

Exercicio Aula 13 Delineamento

Davi Wentrick Feijó

2023-05-08

Exercicio 1:

Uma quimica deseja testar o efeito de quatro agentes na resistencia de um determinado tipo de tecido. Porque pode haver variabilidade de um rolo de tecido para outro, a quimica decide usar um delineamento de blocos casualizados. Sendo que, os rolos de tecido serao considerados como blocos. Ela seleciona cinco rolos e aplica todos os quatro marcas quimicos em ordem aleatoria para cada rolo. Na Tabela 1 estao os resultados das resistencias resultante.

```
#montando o dataframe
marcas = c("A","B","C","D","E")
holandesa = c(122,125,120,150,153)
jersey = c(144,137,134,155,165)
girolando = c(145,144,136,156,171)

dados = data.frame(marcas,holandesa,jersey,girolando)
kable(dados)
```

| marcas | holandesa | jersey | girolando |
|--------|-----------|--------|-----------|
| A | 122 | 144 | 145 |
| B | 125 | 137 | 144 |
| C | 120 | 134 | 136 |
| D | 150 | 155 | 156 |
| E | 153 | 165 | 171 |

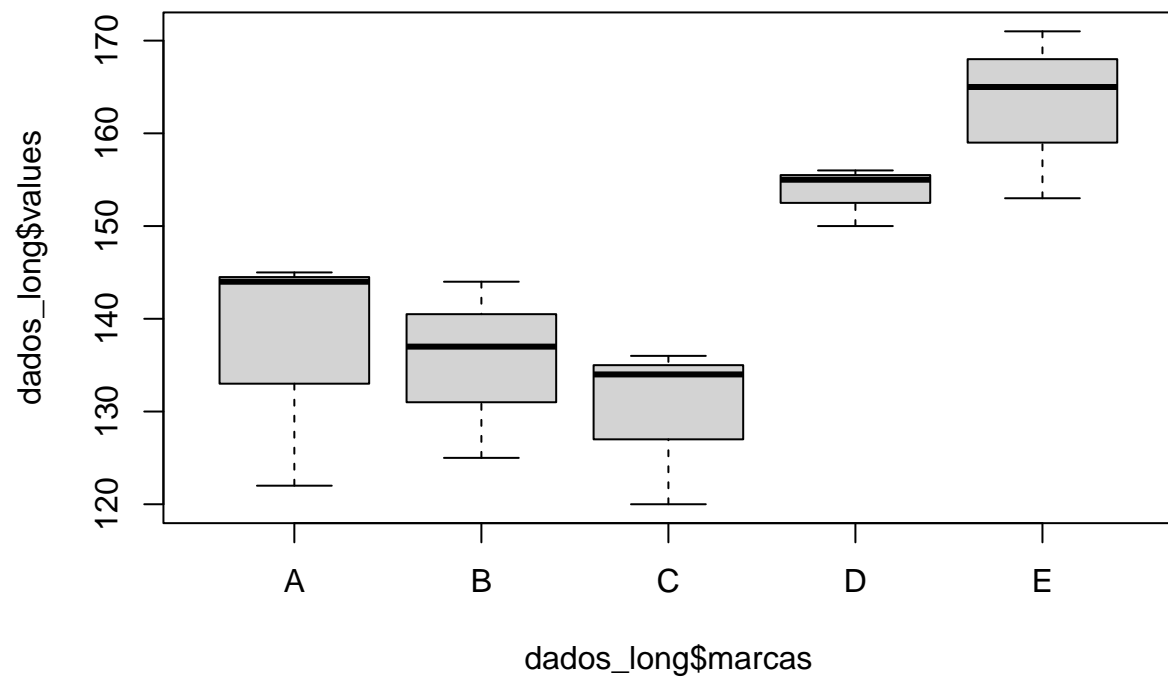
Para trabalhar com os dados precisamos formatar eles de forma que fiquem na forma long por meio da funcao pivot_longer()

```
#padronizando o data frame para analise
```

```
#padronizando o data frame para analise
```

```
dados_long = dados %>%  
  pivot_longer(cols = c(holandesa, jersey, girolando), values_to = "values", names_to = "blocos") %>%  
  mutate(marcas = as.factor(marcas))  
  
kable(dados_long)
```

