Atividade III - Recursos Computacionais (Universidade Federal de Lavras)

Relatório de Estatística Experimental

Antonio Mendes M. Jr

23/05/2019

Sumário

Exercício 5.8.5 (Pimentel-Gomes, Curso de Estatística Experimental)

Ensaio de competição de variedades de mandioca, realizado pelo Instituto de Pesquisas Agronômicas do Leste (atual Centro Nacional de Pesquisa de Mandioca e Fruticultura, da EMBRAPA), em Cruz das Almas, BA. Neste experimento estudou-se a produtividade de variedades de mandioca. O experimento foi delineado em blocos casualizados, provavelmente em razão da não homogeneidade das condições do solo. As produtividades encontradas, em t/ha, são as seguintes:

Variedades	Bloco 1	Bloco 2	Bloco 3	Bloco 4
Aipim Bravo	14.5	15.8	24.0	17.0
Milagrosa	5.7	5.9	10.5	6.6
Sutinga	5.3	7.7	10.2	9.6
Salangó Preta	4.6	7.1	10.4	10.8
Mamão	14.8	12.6	18.8	16.0
Escondida	8.2	8.2	12.7	17.5

Preparação dos dados

Pacotes utilizados

Primeiramente foram importados os pacotes utilizados no relatório.

Para maiores informações sobre os pacotes utilize o comando `?nome_do_pacote` e verifique sua documentação.

Informações sobre a instalação do pacote `labestData` podem ser encontradas em: https://gitlab.c3sl.ufpr.br/pet-estatistica/labestData/blob/devel/README.md#instalao-pelo-repositrio

```
library(labestData)
library(ExpDes.pt)
library(lattice)
library(ggplot2)
library(dplyr)
library(MASS)
library(fBasics)
```

Informações sobre os dados

Os dados utilizados estão presentes no pacote `labestData`. Este pacote foi criado pelo PET Estatística da UFPR com o objetivo de reunir conjuntos de dados para auxiliar no ensino de estatística.

Os dados foram importados e sua disposição verificada por meio das funções `head()` e `tail()`.

```
dados <- PimentelEx5.8.5
```

```
head(dados,6)
```

```
  bloco    variedade prod
1     i   AipimBravo 14.5
2     i    Milagrosa  5.7
3     i      Sutinga  5.3
4     i SalangoPreta  4.6
5     i        Mamao 14.8
6     i    Escondida  8.2
```

```
tail(dados,6)
```

```
   bloco    variedade prod
19    iv   AipimBravo 17.0
20    iv    Milagrosa  6.6
21    iv      Sutinga  9.6
22    iv SalangoPreta 10.8
23    iv        Mamao 16.0
24    iv    Escondida 17.5
```

Utilizando a função `str()`, a estrutura dos dados também foi verificada. As colunas de blocos e tratamentos já são estruturadas como fatores, não sendo necessária a conversão.

```
str(dados)
```

```
'data.frame': 24 obs. of  3 variables:
 $ bloco    : Factor w/ 4 levels "i","ii","iii",..: 1 1 1 1 1 1 2 2 2 2 ...
 $ variedade: Factor w/ 6 levels "AipimBravo","Escondida",..: 1 4 6 5 3 2 1 4 6 5 ...
 $ prod     : num  14.5 5.7 5.3 4.6 14.8 8.2 15.8 5.9 7.7 7.1 ...
```

