Relatório 1

Alunos: André Gomes Regino (RA 230252) e Gustavo Caetano Borges (RA 262883)

## Variáveis de Entrada

a1 <- read.csv(file = "a1.csv", header = FALSE, sep = ";", dec = ".")  
a1 <- a1[,1]  
b1 <- read.csv(file = "b1.csv", header = FALSE, sep = ";", dec = ".")  
b1 <- b1[,1]  
a1

## [1] 1.2323944 7.7595693 14.1414943 11.0545966 2.7511814 13.9885594  
## [7] 8.0588607 6.7699532 13.6981390 9.4972725 9.5659775 4.7655213  
## [13] 14.9342686 0.2779786 1.4296083

b1

## [1] 1.5275066 4.8267059 4.6537708 4.2065837 0.9076421 6.3311087 1.4736370  
## [8] 9.2429929 8.6426045 5.0446648 9.9381620 2.7493410 6.5248544 6.0254456  
## [15] 2.5940320 8.9914781 0.1149999 6.0428365 7.4185673 4.0250625

paired <- read.csv(file = "paired.csv", header = FALSE, sep = ",", dec = ".")  
col1 <- paired[,1]  
col2 <- paired[,2]  
col1

## [1] 12.9 13.5 12.8 15.6 17.2 19.2 12.6 15.3 14.4 11.3

col2

## [1] 13.1 12.2 11.2 13.3 15.2 15.8 12.2 13.4 12.9 11.7

## Exercício 1

Os arquivos a1.csv e b1.csv contém um conjunto de medidas cada um.

Use os seguintes testes

* teste t
* Wilcoxon rank sum

t.test(a1,b1) #teste t

##   
## Welch Two Sample t-test  
##   
## data: a1 and b1  
## t = 2.0123, df = 20.915, p-value = 0.05724  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.09871863 5.96056902  
## sample estimates:  
## mean of x mean of y   
## 7.995025 5.064100

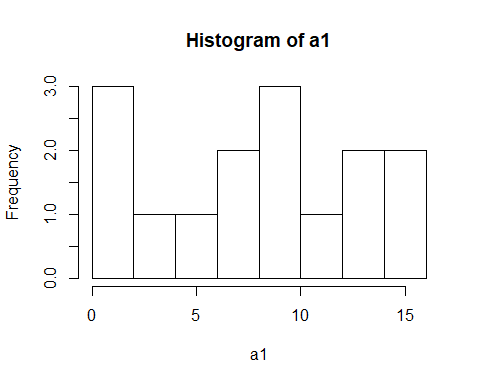
wilcox.test(a1,b1) #wilcoxon rank sum

##   
## Wilcoxon rank sum test  
##   
## data: a1 and b1  
## W = 205, p-value = 0.06887  
## alternative hypothesis: true location shift is not equal to 0

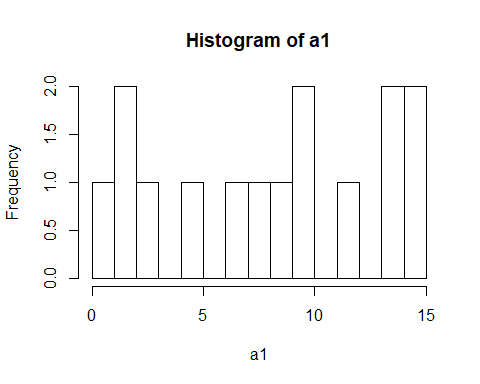
## Exercício 2

Gere um histograma dos dois dados e verifique se eles “se parecem” com Gaussianas. Tendo em vista esse resultado e o tamanho dos dados, quais dos dois valores de p-valor voce deve confiar. Eu não acho que há uma resposta correta para essa pergunta, mas eu gostaria de saber seus argumentos para a sua resposta

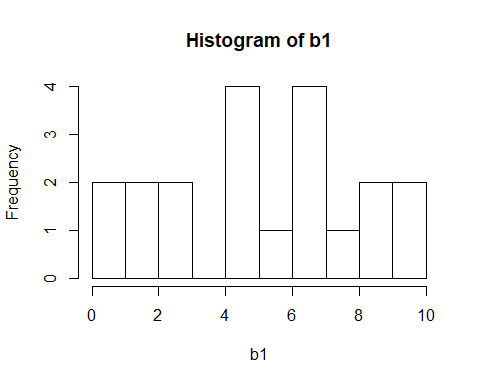
hist(a1, breaks = 9)



hist(a1, breaks = 12)



hist(b1, breaks = 9)  
hist(b1, breaks = 12)



#### Resposta

Dado os histogramas, não há semelhança com Gaussianas. Em nossa opinião, deve-se confiar menos no p-valor do teste T. O teste T assume que os dados são uma gaussiana e que há mais de 30 dados no conjunto. Isso não é verdade para o experimento apresentado, uma vez que há menos do que 30 dados no conjunto (13 e 15 dados, respectivamente) e, segundo o histograma apresentado acima, não há semelhança com gaussianas. O teste Wilcoxon faz menos pressuposições, então, seu valor deve ser mais próxima da realidade do que o teste T, no experimento em questão.

