Pratica ANOVA

Caio Graco-Roza

3/2/2021

Calculando uma Análise de Variância ANOVA passo a passo. Primeiro vamos baixar os pacotes que a gente precisa para manipular e visualizar os dados

```
library(tidyverse) # O tidyverse é uma suite de pacotes que permite tanto a manipulação
 quanto a visualização.
library(kableExtra) # Para gerar a tabela final
options(knitr.kable.NA = '')
```

Agora, vamos inventar uma base de dados fictícia. Nós fizemos um experimento para verificar o efeito da intensidade luminosa (Baixa, Media ou Alta) na taxa de crescimento de uma determinada espécie vegetal. O experimento foi com 10 réplicas para cada tratamento.

> Dica: A função `rnorm` gera `n` valores com distrbuição normal de média e desvio padrao (sd) definidos pelo usuário. Assim fica facil de criar exemplos onde sabemos que nossos dados respeitam as premissas da ANOVA.

```
#Criando a base de dados
set.seed(123) #não vou explicar isso hoje, mas é uma maneira de o R sempre gerar os mesm
os numeros como exemplo.
dados_amplo <- data.frame(Baixo = rnorm(15, mean=14, sd=2),
                Medio = rnorm(15, mean=16, sd=2),
                Alto = rnorm(15, mean= 20,sd=2))

summary(dados_amplo) # comando summary mostra um resumo da planilha
```

```
##      Baixo           Medio            Alto      
##  Min.   :11.47   Min.   :12.07   Min.   :17.47  
##  1st Qu.:13.00   1st Qu.:13.91   1st Qu.:19.40  
##  Median :14.22   Median :15.05   Median :20.85  
##  Mean   :14.30   Mean   :15.51   Mean   :20.59  
##  3rd Qu.:14.86   3rd Qu.:17.20   3rd Qu.:21.70  
##  Max.   :17.43   Max.   :19.57   Max.   :24.34
```

Percebam que a nossa base de dados está no formato amplo, mas nós vamos transformá-la para o formato longo porqu esse é o formato preferível para realizar a nossa Análise.

Dica: A função `pivot_longer` converte tabelas do formato amplo para o formato longo. Você precisa informar quais colunas devem ser combinadas e qual o novo nome que a combinação vai receber (`names_to`). Também é importante informar qual o nome da coluna que vai armazenar os valores (`values_to`) que antes pertenciam às colunas que são combinadas. Aqui, vamos chamar a coluna de `Crescimanto`.

```
dados_longo <- dados_amplo %>%
  pivot_longer(cols=c("Baixo","Medio","Alto"),
               names_to = "Luz",
               values_to = "Crescimento") %>%
  mutate_at("Luz",ordered, levels = c("Baixo","Medio", "Alto")) #Reordernar as categoria
s.

dados_longo
```

Luz <ord>	Crescimento <dbl>
Baixo	12.87905
Medio	19.57383
Alto	20.85293
Baixo	13.53965
Medio	16.99570
Alto	19.40986
Baixo	17.11742
Medio	12.06677
Alto	21.79025
Baixo	14.14102

