Exercicio Aula 11 e 12 Delineamento

Davi Wentrick Feijó

2023-05-06

Exercicio 1:

Uma qumica 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 qumica 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 produtos qumicos em ordem aleatoria para cada rolo. Na Tabela 1 estao os resultados das resistencias resultante.

```r
produtos = c(1,2,3,4)
rolo_1 = c(73,73,75,73)
rolo_2 = c(68,67,68,71)
rolo_3 = c(74,75,78,75)
rolo_4 = c(71,72,73,75)
rolo_5 = c(67,70,68,69)

dados = data.frame(produtos,rolo_1,rolo_2,rolo_3,rolo_4,rolo_5)
kable(dados)
```

produtos	rolo_1	rolo_2	rolo_3	rolo_4	rolo_5
1	73	68	74	71	67
2	73	67	75	72	70
3	75	68	78	73	68
4	73	71	75	75	69

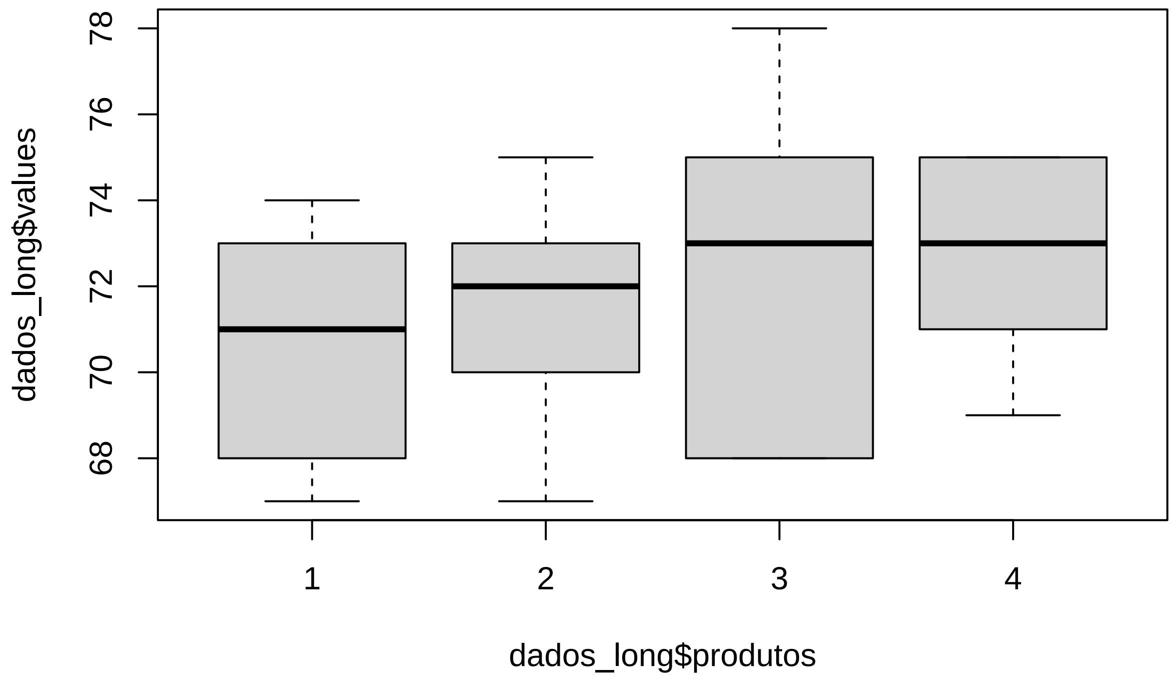
Para trabalhar com os dados precisamos formatar eles de forma que fiquem na forma long por meio da funcao pivot_longer()

```
#padronizando o data frame para analise

dados_long = dados %>%
  pivot_longer(cols = c(rolo_1,rolo_2,rolo_3,rolo_4,rolo_5), values_to = "values", names_to = "blocos")
  mutate(produtos = as.factor(produtos))

kable(dados_long)
```

produtos	blocos	values
1	rolo_1	73
1	rolo_2	68
1	rolo_3	74
1	rolo_4	71
1	rolo_5	67
2	rolo_1	73
2	rolo_2	67
2	rolo_3	75
2	rolo_4	72
2	rolo_5	70
3	rolo_1	75
3	rolo_2	68
3	rolo_3	78
3	rolo_4	73
3	rolo_5	68
4	rolo_1	73
4	rolo_2	71
4	rolo_3	75
4	rolo_4	75
4	rolo_5	69


