UNIVERSIDADE DE SÃO PAULO – USP
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO
DEPARTAMENTO DE MATEMÁTICA APLICADA E ESTATÍSTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

SME0620 - Estatística I

Análise de Inferência Estatística

Professora: Amanda Morales Eudes D´Andrea

Antônio Sebastian Fernandes Rabelo - 10797781
Gabriell Tavares Luna - 10716400
Vitor Oliveira Caires - 10748027

São Carlos
2020

1. Conjunto de dados

Para análise de inferência estatística do conjunto de dados simulados que foi disponibilizado, foram escolhidas 2 informações (variáveis) relacionadas a 50 equipamentos. Os dados faltantes (XX) foram preenchidos de acordo com a tabela 1, portanto todos os dados estão completos.

Tabela 1 -Dados de uma amostra de equipamentos produzidos pela fábrica X.

Identificação do equipamento	Tipo	Custo de fabricação (R$)
1	B	153,7
2	B	163,7
3	B	140,0
4	B	155,2
5	C	155,8
6	C	162,8
7	B	154,7
8	C	149,1
9	B	169,7
10	C	161,4
11	A	142,7
12	B	149,0
13	A	152,8
14	B	151,3
15	B	139,3
16	A	165,9
17	B	154,6
18	C	138,7
19	A	166,0
20	B	156,1
21	A	151,7
22	B	153,2
23	B	163,1
24	A	141,8
25	A	140,4

26	A	132,2
27	B	152,3
28	C	131,7
29	A	173,2
30	B	145,5
31	B	158,9
32	B	157,4
33	C	165,3
34	A	143,7
35	B	157,1
36	A	138,5
37	A	163,8
38	C	153,2
39	B	145,6
40	A	150,7
41	A	133,2
42	B	154,4
43	C	144,6
44	B	139,3
45	A	147,8
46	B	165,1
47	A	135,7
48	B	151,9
49	C	146,0
50	B	153,5

2. Recursos

A análise dos dados foi feita utilizando o software R, versão 4.0.0 (2020-04-24) -- *"Arbor Day".Copyright (C) 2020 The R Foundation for Statistical Computing. Platform: x86_64-w64-mingw32/x64 (64-bit).*

Os comandos utilizados estão disponíveis no GitHub[1] juntamente com o conjunto de dados formatado em tabulações. A análise foi feita com o auxílio do pacote de interface 'R commander' (Rcmdr), que precisa ser instalado no software.

(1) Link para acessar GitHub: https://github.com/Sebastianfrabelo/Inferencia

3. Análise

A partir desse conjunto de dados faremos a inferência de dados da população por meio do **estimador pontual, estimador intervalar** e **do teste de hipóteses**.

Na estimação pontual, o valor de alguma estatística $T(x_1, x_2,...,x_n)$ representa, ou estima, o parâmetro desconhecido. O intervalo de confiança irá definir uma margem de erro que contenha o verdadeiro valor do variável, de acordo com a variância e com o coeficiente de confiança.

O teste de hipóteses é uma afirmação a ser testada sobre o(s) parâmetro(s) da distribuição de probabilidade de uma característica (variável), de acordo com o coeficiente de confiança e variância previamente determinados.

Definidos, para as duas análises, os níveis de significância e de confiança:

Nível de significância: $\alpha = 0,05$

Nível de confiança: $100(1 - \alpha)\% = 95\%$

Além disso, definimos o P-valor do teste: Se dá pelo percentual estatístico pelo qual avaliamos a rejeição ou não rejeição da hipótese nula. Ele é calculado a partir dos dados amostrais, e indica o menor nível de significância com que se rejeitaria a hipótese nula. Em outras palavras, para valores de P calculados menores que o nível de significância fixos do teste, devemos rejeitar tal hipótese.

3.1. "Tipo" - Variável qualitativa nominal

Sendo n = 50 > 30, aproximamos a variável analisada para uma distribuição normal, pois o tamanho da amostra é suficientemente grande para garantir o Teorema Central do Limite.

