MAP 2112 – Introdução à Lógica de Programação e Modelagem Computacional

1º Semestre – 2019

Prof. Dr. Luis Carlos de Castro Santos

lsantos@ime.usp.br/lccs13@yahoo.com

ROTEIRO

Material dos Profs. D.T. Kaplan, R.J; Pruim e N.J.Horton

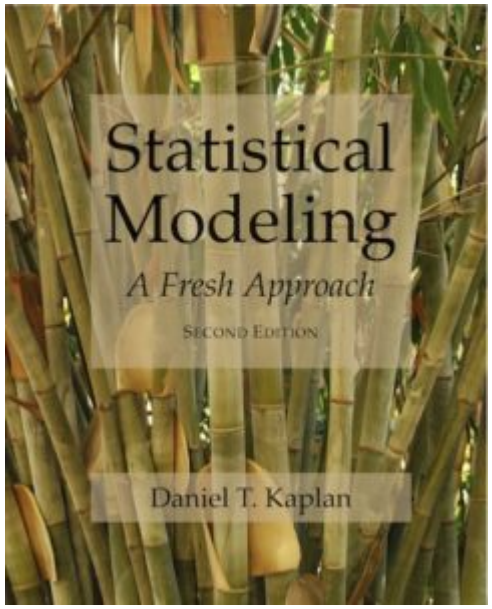
Projeto Mosaic

http://project-mosaic-books.com/?page_id=13

Material disponível html

http://mosaic-web.org/go/SM2-technique/

A partir do capítulo 3



Algum material básico sobre estatística será retirado do curso Genome560 do Prof. Akey da Universidade de Washington (UW)

http://www.gs.washington.edu/academics/courses/akey/56008/lecture.htm

O objetivo desse material é auxiliar na realização do trabalho em grupo. Não haverá cobrança em prova.

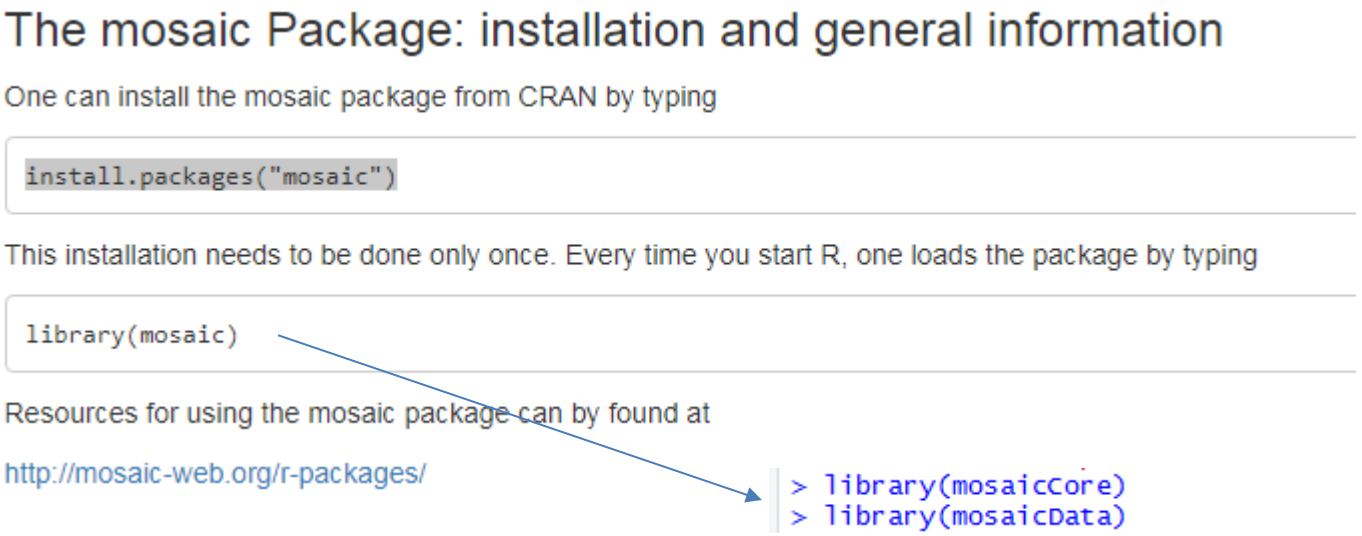
Nas próximas aulas:

- Conceitos básicos de estatística e modelagem

- Representação Gráfica

(com o foco no uso da linguagem R)

O projeto Mosaic desenvolveu alguns pacotes em R com o objetivo de facilitar o aprendizado:



Um outro pacote útil desenvolvido por um dos autores é o fastR (https://cran.r-project.org/web/packages/fastR/index.html) que pode ser carregado da mesma forma.

Chapter 3 Describing Variation

As a setting to illustrate computer techniques for describing variability, take the data that Galton collected on the heights of adult children and their parents. These data, in a modern case/variable format, are made available as `Galton` when the `mosaic` package is used.

```
> head(Galton)
  family father mother sex height nkids
1      1   78.5   67.0   M   73.2     4
2      1   78.5   67.0   F   69.2     4
3      1   78.5   67.0   F   69.0     4
4      1   78.5   67.0   F   69.0     4
5      2   75.5   66.5   M   73.5     4
6      2   75.5   66.5   M   72.5     4
```

Sir Francis Galton

https://en.wikipedia.org/wiki/Francis_Galton

http://mosaic-web.org/go/SM2-technique/describing-variation.html

3.1 Simple Statistical Calculations

Simple numerical descriptions are easy to compute. Here are the mean, median, standard deviation and variance of the children's heights (in inches).

```
> mean( ~ height, data = Galton)
[1] 66.76069
> median( ~ height, data = Galton)
[1] 66.5
> sd( ~ height, data = Galton)
[1] 3.582918
> var( ~ height, data = Galton)
[1] 12.8373
```

Why the tilde? In all these commands the first argument (`~ height`) is recognized by the interpreter as a ***model formula***. Model language involves the tilde, and the tilde must be followed by a variable.

