Introdução ao R com Aplicações

Lupércio França Bessegato
Augusto Carvalho Souza
Dep. de Estatística/UFJF

Roteiro Geral

1. Fundamentos da linguagem R
2. Visualização e descrição de dados
3. Inferência estatística básica
4. Modelos de regressão
5. Análise de dados multivariados
6. Séries temporais
7. Referências

Inferência Estatística Básica

Exemplo

- Influência da escolaridade no rendimento.
  - √ Amostra aleatória com 186 pares de gêmeos monozigóticos
  - √ Fatores de confusão no estudo do efeito:
    - Inteligência inata, condições familiares
    - Busca-se retirar esse efeito estudando-se gêmeos
  - √ Fonte: ALBERT, J., RIZZO, M. *R by Example*. Springer, 2012.
  - √ Dados: *twins.txt* ou *twins.dat.txt*

√ Variáveis:
- DLHRWAGE: diferença no logaritmo do salário por hora, em US$ (gêmeo 1 – gêmeo 2)
- DEDUC1: diferença em escolaridade, em anos (autodeclarada)
- AGE: idade, em anos
- AGESQ: idade ao quadrado, anos^2
- HRWAGEH: salário do gêmeo 2, por hora
- WHITEH: etnia (1 = gêmeo 2 é branco, 0 = c.c.)
- MALEH: sexo (1 = gêmeo 2 é homem, 0 = c.c.)
- EDUCH: escolaridade do gêmeo 2, em anos (autodeclarada)

- HRWAGEL: salário do gêmeo 1, por hora
- WHITEL: etnia (1 = gêmeo 1 é branco, 0 = c.c.)
- MALEL: sexo (1 = gêmeo 2 é homem, 0 = c.c.)
- EDUCL: escolaridade do gêmeo 1, em anos (autodeclarada)
- DEDUC2: diferença de escolaridade de acordo com informações cruzadas, em anos. (escolaridade do gêmeo 1, informada pelo gêmeo 2 e vice-versa)
- DTEN: diferença em permanência no trabalho atual, em anos
- DMARRIED: diferença de estado civil dos gêmeos (1 = casado, 0 = solteiro).
- DUNCOV: diferença em cobertura por trabalho sindicalizado (1 = coberto, 0 = c.c.)

- Dados obtidos por autodeclaração e informação dada pelo irmão
- Identificação de sufixo de variáveis
  - √ L: gêmeo 1; H: gêmeo 2
- Diferenças: (gêmeo1 – gêmeo2)
- Informações sobre arquivo de dados:
  - √ Dados faltantes indicados por ponto '.'.
  - √ Variáveis estão separadas por ','

- Importação e preparação dos dados:

```
> # carregamento direto do pacote
> gemeos1 <- read.table("twins.dat.txt", header=TRUE, sep = ",")
> dim(gemeos1)
[1] 183  16
> str(gemeos1)
''data.frame':   183 obs. of  16 variables:
 $ DLHRWAGE: Factor w/ 125 levels "-0.001153403",..: 81 51 115 53 36 51 122 51 ...
 $ DEDUC1  : int  0 -1 7 0 0 2 -2 -1 -2 0 ...
 $ AGE     : num  33.3 54.1 43.6 31 34.6 ...
 $ AGESQ   : num  1106 2922 1898 959 1200 ...
 $ HRWAGEH : Factor w/ 100 levels ".","1.785714286",..: 12 1 35 31 97 1 53 53 ...
 $ WHITEH  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ MALEH   : int  0 0 0 1 1 0 0 0 1 0 ...
 $ EDUCH   : int  16 9 19 12 14 16 13 13 12 12 ...
 $ HRWAGEL : Factor w/ 112 levels ".","1.666666667",..: 104 101 106 35 40 1 98 ...
 $ WHITEL  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ MALEL   : int  0 0 0 1 1 0 0 0 1 0 ...
 $ EDUCL   : int  16 10 12 12 14 14 15 14 14 12 ...
 $ DEDUC2  : int  0 1 4 0 1 -2 -2 -2 1 0 ...
 $ DTEN    : Factor w/ 91 levels "-0.083","-0.084",..: 53 89 76 24 65 70 76 ...
 $ DMARRIED: int  0 1 -1 0 0 1 1 0 -1 0 ...
 $ DUNCOV  : int  0 0 0 1 -1 0 0 0 0 0 ...
> # qte. dados faltantes
> sum(is.na(gemeos1))
[1] 0
```

Banco não foi carregado corretamente!

- Importação com instrução para NA's:

```
> # carregamento dos dados c/ instrução sobre NA'
> gemeos <- read.table("twins.dat.txt", header=TRUE, sep=",", na.strings=".")
> str(gemeos)
'data.frame':   183 obs. of  16 variables:
 $ DLHRWAGE: num  0.2593 NA 0.7213 0.0116 -0.561 ...
 $ DEDUC1  : int  0 -1 7 0 0 2 -2 -1 -2 0 ...
 $ AGE     : num  33.3 54.1 43.6 31 34.6 ...
 $ AGESQ   : num  1106 2922 1898 959 1200 ...
 $ HRWAGEH : num  11.25 NA 18 16.5 9.62 ...
 $ WHITEH  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ MALEH   : int  0 0 0 1 1 0 0 0 1 0 ...
 $ EDUCH   : int  16 9 19 12 14 16 13 13 12 12 ...
 $ HRWAGEL : num  8.68 7.85 8.75 16.31 16.85 ...
 $ WHITEL  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ MALEL   : int  0 0 0 1 1 0 0 0 1 0 ...
 $ EDUCL   : int  16 10 12 12 14 14 15 14 14 12 ...
 $ DEDUC2  : int  0 1 4 0 1 -2 -2 -2 1 0 ...
 $ DTEN    : num  1.33 8 3 -2 2.92 ...
 $ DMARRIED: int  0 1 -1 0 0 1 1 0 -1 0 ...
 $ DUNCOV  : int  0 0 0 1 -1 0 0 0 0 0 ...
> sum(is.na(gemeos))
[1] 81
```


√ Conjunto de dados carregado:

```
> # conjunto de dados carregados
> head(gemeos)
     DLHRWAGE DEDUC1      AGE     AGESQ HRWAGEH WHITEH MALEH EDUCH HRWAGEL
1  0.25934660      0 33.25120 1105.6422 11.2500      1     0    16    8.68
2          NA     -1 54.05339 2921.7688       NA      1     0     9    7.85
3  0.72131806      7 43.57016 1898.3586 18.0000      1     0    19    8.75
  WHITEL MALEL EDUCL DEDUC2   DTEN DMARRIED DUNCOV
1      1     0    16      0  1.333        0      0
2      1     0    10      1  8.000        1      0
3      1     0    12      4  3.000       -1      0
```

