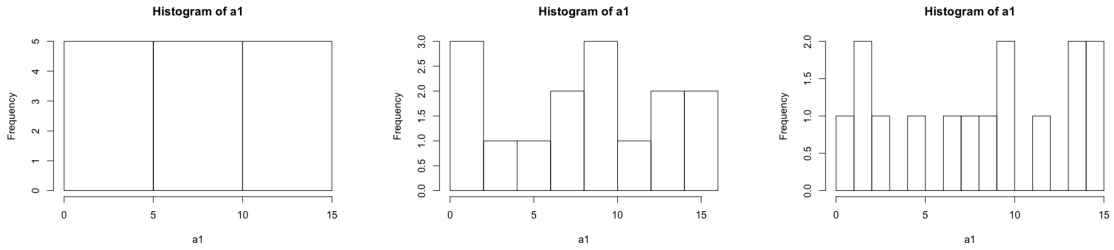
MO850 - Exercício 1

Edgar Tanaka (RA 023577)

1 de Abril de 2018

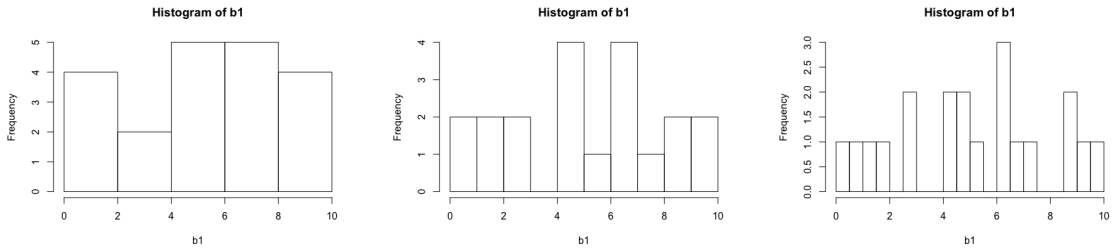
1 Dados não-pareados

Nesta primeira plotamos alguns histogramas dos dados não-pareados A1 e B1 a fim de determinar se a distribuição dos mesmo segue um formato gaussiano. Dado que o número de bins impacta significativamente essa análise visual, decidimos plotar algumas variações dos histogramas. Seguem os resultados nas Figuras 1 e 2.



(a) 3 bins. (b) 8 bins. (c) 15 bins.

Figura 1: Histogramas dos dados A1.



(a) 5 bins. (b) 10 bins. (c) 15 bins.

Figura 2: Histogramas dos dados B1.

Para os dados A1, não foi possível observar de forma clara uma curva de formato gaussiano. O histograma ou segue um padrão constante (como na Figura 1a) ou senão possui vales locais entre o centro e as bordas. A distribuição aparenta ser uniforme e não gaussiana.

Já os histogramas da Figura 2 apresentam um centro mais elevado e bordas que decaem rapidamente. Eles também apresentam vales locais entre o centro e as bordas mas é possível notar uma sutil distribuição gaussiana aqui.

Finalmente, calculamos o p-valor usando os testes de "Wilcoxon signed rank test"e o "Teste T não-pareado". A tabela 1 contém os resultados.

Teste	p-value
Wilcoxon rank sum	0.06887
T-Test não-pareado	0.05724

Tabela 1: Testes Estatísticos em dados não-pareados.

Como não é óbvio que a distribuição de ambos os dados A1 e B1 é normal, não podemos assumir que esses dados são paramétricos. Sendo assim, o teste mais adequado aqui é o teste de Wilcoxon rank sum.

2 Dados Pareados

Nesta segunda parte do exercício, usamos os dados pareados da tabela a1.csv para rodar os testes estatísticos "Wilcoxon signed rank", "Teste T pareado"e "Sign test". A tabela 2 contém os resultados.

Teste	p-value
Wilcoxon signed rank test	0.009766
Teste T pareado	0.00465
Sign test	0.1094

Tabela 2: Testes Estatísticos em dados pareados.

É possível observar que o p-valor do teste T pareado é menor que o p-valor do teste de Wilcoxon. Como foi discutido em aula, esse resultado é esperado dado que o teste T é mais poderoso (ele assume que os dados são paramétricos).