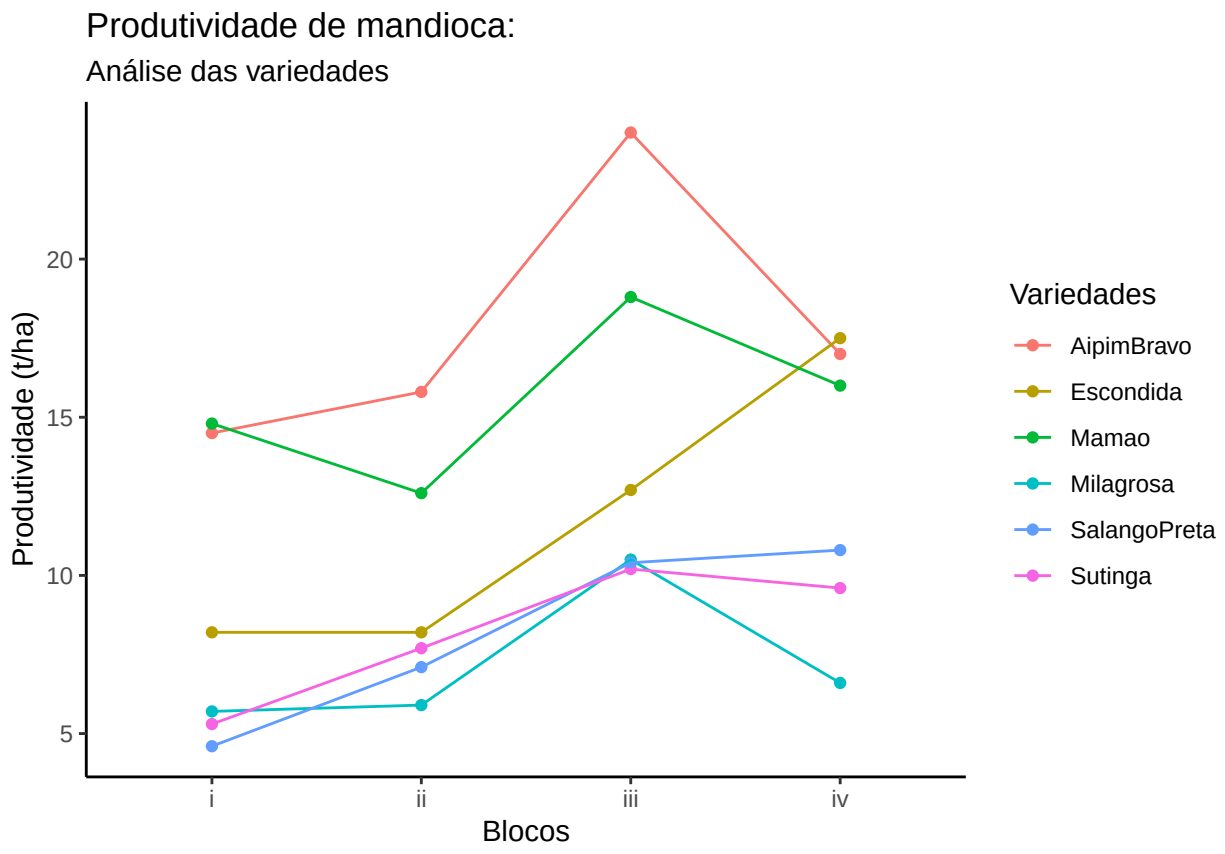
Análise exploratória

Neste tópico foram analisadas algumas características dos dados com a utilização de gráficos.

Diagrama de dispersão

Este gráfico apresenta os níveis de produção de cada uma das variedades em cada um dos blocos.

```
ggplot(dados,
       aes(x=bloco, y=prod, color= variedade, group=variedade)) +
  geom_point() +
  geom_line()+
  theme_classic()+
  labs(
    title = "Produtividade de mandioca: ",
    subtitle = "Análise das variedades",
    x = "Blocos",
    y = "Produtividade (t/ha)",
    color= "Variedades")
```