3.1.1 Aside: The "base" versions

> The built-in, base R version of commands such as `mean()`, `median()`, `sd()`, and so on (that is, in their default non-`mosaic` form) take numeric vectors (columns of numbers) as arguments, but then the convenience of the `data =` designation is not available. For instance:

```
> mean(Galton$height)
[1] 66.76069
> median(Galton$height)
[1] 66.5
> sd(Galton$height)
[1] 3.582918
> var(Galton$height)
[1] 12.8373
```

SINTAXE

R “puro”

```
> mean(Galton$height)
[1] 66.76069
> median(Galton$height)
[1] 66.5
> sd(Galton$height)
[1] 3.582918
> var(Galton$height)
[1] 12.8373
```

Pacote Mosaic

```
> mean( ~ height, data = Galton)
[1] 66.76069
> median( ~ height, data = Galton)
[1] 66.5
> sd( ~ height, data = Galton)
[1] 3.582918
> var( ~ height, data = Galton)
[1] 12.8373
```

Qual o significado e o interesse nesses valores em relação a análise dos dados ?

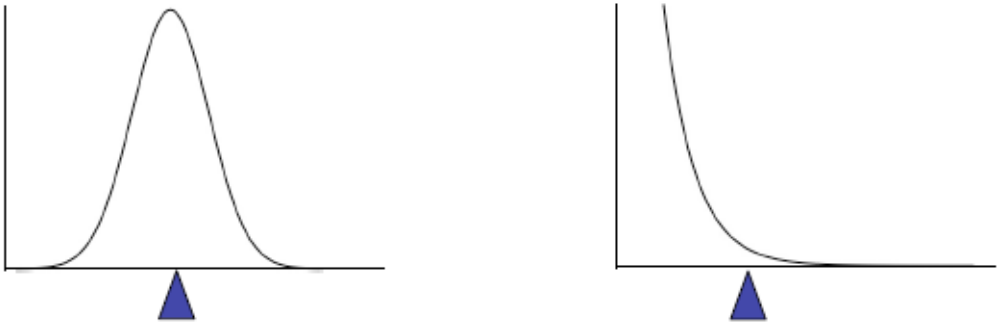

DESVIO

Location: Mean

I. The Mean

To calculate the average $\overline{x}$ of a set of observations, add their value and divide by the number of observations:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture2.pdf

Location: Median

- **Median** – the exact middle value

- **Calculation**:

ímpar

- If there are an odd number of observations, find the middle value

par

- If there are an even number of observations, find the middle two values and average them

- **Example**

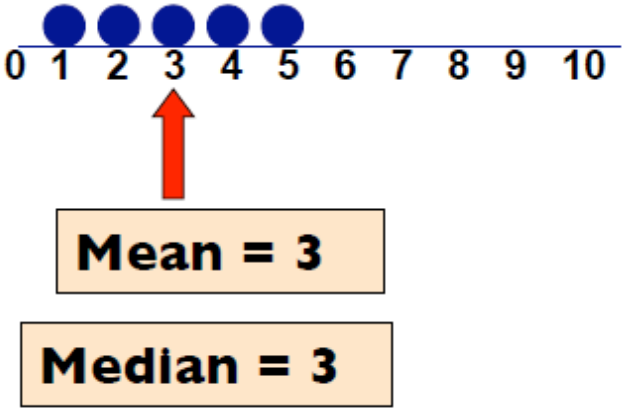
Some data:

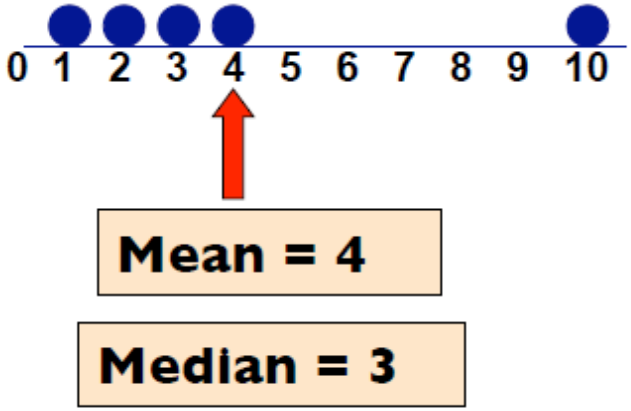
Age of participants: 17 19 21 22 23 23 23 38

Median = (22+23)/2 = 22.5

Which Location Measure Is Best?