√ Distribuição do dados faltantes

```
> # distribuição dos NA's
> sapply(gemeos, function(x) sum(is.na(x)))
DLHRWAGE    DEDUC1       AGE     AGESQ   HRWAGEH    WHITEH     MALEH     EDUCH
      34         0         0         0        22         0         0         0
 HRWAGEL    WHITEL     MALEL     EDUCL    DEDUC2      DTEN DMARRIED   DUNCOV
      21         0         0         0         0         4         0         0
> # Quantidade de individuos s/ dados completos
> sum(!complete.cases(gemeos))
[1] 36
```


√ Localização elementos sem dados completos

```
> # localização dos sujeitos s/ dados completos
> falta <- which(is.na(gemeos), arr.ind = TRUE) # arr.ind=T obtém linha e coluna
> gemeos[unique(as.vector(falta[,1])), ]
     DLHRWAGE DEDUC1      AGE     AGESQ   HRWAGEH WHITEH MALEH EDUCH    HRWAGEL
2          NA     -1 54.05339 2921.7688        NA      1     0     9   7.850000
6          NA      2 71.60301 5126.9913        NA      1     0    16         NA
8          NA     -1 61.45106 3776.2329 35.000000      1     0    13         NA
...
    WHITEL MALEL EDUCL DEDUC2    DTEN DMARRIED DUNCOV
2        1     0    10      1   8.000        1      0
6        1     0    14     -2  24.000        1      0
8        1     0    14     -2  25.500        0      0
...
```


- Variáveis que medem escolaridade:
  - √ EDUCL (gêmeo 2) e EDUCH (gêmeo 2)
  - √ Valores autodeclarados

```
> # tabela de frequência escolaridade
> table(gemeos$EDUCL)

 8 10 11 12 13 14 15 16 17 18 19 20
 1  4  1 61 21 30 11 37  1 10  3  3
> table(gemeos$EDUCH)

 8  9 10 11 12 13 14 15 16 17 18 19 20
 2  1  2  1 65 22 22 15 33  2 11  2  5
```

  - √ Escolaridade apresenta muita variação
  - √ Modas:
    - 12 (*high school*) e 16 (*college*)

- Categorização das variáveis EDU:
  - < 12 = *high school*; 13 a 15 = *some college*; 16 = *college*; > 16 = *graduate school*

```
> # categorização das variáveis de escolaridade
> c.EDUCL <- cut(gemeos$EDUCL, breaks=c(0, 12, 15, 16, 24),
+    labels=c("High School", "Some College", "College Degree",
+ "Graduate School"))
> c.EDUCH <- cut(gemeos$EDUCH, breaks=c(0, 12, 15, 16, 24),
+     labels=c("High School", "Some College", "College Degree",
+   "Graduate School"))
```

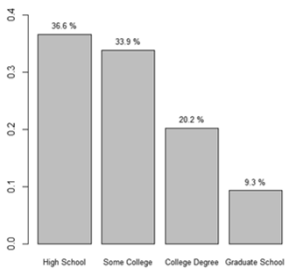
√ Resumo tabular por categoria

```
> table(c.EDUCL)
c.EDUCL
    High School    Some College  College Degree Graduate School
             67              62              37              17
> table(c.EDUCH)
c.EDUCH
    High School    Some College  College Degree Graduate School
             71              59              33              20
```

√ Maioria está entre os níveis *high school* e *some college*

- *Bar plot* da variável c.EDUCL:

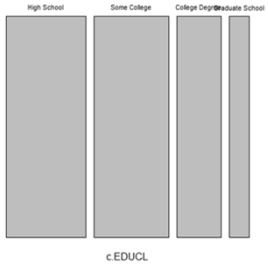
```
> c.EDUCL.tab <- prop.table(table(c.EDUCL))
> EDUCL.bar <- barplot(prop.table(c.EDUCL.tab), cex.names = 0.8,
+    ylim = c(0, 0.40))
> text(EDUCL.bar, EDUCL.tab, labels = paste(round(EDUCL.tab*100, 1),"%"),
+ cex = 0.85, pos = 3, offset = 0.5)
```



√ 70,5% dos gêmeos 1 estão entre os níveis *high school* e *some college*

- Mosaic plot:

```
> mosaicplot(table(c.EDUCL), main = "")
```



√ Largura das barras proporcional à frequência

√ Muito útil para visualizar associação entre duas variáveis categóricas

- Tabela de contingência das escolaridades:

```
> T1 <- table(c.EDUCL, c.EDUCH)
> T1
                c.EDUCH
c.EDUCL          High School Some College College Degree Graduate School
  High School             47           16              2               2
  Some College            18           32              8               4
  College Degree           5           10             18               4
  Graduate School          1            1              5              10
```

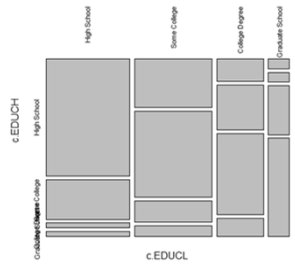
√ Pares com mesmo nível ocupacional

```
> # Pares com mesmo nível educacional
> diag(T1)
 High School    Some College  College Degree Graduate School
            47              32              18              10
> sum(diag(T1)) / sum(T1)
[1] 0.5846995
```

- Proporção de gêmeos com mesmo nível educacional: 58,5%

• Mosaic plot da tabela de contingência:

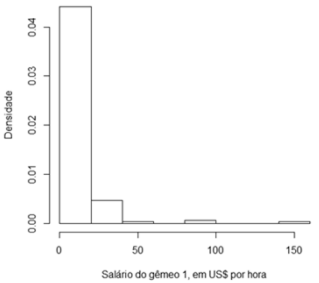
```
> plot(T1, las = 3, main = "")
```



√ Áreas correspondem às contagens na tabela
√ Maiores áreas:
- *High scholl/High scholl*
- *Some college/Some college*

• Histograma do rendimento do gêmeo 1:

```
> hist(gemeos$HRWAGEL, freq = F, main = "", ylab = "Densidade",
       xlab = "Salário do gêmeo 1, em US$ por hora")
```

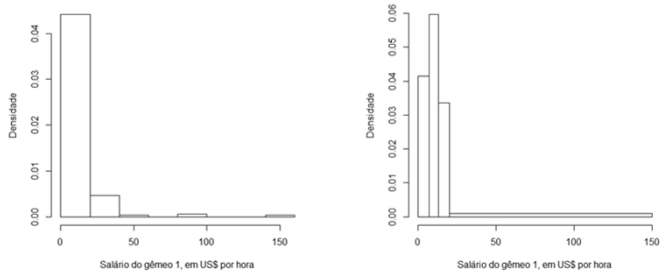


√ Forte assimetria à direita

• Mudança de intervalos de classe:

√ Extremos de classe: 0, 7, 13, 20 e 150

```
> hist(gemeos$HRWAGEL, breaks = c(0, 7, 13, 20, 150), freq = F, main = "",
+ ylab = "Densidade", xlab = "Salário do gêmeo 1, em US$ por hora")
```



