Análise de Dados - UFPE/2019 - Lista 5

Maria Eduarda R. N. Lessa

14 de maio de 2019

LISTA 1

Questão 5:

```
# Criar vetores com os valores da questão:
meses <- c(8,9,4,5,3,6,8,6,6,8,5,5,6,4,4)
setor <- c("C", "C", "I", "I", "I", "C", "C", "I", "I", "C", "C", "I", "C", "I", "I")
tamanho <- c("G", "M", "G", "M", "M", "P", "G", "M", "P", "M", "P", "P", "M", "M", "G")

# Transformar em data frame, nomear base "empresas":
empresas <- data.frame(meses,
                       setor,
                       tamanho)
```

letra a)

```
# Verificar tipo de cada variável:
is.numeric(meses)
```

```
## [1] TRUE
```

```
is.numeric(setor)
```

```
## [1] FALSE
```

```
is.numeric(tamanho)
```

```
## [1] FALSE
```

letra b)

```
# Dividir empresas em dois grupos (C e I):
empresas_c <- subset(empresas, setor == "C")
empresas_i <- subset(empresas, setor == "I")

# Comparar média e mediana dos meses de crescimento:
summary(empresas_c$meses)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   6.000   8.000   7.143   8.000   9.000
```

```r
summary(empresas_i$meses)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.000   4.500   4.625   5.250   6.000
```

letra c)

```r
# Calcular desvio padrão de cada grupo:
sd(empresas_c$meses)
```

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

```r
sd(empresas_i$meses)
```

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

O grupo das empresas da indústria (I) é mais homogêneo em relação ao do comércio (C).

letra d)

```r
# A medida descritiva que fornece a informação sobre 25% das empresas com menor
# crescimento em meses é o primeiro quartil, portanto:
summary(empresas$meses)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0     4.5     6.0     5.8     7.0     9.0
```

O número máximo de meses apresentando crescimento para que a empresa receba o incentivo fiscal seria de 4 meses e meio, nos dados analisados, no entanto, os números de meses são discretos, o que significa que as empresas que apresentam até 4 meses de crescimento receberão o benefício.

letra e)

```r
# Dividir empresas em três grupos (P, M e G):
empresas_p <- subset(empresas, tamanho == "P")
empresas_m <- subset(empresas, tamanho == "M")
empresas_g <- subset(empresas, tamanho == "G")

# Calcular média, mediana e dp dos meses de crescimento das empresas
# de acordo com o tamanho (P, M ou G):

# Empresas P:
summary(empresas_p$meses)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     5.0     5.0     5.5     5.5     6.0     6.0
```

```r
sd(empresas_p$meses)
```

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

```r
# Empresas M:
summary(empresas_m$meses)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.500   6.000   5.857   7.000   9.000
```

```r
sd(empresas_m$meses)
```

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

```r
# Empresas G:
summary(empresas_g$meses)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       4       4       6       6       8       8
```

```r
sd(empresas_g$meses)
```

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

Parece haver uma relação entre o tamanho da empresa e o número de meses com crescimento, já que as médias para as empresas P, M e G foram, respectivamente, 5.5, 5.86 e 6. Nas empresas pequenas houve menor variância entre os meses de crescimento, enquanto nas empresas grandes, maior.

Questão 6:

```r
# Criar vetores com os valores da questão:
cidades <- c("A", "B", "C", "D", "E", "F", "G", "H", "I", "J")
investimento <- c(26, 16, 14, 10, 19, 15, 19, 16, 19, 18)

# Transformar em data frame, nomear base "cid_invest":
cid_invest <- data.frame(cidades,
                         investimento)
```

letra a)

```r
# Calcular média de investimento:
mean(investimento)
```

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

letra b)

```r
# Calcular desvio padrão:
sd(investimento)
```

```
## [1] 4.184628
```

```r
# Calcular valor da média menos 2*dp:
mean(investimento) - (2*(sd(investimento)))
```

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

```r
# Checar se existe algum valor abaixo de 8.830744 na base de dados:
investimento < 8.830744
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

Nenhuma cidade receberá o programa.

letra c)

```r
# Calcular média original (17.2) mais 2 dp e média original menos 2 dp:
mean(investimento) + 2*(sd(investimento))
```

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

```r
mean(investimento) - 2*(sd(investimento))
```

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

```r
# Criar nova base apenas com valores > 8.830744 e < 25.56926:
cid_invest_novo <- subset(cid_invest, investimento <= 25.56926)
# Já foi visto na letra b) que não há valor abaixo de 8.830744 na base, por isso
# é necessário apenas criar filtro para eliminar valores acima de 25.56926.

