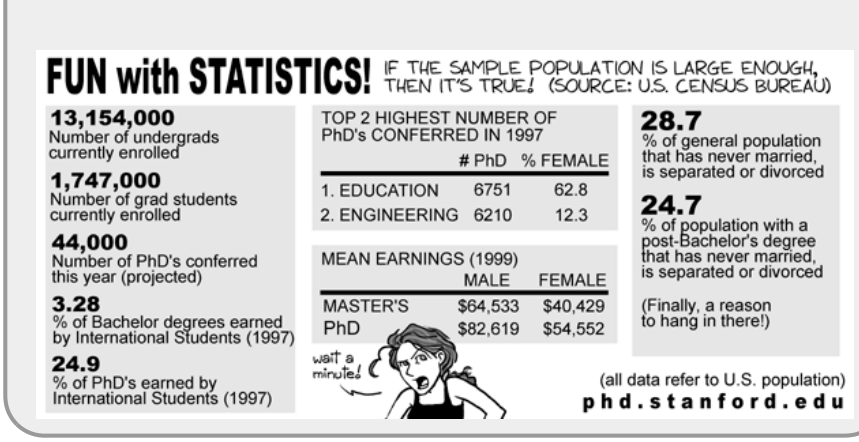
- Tutorial
- Exercícios
- Apostila

6b. Partição da Variação dos Dados

Video gravado pelo Google Meet em aula síncrona no dia 30 de setembro de 2020. Sem edição.

Video



O teste t, apresentado no tutorial 6a, é usado apenas para o caso de termos uma variável resposta numérica contínua e uma preditora categórica com **dois níveis**. Caso a preditora tenha mais do que dois níveis, precisamos usar um outro teste que é uma generalização do teste t, o teste de **Análise de Variância** ou **ANOVA**. O teste está baseado no princípio de partição da variação dos dados. A

variação total dos dados é particionada nos componentes do que é explicado e aquele que não é explicado pela variável preditora categórica. Esse conceito é aplicado de maneira mais ampla na estatística, utilizado em outros tipos de estatística e para a tomada de decisão do modelo que melhor explica a variação nos dados. Por isso, vamos focar este tutorial no **conceito da partição da variação**.

Para exemplificar a partição da variância associada à ANOVA, vamos usar o exemplo de dados de colheita de um cultivar em diferentes tipos de solos, apresentado no livro de Robert Crawley, The R Book, como segue abaixo:

Tradução livre da descrição do livro "*The R Book*" (Crawley, 2007)