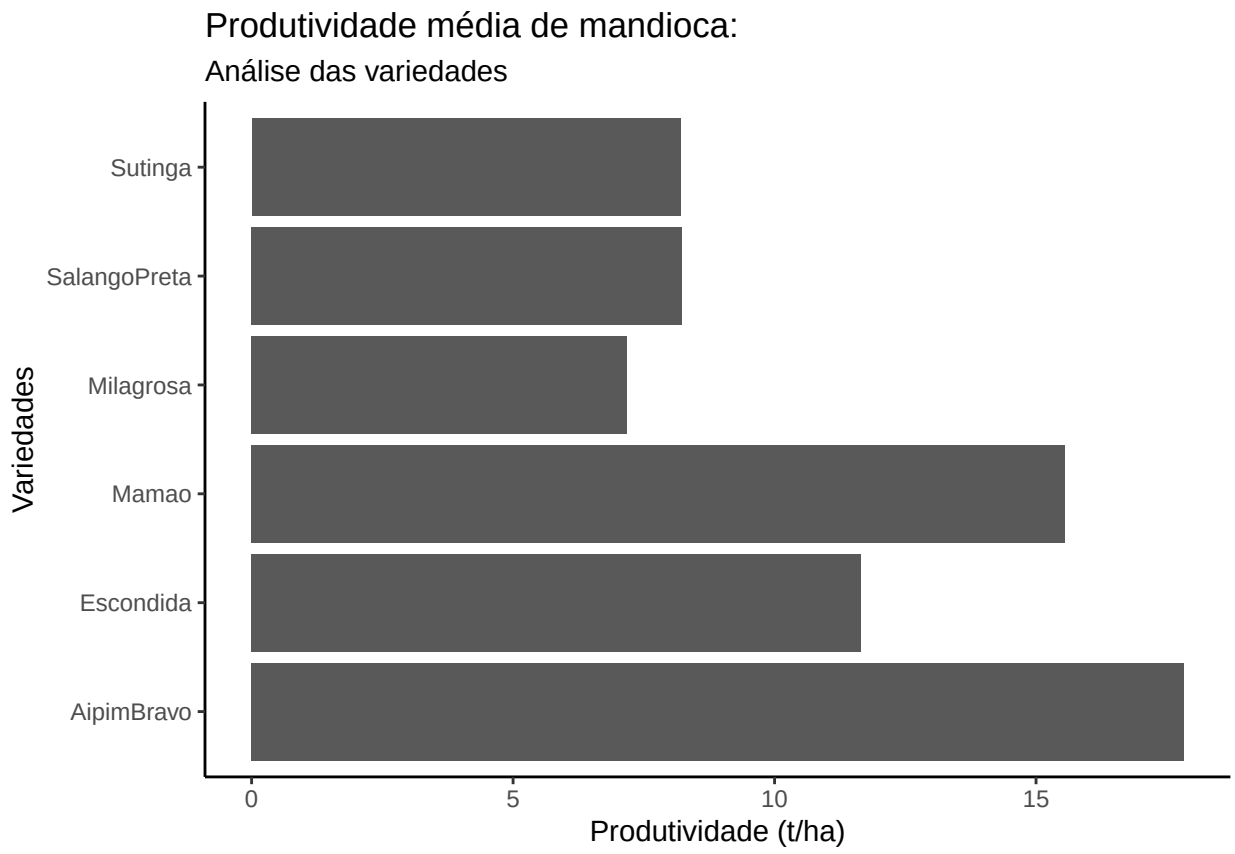
Gráficos de barras

A função `tapply()` torna possível o cálculo da média de produtividade para cada um dos blocos e das variedades presentes no experimento.

Análise das variedades (produtividade média):

```
medvar <- tapply(dados$prod, dados$variedade, mean)
medvar <- data.frame(variedade=names(medvar),media=medvar)


ggplot(medvar, aes(x=variedade, y=media)) +
       geom_col()+
       theme_classic()+
       coord_flip()+
       labs(
       title = "Produtividade média de mandioca: ",
       subtitle = "Análise das variedades",
       x = "Variedades",
       y = "Produtividade (t/ha)")
```



Análise dos blocos (produtividade média):

```
medbloc <- tapply(dados$prod, dados$bloco, mean)
medbloc <- data.frame(bloco=names(medbloc),media=medbloc)


ggplot(medbloc, aes(x=bloco, y=media)) +
       geom_col()+
       theme_classic()+
       labs(
       title = "Produtividade média de mandioca: ",
       subtitle = "Análise dos blocos",
       x = "Blocos",
       y = "Produtividade (t/ha)")
```