√ Facilita a percepção da assimetria

• Categorização do rendimento

```
> c.wage = cut(gemeos$HRWAGEL, c(0, 7, 13, 20, 150))
> table(c.wage)
c.wage
   (0,7]   (7,13]  (13,20] (20,150]
      47       58       38       19
> sum(is.na(gemeos$HRWAGEL))
[1] 21
```

√ Há 21 dados faltantes

- Investigação da relação entre escolaridade e salário
  - √ Tabela de contingência entre categorias

```
> T2 <- table(c.EDUCL, c.wage)
> prop.table(T2, margin = 2)
                    c.wage
c.EDUCL               (0,7]     (7,13]    (13,20]   (20,150]
  High School     0.48936170 0.36206897 0.26315789 0.05263158
  Some College    0.31914894 0.39655172 0.31578947 0.26315789
  College Degree  0.14893617 0.20689655 0.36842105 0.15789474
  Graduate School 0.04255319 0.03448276 0.05263158 0.52631579
```

  - √ `margin = 1`: proporções condicionadas às colunas
  - √ Sugestão de que quanto maior o nível educacional, maior o rendimento

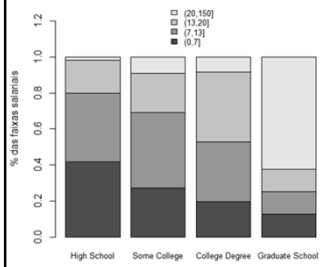
- Proporções das classes de renda, condicionadas à escolaridade

```
> P <- prop.table(T2, margin = 1)
> P
                    c.wage
c.EDUCL               (0,7]     (7,13]    (13,20]   (20,150]
  High School     0.41818182 0.38181818 0.18181818 0.01818182
  Some College    0.27272727 0.41818182 0.21818182 0.09090909
  College Degree  0.19444444 0.33333333 0.38888889 0.08333333
  Graduate School 0.12500000 0.12500000 0.12500000 0.62500000
```

  - √ Linhas somam 1 (100 %)
  - √ Participação maior da classe de rendimento mais elevados nos níveis mais altos de escolaridade

- Visualização bivariada categórica:
  - √ Gráfico de barras empilhado:

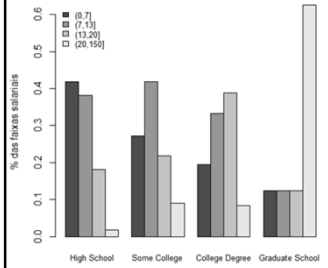
```
> barplot(t(P), ylim = c(0, 1.3), ylab = " % das faixas salariais",
+     legend.text = dimnames(P)$c.wage, cex.names = 0.8,
+     args.legend = list(x = "top", bty = "n", cex = 0.8))
```



  - √ Barplot em matriz: barras estratificadas das colunas da matriz
  - √ Áreas correspondem a proporções
  - √ Gráfico das linhas da tabela P
  - √ Regiões mais claras correspondem aos maiores salários
    - − Maiores à medida que se vai à direita
  - √ Indicação de que níveis educacionais maiores têm maiores salários

- Visualização bivariada categórica:
  - √ Gráfico de barras lado a lado:

```
> barplot(t(P), beside = T, legend.text = dimnames(P)$c.wage, cex.names = 0.8,
+     args.legend = list(x = "topleft", bty = "n", cex = 0.8),
+     ylab = "% das faixas salariais")
```



  - √ Cada barra corresponde a proporção para um nível educacional
  - √ Salários menores:
    - − predominam para níveis *High School* e *Some College*
    - − Improváveis para nível *Graduate School*

Teste de Independência

- Tabela de contingência:
  - √ Se rendimento e nível educacional forem independentes:
    - Probabilidade de um par de gêmeos pertencer às 4 categorias de salários não depende de seu nível educacional
- Conclusão empírica:
  - √ Níveis educacionais mais altos aparentam estar associados com salários mais altos

- Procedimento de teste:
  - √ H_0: nível de educacional e salário são independentes
  - √ Execução do teste – para amostras grandes

```
> S <- chisq.test(T2)
> print(S)
 Pearson's Chi-squared test

data:  T2
X-squared = 54.578, df = 9, p-value = 1.466e-08
```

  - √ Quantidade esperada sob H_0 (independência)

```
> outer(margin.table(T2, 1)/sum(T2), margin.table(T2, 2)/sum(T2))*sum(T2)
c.wage
c.EDUCL              (0,7]    (7,13]   (13,20] (20,150]
  High School     15.956790 19.691358 12.901235 6.450617
  Some College    15.956790 19.691358 12.901235 6.450617
  College Degree  10.444444 12.888889  8.444444 4.222222
  Graduate School  4.641975  5.728395  3.753086 1.876543
> S$expected
```


- √ Estatística de teste

$$X^2 = \sum_{\substack{\text{todas}\\ \text{células}}} \frac{(\text{observado} - \text{esperado})^2}{\text{esperado}}$$

```
> # estatística X^2
> sum((T2 - S$expected)^2 / S$expected)
[1] 54.57759
> S$statistic
X-squared
 54.57759
```

- √ p-valor: $p = \mathrm{P}\{\chi^2_{\mathrm{df}} \geq X^2\}$
  - `df` = (# linhas – 1) x (#colunas – 1)

```
> # p-valor
> > 1 - pchisq(54.57759, df=9)
[1] 1.465839e-08
> S$p.value
[1] 1.465839e-08
```


- Objeto do teste χ^2:

```
> S <- chisq.test(T2)
> print(S)
 Pearson's Chi-squared test

data:  T2
X-squared = 54.578, df = 9, p-value = 1.466e-08
```

```
> names(S)
[1] "statistic" "parameter" "p.value"   "method"    "data.name" "observed"
[7] "expected"  "residuals" "stdres"
```

Resíduos

$$\text{resíduo} = \frac{(\text{observado} - \text{esperado})}{\sqrt{\text{esperado}}}$$

- Informalmente:
  - √ Qualquer resíduo maior que 2, em valor absoluto, indica desvio significativo da hipótese de independência

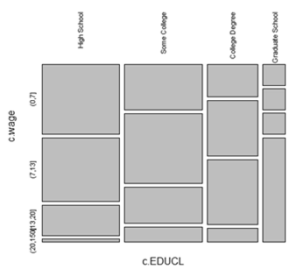
- No exemplo:

```
> S$residuals
 c.wage
c.EDUCL               (0,7]     (7,13]    (13,20]   (20,150]
  High School      1.7631849  0.2949056 -0.8077318 -2.1460758
  Some College    -0.2395212  0.7456104 -0.2509124 -0.5711527
  College Degree  -1.0658020 -0.2475938  1.9117978 -0.5948119
  Graduate School -1.2262453 -1.5577776 -0.9049176  5.9300942
```

  - √ Valores maiores que 2:
    - –2,14: há menos *High school* ganhando salários superiores a 20 do que o antecipado pelo modelo de independência
    - 5,93: há mais *Graduate school* ganhando acima de $20 que o esperado de variáveis independentes
  - √ Nível educacional é mais importante nas faixas salariais mais altas