| marcas | blocos | values |
|--------|-----------|--------|
| A | holandesa | 122 |
| A | jersey | 144 |
| A | girolando | 145 |
| B | holandesa | 125 |
| B | jersey | 137 |
| B | girolando | 144 |
| C | holandesa | 120 |
| C | jersey | 134 |
| C | girolando | 136 |
| D | holandesa | 150 |
| D | jersey | 155 |
| D | girolando | 156 |
| E | holandesa | 153 |
| E | jersey | 165 |
| E | girolando | 171 |



1.1) O modelo considerado e as hipoteses de interesse

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_a = 0$$

$$H_1 : \exists \tau_i \neq 0$$

Queremos verificar se todos os τ_i tem o mesmo efeito de tratamento (ou se nao tem diferenca entre os tratamentos que é nosso H_0) ou se existe pelo menos um τ_i que tem o efeito de tratamento diferente de 0

1.2) A tabela de análise de variancia e suas conclusões Para realizar a tabela da ANOVA é necessário calcular a soma de quadrados, graus de liberdade, quadrados medios para o tratamento, bloco e residuo. Alem do valor F e p-valor do teste anova e do teste para blocos.

A soma de quadrados do modelo que iremos utilizar (com blocos) pode ser escrito como:

$$SQ_{Total} = SQ_{Tratamento} + SQ_{Bloco} + SQ_{Residuo}$$

$$\sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{..})^2 = b \sum_{i=1}^a (Y_{i.} - \bar{Y}_{..})^2 + a \sum_{j=1}^b (Y_{.j} - \bar{Y}_{..})^2 + \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - Y_{.j} - Y_{i.} + \bar{Y}_{..})^2$$

$$SQ_{Total} = \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{..})^2$$

$$SQ_{Tratamento} = b \sum_{i=1}^a (Y_{i.} - \bar{Y}_{..})^2$$

$$SQ_{Bloco} = a \sum_{j=1}^b (Y_{.j} - \bar{Y}_{..})^2$$

$$SQ_{Residuo} = \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - Y_{.j} - Y_{i.} + \bar{Y}_{..})^2$$

Aqui vamos definir nosso a e b que sao a quantiade de tratamentos e de blocos respectivamente

```
# quantidade de blocos e tratamentos
t = length(unique(dados_long$marcas)) #tratamentos
b = length(unique(dados_long$blocos)) #blocos
```

Em seguida podemos calcular a media geral e as somas de quadrados

```
#soma de quadrados
media_total <- mean(dados_long$values)

ssqtot = sum((dados_long$values - media_total)^2)
ssqtrat = sum((tapply(dados_long$values,dados_long$marcas,mean) - media_total)^2) * b
ssqblocos = sum((tapply(dados_long$values,dados_long$blocos,mean) - media_total)^2) * t
ssqres = ssqtot-ssqtrat-ssqblocos
```

Para calcular os quadrados médios do total, tratamento e resíduos podemos seguir as seguinte formulas:

$$QMTratamento = \frac{SQTratamento}{a-1} = \frac{b \sum_{i=1}^a (Y_{i.} - \bar{Y}_{..})^2}{a-1} = \frac{b \sum_{i=1}^a \tau^2}{a-1}$$

$$QMBloco = \frac{SQBloco}{b-1} = \frac{a \sum_{j=1}^b (Y_{.j} - \bar{Y}_{..})^2}{b-1} = \frac{b \sum_{i=1}^a \beta^2}{a-1}$$

$$QMResiduo = \frac{SQResiduo}{(a-1)(b-1)} = \frac{\sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - Y_{.j} - Y_{i.} - \bar{Y}_{..})^2}{(a-1)(b-1)}$$

