INFERÊNCIA ESTATÍSTICA

William Luis Alves Ferreira - NºUSP 9847599

[1]Escola de Engenheira de São Carlos/Instituto de Ciências da Computação e Matemática
Universidade de São Paulo

william.luis.ferreira@usp.br

1. Introdução

Neste documento descreve-se a análise de inferência estatística sobre o conjunto de dados fornecidas para a 4º avaliação da disciplina Estatística I (SME0320) compondo a segunda avaliação semestral. Na seção 2 descreve-se o tratamento inicial das variáveis pertencentes ao conjunto de dados fornecidos, assim como no primeiro trabalho da disciplina (AV3), já na seção 3 e 4 realiza-se a analise solicitada.

2. Preparando dados

Do mesmo modo que no trabalho AV3, para a distinção do banco de dados foi solicitado a inserção aleatória de dados nas lacunas '*XX*', para isso inseriu-se os dados referente ao autor deste trabalho, sendo:

- **Idade:**22;
- **Salários Mínimos:**1;
- **Número de Filhos:**0;
- **Altura:**1.82;
- **Horas Trabalhadas:**50;
- **Peso:**80.1.

Por padrão a importação de dados para o RStudio padroniza a sintaxe dos dados com a pontuação da casa decimal no sistema norte americano, e outro ajuste necessário foi a marcação do tipo de dado que compõem cada variável, sendo:

```r
library(readxl)
Conjunto_de_dados <- read_excel("baseLocal/Conjunto de dados.xlsx",
    col_types = c("numeric", "numeric", "numeric",
        "numeric", "numeric", "text", "text",
        "numeric", "text", "numeric"), skip = 1)
View(Conjunto_de_dados)
```

Listing 1. Código fonte em R

Com os ajustes realizados e a verificação de todos os dados presentes consideramos a amostra como completa e suficiente para iniciar as analises das variáveis selecionadas: **Idade** (Quantitativa discreta) e **Gênero** (Qualitativa nominal); como alvo deste trabalho.

3. Gênero - Inferência da Variável Qualitativa

Como solicitado no enunciado desta avaliação tomaremos o valor "**Feminino**"como sucesso para esta variável, e verifica-se as seguintes características e parâmetros inferidos da amostra:

1. Calcule a estimativa para a proporção de sucessos na população;
2. Faça um intervalo de confiança para a proporção de sucessos na população;
3. Teste a hipótese de que a proporção de sucessos é (ou não) igual a 50%.

Os valores da V.A Gênero (X) são:

M, M, M, F, F, M, M, M, M, F, M, F, F, M, M, F, M, F, F, M, F, M, M, F, F, M, F, M, M, M, F, F, M, F, M, M, F, M, M, M, M, M, M, M, M, F, M, F, M, M

onde M = "Masculino" e F = "Feminino"

Valores	Frequência Absoluta	Frequência Relativa
Feminino	18	0.36
Masculino	32	0.64
Total	50	1.00

Tabela 1. Frequência dos valores da variável Gênero.

Através da discriminação de sucesso perante a variável aleatória Gênero (X) podemos tomar a distribuição de n ensaios de Bernoulli como $X \sim B(n, p)$ para calcularmos o *item 1*. Para isso, devemos encontrar o valor do Estimador Pontual do parâmetro p da população pela função da amostra $\bar{p}$ através do método de substituição, no qual:

$$\bar{p} = \frac{1}{n} \sum_{i=1}^{n} X_i, n = 50$$

No qual X=1 sucesso e X=0 fracasso, temos

$$\sum_{i=1}^{n} X_i = 18$$

Uma outra alternativa seria encontrar um estimador pontual ($\hat{p}$) para o parâmetro p através do método de máxima verossimilhança através da função densidade $f(.; p)$, porém como temos uma distribuição binomial $X \sim B(n, p)$ bem definido e utilizando o método da substituição temos o suficiente para o encontrar a estimativa de p inferida a partir de $\bar{p}$.

Logo o estimador pontual ($\hat{p}$) igual a $\bar{p}$ do parâmetro p possui valor estimado de 0.36.

