2º Estágio / Trabalho Avaliativo

Ítallo Silva - 118110718 | Thiago Nascimento - 118110804 | João Marcelo Junior - 117110448

Questão 1

Um teste de hipótese é uma técnica estatística que nos ajuda na tomada de decisão sobre uma determinada característica de uma população com base em uma amostra desta. Nele formulamos duas hipóteses, que devem ser mutuamente exclusivas, sobre a característica estudada: hipótese nula e hipótese alternativa. Com base nos valores amostrais da característica, tentamos encontrar evidências que comprovem ou rechacem a hipótese nula, neste último caso comprovando a hipótese alternativa.

Por exemplo, desejamos testar se o tempo de resposta de um novo medicamento é inferior ao de outro que já se encontra no mercado. Podemos investigar isso através do tempo médio de resposta. Formulamos então as hipóteses nula e alternativa, respectivamente: o tempo médio do novo medicamento é maior ou igual ao do antigo e o tempo médio do novo medicamento é menor que o do antigo. E utilizando técnicas estatísticas tomamos uma decisão, considerando um certo nível de confiança.

Questão 2

(2) Probabilidade de Rejeitar a Hipótese nula (H_0) quando a mesma é verdadeira.

(4) Não rejeitar a Hipótese nula (H_0) quando a mesma é falsa.

(5) Probabilidade de Rejeitar a Hipótese nula (H_0) quando esta é de fato falsa.

(3) Valor p ou Nível descritivo do Teste ($\hat{\alpha}$).

(1) Rejeitar a Hipótese nula (H_0) quando a mesma é verdadeira.

Questão 3

(2) Chega-se à conclusão por não-rejeitar a hipótese nula (H_0).

(4) Rejeita-se a hipótese nula para qualquer nível de significância maior que 0,034; por exemplo: $\alpha = 0,05$.

(1) Deve-se rejeitar a hipótese nula (H_0).

(3) Diz-se que o teste é altamente significativo pelo fato de que qualquer nível de significância, α, já ser suficientemente maior que o valor p ou; de outra forma; o valor p é tão pequeno que o mesmo será suficientemente menor que qualquer nível de significância α adotado, mesmo que este seja pequeno. Neste caso a hipótese nula deve ser rejeitada com forte evidência estatística (fornecida pelos dados).

Questão 4

a)

Queremos testar a hipótese de que o valor média da característica populacional é igual a 45. Como conhecemos a variância, fazemos o teste de hipóteses com variância conhecida. Temos as seguintes hipóteses:

$$H_0 : \mu = 45$$
$$H_1 : \mu \neq 45$$

Pelo enunciado temos:

$n = 16$, $\overline{x} = 43$, $\sigma = 6$, $\alpha = 0,1$

$$Z_{obs} = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{43 - 45}{\frac{6}{\sqrt{16}}} = -1,34$$

Com o Z_{obs} calculado, precisamos achar o intervalo da região crítica. Como o teste é bilateral, então o valor crítico é de acordo com $\frac{\alpha}{2} = 0,05$. Pela tabela da distribuição normal, temos que o valor crítico $z_c = -1,64$ e $z_c = 1,64$

O intervalo da região crítica é $RC = (-\infty; -1,64] \cup [1,64; +\infty)$

Sendo assim, podemos observar que $Z_{obs} \notin RC \rightarrow aceita\ H_0$. Portanto, ao nível de 10% de signifância há evidências que é possível aceitar a afirmação sobre o valor médio.

b)

Como o teste é bilateral, para encontrar o valor-p $\hat{\alpha}$ temos:

$\frac{\hat{\alpha}}{2} = P(Z < -1,34)$

Utilizando a tabela da distribuição normal, temos: $\hat{\alpha} = 2 * 0,0901 = 0,1802$

Portanto, para níveis de confiança acima de 18,02% conclui-se pela rejeição de H_0. Como o valor-p é razoavelmente grande, concluimos pela não rejeição de H_0 com bastante evidência.