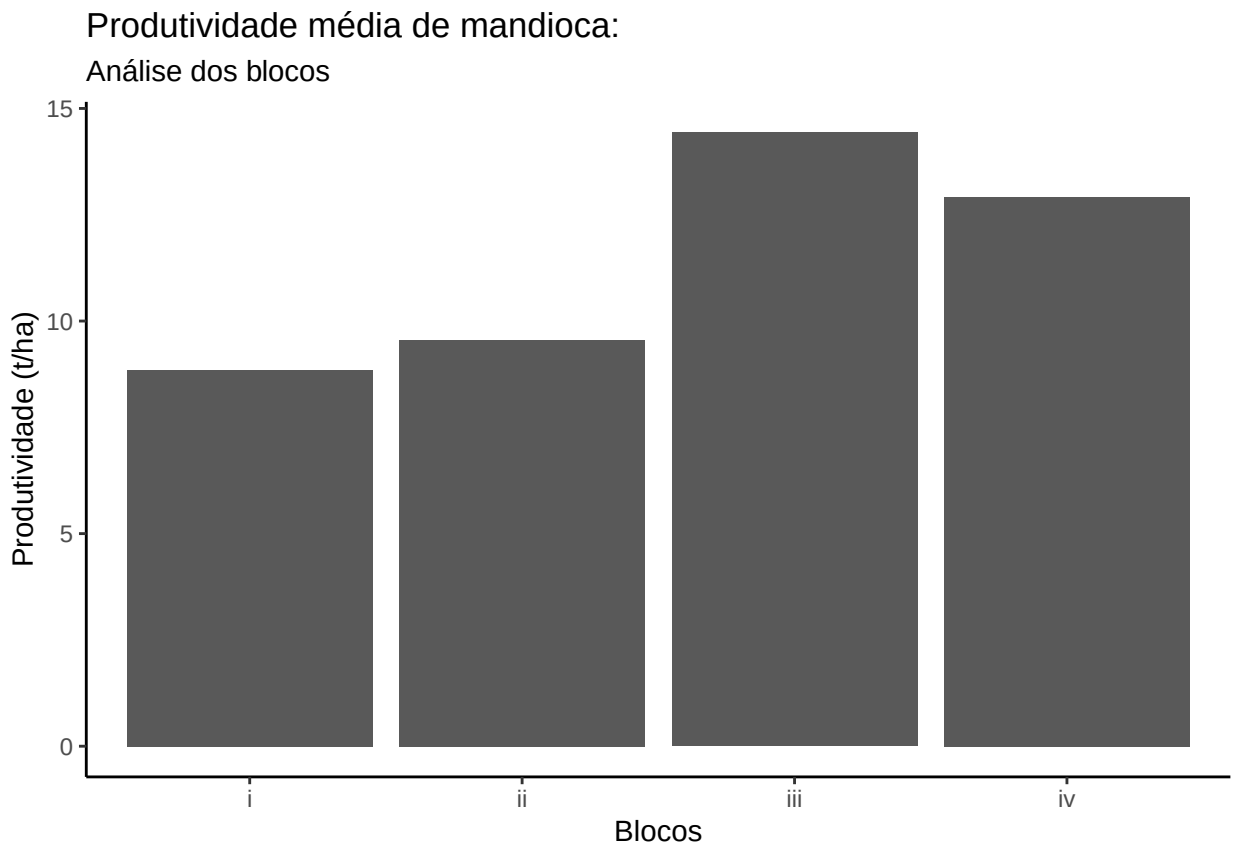
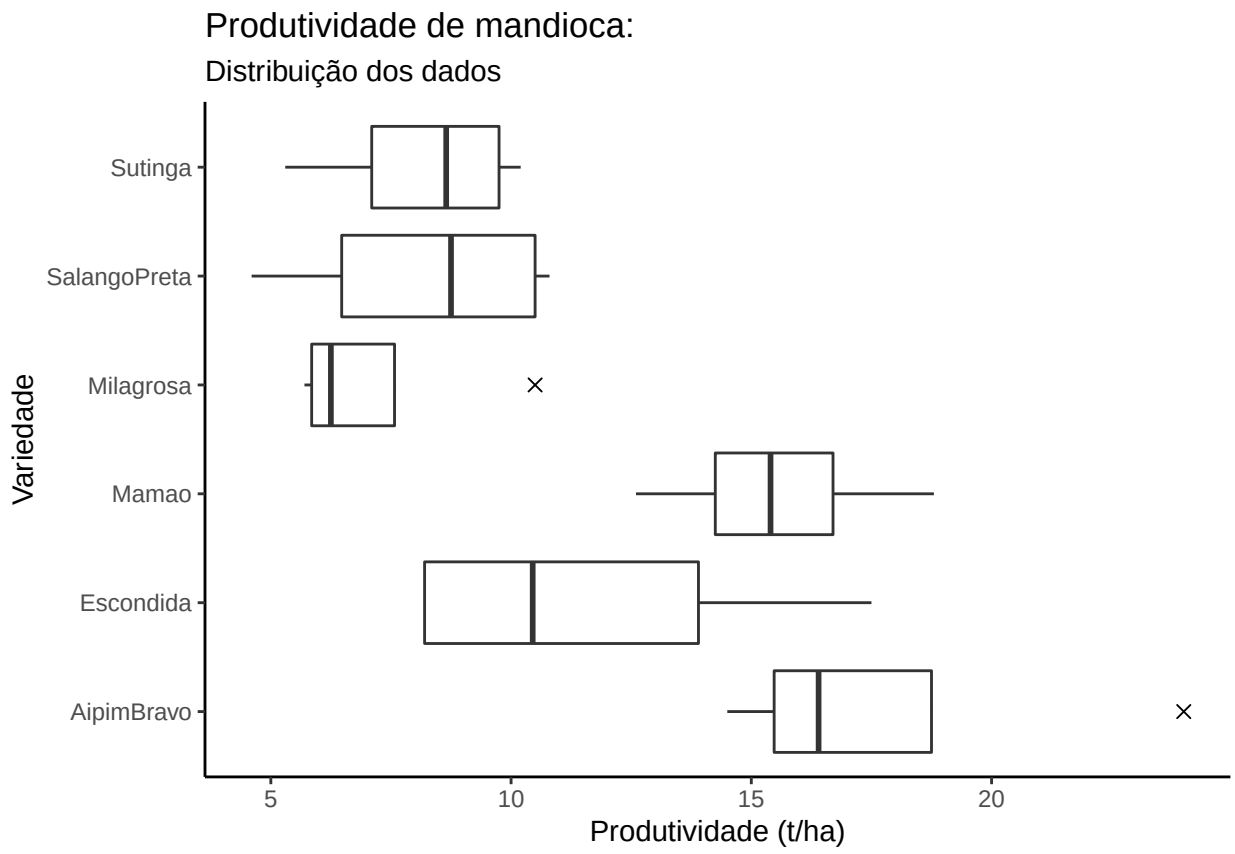


Gráfico boxplot

O gráfico do tipo *boxplot* foi utilizado para verificar características da distribuição dos dados de cada variedade.

```
ggplot(dados, aes(x = variedade, y = prod)) +
  geom_boxplot(outlier.shape  =4, outlier.size = 2, outlier.color = "black") +
  theme_classic()+
  coord_flip()+
  labs(
  title = "Produtividade de mandioca: ",
  subtitle = "Distribuição dos dados",
  x = "Variedade",
  y = "Produtividade (t/ha)")
```