- Mean is best for symmetric distributions without outliers
- Median is useful for skewed distributions or data with outliers





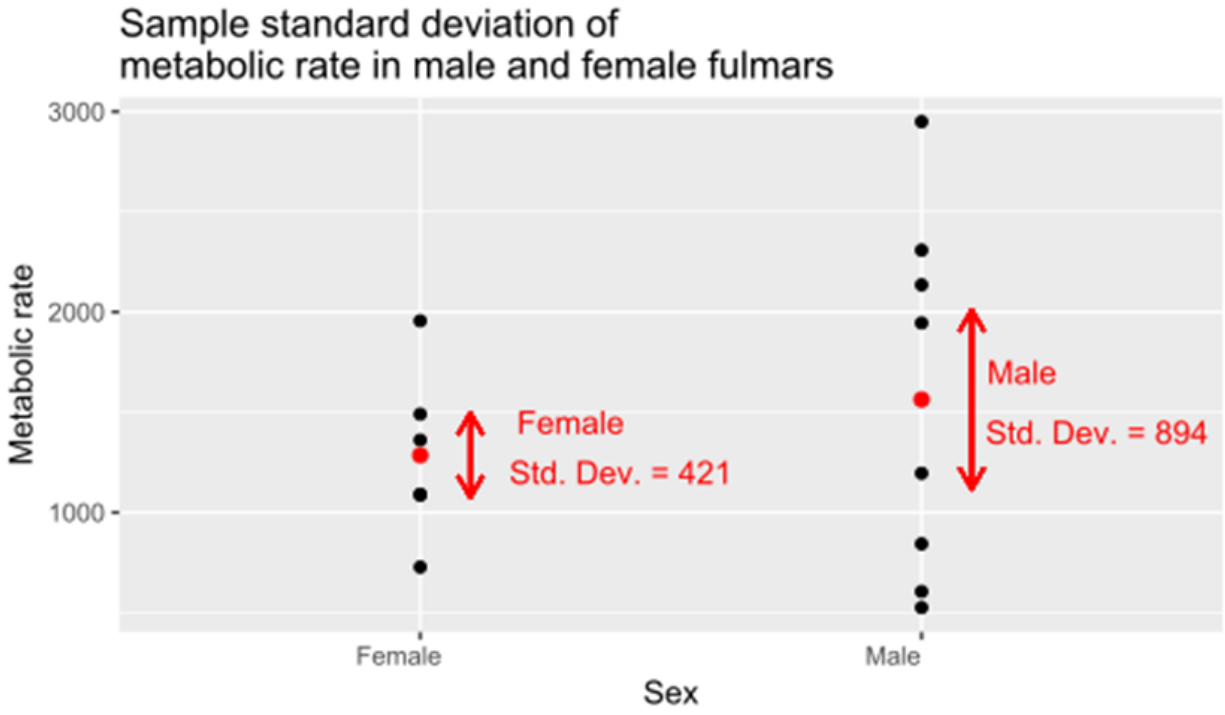
Scale: Standard Deviation

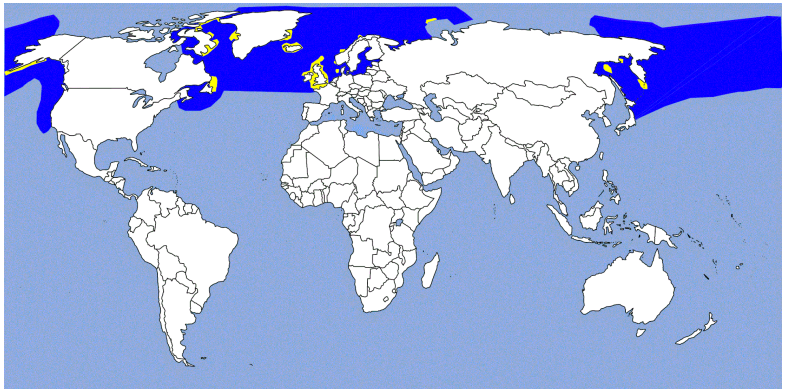
$$\hat{\sigma} = \sqrt{\frac{\sum_{i}^{n}(x_i - \bar{x})^2}{n-1}}$$