Questão 5

a)

Queremos testar se a proporção de dias violentos foi menor que 60%, dessa forma as hipóteses são:

$H_0 : p \geq 0,6$
$H_1 : p < 0,6$

Pelo enunciado temos:

$\overline{p} = 0,33$, $n = 30$

$$z_{obs} = \frac{\overline{p} \ - \ p}{\sqrt{\frac{p\ (1\ -\ p)}{n}}} = \frac{0,33\ -\ 0,6}{\sqrt{\frac{0,6\ *\ 0,4}{30}}} = \frac{-0,27}{\sqrt{\frac{0,24}{30}}} = -3,03$$

Para o nível de significância de 1%, ou seja, $\alpha = 0,01$, temos pela tabela normal o valor crítico $z_c = -2,33$ e o intervalo da região crítica $RC = (-\infty; -2,33]$. Como $z_{obs} \in RC$ rejeitamos H_0, e concluimos que para esse nível de significância há evidências de que o plano emergencial surtiu efeito positivo.

Para o nível de significância de 5%, ou seja, $\alpha = 0,05$, temos pela tabela normal o valor crítico $z_c = -1,64$ e o intervalo da região crítica $RC = (-\infty; -1,64]$. Como $z_{obs} \in RC$ rejeitamos H_0, e concluimos que para esse nível de significância há evidências de que o plano emergencial surtiu efeito positivo.

b)

Sendo $\hat{\alpha}$ o valor-p, temos que $\hat{\alpha} = P(Z < -3,03)$, e pela tabela de distribuição normal $\hat{\alpha} = 0,0012$. Assim, para qualquer nível de significância acima de 0,12% há evidências que permite concluir que o plano surtiu efeito positivo. E sendo o valor-p muito pequeno podemos considerar que o teste é altamente significativo.

c)

```
number_success = 10
tam_amostra = 30

#Teste para 5% de significancia
prop.test(x = number_success, n = 30, p = 0.6, alternative = "less", correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  number_success out of 30, null probability 0.6
## X-squared = 8,9, df = 1, p-value = 0,001
## alternative hypothesis: true p is less than 0,6
## 95 percent confidence interval:
##  0,0000 0,4834
## sample estimates:
##      p
## 0,3333
```

Vemos que o valor do valor-p bem próximo ao obtido manualmente, assim reforçando a conclusão do item anterior.

Questão 6

O teste de hipótese se baseia na afirmação que o candidato A possui mais de 53% dos eleitores favoráveis a ele. Dessa forma as hipóteses que temos são:

$H_0 : p \leq 0,53$ $H_1 : p > 0,53$

Pelo enunciado temos que:

$n = 150$, $\overline{p} = 0,63$

Sendo assim:

$$z_{obs} = \frac{\overline{p} - p}{\sqrt{\frac{p\ (1 - p)}{n}}} = \frac{0,63 - 0,53}{\sqrt{\frac{0,53 * 0,47}{150}}} = 2,45$$

Calculando o valor-p:

$\hat{\alpha} = P(Z > 2,45) = 1 - P(Z < 2,45) = 1 - 0,9929 = 0,0071 = 0,71\%$

Sendo o valor-p igual a 0,71%, para níveis de significância acima dessa porcentagem há evidências que é possível fazer a afirmação do candidato A. E podemos dizer que o teste é altamente significativo.

Questão 7

Temos que, as hipóteses se configuram como:

$H_0 : \mu \geq 5$
$H_1 : \mu < 5$

Os dados sendo:

```
dados = c(2.9, 3.4, 3.5, 4.1, 4.6, 4.7, 4.5, 3.8, 5.3, 4.9,
          4.8, 5.7, 5.8, 5.0, 3.4, 5.9, 6.3, 4.6, 5.5, 6.2)
```

i) Podemos calcular o valor de p através de:

```
mu_a = mean(dados)
sd_a = sd(dados)

t_obs = (mu_a - 5)/(sd_a / sqrt(length(dados)))

p_value = pt(t_obs, length(dados) - 1)

p_value
```

```
## [1] 0,1332
```

ii) Com auxílio de funções R, podemos calcular p através de:

```
t.test(dados, alternative = "less", var.equal=TRUE, mu=5)
```

```
##
##  One Sample t-test
##
## data:  dados
## t = -1,1, df = 19, p-value = 0,1
## alternative hypothesis: true mean is less than 5
## 95 percent confidence interval:
##  -Inf 5,13
## sample estimates:
## mean of x
##     4,745
```

