Análise de Dados - UFPE/2019 - Lista 5

Antonio Fernandes

14 de maio de 2019

Conteúdo

Apresentação

Este documento apresenta as respostas dos exercícios realizados à mão presentes nas listas 1 e 2. Os scripts estão no repositório do GitHub.

Questão 5 - Lista I

Primeiramente foi criado o *Data frame* com as informações necessárias:

```
Emp <- data.frame("EMP" = 1:15, "MES" = c(8,9,4,5,3,6,8,6,6,8,5,5,6,4,4),
                  "SET" = c('C','C','I','I','I','C','C','I','I','C','C','I','C','I','I'),
                  "TAM" = c('G','M','G','M','M','P','G','M','P','M','P','P','M','M','G'))
```

Após isso, o banco foi dividido entre Comércio e Indústria:

```
C <- Emp[ which(Emp$SET == "C"),]
I <- Emp[ which(Emp$SET == "I"),]
```

Em seguida, foram calculadas a média, moda e mediana de cada setor:

```
#Media e mediana de cada grupo

mean(C$MES)
```

```
## [1] 7.142857
```

```r
mean(I$MES)
```

```
## [1] 4.625
```

```r
median(C$MES)
```

```
## [1] 8
```

```r
median(I$MES)
```

```
## [1] 4.5
```

```r
  #Desvio Padrão de cada grupo

sd(C$MES)
```

```
## [1] 1.46385
```

```r
sd(I$MES)
```

```
## [1] 1.06066
```

Onde percebe-se que a média, mediana e desvio padrão dos meses com crescimento é maior no comércio.

O próximo passo é indentificar o número máximo de meses com crescimento para a empresa receber um incentivo fiscal. Nesse caso, serão as empresas com meses menores que o 25 decil:

```r
fivenum(Emp$MES)
```

```
## [1] 3.0 4.5 6.0 7.0 9.0
```

Nesse caso, o máximo é 4.5 meses. Por fim, é necessário verificar as estatísticas descritivas de acordo com o porte da empresa:

```r
G <- Emp[ which(Emp$TAM == "G"),]  ##Tamanho grande
M <- Emp[ which(Emp$TAM == "M"),]  ##Tamanho médio
P <- Emp[ which(Emp$TAM == "P"),]  ##Tamanho pequeno

median(G$MES)
```

```
## [1] 6
```

```r
median(M$MES)
```

```
## [1] 6
```

```r
median(P$MES)
```

```
## [1] 5.5
```

```r
mean(G$MES)
```

```
## [1] 6
```

```r
mean(M$MES)
```

```
## [1] 5.857143
```

```r
mean(P$MES)
```

```
## [1] 5.5
```

```r
sd(G$MES)
```

```
## [1] 2.309401
```

```r
sd(M$MES)
```

```
## [1] 2.115701
```

```r
sd(P$MES)
```

```
## [1] 0.5773503
```

A mediana de meses com crescimento é a mesma para empresas de grande e médio porte (6) e um pouco menor para empresas de pequeno porte (5.5). Já em relação a média, as empresas de grande porte apresentam uma quantidade de meses com crescimento um pouco maior que as empresas de médio e pequeno porte. Do mesmo modo, no tocante ao desvio padrão, as empresas grandes apresentam uma maior variação nos meses com crescimento do que as empresas de médio e pequeno porte.

Questão 6 - Lista I

Primeiramente é utilizado o comando `data.frame` para criar um data frame contendo as informações necessárias:

```r
Inv <- data.frame("CID" = c("A", "B", "C", "D", "E", "F", "G", "H", "I", "J"),
                  "INV" = c(26,16, 14,10, 19, 15, 19, 16, 19, 18))
```

O cálculo do investimento médio ocorre da segunite maneira:

```r
mean(Inv$INV)
```

```
## [1] 17.2
```

Ou seja, o investimento médio é de 17.2

Agora será calculado quais cidades receberão o programa especial:

```r
s <- sd(Inv$INV)

mean(Inv$INV) - (2*s) ##Identificando valor para programa especial
```

```
## [1] 8.830744
```

