Introdução à Estatística em R
Curso de Big Data e Data Science

Guilherme Gomes

10 de Março de 2018
versão 6.1

SUMÁRIO

Introdução

Uma pergunta relevante que surge quando estudamos estatística é o significado da palavra *Dados*. Para o que segue nesse nesse módulo, vamos utilizar a seguinte definição de dados:

Definição: Dados (*Data*) são um conjunto de elementos, cada qual chamada de dado (*Datum*). Um dado é uma medida qualitativa ou quantitativa de uma observação, usado como fonte de informação.

Para fixar ideias considere o seguinte exemplo:

Suponha que um observador registrou a temperatura de todos os dias da semana ao meio dia. Abaixo segue o vetor que representa suas observações, ou seja seu conjunto de dados.

28
27
26
30
37
31
29

Note que esse conjunto de dados por si só, é composto apenas por números. O conjunto de dados não é a informação propriamente dita, ele é usado como fonte de informação. No entanto essa informação não está completa, pois a cada observações devemos associar uma outra variável que representa o dia da semana onde ela foi registrada. Podemos apresentar a informação completa no mesmo conjunto de dados desde que rotulemos as linhas da seguinte forma:

Dia da semana	Temperatura em °C às 12:00
dom	28
seg	27
ter	26
qua	30
qui	37
sex	31
sáb	29

A tabela acima representam uma visualização do conjunto de dados, que são números e variáveis qualitativas, que representam os dias da semana. Essa visualização não é única, como veremos em breve. A interpretação que conseguimos extrair desses números e variáveis qualitativas é que representa a informação existenten no conjunto de dados.

A estatística pode ser dividida em três grandes grupos, de acordo como os dados serão tratados:

1. Descrição: O conjunto de dados é analisado em si, sendo criadas medidas capazes de sumarizar a informação que eles nos fornecem assim como métodos de visualização;
2. Coleta: Se estuda qual a melhor maneira de coletar os dados para criar uma mostra que seja representativa da população;

3. Inferência Estatística: A partir dos dados existentes buscasse fazer afirmações sobre toda a população, que não é observada.

Nos problemas estatísticos atuais, lidamos com uma grande quantidade de dados. No mais, este é um curso de *Big Data* então deveremos ser capazes de lidar com toda essa informação. A análise descritiva é uma primeira abordagem nesse trabalho. Nela buscamos encontrar medidas que resumem grande parte da informação dispersa nos dados existente, assim como estudar ferramentas sobre a melhor forma de visualização desses dados.

Descrição Univariada

Nosso primeiro passo será o de descrever os dados. O `R` possui algumas bases de dados que podem ser acessadas sem a necessidade de importação. Para visualizar essas bases utilize o comando `data()`, este comando irá abrir uma nova aba contendo uma lista com o nome dessas bases. Para acessar a base, basta digitar o nome dela em seu *Console* ou *Script*. Além das bases originais, alguemas bases adicionais podem ser instaladas por meio dos pacotes.

Para esse primeiro exemplo vamos utilizar uma base de dados externa, para isso precisaremos importar os dados. Vamos trabalhar com base de dados de altura em cm e peso em kg de 3000 indivíduos. Utilize o comando abaixo para importar o arquivo para o `R` e salvar em um `data.frame`.

```
setwd('caminho da pasta onde está salvo seu arquivo')
df<-read.csv2('altura_e_peso.csv')
```

Conhecendo seu data.frame

O objeto básico de armazenamento de dados estatísticos no `R` é o `data.frame`. Ele é uma tabela onde as linhas representam as observações e as colunas representam as medidas. [1]

Para visualizar as primeiras linhas em seu *Console* utilize a função `head`.

```
n<-3
head(df,n)
```

```
##   altura  peso
## 1 142.81 63.30
## 2 145.59 52.16
## 3 147.94 62.27
```

Para visualizar as últimas linhas utilize a função `tail`.

```
n<-2
tail(df,n)

##      altura   peso
## 2999 187.82  90.47
## 3000 196.26 100.50
```

Nesse `data.frame` existem 2 variáveis, e 3000 observações (para verificar isso observer o seu painel *Environment* ou use a função `dim` que retorna as dimensões do seu `data.frame`. Um `frame.frame` é uma tabela com alguns metadados, que são nomes das linhas e colunas. Para acessar essas informações utilize os comando `rownames` e `colnames` ou `names`.

```
dim(df)

## [1] 3000    2
```

[1] Para mais informações sobre `data.frame` ver em https://github.com/mobileink/data.frame/wiki/What-is-a-Data-Frame%3F ou https://analisereal.com/2017/02/07/data-frames/

```r
head(rownames(df),5)

## [1] "1" "2" "3" "4" "5"

colnames(df)

## [1] "altura" "peso"

names(df)

## [1] "altura" "peso"
```

Inicialmente vamos trabalhar somente com 1 variável que é a altura. Selecione a coluna que representa a altura. Essa seleção pode ser feita de algumas formas:

```r
altura<-df$altura
altura<-df[,1]
altura<-df[,'altura']
```

Abaixo apresentamos algumas estatísticas que representam a dispersão dos dados. Todas as observações da variável altura estão entre o valor máximo e mínimo.

```r
min(altura)

## [1] 142.81

max(altura)

## [1] 196.26

range(altura)

## [1] 142.81 196.26
```

Tabela de frequência

Uma tabela de frequências resume a informação de quantas observações existem em cada intervalo. Para construir uma devemos nos atentar a dois passos:

1. Primeiramente é preciso definir os intervalos. Nesse passo definimos explicitamente quais são os intervalos que queremos. Para essa tarefa utilize a função `cut`. No primeiro argumento da função é indicada quais os dados que se deseja organizar, no segundo argumento se define os intervalos. Neste exemplos vimos que a dispersão da altura vai de 142.81 até 196.26, desse modo é nececessário se atentar para que os extremos do intervalo contenham esses valores.

   ```r
   intervalo<-cut(altura,breaks = seq(from = 140,to = 200,by = 5))
   ```

   Essa função irá indicar para cada observação em qual intervalo ela está contida. Observe abaixo no `data.frame` auxiliar que foi criado para visualizar esse resultado.

```
head(data.frame(altura,intervalo))
```

```
##   altura intervalo
## 1 142.81 (140,145]
## 2 145.59 (145,150]
## 3 147.94 (145,150]
## 4 148.30 (145,150]
## 5 146.67 (145,150]
## 6 148.19 (145,150]
```

2. Contar quantas observações pertencem a cada intervalo:

```
table(intervalo)
```

```
## intervalo
## (140,145] (145,150] (150,155] (155,160] (160,165] (165,170] (170,175]
##         1        22        96       367       781       878       579
## (175,180] (180,185] (185,190] (190,195] (195,200]
##       223        42        10         0         1
```

Gostariamos de trabalhar com essas frequências para isso vamos salvar essa tabela em um `data.frame`, atribuindo nomes a cada uma das colunas:

```
tab<-data.frame(table(intervalo))
names(tab)<-c('intervalo','freq')
tab
```

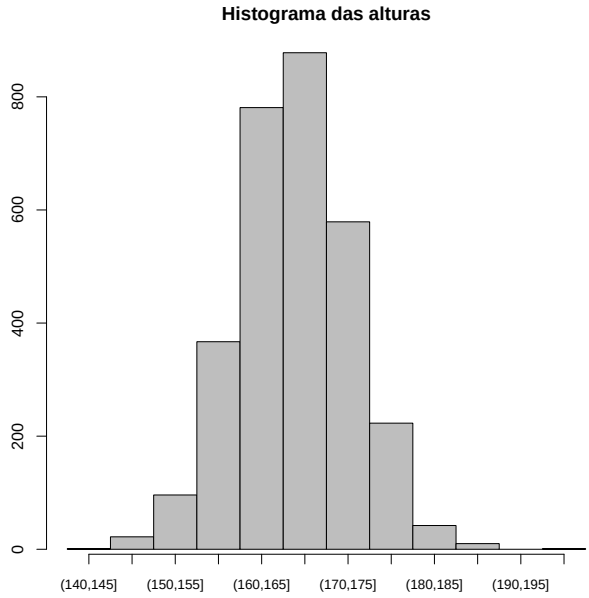
```
##    intervalo freq
## 1  (140,145]    1
## 2  (145,150]   22
## 3  (150,155]   96
## 4  (155,160]  367
## 5  (160,165]  781
## 6  (165,170]  878
## 7  (170,175]  579
## 8  (175,180]  223
## 9  (180,185]   42
## 10 (185,190]   10
## 11 (190,195]    0
## 12 (195,200]    1
```

Histograma

Um histograma é uma visualização gráfica da tabela de frequência. No eixo horizontal estão os intervalos e no eixo vertical a frequência.

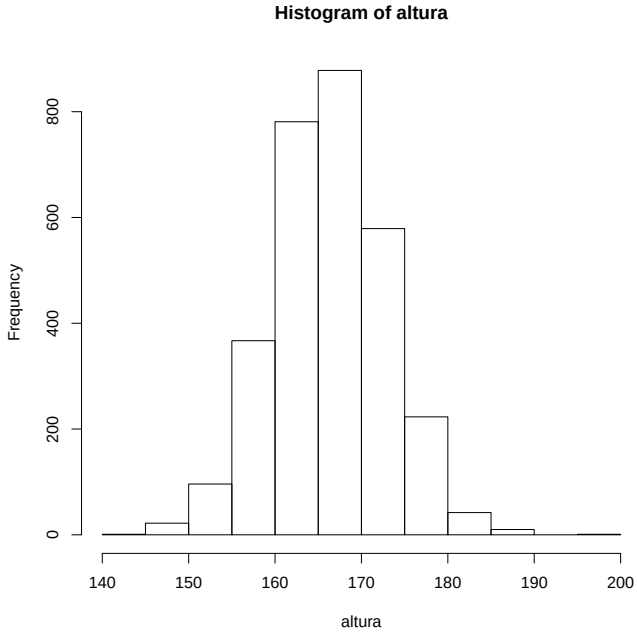
```
bplot<-barplot(height = tab$freq,space = F,
               main='Histograma das alturas')
axis(side = 1,at = bplot
```

```
    ,labels = tab$intervalo
    ,cex.axis=0.8)
```



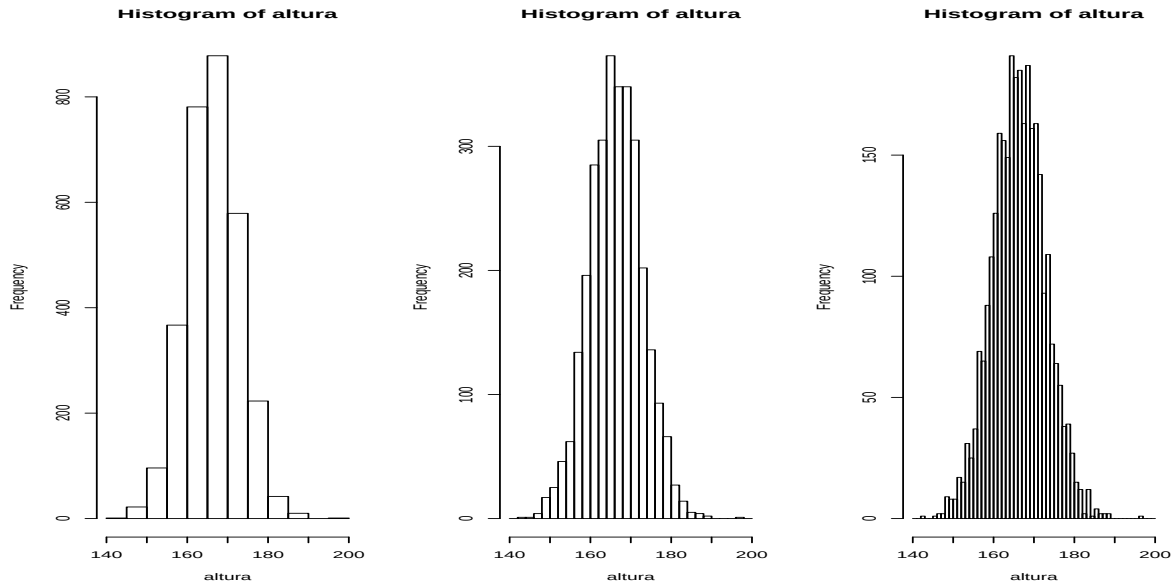
De maneira alternativa podemos pular essa etapa de construir a tabela de frequências e ir direto para o histograma utilizando a função `hist`

```
hist(altura)
```



É possível definir os intervalos do histograma, utilizando o argumento `breaks`. [2]

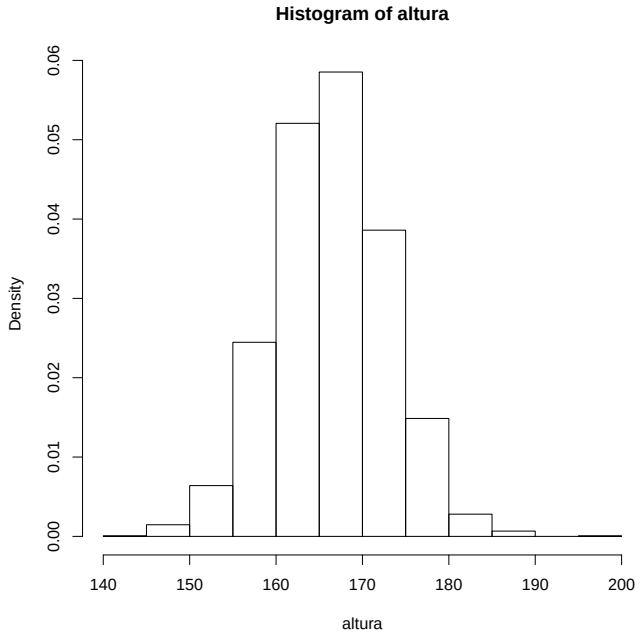
```
par(mfrow=c(1,3))
hist(altura,breaks = seq(from=140,to=200,by=5))
hist(altura,breaks = seq(from=140,to=200,by=2))
hist(altura,breaks = seq(from=140,to=200,by=1))
```



O histograma pode ser apresentado com a frequência relativa, ao invés da frequência absoluta. Nesse caso a área de cada de cada barra representa a frequência relativa daquele intervalo. Isto é feito para que a área total do histograma seja igual a 1. Esse tipo de transformação do eixo tem sentido quando queremos comparar o histograma com distribuições de probabilidade, como veremos mais adiante.

```
hist(altura,prob=T)
```

[2] O número de intervalos k padrão no R é calculado por meio da fórmula de Sturge: $k = log_2 n + 1$, onde n é a quantidade de observações.



Medidas de síntese - Centralidade

São três as medidas básicas de centralidade utilizadas em estatística: média, mediana e moda.

1. Média Aritmética com ponderação Simples

   Exemplo da moeda não viciada: Joguei 11 vezes a moeda, 6 vezes deu Cara e 5 vezes deu Coroa. Qual a média? Não podemos calcular se não estamos associando o evento a uma variável aleatória. Pois essa operação não está bem definida. Afinal de contas o que é a média entre dois eventos? Sabemos realizar operações algébricas com números mas não com eventos. Por isso a importância de trabalhar com variáveis aleatórias.

   $$mean = \frac{6.cara + 5.coroa}{11} = ?$$

   Nesse caso, associando a distribuição Binomial, onde cara representa o evento sucesso (1) e coroa o evento fracasso (0). Assim podemos calcular a média que é 0,6.

   $$mean = \frac{6.1 + 5.0}{11} = \frac{6}{11}$$

   Mas o que significa esse resultado em termos de evento? Nesse caso específico nada, pois o resultado 0,6 não está associado a nenhum dos eventos que definimos. Só sabemos o significado dos números 1 e 0 nesse caso específico. Apesar de não significar nada em termos de eventos, como veremos esse tremo possui um significado em termos probabilístico, sendo usado como um estimador para a probabilidade do resultado cara surgir quando do lançamento de uma moeda.

   Vamos reproduzir um experimento similar:

```r
set.seed(11)
n <- 11      # número de lançamentos
p <- 1/2     # probabilidade do evento sucesso (1), Cara
y<-sample(0:1, size = n, prob = c(1-p,p),replace = TRUE)
y
```

```
##  [1] 1 1 0 1 1 0 1 1 0 1 1
```

```r
mean(y)
```

```
## [1] 0.7272727
```

Como podemos observar essa média pode estar longe do parâmetro que representa a probabilidade. No final dessa aula, veremos a importância do tamanho da amostra para redução desse erro. Na próxima aula vamos estudar as propriedades desse estimador. A seguir, um exemplo com uma amostra maior:

```r
set.seed(11)
n <- 1001      # número de lançamentos
p <- 1/2     # probabilidade do evento sucesso (1), Cara
y<-sample(0:1, size = n, prob = c(1-p,p),replace = TRUE)
# você pode usar a função rbinom() com n=1 também.

head(y)
```

```
## [1] 1 1 0 1 1 0
```

```r
mean(y)
```

```
## [1] 0.5114885
```

Um argumento interessante da função mean do R é o "trimmed" (aparar). Nele você define o quanto deseja aparar das pontas da amostra. Ou seja se você deseja excluir 10% dos elementos maiores e 10% dos elementos menores basta usar o comando:

```r
mean(y,trim = 0.1)
```

```
## [1] 0.5143571
```

Essa opção é interessante para excluir "outliers" que podem ter efeito muito grande na média.

Caso sua base de dados possua NA's, a função mean não irá funcionar a menos que você coloque o argumento para remover os NA's.

```r
x<-c(NA,NA,1,3)

mean(x)
```

```
## [1] NA
```

```r
mean(x, na.rm = T)
```

```
## [1] 2
```

2. Mediana

Uma outra medida pode ser usada para retratar a tendência central dos dados. Esta medida, em muitos casos, tem um evento associado no espaço de eventos. A mediana é a observação que separa a amostra em dois conjuntos com a mesma quantidade de elementos. Ela é encontrada da seguinte maneira: ordenando a amostra pelo resultado numérico da variável aleatória associada ao evento e escolhendo o número que separa a amostra em dois conjuntos de igual tamanho.

Aqui temos o vetor de realização do exemplo do cara ou coroa, ordenado:

$$\Big(0,0,0,0,0,1,1,1,1,1,1\Big)$$

$$\Big(\underbrace{0,0,0,0,0,}_{1^o set}\ \underbrace{\mathbf{1}}_{Mediana}\ ,\underbrace{1,1,1,1,1}_{2^o set}\Big)$$

Observe que, nesse caso, com uma amostra de tamanho ímpar, temos que sempre existirá uma realização que divide a amostra em dois conjuntos de mesmo tamanho. Essa realização de fato representa um evento. Aqui a mediana é 1, que representa o evento cara.

No entanto, se o conjunto tiver um número par de elementos (n) então a mediana é calculada como a média aritmética simples entre os elementos na ordem $n/2$ e $n/2+1$. Nesse caso, pode ser que a mediana não tenha uma interpretação com relação aos eventos, como é o caso da média. Se os dados estão ordenados podemos escrever a mediana como:

- Se n é ímpar:

$$Median(X) = X_j, j = (n+1)/2$$

- Se n é par:

$$Median(X) = \frac{(X_j + X_{j+1})}{2}, j = n/2$$

Utilizando o mesmo exemplo temos:

```
set.seed(11)
n <- 11      # número de lançamentos
p <- 1/2     # probabilidade do evento sucesso (1), Cara
y<-sample(0:1, size = n, prob = c(1-p,p),replace = TRUE)
y
```

```
##  [1] 1 1 0 1 1 0 1 1 0 1 1
```

```
median(y)
```

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

3. Moda

A terceira medida, a moda representa o evento mais recorrente nos dados observados. No caso, a associação é direta com os eventos observados. No exemplo temos que 6 realizações foram 1 e 5 realizações foram 0. Desse modo o evento mais recorrente é aquele associado ao número 1, no caso Cara. Pode ser o caso de uma

amostra ter dois eventos que possuem a mesma quantidade de realizações, dessa forma dizemos que a amostra é bimodal. De forma mais geral, se uma amostra possui mais de uma moda, dizemos que ela é multimodal.

O R não possui uma função raiz para a moda. Vamos então escrever uma:

```
moda<-function(x){
y<-data.frame(table(x))
maior<-max(y[,2])
y[y[,2]==maior,]
}

# Testando a função:
x<-c('ana','maria','joão','ana','maria','rita','rita','rita')
x

## [1] "ana"   "maria" "joão"  "ana"   "maria" "rita"  "rita"  "rita"

moda(x)

##      x Freq
## 4 rita    3
```

Arredondamento

Algo importante de apresentar aqui é a função arredondamento do `R`, a função `round`.

```
round(1.1)

## [1] 1

round(0.94,digits = 1)

## [1] 0.9
```

Caso seja apresentada sem o argumento `digits` ela arredonda os valores decimais para o inteiro mais próximo. Se o argumento `digits` não estiver vazia, ela arredonda para o número decimal com k digitos mais próximo.

O que ocorre porém com os valores intermediários, isto é, quando o último dígito for 5?

```
round(1.5)

## [1] 2

round(2.5)

## [1] 2
```

Esse resultado a primeira vista é estranho. Se o arredondamento de 1.5 é 2 é de se esperar que o arredondamento de 2.5 seja 3, ou seja o arredondamento seja feito para cima. Porém o arredondamento do `R` é feito da seguinte maneira: se for ímpar o algarismo significativo então arredonda para cima, se for par então arredonda para baixo.

Isso ocorre para evitar viés na média quando se arredonda os valores de uma amostra. De fato se ocorresse o arredondamento sempre para cima haveria um problema dee viés no cálculo da média, pois a média entre 1.5 e 2.5 que é 2 seria calculada como sendo 2.5.

```
mean(c(1.5,2.5))

## [1] 2

mean(c(2,3))

## [1] 2.5

mean(round(c(1.5,2.5)))

## [1] 2
```

Medidas de síntese - Dispersão

As medidas de dispersão representam informações de o quanto os dados estão dispersos. Uma medida simples de dispersão de dados pode ser a própria extensão (*range*) das observações. No entanto, essa medida é muito sensível a valores extremos, os chamados *outliers*[3]. Por esse motivo, outras medidas de dispersão dos dados são mais usuais como a variância, desvio padrão e os quantis:

1. Variância

   A variância é uma medida de o quanto os dados estão distantes da média. Seja X_1 uma observação da amostra X, que possui apenas duas observações. Temos que $X_1 - mean(X) = d$ é uma medida de o quanto esta primeira observação está distante da média. Seja a segunda observação X_2, e sua distância da média dada por $X_2 - mean(X) = -d$. Uma ideia de dispersão poderia ser a soma dessas distâncias, no entanto, se somarmos essas duas distâncias o resultado é claramente 0.

   $$X_1 - mean(X) + X_2 - mean(X) = d - d = 0$$

   Porém, a dispersão desses dados não é zero, dado que um está desviado para cima da média e o outro está desviado para baixo. Uma solução pragmática para esse problema é considerar a soma dos quadrados das distâncias da média, pois todo número ao quadrado é positivo.

   $$(X_1 - mean(X))^2 + (X_2 - mean(X))^2 = d^2 + d^2 = 2.d^2$$

   Isso resolve o problema de algumas observações serem maiores que a média e outras menores. No entanto ainda temos um problema relacionado à escala. Quanto mais observações tivermos na amostra, maior será essa soma dos quadrados das distâncias da média. Para resolver esse problema de escala podemos dividir pelo tamanho da amostra.

   $$\frac{(X_1 - mean(X))^2 + (X_2 - mean(X))^2}{2} = \frac{d^2 + d^2}{2} = \frac{2.d^2}{2} = d^2$$

   Assim podemos definir uma medida de dispersão, a chamada variância populacional do seguinte modo:

[3] Esse nome ficará mais intuitivo quando falarmos de regressão. Adiantando um pouco a ideia básica, em regressão vamos ajustar uma linha para os dados, contudo alguns dados podem ficar muito fora dessa linha, por conta disso são chamados de outliers

$$Variance_{pop}(X) = \sum_{i=1}^{n} \frac{(X_i - mean(X))^2}{n}$$

Na próxima aula mostraremos que essa medida pode ser considerada um estimador do parâmetro que representa a variância da população. Contudo, este estimador terá uma propriedade desagradável, ele será viesado. Uma outra medida de variância, que representa um estimador não viesado é a chamada variância amostral, definida como:

$$Variance_{sample}(X) = \sum_{i=1}^{n} \frac{(X_i - mean(X))^2}{n-1}$$

```
set.seed(21)
n = 50
x<-sample(1:10,size = n,replace = TRUE)

sum((x - mean(x))^2)/n

## [1] 8.7796

sum((x - mean(x))^2)/(n-1)

## [1] 8.958776

var(x)

## [1] 8.958776
```

Pelo exemplo acima verificamos que a função var() do R calcula a variância amostral.