Sendo assim, temos que, o valor p é relativamente alto, logo caso seja superior ao nível de significância, não rejeitamos H_0.

iii) Uma vez que $p = 0.1332$, com nível de significância de 5%, temos que $p > 5\%$, podemos afirmar com forte indicio a não rejeição de H_0. Com nível de significância de 1%, a condição $p > 1\%$, permace e ainda podemos afirmar com forte indicio a não rejeição H_0. De forma geral, os resultados não pareciam evidentes, principalmente pelo fato da $\bar{x}$ ser inferior a 5.

Questão 8

Sabendo que as hipóteses:

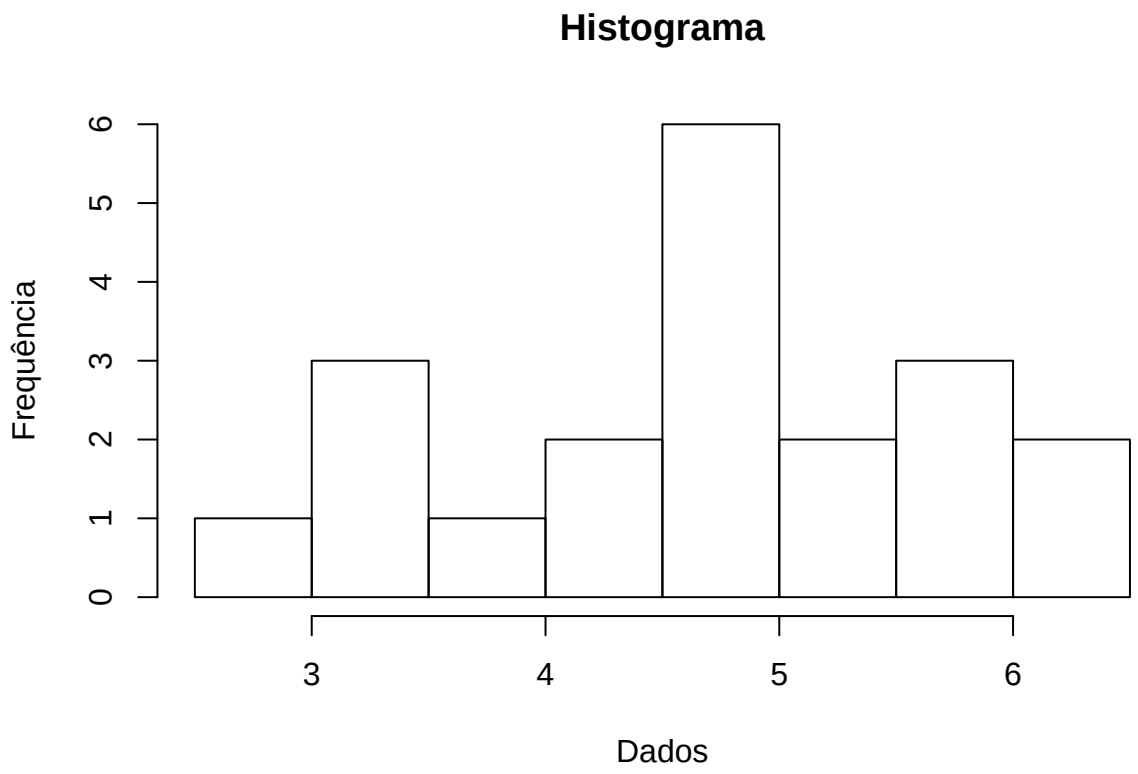
H_0 : os dados provêm de uma distribuição normal H_1 : os dados não têm distribuição normal