Calculando os graus de liberdade dos blocos, tratamentos, e o total!

```
#graus de liberdade  
glb = (b-1)  
glt = (t-1)  
glr = glb*glt  
gltot = (b*t)-1
```

Agora podemos encontrar os quadrados médios

```
#quadrados medios  
qmtrat = ssqtrat/glt  
qmbloco = ssqblocos/glb  
qmres = ssqres/glr
```

Sendo assim, para testar a igualdade das medias de tratamento, a estatística de teste é definida por:

$$F_0 = \frac{QMTratamento}{QMResiduo} \sim F_{a-1,(a-1)(b-1)}$$

É possível seguir a mesma ideia da estatística de teste para efeito de bloco. Então, tem-se que:

$$F_0 = \frac{QMBloco}{QMResiduo} \sim F_{b-1,(a-1)(b-1)}$$

```
# valor F observado
f_obs = qmtrat/qmres
f_obs_blocos = qmbloco/qmres
```

```
#valor critico de 5%
alfa = 0.05
f_crit = qf(alfa,glr,glr)
f_crit_bloco = qf(alfa,glb,glr)
```

```
#p-valor observado
f_value = round(pf(f_obs,glr,glr,lower.tail = FALSE),3)
f_value_blocos = round(pf(f_obs_blocos,glb,glr,lower.tail = FALSE),7)
```

```

# Tabela da ANOVA
anova_table <- data.frame(Fonte_de_variacao = c("Produto", "Rolo de tecido", "Residuos", "Total"),
  GL = c(glt, glb, glr, gltot),
  SS = c(ssqtrat, ssqblocos, ssqres, ssqtot),
  MQ = c(round(qmtrat,2), round(qmbloco,2), round(qmres,2), ''),
  F = c(round(f_obs,3),round(f_obs_blocos,3),'', ''),
  Pf = c(f_value, f_value_blocos, "", ''),
  stringsAsFactors = FALSE)
rownames(anova_table) <- NULL

anova_table

```

```

##  Fonte_de_variacao GL      SS      MQ      F      Pf
## 1      Produto  4 2323.0667 580.77 40.708      0
## 2    Rolo de tecido  2  749.2000  374.6 26.257 0.0003054
## 3      Residuos  8  114.1333  14.27
## 4          Total 14 3186.4000

```


1.3) Os pressupostos necesarios foram atendidos ?

Para verificar os pressupostos temos que calcular os residuos e testar normalidade deles. Vamos aproveitar e calcular nossos estimadores para μ , τ_i e β_j . E isso poder ser feito por meio das seguintes formulas deduzidas por meio dos quadrados médios:

$$\hat{\mu} = \bar{Y}_{..} \quad \hat{\tau} = \bar{Y}_{i.} - \bar{Y}_{..} \quad \hat{\beta} = \bar{Y}_{.j} - \bar{Y}_{..}$$

```
#nossos estimadores
media_total = mean(dados_long$values)

trat_media = tapply(dados_long$values,dados_long$marcas,mean) - media_total

blocos_media = tapply(dados_long$values,dados_long$blocos,mean) - media_total
```

Aqui temos a tabela com os valores observados, valores esperados e os residuos por meio do modelo! e sabendo que os residuos sao definidos por:

$$e_{ij} = Y_{ij} - \hat{Y}_{ij}$$

```
#calculando os residuos
dados_long = dados_long %>%
  mutate(media_trat = ave(values, marcas, FUN = mean) - mean(values),
         media_bloco = ave(values, blocos, FUN = mean) - mean(values),
         y_obs = mean(values) + media_trat + media_bloco,
         residuo = values - y_obs,
         residuo_normalizado = residuo/qmres)

kable(dados_long)
```