Ressaltamos que possuímos apenas uma amostra da população e desejamos inferir um parâmetro da população através de uma função da amostra, com esse intuito é necessário que calculemos a aproximação da distribuição de $\bar{p}$ para definirmos o Intervalo de Confiança (IC) *item 2*. Considerando uma amostra grande suficiente ($n >= 30$) ou $X \sim N(\mu, \sigma^2)$, vale a expressão e a aproximação da distribuição de $X \sim B(n, p)$

para $X \sim N(np, npq)$ e posteriormente transformado em $Z \sim N(0,1)$ munindo-se do Teorema Central do Limite temos as seguintes expressões:

Aproximação pela Teorema Central do Limite:

$$X \sim Binomial(n,p) : X \approx N(np, np(1-p)), n \to \infty$$

Pela transformação da normal para normal padrão:

$$Z = \frac{\sqrt{n}(\bar{p}-p)}{\sqrt{p(1-p)}} \approx N(0,1) \tag{1}$$

Por fim, através da aproximação da distribuição de $\bar{p}$ em Z como na equação 1, e como fornecido no enunciado da atividade $\alpha = 95\%$ e considerando a tabela de probabilidades da Normal Padrão, temos:

$$P\left(-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(-Z_{\alpha/2} \leq \frac{\sqrt{n}(\bar{p}-p)}{\sqrt{p(1-p)}} \leq Z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(\bar{p} - Z_{\alpha/2}\sqrt{\frac{p(1-p))}{n}} \leq p \leq \bar{p} + Z_{\alpha/2}\sqrt{\frac{p(1-p))}{n}}\right) = 1-\alpha$$

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{pq}{n}} \tag{2}$$

$$\bar{p} \pm E, E = z_{\alpha/2}\sqrt{\frac{pq}{n}}$$

$$IC = [\bar{p} - E; \bar{p} + E]$$

Com isso, e considerando que p da população é desconhecida e nenhum estudo piloto foi realizado, assim escolhe-se uma das abordagem a baixo:

Abordagem Otimista

Substituindo através de $pq = p(1-p) = \bar{p}(1-\bar{p})$

$$IC \cong [\bar{p} - Z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p}))}{n}}; \bar{p} + Z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p}))}{n}}]$$

Abordagem Conservativa

Substituindo através de $pq = p(1{-}p) = \frac{1}{4}$, que corresponde ao valor máximo que $p(1{-}p)$ pode assumir.

$$IC \cong [\bar{p} - Z_{\alpha/2} * \frac{1}{\sqrt{4n}}; \bar{p} + Z_{\alpha/2} * \frac{1}{\sqrt{4n}}]$$

Como nenhuma abordagem foi especificada pelo enunciado desta avaliação toma-se a abordagem conservativa, com isso, apresenta-se os cálculos a seguir considerando a tabela dos valores de probabilidades da Normal Padrão:

$$IC \cong [0.36 - Z_{\alpha/2} * \frac{1}{\sqrt{4*50}}; 0.36 + Z_{\alpha/2} * \frac{1}{\sqrt{4*50}}]$$

Consultando a tabela da Normal Padrão temos que $P(z >= 1.96) = \alpha/2 = 0.475$, logo:

$$IC \cong [0.36 - 1.96 * \frac{1}{\sqrt{4*50}}; 0.36 + 1.96 * \frac{1}{\sqrt{4*50}}]$$

$$E \approx 0.1386$$

Com tudo, temos que o intervalo de confiança com $\alpha = 95\%$ é $IC \cong [0.2214; 0.4986]$ com estimativa pontual de 0.36.

Em R, temos:

```
### Trabalhando com a v.a genero
genero <- Conjunto_de_dados$Genero
tabelaexport(genero) #para o relatorio
intervaloConfP <- function(x, conf = 0.95) {
  n <- length(x)
  proporcao <- prop.table(table(x))
  proporcao <- proporcao[1] #Feminino

  qn <- qnorm((1 - conf)/2,mean=0,sd=1)
  ic <- c(proporcao+ sqrt(1/(4*n))*qn, proporcao -sqrt(1/(4*n))*qn)
  return(ic)
}
intervaloConfP(genero)
-------------------------------------------------
> intervaloConfP(genero)
0.2214096 0.4985904
```

Listing 2. Código fonte em R

Por fim, descreve-se a seguir o Teste de Hipótese do *item 3*.

