



Universidade Federal da Bahia  
Instituto Multidisciplinar em Saúde  
Curso de extensão: Introdução ao uso do programa R  
Prof. Leandro Martins de Freitas e Profª Leila Cruz

## Introdução ao uso do programa R

### DIA 3

#### Gráficos - Dados quantitativos

Os gráficos no R podem ser feitos por diversas funções e alguns pacotes trazem suas próprias funções para fazer gráficos. Essas funções compartilham alguns parâmetros. Apresentamos parâmetros que são compartilhados pela maioria das funções que fazem gráficos.

Características compartilhadas por diversos gráficos do R.

**main:** título do gráfico

**xlab:** rótulo do eixo x

**ylab:** rótulo do eixo y

**xlim:** limite inicial e final do eixo x, deve ser apresentado como vetor com o valor mínimo e máximo, exemplo `c(0,10)`

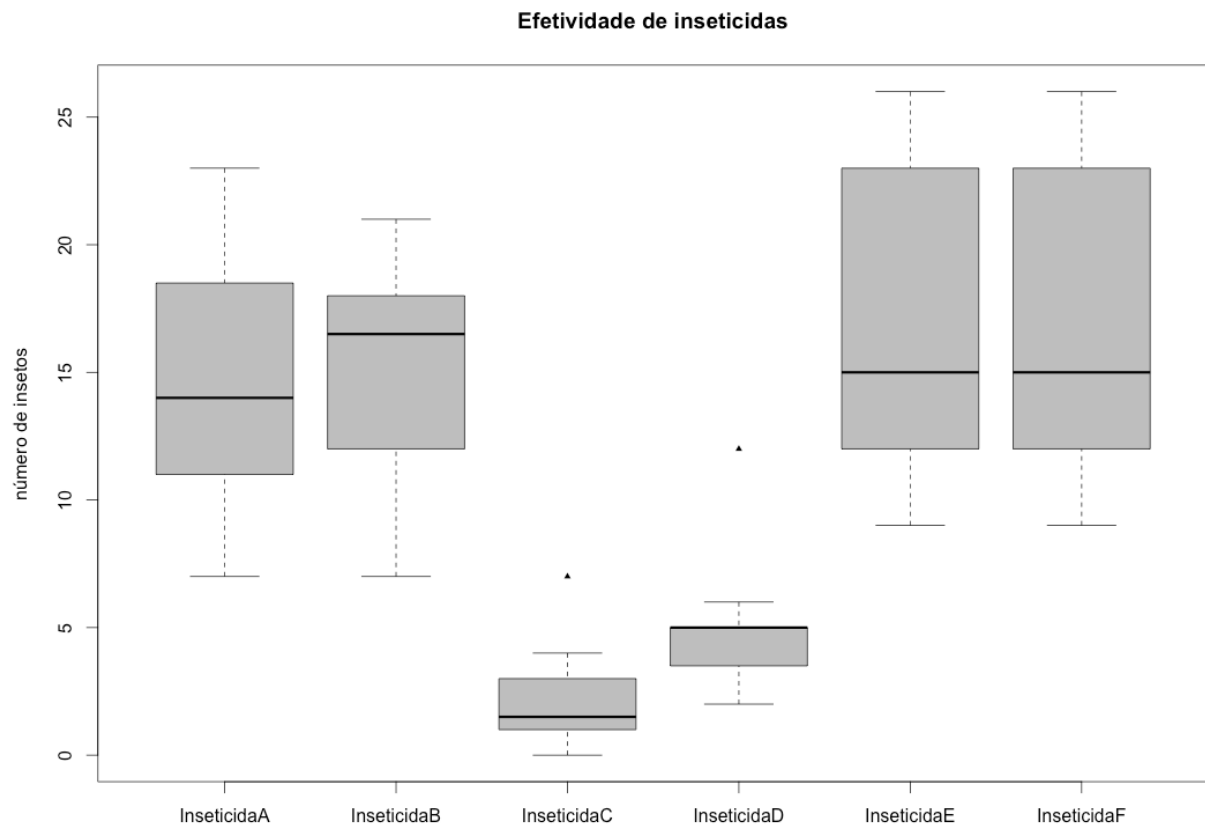
**ylim:** limite inicial e final do eixo y, deve ser apresentado como vetor com o valor mínimo e máximo, exemplo `c(-10,-1)`

**col:** cores que devem ser usadas, pode ser um unico valor tipo character ou um vetor com o nome das cores, exemplo `c("red", "blue", "gray")`

#### Boxplot

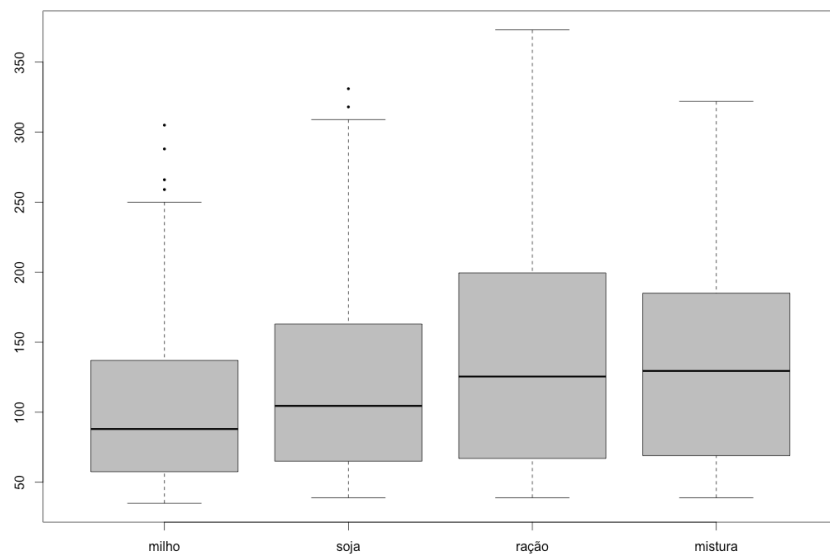
Representação gráfica da dispersão dos dados. O boxplot apresenta os valores mínimo, primeiro quartil, mediana, terceiro quartil, e máximo.

```
#pch: tipo de ponto que será usado
#cex: tamanho dos pontos
#boxplot inseticida
#carregar data set inseto.csv
>spray_A <- inseto$count[inseto$spray == "A"]
>spray_B <- inseto$count[inseto$spray == "B"]
>spray_C <- inseto$count[inseto$spray == "C"]
>spray_D <- inseto$count[inseto$spray == "D"]
>spray_E <- inseto$count[inseto$spray == "E"]
>spray_F <- inseto$count[inseto$spray == "F"]
>boxplot(spray_A,spray_B,spray_C,spray_D,spray_E,spray_F,
         col= "gray", pch=17, cex=0.6,ylab="numero de insetos",
         names = c("InseticidaA","InseticidaB","InseticidaC",
                   "InseticidaD","InseticidaE","InseticidaF"),
         main="Efetividade de inseticidas")
```



