Relatório 3 - Estatística

Fazendo algumas análises em R

Artur Papa - 3886 Erian Alves - 3862 Guilherme Sergio - 3854

Sumário

Contents

Dados

A seguir é feita a leitura do conjunto de dados a ser utilizado na atividade. Os dados são referentes ao contexto de aplicação de fertilantes em produções e a possível presença de pragas.

Obs.: Temos dois caminhos de dados diferentes devido ao fato de um integrante do grupo ter feito o trabalho de um diretório criado no notebook, enquanto que o outro se trata do caminho do repositório clonado do GitHub. Caso dê erro em um dos caminhos, basta comentá-lo e descomentar o que não está selecionado.

```r
df <- read.csv("~/Downloads/dados_atividade1.xlsx - Planilha1.csv")
#head(df)
```

```r
df <- read.csv("~/GitHub/MAF261-Experimental_Statistics/relatorio3/dados_atividade1.xlsx - Planilha1.csv
```

```
## Error in file(file, "rt"): não é possível abrir a conexão
```

```r
head(df)
```

```
##   ID Producao Praga Fertilizante
## 1  1      388     1            A
## 2  2      454     0            A
## 3  3      812     0            A
## 4  4      514     1            A
## 5  5      526     0            A
## 6  6      843     0            A
```

A partir da visualização do conjunto de dados a ser utilizado, é possível observar que o dataset é composto por 4 colunas: uma para o id (ID) da instância do dado, uma referente ao valor da produção (Producao), uma para indicar se houve a presença ou não de pragas (Praga) naquela produção e uma para informar qual o tipo de fertilizante empregado.

Tem-se por intuito, inicialmente, realizar análises com base nos fertilizantes para os valores de produção, logo que há diferentes tratamentos sendo aplicados.

```r
#install.packages("daewr")
library("daewr")
library("kableExtra")
kable(df, align='c')
```

ID	Producao	Praga	Fertilizante
1	388	1	A
2	454	0	A
3	812	0	A
4	514	1	A
5	526	0	A
6	843	0	A
7	1042	0	B
8	697	0	B
9	813	1	B
10	861	0	B
11	1195	1	B
12	1022	1	B
13	280	1	C
14	222	0	C
15	89	0	C
16	557	1	C
17	300	0	C
18	55	0	C
19	640	0	D
20	405	1	D
21	286	1	D
22	456	0	D
23	295	0	D
24	354	0	D

Descritivas

De maneira a se desenvolver um entendimento inicial dos dados dispostos, foi realizada algumas análises descritivas dos dados, apresentadas a seguir.

- Referente a variável de produção:

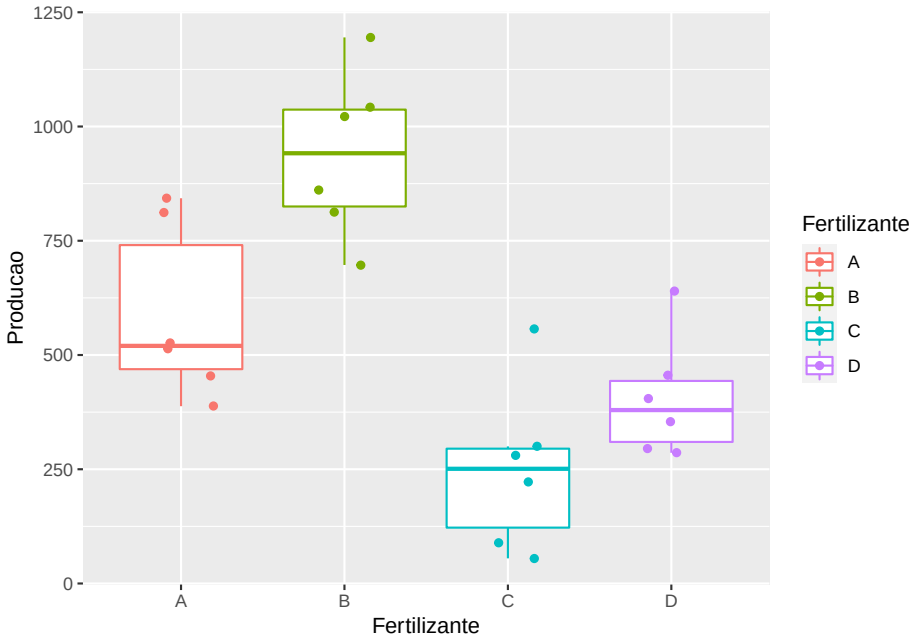
```
library(tidyverse)
df |> group_by(Fertilizante) |>
  summarise(media=mean(Producao), desv.pad=sd(Praga))|>
  kable(align='c', digits=2)
```