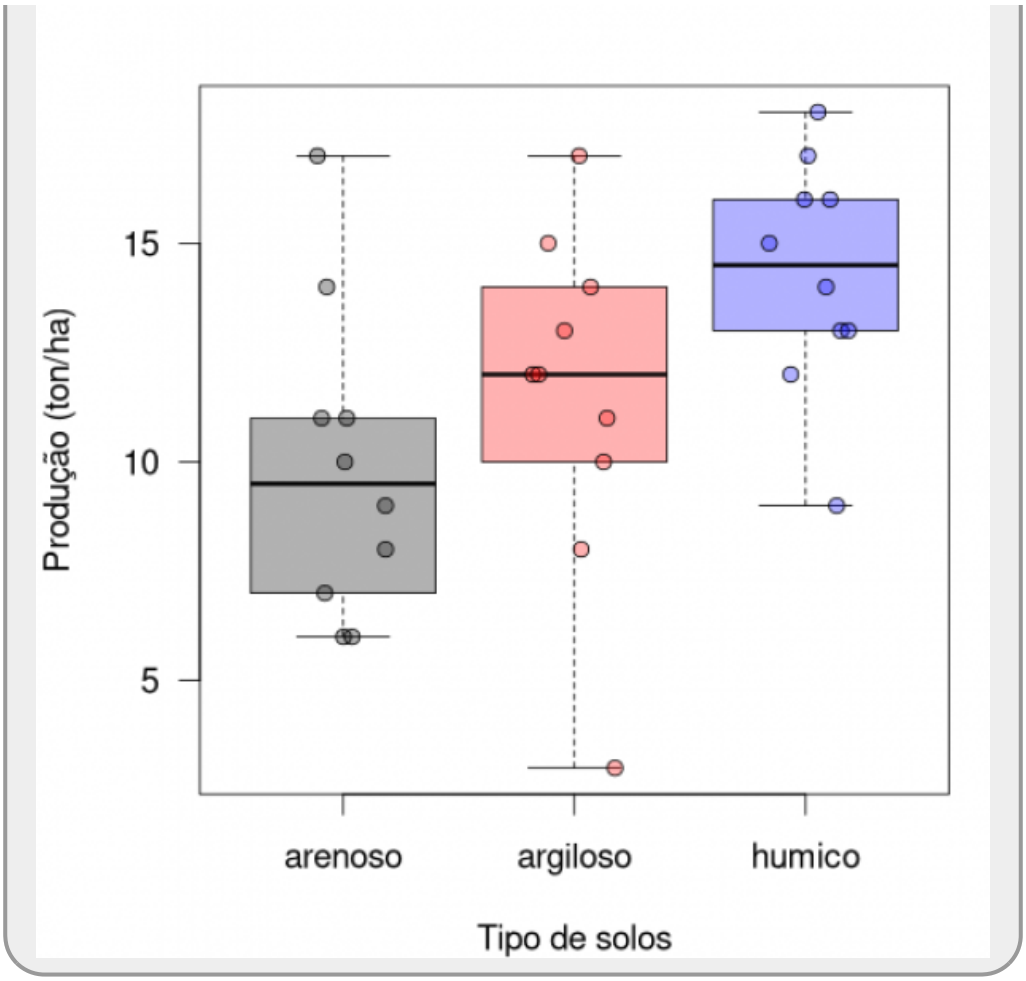
"... a melhor forma de entender o que está acontecendo é trabalharmos um exemplo. Temos um experimento em que a produção agrícola, por unidade de área, é medida em 10 campos de cultivo, selecionados aleatoriamente em cada um de três tipos diferentes de solo. Todos os campos foram semeados com a mesma variedade de semente e manejados com as mesmas técnicas (fertilizantes, controle de pragas). O objetivo é verificar se o tipo de solo afeta significativamente o rendimento de culturas, e caso afete, quanto."[1)]

Vamos organizar os dados apresentados no livro em um objeto no R diretamente:

```
are <- c(6,10,8,6,14,17, 9, 11, 7, 11)
arg <-  c(17, 15, 3, 11, 14, 12, 12, 8, 10, 13)
hum <- c(13, 16, 9, 12, 15, 16, 17, 13, 18, 14)
solo <- rep(c("arenoso", "argiloso", "humico"), each = 10)
cultivar <- data.frame(producao = c(are, arg, hum), solo = solo)
str(cultivar)
```

Primeiro vamos fazer uma inspeção nos dados na forma de um `boxplot`:

```
cols <- c(rgb(0, 0, 0, 0.1), rgb(1, 0,0, 0.3), rgb(0,0,1, 0.3))
par(mar = c(4,4,2,1), las = 1, cex = 1.5)
boxplot(producao ~ solo, data = cultivar, col = cols, xlab = "Tipo de
solos", ylab = "Produção (ton/ha)", range = 0)
points(x = jitter(rep(1:3, each = 10)), y = cultivar$producao, bg =
rep(cols, each = 10), pch = 21)
```



