Exercício 1 - Metodologia de Pesquisa em Computação

Gabriel Luciano Gomes (265673) Gabriel Oliveira Campos (265146)
Paulo Junio Reis Rodrigues (265674)

October 9, 2020

1 Dados não pareados

Para este exercício, o seguinte código foi utilizado para análise dos dados. Note que antes de realizar o cálculo propriamente dito, um pré processamento foi realizado para utilização do T-teste e Wilcoxon.

```
ex1 <- read.csv("D:/dbs/MO430/ex1.csv")

ex1$type <- sapply(ex1$type, as.character)
ex1[which(ex1$type == 'No'),"type"] <- 0
ex1[which(ex1$type == 'Yes'),"type"] <- 1
ex1$type <- sapply(ex1$type, as.numeric)

bpNo <- array()
bpYes <- array()
y <- 1
z <- 1

for(i in 1:length(ex1$type)){
  if(ex1$type[i] == 0){
    bpNo[y] <- ex1$bp[i]
    y <- y+1
  }else{
    bpYes[z] <- ex1$bp[i]
    z <- z+1
  }
}

cat("Mediana das pessoas diabeticas: ", median(bpYes))
cat("Mediana das pessoas nao diabeticas: ", median(bpNo))


if(median(bpYes) < median(bpNo)){
  cat("Pessoas nao diabeticas tem maior pressao.")
}else{
  cat("Pessoas diabeticas tem maior pressao.")
}
```

t.test(bpNo, bpYes)
wilcox.test(bpNo, bpYes)

```
> cat("Mediana das pessoas diabéticas: ", median(bpYes))
Mediana das pessoas diabéticas:  76
> cat("Mediana das pessoas não diabéticas: ", median(bpNo))
Mediana das pessoas não diabéticas:  70
> if(median(bpYes) < median(bpNo)){
+     cat("Pessoas não diabéticas tem maior pressão.")
+ }else{
+     cat("Pessoas diabéticas têm maior pressão.")
+ }
Pessoas diabéticas têm maior pressão.
> t.test(bpNo,bpYes)

        Welch Two Sample t-test

data:  bpNo and bpYes
t = -2.9592, df = 130.28, p-value = 0.003665
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.414080 -1.671482
sample estimates:
mean of x mean of y
 69.54545  74.58824

> wilcox.test(bpNo,bpYes)

        Wilcoxon rank sum test with continuity correction

data:  bpNo and bpYes
W = 3307, p-value = 0.002294
alternative hypothesis: true location shift is not equal to 0
```

Figure 1: Imagem mostra todos os resultados da questão 1

1.1 Qual tem a maior pressão?

Para esta análise, foi realizado uma comparação na mediana dos dados analisados. Com isso, pode-se observar que pessoas diabéticas têm maior pressão em relação às não diabéticas.

1.2 A diferença é significativa?

Para responder esta quentão, é necessário observar os valores obtidos nos testes estatísticos. O teste T apresentou p-valor de 0.003665 e o Wilcoxon 0.002294. Como ambos são inferiores a 0.05, a diferença é significativa.

1.3 Discuta a diferença dos p-valores. Quais dos 2 testes você acha mais apropriado nesse caso.

Ao analisar os p-valores, é possível dizer que ambos os testes são positivos, pois apresentam p-valor inferior a 0.05. Como o Wilcoxon apresentou o menor p-valor, este é o mais apropriado a ser utilizado.