n se populacional

amostral

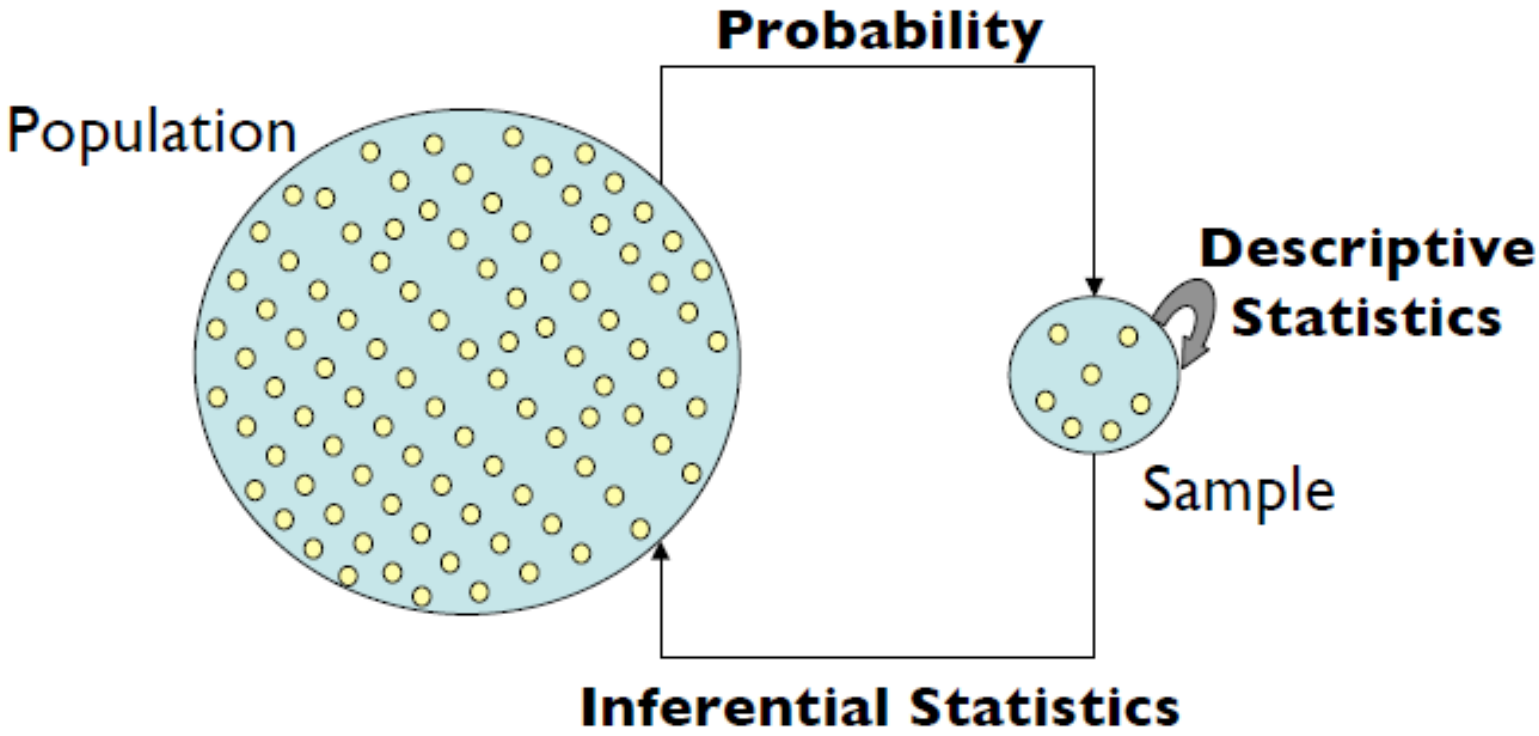
1. Score (in the units that are meaningful)
2. Mean
3. Each score's deviation from the mean
4. Square that deviation
5. Sum all the squared deviations (Sum of Squares)
6. Divide by n-1
7. Square root – now the value is in the units we started with!!!





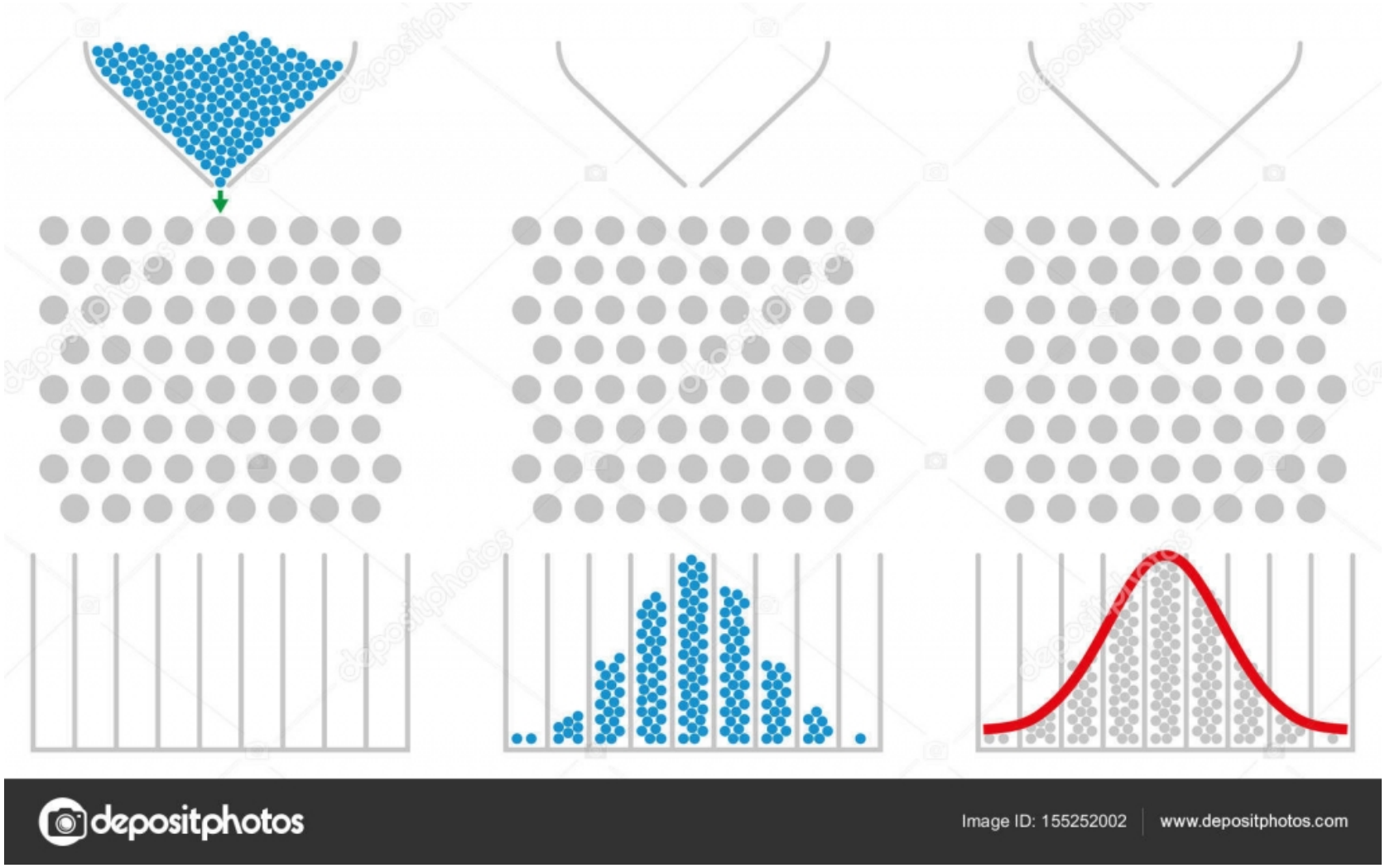
https://en.wikipedia.org/wiki/Standard_deviation