1) Hipótese Semântica

H_0: A proporção de sucessos é igual a 50%.

H_1: A proporção de sucessos não é igual a 50%.

2) Hipótese estatística

A estatística de teste utilizando o estimador pontual $\bar{p}$ com $X \approx N(np, np(1-p))$ considerando a amostra grande o suficiente ($n > 30$) pelo Teorema Central do Limite podemos adotar e expressão 2 e aproximação da distribuição 1, com isso, defini-se a hipótese estatística:

$$H_0 : p = 0.50$$

$$H_1 : p \neq 0.50$$

Configurando-se como Teste de Hipotese bilateral.

3) Desenvolvimento do teste

Com o enunciado desta avaliação considera-se o nível de significância de $\alpha = 5\%$, buscamos definir a região critica, para isso, consideramos a equações 2 e 1 para encontrar $R_c = \{|\bar{p}| > k\}$ onde k delimita a região critica. Primeiro, verificamos as relações entre a distribuição aproximada e a probabilidade associada ao Erro Tipo I, como segue:

$$P(Erro : Tipo(I)) = P(Rejeita : H_0; H_0 : verdadeiro) = \alpha$$

Queremos verificar qual o valor de k define o intervalo do qual exclui o falso negativo, ou seja, quando $\bar{X}$ representa rejeitar H_0 sendo verdadeiro, resultando na expressão $P(|\bar{X}| > k; H_0 : p = p_0)$. Com tudo, podemos utilizar da aproximação da distribuição do estimador pontual de p para encontrar a região crítica e de aceitação através das seguintes expressões:

$$P\left(|\bar{X}| > k\right) = P\left(|Z| > \frac{\sqrt{n}(k-p)}{\sqrt{p(1-p)}}\right) = \alpha$$

$$P\left(|Z| > Z_\alpha\right) = \alpha$$

$$\frac{\sqrt{n}(k-p)}{\sqrt{p(1-p)}} = Z_\alpha \Rightarrow k = p + Z_\alpha \times \sqrt{\frac{p(1-p)}{n}}$$

Pela abordagem conservativa, temos

$$R_c = \left\{ |\bar{p}| > p_0 + Z_\alpha \times \sqrt{\frac{1}{4n}} \right\}$$

Logo, com o formulário estabelecido desenvolve-se os cálculos:

$$k = 0.36 + 1.96 \times \sqrt{\frac{1}{200}} \approx 0.1386$$

$$R_c = \{|\bar{p}| > 0.1386\}$$

Por fim, a região de aceitação é encontrada sendo $R_0 = [0.2214; 0.4986]$.

4) Aplicando Teste de Hipótese

Com o desenvolvimento dos tópicos 1 a 3 conclui-se que $0.50 \notin R_0$, assim rejeita-se a hipótese nula, e de acordo com os dados fornecidos e adotando um nível de significância de 5% e abordagem conservativa, concluí-se que a proporção de sucessos não é de 50%.

4. Idade - Inferência da Variável Quantitativa

Com a variável idade selecionada verificaremos as seguintes características e parâmetros:

1. Calcule a estimativa para a média populacional dessa variável;
2. Faça um intervalo de confiança para a média populacional;
3. Teste a hipótese de que a média populacional é (ou não) igual a 35.

O desenvolvimento dos seguintes itens possuem ferramentas estatísticas e semântica similar aos apresentados na seção 3. Para o calculo da estimativa pontual (*item 1*) do parâmetro μ da população através do método da substituição, usa-se as relações:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, n = 50 \tag{3}$$

$$\sum_{i=1}^{n} Y_i = 1693$$

Novamente, uma alternativa seria utilizar o método da máxima verossimilhança, porém temos que a distribuição possui tamanho suficiente para a aproximação da distribuição de $\bar{Y}$ como exposto na expressão 4 para fazer uso do método da substituição.

Logo o valor do estimador pontual $\hat{\mu}$ através de $\bar{Y}$ do parâmetro μ da população é de 33.86.