78
76
74
72
70
68
dados_long$values
1
2
3
4
dados_long$produtos

1.1) O modelo considerado e as hipoteses de interesse

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

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

Queremos verificar se todos os τ_i tem o mesmo efeito de tratamento (ou se nao tem diferenca entre os tratamentos que é nosso H0) 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 conclusoes Para realizar a tabela da ANOVA é necessario 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:

$$\text{SQTotal} = \text{SQTratamento} + \text{SQBloco} + \text{SQResiduo}$$

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

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

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

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

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

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$produtos)) #tratamentos
b = length(unique(dados_long$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$produtos,mean) - media_total)^2) * b
ssqblocos = sum((tapply(dados_long$values,dados_long$blocos,mean) - media_total)^2) * t
ssqres = ssqtot-ssqtrat-ssqblocos
```

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

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

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

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

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 estatıstica de teste e definida por:

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

E possıvel seguir a mesma ideia da estatıstica de teste para efeito de bloco. Entao, tem-se que:

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

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

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

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

```r
# 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  3  12.95  4.32  2.376    0.121
## 2    Rolo de tecido  4 157.00 39.25 21.606 2.06e-05
## 3          Residuos 12  21.80  1.82
## 4             Total 19 191.75
```

1.3) Os pressupostos necessarios 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:

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

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

trat_media = tapply(dados_long$values,dados_long$produtos,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, produtos, 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)
```

produtos	blocos	values	media_trat	media_bloco	y_obs	residuo	residuo_normalizado
1	rolo_1	73	-1.15	1.75	72.35	0.65	0.3577982
1	rolo_2	68	-1.15	-3.25	67.35	0.65	0.3577982
1	rolo_3	74	-1.15	3.75	74.35	-0.35	-0.1926606
1	rolo_4	71	-1.15	1.00	71.60	-0.60	-0.3302752
1	rolo_5	67	-1.15	-3.25	67.35	-0.35	-0.1926606
2	rolo_1	73	-0.35	1.75	73.15	-0.15	-0.0825688
2	rolo_2	67	-0.35	-3.25	68.15	-1.15	-0.6330275
2	rolo_3	75	-0.35	3.75	75.15	-0.15	-0.0825688
2	rolo_4	72	-0.35	1.00	72.40	-0.40	-0.2201835
2	rolo_5	70	-0.35	-3.25	68.15	1.85	1.0183486
3	rolo_1	75	0.65	1.75	74.15	0.85	0.4678899
3	rolo_2	68	0.65	-3.25	69.15	-1.15	-0.6330275
3	rolo_3	78	0.65	3.75	76.15	1.85	1.0183486
3	rolo_4	73	0.65	1.00	73.40	-0.40	-0.2201835
3	rolo_5	68	0.65	-3.25	69.15	-1.15	-0.6330275
4	rolo_1	73	0.85	1.75	74.35	-1.35	-0.7431193
4	rolo_2	71	0.85	-3.25	69.35	1.65	0.9082569
4	rolo_3	75	0.85	3.75	76.35	-1.35	-0.7431193
4	rolo_4	75	0.85	1.00	73.60	1.40	0.7706422
4	rolo_5	69	0.85	-3.25	69.35	-0.35	-0.1926606

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 = 0.8996, p-value = 0.04054
```

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

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  0.8895 0.4677
##       16
```

```
#pressuposto de modelo aditivo
mod = lm(dados_long$values ~ dados_long$produtos + dados_long$blocos)

ad = (predict(mod))^2

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

anova(mod,admod)
```

```
## Analysis of Variance Table
##
## Model 1: dados_long$values ~ dados_long$produtos + dados_long$blocos
## Model 2: dados_long$values ~ dados_long$produtos + dados_long$blocos +
##     ad
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     12 21.800
## 2     11 21.592  1   0.20815 0.106 0.7508
```

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

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

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

```
## A variancia explicada pelo medole é 0.8863103
```

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

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

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

```
## A variancia explicada pelo medole é 0.8863103
```

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

Para realizar o teste de tukey é necessario seguir alguns passos! Vamos comecar com o calculo da diferenca de médias!

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

Usaremos essa formula para calcular a estatistica do teste e encontrar seu P-Valor. Contudo existe outra forma de identificar uma diferenca significante de médias! Utilizando a seguinte formula:

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

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 coma diferenca absoluta entre as médias. Caso a Diferenca seja maior ou igual ao valor pode se afirmar que existe diferenca significativa entre os respectivos tratamentos.