| marcas | blocos | values | media_trat | media_bloco | y_obs | residuo | residuo_normalizado |
|--------|-----------|--------|------------|-------------|----------|------------|---------------------|
| A | holandesa | 122 | -6.800000 | -9.8 | 127.2000 | -5.2000000 | -0.3644860 |
| A | jersey | 144 | -6.800000 | 3.2 | 140.2000 | 3.8000000 | 0.2663551 |
| A | girolando | 145 | -6.800000 | 6.6 | 143.6000 | 1.4000000 | 0.0981308 |
| B | holandesa | 125 | -8.466667 | -9.8 | 125.5333 | -0.5333333 | -0.0373832 |
| B | jersey | 137 | -8.466667 | 3.2 | 138.5333 | -1.5333333 | -0.1074766 |
| B | girolando | 144 | -8.466667 | 6.6 | 141.9333 | 2.0666667 | 0.1448598 |
| C | holandesa | 120 | -13.800000 | -9.8 | 120.2000 | -0.2000000 | -0.0140187 |
| C | jersey | 134 | -13.800000 | 3.2 | 133.2000 | 0.8000000 | 0.0560748 |
| C | girolando | 136 | -13.800000 | 6.6 | 136.6000 | -0.6000000 | -0.0420561 |
| D | holandesa | 150 | 9.866667 | -9.8 | 143.8667 | 6.1333333 | 0.4299065 |
| D | jersey | 155 | 9.866667 | 3.2 | 156.8667 | -1.8666667 | -0.1308411 |
| D | girolando | 156 | 9.866667 | 6.6 | 160.2667 | -4.2666667 | -0.2990654 |
| E | holandesa | 153 | 19.200000 | -9.8 | 153.2000 | -0.2000000 | -0.0140187 |
| E | jersey | 165 | 19.200000 | 3.2 | 166.2000 | -1.2000000 | -0.0841121 |
| E | girolando | 171 | 19.200000 | 6.6 | 169.6000 | 1.4000000 | 0.0981308 |

Agora podemos realizar nossos testes de normalidade e de igualdade de variancias

```
shapiro.test(dados_long$residuo)
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: dados_long$residuo  
## W = 0.97083, p-value = 0.8701
```

```
leveneTest(residuo ~ marcas, dados_long)
```

```
## Levene's Test for Homogeneity of Variance (center = median)  
##      Df F value Pr(>F)  
## group  4  0.8845 0.5073  
##      10
```

```
#pressuposto de modelo aditivo
```

```
mod = lm(dados_long$values ~ dados_long$marcas + dados_long$blocos)
```

```
ad = (predict(mod))^2
```

```
admod = lm(dados_long$values ~ dados_long$marcas + dados_long$blocos + ad)
```

```
anova(mod, admod)
```

```
## Analysis of Variance Table
```

```
##
```

```
## Model 1: dados_long$values ~ dados_long$marcas + dados_long$blocos
```

```
## Model 2: dados_long$values ~ dados_long$marcas + dados_long$blocos + ad
```

```
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
```

```
## 1      8 114.133
```

```
## 2      7  97.084  1    17.049 1.2293 0.3042
```

1.4) Qual a proporção da variacao total explicada pelo modelo ajustado no item 1.2?

A proporcao de variancia explicada vai ser em relacao a soma de quadrados total do modelo. Ou seja Parte da variancia é explicada pelos tratamentos e outra parte é pelos blocos. O que sobra, os residuos, é a variacao aleatoria que temos, essa que nao tem explicacao dentro do nosso modelo. Logo para saber quanto que nosso modelo explica podemos fazer de duas formas:

Somando a soma de quadrados de tratamento e bloco calculando a porcentagem em relacao ao total

```
R2 = ((ssqtrat+ssqblocos)/ssqtot)

cat("A variancia explicada pelo medole é", R2)
```

```
## A variancia explicada pelo medole é 0.9641811
```

ou podemos ir pelo caminho contrario onde calculamos a variancia explicada pelo residuo e diminuimos do total

```
R2 = 1 - (ssqres/ssqtot)

cat("A variancia explicada pelo medole é", R2)
```

```
## A variancia explicada pelo medole é 0.9641811
```

1.5) Considerando que o objetivo do experimento é maximizar a variável resposta, qual é o elemento químico que deve ser recomendado? Use teste de Tukey para subsidiar sua resposta.