Box blot apresentando a variação da efetividade de seis tipos diferentes de inseticidas em spray.

```
#nome das dietas
#carregar arquivo peso.csv no data frame peso
>boxplot(peso$weight ~ peso$Diet,
          col = "gray", pch = 16, cex=0.5,
          names=c("milho","soja","ração","mistura"),main="Exemplo de
título")
```



Box blot apresentando a variação do peso de galinhas que foram submetidas a 4 tipos diferentes de alimentação.

### Exercício

**Carregar gas\_carbonico.csv no dataframe gas\_carbonico**

**Carregar qualidade\_ar.csv no dataframe qualidade\_ar**

Fazer box plot para os seguintes data set

gas\_carbonico, variável uptake separado nos grupos Type

gas\_carbonico, variável uptake separado nos grupos Treatment

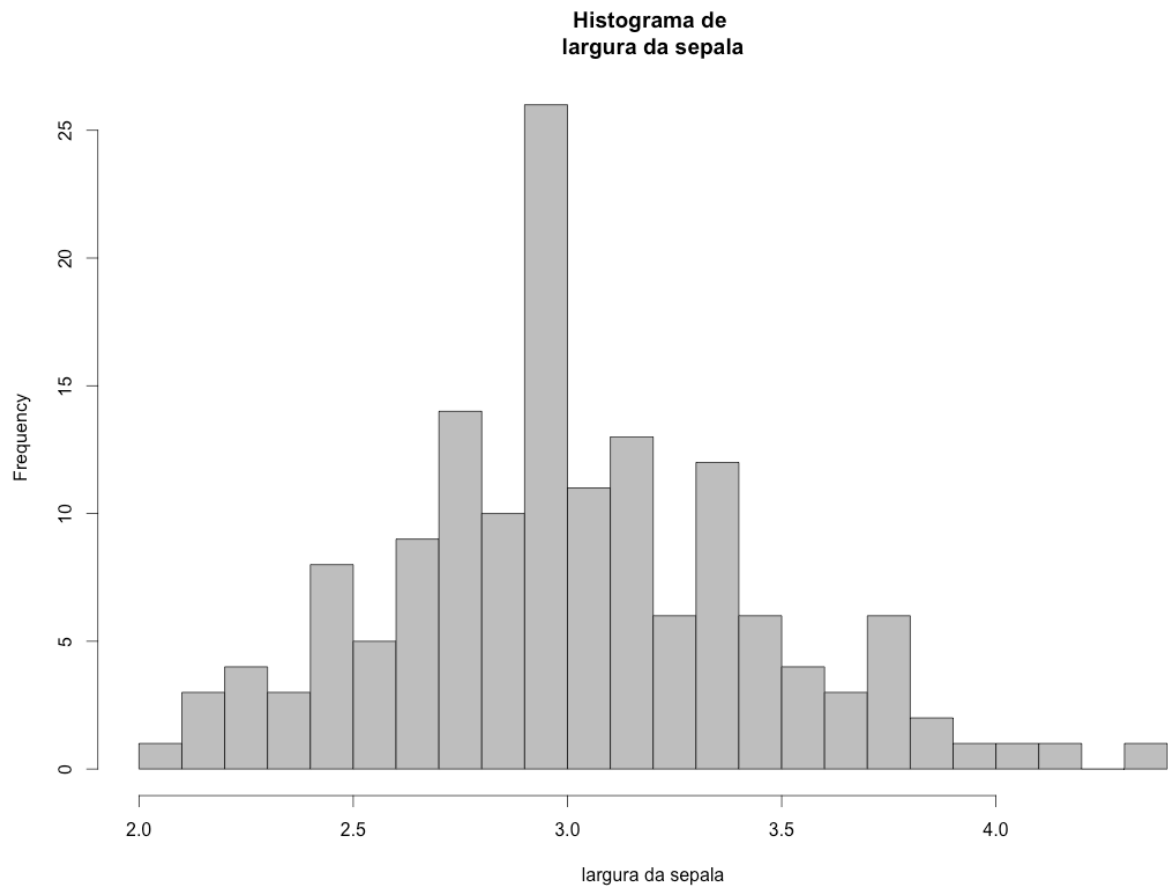
gas\_carbonico, variável uptake separado nos grupos conc