2 Dados pareados

Para este exercício, o seguinte código foi utilizado para análise dos dados.

```
ex1Paired <- read.csv("D:/dbs/MO430/ex1-paired.csv")

Aug <- ex1Paired$August
Nov <- ex1Paired$November

t.test(Aug, Nov, paired = TRUE)
wilcox.test(Aug, Nov, paired = TRUE)

cat("A media do mes de Agosto e: ",mean(Aug))
cat("A media do mes de Novembro e: ",mean(Nov))
cat("A mediana do mes de Agosto e: ",median(Aug))
cat("A mediana do mes de Novembro e: ",median(Nov))

if(mean(Aug) < mean(Nov)){
  cat("A maior media e do  mes Novembro com valor: ",mean(Nov))
}else{
  cat("A maior media e do  mes Agosto com valor: ",mean(Aug))
}
if(meadian(Aug) < meadian(Nov)){
  cat("A maior mediana e do  mes Novembro com valor: ",mean(Nov))
}else{
  cat("A maior mediana e do  mes Agosto com valor: ",mean(Aug))
}
```

```
> t.test(Aug, Nov, paired = TRUE)

        Paired t-test

data:  Aug and Nov
t = -2.3089, df = 12, p-value = 0.03956
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.5239348 -0.2760652
sample estimates:
mean of the differences
                  -4.9

> wilcox.test(Aug, Nov, paired = TRUE)

        Wilcoxon signed rank exact test

data:  Aug and Nov
V = 16, p-value = 0.03979
alternative hypothesis: true location shift is not equal to 0

> cat("A média do mês de Agosto é: ",mean(Aug))
A média do mês de Agosto é:  11.42308
> cat("A média do mês de Novembro é: ",mean(Nov))
A média do mês de Novembro é:  16.32308
> cat("A mediana do mês de Agosto é: ",median(Aug))
A mediana do mês de Agosto é:  10
> cat("A mediana do mês de Novembro é: ",median(Nov))
A mediana do mês de Novembro é:  15.3
> if(mean(Aug) < mean(Nov)){
+     cat("A maior média é do  mês Novembro com valor: ",mean(Nov))
+ }else{
+     cat("A maior média é do  mês Agosto com valor: ",mean(Aug))
+ }
A maior média é do  mês Novembro com valor:  16.32308> if(median(Aug) < median(Nov)){
+     cat("A maior mediana é do  mês Novembro com valor: ",mean(Nov))
+ }else{
+     cat("A maior mediana é do  mês Agosto com valor: ",mean(Aug))
+ }
A maior mediana é do  mês Novembro com valor:  16.32308
>
```

Figure 2: Imagem mostra todos os resultados da questão 2

2.1 Qual o mês com maior mediana ou média?

Após computador os dados, foi identificado que o mês de Novembro possui maior média e também maior mediana.

2.2 A diferença é significativa?

Para responder esta quentão, é necessário observar os valores obtidos nos testes estatísticos. O teste T apresentou p-valor de 0.03956 e o Wilcoxon 0.03979. Como ambos são inferiores a 0.05, a diferença é significativa.

2.3 Discuta a diferença dos p-valores. Quais dos 2 testes você acha mais apropriado nesse caso.

Ambos os testes apresentaram resultados muito semelhantes, 0.03956 para o t-teste e 0.03979 para o Wilcoxon. Sendo assim, ambos os testes podem ser utilizados. Por outro lado, o t-teste apresentou valor inferior, mesmo que seja ínfimo, que o torna a melhor escolha.