Análise de variância

Utilizando a função `aov()`, foi realizada a análise de variância do experimento.

```
dados_anova <- aov(dados$prod ~ dados$variedade + dados$bloco)

# Mostra a tabela ANOVA
summary(dados_anova)
```

```
                Df Sum Sq Mean Sq F value   Pr(>F)
dados$variedade  5  386.9   77.38   19.29 4.73e-06 ***
dados$bloco      3  128.5   42.84   10.68  0.00052 ***
Residuals       15   60.2    4.01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Análise de variância de forma matricial

A análise de variância foi também realizada de forma matricial.

Temos que o modelo de um DBC é dado por:

$$y_{ij} = \mu + b_j + t_i + \varepsilon_{ij}$$

Em que y_{ij} é a observação referente ao tratamento i no bloco j, μ é a constante comum a todas as observação, b_j é o efeito do j-ésimo bloco, t_i é o efeito do i-ésimo tratamento e ε_{ij} é o erro.

Matricialmente temos:

$$Y = X\beta + \varepsilon$$

Em que Y é o vetor resposta, X é a matriz de delineamento, β é o vetor de parâmetros e ε o vetor de erros.

Aplicado ao experimento em estudo, temos:

$$y_{ij} = \begin{bmatrix} 14.5 \\ 5.7 \\ 5.3 \\ 4.6 \\ 14.8 \\ 8.2 \\ 15.8 \\ 5.9 \\ 7.7 \\ 7.1 \\ 12.6 \\ 8.2 \\ 24.0 \\ 10.5 \\ 10.2 \\ 10.4 \\ 18.8 \\ 12.7 \\ 17.0 \\ 6.6 \\ 9.6 \\ 10.8 \\ 16.0 \\ 17.5 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mu \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,1} \\ \varepsilon_{1,2} \\ \varepsilon_{1,3} \\ \varepsilon_{1,4} \\ \varepsilon_{1,5} \\ \varepsilon_{1,6} \\ \varepsilon_{2,1} \\ \varepsilon_{2,2} \\ \varepsilon_{2,3} \\ \varepsilon_{2,4} \\ \varepsilon_{2,5} \\ \varepsilon_{2,6} \\ \varepsilon_{3,1} \\ \varepsilon_{3,2} \\ \varepsilon_{3,3} \\ \varepsilon_{3,4} \\ \varepsilon_{3,5} \\ \varepsilon_{3,6} \\ \varepsilon_{4,1} \\ \varepsilon_{4,2} \\ \varepsilon_{4,3} \\ \varepsilon_{4,4} \\ \varepsilon_{4,5} \\ \varepsilon_{4,6} \end{bmatrix} \tag{1}$$

Em que a primeira coluna da matriz de delineamento corresponde ao efeito da média geral, as quatro seguintes correspondem aos efeitos dos quatro blocos e as seis últimas aos efeitos dos seis tratamentos.

Obtendo os projetores:

```
nobs  <- 24;
nbloc <-  4;
ntrat <-  6;

xmi   <- matrix(c(rep(1,nobs)))

xbloc <- matrix(c(rep(1,6),rep(0,18),
                  rep(0,6),rep(1,6),rep(0,12),
                  rep(0,12),rep(1,6),rep(0,6),
                  rep(0,18),rep(1,6)),ncol=nbloc);

xtrat <- matrix(c(1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,
                  0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,
                  0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,
                  0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,
                  0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,
                  0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1),ncol=ntrat);

Y <- dados$prod;
```

```
X <- cbind(xmi,xbloc,xtrat);

n <- length(Y);

InvXlX <- ginv(t(X)%*%X);

B=InvXlX%*%t(X)%*%Y;

XlY <- t(X)%*%Y

P <- X%*%InvXlX%*%t(X);
E.n <- X[,1]%*%ginv(X[,1]);
I <- diag(rep(1,length(Y)));
X1 <- matrix(c(X[,1]),nrow(X));

Xt <- matrix(cbind(X[,1],X[,6:11]),nrow(X));
Pt <- Xt%*%ginv(t(Xt)%*%Xt)%*%t(Xt);

Xb <- matrix(c(X[,1:5]),nrow(X));
Pb <- Xb%*%ginv(t(Xb)%*%Xb)%*%t(Xb);
```

Calculando as somas de quadrados, graus de liberdade e quadrados médios:

```
SQtrat  <- round(t(Y)%*%(Pt-E.n)%*%Y,2);
SQbloc  <- round(t(Y)%*%(Pb-E.n)%*%Y,2);
SQres   <- round(t(Y)%*%(I-P)%*%Y,2);
SQtotal <- round(t(Y)%*%(I-E.n)%*%Y,2);

GLtrat <- rk(Pt-E.n);
GLbloc <- rk(Pb-E.n);
GLres <- rk(I-P);
GLtotal <- rk(I-E.n);

QMtrat <- round(SQtrat/GLtrat,2);
QMbloc <- round(SQbloc/GLbloc,2);
QMres <- round(SQres/GLres,2);
```

Partindo finalmente para o quadro de análise de variância, temos:

```
Fct <- round((QMtrat)/(QMres),4);
Fcb <- round((QMbloc)/(QMres),4);

Ftab=qf(0.95,GLtrat,GLres);
pvalt=round(1-pf(Fct,GLtrat,GLres),4);

Ftab=qf(0.95,GLbloc,GLres);
pvalb=round(1-pf(Fcb,GLbloc,GLres),4);

analise<-data.frame(FV = c("Tratamento","Bloco","Residuo","Total"),
                    GL = c(GLtrat,GLbloc,GLres,GLtotal),
                    SQ=round(c(SQtrat,SQbloc,SQres,SQtotal),2),
                    QM = c((round(c(QMtrat,QMbloc,QMres),2)),""),
                    F = c(round(Fct,4),round(Fcb,4), " ", " "),
                    Pvalue = c(round(pvalt,6), round(pvalb,6),"", " "))
```