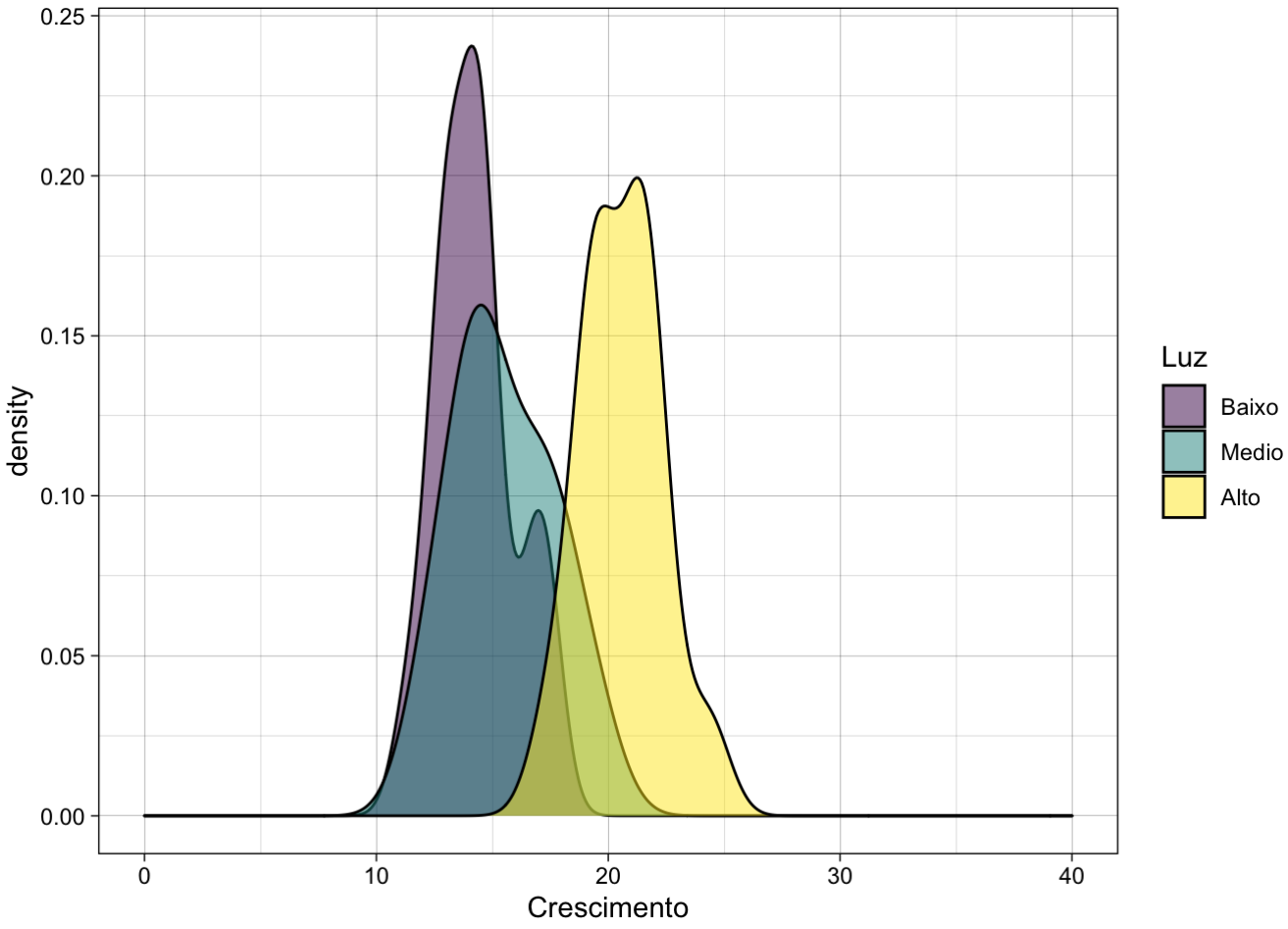
1-10 of 45 rows Previous **1** 2 3 4 5 Next

Antes de começarmos as análises propriamente. É aconselhável plotar os gráficos para ver como os valores estão distribuídos entre os grupos. Para checar distribuição de dado, *histogramas* e graficos de *densidade* são úteis. Vamos usar gráficos de densidade nesse caso:

Dica: Para fazer a visualização gráfica nós usamos o pacote `ggplot2` da suite tidyverse. Esse pacote oferece diferentes temas padrões para alterar a estética do seu gráfico. Aqui, vamos usar o `theme_minimal()`, mas você pode outro tema de seu gosto. Para configurar um tema para as suas imagens use a função `theme_set()`. O site da ggplot2 (https://ggplot2.tidyverse.org/reference/ggtheme.html) tem a lista completa de temas com exemplos.