3 Estudo dos fatores que influenciam no p-valor

Para este exercício, o seguinte código foi utilizado para análise dos dados.

```
# ------ Gerar conjuntos com n = 30, x = 10 e 12 e sd = 5 ------ #
set.seed(1234)
fonteA <- rnorm(n = 30, mean = 10, sd = 5)
fonteB <- rnorm(n = 30, mean = 12, sd = 5)


# ---------------------------------------------------------------- #
# ---------------------- UTILIZANDO O T-TEST ---------------------- #
# ---------------------------------------------------------------- #
# ------------- Media de p valor para 50 repeticoes ------------- #
pVal <- array()

for(i in 1:50) {
  fonteA <- rnorm(n = 30, mean = 10, sd = 5)
  fonteB <- rnorm(n = 30, mean = 12, sd = 5)
  pVal[i] <- t.test(fonteA, fonteB)$p.value
}
cat("Media do valor de p-value:", mean(pVal))

# ------ Media de p valor para 50 repeticoes com n = 60 ------ #
for(i in 1:50) {
  fonteA <- rnorm(n = 60, mean = 10, sd = 5)
  fonteB <- rnorm(n = 60, mean = 12, sd = 5)
  pVal[i] <- t.test(fonteA, fonteB)$p.value
}
cat("Media do valor de p-value para n = 60:", mean(pVal))

# ------ Media de p valor para 50 repeticoes desvio de 10 ------ #
for(i in 1:50) {
  fonteA <- rnorm(n = 30, mean = 10, sd = 10)
  fonteB <- rnorm(n = 30, mean = 12, sd = 10)
  pVal[i] <- t.test(fonteA, fonteB)$p.value
}
cat("Media do valor de p-value para desvio = 10:", mean(pVal))

# ---- Media de p valor para 50 repeticoes media 10 e 15 ---- #
for(i in 1:50) {
```

```
  fonteA <- rnorm(n = 30, mean = 10, sd = 5)
  fonteB <- rnorm(n = 30, mean = 15, sd = 5)
  pVal[i] <- t.test(fonteA, fonteB)$p.value
}
cat("Media do valor de p-value para medias = 10 e 15:", mean(pVal))

# ------------------------------------------------------------ #
# ------------- UTILIZANDO O WILCOXON RANK-SUM --------------- #
# ------------------------------------------------------------ #
# ------------- Media de p valor para 50 repeticoes ---------- #
pVal <- array()

for(i in 1:50) {
  fonteA <- rnorm(n = 30, mean = 10, sd = 5)
  fonteB <- rnorm(n = 30, mean = 12, sd = 5)
  pVal[i] <- wilcox.test(fonteA, fonteB)$p.value
}
cat("Media do valor de p-value:", mean(pVal))

# ------- Media de p valor para 50 repeticoes com n = 60 ------- #
for(i in 1:50) {
  fonteA <- rnorm(n = 60, mean = 10, sd = 5)
  fonteB <- rnorm(n = 60, mean = 12, sd = 5)
  pVal[i] <- wilcox.test(fonteA, fonteB)$p.value
}
cat("Media do valor de p-value para n = 60:", mean(pVal))

# ------ Media de p valor para 50 repeticoes desvio de 10 ------ #
for(i in 1:50) {
  fonteA <- rnorm(n = 30, mean = 10, sd = 10)
  fonteB <- rnorm(n = 30, mean = 12, sd = 10)
  pVal[i] <- wilcox.test(fonteA, fonteB)$p.value
}
cat("Media do valor de p-value para desvio = 10:", mean(pVal))

# ----- Media de p valor para 50 repeticoes media 10 e 15 ----- #
for(i in 1:50) {
  fonteA <- rnorm(n = 30, mean = 10, sd = 5)
  fonteB <- rnorm(n = 30, mean = 15, sd = 5)
  pVal[i] <- wilcox.test(fonteA, fonteB)$p.value
}
cat("Media do valor de p-value para medias = 10 e 15:", mean(pVal))
```

```
> # ----- Gerar conjuntos com n = 30, x = 10 e 12 e sd =5  ----- #
> set.seed(1234)
> fonteA <- rnorm(n = 30, mean = 10, sd = 5)
> fonteB <- rnorm(n = 30, mean = 12, sd = 5)
>
>
> # ------------------------------------------------------------ #
> # ----------------- UTILIZANDO O T-TEST ---------------------- #
> # ------------------------------------------------------------ #
> # ----------- Média de p valor para 50 repetições  ----------- #
> pVal <- array()
>
> for(i in 1:50) {
+     fonteA <- rnorm(n = 30, mean = 10, sd = 5)
+     fonteB <- rnorm(n = 30, mean = 12, sd = 5)
+     pVal[i] <- t.test(fonteA, fonteB)$p.value
+ }
>
> cat("Média do valor de p-value:", mean(pVal))
Média do valor de p-value: 0.2377255> 
> # ------ Média de p valor para 50 repetições com n = 60 ------ #
>
> for(i in 1:50) {
+     fonteA <- rnorm(n = 60, mean = 10, sd = 5)
+     fonteB <- rnorm(n = 60, mean = 12, sd = 5)
+     pVal[i] <- t.test(fonteA, fonteB)$p.value
+ }
>
> cat("Média do valor de p-value para n = 60:", mean(pVal))
Média do valor de p-value para n = 60: 0.1196386> 
> # ----- Média de p valor para 50 repetições desvio de 10 ----- #
>
> for(i in 1:50) {
+     fonteA <- rnorm(n = 30, mean = 10, sd = 10)
+     fonteB <- rnorm(n = 30, mean = 12, sd = 10)
+     pVal[i] <- t.test(fonteA, fonteB)$p.value
+ }
>
> cat("Média do valor de p-value para desvio = 10:", mean(pVal))
Média do valor de p-value para desvio = 10: 0.4966262> 
> # ---- Média de p valor para 50 repetições média 10 e 15 ---- #
>
> for(i in 1:50) {
+     fonteA <- rnorm(n = 30, mean = 10, sd = 5)
+     fonteB <- rnorm(n = 30, mean = 15, sd = 5)
+     pVal[i] <- t.test(fonteA, fonteB)$p.value
+ }
>
> cat("Média do valor de p-value para médias = 10 e 15:", mean(pVal))
Média do valor de p-value para médias = 10 e 15: 0.007673235
```