2. Desvio-padrão

O desvio padrão é apenas uma transformação da variância, para ajustar a unidade de medida. Por exemplo, se tivermos uma amostra que representa as alturas das pessoas na população brasileira, essa altura estará muito provavelmente em metros ou centímetros. Nesse caso, a variância estará em uma outra unidade, no caso metros^2 ou centímetros^2. É estranho apresentar os dados dizendo que a média de altura é de 1.6 metros e a variância é de 0.04 metros^2. Para superar essa estranheza o desvio padrão é definido como a raiz quadrada da variância.

$$Std.Deviation_{sample}(X) = \sqrt{Variance(X)}$$

Desse modo podemos dizer mais facilmente que a média de altura é de 1.6 metros e o desvio padrão é de aproximadamente 0.2 metros. Isso facilita a compreensão dos dados e assim a análise estatística seguinte.

```
set.seed(21)
n = 50
x<-sample(1:10,size = n,replace = TRUE)
var(x)

## [1] 8.958776
```

```r
sqrt(var(x))
```

```
## [1] 2.993121
```

```r
sd(x)
```

```
## [1] 2.993121
```

3. Quantil

   Uma outra medida de dispersão muito utilizada são os quantis (*quantiles*). Eles representam os pontos que dividem os dados em partições de mesmo tamanho. Os quantis são definidos pela quantidade de partições que eles criam. Um k-quantil, divide o conjunto de dados em k partições. Alguns quantis tem nomes especiais: 2-quantil é a mediana, 3-quantis são os tercis, os 4-quantis são os quartis, os 100-quantis são os percentis. Note que existe somente 1 2-quantil, existem 2 tercis, existem 3 quartis e existem 99 percentis. Pela maneira que estamos categorizando os dados o 2-quantil é uma medida de tendência central e não necessariamente uma medida de de dispersão.

```r
set.seed(99)
n = 100
x<-rnorm(n = n,0,1)
round(head(x,3),digits = 3)
```

```
## [1] 0.214 0.480 0.088
```

```r
# sem especificação ele entrega os quartis
quantile(x)
```

```
##          0%         25%         50%         75%        100%
## -3.04093410 -0.64529643  0.04980912  0.45760695  1.86457602
```

```r
# Vamos espeficicar quais queremos
quantile(x,probs = c(0.1,0.5,0.85))
```

```
##         10%         50%         85%
## -1.35800782  0.04980912  0.68691519
```

```r
# Para encontrar todos os percentis
per<-quantile(x,probs = seq(0,1,length.out = 101))
head(per,5)
```

```
##        0%        1%        2%        3%        4%
## -3.040934 -2.513878 -2.298998 -1.761518 -1.681895
```

```r
tail(per,5)
```

```
##      96%      97%      98%      99%     100%
## 1.385415 1.394435 1.405184 1.658724 1.864576
```

4. IQR Definimos a medida IQR (*Interquartile Range*) como:

$$IQR = Q3 - Q1$$

onde:
Q1: primeiro quartil
Q3: terceiro quartil
Podemos calcular essa medida diretamente no R

```
IQR(altura)

## [1] 8.8725

as.numeric(quantile(altura)[4]-quantile(altura)[2])

## [1] 8.8725
```

Box-plot

O box-plot é uma técnica de visualização gráfica que dá ênfase aos quartis. Ele é construído da seguinte forma:

- Uma caixa é desenhada do primeiro quartil até o terceiro. Essa caixa representa 50% dos dados totais que estão mais próximos do centro das observações na amostra.
- Uma linha é desenhada no interior dessa caixa, representando a mediana (ou se preferir o segundo quartil).
- Os 25% dos dados mais espaçados para cima e para baixo, são representados por uma linha pontilhada.

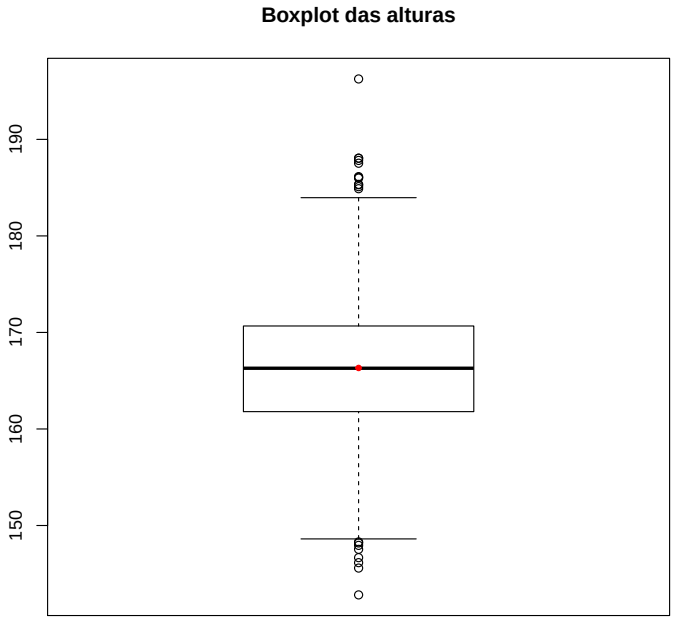
Uma modificação convencional feita nesse tipo de gráfico, para lidar com *outliers* é representar o limite da linha pontilhada por um condicional.

Assim o limite superior será definido como sendo $Q3 + 1,5.IQR$ ou o valor máximo da amostra, dependendo de qual for o maior. Analogamente o limite inferior será definido como $Q1 - 1,5.IQR$ ou o valor mínimo da amostra, dependendo de qual for o menor. Todas as observações que superarem tal limite são representadas por pontos e reconhecidos como *outliers*.[4]

No R o boxplot pode ser gerado pela função `boxplot`:

```
boxplot(altura,main = 'Boxplot das alturas')
points(mean(altura),col = 2,pch = 20)          # introduzindo a média
```

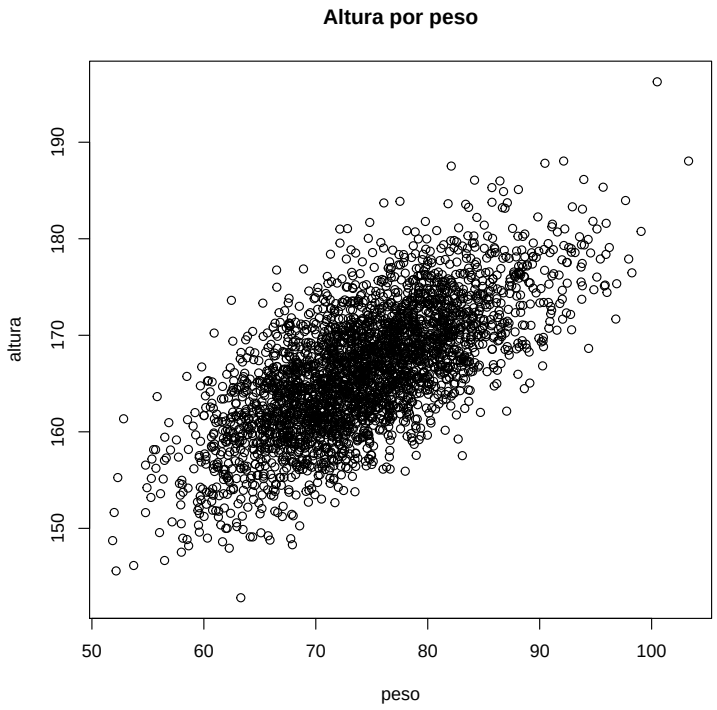
[4]Uma observação a ser feita é que isso é uma convenção, existem outras maneiras de definir *outliers*. Uma delas é que a observação não esteja na mesma linha, em uma regressão linear, por exemplo.


Boxplot das alturas
190
180
170
160
150

Descrição Multivariada

Vamos estudar agora o comportamento multivariado das variáveis. A princípio estudaremos os modelos bivariados, porém os resultados podem ser estendidos para os casos com mais dimensões. A maneira mais direta de obter informações sobre dados bivariados é apresentá-los em um gráfico de pontos ("scatter plot")¿

```
plot(df$peso,df$altura,xlab='peso',ylab='altura',main='Altura por peso')
```



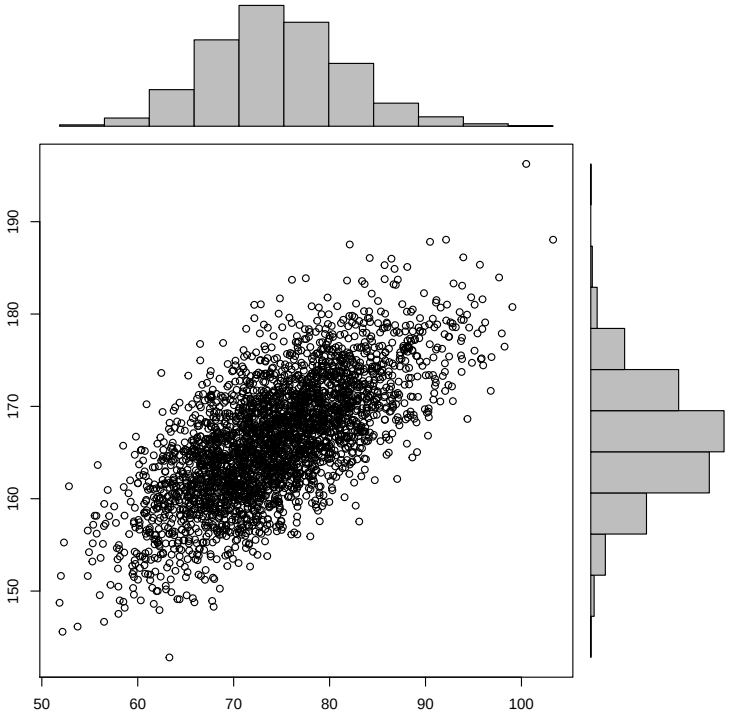
Que bagunça de pontos! Como fazer para melhor visualizar e descrever essa informação?

Histograma Marginal

Nessa visualização os histogramas são calculados de forma a considerar cada uma das dimensões separadamente. No caso da altura o histograma marginal é o mesmo que foi calculado anteriormente. Abaixo segue um script com uma maneira de escrever a função de histograma marginal.

```
scatterhist = function(x, y, xlab="", ylab=""){
  zones=matrix(c(2,0,1,3), ncol=2, byrow=TRUE)
  layout(zones, widths=c(4/5,1/5), heights=c(1/5,4/5))
  xhist = hist(x, plot=FALSE)
  yhist = hist(y, plot=FALSE)
  top = max(c(xhist$counts, yhist$counts))
  par(mar=c(3,3,1,1))
  plot(x,y)
  par(mar=c(0,3,1,1))
  barplot(xhist$counts, axes=FALSE, ylim=c(0, top), space=0)
  par(mar=c(3,0,1,1))
```

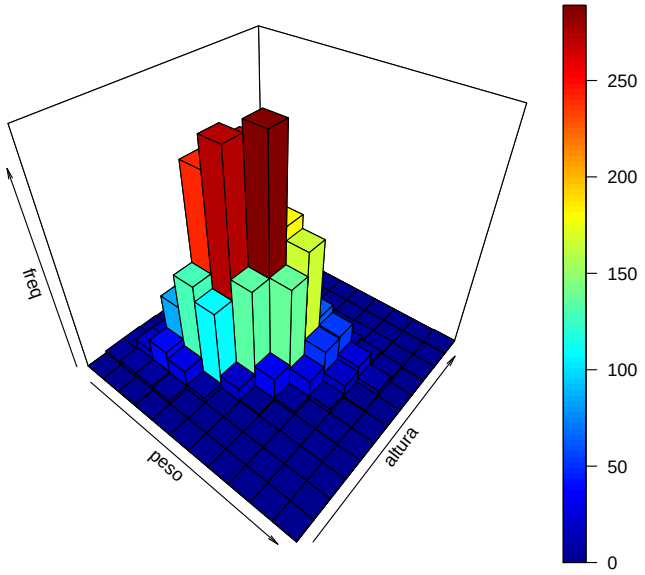
```
  barplot(yhist$counts, axes=FALSE, xlim=c(0, top), space=0, horiz=TRUE)
  par(oma=c(3,3,0,0))
  mtext(xlab, side=1, line=1, outer=TRUE, adj=0,
        at=.8 * (mean(x) - min(x))/(max(x)-min(x)))
  mtext(ylab, side=2, line=1, outer=TRUE, adj=0,
        at=(.8 * (mean(y) - min(y))/(max(y) - min(y))))
}
# autoria: https://www.r-bloggers.com/example-8-41-scatterplot-with-marginal-histograms/
scatterhist(df$peso,df$altura,xlab='',ylab='')
```



Histograma Conjunto

No caso do histograma univariado precisamos de duas dimensões para apresentar os dados, que vem diretamente da tabela de frequências. No histograma bivariado conjunto precisamos montar uma tabela de frequências que de fato é uma matriz onde as colunas e linha representam os intervalos e os elementos do interior da matriz as frequências. Nesse caso o gráfico terá três dimensões (duas para os intervalos e uma para a frequência):

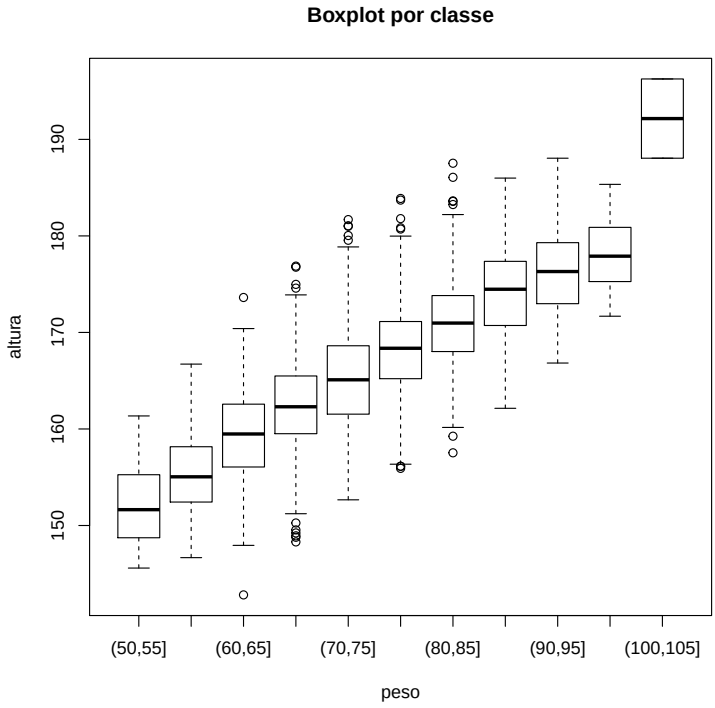
```
library(plot3D)
##  Crio os cortes:
x_c <- cut(df$peso, seq(50,105,by=5))
y_c <- cut(df$altura, seq(140,200,by=5))
##  Tabela de frequências:
z <- table(x_c, y_c)
##  Plot as a 3D histogram:
hist3D(z=z, border="black",xlab='peso',ylab='altura',zlab='freq')
```



Múltiplos Box-plot

Utilizando as funções `cut` para separar os pesos em grupos e `boxplot` podemos criar um boxplot para cada grupo. Essa técnica permite que absorvamos informação em cada grupo de pesos, como por exemplo a dispersão e os *outliers*.

```
df$intervalo_peso<-cut(df$peso,seq(50,105,by=5))
boxplot(altura ~ intervalo_peso,data = df,
             ylab='altura',
             xlab='peso',
             main='Boxplot por classe')
```



Note que a maior quantidade de *outliers* para altura está localizada nos intervalos centrais de peso.

Medidas univariadas para grupos

Um pergunta interessante que surge da análise de dados multivariadas com grupos é a de quais as medidas univaridas para cada grupo. Esse tipo de pergunta é recorrente na *análise de dados* pois nos traz intuições importantes de como os dados estão organizados e como descrevê-los. No nosso exemplo podemos querer responder qual a média de altura em cada grupo de peso.

Uma maneira de fazer isso, que pode ser replicada em inúmeros softwares é utilizar o laço `for`. O índice irá variar nos grupos e utilizaremos a função de interesse em cada grupo:

```
grupos<-unique(df$intervalo_peso)       # defino os grupos
for(i in 1:length(grupos)){
  ind<-df$intervalo_peso==grupos[i]   # indentifico os elementos pertencentes ao grupo
  ind_altura<-df$altura[ind]          # seleciono as alturas desses elementos
  print(mean(ind_altura))             # calculo a medida de interesse para o grupo
}
```

```
## [1] 159.2867
## [1] 152.3456
## [1] 162.5238
## [1] 155.5065
## [1] 165.2363
## [1] 168.3686
## [1] 171.0694
## [1] 174.2129
```

```
## [1] 176.3202
## [1] 178.1381
## [1] 192.155
```

No entanto o R como um software estatístico aplicado tem uma família de funções chamadas de apply que servem justamente para superar a necessidade de usar o laço for. Este tipo de laço tem algumas disvantagens quando estamos trabalhando com uma quantidade grande de dados, pois nele o processo é em fila de modo que para calcular a função média para o índice 3, ele precisa antes ter calculado no índice 1 e 2. A funções da família apply são mais eficientes pois são capazes de processar os comandos simultaneamente, algo que se assemelha a paralelização quando a CPU possui mais de um núcleo. Um outra disvantagem é a quantidade de caracteres que o usuário precisa escrever, o que deixa o código menos legível.[5]

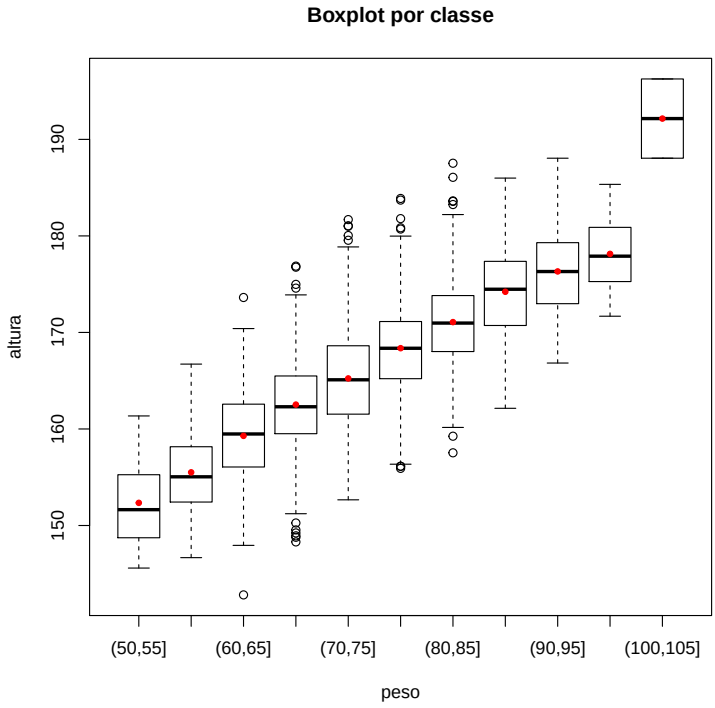
Essa tarefa pode ser realizada de diversas maneiras, podemos utilizar por exemplo as funções tapply ou aggregate:

```
# Qual a média da altura em cada grupo de pesos?
medias<-tapply(df$altura,df$intervalo_peso,mean)

aggregate(df$altura,list(df$intervalo_peso),mean)

##      Group.1        x
## 1    (50,55] 152.3456
## 2    (55,60] 155.5065
## 3    (60,65] 159.2867
## 4    (65,70] 162.5238
## 5    (70,75] 165.2363
## 6    (75,80] 168.3686
## 7    (80,85] 171.0694
## 8    (85,90] 174.2129
## 9    (90,95] 176.3202
## 10  (95,100] 178.1381
## 11 (100,105] 192.1550
```

Para inserir as médias no gráfico com múltiplos boxplot basta utilizar a função points:

```
boxplot(altura ~ intervalo_peso,data = df,
             ylab='altura',
             xlab='peso',
             main='Boxplot por classe')
points(medias,col=2,pch=20)
```

[5]Para saber mais sobre a família de funções apply acesse https://www.datacamp.com/community/tutorials/r-tutorial-apply-family



Medidas de Variação Conjuntas

Futuramente estudaremos modelos de regressão desenvolvidos para auxiliar no entendimento desse tipo de problema. Porém podemos apresentar duas medidas básicas de associação de dados. A covariância, como os dados variam conjuntamente e o coeficiente de correlação.

1. Covariância populacional

$$Cov(X,Y) = \sum_{i=1}^{n} \frac{(X_i - mean(X))(Y_i - mean(Y))}{(n)}$$

Pela equação acima uma condição necessária para realizarmos o cálculo da covariância é que os dois conjuntos de dados X e Y tenham o mesmo tamanho n.

2. Covariância amostral

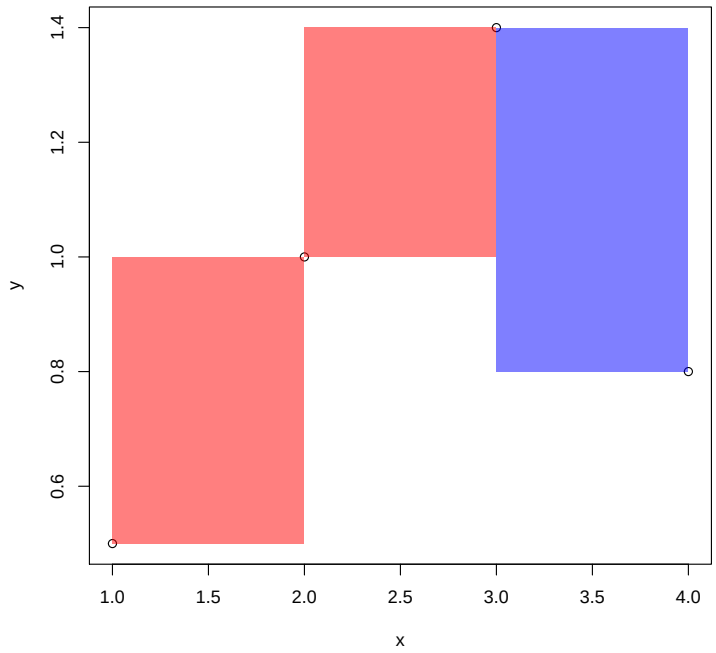
$$Cov(X,Y) = \sum_{i=1}^{n} \frac{(X_i - mean(X))(Y_i - mean(Y))}{(n-1)}$$

Como interpretar a covariância? Considere um par de dados (x, y). Vamos desenhar tal conjunto de dados em um gráfico de pontos.

```
x<-c(1,2,3,4)
y<-c(0.5,1,1.4,0.8)

plot(x,y)
```

x	y
1	0,5
2	1
3	1,4
4	0,8

```
rect(x[1],y[1],x[2],y[2],col=rgb(1,0,0,0.5),border = F)

rect(x[2],y[2],x[3],y[3],col=rgb(1,0,0,0.5),border = F)

rect(x[3],y[3],x[4],y[4],col=rgb(0,0,1,0.5),border = F)
```



Cada par no gráfico define um retângulo. Existem duas possibilidades complementares:

- Existe um ponto no canto inferior-esquerdo e outro no superior-direito (esse será o retângulo positivo);
- ou existe um ponto no canto inferior-direito e outro no cando superior-esquerdo (este será o retângulo negativo).