Para realizar o teste de Tukey é necessário seguir alguns passos! Vamos começar com o cálculo da diferença de médias!

$$q = \frac{\bar{Y}_{max} - \bar{Y}_{min}}{\sqrt{QMRes/n}}$$

Usaremos essa fórmula para calcular a estatística do teste e encontrar seu P-Valor. Contudo existe outra forma de identificar uma diferença significativa de médias! Utilizando a seguinte fórmula:

$$T_{\alpha} = q_{\alpha}(a, f) \sqrt{\frac{QMRes}{n}}$$

Onde a é o número de tratamentos e f é os graus de liberdade associado ao $QMRes$. Agora podemos calcular o valor HSD (Honest Significant Differences) com o qual pode ser comparado com a diferença absoluta entre as médias. Caso a Diferença seja maior ou igual ao valor pode-se afirmar que existe diferença significativa entre os respectivos tratamentos.

Seguindo a mesma lógica podemos definir um intervalo de confiança para nossas estimativas:

$$IC = \bar{Y}_i - \bar{Y}_j \pm q_{\alpha}(a, f) \sqrt{\frac{QMRes}{n}}$$

```

alfa = 0.05
q.value <- qtkey(alfa, t, glr, lower.tail = F) #valor critico do quantil da distribuicao de tukey

hsd = q.value * sqrt(qmres/b) #valor a ser comparado com as diferencas absolutas

combinacoes = combn(unique(dados$marcas),2) #todas as combinacoes de medias
comb_diff = combn(trat_media,2) #vetor com as combinacoes de medias 2 a 2

diferencas <- data.frame(
  comparacao = apply(combinacoes, 2, paste0, collapse = "-"), #vetor que indica quais sao as comparacoes
  diferenca = apply(comb_diff, 2, diff)) %>% #calcula a diferenca de medias
  mutate(lwr = diferenca - q.value * sqrt(qmres/b), #intervalo inferior
         upr = diferenca + q.value * sqrt(qmres/b), #intervalo superior
         pvalor = round(ptukey(abs(diferenca/(sqrt(qmres/b))), t, glr, lower.tail = F),6), #calcula o p
         hsd = abs(diferenca) >= hsd) #comparando a diferenca absoluta com o HSD

```

| comparacao | diferenca | lwr | upr | pvalor | hsd |
|------------|-----------|------------|-----------|----------|-------|
| A-B | -1.666667 | -12.321146 | 8.987813 | 0.980172 | FALSE |
| A-C | -7.000000 | -17.654480 | 3.654480 | 0.247047 | FALSE |
| A-D | 16.666667 | 6.012187 | 27.321146 | 0.004211 | TRUE |
| A-E | 26.000000 | 15.345520 | 36.654480 | 0.000207 | TRUE |
| B-C | -5.333333 | -15.987813 | 5.321146 | 0.468760 | FALSE |
| B-D | 18.333333 | 7.678854 | 28.987813 | 0.002291 | TRUE |
| B-E | 27.666667 | 17.012187 | 38.321146 | 0.000132 | TRUE |
| C-D | 23.666667 | 13.012187 | 34.321146 | 0.000404 | TRUE |
| C-E | 33.000000 | 22.345520 | 43.654480 | 0.000036 | TRUE |
| D-E | 9.333333 | -1.321146 | 19.987813 | 0.089789 | FALSE |