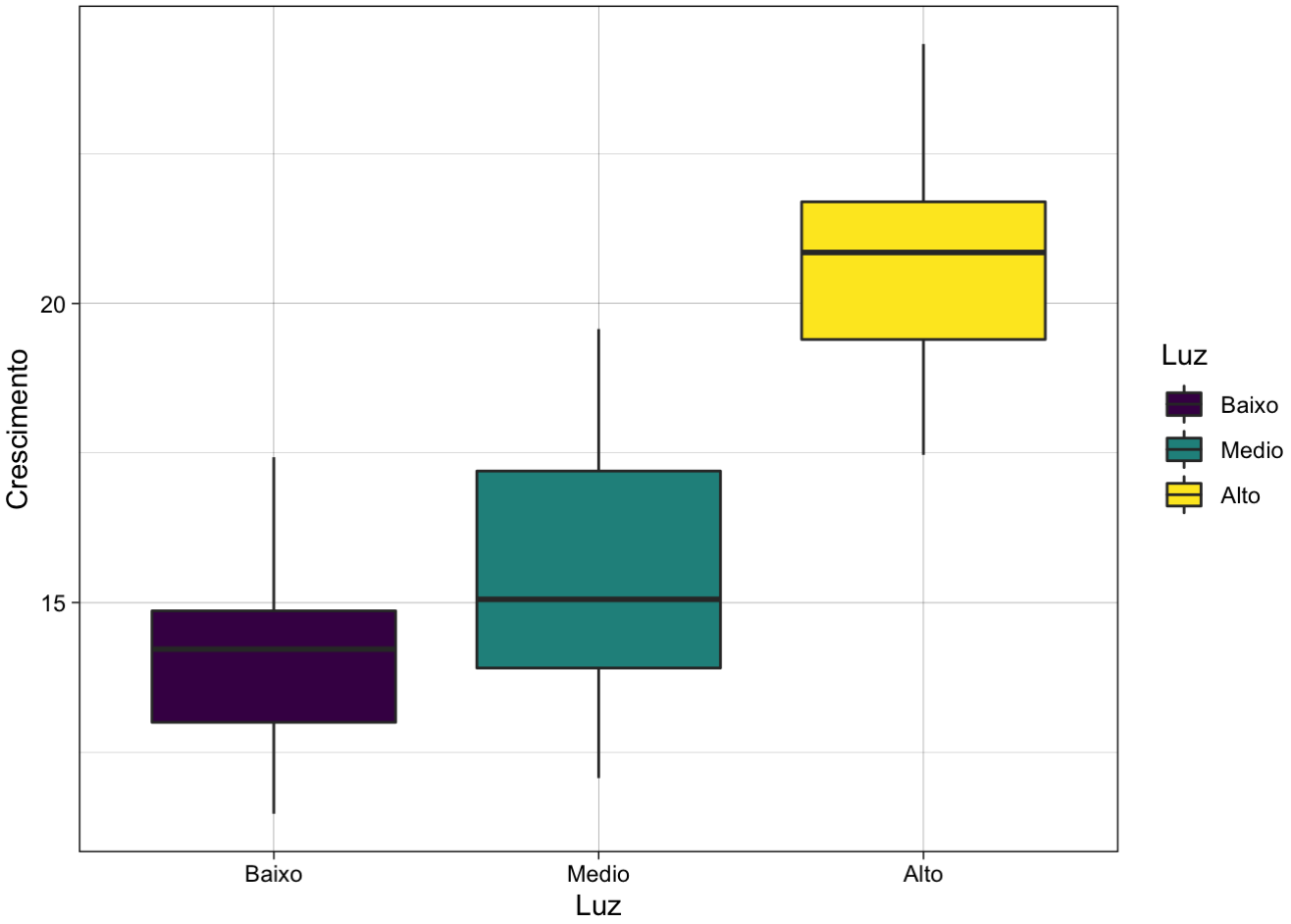
```
theme_set(theme_linedraw())

p1 <- dados_longo %>%
  ggplot(aes(, x = Crescimento, fill=Luz)) +
  geom_density(position="dodge", alpha=.5) +
  xlim(0,40)
p1
```