qualidade\_ar, variável Ozone separado nos grupos Month

qualidade\_ar, variável Solar.R separado nos grupos Month

### Histograma

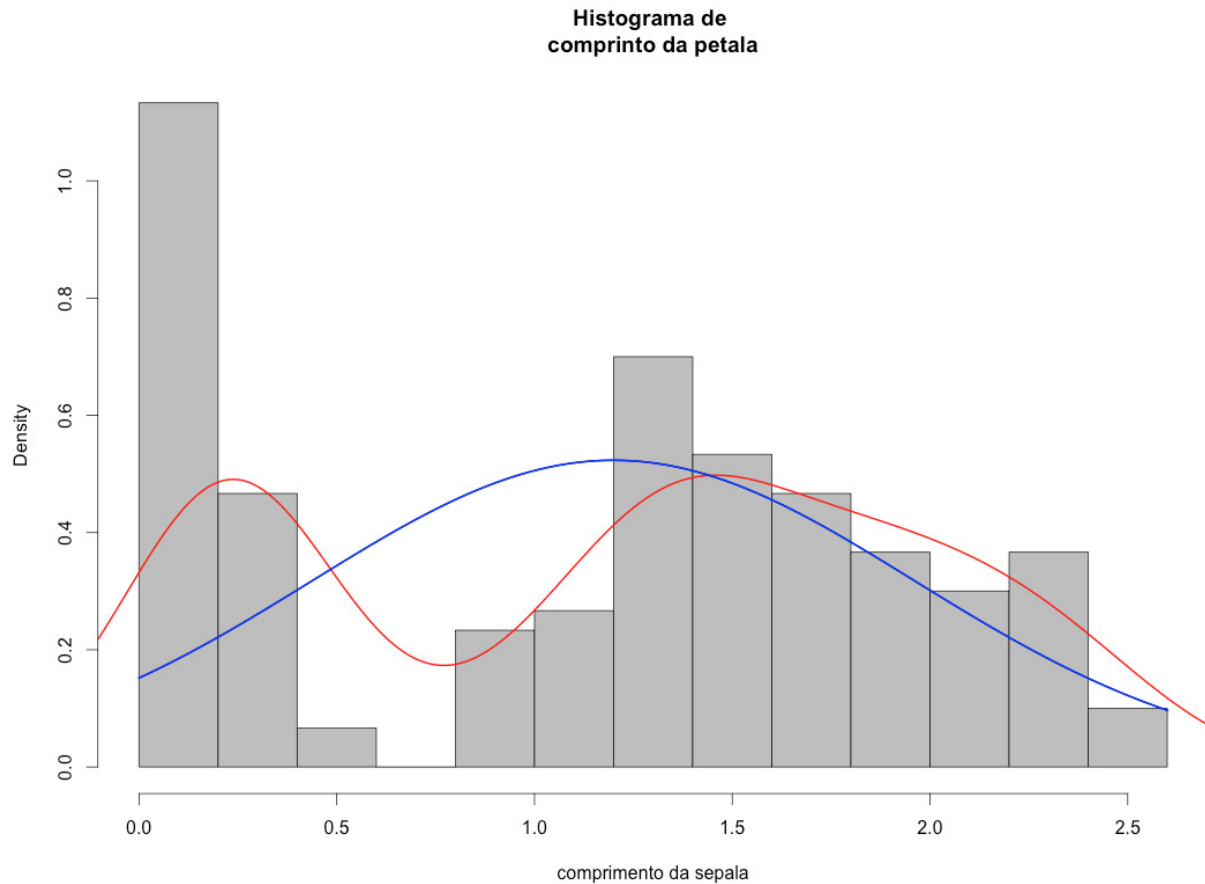
```
>hist(flor$Sepal.Width, breaks = 24, col="gray", main = "Histograma de
\nlargura da sepala")
```



```
>media_Petal_Width <- mean(flor$Petal.Width)
>sd_Petal_Width <- sd(flor$Petal.Width)

>hist(flor$Petal.Width, freq=FALSE,
      xlab="comprimento da sepala",
      breaks = 12, col="grey",
      main = "Histograma de \ncomprimento da petala")

>lines(density(flor$Petal.Width), col="red",lwd=2)
>curve(dnorm(x, mean=media_Petal_Width,
            sd=sd_Petal_Width), add=TRUE,
      col="blue", lwd=2)
```