- *Mosaic plot*:

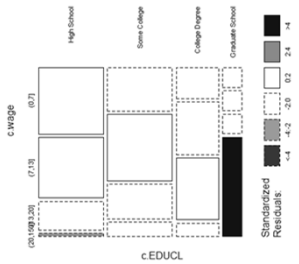
```
> plot(T2, shade = FALSE, las = 3, main = "")
```



  - √ `shade = FALSE`:
    - –Áreas correspondem às contagens nas categorias de níveis educacionais e salários

- *Mosaic plot*:

```
> plot(T2, shade = TRUE, las = 3, main = "")
```



  - √ `shade = TRUE`:
    - –Tipo de sombreamento e de borda relacionam-se com os tamanhos dos resíduos de Pearson.
  - √ Maiores resíduos verificados estão representados pelos retângulos sombreados

Comparação de Variável Quantitativa

- Objetivo:
  - √ Comparar `log(wage)` em dois grupos
    - – (máximo de 12 anos de escolaridade e com mais de 12 anos de escolaridade)
  - √ As médias de `log(wage)` dos dois grupos são iguais?

- Transformação das variáveis:

```
> # Salários na escala log
> log.wages <- log(gemeos$HRWAGEH)
> # categorização da escolarida
> superior <- ifelse(gemeos$EDUCH > 12, "sim", "não")
```

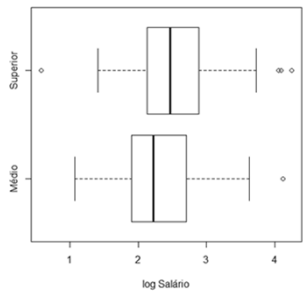
- Tabela de frequência dos grupos

```
> table(superior[complete.cases(log.wages)])

não sim
 58 103
```


- *Box-plot*: `log(wage)` vs. escolaridade:

```
> boxplot(log.wages ~ superior, horizontal = TRUE, xlab = "log Salário",
+ names=c("Médio", "Superior"))
```



- √ Dispersão aparenta ser a mesma nos 2 grupos
- √ Distribuições nos 2 grupos parecem ser simétricas

Escala `log`

- Correção de assimetria à direita
- Ao considerar razões na escala original
- Em modelos com erros multiplicativos
  - √ Ex.: alguns modelos de concentrações ou taxas
- Para considerar ordens de grandeza (usando log base 10)
  - √ Ex.: distâncias astronômicas
- Em geral, contagens são transformadas
  - √ Verificar o problema de contagens zero

Teste t

- H_0: $\mu_M = \mu_S$ ou H_0: $\mu_M - \mu_S = 0$
  - √ Importante:
    - Diferença na escala log é razão na escala original
    - Log(mediana) = mediana dos logs
- Teste t de amostras independentes:
  - √ Assume homocedasticidade
- Teste t corrigido (de Welch)
  - √ Considera heterocedasticidade
  - √ Default do comando `t.test`.

- Teste de Welch
  - √ Teste t para populações heterocedásticas

```
> t.test(log.wages ~ superior)

        Welch Two Sample t-test

data:  log.wages by superior
t = -2.4545, df = 131.24, p-value = 0.01542
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.42999633 -0.04620214
sample estimates:
mean in group não mean in group sim
        2.282119          2.520218
```

  - √ Correção se dá nos graus de liberdade
  - √ IC de 95% para a diferença das médias dos logs de salário nos dois grupos

- Teste t de amostras independentes
  - √ Teste t para populações homocedásticas

```
> t.test(log.wages ~ superior, var.equal = TRUE)

        Two Sample t-test

data:  log.wages by superior
t = -2.3683, df = 159, p-value = 0.01907
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.43665519 -0.03954328
sample estimates:
mean in group não mean in group sim
        2.282119          2.520218
```

  - √ `gl` = 58 + 103 – 2 = 159
  - √ P-valor leva à mesma conclusão
    - 0,01542 (heterocedásticidade)
    - 0,01907 (homocedasticidade)

- Objeto do teste t:

```
> log.wages.t <- t.test(log.wages ~ superior)
> str(log.wages.t)
List of 9
 $ statistic  : Named num -2.45
  ..- attr(*, "names")= chr "t"
 $ parameter  : Named num 131
  ..- attr(*, "names")= chr "df"
 $ p.value    : num 0.0154
 $ conf.int   : atomic [1:2] -0.43 -0.0462
  ..- attr(*, "conf.level")= num 0.95
 $ estimate   : Named num [1:2] 2.28 2.52
  ..- attr(*, "names")= chr [1:2] "mean in group não" "mean in group sim"
 $ null.value : Named num 0
  ..- attr(*, "names")= chr "difference in means"
 $ alternative: chr "two.sided"
```

  - √ Intervalo de 95% de confiança para a diferença da média de `log(wage)`:

```
> log.wages.t$conf.int[1:2]
[1] -0.42999633 -0.04620214
```

- Estimativa pontual da diferença na escala log

```
> diff(log.wages.t$estimate)
mean in group sim
        0.2380992
```

- Estimativa na escala original

```
> exp(diff(log.wages.t$estimate))
mean in group sim
         1.268835
```

√ Interpretação:

– A média geométrica dos salários do grupo com escolaridade superior é 1,27 vezes maior que aquela do grupo sem nível superior

– A razão da média geométrica de salários do grupo com escolaridade superior daquele sem nível superior é 1,27.

41

- IC de 95% da diferença na escala log

```
> log.wages.t$conf.int[1:2]
[1] -0.42999633 -0.04620214
```

- IC de 95% da razão das médias geométrico do grupo sem nível superior e com escolaridade superior, na escala original

```
> exp(log.wages.t$conf.int[1:2])
[1] 0.6505115 0.9548489
```

42

- Outras considerações sobre a escala log

√ Se os dados na escala log forem fortemente simétricos

– A diferença das medianas de salários é ...