# Calcular média da nova base:
mean(cid_invest_novo$investimento)
```

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

A média é uma medida sensível a valores extremos, por isso, com a eliminação do investimento da cidade A, de 26, a nova média obtida é menor do que a original.

Questão 7:

letra a)

```
# Criar vetores com os valores da questão:
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)

# Transformar em data frame, nomear base "estimulos":
estimulos <- data.frame(A,B)

# Calcular média, mediana e desvio padrão do estímulo "A":
summary(A)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   10.00   13.00   16.85   20.00   55.00
```

```
sd(A)
```

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

```
# Calcular média, mediana e desvio padrão do estímulo "B":
summary(B)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    4.00    7.00   10.16   11.00   35.00       1
```

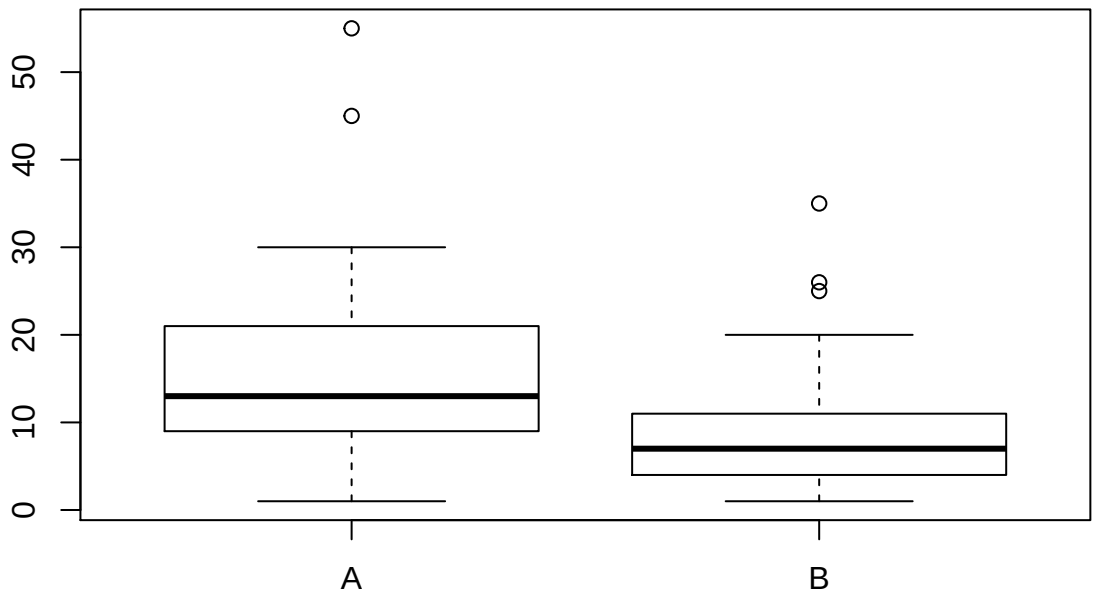
```
sd(B, na.rm = TRUE)
```

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

Em média, para o estímulo "B" o tempo de reação é menor. Os valores de reação para o estímulo "A" variaram mais em relação à média e também estão mais dispersos que os valores de "B", como é possível notar a partir do desvio padrão.

letra b)

```
# gerar boxplot do data frame "estimulos":
boxplot(estimulos)
```




```
# Identificar valores extremos do estímulo "A":
boxplot(estimulos$A, plot=FALSE)$out
```

```
## [1] 55 45
```

```
# Identificar valores extremos do estímulo "B":
boxplot(estimulos$B, plot=FALSE)$out
```

```
## [1] 25 35 26
```

No estímulo "A" os tempos de reação foram maiores e seus valores estãoo mais dispersos do que no estímulo "B". Não existe assimetria acentuada em "A" ou "B", visto que a linha da mediana está próxima do meio da caixa nos dois casos (ainda assim, "A" apresenta maior assimetria do que "B"). Para o estímulo "A" foram observados 2 resultados extremos (45 e 55), enquanto para "B", 3 (25, 26 e 35).

Questão 8:

```r
# Criar vetores com os valores da questão:
fam <- c("A", "B", "C", "D", "E", "F", "G", "H", "I", "J")
renda_sm <- c(12, 16, 18, 20, 28, 30, 40, 48, 50, 54)
saude_perc <- c(7.2, 7.4, 7, 6.5, 6.6, 6.7, 6, 5.6, 6, 5.5)

# Transformar em data frame, nomear base "gasto_fam_saude":
gasto_fam_saude <- data.frame(fam,
                              renda_sm,
                              saude_perc)
```

letra a)