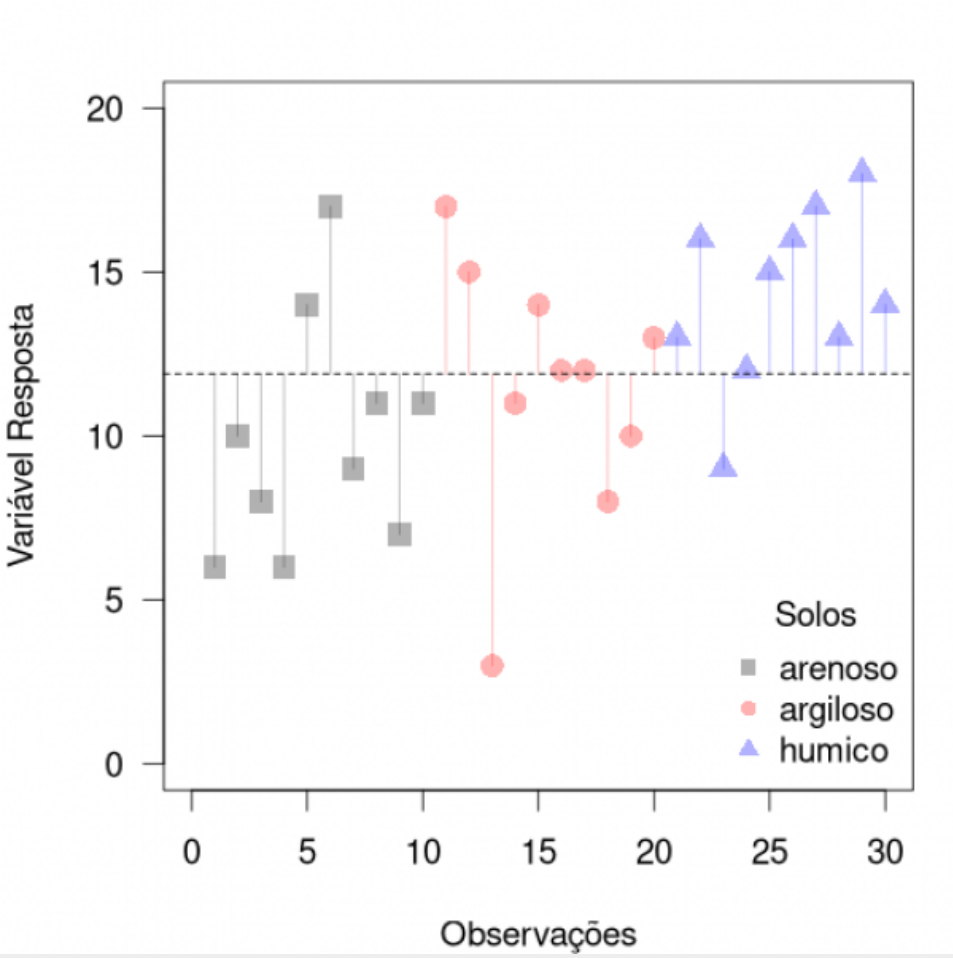
A pergunta aqui é: **Há variação na produtividade média entre solos?**

Variação Total dos Dados

A partição da variação inicia-se pelo reconhecimento e cálculo da variação total associada aos dados. A variação total dos dados é baseada nos desvios das observações em relação à grande média, que no caso, é a média de todos os campos de cultivo. Vamos representar esses desvios:

```
par(mar = c(4,4,2,1), las = 1, cex = 1.5)
colvector <- rep(cols, each= 10)
plot(x = 1:30, y = cultivar$producao , ylim = c(0,20), xlim = c(0, 30),
pch=(rep(c(15,16,17), each=10)), col = colvector, ylab = "Variável
Resposta", xlab = "Observações", cex = 1.5)
    for(i in 1:30)
    {
    lines(c(i,i),c(cultivar$producao[i], mean(cultivar$producao)), col =
colvector[i])
    }
abline(h = mean(cultivar$producao), lty = 2)
legend("bottomright", legend = c("arenoso", "argiloso", "humico"), pch =
15:17 ,col = cols, title = "Solos", bty = "n")
```

Representação dos desvios dos dados: em relação à média geral.



No gráfico esta variação é representada pelos segmentos verticais coloridos. A grande média é definida como a média de produtividade de todos os campos de cultivo (n=30), independente do tipo de solo, e é representada pela linha preta horizontal tracejada.

Medimos essa variação total pela `soma quadrática`: os valores dos desvios dos dados em relação à grande média (segmentos verticais no gráfico) elevados ao quadrado e posteriormente somados. Essa soma quadrática total é nossa medida de variação.