Cidades com menos de 8.83 em investimentos receberão o programa especial. Agora, será verificado qual o valor minimo e máximo de investimento básico e quais cidades estão neste critério.

```r
mean(Inv$INV) - (2*s) #Valor minimo investimento básico
```

```
## [1] 8.830744
```

```r
mean(Inv$INV) + (2*s) #Valor maximo investimento basico
```

```
## [1] 25.56926
```

```r
invba <- Inv[ which(Inv$INV < 25.56
                    & Inv$INV > 8.83), ] #Selecionando casos com investimento básico

mean(invba$INV) #media investimento basico
```

```
## [1] 16.22222
```

Podemos perceber que a média de investimento básico acaba sendo menor do que a média de investimento total ($16.22 < 17.2$).

Questão 7 - Lista I

Primeiro, vamos criar o banco de dados contendo as informações dos estímulos visuais:

```r
Est <- data.frame("IND" = 1:20, "A" = c(55,2,13,11,23,2,15,12,14,28,12,45,19,30,16,12,7,13,1,7),
                  "B" =  c(20,7,6,5,3,25,5,3,3,10,8,5,1,35,9,8,12,2,26,NA))
```

O próximo passo é obter as estatísticas descritivas do banco:

```r
mean(Est$A) ##Media estimulo A = 16.85
```

```
## [1] 16.85
```

```r
mean(Est$B, na.rm = TRUE) ##Media estimulo B = 10.16
```

```
## [1] 10.15789
```

```r
median(Est$A) ##Mediana estimulo A = 13
```

```
## [1] 13
```

```r
median(Est$B, na.rm = TRUE) ##Mediana estimulo B  = 7
```

```
## [1] 7
```

```r
sd(Est$A) ##Desvio padrao estimulo A = 13.80
```

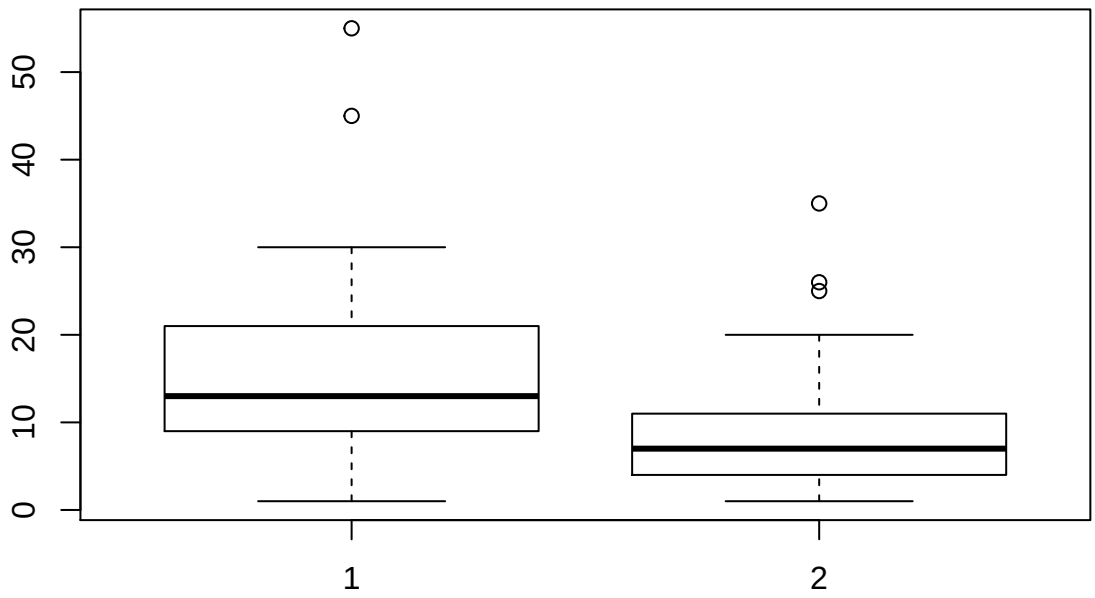
```
## [1] 13.80418
```

```r
sd(Est$B, na.rm = TRUE) ##Desvio padrao estimulo B = 9.45
```

```
## [1] 9.459053
```