```
analise
```

```
           FV GL     SQ    QM       F Pvalue
1 Tratamento  5 386.91 77.38 19.2968      0
2      Bloco  3 128.52 42.84 10.6833  5e-04
3    Residuo 15  60.18  4.01
4      Total 23 575.62
```

Os resultados vão de encontro aos do método anterior, utilizando a função `aov()`.

Pressuposições

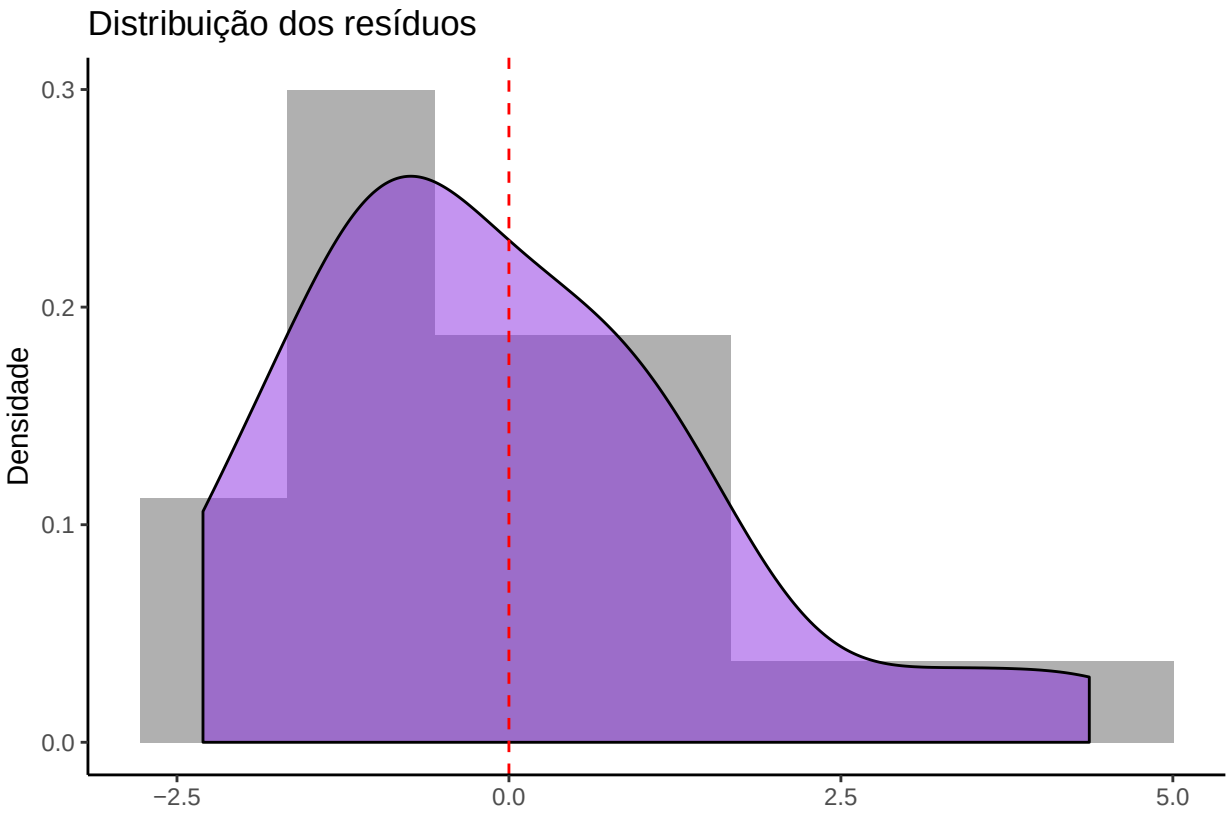
Os resíduos do modelo foram obtidos por meio da função `resid()` e as pressuposições de normalidade dos resíduos e homogeneidade das variâncias foram testadas.

```
res <- resid(dados_anova)
```

Normalidade dos resíduos

Graficamente não é possível afirmar que os resíduos seguem uma distribuição normal.

```
ggplot(data.frame(res), aes(x = res)) +
  geom_histogram(aes(y=..density..), bins = 7, fill="gray69") +
  geom_density(fill = "blueviolet", alpha = 0.5) +
  geom_vline(linetype=2, color = "red", xintercept = mean(res)) +
  theme_classic()+
  labs(
  title = "Distribuição dos resíduos",
  x = " ",
  y = "Densidade")
```



A fim de verificar a normalidade dos resíduos, foi realizado o teste de Shapiro-Wilk. Os resultados sugerem a não rejeição da hipótese nula a um nível de significância de 5%, indicando que os resíduos possuem distribuição normal.

H0: Os resíduos possuem distribuição normal

H1: Os resíduos não possuem distribuição normal