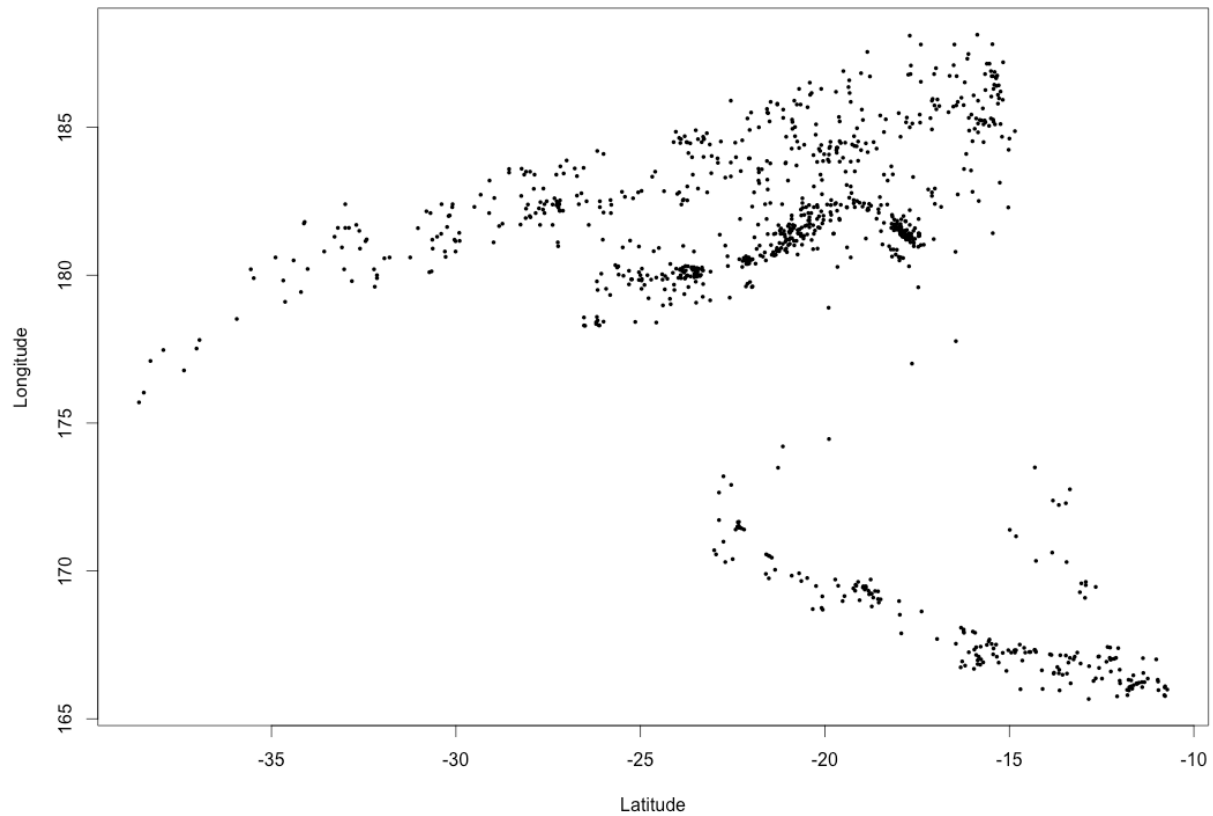
Histograma do comprimento da pétala nas espécies *Iris setosa*, *I. versicolor*, e *I. virginica*. Linha vermelha é curva de densidade. Curva azul, distribuição normal esperada com a média e variação da distribuição do comprimento da pétala.

## Exercício - Histograma

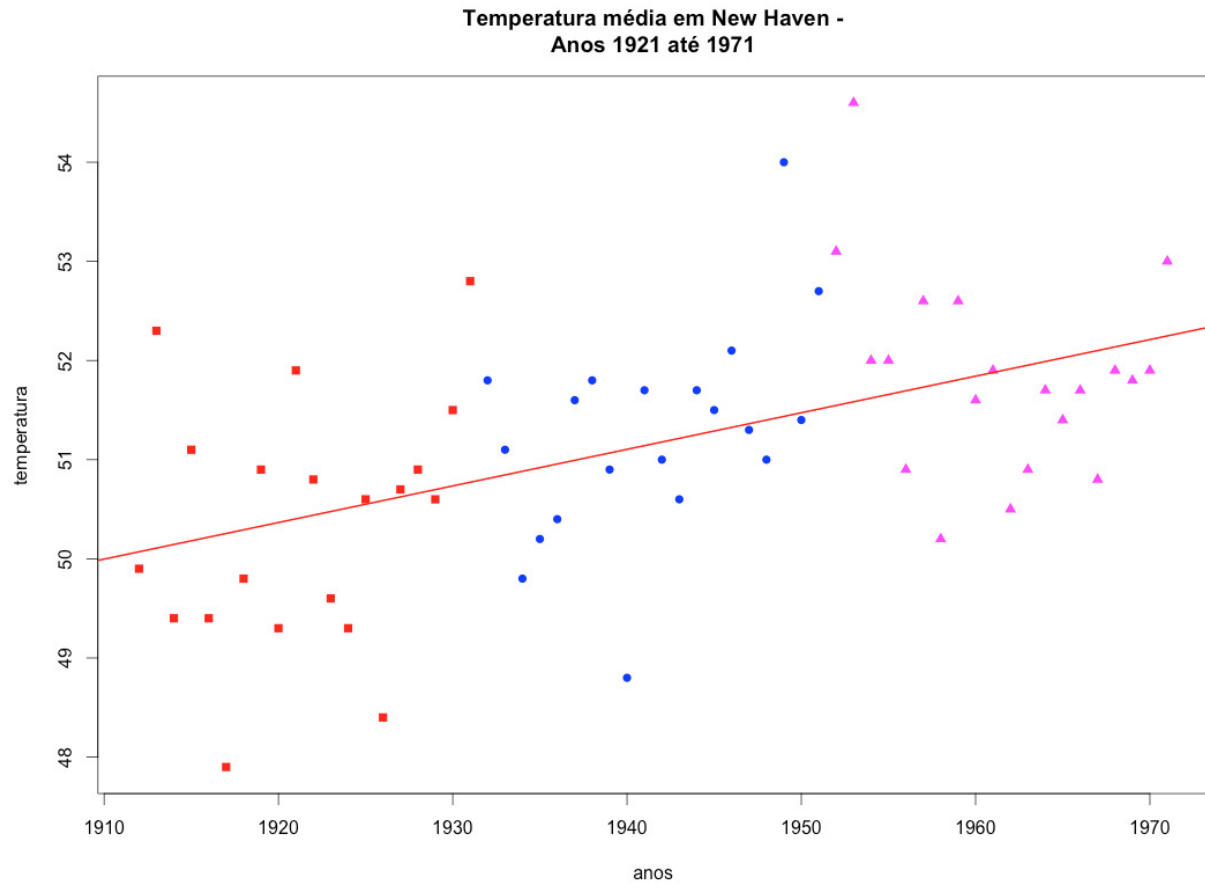
## Dispersão - Linhas

```
#carregar data set terremoto no data frame terremoto
>plot(terremoto$lat, terremoto$long, pch=16, cex=0.5, xlab="Latitude",
ylab="Longitude", main="Terremotos em Fiji, N = 1000, Lat vs Long")
```