"Central Dogma" of Statistics

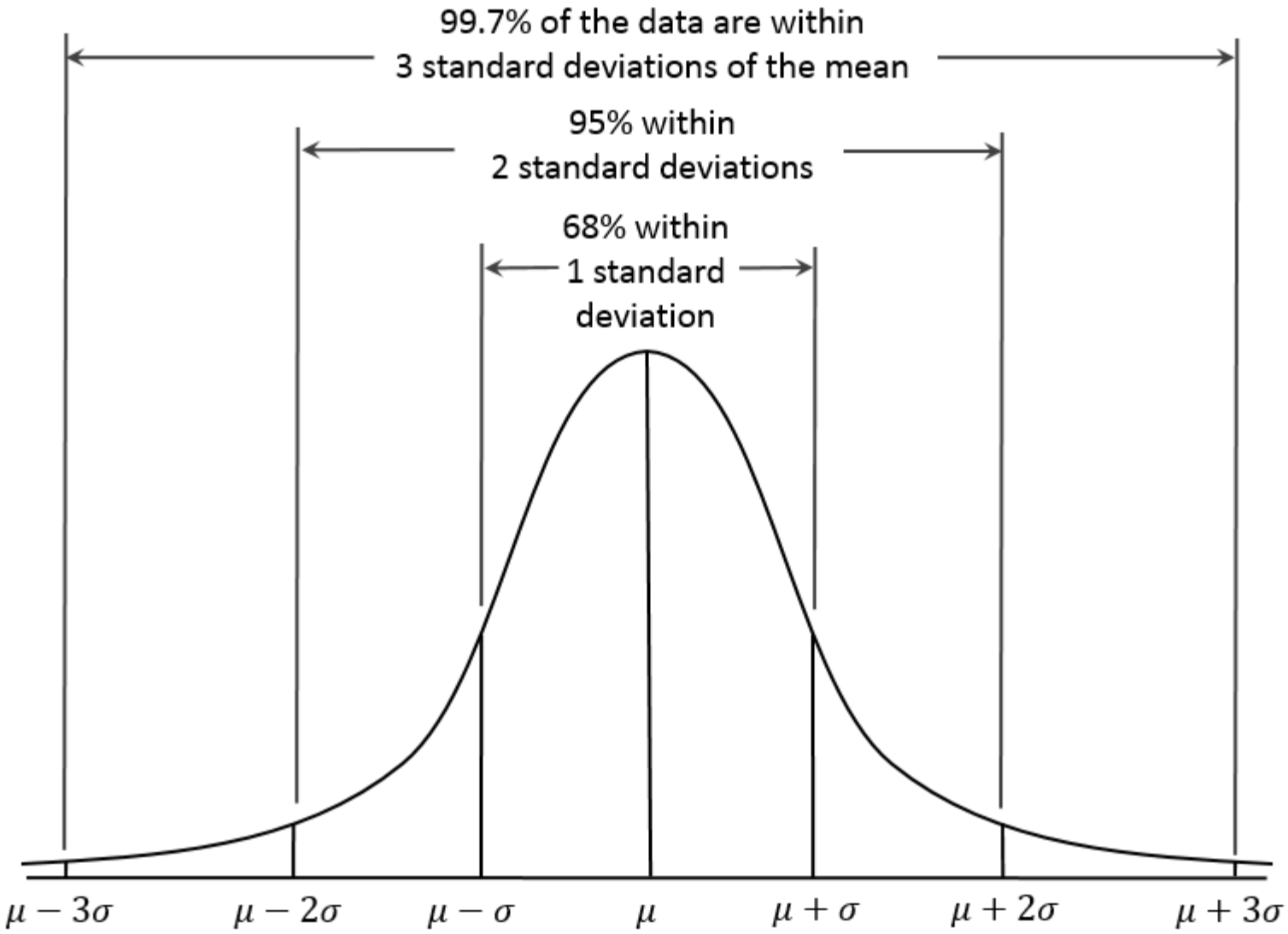


Teorema do limite central

- Quanto maior for o tamanho *n* da amostra, mais a média amostral se aproximará da média da população.
- As propriedades da distribuição amostral asseguram que a média de uma amostra é uma boa estatística para inferir sobre a média da população μ da qual foi extraída.
- Ao mesmo tempo, o teorema do limite central estabelece que se o tamanho da amostra *n* for suficientemente grande a distribuição da média amostral será normal, qualquer que seja a forma da distribuição da população.
- Portanto, o teorema do limite central permite aplicar a distribuição normal para obter respostas da média de uma amostra de tamanho suficientemente grande retirada de uma população qualquer.