Os resultados mostram que a média do tempo de reação do estímulo A é maior do que a média de reação do estímulo B. Do mesmo modo, a média do estímulo A é 13 enquanto que em relação ao estímulo B esse valor é de 7. Ao analisar o desvio padrão dos dois estímulos, o desvio padrão do estímulo A é de 13.80 enquanto que o do estímulo B é de 9.45, sendo possível inferir que o estímulo B possui uma menor variação no tempo de resposta.

Por fim, é possível observamos a distribuição das variáveis por meio de um boxplot:



Podemos visualizar que ambas as distribuições apresentam alguns valores destoantes, que no estímulo A estão acima de 40 (55 e 45) e no estímulo B estão acima de 20 (35,26 e 25).

Questão 8 - Lista I

Na questão 8, também precisamos iniciar criando o banco de dados para realizar as análises:

```
fam <- data.frame("FAM" = c("A", "B", "C", "D", "E", "F", "G", "H", "I", "J"), "REN" =
                    c(12,16,18,20,28,30,40,48,50,54), "SAU" = c(7.2,7.4,7,6.5,6.6,6.7,6,5.6,6,5.5))
```

Após isso, podemos obter as informações solicitadas:

```
mean(fam$REN) #Media renda familia
```

```
## [1] 31.6
```

```
mean(fam$SAU) #Media percentual gasto com saude
```

```
## [1] 6.45
```

```
sd(fam$REN) #sd renda familia
```

```
## [1] 15.42869
```

```
sd(fam$SAU) #sd percentual gasto com saude
```

```
## [1] 0.6570134
```

```
cov(fam$REN, fam$SAU) #cov entre as duas variaveis
```

```
## [1] -9.533333
```

```
cor(fam$REN, fam$SAU) #correlacao entre as duas variaveis
```

```
## [1] -0.9404625
```

Os resultados mostram que a média de renda da família é de 31.6 e a média do gasto percentual com saúde é de 6.45. Já em relação ao desvio padrão, o valor obtido em relação a renda da família foi de 15.42 e o gasto percentual com saúde foi de 0.66. A covariância entre as duas variáveis foi -9.53 e a correlação foi de -0.94.

Ou seja, a correlação mostrou uma relação inversamente proporcional entre as duas variáveis, indicando que o aumento de uma leva a queda da outra. Além disso, o valor obtido muito próximo de 1 mostra que o grau de associação entre as variáveis é forte.

Questão 9 - Lista I

Na questão 9, primeiro será criado um banco contendo as informações dos alunos:

```
not <- data.frame("ALU" = c("A", "B", "C", "D", "E", "F", "G", "H", "I"), "PI" =
                    c(7.5,8.2,8.5,8.7,8.8,9.1,9.2,9.3,10), "P2" = c(8.2, 8,8.3,
                                                                    8.5,9.4,9.6,9,9.3,9.7))
```

Com isso, podemos realizar uma análise de correlação entre as notas da prova I e as notas da prova II:

```r
cor(not$PI, not$P2)
```

```
## [1] 0.8301592
```

O valor obtido na correlação de 0.83 indica uma associação forte entre as variáveis, mostrnado que estas estão altamente correlacionadas. Ou seja, uma nota alta na prova I indica uma nota alta também na prova II.