Fiji Earthquakes, N = 1000, Lat vs Long



```
>min_temp <- min(temperatura)
>max_temp <- max(temperatura)
>plot(NULL,xlab="anos", ylab="temperatura",
      xlim=c(1912,1971),
      ylim=c(min_temp,max_temp),
      main = "Temperatura média em New Haven -\nAnos 1921 até 1971")
>points(1912:1931,temperatura[1:20],pch=15,col="red")
>points(1932:1951,temperatura[21:40],pch=16,col="blue")
>points(1952:1971,temperatura[41:60],pch=17,col="magenta")
>linha_fit <- lsfit(1912:1971,temperatura)
>abline(linha_fit, col="red",lwd=2)
```



### Exercícios

Carregar `carros.csv` no dataframe `carros`

Carregar `qualidade_ar.csv` no dataframe `qualidade_ar`

Carregar `ELISA.csv` no dataframe `ELISA`

Fazer gráfico de dispersão para o data set `carros`, variáveis `seep` x `dist`

Fazer gráfico de dispersão para o data set `ELISA`, variáveis `density` x `conc`

Fazer gráfico de dispersão para o data set `qualidade_ar` Temp vs Ozone

Fazer gráfico de dispersão para o data set `qualidade_ar` Ozone vs Wind

## Análise Exploratória de Dados

A análise exploratória de dados é a primeira aproximação dos dados que foram coletados. Ela não é uma etapa obrigatória da análise estatística, mas é extremamente importante. Sua execução antecedendo a análise principal é aconselhável porque: (1) identificar relações inesperadas ou interessantes entre as variáveis, (2) verificar se existem evidências apoiando ou refutando uma hipótese, (3) verificar a existência de problemas com os dados, (4) identificar se é necessário coletar mais dados para determinada variável e, (5) principalmente, familiarizar-se com o padrão de distribuição de suas variáveis e outras características.

A análise exploratória de dados, ou “Exploratory Data Analysis” (EDA), foi considerada por John Tukey e outros estatísticos como um “conjunto de ações e atitudes” para a exploração dos dados. Esta metodologia consiste em **Revelação** de padrões, uso de testes robustos **Resistentes** a observações extremas mas também na observação de padrões nestas observações, os **Resíduos**, e na **Reexpressão** ou transformação dos valores dos dados (ALBERT; RIZZO 2012).

Neste roteiro, nosso objetivo é mostrar que você pode explorar seus dados com o conhecimento da linguagem R que já aprendeu, pelo menos inicialmente. Após a apresentação dos exemplos, explore os conjuntos de dados indicados da forma mais ampla possível, o que significa que não necessariamente você deva se limitar às análises que fizemos nos exemplos. O propósito é produzir o máximo de informação possível.

## 1 Revelação

A etapa de revelação consiste na elaboração de representações gráficas sobre seus dados. Isso quer dizer que será necessário desenhar gráficos para elucidar as características da distribuição de suas variáveis. Gráficos diferentes podem “revelar” propriedades diferentes; use tantos gráficos quanto necessário para esgotar a apresentação de informações relevantes para responder à sua pergunta de pesquisa.

### 1.1 Dataset: `ToothGrowth`

Resultados de experimento sobre efeito da Vitamina C no crescimento de dentes de cobaias (*Cavia porcellus*).

**Pergunta:** Que modo de suplementação pode fornecer vitamina C em quantidades adequadas para o metabolismo saudável?

**Experimento:** Efeito de diferentes fontes de vitamina C (dieta ou solução de ácido ascórbico) no crescimento dos dentes em cobaias.

Primeiro, familiarize-se com os tipos de variáveis e como estão organizadas.

```
#Exploração preliminar do dataset
#Carrega dados
```

```

> data(ToothGrowth)

#identifica classe do objeto
> class(ToothGrowth)
[1] "data.frame"

#identifica classe das variáveis que compõem o objeto
> lapply(ToothGrowth,class)
> $len
[1] "numeric"

> $supp
[1] "factor"

> $dose
[1] "numeric"

#identifica como as variáveis estão estruturadas
> str(ToothGrowth)
'data.frame': 60 obs. of 3 variables:
 $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
 $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
 $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...

#escreve Head do dataset
> head(ToothGrowth)
  len supp dose
1  4.2   VC  0.5
2 11.5   VC  0.5
3  7.3   VC  0.5
4  5.8   VC  0.5
5  6.4   VC  0.5
6 10.0   VC  0.5

```

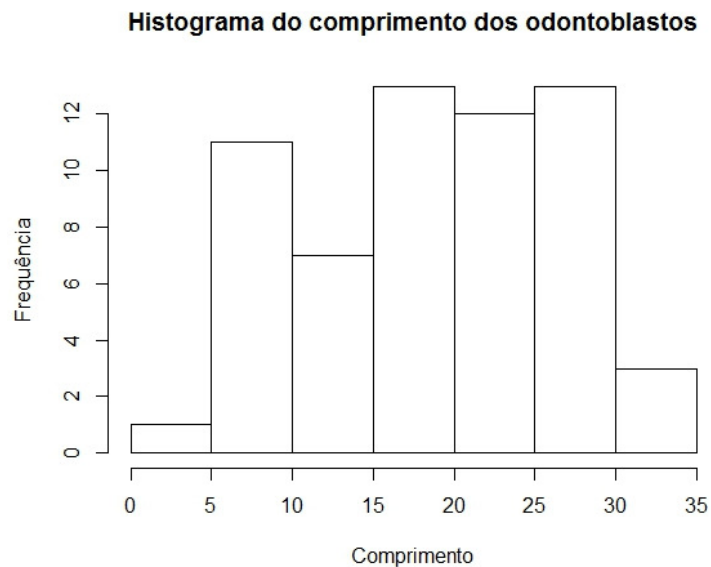
**###Exercício 01:** Pela análise dessas informações, você sabe qual foi o desenho experimental desse estudo? Responda:

- Qual o tamanho total da amostra? Qual o número de indivíduos nos tratamentos “dose” e “supp”?
- Esse experimento teve quantos fatores testados?