```r
# Calcular média e desvio padrão de renda (em salários mínimos ["renda_sm"])
# e percentual gasto com saúde ("saude_perc"):
summary(renda_sm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    12.0    18.5    29.0    31.6    46.0    54.0
```

```r
sd(renda_sm)
```

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

```r
summary(saude_perc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.500   6.000   6.550   6.450   6.925   7.400
```

```r
sd(saude_perc)
```

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

```r
# Calcular covariância e correlação:
cov(renda_sm, saude_perc, method = c("pearson"))
```

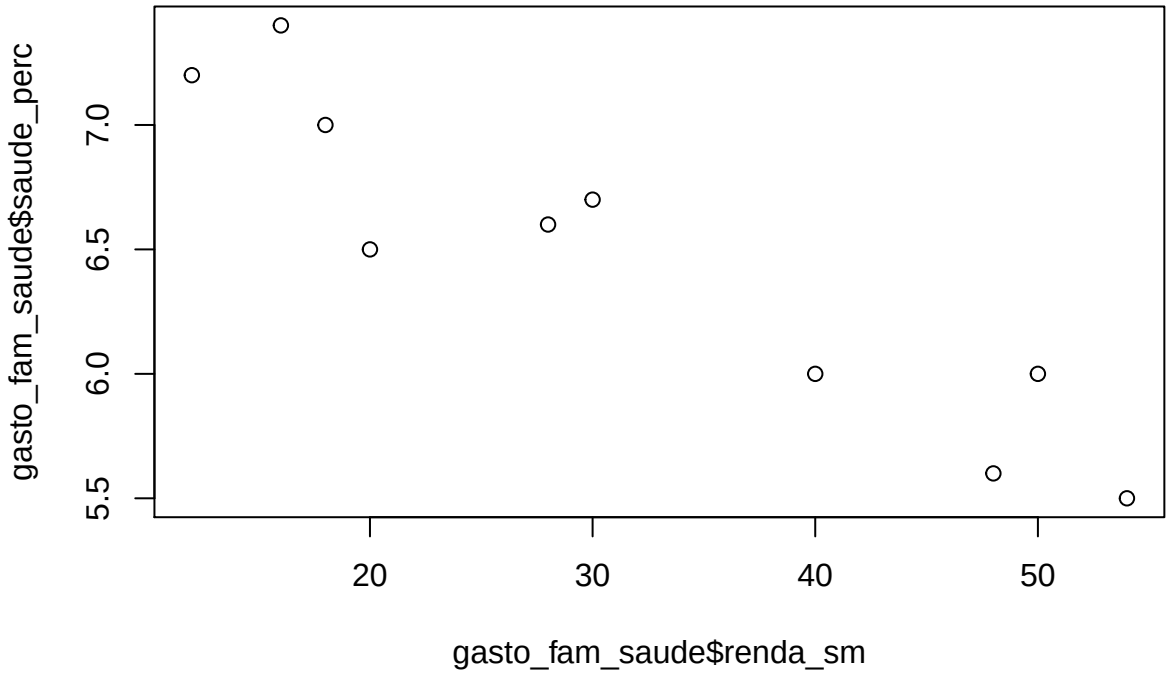
```
## [1] -9.533333
```

```r
cor(renda_sm, saude_perc, method = c("pearson"))
```

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

letra b)

```r
# Gerar diagrama de dispersão:
plot(gasto_fam_saude$renda_sm, gasto_fam_saude$saude_perc)
```



As variáveis apresentam forte correlação negativa, ou seja, é possível observar que quanto maior a renda da família, menor o gasto percentual dessa renda com saúde.

Questão 9:

```r
# Criar vetores com os valores da questão:
alunos <-  c("A", "B", "C", "D", "E", "F", "G", "H", "I")
P1 <- 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)