Nessa ilustração vamos considerar os retângulos positivos como sendo os vermelhos e os negativos como sendo os azuis. A covariância entre essas observações pode ser compreendida intuitivamente como sendo a área vermelha menos a área azul. [6]

No exemplo de altura e peso:

[6] Uma explicação mais detalhada sobre essa intuição pode ser encontrada neste link https://stats.stackexchange.com/questions/18058/how-would-you-explain-covariance-to-someone-who-understands-only-the-mean

```
cov(df$altura,df$peso)
```

```
## [1] 34.2241
```

3. Coeficiente de Correlação O coeficiente de correlação pode ser compreendido como uma medida de o quão próxima duas variáveis estão quando possuem relação linear. Uma correlação perfeita significa que os dados estão perfeitamente alinhados, um gráfico dos dois representa uma linha reta. Correlação nula diz a grosso modo que os dados estão muito espalhados de forma que uma linha não consegue representar o comportamento conjunto.

$$Corr(X,Y) = \frac{Cov(X,Y)}{Std.Deviation(X).Std.Deviation(Y)}$$

Perceba que o coeficiente de correlação é igual se considerarmos nos cálculos, a covariância e desvio padrão amostral ou populacional. Pois os valores os valores n ou $n-1$ serão "cancelados" na divisão. A correlação é diretamente proporcional à covariância, uma vez que o desvio-padrão é sempre um valor positivo. Ele pode ser compreendido como uma padronização da covariância, uma vez que essa depende da unidade de medida das variáveis e da escala, a correlação não depende, pertencendo sempre ao intervalo [0,1].
Aqui calculamos o coeficiente de correlação, com auxílio da função `with` que é utilizada para não precisar repetir o nome do `data.frame` em todos os argumentos:

```
with(df,cor(altura,peso))
```

```
## [1] 0.6961802
```

Suponha um exemplo em que a altura seja apresentada em metros ao invés de centímetros. Note que a correlação se mantém a mesma, ao ponto que a covariância se altera, uma vez que esta não é uma medida invariante a mudança na unidade de medida.

```
# Covariância
with(df,cov(altura,peso))
```

```
## [1] 34.2241
```

```
with(df,cov(altura/100,peso))
```

```
## [1] 0.342241
```

```
# Correlação
with(df,cor(altura,peso))
```

```
## [1] 0.6961802
```

```
with(df,cor(altura/100,peso))
```

```
## [1] 0.6961802
```

Utilizando essa função em um `data.frame` com mais do que duas colunas retorna uma matriz como resultado. Uma matriz simétrica de correlações. Para um exemplo vamos utilizar a base de dados com informações sobre motores de alguns modelos de carros `mtcars`.

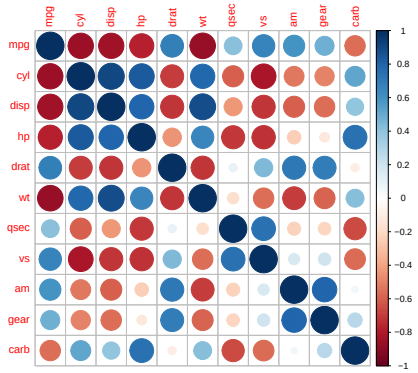
```
cor(mtcars)
```

```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
```

```
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##              qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

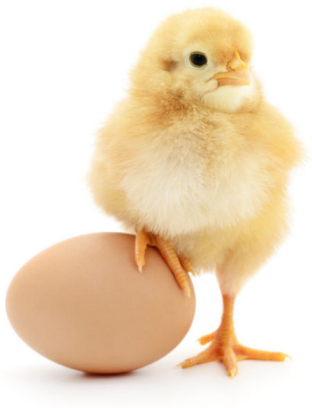
Um pacote que auxilia a visualização dos valores desta matriz é o `corplot`:

```
library(corrplot)

## corrplot 0.84 loaded

M <- cor(mtcars)
corrplot(M)
```



Exercício ChickWeight

Suponha que você foi contratado para prestar consultoria a um produtor de frangos que deseja conhecer a melhor dieta para engordar os seus animais.



Fonte: © Anatolii / Fotolia

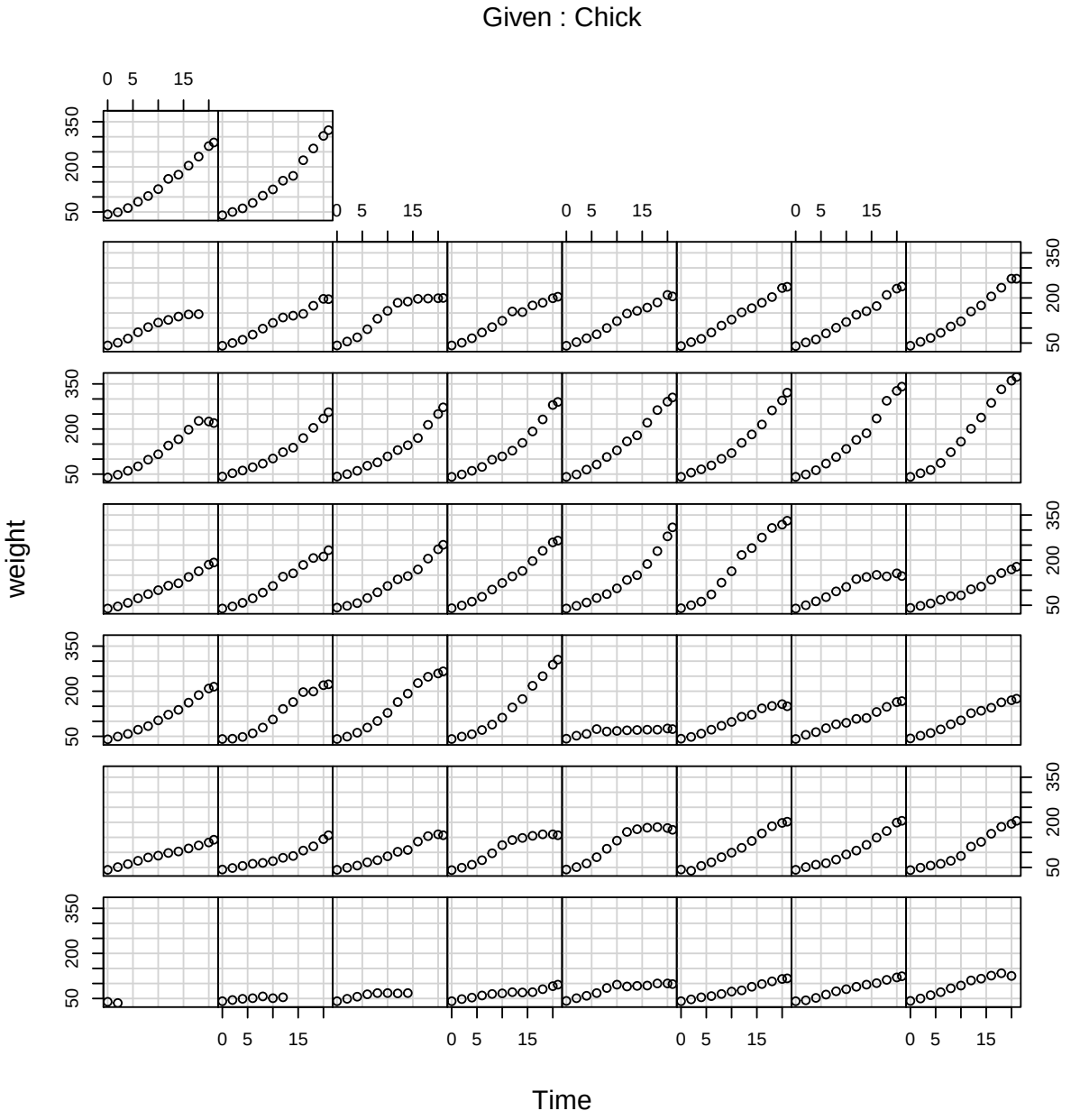
Para ter uma ideia geral do problema, atualmente com a melhora genética, o abate costuma ocorrer entre 28 e 42 dias de idade, quando o peso vivo do animal for de aproximadamente 1 kg. Desse modo é importante para o produtor conhecer qual a forma mais eficiente dos animais ganharem peso.

Utilizaremos uma base disponível no `R` que contém tais dados, `ChickWeight`.
Responda as seguintes perguntas;

1. Quantos frangos são monitorados ao todo?
2. Quantos frangos são monitorados em cada grupo de ração?
3. Qual a média de peso em cada dia?
4. Qual a média de peso em cada dia, dado o tipo de ração?
5. Em um único gráfico apresente quatros curvas contendo a informação obtida no item anterior. No eixo horizontal deve constar o tempo e no eixo vertical a média dos pesos.
6. Qual a correlação existente entre os dias e a média de peso para cada dieta?
7. Compare as medidas de centralidade e dispersão usando um boxplot para cada tipo de dieta no último dia.
8. Conclua se existe alguma ração que é melhor que as demais.

Para ilustração, segue abaixo um gráfico que apresenta a evolução do peso de cada frango, no tempo.

```
coplot(weight ~ Time | Chick, data = ChickWeight,
       type = "b", show.given = FALSE)
```


Given : Chick
weight
Time
0 5 15
50 200 350

Introdução à Probabilidade

Amostra e população

A população é o conjunto de todos os dados que se deseja estudar. Em um determinado estudo a amostra será sempre um subconjunto da população.

$$Sample \subseteq Population$$

Atenção pois estes termos dependem sempre do referencial. Uma amostra em um estudo pode ser considerada a população em outro e vice-versa. Devemos ter na cabeça a ideia de conjuntos. Suponha que temos dois estudos sobre o comportamento do consumidor:um a caráter regional e outro a caráter nacional. Se considerarmos para o primeiro a região como o estado de Alagoas, teremos que a população serão os consumidores de Alagoas. No entanto esse conjunto será considerado um subconjunto se considerarmos a segunda pesquisa que é a caráter nacional, nesse caso os consumidores de Alagoas serão uma amostra da população nacional.

Suponha que temos a seguinte população: os números de 1 a 5. Podemos amostrar essa população de duas maneiras distintas.

- **Sem reposição:** Nesse caso a ordem do sorteio importa. Escolhemos um elemento da população, em seguida retiramos esse elemento para sortear o seguinte.

```
populacao = c(1,2,3,4,5)
amostra = sample(populacao,size = 3,replace = FALSE)
amostra

## [1] 1 3 5
```

- **Com reposição:** Nesse caso sorteamos um elemento da população, "anotamos" o resultado, devolvemos este elemento à população e sorteamos novamente. Assim não alteramos a probabilidade de cada elemento da amostra ser sorteado.

```
populacao = c(1,2,3,4,5)
amostra = sample(populacao,size = 3,replace = TRUE)
amostra

## [1] 3 1 4
```

Esse comando pode ser utilizado com "string" também:

```
populacao = c('João','Maria','José','Henrique','Madalena','Margarida')
amostra = sample(populacao,size = 3,replace = TRUE)
amostra

## [1] "Maria"    "Henrique" "João"
```

Existe um comando no R que traz as letras em minúscula e maiúsculas.

```
letters

##  [1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j" "k" "l" "m" "n" "o" "p" "q"
## [18] "r" "s" "t" "u" "v" "w" "x" "y" "z"

LETTERS

##  [1] "A" "B" "C" "D" "E" "F" "G" "H" "I" "J" "K" "L" "M" "N" "O" "P" "Q"
```

```
## [18] "R" "S" "T" "U" "V" "W" "X" "Y" "Z"

sample(LETTERS,size = 4,replace = FALSE)

## [1] "C" "K" "N" "V"

sample(LETTERS,size = 3,replace = TRUE)

## [1] "S" "K" "D"
```

Aleatoriedade

A estatística trabalha fortemente com a ideia de aleatoriedade. Mas afinal de contas o que é aleatoriedade?

Existe uma teoria matemática cujo principal objetivo é estudar os fenômenos que são ditos aleatórios. Esse campo é a classificado como Teoria da Probabilidade. Muitos trabalhos são desenvolvidos nesse campo, porém nosso objetivo aqui é simplesmente passar os conceitos básicos para entrar de fato no que nos interessa, que é a estatística.

Os dados são gerados de alguma maneira pela natureza. Na estatística assumimos que na maioria das ocasiões a natureza é uma "caixa preta" chamada de processo gerador de dados. Esse processo é como se fosse uma função que recebe algumas variáveis de entrada **x** - também chamadas de inputs - e entrega variáveis de resposta **y** - também chamadas de outputs.

$$x \rightarrow natureza \rightarrow y$$

A análise de dados considera dois problemas:

- **Informação:** "Conhecer" o processo gerador chamado de natureza. Este é o principal interesse dos cientistas naturais e sociais. Eles possuem um modelo de como a natureza funciona e fazem medições coletando dados para verificar se de fato esse modelo é correto;
- **Previsão:** Observando futuros valores para **x** ser capaz de prever **y**, sem necessariamente conhecer a natureza. O cientista de dados se encontra mais nessa categoria. Tipicamente ele não tem um modelo de como a natureza funciona, mas encontra relações entre os dados e é capaz de fazer boas previsões.

Podemos modelar o processo gerador de dados, a natureza, de muitas inúmeras maneiras. Porém uma forma conveniente para fazer medições é utilizando a linguagem formal da matemática. Neste caso podemos separar os fenômenos em dois grandes conjuntos:

1. Determinístico - dado os valores dos inputs os outputs são determinados;
2. Aleatório - dados os valores dos inputs os outputs não são determinados, mas existe um conjunto de valores que eles podem atingir e uma função densidade de probabilidade cujo suporte são esses valores;

Os primeiros fenômenos são aqueles cujo resultado da ação é certa, a consequência de resultados é sempre possível de ser conhecida. Por exemplo, se soltarmos um objeto com massa não nula na superfície da Terra, ele irá cair devido à gravidade. Os fenômenos aleatórios são aqueles que podem ter resultados diversos, que podem ou não ocorrer. Se eu lançar uma moeda, o resultado que é a face virada para cima pode ser cara ou coroa, mas não pode ser as duas. A ideia da teoria matemática é que queremos ser capazes de medir o quão verossímil um

resultado (ou conjunto de resultados) é de ser realizado. No exemplo da moeda esperamos que os dois resultados sejam igualmente prováveis, por conta disso devemos assumir a mesma medida aos dois. A essa medida dá-se o nome de medida de probabilidade. Ela possui características menos gerais do que uma medida genérica. Como um dos exemplos dessas características podemos dizer que a medida total, de todos os eventos possíveis de serem realizados para determinado experimento, deve ser igual a unidade. Assim, se os eventos puderem ser descritos por meio de uma medida de probabilidade são ditos aleatórios.

A **modelagem estatística clássica** assume que os dados são produzidos por um processo gerador aleatório definido por uma função distribuição de probabilidade. Como veremos em breve, essa função é definida por **parâmetros** que são muitas vezes desconhecidos.

Algumas definições são importante para esclarecer as ideias:

Definição: Um **experimento** é um fenômeno que pode ser reproduzido sob as mesmas condições. Um experimento aleatório é aquele cujo resultado não pode ser previsto com certeza.

Definição: O **espaço amostral** é o conjunto de todos os resultados possíveis de um esperimento. Tipicamente associamos a letra grega omega maiúscula Ω para definirmos espaço amostral. Cada elemento do espaço amostral é um resultado possível e associamos omega minúsculo ωa ele.

Definição: Um **evento** é um subconjunto do espaço amostral, ou seja é um conjunto de resultados possíveis.

Por definição um evento pode ser a princípio qualquer coisa, pois é um conjunto de resultados de um experimento qualquer. Como exemplo temos: lançar uma moeda e ter como resultado cara, acertar na loteria, o IBovespa cair 50% em um dia, um meteoro cair na Terra e extinguir toda a vida existente no planeta, entre outros. No entanto trabalhar com eventos é algo matematicamente custoso e gostaríamos de associar números a cada evento. Números são mais tratáveis do que eventos, podemos usar a álgebra neles, por exemplo podemos somar $1 + 2 = 3$. No entanto é estranho somar eventos, o que é o resultado da soma de cara com coroa? Essa operação não está bem definida.

Quando associamos números ordenados a esses eventos e mantemos a probabilidade associada ao número chamamos essa nova relação (função) de variável aleatória. Isso é feito para facilitar o tratamento numérico da aleatoriedade e também para comparar sistemas aleatórios que à primeira vista podem ser diferentes. Por exemplo, comparar o lançamento de uma moeda, ocorrência de acidente, gols de um jogador, etc. Podemos trabalhar com os eventos de maneira mais generalizada e criar modelos probabilísticos com mais facilidade.

Para fixar ideias considere o seguinte exemplo clássico de lançamento de dois dados de seis lados. Seja Ω o espaço amostral deste esperimento:

$$\begin{aligned}\Omega = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ &(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ &(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ &(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ &(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\end{aligned}$$

$$(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}$$

Defina os seguintes eventos:

- **A**: Os dados somarem 3;
- **B**: Os dados somarem 6;
- **C**: O primeiro dado mostrar 1;
- **D**: O segundo dado mostrar 1;

Em notação de conjuntos, podemos escrever os eventos da seguinte forma:

- $A = \{(1,2),(2,1)\}$
- $B = \{(1,5),(2,4),(3,3),(4,2),(1,5)\}$
- $C = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\}$
- $D = \{(1,1),(2,1),(3,1),(4,1),(5,1),(6,1)\}$

Note que o evento A e B não ocorrem simultaneamente:

$$A \cap B = \emptyset$$

Dizemos que A e B são eventos **disjuntos**. No entanto os eventos A e C podem ocorrer simultaneamente se o resultado observado for $(1,2)$, pois:

$$A \cap C = \{(1,2)\}$$

Uma questão que surge naturalmente, já que estamos em um curso computacional, é se o computador consegue replicar experimentos aleatórios. A resposta é sim. Uma segunda questão é: ele faz isso de maneira aleatória? A resposta é não. O computador funciona com comandos claros, mesmo quando é pedido para ele gerar um número aleatório, ele precisa de uma rotina de comandos a seguir. Por conta disso conseguimos replicar exatamente os resultados de um experimento aleatório em diversos computadores. No "R" basta utilizar a mesma 'seed' antes dos códigos.

```
set.seed(1)   # no argumento você pode colocar qualquer número.
```

Considere o lançamento de um dado de seis lados

```
set.seed(1)
sample(x = 1:6,size = 1)

## [1] 2

set.seed(1)
sample(x = 1:6,size = 1)

## [1] 2

set.seed(1)
sample(x = 1:6,size = 1)