Agora podemos partir para a análise gráfica dos resultados. Nossa variável de resposta (y) é o comprimento dos odontoblastos.

```
#Histograma do comprimento dos odontoblastos
```

```
>hist(ToothGrowth$len,      main="Histograma do comprimento dos  
odontoblastos", xlab = "Comprimento", ylab = "Frequência")
```



Nesse histograma parece que a variável “comprimento do odontoblasto” tem formato da distribuição um pouco diferente da forma da curva normal. Será que essa variável tem distribuição normal mesmo? Teste:

```
#Teste de normalidade  
>bell<-shapiro.test(ToothGrowth$len)  
>bell
```

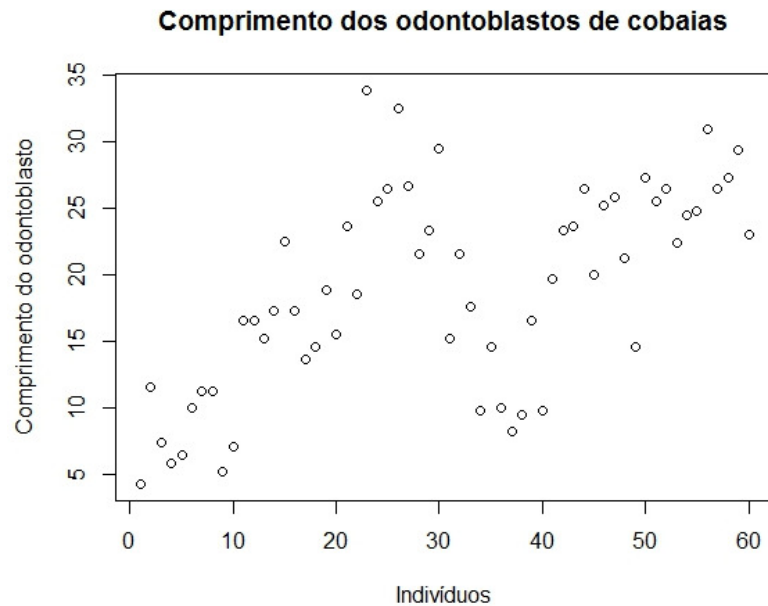
Shapiro-Wilk normality test

```
data: len  
W = 0.9674, p-value = 0.1091
```

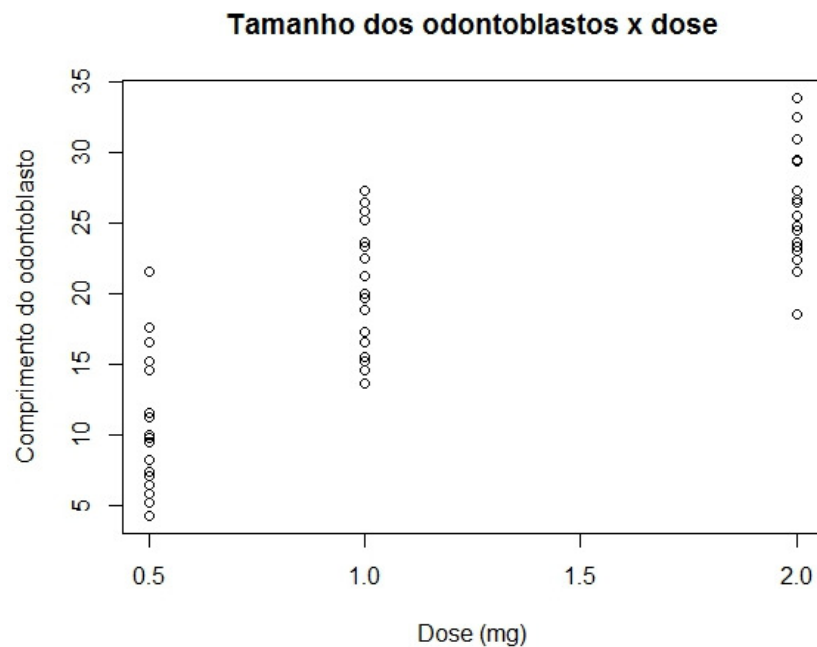
Tudo ok, podemos continuar. Para a análise gráfica, atender ao requisito de normalidade não é tão importante, mas nos testes estatísticos posteriores isso é essencial.

```
#Dispersão: comprimento do odontoblasto de todos os indivíduos  
testados
```