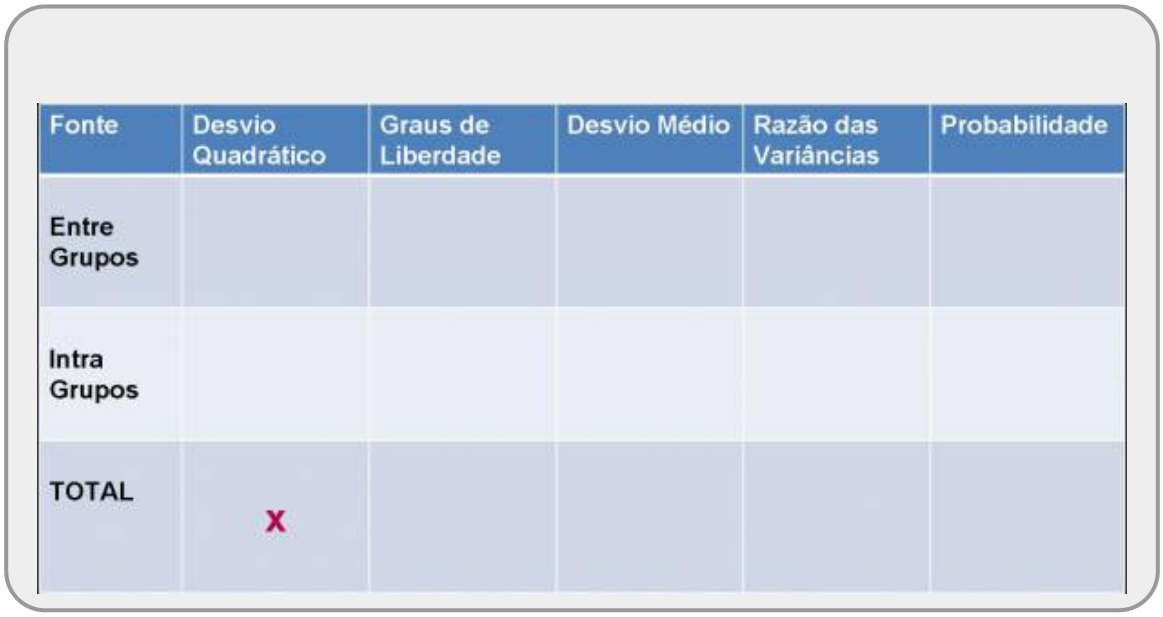
$$ SQ_{"total"} = \sum_{i=1}^k\sum_{j=1}^n (y_{ij} - \bar{\bar{y}})^2 $$

Aqui calculamos o valor de somatória quadrática que é a base da análise de ANOVA.

```
(mGeral <- mean(cultivar$producao))
(dGeral <- cultivar$producao - mGeral)
(dqGeral <- dGeral^2)
(sqGeral <- dqGeral)
(sdqGeral <- sum(sqGeral))
```

Fizemos acima todos os passos isoladamente, pois, alguns desse valores intermediários iremos utilizar mais à frente.

Vamos iniciar a construção da nossa tabela de ANOVA, incluindo a medida de variação total na sua posição:



Fonte	Desvio Quadrático	Graus de Liberdade	Desvio Médio	Razão das Variâncias	Probabilidade
Entre Grupos					
Intra Grupos					
TOTAL	X				

Particionando a Variação

Um ponto importante da soma quadrática é que esta variação é aditiva e pode ser decomposta. Uma parte dessa variação é explicada pelo tratamento, que são nossas variáveis preditoras [2)]. A porção não explicada pelas variáveis preditoras é o **resíduo**, algumas vezes chamado de *erro*. A porção não explicada está associada à variação aleatória dos dados ou a algum, ou vários fatores que não foram incluídos ou controlados no nosso experimento.

