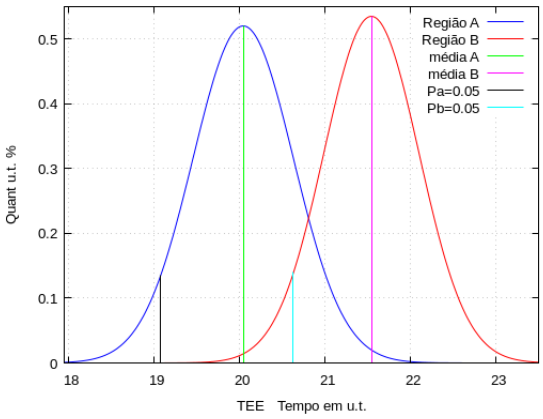

Figura 7: TEE Normal Média

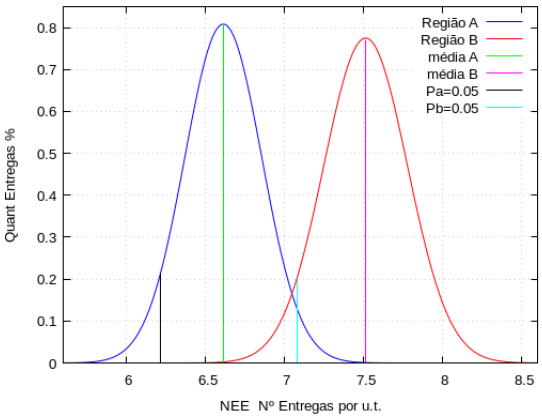

Figura 8: NEE Normal Média

4 Resultados e interpretação

A mediana é o ponto de equilibriu da distribuição nos informa o ponto na qual o pesso em ambos os lados é igual, em conjunto com a média e a moda nos pode dar mais informação sobre sua identidade, sobre sua calda e sua forma. Neste trabalho temos quatro distribuições Normais que diferem uma das outras, ou seja cada região tem um comportamento que lhe é próprio.
Algumas deduções estão descritas nos problemas propostos, mas tudo indica que o objetivo é obter uma média representativa dos acontecimentos de forma a poder inferir com uma certa incerteza, o procedimento de normalizar os dados empíricos pressupõe sempre uma perda de alguma informação residual, mas nesta situação como são variáveis imprevisíveis dependendo de muitas variáveis seria melhor ter critérios mais rigorosos e maior numero de amostras.
Comparando o histograma com a distribuição normal hipotética da para perceber muitas diferenças.

5 Conclusões

Este relatório foi feito recorrendo ao excell do libreoffice, em conjunto com a calculadora Casio fx-9860GII e o WxMaxima, portanto todos os resultados não estão apresentados no excell devido aos cálculos auxiliares terem sido feitos aparte.

A ortografaria do relatório pode ter erros também suas soluções, os exercícios propostos são muito abrangentes e o tempo definido curto para sua exploração mais detalhada, sendo que podia ser muito mais elaborado e feito mais testes para ter um estudo mais aprofundado.
Muitas das questões levam a ter dúvidas de forma a aprofundar a matéria, dando a sensação na qual não conseguimos obter uma completa perceção no seu todo.

O relatório é um estudo acerca da estatística das variáveis aleatórias expressas ao redor da **Distribuição Normal** em que sua Média = Mediana = Moda, é uma distribuição simétrica, quando estamos a analisar valores discretos e contínuos no mundo real isto não acontece devido a não ser simétrico podendo ter vários casos diferentes, e quanto menor o numero de amostras da população maior a dificuldade de se poder inferir e estimar valores.

Fazer o estudo de uma população para poder inferir seu comportamento através de tiros no escuro, ou seja, hipóteses tomadas como verdades e comparar com os resultados de forma a poder tirar uma decisão das suas preposição.

No caso do χ^2 podermos averiguar qual o grau de proximidade da distribuição proposta para representar nossos dados, para podermos depois analisar o desconhecido pelo já adquirido, sempre com uma margem de incerteza.
Fazer inferências acerca de uma população atravez de amostras há sempre a possibilidade de erro, neste caso são dois os tipos identificados. O primeiro tipo é quando se rejeita a hipótese imposta quando ela é verdade, e a segunda aceitar uma hipótese que é falsa, sendo que a segunda no meu ver é mais grave, dado que errar e estar tudo bem é sempre uma boa surpresa, caso contrario um desastre.

Mais uma nota que o pré-requisito o relatório ser inferior a 10 páginas, terá de se retirar o titulo a Lista de figuras, bibliografia e indexação.

Lista de Figuras

Bibliografia

[1] *Probabilidades e estatística Volume 1.* McGraw-Hill, 1990.

[2] *Probabilidades e Processos Estocásticos.* Universidade de Aveiro, 2002.

[1]

[1] Apontamentos Estatistica