Os dados sendo:

```
dados = c(2.9, 3.4, 3.5, 4.1, 4.6, 4.7, 4.5, 3.8, 5.3, 4.9,
          4.8, 5.7, 5.8, 5.0, 3.4, 5.9, 6.3, 4.6, 5.5, 6.2)
```

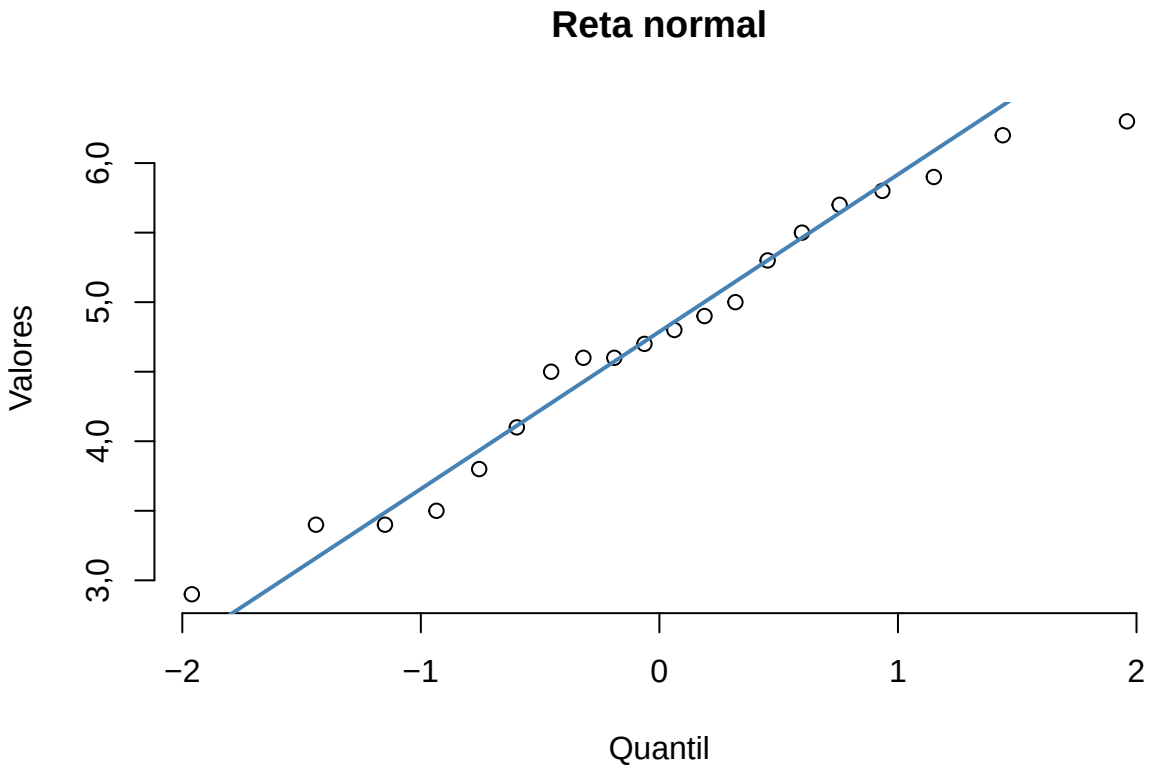
Temos que o gráfico se mostra:

```
hist(dados, xlab="Dados", ylab="Frequência", main="Histograma")
```



Além disso, podemos plotar de formar mais evidente a existência de normalidade através do Q-Q plot.

```
qqnorm(dados, pch = 1, frame = FALSE, main="Reta normal", xlab="Quantil", ylab="Valores")
qqline(dados, col = "steelblue", lwd = 2)
```



O histograma em formato similar ao de um sino e o grande número de pontos próximos a reta de 45º do Q-Q plot aponta para a existência de normalidade, mas é preciso testarmos para sua confirmação:

```
shapiro.test(dados)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dados
## W = 0,96, p-value = 0,6
```

O p-valor foi de 0.5986. Sendo assim, quer dizer que que há forte indicio de que os dados são normais, pois não diferem de uma curva normal.