Figure 3: Imagem mostra todos os resultados da questão 3 para o teste t

3.1 Discuta a influencia dos 3 fatores no p-valor: número de dados, ruído dos dados (o desvio padrão das fontes) e "tamanho da diferença" entre as fontes (diferença entre as médias)

Como o primeiro parâmetro, tem-se o número de dados. Para este teste, variando o valor de n, é possível observar que o p-valor reduz e, quanto maior o conjunto, maior a probabilidade das amostras virem de conjuntos distintos. Entretanto, para o exercício o valor de $n = 60$, não foi suficiente para satisfazer a condição de $p < 0.05$, falhando no t-test.

Para o segundo parâmetro, tem-se o ruído dos dados (desvio padrão das fontes). Para este teste, é possível observar que o p-valor aumenta, e muito, quando se aumenta o desvio padrão. Isto poque os dados estarão mais distribuídos (distantes da média), que possibilita a retirada de elementos semelhantes da fonte.

Por fim, tem-se o tamanho da diferença entre as fontes. Para este teste, é possível observar que é o teste mais agravante a ser observado, pois o p-valor reduz ao ponto de satisfazer a condição do t-test (ser inferior a 0.05). Entretanto, este valor não passa de um erro de amostragem, pois os conjuntos foram amostrados de uma mesma fonte.

3.2 Exemplos utilizando o teste Wilcoxon

Após realizar os teste de Wilcoxon, foi verificado que os resultados obtidos foram bem próximos aos do teste T. Demonstrando que as propriedades do p-valor, definem o seu resultado, independente do teste utilizado.

```
> # ------------------------------------------------------------ #
> # ------------ UTILIZANDO O WILCOXON RANK-SUM  --------------- #
> # ------------------------------------------------------------ #
> # ----------- Média de p valor para 50 repetições  ----------- #
> pVal <- array()
>
> for(i in 1:50) {
+     fonteA <- rnorm(n = 30, mean = 10, sd = 5)
+     fonteB <- rnorm(n = 30, mean = 12, sd = 5)
+     pVal[i] <- wilcox.test(fonteA, fonteB)$p.value
+ }
>
> cat("Média do valor de p-value:", mean(pVal))
Média do valor de p-value: 0.2134911>
> # ------ Média de p valor para 50 repetições com n = 60 ------ #
>
> for(i in 1:50) {
+     fonteA <- rnorm(n = 60, mean = 10, sd = 5)
+     fonteB <- rnorm(n = 60, mean = 12, sd = 5)
+     pVal[i] <- wilcox.test(fonteA, fonteB)$p.value
+ }
>
> cat("Média do valor de p-value para n = 60:", mean(pVal))
Média do valor de p-value para n = 60: 0.1011155>
> # ----- Média de p valor para 50 repetições desvio de 10 ----- #
>
> for(i in 1:50) {
+     fonteA <- rnorm(n = 30, mean = 10, sd = 10)
+     fonteB <- rnorm(n = 30, mean = 12, sd = 10)
+     pVal[i] <- wilcox.test(fonteA, fonteB)$p.value
+ }
>
> cat("Média do valor de p-value para desvio = 10:", mean(pVal))
Média do valor de p-value para desvio = 10: 0.3906538>
> # ---- Média de p valor para 50 repetições média 10 e 15 ---- #
>
> for(i in 1:50) {
+     fonteA <- rnorm(n = 30, mean = 10, sd = 5)
+     fonteB <- rnorm(n = 30, mean = 15, sd = 5)
+     pVal[i] <- wilcox.test(fonteA, fonteB)$p.value
+ }
>
> cat("Média do valor de p-value para médias = 10 e 15:", mean(pVal))
Média do valor de p-value para médias = 10 e 15: 0.006907904
```

Figure 4: Imagem mostra todos os resultados da questão 3 para o teste Wilcoxon