CLT: samples of observations of random variables independently drawn from independent distributions converge in distribution to the normal

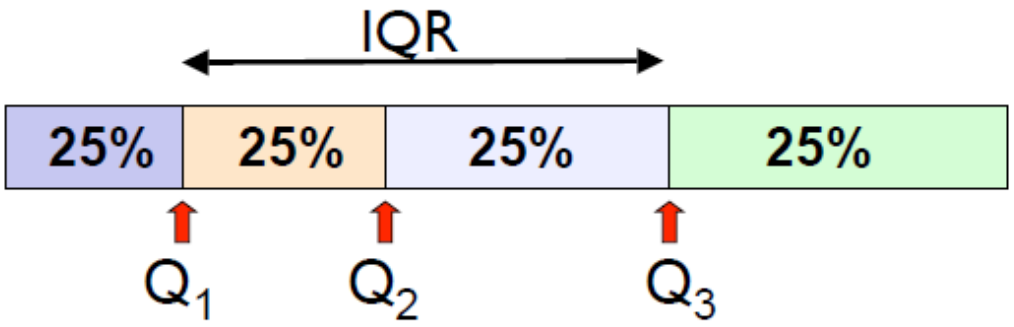

99.7% of the data are within
3 standard deviations of the mean
95% within
2 standard deviations
68% within
1 standard
deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Scale: Variance

- Average of squared deviations of values from the mean

$$\hat{\sigma}^2 = \frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n-1}$$

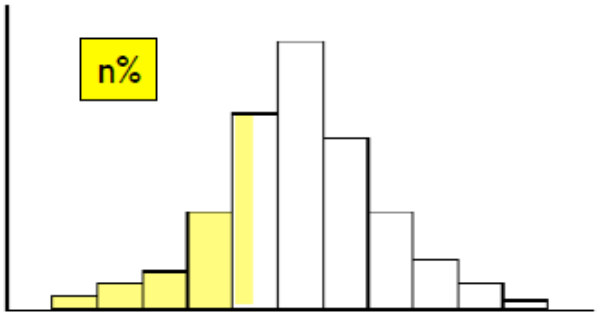
Scale: Quartiles and IQR



- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

Percentiles (aka Quantiles)

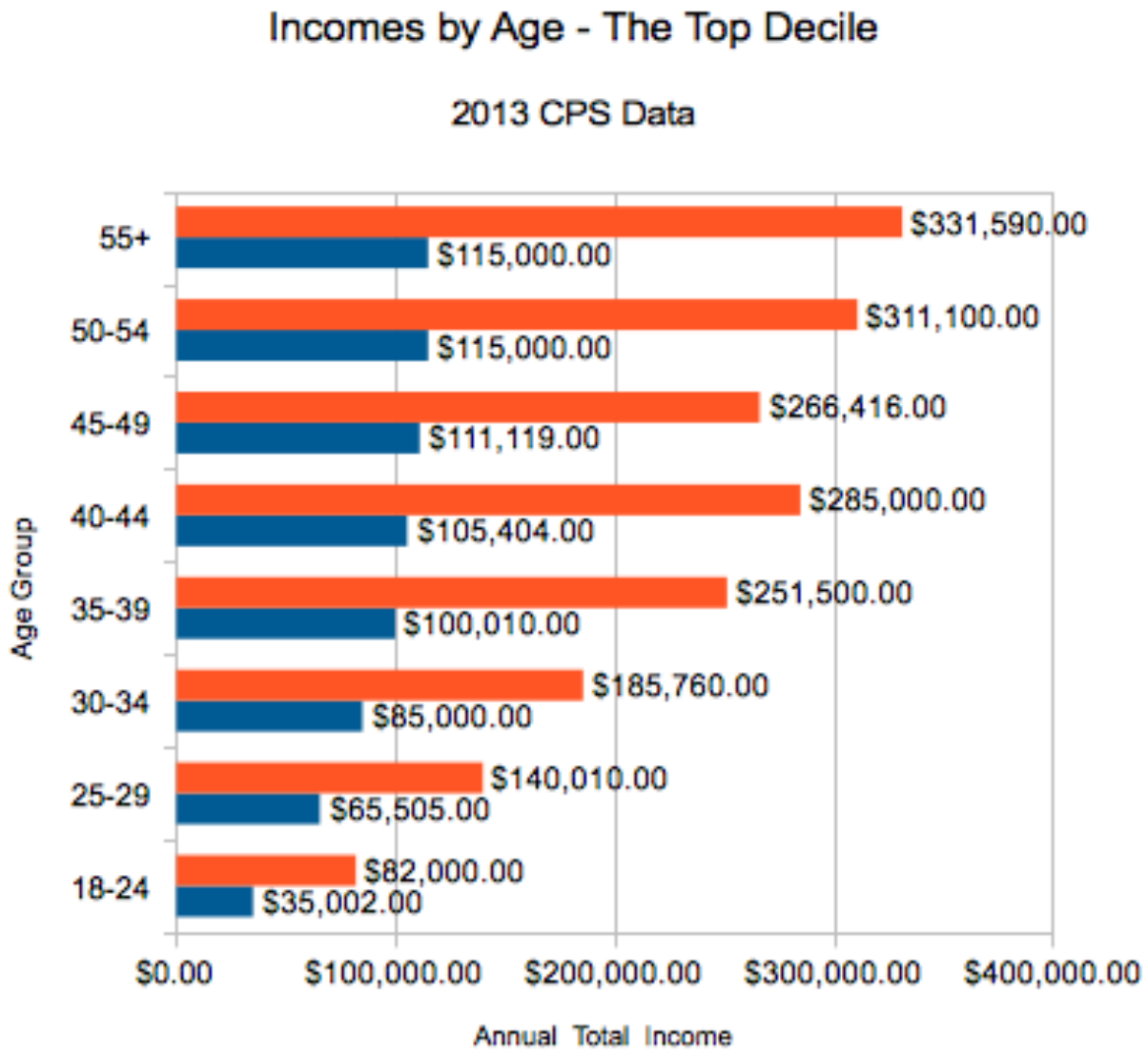
In general the **n^{th} percentile** is a value such that n% of the observations fall at or below or it