## [1] 2
```

Sempre que utilizarmos esse número no `set.seed()` o resultado será o mesmo, desse modo, como sabemos qual vai ser o resultado ele não é aleatório, porém pseudo-aleatório. Há um ramo da computação que se consagrou para criar algorítmos geradores de números aleatórios, eles ficaram tão bons que hoje simplesmente os ignoramos (mas eles estão lá gerando pseudo-aleatoriedade para nós).

Medida de probabilidade

Definição: Probabilidade é a medida de o quão verossímil um evento ocorrer. Definimos como sendo uma função $P(.)$ cujo domínio é o conjunto dos subconjuntos de Ω e o contradomínio é o intervalo $[0, 1]$. Aqui vale enfatizar que a probabilidade é tomada sobre o evento e não sobre os resultados possíveis. Tal que $\forall E \subseteq \Omega$:

- $P(E) \geq 0$
- $P(\Omega) = 1$
- $P(E \cup F) = P(E) + P(F)$, se $F \subseteq \Omega$ tivermos $E \cap F = \emptyset$ (E e F são disjuntos).

Segue abaixo as propriedades da medida de probabilidade:

- $P(\emptyset) = 0$
- $E, F \subseteq \Omega \Rightarrow P(E \cup F) = P(E) + P(F) - P(E \cap F)$
- $E \subseteq F \Rightarrow P(E) \leq P(F)$

Retomando ao nosso exemplo anterior do lançamento de dois dados vamos dizer que cada fazer tem a probabilidade de $1/6$ de ser observada. Desse modo cada combinação de pares ordenados tem a probabilidade de $1/36$ de ser observada. Podemos então calcular a probabilidade de cada um dos eventos listados:

- $P(A) = P(\{(1,2),(2,1)\}) = 2/36$
- $P(B) = P(\{(1,5),(2,4),(3,3),(4,2),(1,5)\}) = 5/36$
- $P(C) = P(\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\}) = 6/36$
- $P(D) = P(\{(1,1),(2,1),(3,1),(4,1),(5,1),(6,1)\}) = 6/36$

Seguindo das propriedades podemos calcular a propabilidade de ocorrência do evento A ou do evento C, istó é, a probabilidade de a soma dar 3 ou de cair 1 no primeiro dado:

$$P(A \cup C) = P(A) + P(C) - P(A \cap C)$$

Note que:

$A \cup C = \{(1,2),(2,1)\} \cup \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\} = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1)\}$

e $A \cap C = \{(1,2)\}$

Desse modo:

$$P(A \cup C) = P(\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1)\}) =$$
$$= P(\{(1,2),(2,1)\}) + P(\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\}) - P(\{(1,2)\})$$
$$= 2/36 + 6/36 - 1/36 = 7/36$$

Suponha agora o caso onde temos alguma informação sobre o resultado do esperimento, mas essa informação é incompleta. Podemos utilizar o conhecimento que temos para recalcular as probabilidades. Suponha que saibamos que o evento A tenha ocorrido, nessas condições qual a probabidade do evento C também ter ocorrido? Nossa intuição pode nos ajudar nessa resposta. Como o evento ocorreu, então sabemos que ou o resultado foi $(1, 2)$ ou foi $(2, 1)$. Se $(1, 2)$ foi observado então o evento C ocorreu, por outro lado se a observação foi $(2, 1)$, então o evento C não ocorreu. Uma vez que os dois resultados são equiprováveis podemos constatar que a probabidade de que C ocorreu dado que A ocorreu é de 50%. Podemos formalizar e generalizar esse resultado com a definição da probabilidade condicional.

Definição: A **probabilidade condicional** de F dado E, $P(F|E)$, é definida como:

$$P(E).P(F|E) = P(F \cap E)$$

Considerando que um evento E ocorra, $P(F|E)$ é a probabilidade do evento F também ocorrer.

Trazendo para o nosso caso $P(A) = 2/36$, $P(C) = 6/36$ e $P(A \cap C) = 1/36$, desse modo:

$$P(A).P(C|A) = P(A \cap C) \Leftrightarrow 2/36.P(C|A) = 1/36 \Leftrightarrow P(C|A) = 50\%$$

Caso um evento não adicione informação alguma sobre a realização de outro evento dizemos que eles são **independentes**. No nosso exemplo C e D são independentes, o fato de saber que o primeiro dado resultou em 1 não altera em nada nossa informação de qual foi o resultado do segundo dado. Mais formalmente.

Definição: Dois eventos A e B são independentes se:

$$P(A \cap B) = P(A).P(B) \Leftrightarrow P(A|B) = P(A) \Leftrightarrow P(B|A) = P(B)$$

Variável Aleatória

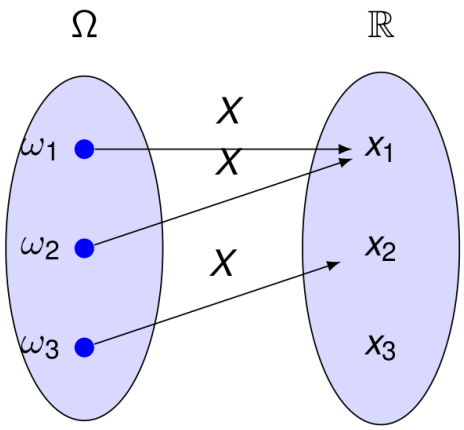
Definição: Uma variável aleatória (v.a.) é uma função $f : \Omega \to \mathbb{R}$ que associa a cada resultado possível de um experimento um número real.

Seja então $X : \Omega \to \mathbb{R}$ uma variável aleatória:

$$P(X = x_1) = P(\{\omega_1, \omega_2\})$$
$$P(X = x_2) = P(\{\omega_3\})$$
$$P(X = x_3) = P(\emptyset) = 0$$

Para fixar ideias considere outro exemplo clássico que é o lançamento de moedas. Seguindo a notação em inglês utilizaremos H para nos referirmos ao resultado cara (head) e T para nos referirmos a coroa (tail). Seja Ω o espaço amostral deste esperimento.

$$\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$$



Seja X a variável aleatória que representa o número de caras observadas no lançamento de duas moedas:
$X(\omega) = 0$ se $\omega = (T, T)$
$X(\omega) = 1$ se $\omega \in \{(T, H), (H, T)\}$
$X(\omega) = 2$ se $\omega = (H, H)$

A probabilidade associada a cada valor da v.a. é aquela associada ao evento que ela representa:

$$P(X = 0) = P(\{(T, T)\}) = \frac{1}{4}$$

$$P(X = 1) = P(\{(T, H), (H, T)\}) = \frac{1}{2}$$

$$P(X = 2) = P(\{(H, H)\}) = \frac{1}{4}$$

A esperança matemática de uma variável aleatória X, representada pelo símbolo $E[X]$ é a **média ponderada dos valores que essa v.a. pode assumir**, onde os pesos são as probabilidade.

Se X pode assumir os possíveis n valores $x_1, x_2, ..., x_n$, então a esperança é dada por:

$$E[X] = x_1.P(X = x_1) + x_2.P(X = x_2) + ... + x_n.P(X = x_n)$$

No exemplo da moeda temos:

$$E[X] = 0.P(X = 0) + 1.P(X = 1) + 2.P(X = 2)$$
$$E[X] = 0.\frac{1}{4} + 1.\frac{1}{2} + 2.\frac{1}{4} = 1$$

Distribuição

Uma distribuição de probabilidade é um modelo probabilístico, no sentido que é capaz de descrever completamente o comportamento de uma variável aleatória. Desse modo, quando estivermos falando em modelar um evento aleatório na grande parte dos casos estaremos falando em "escolher" uma distribuição de probabilidade que melhor descreve a v.a. associada ao evento. Grosso modo, a distribuição de probabilidade de uma variável aleatória X é uma função que associa a cada conjunto de possíveis valores de X uma probabilidade. Conforme apresentamos na subseção anterior, as v.a.s podem ser discretas ou contínuas. Abaixo seguem alguns exemplos de distribuição de probabilidade, que serão mais usadas no curso.

A distribuição de probabilidade e uma variável aleatória que assume um número finito de valores pode ser compreendida intuitivamente em um gráfico:

- No eixo horizontal são apresentados os valores que a v.a. pode assumir;
- No eixo vertical são apresentados as probabilidades associadas a cada um desses valores.

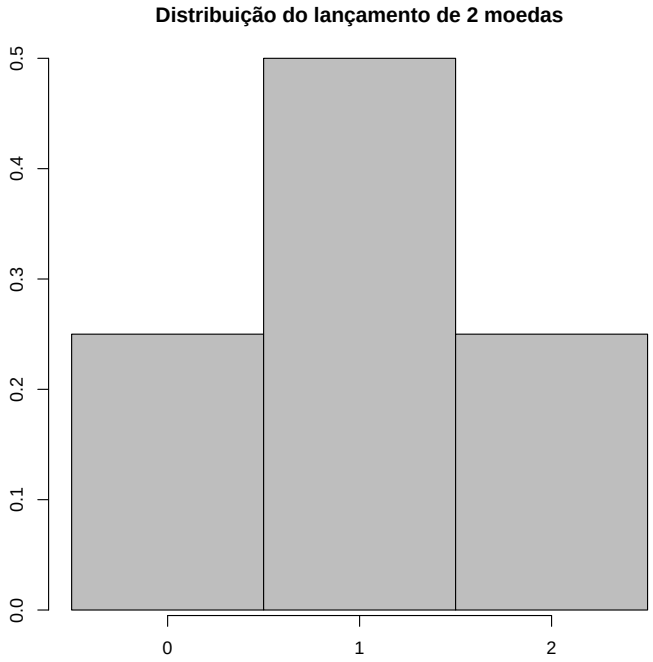
Abaixo segue o exemplo da distribuição de probabilidade de uma variável aleatória que representa o número de caras observadas em um lançamento de duas moedas. Essa variável aleatória pode ser descrita como uma função:

$$X : \Omega \to \mathbb{R}$$

onde:

$X(\{coroa, coroa\}) = 0$ e $P(X(\{coroa, coroa\})) = 0.25$
$X(\{cara, coroa\}) = X(\{coroa, cara\}) = 1$ e $P(X(\{cara, coroa\})) = 0.5$
$X(\{cara, cara\}) = 2$ e $P(X(\{cara, cara\})) = 0.25$

```
b<-barplot(c(0.25,0.5,0.25),space = F,main ='Distribuição do lançamento de 2 moedas')
axis(side=1,at = b,labels = c(0,1,2))
```



Podemos generalizar essa variável aleatória para n lançamentos de moeda, ou qualquer fenômeno que possa ser representado por uma sequência de resultados binários em uma variável aleatória chamada de **binomial**. A seguir vamos apresentar algumas definições de variáveis aleatórias que são as mais comuns.

Distribuição de Bernoulli

Esse é o modelo probabilístico mais básico e recorrentemente utilizado. Ele é usado quando queremos modelar qualquer evento que possa ser caracterizado por dois únicos resultados: sucesso, associado ao número 1 e, fracasso associado ao número 0.

Suporte: $\{0, 1\}$
Parâmetros: $\{p\}$
Função massa de probabilidade:

$$f(k; p) = Pr(X = k) = p^k.(1 - p)^{1-k} \quad k \in \{0, 1\} \tag{1}$$

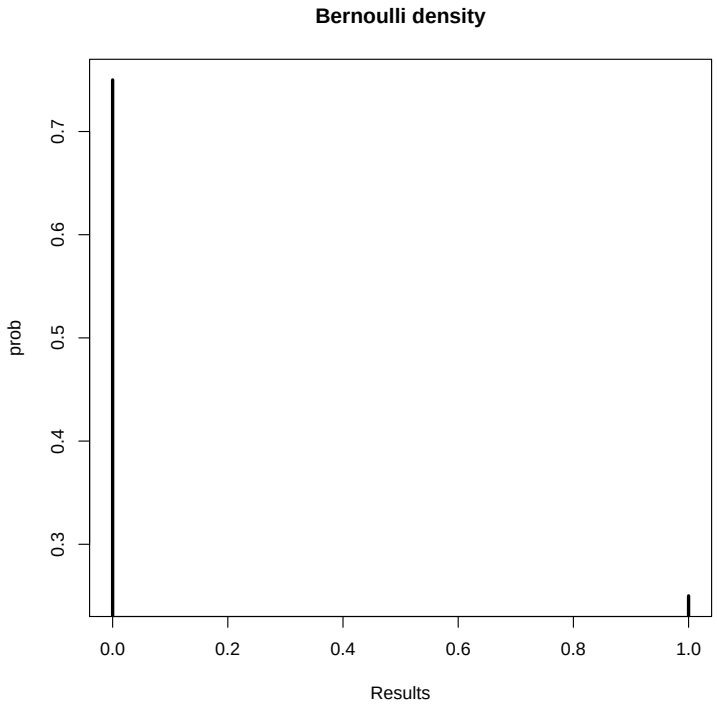
Tomando a esperança (média populacional, considerando a distribuição como o modelo):

$$E(X) = \int k.f(k; p)dk = \int k.p^k.(1 - p)^{1-k}dk = 1.p^1.(1 - 1)^{1-1} + 0.p^0.(1 - p)^{1-0} = p = P(X = 1) \tag{2}$$

$$E(X) = P(X = 1) \tag{3}$$

Desse modo temos que a média da Bernoulli representa a probabilidade do evento sucesso (1) ocorrer.
Gerando uma amostra com 10 obervações de uma distribuição Bernoulli, com parâmetro $1/4$.

```r
n <- 10       # número de observações
p <- 1/4      # defino o parâmetro para a probabilidade

sample(0:1, size = n, prob = c(1-p,p),replace = TRUE)

##  [1] 0 0 1 0 1 1 0 0 0 0

# Bernoulli
plot(c(0,1),c(1-p,p),
     main = "Bernoulli density",
     type = 'h',
     lwd = 3, ylab = 'prob',xlab = 'Results')  # somente 2 pontos
```



Distribuição Binomial

O modelo binomial é associado diretamente ao de Bernoulli, no sentido de que o resultado possível da variável aleatória representa o número de sucesso (k) em um número de tentativas (n).

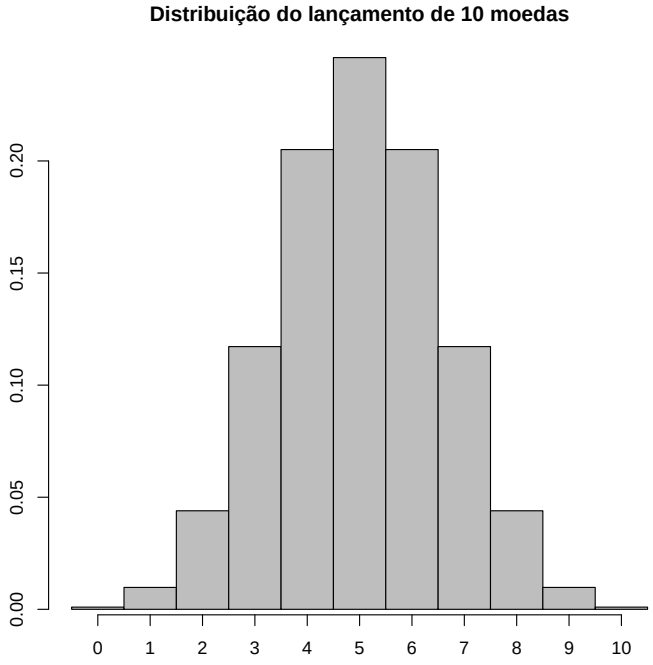
Suporte: $\{0, 1, ..., n\}$
Parâmetros: $\{n, p\}$
Função massa de probabilidade:

$$f(k; n, p) = Pr(X = k) = nkp^k(1 - p)^{n-k} \tag{4}$$

Vamos exemplificar o lançamento de 10 moedas não viciadas. Seja X a variável aleatória que representa o número de caras. Desse modo X tem uma distribuição binomial. Para fixar ideias responda: Qual a probabilidade

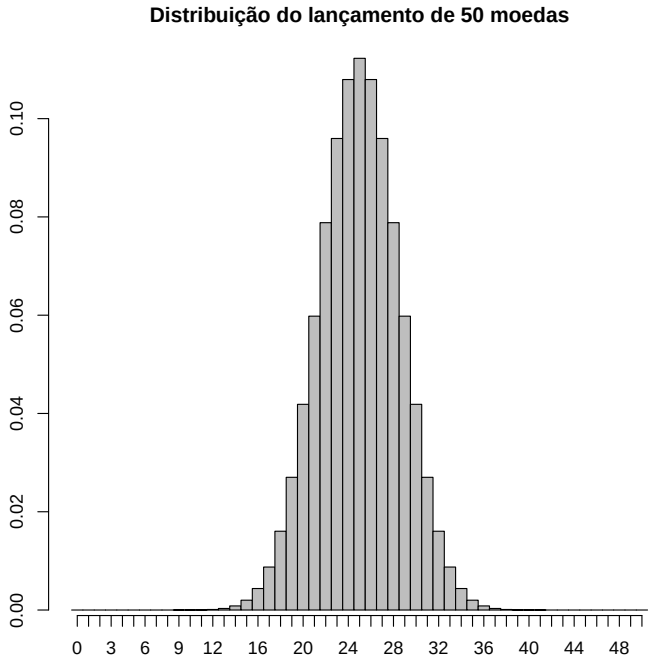
de que em 10 lançamentos 5 sejam cara?

```
# Binomial
p<-0.5
n <- 10
k <- seq(0, n, by = 1)
y <- dbinom(k, n, p)
b<-barplot(y,space = F,main ='Distribuição do lançamento de 10 moedas')
axis(side=1,at = b,labels = k)
```



Abaixo geramos a distribuição do lançamento de 50 moedas.

```
# Binomial
p<-0.5
n <- 50
k <- seq(0, n, by = 1)
y <- dbinom(k, n, p)
b<-barplot(y,space = F,main ='Distribuição do lançamento de 50 moedas')
axis(side=1,at = b,labels = k)
```



Conforme o número de lançamentos aumenta a distribuição da binomial vai ganhando cada vez mais o formato de sino. Este é um resultado teórico importante na estatística que é a convergência da binomial para uma variável aleatória Normal, que apresentaremos em breve. Seja X uma variável aleatória com distribuição binomial e parâmetros p e n, se $n \to \infty$ temos que:

$$\frac{X - n.p}{\sqrt{n.p.(1-p)}} \to Normal(0,1)$$

Distribuição Normal

A distribuição Normal é de longe a mais utilizada em estatística. Conforme veremos adiante, ela tem propriedades desejáveis para grandes amostras.

Suporte: $(-\infty, +\infty)$

Parâmetros:

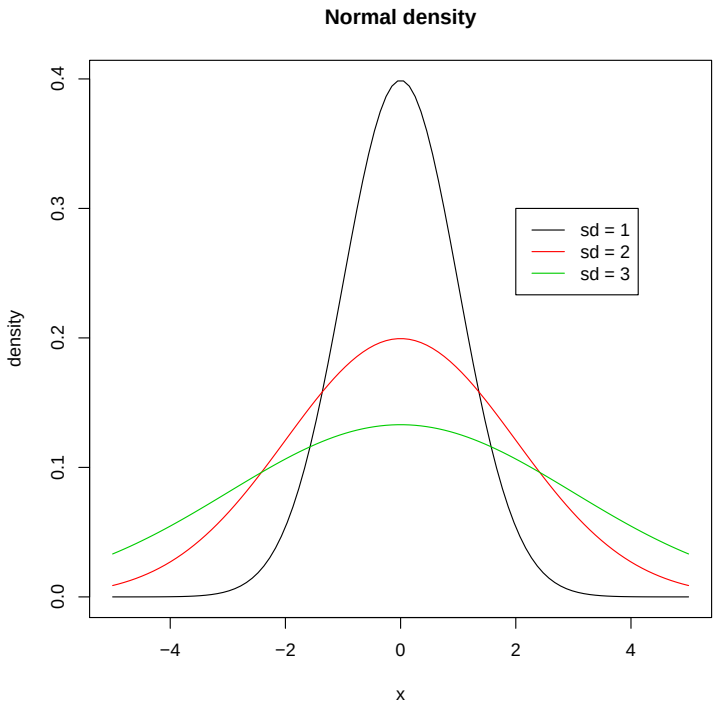
- μ: é o parâmetro que representa a média da distribuição (o valor no eixo horizontal associado ao topo do sino);
- σ: é o parâmetro que representa o desvio padrão (a dispersão dos dados, ou o tamanho da boca do sino).

Função densidade de probabilidade:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2.\pi\sigma^2}}.exp\Big(-\frac{(x-\mu)^2}{2.\sigma^2}\Big) \tag{5}$$

Observe a distribuição normal para diversas variâncias:

```r
# Normal
mu<-0
sd1<-1
sd2<-2
sd3<-3
n<-100
x <- seq(-5,5, length.out = n)
y1 <- dnorm(x, mu, sd1)
y2 <- dnorm(x, mu, sd2)
y3 <- dnorm(x, mu, sd3)
plot (x, y1,ylab = "density", main = "Normal density",type = 'l')
lines(x,y2,col=2)
lines(x,y3,col=3)
legend(x = 2,y = 0.3,
       legend = c('sd = 1','sd = 2','sd = 3'),
       lty = c(1,1),
       col = c(1,2,3))
```



Diferentemente das distribuições acima, a Normal é uma distribuição com suporte infinito e não enumerável. Nesse sentido a probabilidade associada a cada ponto deve ser zero, caso contrário, a probabilidade de todo o suporte seria maior do que 1. A probabilidade para esse tipo de distribuição é calculada para um intervalo. Qual a probabilidade de um evento associado ao intervalo $[a, b]$ ocorre?

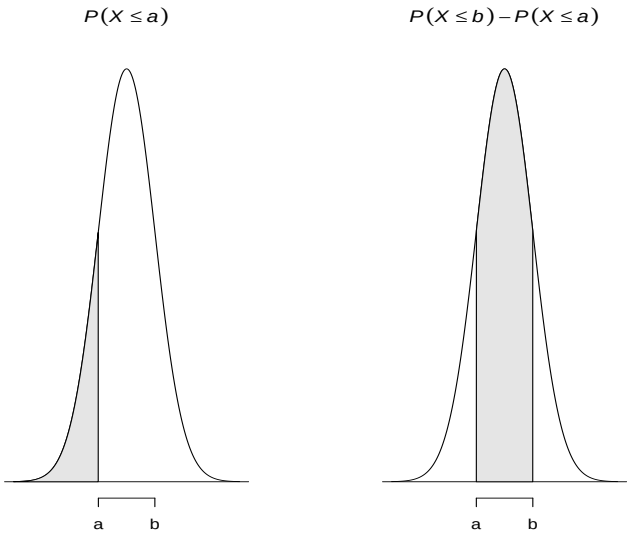
$$P(a \leq X \leq b) = \int_a^b f(y; \mu, \sigma) dy \tag{6}$$

Aproveitando a deixa, definimos a função de probabilidade acumulada de uma v.a. com parâmetros θ da seguinte maneira:

$$F(x;\theta) = P(X \leq x) = \int_{-\infty}^{x} f(y;\theta)dy \tag{7}$$

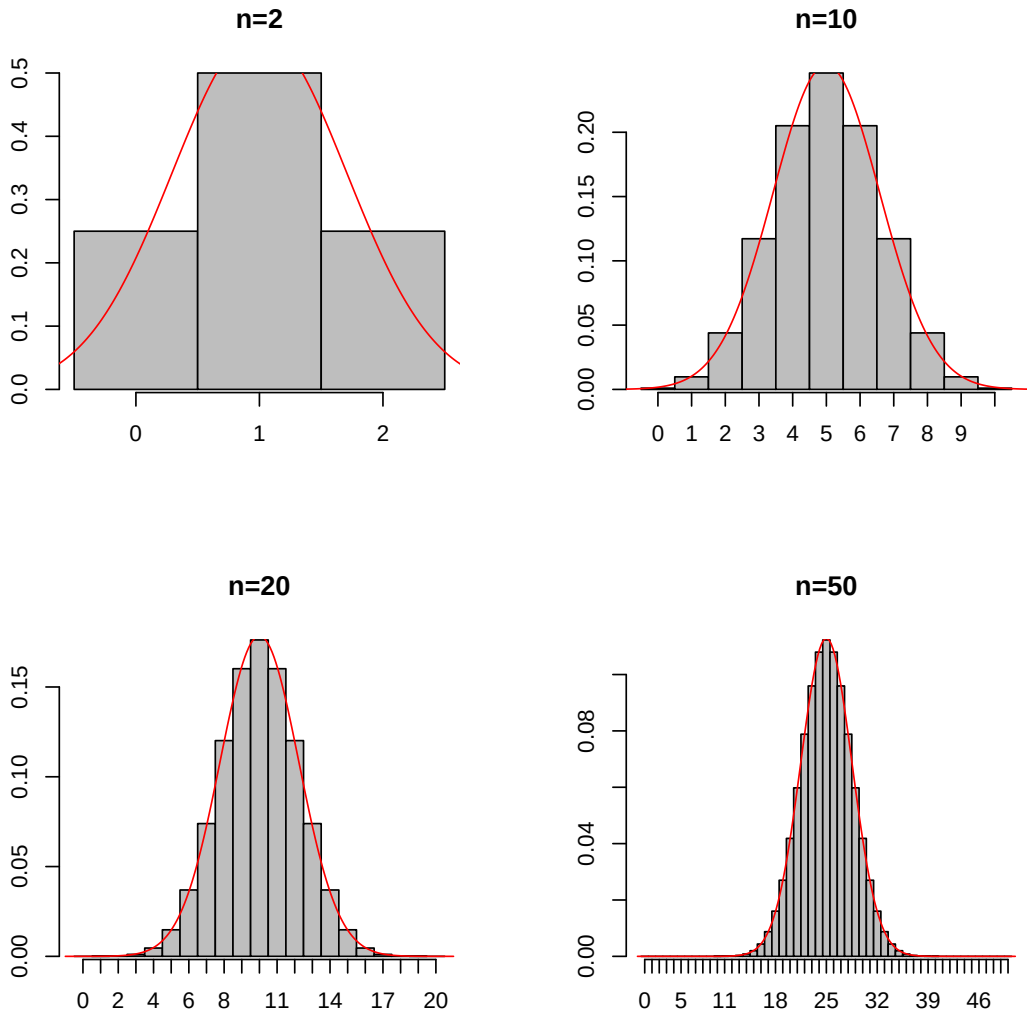
A seguir apresentamos um gráfico para ilustrar como a probabilidade de uma distribuição contínua é calculada.

```
par(mfrow = c(1,2))

x <- seq(-4, 4, length = 1000)
y <- dnorm(x)
plot(x, y, axes = FALSE,type = 'l', xlab = '', ylab = '',
main = expression(italic(P(X<=a))))
abline(h=0)
x1 <- x[x <= -1] ; y1 <- dnorm(x1)
x2 <- c(-4, x1, x1[length(x1)], -4) ; y2 <- c(0, y1, 0, 0)
polygon(x2, y2, col = 'grey90')
axis(1, at = c(-1, 1), font = 8, labels = c('a', 'b'))

plot(x, y, axes = FALSE,type = 'l', xlab = '', ylab = '',
main = expression(italic(P(X<=b)-P(X<=a))) )

abline(h=0)
x1 <- x[-1<= x & x<= 1] ; y1 <- dnorm(x1)
x2 <- c(-1, x1, x1[length(x1)], -1) ; y2 <- c(0, y1, 0, 0)
polygon(x2, y2, col = 'grey90')
axis(1, at = c(-1, 1), font = 8, labels = c('a', 'b'))
```



Abaixo segue um exemplo da convergência da Binomial para a Normal quando o tamanho da amostra cresce. Observe que para os valores mais baixos de n a diferença das duas distribuições é marcante. No entanto, com o crescimento do tamanho da amostra a convergência é alcançada.