Questão 9

Leitura dos dados

```
amamentacao_dataset <- read_xlsx('../02_Dados_Amamentacao_Cancer_para_Exportar.xlsx')
head(amamentacao_dataset)
```

```
## # A tibble: 6 x 3
```

```
##      id cancer amamentacao
##   <dbl>  <dbl>       <dbl>
## 1     1      1           1
## 2     2      1           1
## 3     3      1           1
## 4     4      1           1
## 5     5      1           1
## 6     6      1           1
```

Vemos que existem três colunas: id, cancer e amamentacao. A primeira indica o id da observação, a segunda é uma variável binária indicando a ocorrência de câncer (1) ou não (0), a terceira é também uma variável binária indicando se a mulher foi amamentada pela mãe (1) ou não (0).

O teste de hipótese

Deseja-se investigar se o fato de ter sido amamentada pela mãe é um fator de proteção para o câncer de mama. Para tanto, seja p_a a proporção de mulheres com câncer dentre as foram amamentadas e p_{na} a proporção de mulheres com câncer dentre as que não foram amamentadas. Vale ressaltar que ambas as populações são independentes.

Podemos então formular as hipóteses:

$$H_0 : p_a = p_{na} \rightarrow H_0 : p_a - p_{na} = 0$$
$$H_1 : p_a \neq p_{na} \rightarrow H_1 : p_a - p_{na} \neq 0$$

Separando as populações, temos:

```
pop_ama <- amamentacao_dataset %>% filter(amamentacao == 1) %>% pull(cancer)

pop_nama <- amamentacao_dataset %>% filter(amamentacao != 1) %>% pull(cancer)
```

Temos que as proporções na amostra, que denotaremos por, $\hat{p}_a$ e $\hat{p}_{na}$ são respectivamente dadas por:

```
p_a <- length(which(pop_ama == 1)) / length(pop_ama)
p_na <- length(which(pop_nama == 1)) / length(pop_nama)
```

$\hat{p}_a = 0{,}4401$ e $\hat{p}_{na} = 0{,}5335$.

A diferença estimada entre as probabilidades é $\hat{p}_a - \hat{p}_{na}$ = -0,0934, e o desvio padrão é dado por:

$$\sigma(\hat{p}_a - \hat{p}_{na}) = \sqrt{\frac{0.44 \times 0.56}{802} + \frac{0.534 \times 0.466}{328}} = 0,033$$

Calculemos então o z_{obs}, para tanto precisamos da proporção total da amostra, daqueles que tiveram câncer:

$$\hat{p} = \frac{353 + 175}{802 + 328} = \frac{528}{1130} = 0.467$$

E então o z_{obs}:

$$z_{obs} = \frac{0.44 - 0.534}{\sqrt{0.467(1 - 0.467)\left[\frac{1}{802} + \frac{1}{328}\right]}} = -2.875$$

Para encontrar o p-value, como é um teste bilateral procuraremos pela probabilidade $2P(Z \geq |z|) = 2P(Z \geq 2.875)$, calculemos:

```
p.value <- 2 * (1 - pnorm(2.875))
```

Temos então que o valor p é de $0,004$ ou $0,4\%$. Como o p-value encontrado é muito baixo, podemos com certa significância rejeitar a hipótese nula, sendo assim a amostra fornece evidências de que a amamentação protege contra o câncer de mama.

Por fim, vamos utilizar a função pronta do teste Chi-Quadrado para estimar o valor p.

Usando o teste Chi-Quadrado

```
chisq.test(amamentacao_dataset$amamentacao, amamentacao_dataset$cancer, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  amamentacao_dataset$amamentacao and amamentacao_dataset$cancer
## X-squared = 8,2, df = 1, p-value = 0,004
```

Como podemos ver, o valor p obtido pelo teste é bem próximo a nossa estimativa manual, assim corroborando nossa rejeição a hipótese nula.