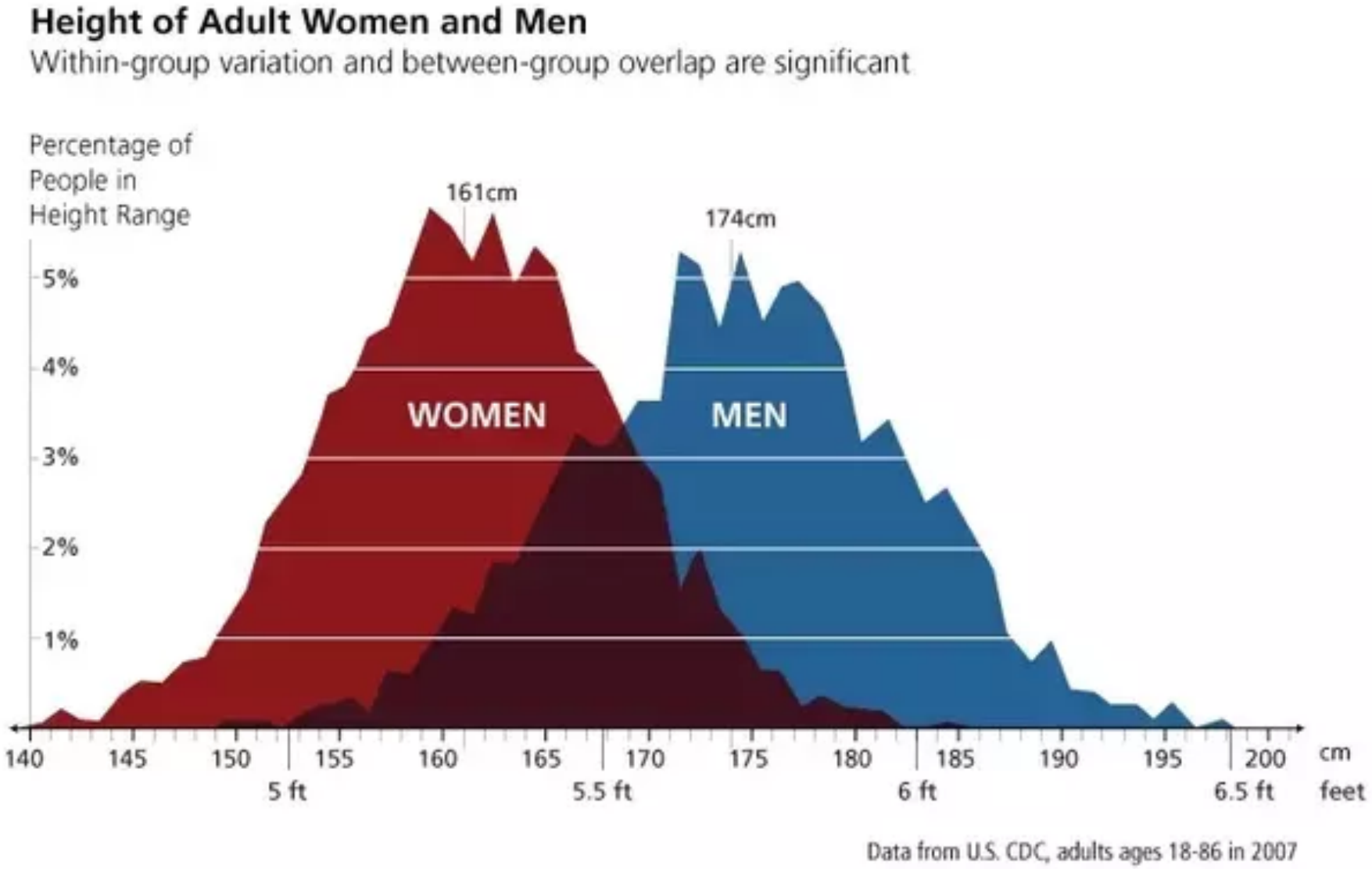
Q_1 = 25th percentile

Median = 50th percentile

Q_2 = 75th percentile



Orange: The 1%
Blue: The 10%
(99th quantile & 90th quantile)


Height of Adult Women and Men
Within-group variation and between-group overlap are significant
Percentage of People in Height Range
161cm
174cm
5%
4%
3%
2%
1%
WOMEN
MEN
140
145
150
155
160
165
170
175
180
185
190
195
200
cm
5 ft
5.5 ft
6 ft
6.5 ft
feet
Data from U.S. CDC, adults ages 18-86 in 2007


DESVIO

```
> head(Galton)
  family father mother sex height nkids
1      1   78.5   67.0   M   73.2     4
2      1   78.5   67.0   F   69.2     4
3      1   78.5   67.0   F   69.0     4
4      1   78.5   67.0   F   69.0     4
5      2   75.5   66.5   M   73.5     4
6      2   75.5   66.5   M   72.5     4
```

As we have seen before, R's model language also allows more sophisticated model arguments, as in

```
median(height ~ sex, data = Galton)
```

```
##    F    M
## 64.0 69.2
```

A percentile tells where a given value falls in a distribution. For example, a height of 63 inches is on the short side in Galton's data:

```
pdata( ~ height, 63, data = Galton)
```