```
shapiro.test(res)
```

```
    Shapiro-Wilk normality test

data:  res
W = 0.93298, p-value = 0.1137
```

Homocedasticidade dos resíduos

Foi realizado o teste de O'Neill e Mathews a fim de se verificar se as variâncias são homogênas. Os resultados sugerem a não rejeição da hipótese nula a um nível de significância de 5%, indicando que as variâncias não se diferem.

H0: As variâncias são homogêneas

H1: As variâncias não são homogêneas

```
oneilldbc(dados$prod, dados$variedade, dados$bloco)
```

```
[1] 0.5083825
```

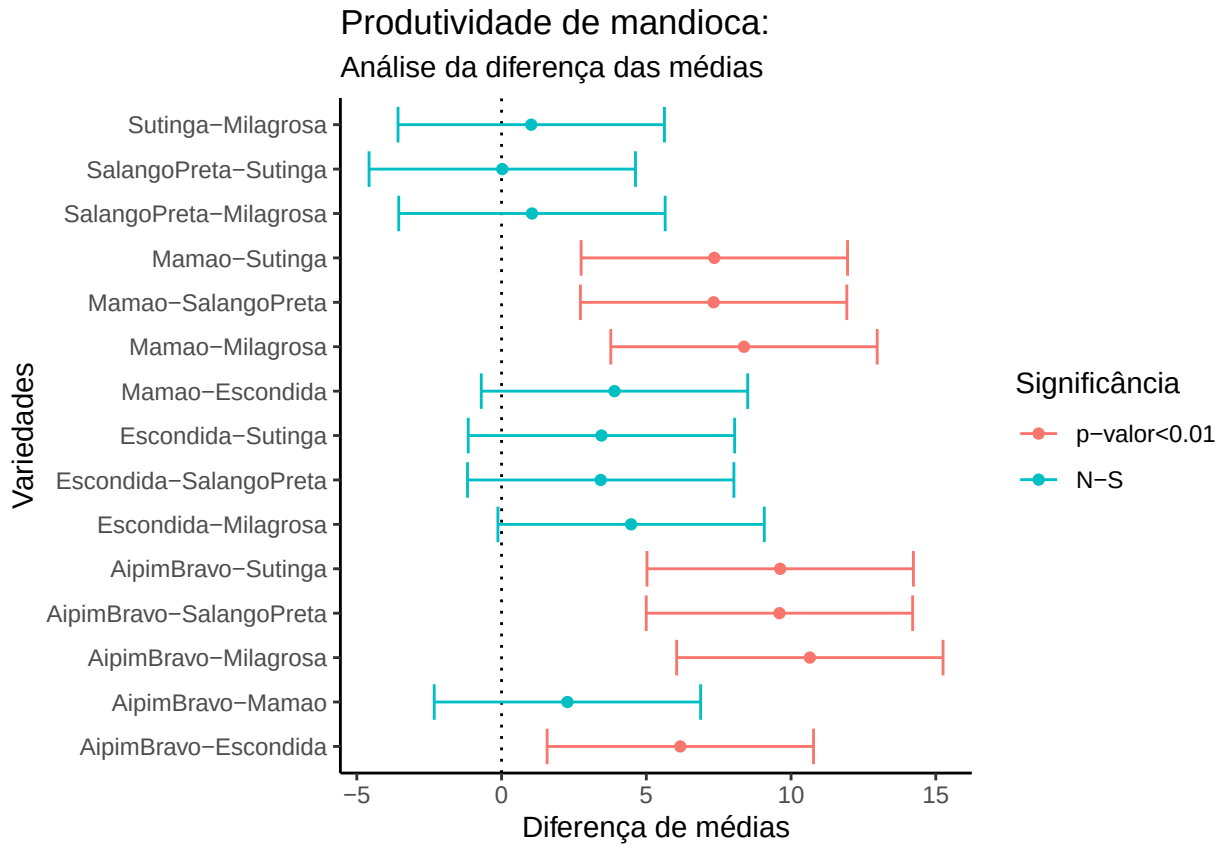
Teste de Tukey

Por fim, dado que as pressuposições foram atendidas, foi realizado o teste de Tukey para as médias das variedades.

```
resTukey <- TukeyHSD(dados_anova,"dados$variedade",ordered=T)

tky = as.data.frame(resTukey$`dados$variedade`)
tky$pares = rownames(tky)

# Gráfico mostrando a comparação das varidades pelo teste de Tukey
ggplot(tky, aes(color=cut(`p adj`, c(0, 0.01, 0.05, 1), # 0-0.01, 0.01-0.05 e 0.05-1
                          label=c("p-valor<0.01","valor-p < 0.05","N-S")))) +
  theme_classic()+
  coord_flip()+
  geom_hline(yintercept=0, lty="13", colour="black") + # linha em 0
  geom_errorbar(aes(pares, ymin=lwr, ymax=upr), width=0.8) +
  geom_point(aes(pares, diff)) +
  labs(
    title = "Produtividade de mandioca: ",
    subtitle = "Análise da diferença das médias",
    x = "Variedades",
    y = "Diferença de médias",
    colour="Significância")
```



Utilização do pacote `expDes.pt`

Os procedimentos anteriores foram novamente realizados com o auxílio do pacote `expDes.pt` a fim de verificar se os resultados obtidos seriam equivalentes. Esse pacote possui a função `dbc()` que realiza a análise completa dos dados, fornecendo não só os resultados das análises de variância como também os resultados dos testes das pressuposições e das médias.