```r
# Normal
par(mfrow=c(2,2))
p<-0.5
n <- 2
k <- seq(0, n, by = 1)
y <- dbinom(k, n, p)
b<-barplot(y,space = F,main ='n=2')
axis(side=1,at = b,labels = k)
lines(seq(min(k)-1,max(k)+1,length.out = 200)+0.5,
      dnorm(seq(min(k)-1,max(k)+1,length.out = 200),
            mean = n*p,sd = sqrt(n*p*(1-p))),col=2)

p<-0.5
n <- 10
k <- seq(0, n, by = 1)
y <- dbinom(k, n, p)
b<-barplot(y,space = F,main ='n=10')
axis(side=1,at = b,labels = k)
lines(seq(min(k)-1,max(k)+1,length.out = 200)+0.5,
      dnorm(seq(min(k)-1,max(k)+1,length.out = 200),
            mean = n*p,sd = sqrt(n*p*(1-p))),col=2)

p<-0.5
n <- 20
k <- seq(0, n, by = 1)
y <- dbinom(k, n, p)
b<-barplot(y,space = F,main ='n=20')
axis(side=1,at = b,labels = k)
lines(seq(min(k)-1,max(k)+1,length.out = 200)+0.5,
      dnorm(seq(min(k)-1,max(k)+1,length.out = 200),
            mean = n*p,sd = sqrt(n*p*(1-p))),col=2)

p<-0.5
n <- 50
k <- seq(0, n, by = 1)
y <- dbinom(k, n, p)
b<-barplot(y,space = F,main ='n=50')
axis(side=1,at = b,labels = k)
lines(seq(min(k)-1,max(k)+1,length.out = 200)+0.5,
      dnorm(seq(min(k)-1,max(k)+1,length.out = 200),
            mean = n*p,sd = sqrt(n*p*(1-p))),col=2)
```



Observe que esse gráfico lembra um histograma.
De fato existe associação entre esses dois conceitos.

- A **distribuição** representa a chance de observar valores que pertencem a certo intervalo.
- O **histograma** representa a frequência de valores observados que pertencem a determinado intervalo. Pode ser pensando como uma distribuição empírica.

Quanto maior o tamanho da amostra, mais próximo o histograma será da distribuição teórica. (Se você lançar uma moeda muitas vezes irá observar que aproximadamente metade delas deu cara e metade coroa)

Simulação

A partir de um modelo probabilístico adotado, representado por uma variável aleatória com uma distribuição de probabilidade, podemos gerar dados simulados. Vamos reproduzir alguns exemplos. No início dessa seção vimos como amostrar a partir de coleção de dados. Utilizando o comando `sample` e dois valores possíveis podemos reproduzir um modelo **Bernoulli**. Nos próximos exemplos vamos verificar algumas especificações de modelos e como gerar os dados simulados a partir deles:

Exemplo 1: Simular o lançamento de 5 moedas.

```
set.seed(1)
n<-5
sample(x = c('Cara','Coroa'),size = n,replace = T)

## [1] "Cara"  "Cara"  "Coroa" "Coroa" "Cara"
```

Seja a v.a. de Bernoulli que associa 1 quando o resultado é Cara e 0 quando o resultado é Coroa.

$$X = (X_1, X_2, ..., X_5)$$

$$x = (x_1, x_2, x_3, x_4, x_5) = (1, 1, 0, 0, 1)$$

Exemplo 2: Simular o resultado de caras (1) em um lançamento de 10 moedas, utilizando a binomial.

```
set.seed(1)
n<-10
p<-0.5
rbinom(1,size = n,prob = p)

## [1] 4
```

A frequência de Caras na amostra é dado por $\hat{p} = x/n = 4/10$ Chamamos $\hat{p} = 0.4$ de **estimativa** do **parâmetro** $p = 0.5$ Aqui conhecemos o parâmetro, pois ele foi necessário para gerar os dados simulados, porém esse não será o caso na grande parte das pesquisas não simuladas.

Note que $\hat{p}$ está próximo de p, porém o quão bom ele é?

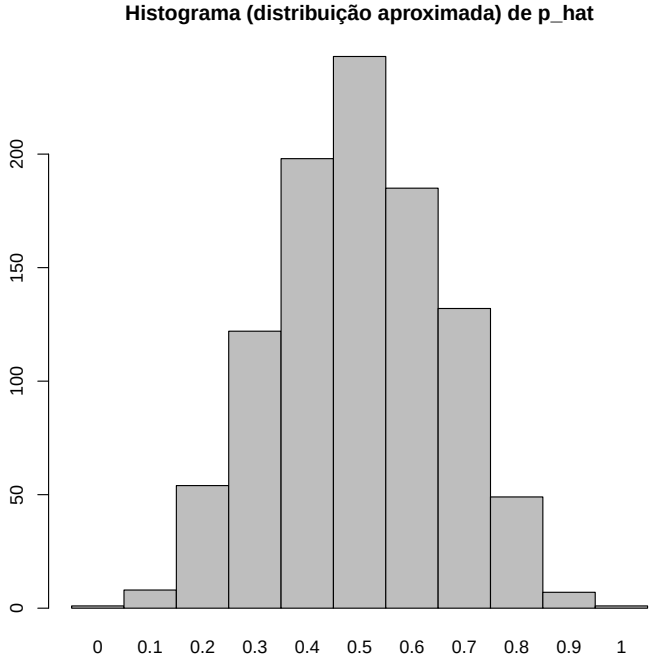
O valor $\hat{p} = 0.4$ depende dos valores observados da amostra. Caso a amostra ainda não seja observada, sendo tratada como um conjunto ordenado de variáveis aleatórias, teremos que $\hat{P}$ (P chapéu maiúsculo) é uma variável aleatória, chamada de **estimador**.

- $\hat{P} = \frac{X}{n}$ é um **estimador**
- $\hat{p} = \frac{x}{n}$ é uma **estimativa**

Exemplo 3: Simular o resultado de caras (1) em 1000 lançamentos de 10 moedas, utilizando a binomial.

```
set.seed(2)
n<-10
p<-0.5
s<-rbinom(1000,size = n,prob = p)
p_hat<-s/n
t<-table(p_hat)
t
```

```
## p_hat
##   0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9   1
##   1   8  54 122 198 243 185 132  49   7   1

barplot(t,space = F,main = 'Histograma (distribuição aproximada) de p_hat')
```



O resultado mais observado de $\hat{p}$ é exatamente o valor verdadeiro p. No entanto $\hat{p}$ também assume diversos outros valores, entre eles valores extremos como 0 e 1. Mesmo com uma moeda não viciada, existe a chance de observar cara em todos os 10 lançamentos, apesar de ser pouco provável. Assim como observar 0 caras na mesma quantidade de lançamentos.

No exemplo anterior simulamos 1000 amostras que representam 10 lançamentos de moeda. No entanto se tivermos somente uma amostra, poderemos afirmar algo a respeito do parâmetro verdadeiro? Iremos verificar como responder tal pergunta na próxima seção, com o que é chamado de teste de hipótese.

Teste de hipótese

Ideia básicas

Vamos motivar a ideia de teste de hipótese por meio de um exemplo. Suponha que Joaquim está sendo acusado de ser trapaceiro em um jogo de moeda onde ele ganha um real se der cara e perde um real se der coroa.

Fonte: http://www.pbs.org/newshour/making-sense/cant-decide-just-flip-coin/

Ele tem uma única chance de provar sua inocência lançando a moeda 100 vezes em frente do juíz.

O teste para julgar Joaquim é o seguinte:
H_0: Joaquim é inocente
H_1: Joaquim não é inocente.
Caso $\hat{p}$ seja maior que 0.5 Joaquim será condenado, caso $\hat{p}$ seja menor ou igual a 0.5 Joaquim será considerado inocente.

Note pela simulação feita na seção anterior que existe uma chance de Joaquim ser condenado mesmo sendo inocente, do mesmo modo que existe a chance de Joaquim ser inocentado mesmo sendo trapaceiro. Estes são erros que recebem nomes especiais na estatística.

Tipos de erro:
- **erro do tipo I**: rejeitar a hipótese nula quando ela é verdadeira;
  (Culpar um inocente)
- **erro do tipo II**: aceitar a hipótese nula quando ela é falsa.
  (Deixar livre um culpado)

Qual a probabilidade de cometer o erro do tipo I no exemplo anterior? Ou seja, qual a probabilidade de Joaquim ser considerado culpado mesmo sendo inocente. Joaquim será sempre culpado $\hat{p} > 0.5$, assim queremos calcular qual a probabilidade desse evento ocorrer, ou seja $P(\hat{p} > 0.5)$. Chamamos o valor da estatística de teste que é determinante para saber se rejeitamos ou não a hipótese nula de valor crítico ($\hat{p}_c$).

Já vimos que a Binomial padronizada converge para uma Normal(0,1), desse modo podemos calcular essa probabilidade aproximada em termos da normal padrão. Para isso basta encontrar o valor crítico associada a normal. Pela equação anterior apresentada:

$$Z_c = \frac{n \times \hat{p}_c - n \times p}{\sqrt{n \times p \times (1-p)}} = \frac{100 \times \hat{p}_c - 100 \times 0.5}{\sqrt{100 \times 0.5 \times (1-0.5)}} = 0 \tag{8}$$

A probabilidade de $\hat{p}$ ser maior que 0.5 é aproximadamente igual à probabilidade de Z, uma variável aleatória normal padrão (com média 0 e desvio-padrão 1), ser maior que 0

$$P(\hat{p} > 0,5) = P(Z > 0) = ? \tag{9}$$

A função `pnorm` do `R` nos entrega a probabilidade acumulada até certo ponto da normal com os parâmetros definidos. Assim `pnorm(Zc,mean = 0,sd = 1)`$= P(Z \leq Zc)$. Para calcular a probabilidade de ser maior, basta calcular o complementar, `1 - pnorm(Zc,mean = 0,sd = 1)`$= P(Z > Zc)$.

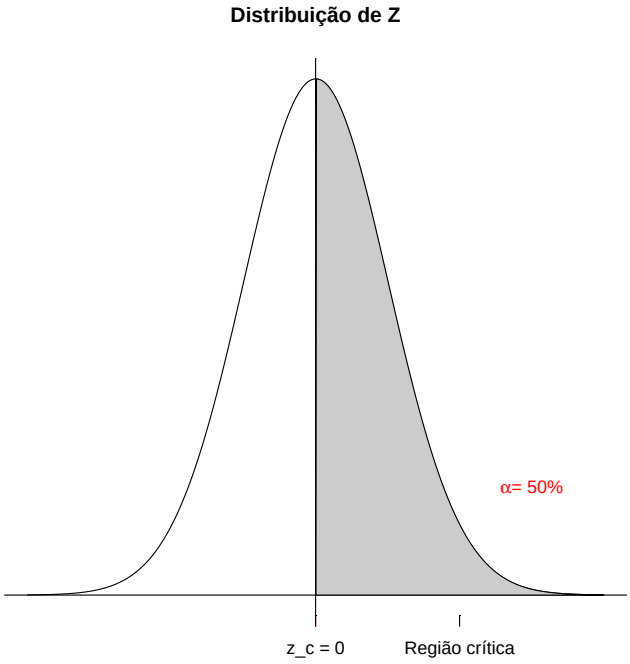
```
1 - pnorm(q = 0,mean = 0,sd = 1)

## [1] 0.5
```

$$P(Z > 0) = 50\% \tag{10}$$

Assim a probabilidade de condenar Joaquim, mesmo sendo inocente é 50%!

Abaixo aprentamos a representação gráfica dessa probabilidade. Ela é a área representada pelo cinza escuro. Como equivale a metade da área abaixo da distribuição representa 50% da probabilidade.

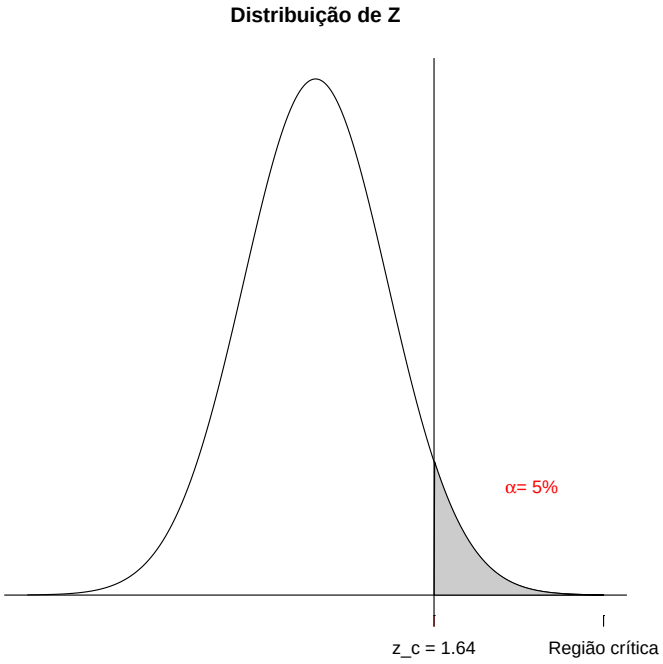
```
x <- seq(-4, 4, length = 1000)
y <- dnorm(x)
plot(x, y, axes = FALSE,type = 'l', xlab = '', ylab = '',
     main='Distribuição de Z')
abline(h=0)
x1 <- x[x >= 0] ; y1 <- dnorm(x1)
x2 <- c(4, x1, x1[length(x1)], 4) ; y2 <- c(0,0, y1, 0)
polygon(x2, y2, col = 'grey80')
axis(1, at = 0, labels='',col=2)
abline(v = 0)
axis(1, at = 0, labels = 'z_c = 0')
axis(1, at = 2, labels = 'Região crítica')
text(3,0.07,labels = expression(paste(alpha,'= 50%')),pos=3,col=2)
```



- A probabilidade de cometer o **erro do tipo I** é chamado de **nível de significância** de um teste, representado pela letra grega α.)
- O complementar dessa probabilidade, ou seja $1 - \alpha$ é chamado de **nível de confiança** de um teste.

A probabilidade de 50% de condenar um inocente é muita alta. Para superar tal problema, o direito parte da **hipótese de presunção de inocência** em um julgamentos e somente com **fortes evidências contrárias** tal hipótese é rejeitada. A estatística busca ser parcimoniosa assim como o direito de forma a trabalhar com valores pequenos (tipicamente 10%, 5% e 1%) para a probabilidade de condenar um inocente, isto é o nível de significância.

Isso ocorre porque tanto no direito e quanto na estatística condenar um inocente é mais grave do que deixar livre um culpado.



Abaixo segue a sequência de passos para um teste de hipótese genérico:

1. **Definir a hipótese nula**
   Este passo é fundamental pois representa justamente a hipótese que está sendo testada. Uma vez definida a hipótese nula conseguimos encontrar qual a hipótese alternativa bastando considerar o evento que é complementar. Um exemplo: Se H_0 representa a hipótese de que $\mu = 7$ então concluímos que H_1 representa $\mu \neq 7$.
2. **Definir um nível de significância**
   Aqui é importante lembrar que o nível de significância α é definido antes de a estatística de teste ser calculada. Teoricamente temos que definir esse valor antes de irmos para os dados.
3. **Encontrar a região crítica**
   A partir do nível de significância definido no item anterior encontra-se qual a região crítica, ou seja, para quais valores assumidos pela estatística de teste podemos rejeitar a hipótese nula.
4. **A estatística de teste é calculada**
   Somente depois dos demais passos serem definidos a estatística é calculada.
5. **Conclusão do teste**
   Observado o valor da estatística de teste rejeito ou não a hipótese nula.
   - Se a estatística estiver dentro da região crítica, rejeito H_0;
   - Se a estatística não pertencer a região crítica, não rejeito H_0.

Reformulando o exemplo anterior considerando agora um nível de significância igual a 5%.

1. Defino $H_0 : p \leq 0.5$.
2. Defino $\alpha = 5\%$;
3. Região crítica;
   $P(Z > z_c) = 0.05$

Resolvendo para z_c

```
round(qnorm(p = 0.95,mean = 0,sd = 1),2)

## [1] 1.64
```

A região crítica nesse caso são todos os valores maiores que 1.64 isto é $(1.64, \infty)$

4. Calcular a estatística de teste; Isso depende da amostra.
   (a) Vamos simular o lançamento de Joaquim, supondo nesse primeiro caso que de fato a moeda dele **não é viciada**, ou seja $p = 0.5$.

```
set.seed(321)
n<-100
p<-0.5
s<-rbinom(1,size = n,prob = p)
p_hat<-s/n
p_hat

## [1] 0.59
```

   Como $\hat{p} = 0.59$, então $z = (59 - 50)/\sqrt{100 \times 0.5 \times 0.5} = 1.8$ Nesse caso $1.8 \in (1.64, \infty)$ e a hipótese nula é rejeitada, ou seja **Joaquim é tido como culpado**.

   (b) Considere agora uma simulação em que Joaquim tem de fato uma **moeda viciada**, com $p = 0.55$

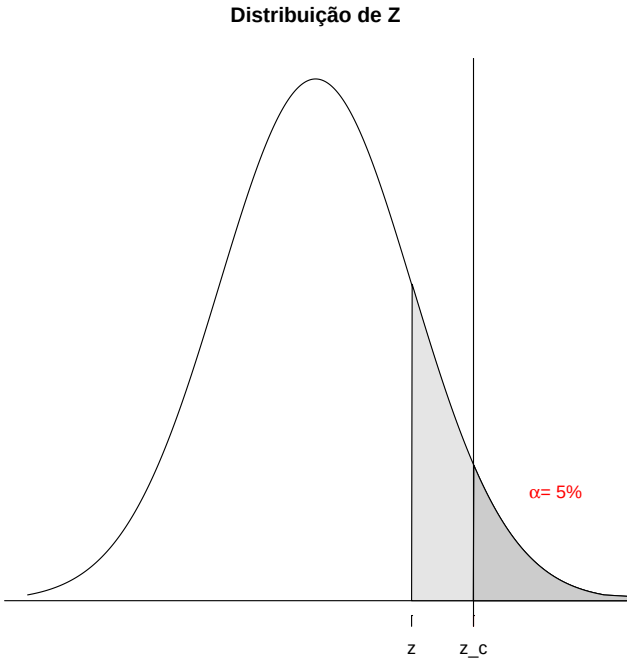
```
set.seed(321)
n<-100
p<-0.55
s<-rbinom(1,size = n,prob = p)
p_hat<-s/n
p_hat

## [1] 0.46
```

   Como $\hat{p} = 0.46$, então $z = (46 - 50)/\sqrt{100 \times 0.5 \times 0.5} = -0.8$ Nesse caso $-0.8 \notin (1.64, \infty)$ e a hipótese nula não é rejeitada, ou seja **Joaquim é tido como inocente**.

Os resultados acima são um pouco decepcionantes pois os resultado dos testes contrariam a realidade. Estes são exemplos do que pode ocorrer na prática quando consideramos amostras relativamente pequenas que é o caso de 100 lançamentos. No entanto quando temos muitos dados a probabilidade de observarmos estimativas muito distantes dos parâmetros verdadeiros é pequena e assim é cada vez mais difícil cometermos esses tipos de erro.

Valor p

A unidade de medida do valor p é a mesma da medida de probabildade, isto é, temos que o valor p está contido no intervalo $[0, 1]$. Isso posto, qual a interpretação do valor p? Como vimos acima um teste de hipótese é construído sobre uma estatística de teste, que por sua vez é uma variável aleatória com uma determinada distribuição. Nos exemplos acima nossa estatística de teste tinha distribuição Normal(0,1). O valor p é calculado como sendo a probabilidade de observarmos valores mais extremos para a estatística de teste do que aquele observado, considerando H_0 verdadeira. Abaixo o valor p é identificado pela área cinza.



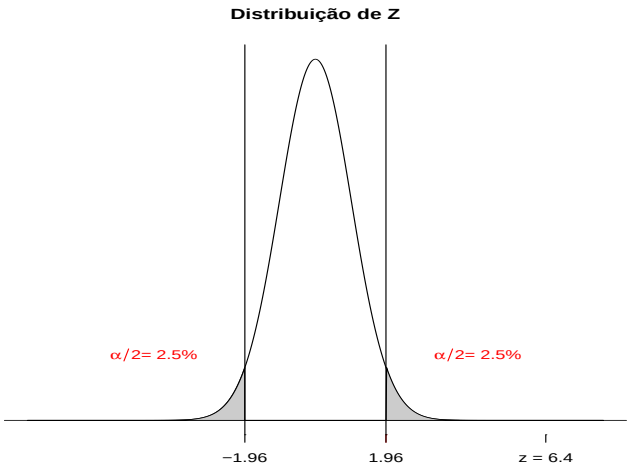
Existe uma relação direta entre o valor p e o nível de significancia, α, do teste de hipótese. Estipulado que a probabilidade de cometer o erro do tipo I seja α, calculamos a estatística de teste. Caso o valor p associado a esta estatística seja maior do que α isso significa que a estatística está dentro do intervalo de confiança, desse modo não rejeitamos a hipótese nula. Caso o valor p seja menor do que α então a estatística de teste está dentro da região crítica, desse modo rejeitamos a hipótese nula. Resumindo:

- Se valor $p > \alpha$, então não rejeito H_0.
- Se valor $p < \alpha$, então rejeito H_0.

Teste bilateral

No exemplo anterior, Joaquim ganha somente se der cara, por conta disso fazemos um teste considerando apenas um lado da distribuição, o que é chamado de *Teste unilateral*. No entanto, podemos querer saber se uma moeda é viciada para algum dos lados, sem necessariamente explicitar qual. Nesse caso iremos considerar a região crítica valores distantes para os dois lados da distribuição da estatística de teste.

1. Defino $H_0 : p = 0.5$.
2. Defino $\alpha = 5\%$;
3. Região crítica;
   $P(Z > z_c) + P(Z < -z_c) = 0.05$



Resolvendo para z_c

```
round(qnorm(p = 0.975,mean = 0,sd = 1),2)

## [1] 1.96
```

A região crítica nesse caso são todos os valores maiores que 1.96 ou menores que -1.96, o conjutno desse região representado por $(-\infty, -1.96) \cup (1.96, \infty)$

4. Calcular a estatística de teste;
   Considere o lançamento de uma moeda não viciada

```
set.seed(123)
n<-100
p<-0.5
s<-rbinom(1,size = n,prob = p)
p_hat<-s/n
p_hat

## [1] 0.49
```

   Se $\hat{p} = 0.49$, então $z = (49 - 50)/\sqrt{100 \times 0.5 \times 0.5} = -0.2$
5. Nesse caso $-0.2 \notin (-\infty, -1.96) \cup (1.96, \infty)$ e a hipótese nula não é rejeitada, ou seja **a moeda é tida como não viciada**.

Teste t

Os testes de hipótese apresentados acima são feitos para um tipo de problema onde a **variância é conhecida** sob H_0, esses são testes chamados de **teste Z**, pois o estimador segue uma distribuição Normal padrão. No entanto, um outro tipo de teste é feito quando a **variância é desconhecida**. O que muda na prática é somente a distribuição do estimador e a maneira como o intervalo de confiança é calculado.

Vamos motivar o teste t por um exemplo clássico. Considere uma fábrica de cervejas em Dublin na Irlanda que produz cervejas escuras, qualquer.

Em um controle de qualidade se deseja saber se o PH médio de determinado lote de cerveja é igual a 5.

Suponha que uma amostra de n unidades foi coletada de tal lote. A estatística de teste t, apresentada a seguir

Fonte: http://www.dailymail.co.uk/health/article-4073588/Don-t-want-deaf-pint-Guinness-day-High-levels-iron-helps-prevent-hearing-loss-study-finds.html

segue uma distribuição t-Student com (n-1) graus de liberdade.

$$t = \sqrt{n}\frac{(\bar{X} - \mu_0)}{s} \sim t - Student_{(n-1)} \tag{11}$$

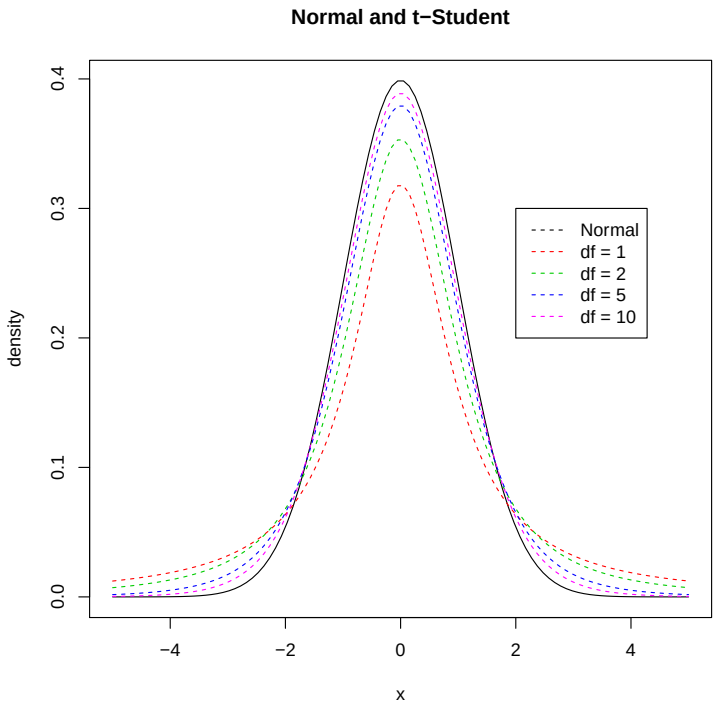
onde:

$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$

$s = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}}$

A distribuição t-Student possui um formato de sino, muito similar a Normal padrão, porém com caudas mais pesadas. No entanto, assintoticamente a distribuição t-Student converge para a Normal padrão. Abaixo apresentamos um gráfico para ilustrar tal convergência, com uma Normal padrão e 4 t-Students com graus de liberdade ("degrees of freedom") respectivamente iguais a 1, 2, 5 e 10.