Por outro lado, nós também estamos interessados em valores não usuais (outliers) e o uso de **boxplots** auxilia na busca por esses valores.

```
p2<- dados_longo %>%
  ggplot(aes(y = Crescimento, x= Luz,  fill=Luz)) +
  geom_boxplot(alpha=1)
p2
```



Vamos ao que interessa de fato. A análise de variância é basea-se na *soma dos quadrados das diferenças* (SQD) **dentro** dos grupos (ERRO) e **entre** os grupos (EFEITO) que estamos comparando. O cálculo da soma dos quadrados **dentro** dos grupos é expresso pela seguinte equação.

$$SQD_{dentro} = \sum_{i=1}^{a} \sum_{j=1}^{n} (Y_j - \overline{Y_i})^2$$

Que corresponde aos seguintes passos no R:

```
SQD_dentro <- dados_longo %>%
  group_by(Luz) %>% #Agrupa os calculos por Tratamento
  mutate(Media_trat = mean(Crescimento)) %>% #Calcula a Media de crescimento em cada gru
po
  mutate(QD = (Crescimento - mean(Media_trat))^2) %>% # quadrado da diferença (QD) entre
valores observados e valores médios por grupo
  ungroup() %>% #Desfaz os grupos para fazer a soma final
  summarise(SQD_dentro = sum(QD)) #Soma dos quadrados das diferenças
SQD_dentro
```

SQD_dentro <dbl>
148.9543

1 row

Isso pode ser simplificado pela seguinte formula

$$SQD_{dentro} = \sum_{i=1}^{n}(n-1)s_i^2$$

Onde s^2 é a variância do grupo. Talvez assim fique mais facil de entender o porquê de o nome ser análise de variância.

```
dados_longo %>% group_by(Luz) %>%
  summarise(sd_trat = var(Crescimento), n = n() -1 ) %>%
  ungroup() %>%
  summarise(SQD_dentro = sum(sd_trat*n))
```