Fertilizante	media	desv.pad
A	589.50	0.52
B	938.33	0.55
C	250.50	0.52
D	406.00	0.52

- A partir das informações das médias das produções para cada tipo de tratamento expostas acima, tem-se uma suspeita de que, aparentemente, existe alguma diferença entre os tratamentos.
- Ao se observar o gráfico abaixo, tal suspeita torna-se mais concreta, visto que as médias entre os experimentos estão bem diferentes. É possível observar que o tratamento B apresenta valores de produção superiores aos restantes, enquanto o tratamento C consiste dos valores mais baixos.

- Uma maneira de identificar se existe alguma diferença significativa entre as médias é por meio da realização de um teste ANOVA.

```
df |> ggplot(aes(x=Fertilizante, y=Producao, color=Fertilizante)) +
  geom_boxplot(outlier.shape=NA) +
  geom_jitter(width=0.2)
```



ANOVA

Como queremos verificar se existe alguma diferença entre os tratamentos, utilizando a notação de testes de hipótese, o que desejamos testar pode ser representado da seguinte maneira:

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$$

Ou seja, a hipótese nula dos dados é que todas as médias de produção para os diferentes fertilizantes são iguais. Aplicando a anova:

```
m <- aov(Producao ~ Fertilizante, data = df)
summary(m)
```

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## Fertilizante  3 1576427  525476   17.66 7.66e-06 ***
## Residuals    20  595078   29754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A partir das informações fornecidas com a aplicação da ANOVA, foi encontrado como p-valor o resultado 7.66e-06. Trata-se de um valor bem menor ao nível de significância estabelecido para o teste, indicado que há evidência para rejeitar a hipótese nula. Considerando isso, a suspeita anterior de que haveria uma diferença entre as médias das produções para os diferentes tratamentos é confirmada pelo teste.

Pressupostos

Em relação ao teste aplicado previamente, é importante comentar sobre as suposições realizadas para sua aplicação seja válida: - Populações normalmente distribuídas; - Populações tem mesma variância; - Amostras são aleatórias e mutuamente independentes;

Já abordando sobre os resíduos, podemos observar como eles se comportam.

O modelo que estamos considerando diz que uma observação de uma produção para um dado tratamento é igual a média para aquele tratamento somado ao seu respectivo resíduo, como é mostrado a seguir:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

Partindo disto, pode-se definir o resíduo como o valor da observação menor a média amostral do tratamento. A expressão é apresentada a seguir:

$$e_{ij} = y_{ij} - \bar{y}_{i\cdot}$$

```
residuo <- m$residuals
dado <- data.frame(id=1:12, residuo=residuo)
kable(dado)
```

id	residuo
1	-201.50000
2	-135.50000
3	222.50000
4	-75.50000
5	-63.50000
6	253.50000
7	103.66667
8	-241.33333
9	-125.33333
10	-77.33333
11	256.66667
12	83.66667
1	29.50000
2	-28.50000
3	-161.50000
4	306.50000
5	49.50000
6	-195.50000
7	234.00000
8	-1.00000
9	-120.00000
10	50.00000
11	-111.00000
12	-52.00000

Na exposição acima é possível realizar uma observação inicial quanto aos resíduos encontrados pela ANOVA.

Resíduo versus tratamento

```
plot(m, which=5, pch=19)
```

O gráfico acima apresenta a relação entre os tratamentos e os valores dos resíduos. A partir dele, esperava-se observar que os valores para todos os tratamentos estivessem próximos de zero caso a hipótese nula fosse verdadeira.