1.6) Refaca as contas necessarias para responder os itens (1.2), (1.3) e (1.6) utilizando as funcoes aov e TukeyHSD e confira com os resultados obtidos.

```
# anova no r para compara os valores
```

```
aov_res = aov(dados_long$values ~ dados_long$marcas+dados_long$blocos)
```

```
aov_res %>% summary()
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## dados_long$marcas  4 2323.1    580.8    40.71 2.31e-05 ***
## dados_long$blocos  2  749.2    374.6    26.26 0.000305 ***
## Residuals         8  114.1     14.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#teste de tukey no R
```

```
TukeyHSD(aov_res)
```

```
##      Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = dados_long$values ~ dados_long$marcas + dados_long$blocos)
##
## $'dados_long$marcas'
##           diff          lwr          upr          p adj
## B-A -1.666667 -12.321146   8.987813  0.9801720
## C-A -7.000000 -17.654480  3.654480  0.2470465
## D-A 16.666667   6.012187 27.321146  0.0042113
## E-A 26.000000  15.345520 36.654480  0.0002071
## C-B -5.333333 -15.987813   5.321146  0.4687601
## D-B 18.333333   7.678854 28.987813  0.0022908
## E-B 27.666667  17.012187 38.321146  0.0001321
## D-C 23.666667  13.012187 34.321146  0.0004044
## E-C 33.000000  22.345520 43.654480  0.0000359
## E-D  9.333333  -1.321146 19.987813  0.0897894
##
## $'dados_long$blocos'
##           diff          lwr          upr          p adj
## holandesa-girolando -16.4 -23.226043 -9.573957 0.0003335
## jersey-girolando    -3.4 -10.226043  3.426043 0.3747816
## jersey-holandesa     13.0   6.173957 19.826043 0.0015682
```

Como nao temos normalidade dos residuos, o certo é fazer o kruskall-wallis no lugar da ANOVA

```
#teste de kruskall-wallis
```

```
kruskal.test(dados_long$values ~ dados_long$marcas)
```

```
##
```

```
## Kruskal-Wallis rank sum test
##
## data:  dados_long$values by dados_long$marcas
## Kruskal-Wallis chi-squared = 11.011, df = 4, p-value = 0.02644
```

1.7) Determine a probabilidade do erro tipo 2 para o caso de: ($\tau_1 = -1.5$, $\tau_2 = 0$, $\tau_3 = 0$, $\tau_4 = 1.5$).

Pode-se mostrar que, se H_0 for falsa, a estatística $F_0 = QM_{Trat}/QM_{Res}$ tem distribuição F não central com $a - 1$ e $(a - 1)(b - 1)$ graus de liberdade e parâmetro de não centralidade δ . Se $\delta = 0$, a distribuição F não central torna-se a distribuição F usual (central). O parâmetro de não centralidade da distribuição F pode ser obtido ao calcular:

$$\delta = \frac{b \sum_{i=1}^a \tau_i^2}{\sigma^2}$$

Sabemos que um bom estimador para σ^2 é o quadrado médio do resíduo, ou seja:

$$\sigma^2 = QM_{Residuo} = \frac{SQ_{Residuo}}{(a - 1)(b - 1)} = \frac{\sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - Y_{.j} - Y_{i.} - \bar{Y}_{..})^2}{(a - 1)(b - 1)}$$

```
alpha = 0.05

taui = c(-1.5,0,0,1.5) #taui dados pela questao

sigma2 = qmres #estimador da variancia

delta = b*sum(taui^2/sigma2) #nosso delta de nao centralidade da F

fcrit = qf(1-alpha,t-1,(t-1)*(b-1)) #calculando o valor F critico em relacao ao erro de tipo II

beta = pf(q = fcrit,df1 = t-1,df2 = (t-1)*(b-1), ncp = delta)

poder = 1 - beta
poder

## [1] 0.080143
```

1.8) Para os valores de taus considerados no item anterior, determine qual deve ser o numero de blocos para que o erro tipo 2 seja inferior a 10%?

```
#vamos aumentando o numero de blocos (b) ate mudar a porcentagem
b = 7 #caso os experientno tenha 10 blocos

alpha = 0.05

taui = c(-1.5,0,0,1.5) #taui dados pela questao

sigma2 = qmres #estimador da variancia

delta = b*sum(taui^2/sigma2) #nosso delta de nao centralidade da F

fcrit = qf(1-alpha,t-1,(t-1)*(b-1)) #calculando o valor F critico em relacao ao erro de tipo II

beta = pf(q = fcrit,df1 = t-1,df2 = (t-1)*(b-1), ncp = delta)

poder = 1 - beta
poder

## [1] 0.1596157
```