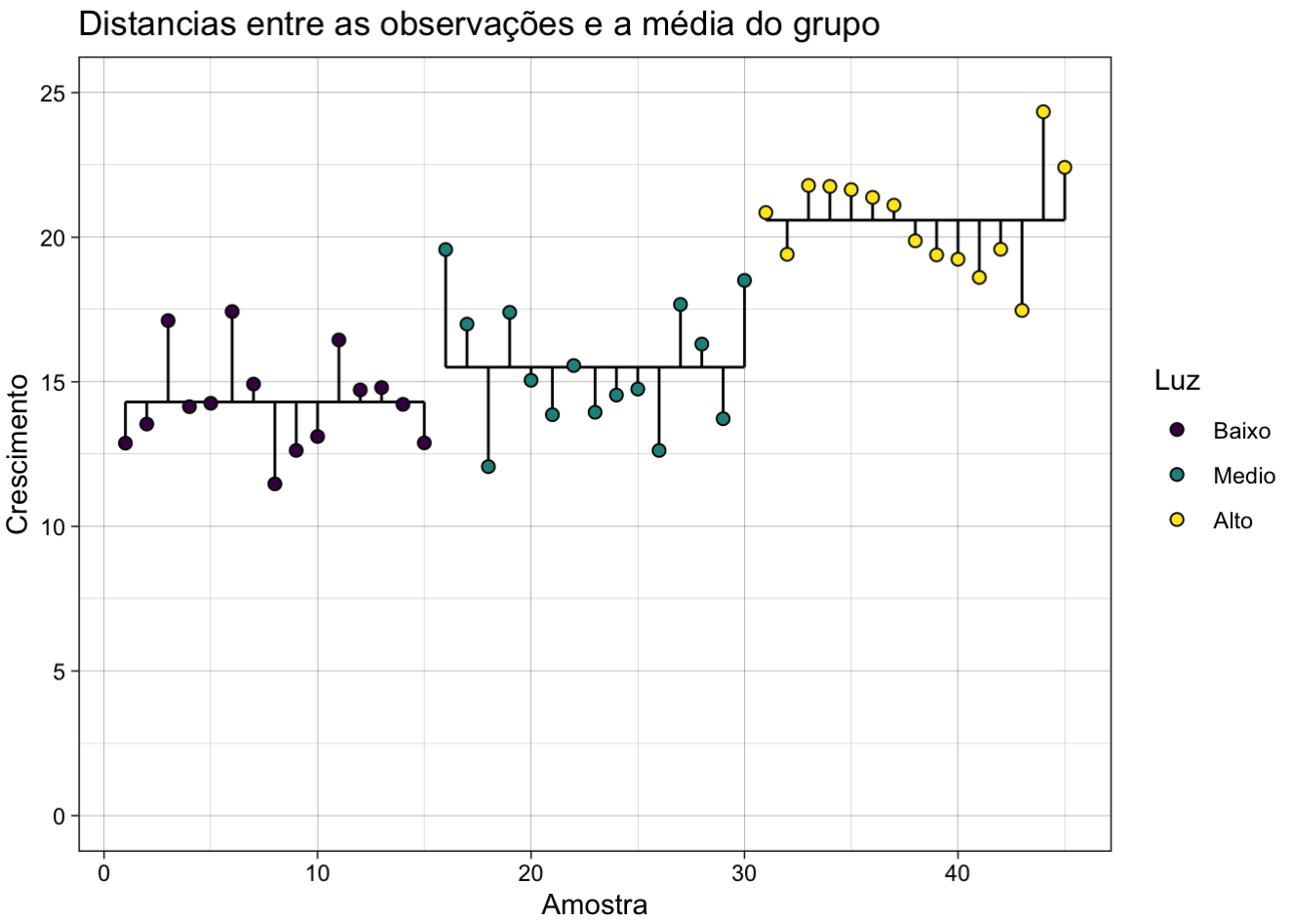
SQD_dentro <dbl>
148.9543

1 row

Basicamente o que estamos fazendo é calcular a distancia entre as nossas observações e a média de cada grupo. Visualizando isso gráficamente fica mais simples.

```
p3<- dados_longo %>%
  arrange(Luz) %>%
  rowid_to_column(var = "Amostra") %>%
  group_by(Luz) %>%
  mutate(Media_trat = mean(Crescimento)) %>%
  ggplot(aes(y=Crescimento, x=Amostra, group=Luz, fill=Luz)) +
  geom_line(aes(y=Media_trat, x=Amostra)) +
  geom_segment(aes(x=Amostra, xend=Amostra, y=Crescimento, yend=Media_trat)) +
  geom_point(shape=21, size=2) +
  ylim(0,25)+
  ggtitle("Distancias entre as observações e a média do grupo")

p3
```



Simplificando, o SQD_{Dentro} representa a soma do tamanho das linhas verticais do gráfico acima. Como tamanho não é uma medida que pode ter valores negativos (Ou você conhece alguém que mede -1.7 metros?), os valores de SQD sempre serão positivos.

Uma vez que temos a soma dos quadrados das diferenças (SQD) dentro dos grupos, vamos calcular *entre* grupos. A equação é um pouco diferente:

$$SQ_{dentro} = \sum_{i=1}^{a} \sum_{j=1}^{n} n_j(\overline{Y_j} - \overline{Y})^2$$

Aqui nós vamos primeiro calcular a média dos grupos e depois subtrair da média **total** do experimento. Esse valor é ponderado pelo numero de observações em cada grupo. Abaixo tem um exemplo de como executar a equação no R.