```
mu<-0
sd1<-1
n<-100
x <- seq(-5,5, length.out = n)
y1 <- dnorm(x, mu, sd1)
y2 <- dt(x, 1)
y3 <- dt(x, 2)
y5 <- dt(x, 5)
y10 <- dt(x, 10)
plot (x, y1,ylab = "density", main = "Normal and t-Student",type = 'l')
lines(x,y2,col=2,lty=2)
lines(x,y3,col=3,lty=2)
lines(x,y5,col=4,lty=2)
lines(x,y10,col=6,lty=2)
legend(x = 2,y = 0.3,
       legend = c('Normal','df = 1','df = 2','df = 5','df = 10'),
       lty = 2,
       col = c(1,2,3,4,6))
```



Suponha que para uma determinada amostra de 100 observações temos que a média encontrada é igual a 4.8 e o desvio-padrão amostral é igual a 0.1, desse modo:

$n = 100$
$\bar{X} = 4.8$
$s = 0.1$

Para o teste de hipótese seguiremos os mesmos passos anteriores.
Porém considerando que a distribuição é uma t-Student com $(n-1) = 100 - 1 = 99$ graus de liberdade.

1. Defino $H_0 : \mu = 5$.
2. Defino o nível de significância $\alpha = 5\%$;
3. Região crític

$$P(t > t_c) + P(t < -t_c) = 0.05$$

Note que agora o nosso estimador tem distribuição t-Student e não mais Normal como antes. Basta então considerar quais são os valores críticos dessa distribuição, ao nível de significância adotado. Resolvendo para t_c temos que:

```
round(qt(p = 0.975,df = 99),2)

## [1] 1.98
```

A região crítica nesse caso é $(-\infty, -1.98) \cup (1.98, \infty)$

4. Calcular a estatística de teste;

$$t = \sqrt{n}\frac{(\bar{X} - \mu_0)}{s} = \sqrt{100}\frac{(4.8 - 5)}{0.1} = -20$$

5. Como $-20 \in (-\infty, -1.96) \cup (1.96, \infty)$ a hipótese nula é rejeitada.

Em resumo **qual teste devo usar para a média, t ou z?** Aí vão alguns comentários gerais sobre esses dois tipos de teste.

- O **teste t** é uma generalização do **teste z** quando a variância populacional não é conhecida sob a hipótese nula;
- Quando o tamanho da amostra é grande $n \to \infty$, então esses dois testes são iguais;
- Um **teste t** é mais parcimonioso do que um **teste z** no sentido de que permite observar valores mais extremos sem rejeitar H_0.
  Na dúvida, utilize um teste t.

Aplicação no R

O teste z não possui uma função nativa no R para er executado, no entando existe um pacote chamado BSDA do livro *Basic Statistics and Data Analysis* de Larry J. Kitchens. Basta instalar o pacote usando a linha de comando:

```
install.package('BSDA')
```

Exemplo 1:
Simule uma amostra aleatória com 30 observações de uma distribuição Normal com média 2 e desvio padrão 1.

```
set.seed(99)
x<-rnorm(n = 30,mean = 2,sd = 1)
```

Utilize somente as observações, e faça um exercício de imaginação supondo que não conhecemos o valor real do parâmetreo que representa a média. Agora teste a partir dessa amostra simulada se a média é estatisticamente diferente de 2.1.

```
library(BSDA)

## Loading required package:  lattice
##
## Attaching package:  'BSDA'
## The following object is masked from 'package:datasets':
##
##    Orange

z.test(x,mu=2.1,sigma.x = 1)

##
##  One-sample z-Test
##
## data:  x
```

```
## z = -1.3963, p-value = 0.1626
## alternative hypothesis: true mean is not equal to 2.1
## 95 percent confidence interval:
##  1.487228 2.202905
## sample estimates:
## mean of x
##  1.845067
```

Para os valores convencionais de nível de significância, entre 10% e 1% não podemos rejeitar a hipótese nula (Observe a informação do **p-value**). Ou seja não podemos rejeitar que a média é 2.1, caso tenhamos somente essa amostra e conheçamos a variância.

Exemplo 2:
Suponha que você possui a mesma amostra, porém agora desconhece a variância da variável aleatória. Nesse caso devemos utiliza o teste t.

```
t.test(x,mu = 2.1)

##
##  One Sample t-test
##
## data:  x
## t = -1.3598, df = 29, p-value = 0.1844
## alternative hypothesis: true mean is not equal to 2.1
## 95 percent confidence interval:
##  1.461643 2.228491
## sample estimates:
## mean of x
##  1.845067
```

Como no caso anterior aqui não podemos rejeitar a hipótese nula de que $\mu = \mu_0$ a níveis de confiança convencionais.

Exemplo 3:
Continue supondo que a variância é desconhecida. Agora queremos testar a hipótese nula de que a média é maior ou igual a 2.3. Perceba que nesse caso devemos realizar um teste unilateral.

```
t.test(x,mu = 2.3,alternative = 'less')

##
##  One Sample t-test
##
## data:  x
## t = -2.4267, df = 29, p-value = 0.01084
## alternative hypothesis: true mean is less than 2.3
## 95 percent confidence interval:
```

```
##       -Inf 2.163606
## sample estimates:
## mean of x
##  1.845067
```

Nesse teste a um nível de significância de 5% a hipótese nula é rejeitada, de modo que não é possível afirmar a partir da amostra que a média do processo gerador é maior ou igual a 2.3.

Exemplo 4:

Simule uma amostra com 40 observações a partir de uma distribuição normal com média 2.5 e desvio padrão 2.

```
set.seed(88)
y<-rnorm(n = 40,mean = 2.5,sd = 2)
```

Teste se a média do processo gerador de X é diferente da média do processo gerador de Y.[7] Aqui conhecemos o processo gerador dos dois processos e sabemos que as médias são diferentes, porém na prática não conhecemos tais parâmetros. Novamente, vamos fazer um exercício de imaginação, supondo que possuímos apenas as amostras.

```
t.test(x,y)

##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -0.9097, df = 63.09, p-value = 0.3664
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.0149649  0.3799453
## sample estimates:
## mean of x mean of y
##  1.845067  2.162576
```

Observando o valor p temos que a hipótese nula não pode ser rejeitada a níveis de significância convencionais. Ou seja, mesmo as médias amostrais sendo diferentes, não temos evidência suficiente para rejeitar a hipótese nula de que as médias populacionais são iguais. Esse resultado tem relação com o poder do teste. Devido ao tamanho da amostra relativamente pequeno o poder do teste é baixo, para aumentar o poder e assim a precisão desse teste podemos aumentar o tamanho da amostra.

Exemplo 5:

Realize o mesmo exemplo anterior, porém agora com amostras de tamanho 200.

```
set.seed(32)
x<-rnorm(n = 200,mean = 2,sd = 1)
y<-rnorm(n = 200,mean = 2.5,sd = 2)
t.test(x,y)
```

[7] Para mais detalhes sobre o cálculo para a estatística desse teste consultar o material complementar

```
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -4.2843, df = 281.21, p-value = 2.52e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.9603775 -0.3557000
## sample estimates:
## mean of x mean of y
##  1.944300  2.602339
```

A conclusão do teste é que podemos rejeitar a hipótese nula de que as médias são iguais a qualquer nível de confiança convencional, o que é esperado dado que pelo processo gerador esses parâmetros são de fato diferentes. O resultado de aumentar a amostra tem efeito na redução da variância da estatística de teste, deixando esta mais precisa.

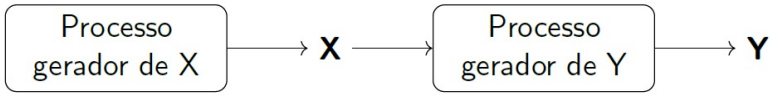
Exercício PNAD e continuação ChickWeight

Para fixar ideias resolva os seguintes exercícios:

1. **Exercício 1:** Considere a base de dados `pnad2015-sub.csv`, ela representam uma subamostra com 10000 observações da PNAD do ano de 2015. Pede-se:
   (a) Calcule a média da renda total para os homens e para as mulheres;
   (b) Faça um teste unilateral para a renda de homens, considerando a hipótese nula de que a média é maior ou igual que 1500, a um nível de significância de 5%.
   (c) Faça um teste para verificar se a média da renda de homens é igual ao de mulheres.
2. **Exercício 2:** Retome ao exemplo do final da aula anterior sobre a melhor ração para frangos.
   (a) Teste se no dia 21, os frangos alimentados com a dieta 3 são mais pesados do que os frangos alimentados com a dieta 4;
   (b) Faça o mesmo teste para os alimentados com a dieta 1 e a dieta 2;
   (c) Faça o mesmo teste para os alimentados com a dieta 2 e a dieta 3;
   (d) Com esses resultados é possível concluir que todas as dietas tem o mesmo efeito?

Regressão Linear

Até o momento consideramos que os processos geradores de uma variável X são **independentes** de qualquer outra variável. Agora iremos considerar modelos EM que o processo gerador de uma variável Y **depende** da variável X.



Fonte: Elaboração própria

As variáveis X e Y recebem nomes especiais:

- **Variável** Y: dependente, de resposta, explicada, prevista ou regressando;
- **Variável** X: independente, de controle, explicativa, previsora ou regressor.

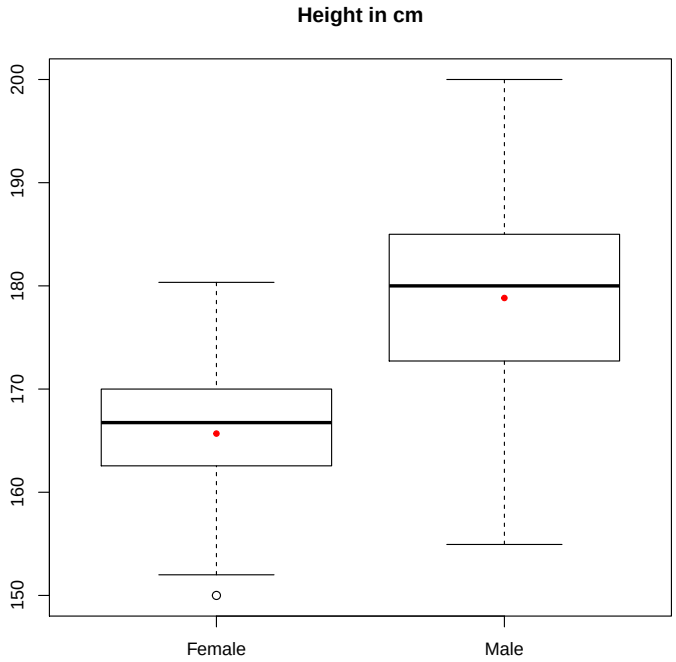
Para motivar o tema utilizaremos nessa aula a base de dados `survey` do pacote `MASS`. Inicialmente vamos rever as duas aulas passadas com dois exercícios básicos.

1. Instale o pacote `MASS` e leia a base survey;

```
install.packages('MASS')
```

```
library(MASS)
data(survey)
```

2. Em gráficos de boxplot com as médias, apresente a diferença existente entre a variável `Height` entre homens e mulheres;



3. Teste a hipótese de que as médias são diferentes a um nível de significância de 5%.

```
##
##  Welch Two Sample t-test
##
## data:  survey$Height[survey$Sex == "Male"] and survey$Height[survey$Sex == "Female"]
## t = 12.924, df = 192.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  11.13420 15.14454
## sample estimates:
## mean of x mean of y
##  178.8260  165.6867
```

Note que algumas das observações são `NA` para a variável `Height`. Como você faria para contar a quantidade de `NA`? Eu gosto do comando abaixo.

```
sum(is.na(survey$Height))
```

```
## [1] 28
```

Dado que faltam 28 dados de altura para a amostra, qual a melhor previsão que podemos fazer para esses valores que são `NA`? Essa é uma das questões mais importantes que um cientista de dados terá que responder. Até agora pelo que vimos temos duas escolhas:

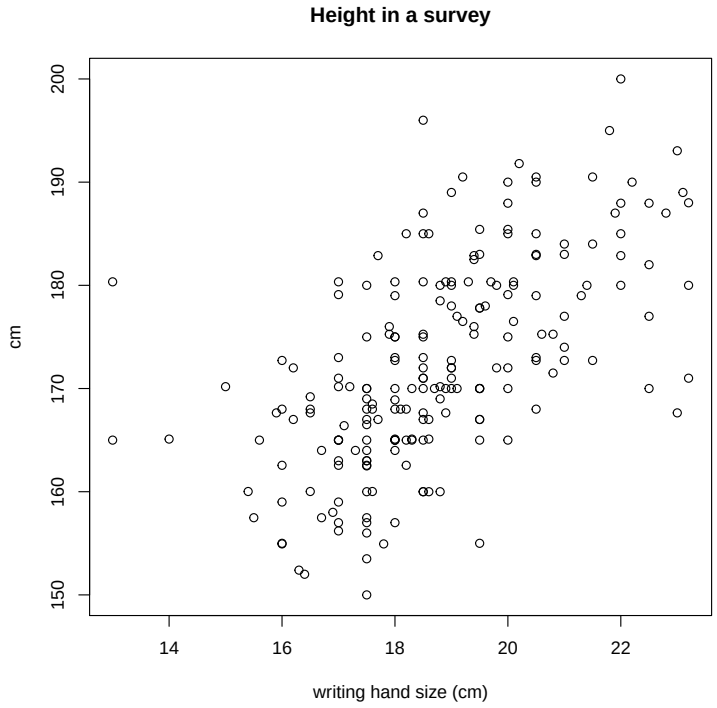
- podemos usar uma medida de centralidade global, como a média, ou;
- podemos usar uma medida de centralidade para cada grupo. [8]

Mínimos Quadrados Ordinários

Essas são **medidas unitárias** para calcular a relação entre duas variáveis. Considere agora que você deseja utilizar a informação do tamanho da mão destra para prever a altura dos indivíduos. Conforme observamos pelo gráfico de pontos abaixo existe uma relação positiva entre essas duas medidas, como deveríamos esperar.

```
plot(survey$Wr.Hnd,survey$Height,
     main = 'Height in a survey',
     xlab = 'writing hand size (cm)',
     ylab = 'cm')
```

[8] Aqui os grupos podem ser classificados de diferentes maneiras, a primeira vista existem grupos evidentes representados por variáveis categóricas, mas também existem grupos não tão evidentes assim. Um tópico importante de *Machine Learning* é o estudo de grupos nos dados ou se preferir, análise de *clusters*



Uma a relação entre elas é por meio de uma **função afim**, que é aproximadamene a ideia de grupos quando existe um número infinito de grupos representados por cada valor possível de uma variável contínua. De maneira mais geral queremos encontrar uma reta que resuma a informação do gráfico.

Equação da reta:

$$\hat{y} = b_0 + b_1.x$$

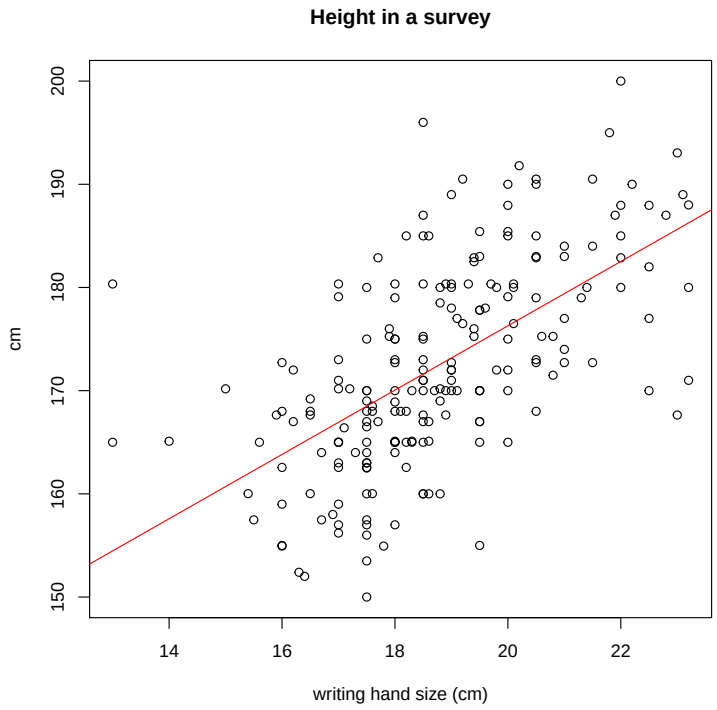
onde:
b_0 é o coeficiente linear (ou intercepto), e
b_1 é o coeficiente angular.

A função `lm` de *Linear Model* nos retorna exatamente os coeficientes que desejamos para definir tal reta.

```
lm(Height ~ Wr.Hnd,data = survey)

##
## Call:
## lm(formula = Height ~ Wr.Hnd, data = survey)
##
## Coefficients:
## (Intercept)       Wr.Hnd
##     113.954        3.117
```

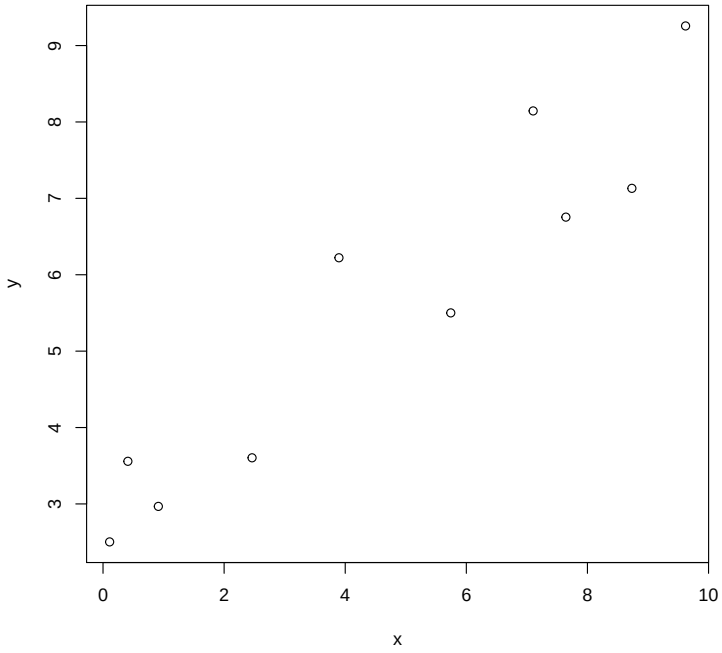
Traçando essa reta nos pontos temos o seguinte modelo:

```
reg<-lm(Height ~ Wr.Hnd,data = survey)
plot(survey$Wr.Hnd,survey$Height,
     main = 'Height in a survey',
     xlab = 'writing hand size (cm)',
     ylab = 'cm')
abline(reg$coefficient,col=2)
```



A título de ilustração para entender o método vamos utilizar dados simulados. A seguir simulei um conjunto de dados com duas variáveis x e y e 10 observações.Tais dados são desenhados em um gráfico de pontos, note que eles estão espalhados em torno de uma linha imaginária.