Analisando a tabela de frequência (Tabela 2), criada com o auxílio do software, decidimos considerar o **"Tipo B"** como **"Sucesso"** para o evento, podendo calcular sua proporção diretamente pelo software. Os tipos "A" e "B" são considerados insucesso.

Foi necessário inserir uma nova coluna no conjunto de dados para classificação de cada evento como **"Sucesso (B)"** ou **"Insucesso (F)"**.

Tabela 2 - Tabela de frequência da variável "Tipo".

Tipo	f_i	F_i	f_{r_i}	F_{r_i}
A	16	16	0,32	0,32
B	24	40	0,48	0,80
C	10	50	0,20	1
Total	50	-	1	-

f_i: frequência absoluta do "Tipo i".

F_i: frequência acumulada para o "Tipo i".

f_{r_i}: frequência relativa do "Tipo i".

F_{r_i}: frequência relativa acumulada do tipo i

3.1.1. Teste de Hipóteses

Nesse primeiro teste, vamos analisar a hipótese nula de que a proporção populacional, representada pela letra **"p"**, entre "Sucessos" e "Insucesso" seja:

$$\mathbf{p_0 = 0,5}$$

Assim, temos duas hipóteses:

Hipótese nula: $\mathbf{p = p_0}$

Hipótese alternativa: $\mathbf{p \neq p_0}$

3.1.2. Resolução no software R:

Usando a mesma aproximação para uma distribuição normal, com o auxílio da interface R commander, obtém-se o comando necessário para calcular a estimação pontual, estimação intervalar e o teste de hipótese, fornecendo o nível de confiança de 95%, apresenta os resultados:

Proporção amostral de sucessos: $p^* = 0{,}48$

Intervalo de confiança: $IC \cong [0.3479714 ; 0.6148826]$

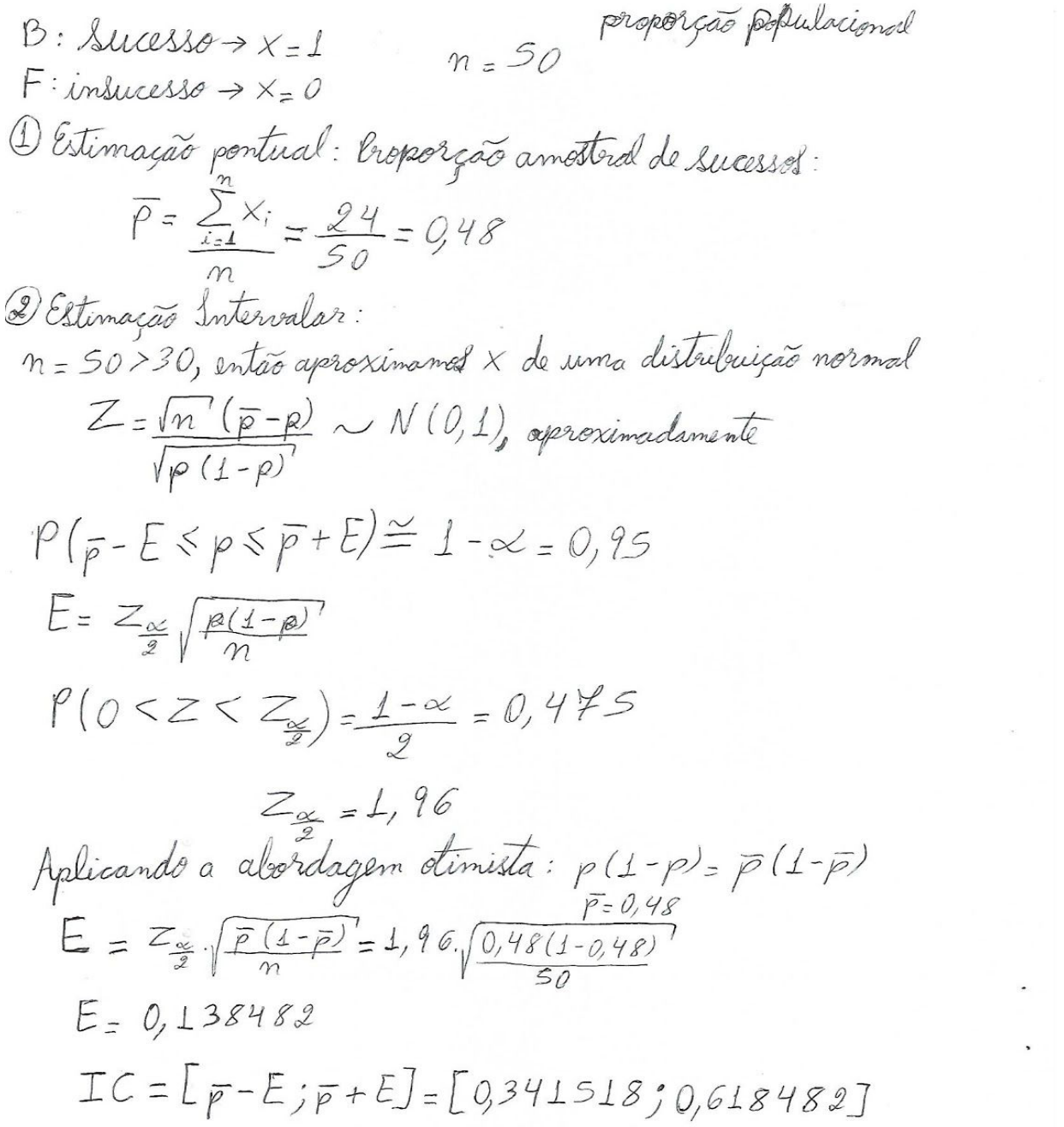
Teste de hipóteses:

Nível descritivo: $p_{value} = 0.7773$

$p_{value} > \alpha$

Interpretando, como $p > \alpha$, e acordo com os dados, ao nível de significância de 5%, **há evidências suficientes para aceitar a hipótese de que a proporção populacional é de 50%.**

3.1.3. Resolução manuscrita



③ Teste de Hipóteses $\quad p_0 = 0,5 \quad \bar{p} = 0,48$

① Formulação

$$\begin{cases} H_0: p = p_0 = 0,5 \\ H_1: p \neq 0,5 \end{cases}$$

② Estatística de teste: $n = 50 > 30$, portanto:

$$Z = \frac{\sqrt{n}(\bar{p} - p_0)}{\sqrt{p_0(1-p_0)}} = \frac{\sqrt{50}(\bar{p} - 0,5)}{\sqrt{0,5(1-0,5)}} \underset{\text{sob } H_0}{\sim} N(0,1), \text{ aproximadamente}$$

③ Região crítica:

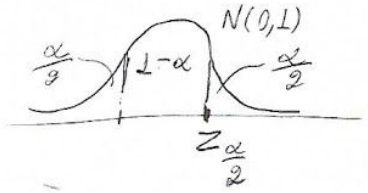
$$P(|Z| > Z_{\frac{\alpha}{2}}) = \alpha = 0,05$$

$$P(Z > Z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

$$P(0 < Z < Z_{\frac{\alpha}{2}}) = \frac{1}{2} - \frac{\alpha}{2} = 0,475$$

$$Z_{\frac{\alpha}{2}} = 1,96$$

$$R_c = \{|Z| > 1,96\} = \{Z < -1,96 \text{ ; } Z > 1,96\}$$



④ Decisão: $\bar{p} = 0,48$:

$$z = \frac{\sqrt{50}(0,48 - 0,5)}{\sqrt{0,5(1-0,5)}} = -0,2828$$

$z \in R_a$

$z \notin R_c$

Ao nível de significância de 5%, não há evidência para rejeitar a hipótese nula.