# Transformar em data frame, nomear base "alunos_provas":
alunos_provas <- data.frame(alunos,
                            P1,
                            P2)
```

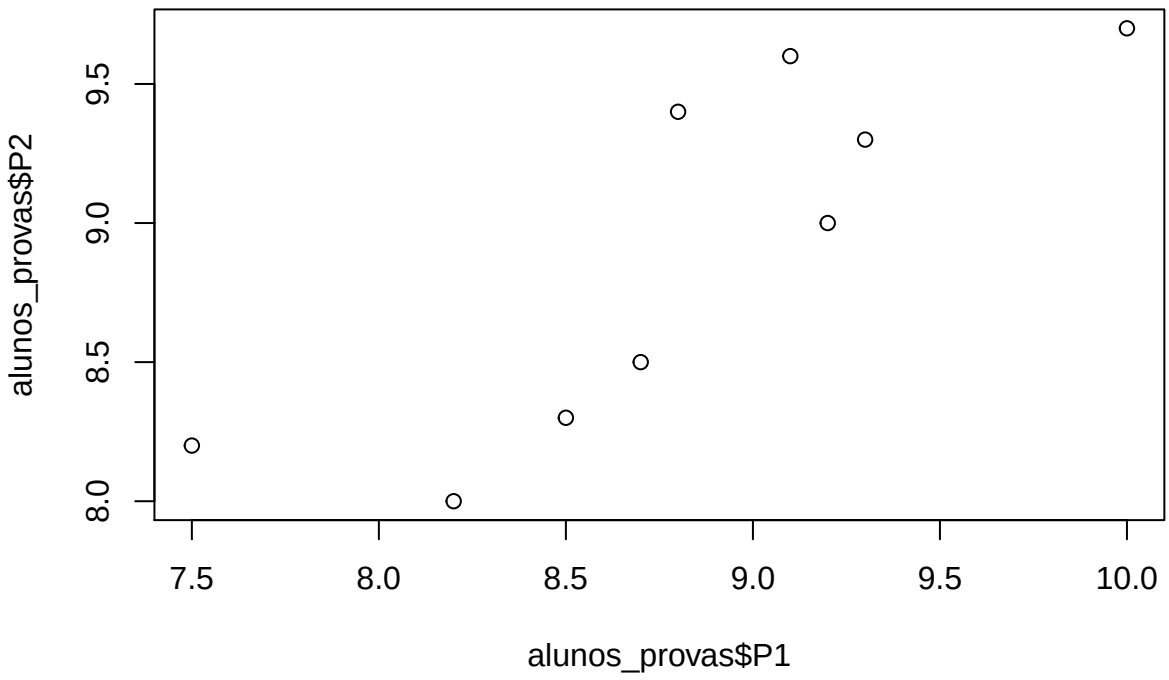
letra a)

```r
# Calcular coeficiente de correlação entre P1 e P2:
cor(P1, P2, method = c("pearson"))
```

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

letra b)

```r
# Gerar diagrama de dispersão:
plot(alunos_provas$P1, alunos_provas$P2)
```



É possível observar uma forte correlação positiva (0.83) entre as notas dos estudantes nas duas provas.

LISTA 2

Questão 5:

letra a)

```
# 1000 eleitores
# 620 afirmaram jamais votar no candidato x (sucesso = 1)
# 380 não responderam ou votariam em x (fracasso = 0)

# Criar vetor com a resposta dos eleitores:
votos <- rep(c(1,0), times = c(620,380))

# Calcular a média da amostra:
mean(votos)
```

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

letra b)

```
# Calcular desvio padrão da amostra:
sd(votos)
```

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

letra c)

```
# Verificar intervalo de confiança de 95%:
t.test(votos)
```

```
##
##  One Sample t-test
##
## data:  votos
## t = 40.373, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.5898644 0.6501356
## sample estimates:
## mean of x
##      0.62
```

Questão 6:

letra a)

```
# Instalar e requerer pacote "samplingbook":

# install.packages("samplingbook")
# require(samplingbook)

# Utilizar função "sample.size.prop" para descobrir número necessário de eleitores
# consultados para erro de 0.05:

# sample.size.prop(0.05, P = 0.5, N = Inf, level = 0.95)

# Resultado - Sample size needed: 385
```

letra b)

```
# Utilizar função "sample.size.prop" para descobrir número necessário de eleitores
# consultados para erro de 0.02:

# sample.size.prop(0.02, P = 0.5, N = Inf, level = 0.95)

# Resultado - Sample size needed: 2401
```

Comparando resultados de a) e b): Quanto maior o tamanho da amostra, mais próximas suas medidas serão dos parâmetros populacionais, por isso, quanto menor o erro, maior o tamanho da amostra.

letra c)

```
# Utilizar função "sample.size.prop" para descobrir número necessário de eleitores
# consultados para erro de 0.02, quando já sabemos o posicionamento de 25% dos indivíduos da amostra:
# sample.size.prop(0.02, P = 0.75, N = Inf, level = 0.95)

# Resultado - Sample size needed: 1801
```

Sim, é possível diminuir o tamanho da amostra em aproximadamente 25%.

Questão 11:

```r
# Criar tabela "descr_partido" com os dados da questão:
descr_partido <- matrix(c(450, 150,
                          100, 300), ncol = 2, byrow = TRUE)
colnames(descr_partido) <- c("Esquerda",
                             "Direita")
rownames(descr_partido) <- c("Favoravel",
                             "Contrario")
descr_partido <- as.table(descr_partido)