Questão 5 - Lista II

Primeiramente, vamos criar um banco com 1000 casos e atribuir valor 1 até 620 (os respondentes que não votaram) e 0 para os 380 casos restantes.

```r
Banco <- data.frame("RES" = 1:1000)

Banco$valor <- ifelse(Banco$RES <= 620, 1, 0) ##Até 620 atribuir valor 1 e depois valor 0
```

Agora podemos obter o valor da média e desvio padrão da amostra

```r
mean(Banco$valor)
```

```
## [1] 0.62
```

```r
sd(Banco$valor)
```

```
## [1] 0.4856293
```

Para estimar a média populacional com um I.C de 95%, vamos obter o valor Z

```r
error <- qnorm(0.975)
```

Com isso, é possível calcular o valor minimo e máximo do I.C onde está a média da população

```r
0.62 + (error*0.486/sqrt(1000)) ##Valor máximo dentro do IC de 95
```

```
## [1] 0.650122
```

```r
0.62 - (error*0.486/sqrt(1000)) ##Valor minimo dentro do IC de 95
```

```
## [1] 0.589878
```

Questão 6 - Lista II

(a) Para obter a quantidade de eleitores que devem ser consultados com um erro de 0.05 e probaiblidade de 0.95 e p de 0.05 fazemos o seguinte:

```r
Zs <- 1.96^2 ##Calculando Z ao quadrado
E <- 0.05^2  ##Calculando erro ao quadrado

0.5*(1-0.5) ##Executando parte da fórmula para obter valor final com p
```

```
## [1] 0.25
```

```r
n <- Zs* 0.25/E ##Obtendo o valor de N com erro de 0.05

print(n)
```

```
## [1] 384.16
```

Ou seja, precisamos de aproximadamente 385 respondentes

(b) Com um erro menor (0.02) fazemos o seguinte:

```r
E <- 0.02^2 ##Caclulando erro ao quadro de 0.02

n <- Zs* 0.25/E ##Obtendo o valor de N com erro de 0.02

print(n)
```

```
## [1] 2401
```

Com um erro menor, o n amostral aumenta, sendo necessário 2401 respondentes.

(c) Com a informação de que um candidato pode ter 25% dos votos, podemos reduzir o tamanho da amostra necessária:

```r
0.25*(1-0.25) ##Executando parte da fórmula para obter valor final com p
```

```
## [1] 0.1875
```

```r
n <- Zs* 0.1875/E ##Obtendo o valor de N com erro de 0.02 e informação de 25% dos eleitores

print(n)
```

```
## [1] 1800.75
```

Com essa informação, podemos reduzir o tamanho da amostra para 1801 respondentes. Isso se deve ao fato de termos um valor da população já conhecido (25%)

(d) Sabendo que 564 respondentes de acordo com a amostra obtida em (b) votaram no partido conservador, podemos obter o I.C a 95% de confiança executando o seguinte comando:

```r
Banco <- data.frame("Elet" = 1:2401) ##Criando o banco

Banco$valor <- ifelse(Banco$Elet <= 564, 1, 0) ##Até 564 atribuir valor 1 e depois valor 0

mean(Banco$valor) ##media
sd(Banco$valor) ##Desvio padrao

error <- qnorm(0.975) ##identificando valor Z para 95%
```

```r
0.235 + (error*0.424/sqrt(2401))
```

```
## [1] 0.2519597
```

```r
##Valor máximo dentro do IC de 95
0.235 - (error*0.424/sqrt(2401))
```

```
## [1] 0.2180403
```

```r
##Valor minimo dentro do IC de 95
```

Questão 11 - Lista II

Nesta questão, podemos formular as seguintes hipóteses nula e alternativa:

H0: Não há associação entre ideologia e voto do parlamentar no tema;

H1: Há associação entre ideologia e voto do parlamentar no tema

Para efetuar o teste de hipóteses, vamos criar uma matriz de dados contento a quantidade de votos de acordo com a posição do parlamentar em relação ao tema e a ideologia do partido do qual o parlamentar faz parte:

```r
droga <- matrix(cbind(c(450,100),c(150,300)), nrow = 2,
                dimnames = list(c("FAV","CONT"), c("ESQ","DIR")))

options(scipen=999) ##desativando a função de notação científica no output
```

Com a matriz criada, podemos efetuar o teste de qui-quadrado para testar as hipóteses:

```r
chisq.test(droga)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  droga
## X-squared = 240.41, df = 1, p-value < 0.00000000000000022
```

Como é possível verificar nos resultados, o valor do X obtido foi de 240 com um *p-valor* menor que 0.05. Ou seja, podemos rejeitar a hipótese nula apresentada e considerar que há uma relação entre a ideologia e o voto do parlamentar no tema.