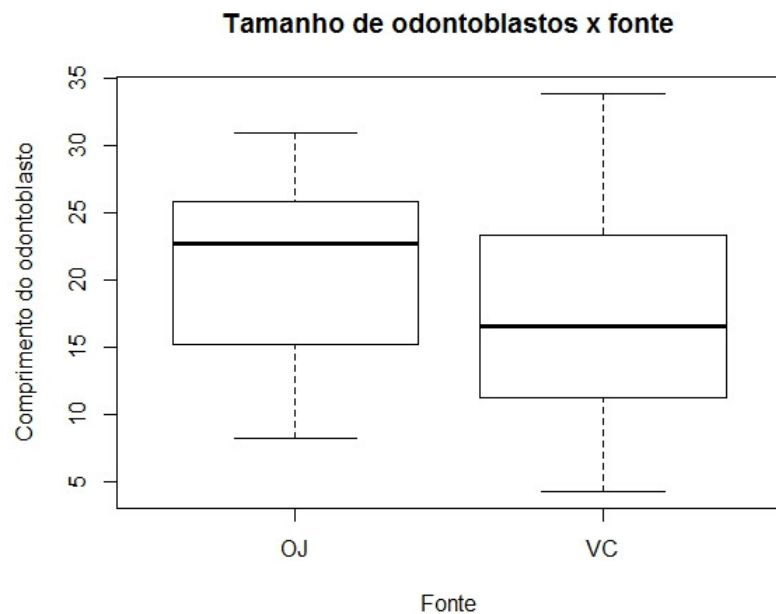
```
>dispersal<-plot(ToothGrowth$len,      main      =      "Comprimento      dos  
odontoblastos de cobaias", xlab = "Indivíduos", ylab = "Comprimento do  
odontoblasto")
```



```
#Dispersao: comprimento do odontoblasto x DOSE  
>dispersa2<-plot(ToothGrowth$dose, ToothGrowth$len ,main="Tamanho dos  
odontoblastos x dose", xlab="Dose (mg)",ylab="Comprimento do  
odontoblasto")
```



```
#Dispersao: comprimento do odontoblasto x FONTE
>dispersa3<-plot(ToothGrowth$supp,ToothGrowth$len,main="Tamanho de
odontoblastos x fonte",xlab="Fonte",ylab="Comprimento do
odontoblasto")
```



**###Exercício 02:** A interpretação dos dados é parte crucial da EDA. Responda:

- O que você pode observar nos gráficos de dispersão acima?
- Mas espere: `dispersa3` é um boxplot, e não um gráfico de dispersão convencional, ainda que o comando que o produziu (`plot`) seja exatamente o mesmo dos gráficos anteriores. Por que isso ocorreu?

**###Exercício 03:** O painel com a comparação dos gráficos dos efeitos de diferentes doses de vit.C ministradas pela dieta ("OJ", "Orange juice": suco de laranja) ou suplementação de vit.C ("VC", "Vitamin C": solução de ácido ascórbico) corroborou suas expectativas a respeito dos resultados do experimento? Por quê?

## 2. Testes Robustos e Resíduos

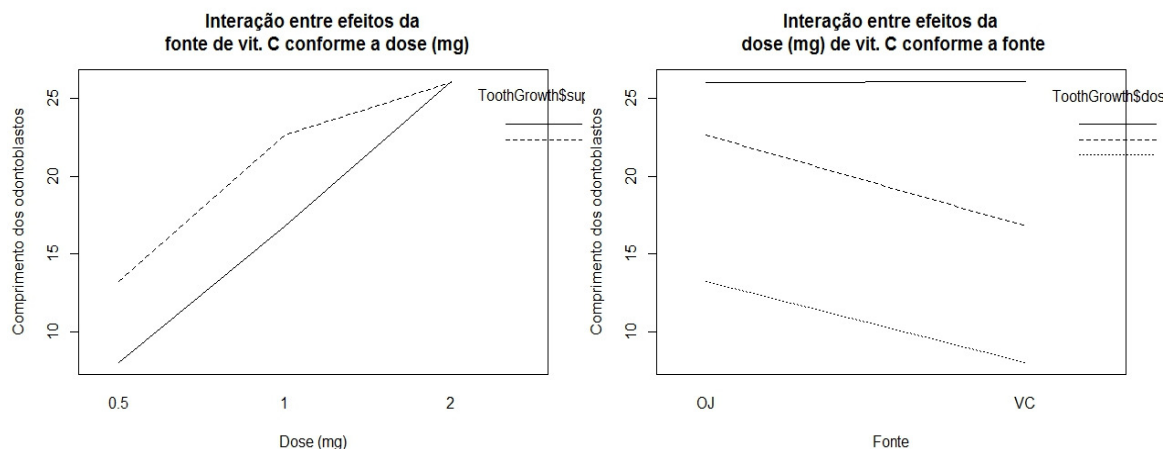
Agora que temos uma ideia da relação entre as variáveis, podemos aplicar testes robustos para verificar se essas relações hipotéticas são corroboradas. Os gráficos abaixo exploram mais uma vez os efeitos dos diferentes tratamentos no comprimento dos odontoblastos.