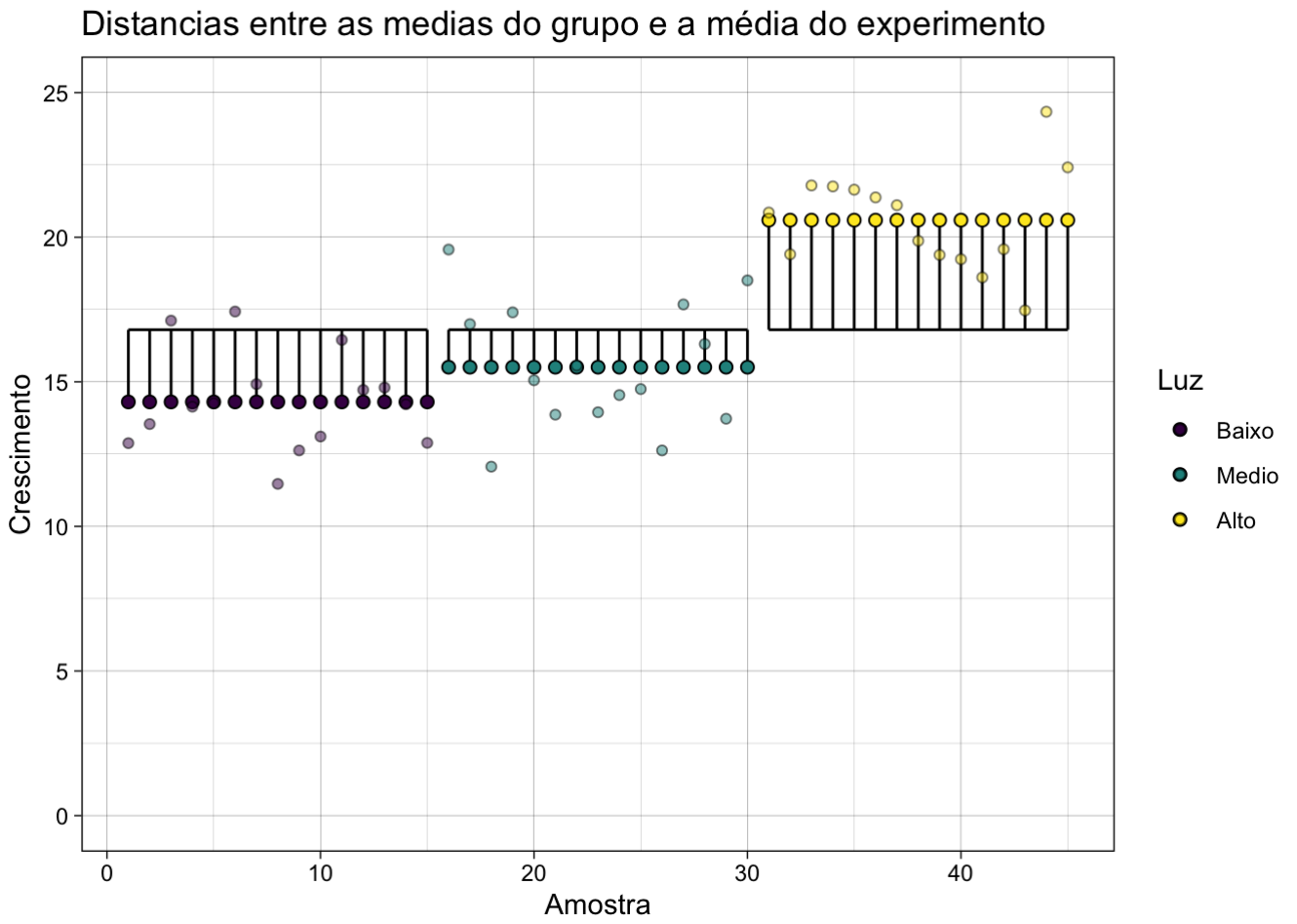
```
SQD_entre<- dados_longo %>%
  group_by(Luz) %>% #Agrupa os calculos por Tratamento
  summarise(Media_trat = mean(Crescimento), n = n()) %>% #Calcula a Media de crescimento e numero de amostras em cada grupo
  summarise(QD = n*(Media_trat - mean(Media_trat))^2) %>% #Essa parte aqui representa toda a equação, com exceção da soma final.
  summarise(SQD_entre = sum(QD)) #Soma dos quadrados da diferença entre grupos
SQD_entre
```