Resíduo versus valor ajustado

Como o modelo é

$$y_{ij} = \mu_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

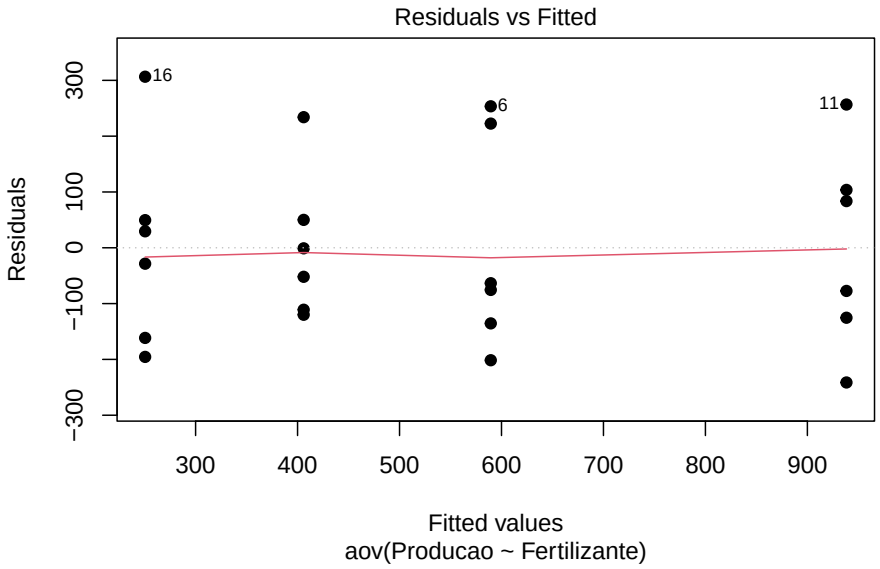
então o valor ajustado é

$$\hat{y}_{ij} = \bar{y}_{i\cdot}$$

```
library(tidyverse)
df |> group_by(Fertilizante) |>
  summarise(media = mean(Producao), desv.pad = sd(Producao)) |>
  kable(align='c', digits=2)
```

Fertilizante	media	desv.pad
A	589.50	191.03
B	938.33	180.86
C	250.50	180.21
D	406.00	131.67

```
plot(m, which=1, pch=19)
```

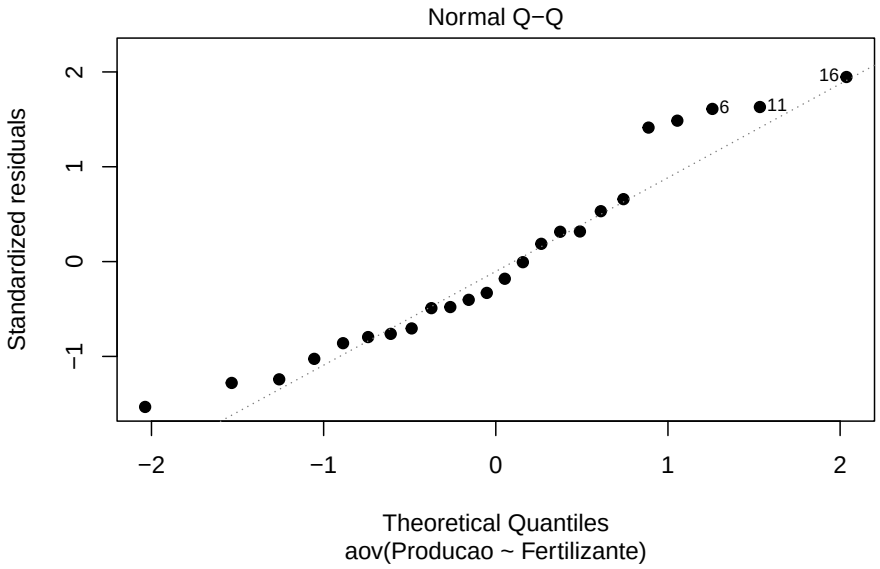


Analisando o gráfico é possível observar levemente que os resíduos estão distribuidos aleatoriamente e também que aparentemente a variância deles é parceida. Os pontos apresentam uma certa aleatoriedade nos dois "lados" de 0, um bom comportamento.