√ O logaritmo é a única transformação não linear que produz resultados que podem ser expressos de forma clara em termos dos dados originais

43

Intervalo de Confiança

- Tem-se confiança no procedimento de construção do intervalo

√ Não se sabe se o intervalo contém ou não a média verdadeira

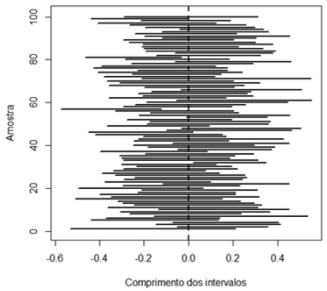
- Construção de 100 intervalos:

```
> n.rep <- 100
> mi <- 0
> n <- 24
> dp <- sd(log.wages, na.rm = TRUE)
> dp
[1] 0.6211594
> amostras <- matrix(rnorm(n.rep * n, mi, dp), n)
> int.conf <- function(x) t.test(x)$conf.int
> intervalos <- apply(amostras, 2, int.conf)
> plot(range(intervalos), c(0, 1 + n.rep), type = "n", ylab = "Amostra",
+ xlab = "Comprimento dos intervalos")
> for (i in 1:n.rep) lines(intervalos[, i], rep(i, 2), lwd = 2)
> abline(v = mi, lwd = 2, lty = 2)
```

44

- Representação dos 100 IC's:

```
> sum(intervalos[1, ] <= mi & intervalos[2, ] >= mi)
[1] 96
```



√ 4 dos 100 intervalos não contêm a média veradeira

45

Teste de Mann-Witney para 2 Amostras

- Não necessita de hipótese sobre a distribuição subjacente dos dados
  - √ Trabalha com a ordem das observações
- Pode-se trabalhar com variáveis ordinais
- Há perda de informações
  - √ Teste t é mais poderoso se suas suposições forem verdadeira

46

- Teste de Mann-Witney

```
> wilcox.test(log.wages ~ superior, conf.int = TRUE)

        Wilcoxon rank sum test with continuity correction

data:  log.wages by superior
W = 2264, p-value = 0.01093
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -0.44266384 -0.06455011
sample estimates:
difference in location
            -0.2575775
```

√ IC de 95% é similar ao obtido para a diferença das médias populacionais (escala log)

47

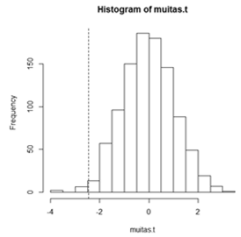
Teste de Permutação

- Outro procedimento para comparar 2 amostras
- Procedimento:
  - √ Alocação aleatória dos 71 gêmeos com ensino superior e dos 112 que não passaram
  - √ Calcula-se a estatística de teste dos dados permutados
  - √ Repetir para um grande número de iterações
  - √ Valor observado da estatística é comparado com a distribuição das réplicas da permutação

48

- Execução de teste de permutação:

```
> # cria função da estatística de teste
> resample <- function() t.test(log.wages ~ sample(superior))$statistic
> # executa as réplicas
> muitas.t = replicate(1000, resample())
> # estatística observada - dados
> t.obs = t.test(log.wages ~ superior)$statistic
> hist(muitas.t)
> abline(v = t.obs, lty = 2)
> # p-valor
> 2 * mean(muitas.t < t.obs)
[1] 0.018
```



√ P-valor é similar àqueles calculados usando o teste t e o de Mann-Witney

Comparação Pareada das Médias

- Objetivo do estudo:
  - √ Comparar os salários dos dois grupos de pessoas com níveis educacionais diferentes
- Há dificuldades em se obter estimativas precisas do efeito da educação no salário
  - √ Há muitas variáveis de confusão que também podem explicar a diferença entre os 2 grupos
    - Situação familiar, inteligência inata, habilidades naturais.

- Uma possível solução
  - √ Comparar a diferença de salários entre gêmeos.
- Criação de data frame apenas com gêmeos que têm diferentes níveis de educação

```
> gemeos.dif <- subset(gemeos, EDUCL != EDUCH)
> gemeos.dif <- gemeos.dif[complete.cases(gemeos.dif), ])
```


- Criação de variáveis:
  - √ `log.wage.baixo`: log(salário) para o gêmeo com menor nível educacional
  - √ `log.wage.alto`: log(salário) para o gêmeo com maior nível educacional

```
> # log salário - gêmeo com menor escolaridade
> log.wage.baixo <- with(gemeos.dif,
+ ifelse(EDUCL < EDUCH, log(HRWAGEL), log(HRWAGEH)))
> # log salário - gêmeo com maior escolaridade
> log.wage.alto <- with(gemeos.dif,
+ ifelse(EDUCL < EDUCH, log(HRWAGEH), log(HRWAGEL)))
> head(cbind(log.wage.baixo, log.wage.alto))
     log.wage.baixo log.wage.alto
[1,]       2.169054      2.890372
[2,]       3.555348      2.032088
[3,]       2.484907      2.708050
[4,]       2.847812      2.796061
[5,]       2.748872      3.218876
[6,]       2.079442      2.708050
```

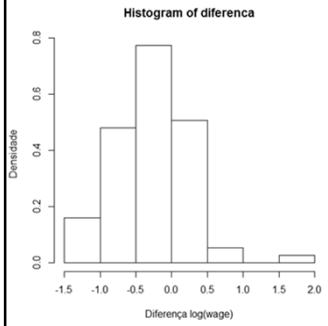
- Médias populacionais:
  - √ μ_B: média de log-salário do gêmeo com menor nível educacional
  - √ μ_B: média de log-salário do gêmeo com maior nível educacional
  - √ Diferença de interesse: $\mu_B - \mu_A$.
- Diferenças observadas

```
> # diferenças entre os log salários
> diferenca <- log.wage.baixo - log.wage.alto
```

53

- Histograma: diferenças de `log(wage)`

```
> # diferenças entre os log salários
> diferenca <- log.wage.baixo - log.wage.alto
```

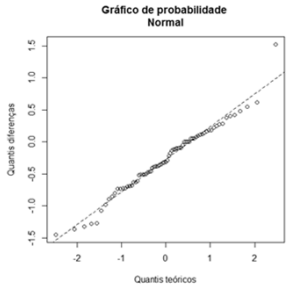


- √ Diferenças emparelhadas de `log(wage)` parecem ser aproximadamente normais

54

- Verificação da normalidade das diferenças

```
> qqnorm(diferenca, ylab = "Quantis diferenças", xlab = "Quantis teóricos",
+  main="Gráfico de probabilidade \nNormal")
> qqline(diferenca, lty = 2)
> shapiro.test(diferenca)
        Shapiro-Wilk normality test
data:  diferenca
W = 0.97627, p-value = 0.1711
```



- √ Não há evidências amostrais para se rejeitar a hipótese de normalidade das diferenças

55

- Teste t emparelhado

```
> log.wage.par <- t.test(log.wage.baixo, log.wage.alto, paired = TRUE)
> print(log.wage.par)

        Paired t-test
data:  log.wage.baixo and log.wage.alto
t = -4.5516, df = 74, p-value = 2.047e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3930587 -0.1537032
sample estimates:
mean of the differences
              -0.273381
```

- √ Diferença da média de log(wage) é estatisticamente significante
- √ Estimativa de média geométrica na escala original

```
> exp(log.wage.par$estimate
mean of the differences
              0.7608029
```

56

- Intervalo de confiança para a diferença de médias aritméticas – log escala

```
> > log.wage.par$conf.int[1:2]
[1] -0.3930587 -0.1537032
```

√ Comentário

- Intervalo de confiança para a razão das médias geométricas – escala original

```
> exp(log.wage.par$conf.int[1:2])
[1] 0.6749891 0.8575265
```