Também é possível observar que o Sign Test teve o maior p-valor de todos os testes estatísticos. Isso pode ser explicado pelo fato do Sign Test ser o menos poderoso dos três. De acordo com a wikipedia, o Sign Test é recomendado apenas quando estamos restritos às operações de comparação $<$, $>$ e $=$. Para dados numéricos, podemos usar os testes de "Wilcoxon signed-rank test"e "Teste T"que são mais poderosos.

3 Fatores que impactam o p-valor

Nesta última parte do exercício, dois grupos de dados foram sintetizados utilizando a função rnorm da linguagem R. O primeiro tinha 15 pontos de dados criados a partir de uma distribuição normal (média 10 e desvio padrão 5). O segundo seguiu uma outra distribuição normal (média 13 e desvio padrão 5). Finalmente, foi calculado o p-valor a partir do teste T não-pareado com estes dois grupos de dados. Este processo foi repetido 10 vezes e uma média dos p-valores foi calculada. Algumas variações deste experimento foram executadas em seguida mas a média dos p-valores foi calculada sempre a partir de 10 execuções. A tabela 3 apresenta os resultados.

Var	Pontos de dados	Média Grupo 1	Média Grupo 2	Desvio-padrão	P-valor médio
1	15	10	13	5	0.228249
2	25	10	13	5	0.08051806
3	15	10	17	5	0.005384237
4	15	10	13	8	0.3721346

Tabela 3: Variações do experimento.

Quando aumentamos o número de pontos de dados por grupo de 15 para 25 (variação 1 para 2), o p-valor médio caiu de 0.22 para 0.08. Isso pode ser explicado por dados paramétricos necessitarem de mais de 30 pontos de dados, algo que não era verdade na variação 1. Mesmo assim, como o p-valor não é menor que 0.05, não é possível afirmar que as duas fontes de dados são distintas.

Se agora compararmos a variação 1 e 3, o p-valor médio caiu de 0.22 para 0.005. O resultado é esperado dado que agora as curvas normais estão mais separadas e portanto fica mais evidente que os dados do grupo 1 e 2 vieram de fontes independentes. Essa diferença traz o p-valor médio a um valor menor que 0.05 permitindo concluir que as fontes dos dados do grupo 1 e 2 são distintas.

Finalmente, se compararmos a variação 1 e 4, notamos que o aumento no desvio padrão de 5 para 8 também causou um aumento no p-valor médio. Isso faz sentido uma vez que as duas distribuições do grupo 1 e 2 agora ficaram mais sobrepostas, ou seja, há uma sobreposição maior entre os dois grupos levando então a uma maior confusão sobre a independência dessas duas fontes de dados.

4 Código R

Todo o código R para este trabalho segue abaixo:

```
# Non Paired Data - histogram, wilcox and t-test
a1 = read.csv("a1.csv", stringsAsFactors=FALSE, header=FALSE)$V1
b1 = read.csv("b1.csv", stringsAsFactors=FALSE, header=FALSE)$V1
wilcox.test(a1, b1, paired=FALSE)
t.test(a1, b1, paired=FALSE)
hist(a1, breaks=10)
hist(b1, breaks=10)


# Paired Data - wilcox and t-test
a2 = read.csv("paired.csv", header=FALSE)
x = a2$V1
y = a2$V2
wilcox.test(x, y, paired=TRUE)
t.test(x, y, paired=TRUE)

# Paired data - Sign test
successes = sum(x > y)
failures = length(x) - successes
binom.test(c(successes, failures))


# Study on the factors that impact the p-value
mean_p_values <- function(num_pairs, n, mean1, mean2, stddev) {
  p_values = c()
  for (i in 1:num_pairs) {
    a = rnorm(n, mean1, stddev)
    b = rnorm(n, mean2, stddev)
    t_test = t.test(a, b, paired=FALSE)
    p_values[i] = t_test$p.value
  }

  return(mean(p_values))
}

mean_p_values(num_pairs=10, n=15, mean1=10, mean2=13, stddev=5)
mean_p_values(num_pairs=10, n=25, mean1=10, mean2=13, stddev=5)
mean_p_values(num_pairs=10, n=15, mean1=10, mean2=17, stddev=5)
mean_p_values(num_pairs=10, n=15, mean1=10, mean2=13, stddev=8)
```