## Exercício 3

O arquivo paired.csv contem um conjunto de dados pareados, onde cada coluna é um grupo e as linhas o pareamento.

Rode os seguintes algoritmos

* teste t pareado
* Wilcoxon signed rank (a versão pareada do Wilcoxon sum rank).

Compare os dos p-valores. Discuta se voce tem uma opinião sobre quais dos dois usar. De novo eu não sei se há uma resposta certa para essa pergunta.

t.test(col1, col2, paired = TRUE) #teste t pareado

##   
## Paired t-test  
##   
## data: col1 and col2  
## t = 3.7366, df = 9, p-value = 0.00465  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.5445317 2.2154683  
## sample estimates:  
## mean of the differences   
## 1.38

wilcox.test(col1, col2, paired = TRUE) # wilcoxon pareado

##   
## Wilcoxon signed rank test  
##   
## data: col1 and col2  
## V = 52, p-value = 0.009766  
## alternative hypothesis: true location shift is not equal to 0

#### Resposta

Em nossa opinião, novamente não podemos confiar no teste T, pois os dados não satisfazem as pressuposições deste teste. O histograma mostra que não se forma curva gaussiana com os dados e o conjunto de dados é muito pequeno. Dessa forma, o wilcoxon pode ter um resultado de p-value um pouco mais aproximado da realidade.

## Exercício 4

Rode a versão não pareada do teste t e do Wilcoxon. A versão não pareada deve ser mais fraca (poder menor - maior p-valor) que as versões pareadas dos algoritmos. Verifique que isso é verdade.

t.test(col1, col2, paired = FALSE)

##   
## Welch Two Sample t-test  
##   
## data: col1 and col2  
## t = 1.5582, df = 14.856, p-value = 0.1402  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.5093412 3.2693412  
## sample estimates:  
## mean of x mean of y   
## 14.48 13.10

wilcox.test(col1, col2, paired = FALSE)

## Warning in wilcox.test.default(col1, col2, paired = FALSE): cannot compute exact  
## p-value with ties

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: col1 and col2  
## W = 67.5, p-value = 0.1984  
## alternative hypothesis: true location shift is not equal to 0

#### Resposta

Os p-valores das versões não pareadas (t = 0,1402 e wilcoxon = 0.1984) são, de fato, maiores do que os p-valores das versões pareadas (t = 0,00465 e wilcoxon = 0,00976).

## Exercício 5

Gere 2 conjuntos de 15 dados amostrados de uma normal de media 10 e 13, ambos com desvio padrão de 5.

Calcule a media do p-valor usando o teste t para 50 repetições dos pares descritos acima.

mediap=0  
for(i in 1:50){  
 set1 = rnorm(15, 10, 5)  
 set2 = rnorm(15, 13, 5)  
 mediap[i] = t.test(set1,set2)[3]  
}  
mediap <- unlist(mediap)  
mean(mediap)

## [1] 0.2090531

## Exercício 6

Calcule a media do p-valor para o teste t para 50 repetições dos pares acima, mas com 25 dados cada

mediap1=0  
for(i in 1:50){  
 set1 = rnorm(25, 10, 5)  
 set2 = rnorm(25, 13, 5)  
 mediap1[i] = t.test(set1,set2)[3]  
}  
mediap1 <- unlist(mediap1)  
mean(mediap1)

## [1] 0.1548552

## Exercício 7

Calcule a media do p-valor para o teste t para 50 repetições dos pares acima, com 15 dados cada mas com 10 como desvio padrão

mediap2=0  
for(i in 1:50){  
 set1 = rnorm(15, 10, 10)  
 set2 = rnorm(15, 13, 10)  
 mediap2[i] = t.test(set1,set2)[3]  
}  
mediap2 <- unlist(mediap2)  
mean(mediap2)

## [1] 0.4145695

## Exercício 8

Calcule a media do p-valor para o teste t para 50 repetições dos pares acima, com 15 dados, 5 de desvio padrão mas com medias 10 e 17.

mediap3=0  
for(i in 1:50){  
 set1 = rnorm(15, 10, 5)  
 set2 = rnorm(15, 17, 5)  
 mediap3[i] = t.test(set1,set2)[3]  
}  
mediap3 <- unlist(mediap3)  
mean(mediap3)

## [1] 0.005628803

## Exercício 9

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

#### Resposta

* **Número de dados:** Identificamos que quanto maior o número de dados (no exercício em questão), menor o p-valor. Exemplo: em uma das execuções, com 15 dados, obtivemos média de p-valores 0.1884612 e com 25 dados obtivemos média de p-valores 0.134518.
* **Rúido dos dados:** Identificamos que quanto maior o desvio padrão, maior o p-valor. Exemplo: em uma das execuções, com desvio padrão de 5, obtivemos média de p-valores 0.1884612 e com desvio padrão de 10 obtivemos média de p-valores 0.3657076.
* **Diferença entre fontes:** Identificamos que quanto maior a diferença entre as médias, menor o p-valor. Exemplo: em uma das execuções, com diferença entre as médias de 3 (10 e 13), obtivemos média de p-valores 0.1884612 e com diferenças de médias de 7 (10 e 17), obtivemos média de p-valores 0.003102274.