```
# Simulando um processo gerador
set.seed(13)   # Valor arbitrário
n <- 10
x<-runif(n,min = 0,max = 10)
x<-sort(x)
beta0 <- 2.0   # não conhecemos esse parâmetro
beta1 <- 0.8   # não conhecemos esse parâmetro
eps<- rnorm(n)
y = beta0 + beta1*x + eps
plot(x,y)
```

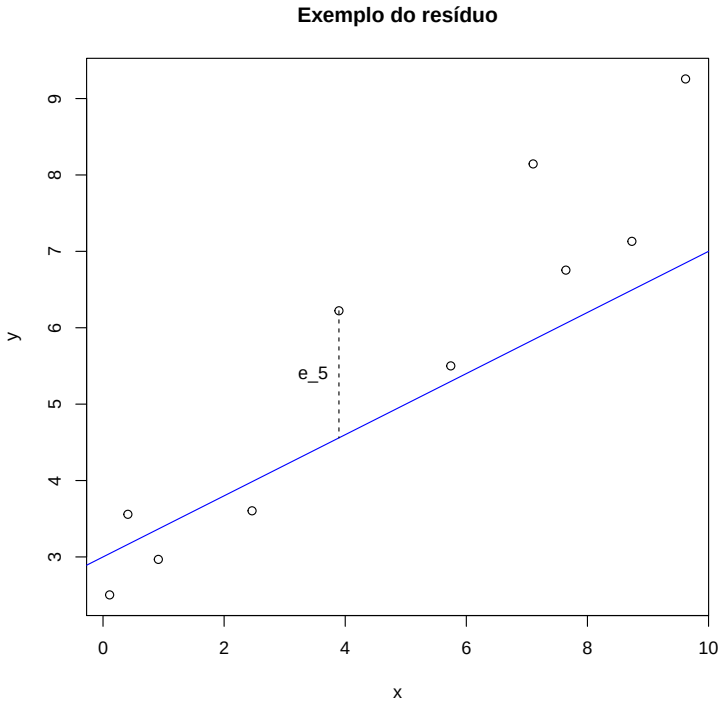


Qual a melhor reta que podemos encontrar?

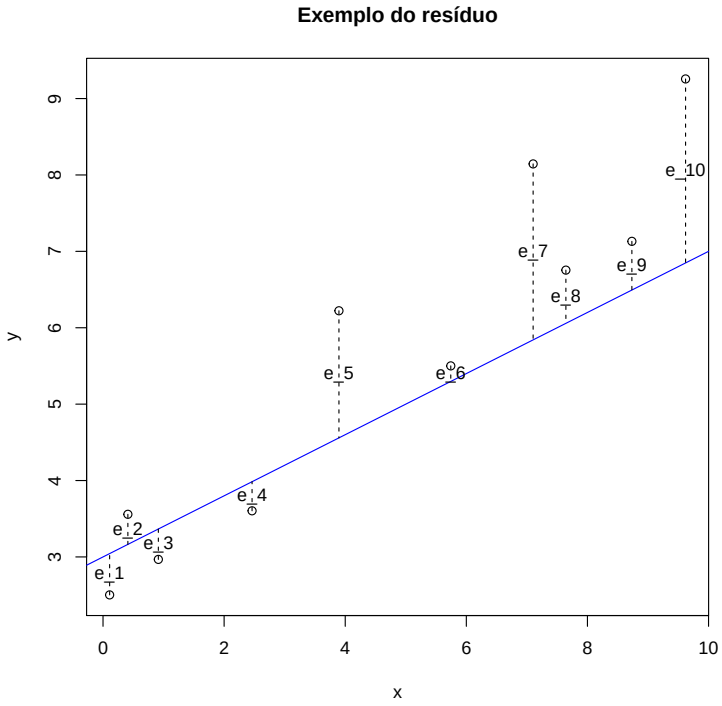
Aquela que está mais próxima de todos os pontos. Defina a distância da reta ao ponto i como sendo e_i, chamado de **resíduo**. Desse modo temos que $e_i = y_i - \hat{y}_i$. Apesar de termos gerado o processo e conhecermos os parâmetros, vamos assumir que eles são desconhecidos, pois nossa tarefa aqui é estimá-los e quando estivermos trabalhando com dados reais não conhecemos o processo.

Definimos uma reta com os coeficientes $b0 = 3$ e $b1 = 0.4$:

```
# Suponha que usamos esses valores para fitar a linha
b0 = 3
b1 = 0.4
yhat = b0 + b1*x
plot(x,y,main = 'Exemplo do resíduo')
abline(a = b0,b = b1,col=4)
lines(c(x[5],x[5]),c(y[5],yhat[5]),lty = 2)
text(x[5],(y[5]+yhat[5])/2,labels = 'e_5',pos = 2)
```




```
plot(x,y,main = 'Exemplo do resíduo')
abline(a = b0,b = b1,col=4)
for (i in 1:n){
lines(c(x[i],x[i]),c(y[i],yhat[i]),lty = 2)
text(x[i],(y[i]+yhat[i])/2,labels = paste0('e_',i))
}
```



Essa reta não parece estar muito próxima de todos os pontos, podemos fazer melhor. Os coeficiêntes linear e angular - (b_0^*, b_1^*) - para melhor reta são dados pela **minimização dos quadrados dos resíduos**:

$$[b_0^*, b_1^*] = argmin_{b0,b1} SQR = argmin_{b0,b1} \sum_{i=1}^{n} e_i^2$$

$$= argmin_{b0,b1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= argmin_{b0,b1} \sum_{i=1}^{n} (y_i - b0 - b1.x_i)^2$$

A solução do problema de minimização acima é dada por:

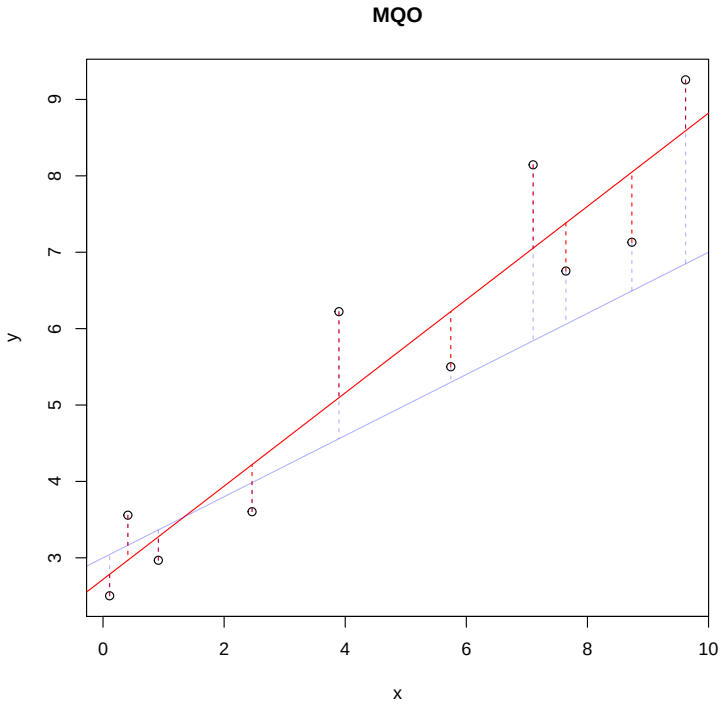
$$b_1^* = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{cov(x,y)}{var(x)}$$

$$b_0^* = \bar{y} - b_1^*.\bar{x}$$

que são exatamente os coeficientes calculados pela função `lm`. Segue a seguir o melhor ajuste de reta aos pontos.

```
modelo<-lm(y ~ x)
plot(x,y,main = 'MQO')
abline(reg = modelo, col=2)
abline(a = b0,b = b1,col=rgb(0,0,1,0.3))
for (i in 1:n){
```

```
lines(c(x[i],x[i]),c(y[i],modelo$fitted.values[i]),lty = 2,col=2)
lines(c(x[i],x[i]),c(y[i],yhat[i]),lty = 2,col=rgb(0,0,1,0.3))
}
```

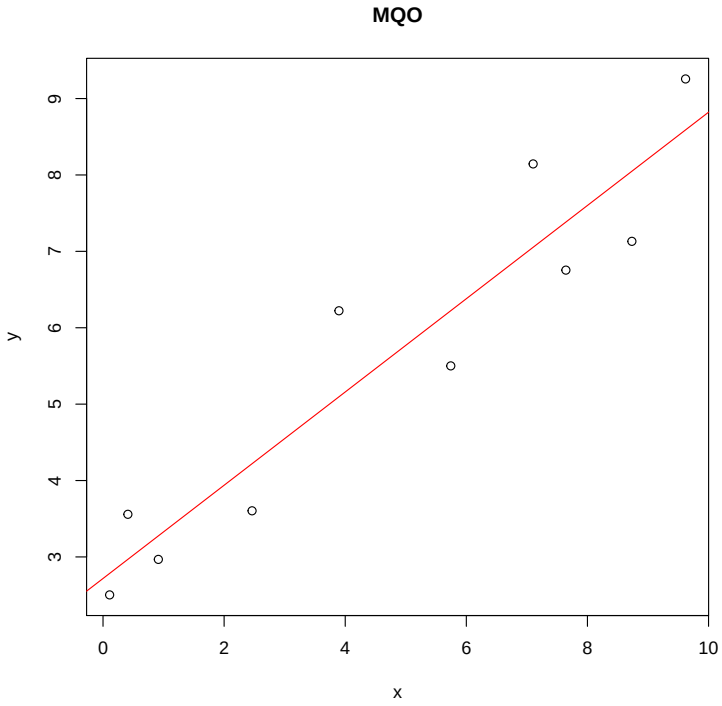


Qualidade do ajuste

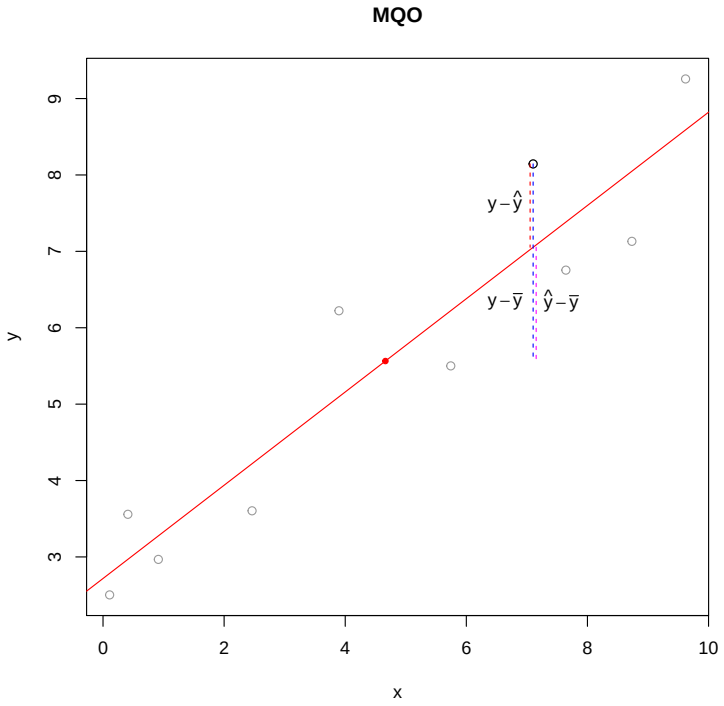
ANOVA é uma sigla para análise de variância. Queremos identificar quanto da variância nos dados pode ser explicada por cada componente do modelo. Considere:

- y_i: o valor observado da variável;
- $\hat{y}_i$: o valor previsto da variável;
- $\bar{y}$: a média dos valores observados.

```
modelo<-lm(y ~ x)
plot(x,y,main = 'MQO')
abline(reg = modelo, col=2)
```




```
modelo<-lm(y ~ x)
plot(x,y,main = 'MQO',col=rgb(0,0,0,0.4))
abline(reg = modelo, col=2)
i<-7
points(x[i],y[i])
points(mean(x),mean(y),pch=20,col=2)

ee<-0.05

lines(c(x[i]-ee,x[i]-ee),c(y[i],modelo$fitted.values[i]),lty = 2,col=2)
text(x[i],(y[i]+modelo$fitted.values[i])/2,labels = bquote(y-hat(y)),pos = 2)

lines(c(x[i],x[i]),c(y[i],mean(y)),lty = 2,col=4)
text(x[i],(modelo$fitted.values[i]+mean(y))/2,labels = bquote(y-bar(y)),pos = 2)

lines(c(x[i]+ee,x[i]+ee),c(modelo$fitted.values[i],mean(y)),lty = 2,col=6)
text(x[i],(modelo$fitted.values[i]+mean(y))/2,labels = bquote(hat(y)-bar(y)),pos = 4)
```



A seguinte identidade é válida para todo i:

$$y_i - \bar{y}_i = (\hat{y}_i - \bar{y}_i) + (y_i - \hat{y}_i)$$

O que não é imediato, é que a igualdade abaixo também é válida:

$$\sum_i^n (y_i - \bar{y}_i)^2 = \sum_i^n (\hat{y}_i - \bar{y}_i)^2 + \sum_i^n (y_i - \hat{y}_i)^2$$

Porém esse resultado é obtido pois os dois termos do lado direito da equação não são correlacionados.

Seja:

- Soma dos quadrados totais: $SQT = \sum_i^n (y_i - \hat{y}_i)^2$
- Soma dos quadrados explicados: $SQE = \sum_i^n (\hat{y}_i - \bar{y}_i)^2$
- Soma dos quadrados dos resíduos: $SQR = \sum_i^n (y_i - \hat{y}_i)^2$

$$SQT = SQE + SQR$$

Defino a variável R^2:

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}$$

Essa variável é chamada e coeficiente de determinação e representa o quanto o modelo linear é um bom ajuste para os dados.

Quanto o modelo explica dos dados?

1. R^2 é um valor entre 0 e 1;
2. Quanto mais próximo de 1, mais próximos os pontos estão da reta;
3. Quanto mais próximo de 0, mais distantes os pontos da reta.

Além disso, em um modelo linear com apenas uma variável explicativa temos que $R^2 = (correlação)^2$. Desse modo quanto mais próximo de 0 a correlação menor o poder explicativo de um modelo linear para essas variáveis.

Voltando para nosso exercício com os dados da base `survey`:

```
reg<-lm(Height ~ Wr.Hnd,data=survey)
SQR<-sum(reg$residuals^2)
SQT<-sum((survey$Height[!is.na(survey$Height)]
          -mean(survey$Height,na.rm = T))^2)
1-SQR/SQT

## [1] 0.3611947
```

Calculando pela correlação:

```
with(survey,cor(Height,Wr.Hnd,use = 'complete.obs'))^2

## [1] 0.3611901
```

Assim aproximadamente 36,12% da variância da altura é explicada em um modelo linear pela variável amplitude da mão.

Vamos olhar para outras variáveis que parecem explicar a altura, uma delas é o sexo.

Múltiplos regressores e variável dummy

Um modelo com múltiplos regressores tem as mesmas características do modelo com apenas um regressor, queremos encontrar a melhor reta que se ajusta aos dados. Porém essa reta não se encontra mais no espaço euclideano bidimensional, mas sim multidimendional. Se considerarmos 3 variáveis (a explicada e 2 explicativas), teremos então um espaço 3D para encontrar a reta. Se considerarmos n variáveis explicativas teremos então um plano com $n + 1$ dimensões.

Equação da reta:

$$\hat{y} = b_0 + b.x$$

porém agora b e x são vetores n dimensionais e . representa agora um produto interno [9]. É possível escrever de outra forma a equação acima, evidenciando os termos do produto interno:

Equação da reta:

$$\hat{y} = b_0 + b_1.x_1 + b_2.x_2 + ...b_n.x_n$$

onde b_i e x_i são os valores das coordenadas i dos respectivos vetores b e x.

Recuperando o exemplo anterior vamos introduzir a variável sexo como regressor. Desse modo, além do tamanho da mão destra, a variável sexo também será utilizada como explicativa. No R basta acrescentar a variável de interesse na função de maneira aditiva.

[9] O produto interno pode ser compreendido como o somatório do produto de cada termo de um vetor pela coordenada correspondente em outro vetor)

```
reg<-lm(Height ~ Wr.Hnd + Sex,data = survey)
reg

##
## Call:
## lm(formula = Height ~ Wr.Hnd + Sex, data = survey)
##
## Coefficients:
## (Intercept)       Wr.Hnd      SexMale
##     137.687        1.594        9.490
```

Como estamos estimando um modelo linear com variáveis binárias (*dummy*) estamos de fato encontrando duas retas com diferentes coeficientes de lineares. A variável sexo é binária, sendo no R considerada por ordem alfabética $Female = 0$ e $Male = 1$. Os interceptos das retas são:

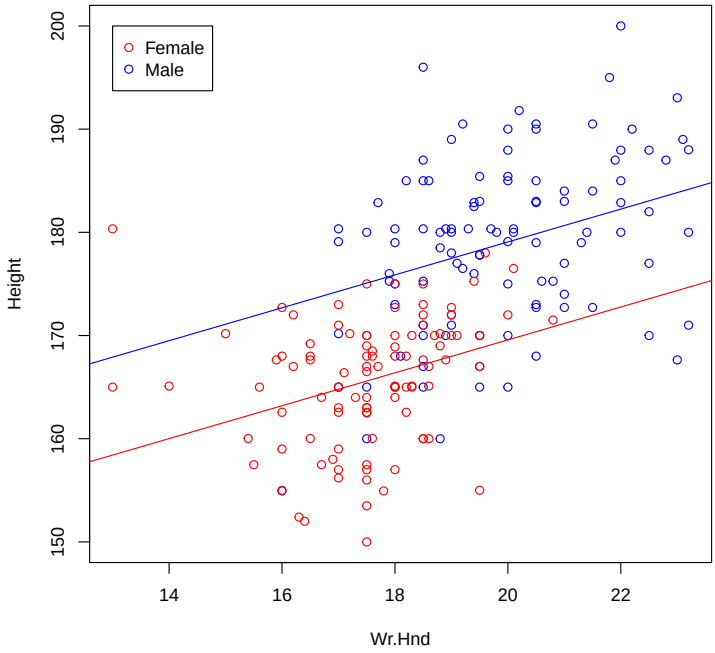
- Para mulheres:

$$137.687 + 0 * 9.49 = 137.687$$

- Para homens:

$$137.687 + 1 * 9.49 = 147.177$$

```
with(survey,plot(Wr.Hnd[survey$Sex=='Male'],
                 Height[survey$Sex=='Male'],
                 xlab = 'Wr.Hnd',
                 ylab = "Height",
                  xlim = range(Wr.Hnd,na.rm = T),
                  ylim = range(Height,na.rm=T),col=4))
with(survey,points(Wr.Hnd[survey$Sex=='Female'],
                 Height[survey$Sex=='Female'],
                 xlab = 'Wr.Hnd',
                 ylab = "Height",
                  xlim = range(Wr.Hnd,na.rm = T),
                  ylim = range(Height,na.rm=T),col=2))

abline(c(reg$coefficients[1]+reg$coefficients[3],reg$coefficients[2]),col=4)
abline(c(reg$coefficients[1],reg$coefficients[2]),col=2)

legend(x = 13,y = 200,
       legend = c('Female','Male'),
       pch = c(1,1),
       col = c(2,4))
```



Vamos calcular o coeficiente de determinação (R^2)desse modelo.

```
reg<-lm(Height ~ Wr.Hnd + Sex,data=survey)
SQR<-sum(reg$residuals^2)
SQT<-sum((survey$Height[!is.na(survey$Height)]
          -mean(survey$Height,na.rm = T))^2)
1-SQR/SQT

## [1] 0.5062514
```

O modelo fica claramente melhor, conseguindo explicar mais 50% da variação dos dados.

Significância individual

O processo gerador da variável Y será modelado como sendo uma função "linear" da variável X. Assim como fizemos na simulação acima.

$$Y = \beta_0 + \beta_1.X + \epsilon$$

Onde ϵ é uma variável aleatória tal que:

- $cov(X, \epsilon) = 0$
- $E[\epsilon] = 0$

Assim a média condicional da variável aleatória Y é modelada como sendo dependente da variável X.

$$E[Y|X] = E[\beta_0 + \beta_1.X + \epsilon|X]$$

$$E[Y|X] = \beta_0 + \beta_1.E[X|X]$$

$$E[Y|X] = \beta_0 + \beta_1.X$$

É possível mostrar que os melhores estimadores para β_0 e β_1 são obtidos pelo MQO. Assim:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1.X$$

Uma vez estimados os parâmetros do modelo queremos ser capazes de dizer se eles são estatisticamente significativos, ou seja se eles são diferentes de 0. Caso o coeficiente seja significativo, isso é evidência de que a variável explicativa de fato tem poder para explicar o modelo. O teste é o seguinte.

- H0: o parâmetro é igual a zero;
- H1: o parâmetro é diferente de zero.

```
reg<-lm(Height ~ Wr.Hnd + Sex,data=survey)
summary(reg)

##
## Call:
## lm(formula = Height ~ Wr.Hnd + Sex, data = survey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7479  -4.1830   0.7749   4.6665  21.9253
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 137.6870     5.7131  24.100  < 2e-16 ***
## Wr.Hnd        1.5944     0.3229   4.937 1.64e-06 ***
## SexMale       9.4898     1.2287   7.724 5.00e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.987 on 204 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.5062,Adjusted R-squared:  0.5014
## F-statistic: 104.6 on 2 and 204 DF,  p-value: < 2.2e-16
```

Com a função `summary` conseguimos as estatísticas calculadas e o teste de hipótese, assim como o p-valor.

Previsão

Com o modelo linear podemos fazer previsão de valores não observados e responder a questão levantada no início dessa seção. Essas previsões são os valores sobre a reta estimada. Utilizando como regressores `Wr.Hnd` e `Sex`, calcule a melhor previsão para os dados que estão missings (`NA`).