Questão 12 - Lista II

Na questão 12, também precisamos inicialmente elaborar as duas hipóteses para realizarmos o teste T:

H0: Taxa média de reeleição dos parlamentares foi a mesma antes e depois do *Watergate scandal*;

H1: Taxa média de reeleição dos parlamentares não foi a mesma antes e depois do *Watergate scandal*

Agora criamos um *Data frame* com as informações da questão:

```
election <- data.frame("YEAR" = seq(from = 1964, to = 2006, by = 2 ), "HOUSE" =
                       c(87,88,97,85,94,88,96,94,91,90,95,98,98,96,88,90,94,98,98,96,98,94),
                       "SEN" = c(85,88,71,77,74,85,64,60,55,93,90,75,85,96,83,92,91,90,79,86,96,79))
```

E criamos uma variável atribuindo um valor 0 e 1 para antes e depois do escândalo:

```
election$SCANDAL <- ifelse(election$YEAR <= 1972, 0,1)
```

Com isso, podemos fazer o teste T para as duas casas do congresso americano comparando antes e depois:

```
t.test(election$HOUSE ~ election$SCANDAL)
```

```
##
##  Welch Two Sample t-test
##
## data:  election$HOUSE by election$SCANDAL
## t = -1.6622, df = 5.2313, p-value = 0.1548
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -10.193900   2.123311
## sample estimates:
## mean in group 0 mean in group 1
##        90.20000        94.23529
```

```
##teste t comparando antes e depois do escândalo para a câmara

t.test(election$SEN ~ election$SCANDAL)
```

```
##
##  Welch Two Sample t-test
##
## data:  election$SEN by election$SCANDAL
## t = -0.74516, df = 11.689, p-value = 0.4709
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.954467   6.366231
## sample estimates:
## mean in group 0 mean in group 1
##        79.00000        82.29412
```

```
##teste t comparando antes e depois do escândalo para o senado
```

Ao verificarmos o p-valor em ambos os testes, verificamos que estes são maiores que 0.05. Ou seja, não é possível rejeitar a hipótese nula de que a taxa média de reeleição foi a mesma tanto antes quanto depois do escândalo.

Questão 13 - Lista II

Na questão 13, podemos apresentar as seguintes hipóteses:

H0: Não há associação entre as variáveis PIB e Votos;

H1: Há associação entre as variáveis PIB e Votos.

Para testar essas hipóteses, vamos criar o banco com as informações necessárias:

```
vote <- data.frame("YEAR" = seq(from = 1876, to = 1932, by = 4 ),
                   "GROWTH" =  c(5.11,3.879,1.589,-5.553, 2.763,-10.024,-1.425,-2.421,
                                 -6.281,4.164, 2.229,-11.463,-3.872,4.623, -14.586),
                   "VOTES" = c(48.516,50.22,49.846,50.414,48.268,47.76,53.171,
                               60.006,54.483,54.708,51.682,36.148,58.263,58.756,40.851))
```

Com o *Data frame* criado, podemos então efetuar o teste de correlação:

```
cor.test(vote$GROWTH, vote$VOTES)
```

```
##
##  Pearson's product-moment correlation
##
## data:  vote$GROWTH and vote$VOTES
## t = 2.2546, df = 13, p-value = 0.04205
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02461898 0.81980079
## sample estimates:
##       cor
## 0.5301953
```

O resultado do teste de correlação mostra que existe uma associação entre as variáveis com força média e significativa (p-valor < 0.05). Ao comparamos com a série histórica maior, veiricamos que a associação entre as variáveis persiste e também é estatisticamente significativa.