√ Comentário

Exemplo

- Frequência crítica de cintilação:
  - √ Amostra com 19 sujeitos com diferentes cores de olhos.
  - √ Variáveis:
    - Colour: cor dos olhos. (Blue = azuis, Brown = castanhos, Green = verdes)
    - Flicker: Frequência de cintilação, em ciclos/s
  - √ Fonte: ALBERT, J., RIZZO, M. *R by Example*. Springer, 2012.
  - √ Dados: *flicker.txt*

Pergunta de Interesse

- A frequência de cintilação está relacionada com as cores dos olhos?
  - √ Variável resposta:
    - `flicker`: variável contínua
  - √ Variável explicativa:
    - `Colour`: fator
- Procedimento:
  - √ Comparação de médias de 3 populações

Análise de Variância

- Comparação de médias da resposta de duas ou mais populações
  - √ H_0: Resposta média é igual para todos os grupos
  - √ H_1: Há diferença entre pelo menos duas médias dos grupos
  - √ Rejeitar H_0 não implica conseguir localizar as diferenças

- Importação dos dados:

```
> # carregamento direto da web
> cintila <- read.table(file="http://www.statsci.org/data/general/flicker.txt",
+   header=TRUE)
> # carregamento de arquivo de dados
> cintila <- read.table("flicker.txt", header = TRUE)
> dim(cintila)
[1] 19  2
> str(cintila)
'data.frame':   19 obs. of  2 variables:
 $ Colour : Factor w/ 3 levels "Blue","Brown",..: 2 2 2 2 2 2 2 3 3 ...
 $ Flicker: num  26.8 27.9 23.7 25 26.3 24.8 25.7 24.5 26.4 24.2 ...
> attach(cintila)
```

√ Fator `Colour`:

```
> # fator Colour
> is.factor(cintila$Colour)
[1] TRUE
> levels(cintila$Colour)
[1] "Blue"  "Brown" "Green“
> unclass(cintila$Colour)
 [1] 2 2 2 2 2 2 2 2 3 3 3 3 3 1 1 1 1 1 1
attr(,"levels")
[1] "Blue"  "Brown" "Green“
```

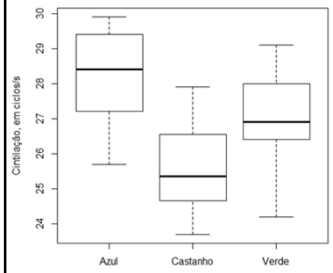

- Análise descritiva:

√ Fator `Colour`:

```
> attach(cintila)
> table(Colour)
Colour
 Blue Brown Green
    6     8     5
```


- *Boxplot* de `flicker` vs. `Colour`:

```
> # Boxplot flicker vs. Colour
> nomes <- c("Azul", "Castanho", "Verde")
> boxplot(Flicker ~ Colour, ylab = "Cintilação, em ciclos/s",
+  names = nomes)
> # boxplot - alternativa com plot
> plot(Colour, Flicker, ylab = "Cintilação, em ciclos/s", xaxt = "n")
> axis(1, at = c(1, 2, 3), labels = nomes)
```

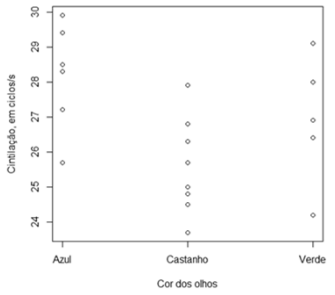


√ Amostras aparentam ser similares com relação à variância

√ Locações parecem ser diferentes

- *Stripchart* de `flicker` vs. `Colour`:

```
> stripchart(Flicker ~ Colour, vertical = TRUE, group.names = nomes, pch = 1,
+ xlab = "Cor dos olhos", ylab = "Cintilação, em ciclos/s")
```



√ *Scatter plot* com fator

- Visualização gráfica – Comentários:
  - √ Gráficos sugerem que as amostras são similares com relação à variância, mas possivelmente com locações distintas
  - √ Grupos ‘Azul’ e ‘Castanho’ parecem estar mais afastados que ‘Azul’ e ‘Verde’ ou ‘Castanho’ e ‘Verde’
  - √ Não é simples dizer se as diferenças são significativas

- Médias e desvios padrão por grupo:

```
> # Comparação das médias e desvios padrão por grupos
> media.dp = function(x) c(mean=mean(x), sd=sd(x))
> by(Flicker, Colour, FUN = media.dp)
Colour: Blue
     mean        sd
28.166667  1.527962
------------------------------------------------------------
Colour: Brown
     mean        sd
25.587500  1.365323
------------------------------------------------------------
Colour: Green
     mean        sd
26.920000  1.843095
```

  - √ Desvios padrão dos 3 grupos são próximos
  - √ Médias dos 3 grupos aparentam estar próximas
  - √ Há diferenças significantivas entre elas?

Anova a um Fator

- Hipóteses
  - √ H_0: $\mu_1 = \mu_2 = ... = \mu_a$
  - √ H_1: há diferença entre pelo menos 2 médias
    - μ_j: média populacional do j-ésimo nível do fator
  - √ Se H_0 é verdadeira, se não há diferenças entre as médias:
    - Erro quadrático médio entre as amostras e dentro das amostras estimam o parâmetro

- Funções para condução Anova a 1 fator
  - √ `oneway.test`
  - √ `lm`
  - √ `aov`
- Anova é válida se:
  - √ Erros são i.i.d. N(0, σ^2)
    - Normais
    - Independentes
    - Mesma variância (homocedásticos)

- Comando `oneway.test`:
  - √ Caso heterocedástico

```
> # comando oneway - heterocedástico
> oneway.test(Flicker ~ Colour)

        One-way analysis of means (not assuming equal variances)

data:  Flicker and Colour
F = 5.0505, num df = 2.0000, denom df = 8.9259, p-value = 0.03412
```

- √ Evidências amostrais de haver diferenças significativa entre as médias
- √ Graus de liberdade do denominador não inteiro
  - Ajustados para corrigir variâncias desiguais

- Comando `oneway.test`:
  - √ Caso homocedástico:

```
> # comando oneway - homocedástico
> oneway.test(Flicker ~ Colour, var.equal = TRUE)

        One-way analysis of means

data:  Flicker and Colour
F = 4.8023, num df = 2, denom df = 16, p-value = 0.02325
```

- √ Conclusão não muda
- √ Há evidências amostrais de ocorrer diferenças significativas entre as médias
  - Não há informações sobre quais delas diferem
- √ Saída não apresenta tabela ANOVA, resíduos ou valores ajustado

Especificação do Modelo

$$Y_{ij} = \mu_j + \epsilon_{ij} \begin{cases} i = & 1, 2, \ldots, n_j \\ j = & 1, 2, \ldots, a \end{cases}$$

- √ Média do j-ésimo tratamento: μ_j
- √ Erros do modelo: $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$
- √ Resíduos: $(y_{ij} - \bar{y}_{ij})$
- √ Parâmetros do modelo: $\mu_1, \mu_2, \ldots, \mu_a, \sigma^2$
  (a + 1) parâmetros
- √ Erro quadrado médio (EQM) estima σ^2.