```
previsao<-predict(reg,newdata = survey[is.na(survey$Height),])

# Apresentando a previsão junto com os valores utilizados como regressores.
cbind(survey[is.na(survey$Height),c('Sex','Wr.Hnd')],previsao)
```

```
##        Sex Wr.Hnd previsao
## 3     Male   18.0 175.8768
## 12    Male   21.0 180.6601
## 15    Male   16.0 172.6879
## 25  Female   17.0 164.7925
## 26    Male   18.5 176.6740
## 29    Male   17.8 175.5579
## 31  Female   18.5 167.1842
## 35    Male   18.0 175.8768
## 58    Male   19.5 178.2685
## 68  Female   18.5 167.1842
## 70    Male   21.0 180.6601
## 81    Male   19.5 178.2685
## 83  Female   17.5 165.5897
## 84  Female   17.0 164.7925
## 90  Female   18.0 166.3870
## 92  Female   17.5 165.5897
## 96  Female   19.0 167.9814
## 108 Female   17.0 164.7925
## 121   Male   20.0 179.0657
## 133 Female   18.9 167.8220
## 157   Male   14.0 169.4990
## 173 Female   15.5 162.4009
## 179 Female   20.5 170.3731
## 203 Female   18.8 167.6625
## 213   Male   18.0 175.8768
## 217 Female   16.3 163.6764
## 225 Female   17.6 165.7492
## 226 Female   17.5 165.5897
```

Exercício ToothGrowth

Um laboratório farmaceutico tem interesse em estudar os benefícios da vitamina C. Mais especificamente gostariam de estudar a conjectura de que a ingestão de vitamina C auxilia o crescimento dos ossos.

Para desenvolver tal pesquisa foram utilizados porcos da índia. Cada animal recebeu doses diferentes de Vitamina C e foram registrados os tamanhos de seus dentes.

Utilizando a base de dados `ToothGrowth`.:

1. Estime um modelo linear com a variável dependente sendo o tamanho do dente - `len` - e a variável explicativa sendo a quantidade ingerida - `dose`;
2. Verifique se o regressor é significativo a 95% de confiança;
3. Em um gráfico apresente os dados e a melhor reta;

   temos outra variável que é a maneira como a Vitamina C é aplicada - `supp`:
4. Estime o modelo linear com a variável dependente sendo o tamanho do dente - `len` - e as variável explicativas sendo a quantidade ingerida - `dose` - e o tipo de ingestão - `supp`;

Fonte: http://www.mdig.com.br/index.php?itemid=11670

5. Verifique se os regressores são significativos a 95% de confiança;

Modelo linear generalizado

Um **modelo linear generalizado** pode ser escrito da seguinte forma:

$$f(Y;\theta,\gamma) \tag{12}$$

$$\theta = m(\eta) = m(X.\beta) \tag{13}$$

onde:
Y é a variável dependente com distribuição dada pela função densidade $f(Y;\theta,\gamma)$, da família exponencial;
θ é conjunto de parâmetros da distribuição que desejamos modelar;
γ é o conjunto de parâmetros conhecidos;
X é o conjunto de variáveis dependentes;
β é o conjunto de parâmetros que associa de maneira linear as variáveis dependentes;
m é uma função com domínio em $\mathbb{R}$.

Tendo essa definição podemos incluir o modelo de regressão linear que vimos anteriormente nesse arcabouço, basta definirmos $Y \sim N(\mu_{Y|X},\sigma)$ e m sendo a função identidade. Porém ele generaliza no sentido de que outros tipos de relação podem ser encontradas, como por exemplo se tivermos uma variável resposta discreta.

Suponha que a variável resposta é binária e assume valores 0 ou 1. A média da Bernoulli é uma função limitada, deve estar entre 0 e 1. Como os modelos de regressão linear são irrestritos, podemos ter valores maiores do que 1, ou menores do que 0 para a média da Bernoulli. Isto não está de acordo com a teoria de probabilidade que vimos. Por conta disso fazemos uma transformação da média para um parâmetro não limitado, usando por exemplo a função logit e assim modelamos o parâmetro tranformado por meio de um modelo linear. Esse tipo de modelo apesar de ser não linear é considerado um modelo linear generalizado, pois a transformação da variável dependente é tratada em um modelo linear.

Para ilustra esse tópico, vamos iniciar com um exemplo. Suponha um jogo entre Joaquim e Roberto. Cada um possui sua moeda e queremos modelar a probabilidade de dar Cara quando Joaquim lança e a mesma probabilidade quando Roberto lança. Utilizando uma abordagem de regressão linear nos moldes do que vimos na seção anterior, teremos o seguinte modelo.

$$Y = \beta_0 + \beta_1.X + \epsilon \tag{14}$$

$$E[Y|X] = \beta_0 + \beta_1.X \tag{15}$$

onde:
Y pode assumir valores em $\{0,1\}$, onde 0 é Coroa e 1 é Cara;
X pode assumir valores em $\{0,1\}$, onde 0 representa o lançamento de Roberto e 1 representa o lançamento de Joaquim;
$\epsilon \sim N(0,\sigma)$.

Note que com essa especificação temos que a probabilidade de um lançamento ser Cara é dado por uma Normal com média $\beta_0 + \beta_1.X$ e desvio padrão σ. No entanto esse modelo não parece ser muito razoável, já discutimos no começo do curso que o melhor modelo para tratar variáveis binárias é o Bernoulli. Desse modo queremos ser

capazes de escrever um modelo tal que a distribuição de Y seja uma Bernoulli com parâmetro p.

Basta para isso definir uma função $m : \mathbb{R} \to [0, 1]$, ou seja uma função que tenha no domínio todos os números reais e na imagem apenas o intervalo $[0, 1]$. Desse modo conseguimos definir um modelo linear para uma variável auxiliar, digamos η, que será o argumento da função m, que por sua vez retorna a probabilidade de sucesso da variável aleatória Y. Dessa forma podemos definir um **modelo linear generalizado** para variável dependente binária como sendo:

$$Y \sim Bernoulli(p) \tag{16}$$

$$p = E[Y|X] = m(\eta) = m(\beta_0 + \beta_1.X) \tag{17}$$

Restrição para a probabilidade estar entre 0 e 1:

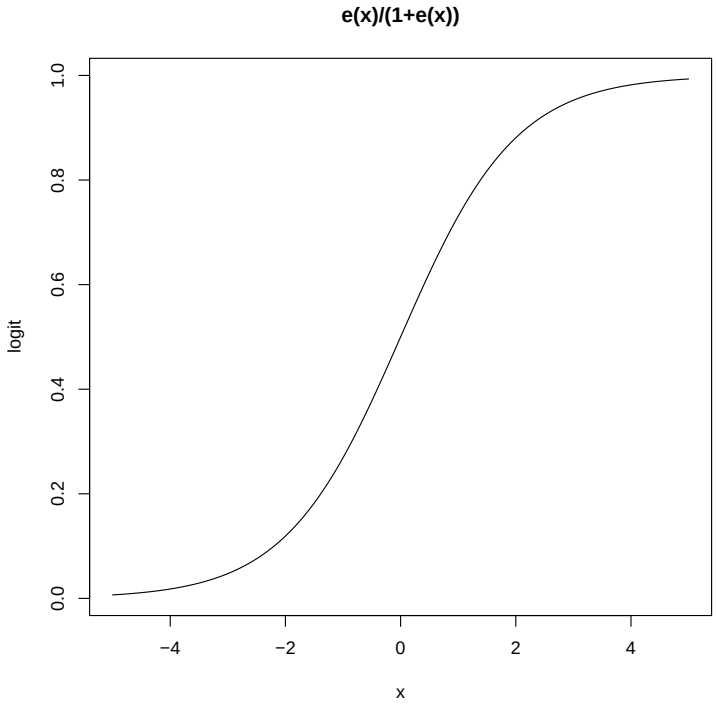
$$m : \mathbb{R} \to [0, 1] \tag{18}$$

Modelo de regressão logística - Logit

Existe uma grande família de funções que consegue transformar a qualquer valor da reta em um número entre 0 e 1. Uma função com características interessantes é a **função logística**. Ela é contínua, diferenciável e simétrica ímpar em torno de 0.5. Quando consideramos um modelo linear generalizado para variável dependente de Bernoulli e a função de ligação m(.) sendo a logística, estamos trabalhando com um modelo Logit.

$$m(x) = \frac{e^x}{(1 + e^x)} \tag{19}$$

```
x<-seq(-5,5,length.out = 1000)
logit<-exp(x)/(1+exp(x))
plot(x,logit,type = 'l',
     main = 'e(x)/(1+e(x))')
```



A função inversa da logística é dada por:

$$m^{-1}(y) = log\left(\frac{y}{(1-y)}\right) \tag{20}$$

Interpretação dos coeficientes:
Na equação acima, se considerarmos $x = X.\beta$ temos que a probabilidade do evento sucesso ocorrer é dada por:

$$\mu_{Y|X} = \frac{e^{X.\beta}}{(1+e^{X.\beta})} \tag{21}$$

desse modo temos que o modelo logit pode ser escrito, como: A função inversa da logística é dada por:

$$log\left(\frac{\mu_{Y|X}}{1-\mu_{Y|X}}\right) = X.\beta = \beta_0 + \beta_1.X_1 + ... + \beta_k.X_k \tag{22}$$

Um novo produto será lançado e uma pesquisa de opinião foi realizada para verificar qual o público será mais atraído por ele. Desse modo facilitar o direcionamento da campanha de marketing. Considere a base de dados `tastesgreat` do pacote `UsingR`.

Estime um modelo logit para a variável `enjoyed` dado a idade (`age`)

```
library(UsingR)
```

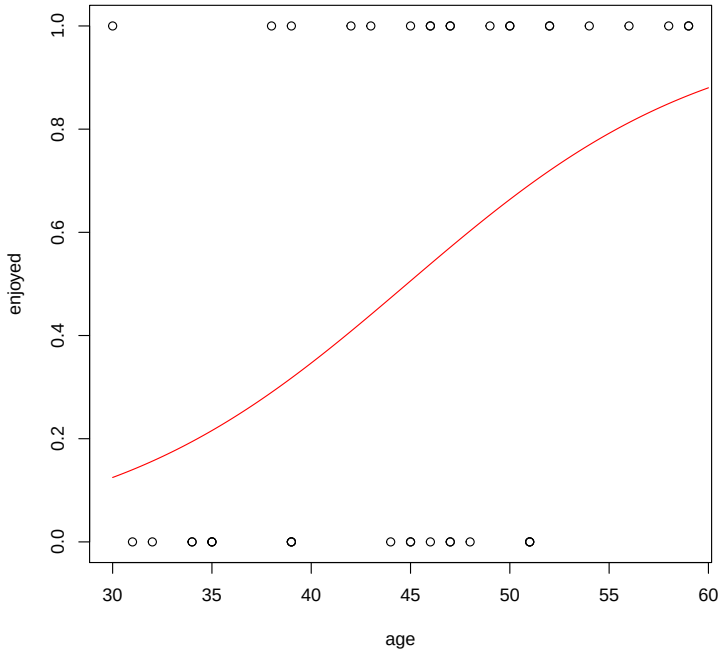
```
data(tastesgreat)
head(tastesgreat)
```

Fonte: http://cdn2.hubspot.net/hub/215841/file-3945428210-jpg/blog-files/new-product-blog-602x347pix.jpg

```
##   gender age enjoyed
## 1 Female  35       0
## 2   Male  44       0
## 3   Male  45       1
## 4 Female  47       1
## 5 Female  51       0
## 6 Female  47       0
```

Para estimar a regressão logística utiliza-se a função `glm` com o parâmetro `family = binomial()`. Segue abaixo a estimação e o ajuste da curva aos dados. Aqui a curva representa a probabilidade de observar uma observação 1 na variável de interesse.

```
reg<-glm(enjoyed ~ age,data = tastesgreat,family=binomial())
plot(enjoyed ~ age,data=tastesgreat)
x<-seq(30,60,length.out = 1000)
logit<-exp(reg$coefficients[1]+reg$coefficients[2]*x)/
  (1+exp(reg$coefficients[1]+reg$coefficients[2]*x))
lines(x,logit,col=2)
```




```
reg

##
## Call:  glm(formula = enjoyed ~ age, family = binomial(), data = tastesgreat)
##
## Coefficients:
## (Intercept)          age
##     -5.8876       0.1314
##
## Degrees of Freedom: 39 Total (i.e. Null);  38 Residual
## Null Deviance:     55.45
## Residual Deviance: 47.36  AIC: 51.36
```

No modelo linear temos que a interpretação dos coeficientes é direto, sendo os coeficientes lineares e angulares. Quando temos uma variável *dummy* vimos que a intepretação ainda se mantém, com o coeficiente da variável *dummy* sendo somado ao coeficiente linear. Mas qual a interpretação em um modelo logístico?

Interpretação dos coeficientes

A interpretação dos coeficientes no caso da regressão logística não é imediato como nos casos da regressão linear. De fato, conseguimos interpretar o aumento na probabilidade do evento sucesso ($Y = 1$) por um aumento unitário em uma variável de controle X_1 por meio da transformação logística:

$$\text{P(Y=1—X}_1, ..., X_k) = m(\beta_0 + \beta_1.X_1 + ... + \beta_k.X_k) = \frac{e^{\beta_0+\beta_1.X_1+...+\beta_k.X_k}}{1+e^{\beta_0+\beta_1.X_1+...+\beta_k.X_k}}$$

$$P(Y=1|(X_1+1),...,X_k) = m(\beta_0+\beta_1.X_1+...+\beta_k.X_k) = \frac{e^{\beta_0+\beta_1.(X_1+1)+...+\beta_k.X_k}}{1+e^{\beta_0+\beta_1.(X_1+1)+...+\beta_k.X_k}}$$

$$\Delta Probabilidade = P(Y=1|(X_1+1),...,X_k) - P(Y=1|X_1,...,X_k)$$

Note que esse aumento depende do nível das váriáveis de controle, isto é, depende de todos os valores $(X_1,...,X_k)$. Assim a interpretação dos coeficientes poderia ser feita individualmente, considerando variações unitárias em cada variável de controle para todos os seus níveis.
Esse método é custoso em termos de visualização dos resultados. Para superar tal dificuldade é usado comumentemente o que é chamado de **razão de chances** (ou odds ratio). Ela é definido do seguinte modo:

Definição: A chance do evento A é igual a probabilidade do evento A ocorrer, sobre a probabilidade dele não ocorrer.

$$O_A = \frac{P(A)}{1-P(A)}$$

Definição: A razão de chances de um evento A dado o evento B é definido como a chance do evento A ocorrer dado B $(A|B)$ sobre a chance do evento A ocorrer quando B não é observado $(A|\bar{B})$:

$$OR_{A|B} = \frac{\frac{P(A|B)}{1-P(A|B)}}{\frac{P(A|\bar{B})}{1-P(A|\bar{B})}} = \frac{P(A|B)}{1-P(A|B)}.\frac{1-P(A|\bar{B})}{P(A|\bar{B})}$$

Desse modo podemos interpretar o coeficiente β_1 de uma regressão logística de maneira mais direta. No caso, e^{β_1} representa a razão de chances entre o evento $Y=1$ dado $X=(X_1+1,X_2,...,X_k)$ e $Y=1$ dado $X=(X_1,X_2,...,X_k)$, que é uma constante. Assim, e^{β_1} representa o aumento nas chances de observar $Y=1$ dado que houve uma aumento de uma unidade na variável X_1. Abaixo segue a derivação desse resultado:

$$O_{Y|X_1+1,X_2,...,X_k} = \frac{P(Y|X_1+1,X_2,...,X_k)}{1-P(Y|X_1+1,X_2,...,X_k)} = \frac{m(X\beta)}{1-m(X\beta)} = \frac{e^{X\beta}}{1+e^{X\beta}}.\frac{1}{1-\frac{e^{X\beta}}{1+e^{X\beta}}}$$

$$O_{Y|X_1+1,X_2,...,X_k} = \frac{e^{X\beta}}{1+e^{X\beta}}.\frac{1+e^{X\beta}}{1+e^{X\beta}-e^{X\beta}} = e^{X\beta} = e^{\beta_0+\beta_1.(X_1+1)+\beta_2.X_2+...+\beta_k.X_k}$$

De forma análoga:

$$O_{Y|X_1,X_2,...,X_k} = e^{\beta_0+\beta_1.X_1+\beta_2.X_2+...+\beta_k.X_k}$$

A razão de chances entre os eventos $Y=1$ dado um aumento de uma unidade na variável X_1 e $Y=1$ dado que as variáveis de controle permanecem no mesmo nível é dada por:

$$OR_{Y|X_1+1} = \frac{O_{Y|X_1+1,X_2,...,X_k}}{O_{Y|X_1,X_2,...,X_k}} = \frac{e^{\beta_0+\beta_1.(X_1+1)+\beta_2.X_2+...+\beta_k.X_k}}{e^{\beta_0+\beta_1.X_1+\beta_2.X_2+...+\beta_k.X_k}} = e^{\beta_1}$$

$$OR_{Y|X_1+1} = e^{\beta_1}$$

Assim para calcular a razão de chances de um aumento em uma unidade na variável de controle X_i devemos calcular o exponencial do coeficiente associado a tal variável e^{β_i}.

Voltando ao nosso exemplo. A razão de chances entre o $Y = 1$ dado o aumento de uma unidade da variável X_i e $Y = 1$ considerando todos os regressores constantes é dado por:

$$exp(\hat{\beta}_i)$$

Se esse valor for maior do que 1, então a variável X_i aumenta a chance de $Y_i = 1$ ocorrer, se for negativo, então reduz a chance desse evento ocorrer.

Em nosso exemplo:

$$exp(\hat{\beta}_1) = exp(0.1314) = 1.14$$

Para cada aumento em um ano de idade, a chance do indivíduo gostas do produto aumenta em 14%.

Vamos agora acrescentar mais regressores.

```
reg<-glm(enjoyed ~ age + gender,data = tastesgreat,family=binomial())
reg
```

```
##
## Call:  glm(formula = enjoyed ~ age + gender, family = binomial(), data = tastesgreat)
##
## Coefficients:
## (Intercept)          age   genderMale
##     -8.1844       0.1649       2.4224
##
## Degrees of Freedom: 39 Total (i.e. Null);  37 Residual
## Null Deviance:     55.45
## Residual Deviance: 38.98  AIC: 44.98
```

O fato de ser homem aumenta a chance de gostar do produto em:

$$exp(\hat{\beta}_2) = exp(2.4224) = 11.27$$

Esse produto parece então ser mais indicado para homens.
Podemos usar essa estimação da amostra para fazer previsão por meio da função `predict`.Por exemplo.

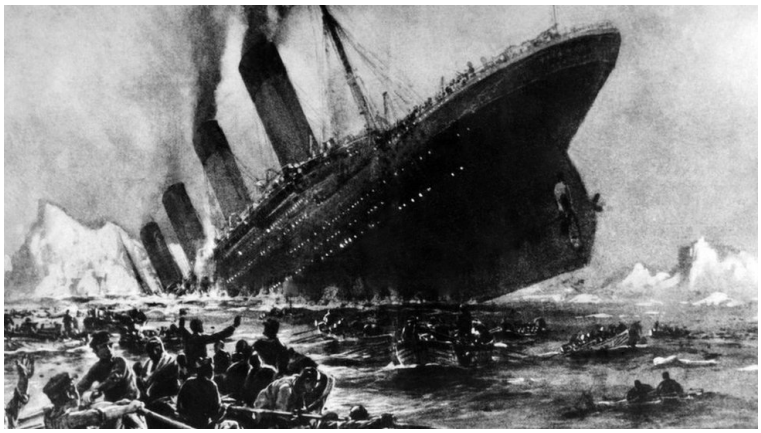
1. Qual a probabilidade de uma mulher de 34 anos gostar do produto?
2. E de um homem de $25,2$ anos?

```
new<-data.frame(gender = c('Female','Male'),age = c(34,25.5))
predict(reg,new,type='response')
```

```
##          1          2
## 0.07058988 0.17408497
```

Exercício Titanic

Considere a base de dados `train.csv` com os dados do Titanic do Kaggle.



Fonte: http://news.bbcimg.co.uk/media/images/59467000/jpg/$_59467081_illustrationofsinking96518519.jpg$

Estime a probabilidade de um indivíduo morrer dado o preço que pagou pela passagem, sua idade e seu sexo por meio de um modelo logístico.

Exercícios adicionais

Exercício 12.1.

Considere a base de dados `pnad2015_sub.csv`. Ela representa uma subamostra da PNAD do ano de 2015, com variáveis selecionadas.

1. Apresente a informação `renda_total` em um box-plot e em um histograma. Note que a informação fica muito espalhada, para visualizar melhor vamos fazer uma transformação monotônica utilizando a função logarítima.
2. Defina a variável `log_rt` como sendo o logarítmo neperiano de renda_total, atenção pois o logarítmo não está definido para valores não positivos. Apresente a informação de `log_rt` em um box-plot e em um histograma.
3. Apresente em um único gráfico um boxplot com `log_rt` para cada U.F. Apresente também a média nesses boxplots.
4. Repita o exercício anterior porém ao invés de U.F. faça um boxplot para cada elemento da variável `anos_estudo`. Calcule a correlação entre essas duas variáveis. (Dica: A função de correlação do `R` não reconhece o argumento `na.rm = T` como a função da média reconhece, para desconsiderar os `NA` use o argumento `use = 'complete.obs'`)
5. Apresente um gráfico com histogramas marginais onde no eixo x esteja a variável `anos_estudo` e no eixo y a variável `log_rt`.
6. Apresente um gráfico 3D com histograma conjunto onde no eixo x esteja a variável `anos_estudo` e no eixo y a variável `log_rt`.

Exercício 12.2.

Utilizando a base de dados `pnad2015_sub.csv` faça os seguintes testes de hipótese.

1. A variável `reg_nascimento` é uma variável binária que assume o valor (1) se a pessoa habita na UF de nascimento e (0) caso contrário. Verifique se a `renda_total` média é diferente nesses dois grupos. (Dica: utilize a função `t.test`)
2. Dentre o grupo de migrantes que não moram na região onde nasceram (isto é `reg_nascimente` $= 0$) verifique se o salário é diferente dos nascidos no nordeste e no sudeste. (Dica: Crie uma variável `sudeste <- c('SP','RJ','MG,'ES')` contendo as siglas dos estados. Use o comando `pnad2015$UF_nasc %in% sudeste` para selecionar somente as observações que pertencem à região Sudeste. Por fim utilize o operador lógico `&` para combinar essa condição com a condição de ser migrante.)
3. Teste se a proporção de indivíduos analfabetos no Amapá é diferente da proporção de indivíduos analfabetos em Sergipe. Teste também se a proporção de indivíduos analfabetos no Amapá é diferente da proporção nacional.

Exercício 12.3.

Nos exercícios anteriores verificamos que a média de salários é estatisticamente diferente entre homens e mulheres. Com a regressão linear podemos dizer mais, por exemplo se o prêmio de educação é diferente entre homens ou mulheres. Utilizando a base de dados `pnad2015_sub.csv` faça os seguintes exercícios:

1. Faça um regressão linear com a variável dependente sendo a `renda_total` e a variável independente sendo `anos_estudo`. Quanto é em média o prêmio, por ano de educação (isto é, se aumentar anos de estudo em 1 ano, quando aumenta a renda total)?
2. Acrescente um regressor que é `sexo`. Verifique se a média de salários de homens e mulheres é estatisticamente diferente, agora controlado por educação.
3. Acrescente a iteração de `sexo` com `anos_estudo` para verificar se o prêmio de estudo é diferente entre homens e mulheres.
4. Desta última regressão, qual seria a melhor previsão para a renda total de um homem com 5 anos de estudo? E para uma mulher com 11 anos de estudo?

5. Refaça os itens anteriores agora com a base de dados `pnad2005_sub.csv`. Quais as principais mudanças ocorridas nesses últimos 10 anos, com relação ao prêmio de estudo e "gap" salarial entre homens e mulheres que conseguimos cosntatar a partir das regressões feitas?

Exercício 12.4.

Voltando ao exemplo da melhor ração para alimentar os frangos da base de dados `ChickWeight`:

1. Estime um modelo linear com a variável dependente sendo o peso - `weigth` - e a variável independente sendo o dia - `Time`;
2. Apresente o gráfico com a variável `weight` no eixo vertical e a variável `Time` no eixo horizontal;
3. Estime um modelo linear acrescentando o tipo de alimentação - `Diet` - como regressor.

Bibliografia e Material complementar

Existe um volume imenso de material de estatística voltada a ciência de dados. Abaixo seguem algumas sugestões de livros e de materiais complementares ao curso para os que interessados continuem sua especialização.

1. Para teoria de estatística
   - Bussab, W. D. O., Morettin, P. A. (2010). Estatística básica. Saraiva.
   - George Casella, Roger L. Berger. (2002) Statistical inference, Ed. 2nd
   - Meyers, Paul L. (2013) Probabilidade: aplicações à estatística.
   - Stachurski, John. (2016) A Primer in Econometric Theory. MIT Press.
   - Introdução à Estatística: Tomando Decisões Com Base Em Dados in Udacity. https://br.udacity.com/course/intro-to-statistics--st101
   - Introduction to Statistics: Descriptive Statistics An introduction to descriptive statistics, emphasizing critical thinking and clear communication. in edx. https://www.edx.org/course/introduction-statistics-descriptive-uc-berkeleyx-stat2-1x
2. Estatística no R
   - Kerns, G. Jay. (2011) Introduction to Probability and Statistics Using R https://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf
   - Cohen, Y. and Cohen, J. Y. (2008). Statistics and Data with R: An applied approach through examples.John Wiley Sons.
   - Verzani, John. (2005) Using R for Introductory Statistics. HAPMAN HALL/CRC ftp://cran.r-project.org/pub/R/doc/contrib/Verzani-SimpleR.pdf
   - Statistics with R in Coursera. Duke University https://www.coursera.org/specializations/statistics
   - Statistics and R in edx. Harvard University https://www.edx.org/course/statistics-r-harvardx-ph525-1x-0
   - A Hands-on Introduction to Statistics with R in datacamp. https://www.edx.org/course/statistics-r-harvardx-ph525-1x-0
3. Data Science
   - R for Data Science. (2017). Grolemund, G. and Wickham, H. O'Reilly. http://r4ds.had.co.nz/
   - Curso do MIT no edx: Data Analysis for Social Scientists: Learn methods for harnessing and analyzing data to answer questions of cultural, social, economic, and policy interest. (Esther Duflo and Sara Fisher) https://www.edx.org/course/data-analysis-social-scientists-mitx-14-310x-3#!