Variação Não Explicada

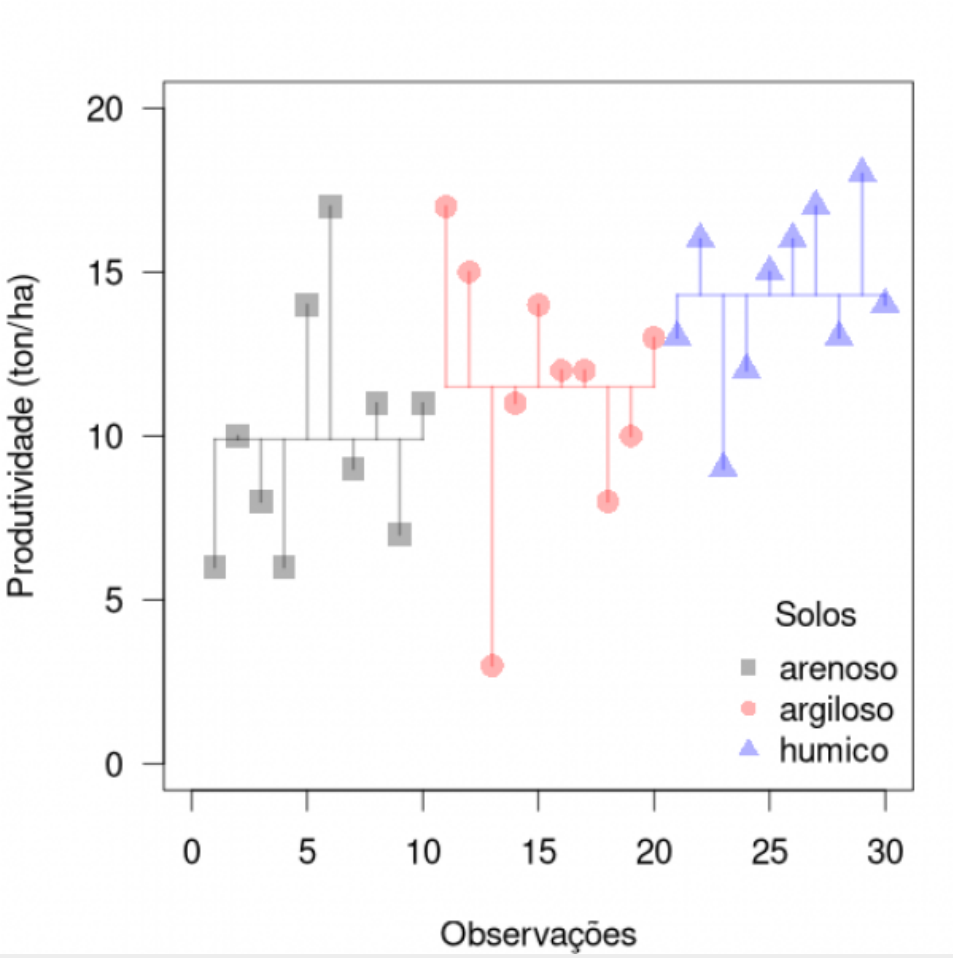
Vamos primeiro representar essa variação não explicada, ou variação interna aos níveis da variável categórica:

```
(mSolos <- tapply(cultivar$producao, cultivar$solo, mean))
mSolosVetor <- rep(mSolos, each = 10)
par(mar = c(4,4,2,1), las = 1, cex = 1.5)
plot(x = 1:30, y = cultivar$producao , ylim = c(0,20), xlim = c(0, 30),
pch=(rep(c(15,16,17), each=10)), col = colvector, ylab = "Produtividade
(ton/ha)", xlab = "Observações", cex = 1.5)
segments(x0 = 1:30, y0 = cultivar$producao, y1= mSolosVetor, col =
colvector, lwd = 2)
segments(x0 = c(1,11, 21), y0 = mSolos, x1 = c(10, 20, 30), col = cols, lwd
=1.5)
```

```
legend("bottomright", legend = c("arenoso", "argiloso", "humico"), pch =
15:17 ,col = cols, title = "Solos", bty = "n")
```

No código acima utilizamos a função `segments` e não foi necessário utilizar a iteração. Além disso, criamos um vetor de médias com a repetição das médias dos respectivos solos para cada observação no `mSolosVetor`, para facilitar a construção do gráfico.