Normalidade dos resíduos

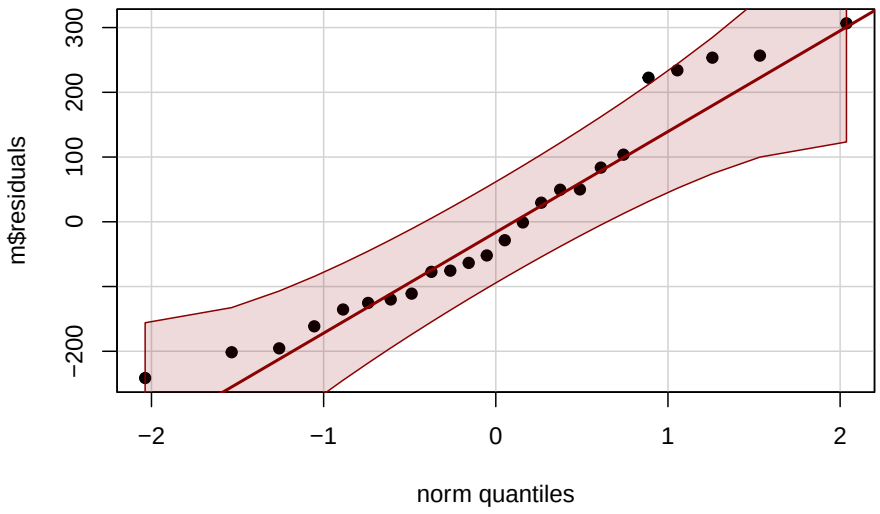
A normalidade dos resíduos é um pressuposto importante para o estudo das médias dos tratamentos utilizando a ANOVA. Para realizar uma análise referente a eles, pode-se utilizar um gráfico ggplot. Nele, para se observar a normalidade dos resíduos, os valores devem seguir uma relação linear e estarem próximos da reta presente no gráfico.

```
plot(m, which=2, pch = 19)
```



Para auxiliar na análise, é possível utilizar o recurso de um envelope a ser desenhado no gráfico que irá indicar uma "faixa de normalidade". Os pontos que fugirem da faixa podem apontar para uma falha na normalidade.

```
library("car") # se necessário: install.packages("car")
qqPlot(m$residuals, pch=19, col.lines="darkred", id=F)
```



Testando a normalidade dos resíduos

O Teste de Shapiro-Wilk tem como objetivo avaliar se uma distribuição é semelhante a uma distribuição normal. A distribuição normal também pode ser chamada de gaussiana e sua forma assemelha-se a de um

sino. Esse tipo de distribuição é muito importante, por ser frequentemente usada para modelar fenômenos naturais. Além disse, vale observar que para dizer que uma distribuição é normal, o valor p precisa ser maior do que 0.05.

Para testar a normalidade é possível aplicar o teste de Shapiro-Wilks. As hipóteses do teste são:

$$H_0 = \text{Testes seguem uma distribuição normal} H_1 = \text{Testes não seguem uma distribuição normal}$$

Aplicando o teste:

```
shapiro.test(m$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m$residuals
## W = 0.93551, p-value = 0.1295
```

Com um nível de relevância p = 0.05, é possível observar que não temos evidências suficientes para rejeitar a hipótese nula, dessa forma podemos assumir que os dados são normais. O pressuposto para a normalidade dos resíduos é válido.

ANOVA para dados binários

Regressão Logística

Geralmente a regressão logística é usada para ajudar a criar previsões precisas. É semelhante à regressão linear, exceto que, em vez de um resultado gráfico, a variável de destino é binária, o valor é 0 ou 1.

Nesse caso, a regressão logística será usada para fazermos os dados da ANOVA de maneira binária, vale notar que além da regressão utilizada, quando aplicamos a ANOVA no nosso modelo usamos como teste "Chisq" que quer dizer "chi-squared", ou seja, nessa análise de dados utilizamos do teste qui-quadrado para validarmos nosso resultado.

```
modelo <- glm(Praga ~ Fertilizante, data = df, family = binomial(link = "logit"))
summary(modelo)
```

```
##
## Call:
## glm(formula = Praga ~ Fertilizante, family = binomial(link = "logit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1774  -0.9005  -0.9005   1.2536   1.4823
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.931e-01  8.660e-01  -0.800    0.423
## FertilizanteB  6.931e-01  1.190e+00   0.582    0.560
## FertilizanteC  4.079e-16  1.225e+00   0.000    1.000
```

```
## FertilizanteD  8.158e-16  1.225e+00   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 31.755  on 23  degrees of freedom
## Residual deviance: 31.232  on 20  degrees of freedom
## AIC: 39.232
##
## Number of Fisher Scoring iterations: 4
```

ANOVA

```
anova(modelo, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Praga
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                             23     31.755
## Fertilizante  3  0.52276        20     31.232   0.9139
```

Analisando os dados, pode-se notar que atingimos o resultado esperado haja vista que tivemos uma resposta semelhante quando comparado com a ANOVA. Observando o resultado, podemos assumir que não temos evidências para rejeitar a hipótese nula. Assim, assumimos que os fertilizantes não interferem nas pragas durante a produção.