Com isso, verifica-se que a variável aleatória Idade (Y) possui distribuição desconhecida e amostra grade o suficiente para aproximação pela Teoria Central do Limite (TCL), logo se $n > 30$ podemos aproximar como:

$$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n}) \tag{4}$$

Pela transformação da normal para normal padrão:

$$Z = \frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma} \sim N(0, 1) \tag{5}$$

Como a variância da população é desconhecida, utilizaremos a aproximação de t-student:

$$T = \frac{\sqrt{n}(\bar{Y} - \mu)}{S} \sim t_{n-1} \tag{6}$$

Sendo S = desvio padrão da amostra.

Considerando que a variância da população é desconhecida, e tamanho da amostra suficiente para aplicar a aproximação da distribuição de $\bar{Y}$ pelo TCL, com isso, apresenta-se as relações entre as probabilidade da distribuição aproximada à t-student e os valores de L e U:

$$P\left(-t_{\alpha/2,49} \leq T \leq t_{\alpha/2,49}\right) = 1 - \alpha$$

$$P\left(-t_{\alpha/2,49} \leq \frac{\sqrt{n}(\bar{Y} - \mu)}{S} \leq t_{\alpha/2,49}\right) = 1 - \alpha$$

$$P\left(\bar{Y} - t_{\alpha/2,49}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{\alpha/2,49}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

$$IC = [L; U] = [\bar{Y} - E; \bar{Y} + E], E = t_{\alpha/2,49}\frac{S}{\sqrt{n}}$$

Por fim, calcula-se o IC assumindo as expressões anteriores e com uso da tabela de probabilidades de t-student:

$$P(t >= 2.009) = \alpha/2 = 0.475 \Rightarrow t_{\alpha/2,49} = 2.009$$

$$IC = [L; U] = [33.86 - 2.009\frac{S}{\sqrt{50}}; 33.86 + 2.009\frac{S}{\sqrt{50}}]$$

$$S = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n - 1}} = \sqrt{\sum_{i=1}^{50} \frac{(y_i - 33.86)^2}{50 - 1}} = \sqrt{22.3269} \approx 4.7251$$

$$IC = [L; U] = [33.86 - 2.009\frac{4.7251}{\sqrt{50}}; 33.86 + 2.009\frac{4.7251}{\sqrt{50}}]$$

$$E \approx 1.3424$$

Com tudo, temos que o intervalo de confiança com 95% de confiança é $IC \approx [32.51; 35.20]$ com estimador pontual de 33.86.

Em R temos:

```r
### Trabalhando com a v.a idade
idade <- Conjunto_de_dados$Idade

intervaloConfM <- function(x, conf = 0.95) {
  n <- length(x)
  media <- mean(x)
  variancia <- var(x)
  quantis <- qt(c((1 - conf)/2, 1 - (1 - conf)/2), df = n - 1)
  ic <- media + quantis * sqrt(variancia/n)
  return(ic)
}
intervaloConfM(c(idade))
------------------------------------------------------------
> intervaloConf(c(idade))
[1] 32.51713 35.20287
```

Listing 3. Código fonte em R

Por fim, descreve-se a seguir o Teste de Hipótese do *item 3*.

1) Hipótese Semântica

H_0: A média populacional é igual a 35 anos.

H_1: A média populacional não é igual a 35 anos.

2) Hipótese estatística

A estatística de teste utilizando o estimador pontual $\bar{Y} \approx N(\mu, \frac{\sigma^2}{n})$ considerando a amostra grande o suficiente ($n > 30$) pelo Teorema Central do Limite podemos adotar a expressão 3 e aproximação da distribuição 6, com isso, defini-se a hipótese estatística:

$$H_0 : \mu = 35$$

$$H_1 : \mu \neq 35$$

Configurando um teste de hipótese bilateral.