Representação dos resíduos dos modelo: desvios dos dados em relação às médias dos tratamentos.



Para calcular a variação não explicada precisamos usar o mesmo procedimento da variação total: elevar os resíduos ao quadrado e somar, resultando também em uma somatória quadrática.

```
(sqIntra <- sum((cultivar$producao - mSolosVetor)^2))
```

Incluindo esse valor na nossa tabela:

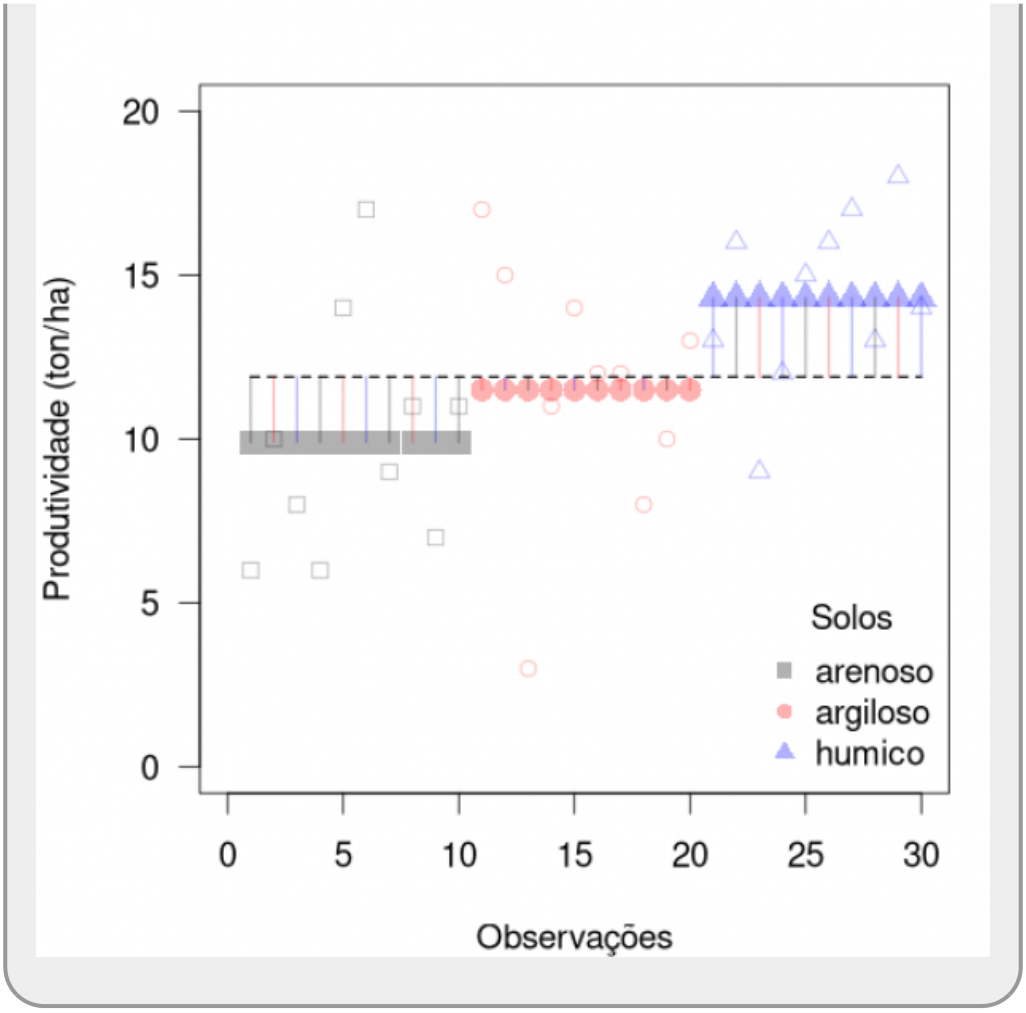
Fonte	Desvio Quadrático	Graus de Liberdade	Desvio Médio	Razão das Variâncias	Probabilidade
Entre Grupos	X				
Intra Grupos	315.5				
TOTAL	414.7				