Seguindo a mesma logica podemos definir um intervalo de confianca ppara nossas estimativas:

$$IC = \bar{Y}_{i.} - \bar{Y}_{j.} \pm q_\alpha(a, f)\sqrt{\frac{QMRes}{n}}$$

```
alfa = 0.05
q.value <- qtukey(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$produtos),2) #todas as combinacoes de médias
comb_diff = combn(trat_media,2) #vetor comas combinacoes de medias 2 a 2

diferencas <- data.frame(
  comparacao = apply(combinacoes, 2, paste0, collapse = "-"), #vetor que indica quais sao as comparacoe
  diferenca = apply(comb_diff, 2, diff)) %>% #calcula a diferenca de médias
  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), #calculo do p
         hsd = abs(diferenca) >= hsd) #comparando a diferenca absoluta com o HSD
```

comparacao	diferenca	lwr	upr	pvalor	hsd
1-2	0.8	-1.7308322	3.330832	0.785273	FALSE
1-3	1.8	-0.7308322	4.330832	0.204259	FALSE
1-4	2.0	-0.5308322	4.530832	0.141733	FALSE
2-3	1.0	-1.5308322	3.530832	0.654014	FALSE
2-4	1.2	-1.3308322	3.730832	0.518273	FALSE
3-4	0.2	-2.3308322	2.730832	0.995203	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.

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

aov_res = aov(dados_long$values ~ dados_long$produtos+dados_long$blocos)

aov_res %>% summary()
```

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)    
## dados_long$produtos  3  12.95    4.32   2.376    0.121    
## dados_long$blocos    4 157.00   39.25  21.606 2.06e-05 ***
## Residuals           12  21.80    1.82                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#teste de tukey no R

TukeyHSD(aov_res)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = dados_long$values ~ dados_long$produtos + dados_long$blocos)
##
## $`dados_long$produtos`
##     diff        lwr      upr     p adj
## 2-1  0.8 -1.7308322 3.330832 0.7852734
## 3-1  1.8 -0.7308322 4.330832 0.2042593
## 4-1  2.0 -0.5308322 4.530832 0.1417326
## 3-2  1.0 -1.5308322 3.530832 0.6540138
## 4-2  1.2 -1.3308322 3.730832 0.5182726
## 4-3  0.2 -2.3308322 2.730832 0.9952030
##
## $`dados_long$blocos`
##                diff        lwr        upr     p adj
## rolo_2-rolo_1 -5.00  -8.037831 -1.9621691 0.0015656
## rolo_3-rolo_1  2.00  -1.037831  5.0378309 0.2814173
## rolo_4-rolo_1 -0.75  -3.787831  2.2878309 0.9295872
## rolo_5-rolo_1 -5.00  -8.037831 -1.9621691 0.0015656
## rolo_3-rolo_2  7.00   3.962169 10.0378309 0.0000717
## rolo_4-rolo_2  4.25   1.212169  7.2878309 0.0056966
## rolo_5-rolo_2  0.00  -3.037831  3.0378309 1.0000000
## rolo_4-rolo_3 -2.75  -5.787831  0.2878309 0.0830636
## rolo_5-rolo_3 -7.00 -10.037831 -3.9621691 0.0000717
## rolo_5-rolo_4 -4.25  -7.287831 -1.2121691 0.0056966
```

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

```r
#teste de kruskall-wallis

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

```
##
##  Kruskal-Wallis rank sum test
##
## data:  dados_long$values by dados_long$produtos
## Kruskal-Wallis chi-squared = 1.5008, df = 3, p-value = 0.6821
```

1.7) Determine a probabilidade do erro tipo 2 para o caso de: (τ_1 = -1.5, τ_2 = 0, τ_3 = 0, τ_4 = 1.5).

Pode-se mostrar que, se H0 for falsa, a estatıstica F0 = QMTrat/QMRes tem distribuicao F nao central com a - 1 e (a - 1)(b - 1) graus de liberdade e parametro de nao centralidade δ,Se $\delta = 0$, a distribuicao F nao central torna-se a distribuicao F usual (central). O parametro de nao centralidade da distribuicao F pode ser obtido ao calcular:

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

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

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

```
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] 0.7077324
```

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 experiemtno 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] 0.8937297
```