# Checar tabela:
descr_partido
```

```
##           Esquerda Direita
## Favoravel      450     150
## Contrario      100     300
```

letra a)

Hipótese nula (H0): Não existe relação entre a ideologia do partido e o posicionamento dos seus integrantes quanto a descrimininalização das drogas.

Hipótese alternativa (H1): Existe relação entre a ideologia do partido e o posicionamento dos seus integrantes quanto a descrimininalização das drogas.

letra b)

Erro do tipo 1: A hipótese nula é rejeitada quando é, de fato, verdadeira. Nesse caso, um erro do tipo 1 seria cometido se, de fato, não houvesse relação entre a ideologia do partido e o posicionamento dos seus integrantes quanto a descriminalização das drogas e fosse assumido no estudo que há uma relação.

Erro do tipo 2: A hipótese nula NÃO é rejeitada, quando, na verdade, deveria ser. No exemplo da questão o erro do tipo 2 seria assumir que a ideologia do partido não tem relação com o posicionamento de seus representantes quanto a descriminalização das drogas, quando, na verdade, esta relação existe.

letra c)

```r
# Realizar teste do qui-quadrado:
chisq.test(descr_partido, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  descr_partido
## X-squared = 242.42, df = 1, p-value < 2.2e-16
```

Ao analisar o valor do qui-quadrado (242,42) e compará-lo com o seu valor crítico (3,841), levando em conta os graus de liberdade (neste caso igual a 1), é possível concluir que existe forte associação entre as variáveis, com o p valor $< 0,05$.

Questão 12:

```
# Criar vetores com os valores da questão, para a "câmara" (house):
pre_water_house <- c(87,88,97,85,94)
pos_water_house <- c(88,96,94,91,90,95,98,98,96,88,90,94,98,98,96,98,94)

# Realizar Teste-t para comparar as médias anteriores e posteriores ao
# escândalo de Watergate:
t.test(pre_water_house, pos_water_house, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  pre_water_house and pos_water_house
## t = -2.0217, df = 20, p-value = 0.05679
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.1988977  0.1283095
## sample estimates:
## mean of x mean of y
##  90.20000  94.23529
```

```
# Criar vetores com os valores da questão, para o senado (senate):
pre_water_senate <- c(85,88,71,77,74)
pos_water_senate <- c(85,64,60,55,93,90,75,85,96,83,92,91,90,79,86,96,79)

# Realizar Teste-t para comparar as médias anteriores e posteriores ao
# escândalo de Watergate:
t.test(pre_water_senate, pos_water_senate, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  pre_water_senate and pos_water_senate
## t = -0.56045, df = 20, p-value = 0.5814
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.55470   8.96646
## sample estimates:
## mean of x mean of y
##  79.00000  82.29412
```

É possível observar que em ambos os casos o valor p é > 0.05, por isso NÃO é possível rejeitar a hipótese nula.

Questão 13:

```
# Criar vetores com os valores da questão:
ano <- c(1876,1880,1884,1888,1892,1896,1900,1904,1908,1912,1916,1920,1924,1928,1932)

var_PIB <- c(5.11,3.879,1.589,-5.553,2.736,-10.024,-1.425,-2.421,-6.281,
             4.164,2.229,-11.463,-3.872,4.623,-14.586)

var_votos <- 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)

# Transformar em data frame "ano_PIB_votos":
ano_PIB_votos <- data.frame(ano, var_PIB, var_votos)
```

letra a)

Hipótese nula (H0): Não existe relação entre a variação do PIB e o percentual de votos recebidos pelo candidato do partido incumbente.

Hipótese alternativa (H1): Existe relação entre a variação do PIB e o percentual de votos recebidos pelo candidato do partido incumbente.

letra b)

```
# Calcular o coeficiente de correlação, a estatística-t, os graus de liberdade e o p-valor:
cor.test(var_PIB, var_votos, method = c("pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  var_PIB and var_votos
## t = 2.2561, df = 13, p-value = 0.04193
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02495982 0.81991260
## sample estimates:
##       cor
## 0.5304405
```

É possível concluir, a partir dos dados apresentados acima, que há relação entre as variáveis.

letra c)

É possível observar que a análise da relação entre a variação do PIB e o percentual de votos recebidos pelo candidato do partido incumbente, entre os anos de 1876 e 2008, permitiu rejeitar a hipótese nula, apresentando maior coeficiente de correlação e menor p-valor do que a análise dos anos 1876 a 1932.