###interação entre variáveis

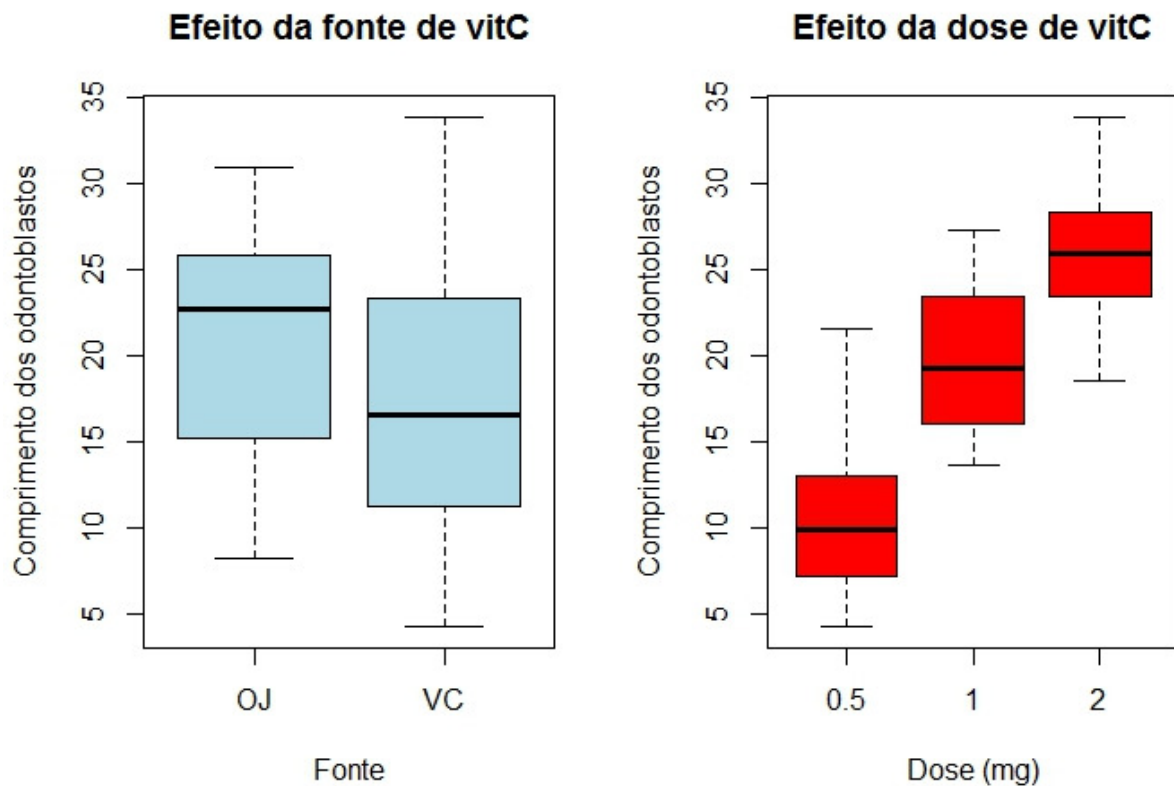
```
> dev.off    #esse comando "limpa" as definições gráficas
> par(mfrow=c(1,2))    #organiza a exibição de 2 gráficos lado a lado
```

```
> interage1<-interaction.plot(ToothGrowth$dose,      ToothGrowth$supp,
ToothGrowth$len, main="Interação entre efeitos da\
 fonte de vit. C conforme a dose (mg)",xlab="Dose (mg)",ylab="Comprimento dos odontoblastos")
```

```
> interage2<-interaction.plot(ToothGrowth$supp,      ToothGrowth$dose,
ToothGrowth$len, main="Interação entre efeitos da\
 dose (mg) de vit. C conforme a fonte",xlab="Fonte",ylab="Comprimento dos odontoblastos")
```



Estes gráficos permitem observar que o efeito da fonte de vitamina D é menos importante que o efeito da dosagem: os efeitos da fonte acabam se igualando a medida que a dose aumenta, mas os efeitos da dose são máximos independentemente da fonte. O efeito da dose é mais forte. Compare os boxplots:



Para testar essa hipótese, podemos realizar um teste bastante robusto para comparar os tratamentos, a análise de variância (ANOVA). Esta análise indicará se os tratamentos de fato diferem significativamente.

```
#ajuste de modelo linear
>resultado <- lm( len ~ supp + dose + supp*dose, data=ToothGrowth)
>summary(resultado)
```

Call:

```
lm(formula = len ~ supp + dose + supp * dose, data = ToothGrowth)
```

Residuals:

| Min     | 1Q      | Median | 3Q     | Max    |
|---------|---------|--------|--------|--------|
| -8.2264 | -2.8462 | 0.0504 | 2.2893 | 7.9386 |

Coefficients:

| Estimate       | Std. Error | t value | Pr(> t ) |
|----------------|------------|---------|----------|
| (Intercept)    |            |         |          |
| suppVC         |            |         |          |
| dose0.5        |            |         |          |
| dose1          |            |         |          |
| suppVC:dose0.5 |            |         |          |
| suppVC:dose1   |            |         |          |

```

(Intercept)  11.550      1.581    7.304 1.09e-09 ***
suppVC       -8.255      2.236   -3.691 0.000507 ***
dose         7.811      1.195    6.534 2.03e-08 ***
suppVC:dose   3.904      1.691    2.309 0.024631 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.083 on 56 degrees of freedom
Multiple R-squared:  0.7296, Adjusted R-squared:  0.7151
F-statistic: 50.36 on 3 and 56 DF,  p-value: 6.521e-16

```

```

#computa ANOVA
>anova(resultado)
Analysis of Variance Table

```

```

Response: len
      Df Sum Sq Mean Sq  F value    Pr(>F)
supp    1  205.35   205.35   12.3170 0.0008936 ***
dose    1 2224.30  2224.30  133.4151 < 2.2e-16 ***
supp:dose 1   88.92    88.92    5.3335 0.0246314 *
Residuals 56  933.63   16.67
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Sabendo que existem diferenças significativas, precisamos realizar o teste de comparação múltipla para localizar os grupos que efetivamente diferem dos demais.

```

#comparação múltipla
>multic<-TukeyHSD(aov(len ~ supp + dose + supp*dose, data=ToothGrowth))

```

```

Tukey multiple comparisons of means
 95% family-wise confidence level

```

```

Fit: aov(formula = len ~ supp + dose + supp * dose, data = ToothGrowth)

```

```

$supp
      diff      lwr      upr    p adj
VC-0J -3.7 -5.811942 -1.588058 0.0008936

```

```

>ToothGrowth$dose <- factor(ToothGrowth$dose)
>resultado2 <- aov(len ~ dose + supp, data=ToothGrowth)
>TukeyHSD(resultado2, "supp")

```

Tukey multiple comparisons of means  
95% family-wise confidence level

Fit: aov(formula = len ~ dose + supp, data = ToothGrowth)

```
$supp
      diff      lwr      upr    p adj
VC-0J -3.7 -5.679762 -1.720238 0.0004293
```

>TukeyHSD(resultado2,"dose")

Tukey multiple comparisons of means  
95% family-wise confidence level

Fit: aov(formula = len ~ dose + supp, data = ToothGrowth)

```
$dose
      diff      lwr      upr p adj
1-0.5  9.130  6.215909 12.044091 0e+00
2-0.5 15.495 12.580909 18.409091 0e+00
2-1    6.365  3.450909  9.279091 7e-06
```