3.2. "Custo de Fabricação (em reais)" - Variável quantitativa contínua

Utilizamos o mesmo de nível de confiança e significância da análise anterior (5%).

Para o caso da variância da variável ser desconhecida, utilizaremos para análise uma distribuição "t de Student" com *n-1* graus de liberdade, portanto utilizaremos o teste T:

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{s} \sim t_{n\text{-}1}$$

Sendo:

$\overline{X}$ é a média amostral
μ é a média populacional testada (sob H_0)
s é o desvio padrão amostral
n é o tamanho da amostra

3.2.1. Teste de Hipóteses

Vamos analisar a hipótese nula de que a média populacional, representada por "μ", seja:

$$\mu_0 = 0{,}5$$

Temos uma hipótese nula simples e uma hipótese alternativa composta:

Hipótese nula: $\mu = \mu_0$

Hipótese alternativa: $\mu \neq \mu_0$

3.2.2. Resolução no software R:

Com o auxílio da interface R commander, obtém-se o comando necessário para calcular a estimação pontual, estimação intervalar e o teste de hipótese, fornecendo o nível de confiança de 95%, apresenta os resultados:

Proporção amostral de sucessos: $\overline{x} = 151{,}466$

Intervalo de confiança: IC (μ, 95%) ≅ [148.5616 ; 154.3704]

Teste de hipóteses:

Nível descritivo: $p_{value} = 0.01812$

$p_{value} < \alpha$

Interpretando a solução, como $p_{value} < \alpha$, podemos afirmar que de acordo com a análise dos dados ao nível de significância de 5%, **há evidências para rejeitar a hipótese nula H_0**, portanto não é possível afirmar que a média do *"Custo de Fabricação"* é igual a R$ 155,00.

3.2.3. Resolução manuscrita

$X \rightarrow$ custo de fabricação (em reais)

① Estimação pontual: $n = 50$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = 151,466$$

② Estimação Intervalar:

Considerando a variância desconhecida, trata-se de uma distribuição t de Student:

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t_{n-1}$$

$$E = t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \qquad S = 10,21964$$

$$P(\bar{X} - E \leq t \leq \bar{X} + E) = 1 - \alpha \qquad \alpha = 0,5 \quad n = 50$$