```
dbc(dados$variedade,dados$bloco,dados$prod)
```

```
------------------------------------------------------------------------
Quadro da analise de variancia
------------------------------------------------------------------------
           GL     SQ     QM     Fc      Pr>Fc
Tratamento  5 386.91 77.383 19.288 0.00000473
Bloco       3 128.52 42.842 10.679 0.00052043
Residuo    15  60.18  4.012
Total      23 575.62
------------------------------------------------------------------------
CV = 17.51 %

------------------------------------------------------------------------
Teste de normalidade dos residuos
valor-p:  0.1136566
De acordo com o teste de Shapiro-Wilk a 5% de significancia, os residuos podem ser considerados norma
------------------------------------------------------------------------

------------------------------------------------------------------------
Teste de homogeneidade de variancia
valor-p:  0.5083825
De acordo com o teste de oneillmathews a 5% de significancia, as variancias podem ser consideradas ho
------------------------------------------------------------------------

Teste de Tukey
------------------------------------------------------------------------
Grupos Tratamentos Medias
a      AipimBravo      17.825
ab     Mamao   15.55
 bc    Escondida   11.65
  c    SalangoPreta    8.225
  c    Sutinga     8.2
  c    Milagrosa   7.175
------------------------------------------------------------------------
```

Conclusões

As pressuposições para a realização da análise de variância foram satisfeitas, tendo as variâncias homogêneas e os resíduos apresentando distribuição normal.

O teste de Tukey, com significância de 5%, indica que a variedade Aipim Bravo possui a maior produtividade, entretanto ela não apresenta diferença estatisticamente significante quando comparada à variedade Mamão. Por sua vez, a variedade Mamão não se diferencia da variedade Escondida. Por fim, todas essas variedades já citadas se diferenciam das variedades Salangó Preta, Sutinga e Milagrosa, que não se diferenciam entre sim.

Referências

Boxplot. *Portal Action*. Disponível em http://www.portalaction.com.br/estatistica-basica/31-boxplot. Acessado em: 16 mai. 2019.

Diethelm Wuertz, Tobias Setz and Yohan Chalabi (2017). *fBasics: Rmetrics - Markets and Basic Statistics*. R package version 3042.89. https://CRAN.R-project.org/package=fBasics.

Entendendo o uso das funções apply, lapply, sapply, tapply, mapply. *Produção Animal*. Disponível em https://producaoanimalcomr.wordpress.com/2015/12/10/entendendo-o-uso-das-funcoes-apply-lapply-sapply-tapply-mapply/. Acessado em: 16 mai. 2019.

Eric Batista Ferreira, Portya Piscitelli Cavalcanti and Denismar Alves Nogueira (2018). *ExpDes.pt: Pacote Experimental Designs (Portuguese)*. R package version 1.2.0. https://CRAN.R-project.org/package=ExpDes.pt.

Geom_density in ggplot2. *Ploty*.Disponível em https://plot.ly/ggplot2/geom_density/. Acessado em: 16 mai. 2019.

Ggplot2 add straight lines to a plot: horizontal, vertical and regression lines. *STHDA*. Disponível em http://www.sthda.com/english/wiki/ggplot2-add-straight-lines-to-a-plot-horizontal-vertical-and-regression-lines. Acessado em: 16 mai. 2019.

Ggplot2 Quick Reference: colour (and fill). *SAPE*. Disponível em http://sape.inf.usi.ch/quick-reference/ggplot2/colour. Acessado em: 16 mai. 2019.

Gráficos com estilo - ggplot2. Disponível em https://curso-r.github.io/posts/aula05.html. Acessado em: 16 mai. 2019.

Hadley Wickham, Jim Hester and Winston Chang (2019). *devtools: Tools to Make Developing R Packages Easier*. R package version 2.0.2. https://CRAN.R-project.org/package=devtools.

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.0.1. https://CRAN.R-project.org/package=dplyr.

PET Estatística UFPR (2016). *labestData: Biblioteca de Dados para Ensino de Estatística*. R package version 0.0-17.458.

Pimentel-Gomes, F. (2009). *Curso de Estatística Experimental (15th ed.)*. Piracicaba, SP: FEALQ.

Sarkar, Deepayan (2008) *Lattice: Multivariate Data Visualization with R*. Springer, New York. ISBN 978-0-387-75968-5.

Teste de Bartlett. *Portal Action*. Disponível em http://www.portalaction.com.br/manual-anova/teste-de-bartlett-para-analise-de-resistencia-da-fibra. Acessado em: 16 mai. 2019.

Teste de Shapiro-wilk. *Portal Action*.Disponível em http://www.portalaction.com.br/inferencia/64-teste-de-shapiro-wilk. Acessado em: 16 mai. 2019.