Testes para homogeneidade na variância

F

É possível utilizar um teste F para comparar duas variâncias. As hipóteses para o teste F são:

$$H_0 = \text{As variâncias são iguais} H_1 = \text{As variâncias são diferentes}$$

Par a par A e B

```
var.test(Producao ~ Fertilizante, data=df[df$Fertilizante=="A" | df$Fertilizante=="B",])
```

```
##
##  F test to compare two variances
##
## data:  Producao by Fertilizante
```

```
## F = 1.1156, num df = 5, denom df = 5, p-value = 0.9074
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.156108 7.972555
## sample estimates:
## ratio of variances
##           1.115607
```

A e C

```
var.test(Producao ~ Fertilizante, data=df[df$Fertilizante=="A" | df$Fertilizante=="C",])
```

```
##
##  F test to compare two variances
##
## data:  Producao by Fertilizante
## F = 1.1237, num df = 5, denom df = 5, p-value = 0.9013
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1572403 8.0303850
## sample estimates:
## ratio of variances
##           1.123699
```

A e D

```
var.test(Producao ~ Fertilizante, data=df[df$Fertilizante=="A" | df$Fertilizante=="D",])
```

```
##
##  F test to compare two variances
##
## data:  Producao by Fertilizante
## F = 2.105, num df = 5, denom df = 5, p-value = 0.4334
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   0.2945512 15.0429598
## sample estimates:
## ratio of variances
##           2.104976
```

B e C

```
var.test(Producao ~ Fertilizante, data=df[df$Fertilizante=="C" | df$Fertilizante=="B",])
```

```
##
##  F test to compare two variances
##
## data:  Producao by Fertilizante
## F = 1.0073, num df = 5, denom df = 5, p-value = 0.9939
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
```

```
##  0.140946 7.198219
## sample estimates:
## ratio of variances
##           1.007254
```

B e D

```
var.test(Producao ~ Fertilizante, data=df[df$Fertilizante=="D" | df$Fertilizante=="B",])
```

```
##
##  F test to compare two variances
##
## data:  Producao by Fertilizante
## F = 1.8868, num df = 5, denom df = 5, p-value = 0.5027
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   0.2640277 13.4841012
## sample estimates:
## ratio of variances
##           1.886843
```

C e D

```
var.test(Producao ~ Fertilizante, data=df[df$Fertilizante=="C" | df$Fertilizante=="D",])
```

```
##
##  F test to compare two variances
##
## data:  Producao by Fertilizante
## F = 1.8733, num df = 5, denom df = 5, p-value = 0.5075
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   0.2621264 13.3869963
## sample estimates:
## ratio of variances
##           1.873255
```

Observando todos os testes par a par acima, é possível observar que aparentemente existe homogeniedade entre as variâncias. isso satisfaz um pressuposto importante da anova.

Bartlett

```
bartlett.test(Producao ~ Fertilizante, data=df)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Producao by Fertilizante
## Bartlett's K-squared = 0.70949, df = 3, p-value = 0.871
```

O teste de barlett ainda confirma isso.

Teste de tukey

Feita a anova e visto que existem pelo menos dois tratamentos com diferença significativa, é possível aplicar um teste para encontrar onde vem a diferença. O gráfico plotado na análise descritiva sugere uma diferença entre quase todos os tratamentos, resta saber quais são significantes.

```
TukeyHSD(m)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Producao ~ Fertilizante, data = df)
##
## $Fertilizante
##          diff        lwr       upr     p adj
## B-A  348.8333   70.09003  627.5766 0.0110653
## C-A -339.0000 -617.74330  -60.2567 0.0137652
## D-A -183.5000 -462.24330   95.2433 0.2836445
## C-B -687.8333 -966.57663 -409.0900 0.0000059
## D-B -532.3333 -811.07663 -253.5900 0.0001702
## D-C  155.5000 -123.24330  434.2433 0.4218483
```

Analisando o resultado do teste de tukey, é possível perceber diversas diferenças. Com um nível de signifcância de 5%, o fertilizante B apresentou diferença significativa para todos os outros fertilizantes. Nesse sentido, podemos que o tratamento B apresentou uma melhora na produção estatisticamente significativa. O inverso não pode ser dito para o tratamento C. Ele apresenta a menor média, mas não temos evidências suficientes para mostrar uma diferença com o tratamento D.