Variação Explicada

A variação explicada pelas variáveis preditoras é a diferença entre a variação total e a não explicada, devido à característica aditiva das somatórias quadráticas. Vamos representar a variação que foi explicada, ou variação entre os níveis da variável preditora categórica, em um gráfico:

```
par(mar = c(4,4,2,1), las = 1, cex = 1.5)
plot(x = 1:30, y = cultivar$producao , ylim = c(0,20), xlim = c(0, 30),
pch=(rep(c(0, 1 ,2), each=10)), col = colvector, ylab = "Produtividade
(ton/ha)", xlab = "Observações", cex = 1)
points(x = 1:30, y = mSolosVetor, pch = rep(c(15,16,17), each=10),  col =
colvector, cex = 1.5)
segments(x0 = 1, y0 = mGeral, x1= 30, col = 1, lty = 2, lwd = 1.5)
segments(x0 = 1:30, y0 = mSolosVetor, y1 = rep(mGeral, 30), col = cols, lwd
=1.5)
legend("bottomright", legend = c("arenoso", "argiloso", "humico"), pch =
15:17 ,col = cols, title = "Solos", bty = "n")
```

Representação da variação explicada dos dados



Vamos calcular e incluir na nossa tabela essa variação, calculada pelo soma quadrática dos segmentos representados na figura:

```
(sqEntre <- sum((mSolosVetor - mGeral)^2))
```

Incluindo esse valor na nossa tabela:

Fonte	Desvio Quadrático	Graus de Liberdade	Desvio Médio	Razão das Variâncias	P-valor
Entre Grupos	99.2				
Intra Grupos	315.5				
TOTAL	414.7				

Graus de Liberdade

Precisamos agora calcular os **desvios quadráticos médios** que são as somas quadráticas dividido pelos graus de liberdade (`gl`). Vamos utilizar a soma quadrática total para exemplificar o cálculo dos graus de liberdade: para calcular essa soma utilizamos os `30` valores de produtividade e calculamos a média geral. No caso, por termos calculado a média, perdemos um grau de liberdade e ficamos com `29 gl`. Utilizando a mesma lógica para a soma quadrática não explicada, partimos das mesmas `30` informações e calculamos as 3 médias de produtividade dos solos, portanto, ficamos com `27 gl`. Por fim, na soma quadrática explicada temos 3 informações, as médias de cada solo, e calculamos uma estatística, a média geral, ficando com 2 graus de liberdade. Com esses informação podemos então, calcular esses desvios quadráticos médios:

```
(sq <- c(sdqGeral, sqIntra, sqEntre))
glib <- c(29, 27, 2)
(msq <- sq/glib)
```

Agora nossa tabela está quase completa, se calcularmos a razão entre os desvios médios explicado pelo não explicado, temos o cálculo da estatística da ANOVA, o F-Fisher:

```
(fcultiva <-  msq[3]/msq[2])
```

Fonte	Desvio Quadrático	Graus de Liberdade	Desvio Médio	Razão das Variâncias	Probabilidade
Entre Grupos	99.2	2	49.6	4.24	XX
Intra Grupos	315.5	27	11.7		
TOTAL	414.7	29			

Só falta agora o cálculo do p-valor associado à estatística F. O F-Fisher é uma distribuição probabilística que tem dois parâmetros: os graus de liberdade dos cálculos da (1) variação média entre e (2) intra.

```
pf(q = fcultiva, df1 = 2, df2 = 27, lower.tail = FALSE)
```