Modelo de Efeitos

$$Y_{ij} = \mu + \tau_j + \epsilon_{ij} \begin{cases} i = & 1, 2, \ldots, n_j \\ j = & 1, 2, \ldots, a \end{cases}$$

- √ Efeito do j-ésimo tratamento: μ_i
- √ H_0: $\tau_1 = \tau_2 = ... = \tau_a = 0$
- √ Parâmetros do modelo: $\mu, \tau_1, \tau_2, \ldots, \tau_a, \sigma^2$
  (a + 2) parâmetros
- √ No R, modelo de efeitos tem a restrição $\tau_1 = 0$
  - Estima (a + 1) parâmetros: $\mu, \tau_2, \ldots, \tau_a, \sigma^2$

- Anova usando lm ou aov:
  - √ Fórmula: `resposta ~ grupo`
  - √ Ajustam o mesmo modelo, mas são apresentadas de maneira diferente
- Coeficientes:
  - √ Estimados por máxima verossimilhança
  - √ Intercepto: média amostral para 1º nível do fator

- Comando `lm`:

```
> # comando lm
> cintila.lm <- lm(Flicker ~ Colour)
> cintila.lm
Call:
lm(formula = Flicker ~ Colour)
Coefficients:
(Intercept)  ColourBrown  ColourGreen
     28.167       -2.579       -1.247
```

$\beta_1 = (25,5875 - 28,1667) = -2,5792$

$\beta_2 = (26,9200 - 28,1667) = -1,2467$

- √ Estimativas de MQ de: μ, τ_2, τ_3.
- √ Intercepto: 28,167
  - – Média amostral para olhos azuis
- √ β_1: diferença entre médias ‘Castanho’ e ‘Azul’
- √ β_2: diferença entre médias ‘Verde’ e ‘Azul’

- Valores ajustados ($\hat{y}_{ij}$):

```
> # valores ajustados
> cintila.lm$fitted.values
Call:
       1        2        3        4        5        6        7        8
25.58750 25.58750 25.58750 25.58750 25.58750 25.58750 25.58750 25.58750
       9       10       11       12       13       14       15       16
26.92000 26.92000 26.92000 26.92000 26.92000 28.16667 28.16667 28.16667
      17       18       19
28.16667 28.16667 28.16667
```

- √ Comandos alternativos

```
> cintila.lm$fit
Call:
       1        2        3        4        5        6        7        8
25.58750 25.58750 25.58750 25.58750 25.58750 25.58750 25.58750 25.58750
> predict(cintila.lm)
Call:
       1        2        3        4        5        6        7        8
25.58750 25.58750 25.58750 25.58750 25.58750 25.58750 25.58750 25.58750
```


- Valores ajustados ($\hat{y}_{ij}$):

```
> # Valores ajustados
> table(predict(cintila.lm))

         25.5875            26.92 28.1666666666667
               8                5                6
> unique(predict(cintila.lm))
[1] 25.58750 26.92000 28.16667
```

$\hat{y}_{ij} = \bar{y}_j$

- √ São as médias dos grupos

- Comando `aov`:
  - √ Modelo das médias dos grupos

```
> # comando aov
> cintila.aov <- aov(Flicker ~ Colour)
> # apresenta modelo das médias
> model.tables(cintila.aov, type = "means")
Tables of means
Grand mean

26.75263

 Colour
    Blue Brown Green
   28.17 25.59 26.92
rep  6.00  8.00  5.00
```

√ Tabela das médias dos grupos e da quantidade de réplicas para cada grupo

- Comando `aov`:
  - √ Modelo dos efeitos em cada grupo

```
> # apresenta modelo de efeitos
> model.tables(cintila.aov, type = "effects")
Tables of effects

 Colour
    Blue  Brown  Green
   1.414 -1.165 0.1674
rep 6.000  8.000 5.0000
```

$\hat{\tau}_j = \bar{y}_j - \bar{\bar{y}}_j$

√ Estimativa dos efeitos:
(médias dos grupos – média global)

Notação

x_{ij} : i-ésima observação no j-ésimo grupo
$\bar{x}_{.j}$: média amostral do grupo j
s_j: desvio padrão do grupo j
n_j: quantidade de observações do grupo j
$n = \sum n_i$: total global de observações
$\bar{x}_{..}$: média global

Tabela ANOVA

Fonte de Variação	Soma de Quadrados	gl	Média Quadrática[a]	F_0
Entre	$\sum n_i(\bar{x}_{.j} - \bar{x}_{..})^2.$	a – 1	s_B^2	$f_0 = \frac{s_B^2}{s_W^2}$
Dentro	$\sum (n_j - 1)s_j^2.$	n – a	s_W^2	
Total	$\sum\sum (x_{ij} - \bar{x}_{..})^2.$	n – 1		

[a]Média quadrática = (soma de quadrados)/gl.

- P-valor:
  - √ $p = \Pr\{F \geq f_0\}$

- Obtenção da tabela Anova
  - √ Comando `anova` aplicado a objeto `lm`
  - √ Comando `summary` aplicado a objeto `aov`

- Comando `anova`:

```
> # comando anova
> anova(cintila.lm)
Analysis of Variance Table

Response: Flicker
          Df Sum Sq Mean Sq F value  Pr(>F)
Colour     2 22.997 11.4986  4.8023 0.02325 *
Residuals 16 38.310  2.3944
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

  - √ Saída:
    - Estimativa de σ^2: $\mathrm{EQM} = 2,3944$
    - Estatística F: $F = 4,8023 \sim \mathrm{F}_{2,16}$
    - p-valor: $p = 0,0235$

```
> 1-pf(4.8023, 2, 16)
[1] 0.02324962
```

  - √ Há evidência amostral de haver diferenças significativas entre as médias de cintilação dos grupos

- Comando `anova`:
  - √ Supressão de estrelas

```
> # opção para suprimir estrelas
> options(show.signif.stars=FALSE)
> anova(cintila.lm)
Analysis of Variance Table

Response: Flicker
          Df Sum Sq Mean Sq F value  Pr(>F)
Colour     2 22.997 11.4986  4.8023 0.02325
Residuals 16 38.310  2.3944
```

- Comando `summary`:

```
> summary(cintila.aov)
 Df Sum Sq Mean Sq F value Pr(>F)
Colour       2  23.00  11.499   4.802 0.0232
Residuals   16  38.31   2.394
```

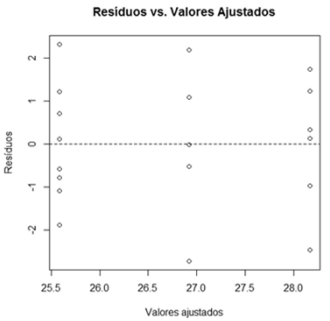

- Para o teste F ser válido
  - √ Amostras devem ser independentes
  - √ Erros independentes e normalmente distribuídos com média 0 e variância constante σ^2.
- Tamanhos amostrais são pequenos:
  - √ Difícil verificar hipóteses de normalidade e de homocedasticidade
  - √ É possível verificar desvios severos às hipóteses do modelo usando gráficos de resíduos

Resíduos

- Erros e resíduos:
  - √ $\epsilon_{ij} = Y_{ij} - \mu_j$: erro (ruído) do modelo
    - distância entre as observações e suas médias
  - √ $e_{ij} = y_{ij} - \bar{y}_{j}$: resíduo do ajuste do modelo
    - Estima a distância verdadeira
- Análise dos resíduos:
  - √ $\hat{u}_{ij}$ vs. grupo i
  - √ Gráfico de normalidade de todos os resíduos

- *Plot* dos resíduos vs. valores ajustados:

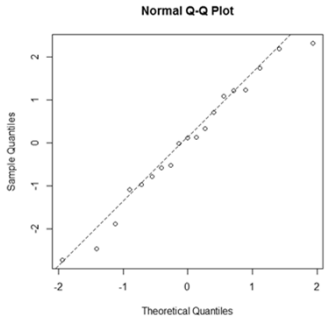
```
> # plot resíduos vs valores ajustados
> plot(cintila.lm$fit, cintila.lm$res, xlab = "Valores ajustados",
+ ylab = "Resíduos", main = "Resíduos vs. Valores Ajustados")
> abline(h = 0, lty = 2)
```



√ Resíduos devem ser aproximadamente simétricos em torno de zero e ter variância aproximadamente igual

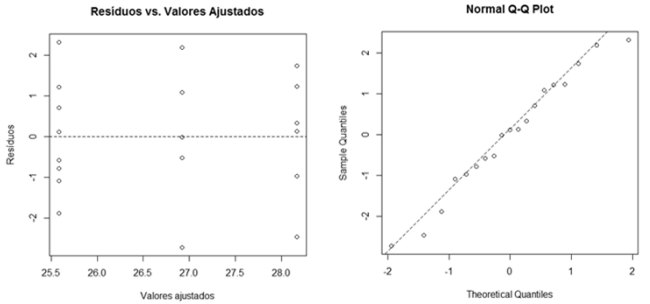
- Gráfico de normalidade dos resíduos:

```
> # Gráfico de normalidade dos resíduos
> qqnorm(cintila.lm$res)
> qqline(cintila.lm$res, lty = 2)
```



√ Resíduos devem estar dispostos aproximadamente ao longo da reta

- Gráficos dos resíduos:



√ Os plots não revelam qualquer desvio severo das hipóteses do modelo

Comparação das Médias dos Tratamentos

- Se H_0 é rejeitada:
  - √ Há alguma diferença entre as médias dos tratamentos
  - √ Essas diferenças não estão determinadas
  - √ Pares de médias diferentes no modelo ANOVA a 1 fator: $\binom{a}{2} = \frac{a(a-1)}{2}$
- São necessárias comparações múltiplas:
  - √ Teste de Tukey
  - √ Método das diferenças significantes mínimas (Fisher Least Significant Difference – LSD)

Fisher Least Significant Difference - LSD

- H_0: $\mu_i = \mu_j$ vs. H_1: $\mu_i \neq \mu_j$, $1 \leq i < j \leq a$
- H_0 é testada com o teste t para 2 amostras

$$T = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{\text{EQM}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \quad T \sim t_{n-a}, \text{ com } n = \sum_{i=1}^{a} n_i$$

√ H_1 é bilateral e $\mu_i = \mu_j$ são significativamente diferentes se: $|\bar{y}_i - \bar{y}_j| > t_{1-\alpha/2,n-a}\sqrt{\text{EQM}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$

$$\text{LSD} = t_{1-\alpha/2,n-a}\sqrt{\text{EQM}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

- Comparação múltipla – LSD de Fisher:

```
> # Comparações múltiplas - LSD de Fisher
> # EQM da ANOVA
> eqm <- anova(cintila.lm)$"Mean Sq"[2]
> eqm
[1] 2.39438
> # graus de liberdade
> gl <- anova(cintila.lm)$Df[2]
> gl
[1] 16
> # Percentil 97.5% de t
> t97.5 <- qt(.975, df = gl)
> # tamanho dos grupos
> n <- table(Colour)
> # médias dos tratamamentos
> medias <- by(Flicker, Colour, mean)
> medias
Colour: Blue
[1] 28.16667
------------------------------------------------------------
Colour: Brown
[1] 25.5875
------------------------------------------------------------
Colour: Green
[1] 26.92
```


√ LSD de Fisher – continuação:

```
> # aplica operador elemento a elemento
> dif.mat <- outer(medias, medias, "-")
> dif.mat
       Colour
Colour       Blue    Brown     Green
  Blue   0.000000 2.579167  1.246667
  Brown -2.579167 0.000000 -1.332500
  Green -1.246667 1.332500  0.000000
> diferenca <- abs(dif.mat[lower.tri(dif.mat, diag = FALSE)])
> # valor de LSD
> lsd.mat <- t97.5 * sqrt(eqm * outer(1/n, 1/n, "+"))
> LSD <- lsd.mat[lower.tri(lsd.mat, diag = FALSE)]
> comparacao <- cbind(diferenca, LSD)
> rownames(comparacao) <- c("Blue-Brown", "Blue-Green", "Brown-Green")
> comparacao
            diferenca      LSD
Blue-Brown   2.579167 1.771562
Blue-Green   1.246667 1.986318
Brown-Green  1.332500 1.870056
```

√ Diferença significativa a um nível de 5%:

- Azul – Castanho

- Conclusões:
  - √ Evidências para concluir que a frequência crítica média de cintilação para olhos azuis é significativamente maior que a média para os olhos castanhos
  - √ Nenhum outro par é significativamente diferente a um nível de 5%

- Comentários:
  - √ Teste t múltiplo ou os intervalos t do método LSD introduzem um problema ao inflar o risco de erro tipo I.
  - √ Teste LSD de Fisher refere-se à situação em que as comparações são efetuadas apenas após um teste t significante.

√ LSD de Fisher – comando `pairwise.t`:

```
> pairwise.t.test(Flicker, Colour)

        Pairwise comparisons using t tests with pooled SD

data:  Flicker and Colour

      Blue  Brown
Brown 0.021 -
Green 0.301 0.301

P value adjustment method: holm
```

  - √ p-valores do teste:
    - Apenas o par (blue, brown) tem diferença de médias significativa à $\alpha = 0,05$.

Referências

Bibliografia Recomendada

- ALBERT, J.; RIZZO, M. *R by Example*. Springer, 2012.
- CHAPMAN, C.; FEIT, E. M. *R for marketing research and analytics*. Springer, 2015.
- KLEIBER, C.; ZEILEIS, A. *Applied econometrics with R*. Springer, 2008.
- DALGAARD, P. *Introductory statistics with R*. Springer, 2008.