$$\therefore t_{\frac{0,5}{2}} = 2,009$$

$$E = 2,009 \cdot \frac{10,21964}{\sqrt{50}} = 2,903559$$

$$IC(\mu, 95\%) = [148,5624; 154,3696]$$

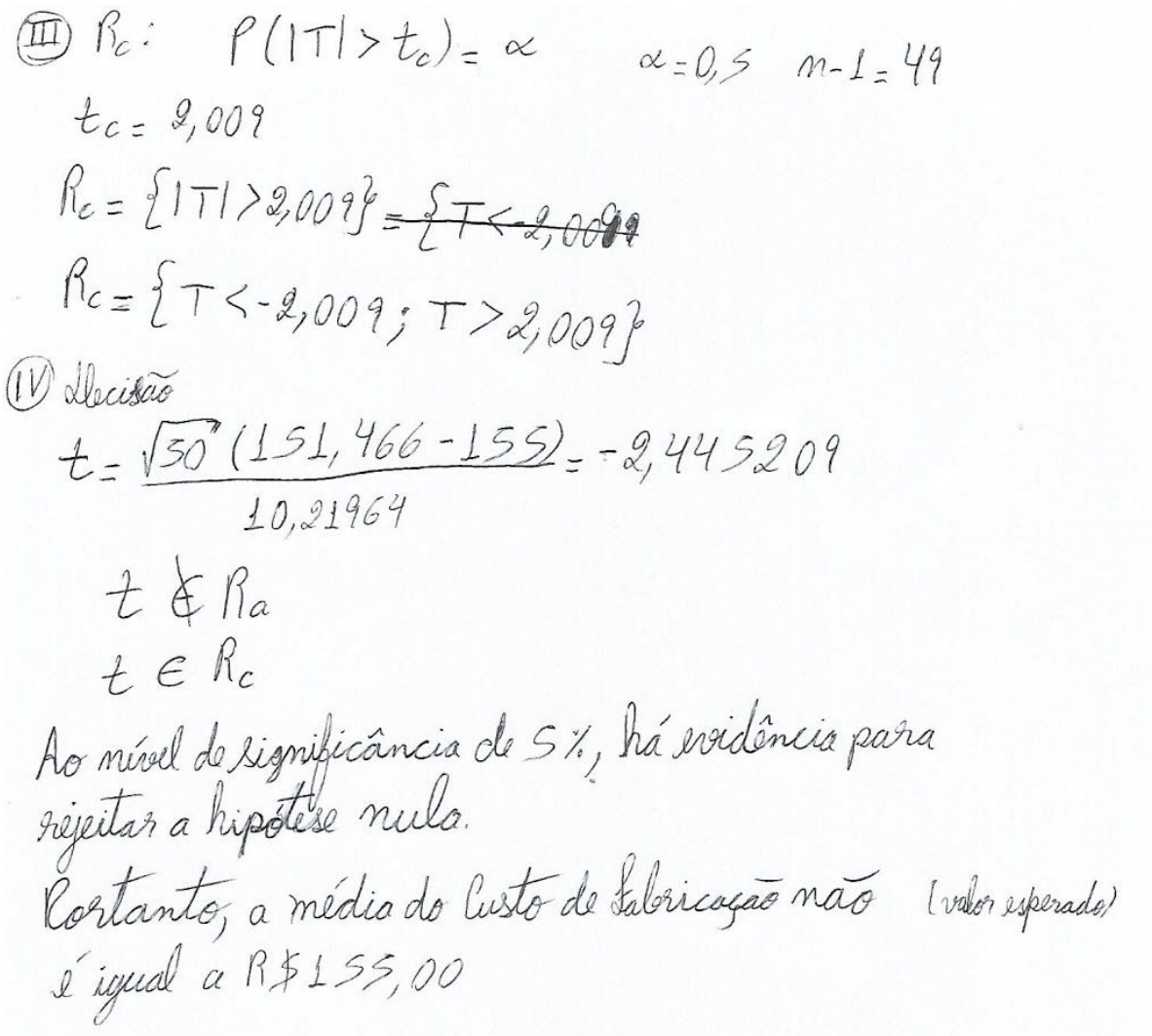
③ Teste de Hipóteses: $\mu_0 = 155 \qquad \bar{x} = 151,466$

Ⅰ $\begin{cases} H_0 : \mu = \mu_0 = 155 \\ H_1 : \mu \neq 155 \end{cases}$

Ⅱ Estatística de testes: t-student, pois $n > 30$, σ desconhecido

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \underset{\text{sob } H_0}{\sim} t_{n-1}$$

$$T = \frac{\sqrt{50} \cdot (\bar{X} - 155)}{10,21964} \underset{\text{sob } H_0}{\sim} t_{n-1}$$



3.2.4. Observações

A discordância entre os Intervalos de Confiança obtidos pelas diferentes resoluções deve-se ao fato do software utilizar um método diferente do utilizado na resolução manuscrita.

Para a variável "Tipo" é possível observar que o Intervalo de Confiança obtido na resolução pelo software é menor e está contido no obtido na resolução manuscrita. Conferindo confiabilidade para ambas as análises.

Para a variável "Custo em reais" é possível observar que o Intervalo de Confiança obtido na resolução manuscrita é menor e está contido no obtido na resolução pelo software. Conferindo também confiabilidade para ambas as análise, pois as estatísticas **Z** e **t-student**, ainda diferem um pouco para quantidade de 50 amostras.