SQD_entre
<dbl>

SQD_entre <dbl>
334.0049

1 row

```
p4 <- dados_longo %>%
  arrange(Luz) %>%
  rowid_to_column(var = "Amostra") %>%
  mutate(Media_total = mean(Crescimento)) %>% #media total sem agrupamento
group_by(Luz) %>%
  mutate(Media_trat = mean(Crescimento)) %>% #Media do grupo, com agrupamento.
  ggplot(aes(
    y = Media_trat,
    x = Amostra,
    group = Luz,
    fill = Luz
  )) +
  geom_line(aes(y = Media_total, x = Amostra)) +
  geom_segment(aes(
    x = Amostra,
    xend = Amostra,
    y = Media_trat,
    yend = Media_total
  )) +
  ylim(0,25)+
  geom_point(shape = 21, size = 2) +
  geom_point(aes(y=Crescimento,x=Amostra, fill=Luz),shape=21, alpha=0.5)+
  ylab("Crescimento")+
  ggtitle("Distancias entre as medias do grupo e a média do experimento")
p4
```



Agora fica tudo mais facil. Uma vez que temos tanto o SQD_{Dentro} e o SQD_{Entre} nós podemos começar a calcular a nossa análise de variância. Mas pra isso vamos estimar os graus de liberdade **dentro** dos tratamentos $GL_{entre} = n_{amostras} - n_{tratamentos}$ e **entre** os tratamentos $GL_{dentro} = (n_{tratamentos} - 1)$.

```
GL_dentro <-  45 - 3
GL_entre <- 3 - 1
```

Com isso podemos calcular o nosso Quadrado médio **entre** e **dentro** dos tratamentos. Os quadrados médios são a razão entre o SQD e os graus de liberdade dentro e entre tratamentos.

```
QM_dentro <- as.numeric(SQD_dentro/GL_dentro)
QM_entre <-  as.numeric(SQD_entre/GL_entre )
```

Para estimar a significancia das nossas differenças, nós utilizamos a estatística F (não vamos ir a fundo nisso agora). O valor de F pode ser obtido como a razão entre o QM_{Entre} e o QM_{Dentro}. Em outras palavras, nós estamos verificando se as diferenças observadas entre os tratamentos é maior do que a variação medida dentro de cada grupo. Quanto maior essa razão maior vai ser a significancia da nossa análise.

No R:

```
Valor_F <- as.numeric(QM_entre/QM_dentro)
Valor_F
```

```
## [1] 47.08894
```

Seguindo a lógica acima, valores de F menores que 1 indicam que a variação dentro dos grupos é maior do que a variação gerada pelos nossos tratamentos e por isso nosso teste não vai ser significativo. Entretanto, a maneira correta de fazer essa ponderação é estimando o P baseando-se na posição do nosso valor de F na distribuição de Fisher. Isso nós vamos fazer usando uma função do R, uma vez que isso não faz parte do escopo dessa aula.

```
valor_P <- pf(Valor_F, GL_entre, GL_dentro, lower.tail = FALSE)
```

Vamos gerar agora uma tabela com os resultados calculados por nós.

```
tribble(~"Fonte de variação", ~"GL", ~"SQD", ~"QM", ~ "F", ~"p",
         "Entre tratamentos",GL_entre, round(SQD_entre,2), round(QM_entre,2), round(Valo
r_F,2), round(valor_P,11),
         "Dentro dos tratamentos", GL_dentro,round(SQD_dentro,2),round(QM_dentro,2), NA,
NA) %>%

  kbl(caption = "Resultados da Análise de Variância", digits=16) %>%
  kable_classic(full_width = F, html_font = "Cambria")
```

Resultados da Análise de Variância

Fonte de variação	GL	SQD	QM	F	p
Entre tratamentos	2	334	167.00	47.09	2e-11
Dentro dos tratamentos	42	148.95	3.55		

Agora vamos comparar nossos resultados com o de uma ANOVA usando a função `aov()` nativa do R.

```
summary(aov(Crescimento~Luz, data=dados_longo))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Luz          2    334  167.00   47.09 1.87e-11 ***
## Residuals   42    149    3.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```