3) Desenvolvimento do teste

Com o enunciado desta avaliação considera-se o nível de significância de $\alpha = 5\%$, buscamos definir a região critica, para isso, consideramos a equações 3 e 6 para encontrar $R_c = \{|\bar{Y}| > k\}$ onde k delimita a região critica. Primeiro, verificamos as relações entre a distribuição aproximada e a probabilidade associada ao Erro Tipo I, como segue idem a seção 3:

$$P(Erro : Tipo(I)) = P(Rejeita : H_0; H_0 : verdadeiro) = \alpha$$

Queremos verificar qual o valor de k define o intervalo do qual exclui o falso negativo, ou seja, quando $\bar{Y}$ representa rejeitar H_0 sendo verdadeiro, resultando na expressão $P(|\bar{Y}| > k; H_0 : \mu = \mu_0)$. Com tudo, podemos utilizar da aproximação da distribuição ($\bar{Y}$) do estimador pontual ($\hat{\mu}$) para encontrar a região crítica e de aceitação através das seguintes expressões:

$$P\left(|\bar{Y}| > k\right) = P\left(|T| > \frac{\sqrt{n}(k - \mu_0)}{S}\right) = \alpha$$

$$P\left(|T| > t_{\alpha/2,49}\right) = \alpha$$

$$\frac{\sqrt{n}(k - \mu_0)}{S} = t_{\alpha/2,49} \Rightarrow k = \mu_0 + \frac{t_{\alpha/2,49} \times S}{\sqrt{n}}$$

$$R_c = \left\{\mu_0 - \frac{t_{\alpha/2,49} \times S}{\sqrt{n}} > \bar{Y}; \bar{Y} > \mu_0 + \frac{t_{\alpha/2,49} \times S}{\sqrt{n}}\right\}$$

$$R_0 = \left\{\mu_0 - \frac{t_{\alpha/2,49} \times S}{\sqrt{n}} \leq \bar{Y} \leq \mu_0 + \frac{t_{\alpha/2,49} \times S}{\sqrt{n}}\right\}$$

Logo, com o formulário estabelecido desenvolve-se os cálculos:

$$R_0 = \left\{33.86 - \frac{2.009 \times 4.7251}{\sqrt{50}} \leq \bar{Y} \leq 33.86 + \frac{2.009 \times 4.7251}{\sqrt{50}}\right\}$$

Por fim, a região de aceitação encontrada é $R_0 = [32.51; 35.20]$

4) Aplicando Teste de Hipótese

Com o desenvolvimento dos tópicos 1 a 3 conclui-se que $35 \in R_0$, aceita-se a hipótese nula, logo de acordo com os dados fornecidos e adotando um nível de significância de 5%, concluí-se que a média populacional é 35 anos.

5. Script R

Nesta seção apresenta o *script* desenvolvido para este trabalho que fomentou as tabelas e figuras utilizadas neste relatório em Latex - Overleaf. Todos os arquivos resultantes deste trabalho podem ser acessos em illiamw/SME0320_AV4

```r
#data
require(xtable)

library(readxl)
Conjunto_de_dados <- read_excel("./Conjunto de dados.xlsx",
```

```r
                                        col_types = c("numeric", "numeric", "numeric",
                                                      "numeric", "numeric", "text", "text",
                                                      "numeric", "text", "numeric"), skip = 1)
View(Conjunto_de_dados)
Conjunto_de_dados

##Frequencia tabela
tabelaexport <- function(x){
  freq <- table(x)
  freq_abs<-data.frame(freq)
  freq_rel<-data.frame(prop.table(freq))
  xtable(data.frame(VAlores=  freq_abs$x,
                    FrequenciaAbsoluta = freq_abs$Freq,
                    FrequenciaRelativa = freq_rel$Freq),caption = "title")
}

### Trabalhando com a v.a genero
genero <- Conjunto_de_dados$G nero
tabelaexport(genero) #para o relat rio
intervaloConfP <- function(x, conf = 0.95) {
  n <- length(x)
  proporcao <- prop.table(table(x))
  proporcao <- proporcao[1] #Feminino

  qn <- qnorm((1 - conf)/2,mean=0,sd=1)
  ic <- c(proporcao+ sqrt(1/(4*n))*qn, proporcao -sqrt(1/(4*n))*qn)
  return(ic)
}
intervaloConfP(genero)


### Trabalhando com a v.a idade
idade <- Conjunto_de_dados$Idade

intervaloConfM <- function(x, conf = 0.95) {
  n <- length(x)
  media <- mean(x)
  variancia <- var(x)
  quantis <- qt(c((1 - conf)/2, 1 - (1 - conf)/2), df = n - 1)
  ic <- media + quantis * sqrt(variancia/n)
  return(ic)
}
intervaloConfM(c(idade))
```

Listing 4. Código fonte em R