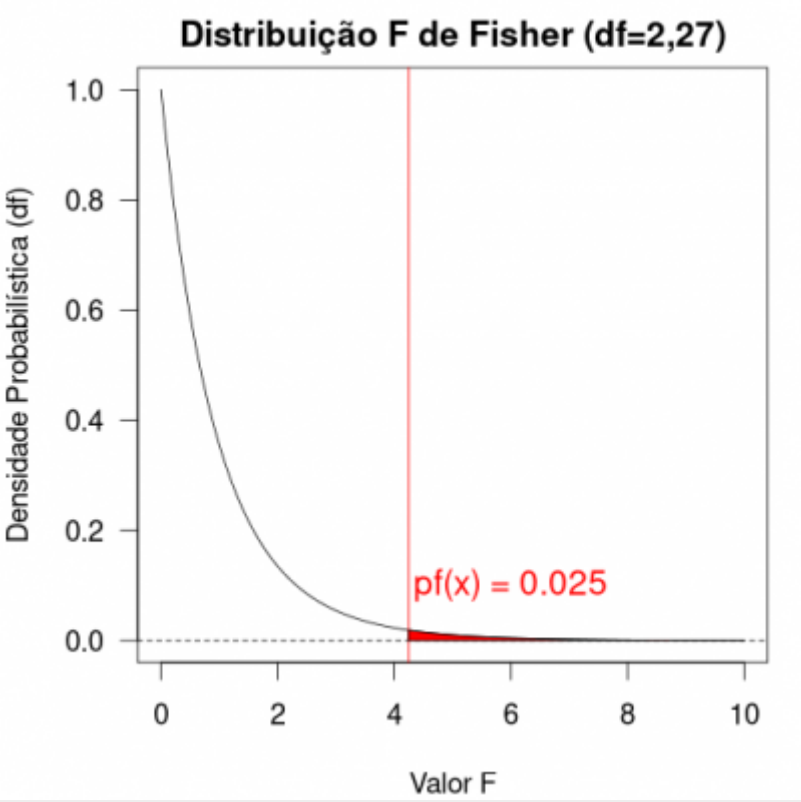
Pronto! Nosso almejado **p-valor** e tabela completa!

Fonte	Desvio Quadrático	Graus de Liberdade	Desvio Médio	Razão das Variâncias (F)	P-valor
Entre Grupos	99.2	2	49.6	4.24	0.025
Intra Grupos	315.5	27	11.7		
TOTAL	414.7	29			

Distribuição de F

Os gráficos de outras aulas apresentaram a distribuição de densidade probabilística, onde a variável y é relacionada à probabilidade de cada valor em intervalos muito pequenos. O valor da probabilidade cumulativa é a área da curva até o valor fornecido, o que é retornado pela função `pf`. No caso, como utilizamos o argumento `lower.tail = FALSE`, a função retorna a outra área da curva, representada pela figura a seguir:

```
curve(expr=df(x, 2,27), main="Distribuição F de Fisher
(df=2,27)", xlab="Valor FALSE",ylab="Densidade
Probabilística (df)", xlim=c(0,10))
abline(v = fcultiva, col="red")
abline(h = 0, lty = 2)
xf <- seq (fcultiva, 10, 0.01)
ydf <- df(xf, 2, 27)
polygon(c(fcultiva, xf), c(0, ydf), col="red")
text(6, 0.1,paste("pf(x)
=",round(pf(fcultiva,2,27,lower.tail=F),4)), cex = 1.2,
col="red")
```



Anova no R

A tabela de anova é tão importante que a função anova no R retorna a tabela para objetos de modelos. Vamos ver isso no próximo tópico. Para o teste da

análise de variância, propriamente dito, a função utilizada é `aov`. No caso do objeto produzido pela função `aov` a tabela aparece se aplicarmos a função `anova` ou `summary` ao objeto `aov`, que também é um objeto da classe modelo linear `lm`.

```
(cultAov <- aov(producao ~ solo, data = cultivar))
class(cultAov)
anova(cultAov)
summary(cultAov)
```

1)

The best way to see what is happening is to work through a simple example. We have an experiment in which crop yields per unit area were measured from 10 randomly selected fields on each of three soil types. All fields were sown with the same variety of seed and provided with the same fertilizer and pest control inputs. The question is whether soil type significantly affects crop yield, and if so, to what extent.

2)

podemos ter mais de uma