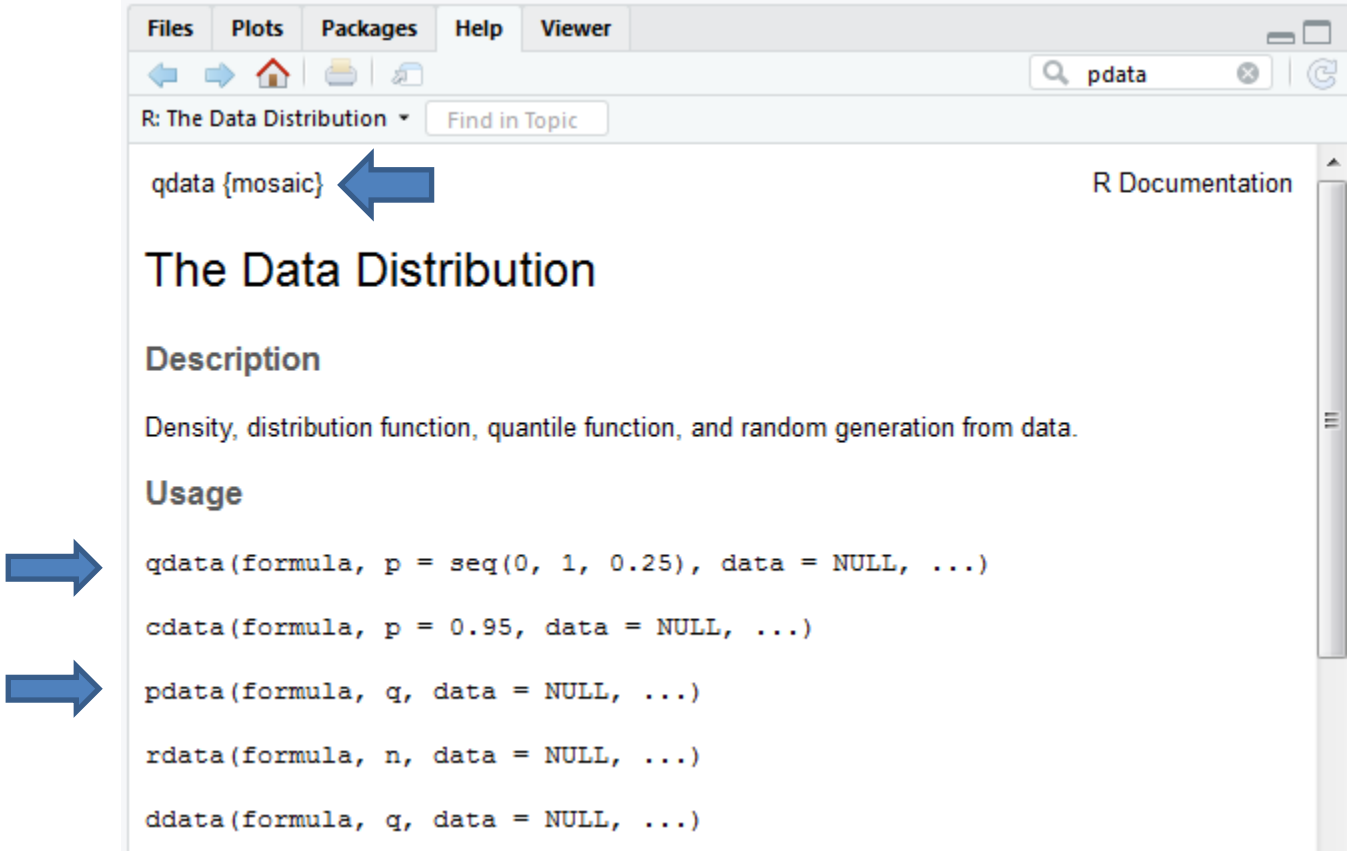
```
## [1] 0.1915367
```

Only about 19 % of the cases have a height less than or equal to 63 inches. The `pdata` operator takes one or more values as a second argument (here, 63) and finds where they fall in the distribution of values in the first argument (`height`).

A quantile refers to the same sort of calculation, but inverted. Instead of giving a value in the same units as the distribution, you give a probability: a number between 0 and 1. The `qdata` operator then calculates the value whose percentile would be that value. What's the 20th percentile of Galton's heights?

```
qdata( ~ height, 0.2, data = Galton)
```

```
##         p quantile
##       0.2     63.5
```


Files
Plots
Packages
Help
Viewer
pdata
R: The Data Distribution
Find in Topic
qdata {mosaic}
R Documentation
The Data Distribution
Description
Density, distribution function, quantile function, and random generation from data.
Usage
qdata(formula, p = seq(0, 1, 0.25), data = NULL, ...)
cdata(formula, p = 0.95, data = NULL, ...)
pdata(formula, q, data = NULL, ...)
rdata(formula, n, data = NULL, ...)
ddata(formula, q, data = NULL, ...)

Remember that the probability is given as a number between 0 and 1, so use 0.50 to indicate that you want the value which falls at the 50th percentile.

```
qdata( ~ height, 0.5, data = Galton)
```

```
##         p quantile
##       0.5     66.5
```

```
median( ~ height, data = Galton)
```

```
## [1] 66.5
```

1. The 25th and 75th percentile in a single command — in other words, the 50 percent coverage interval:

```
qdata( ~ height, c(0.25, 0.75), data = Galton)
```

```
##     quantile    p
## 25%     64.0 0.25
## 75%     69.7 0.75
```

1. The 2.5th and 97.5th percentile — in other words, the 95 percent coverage interval:

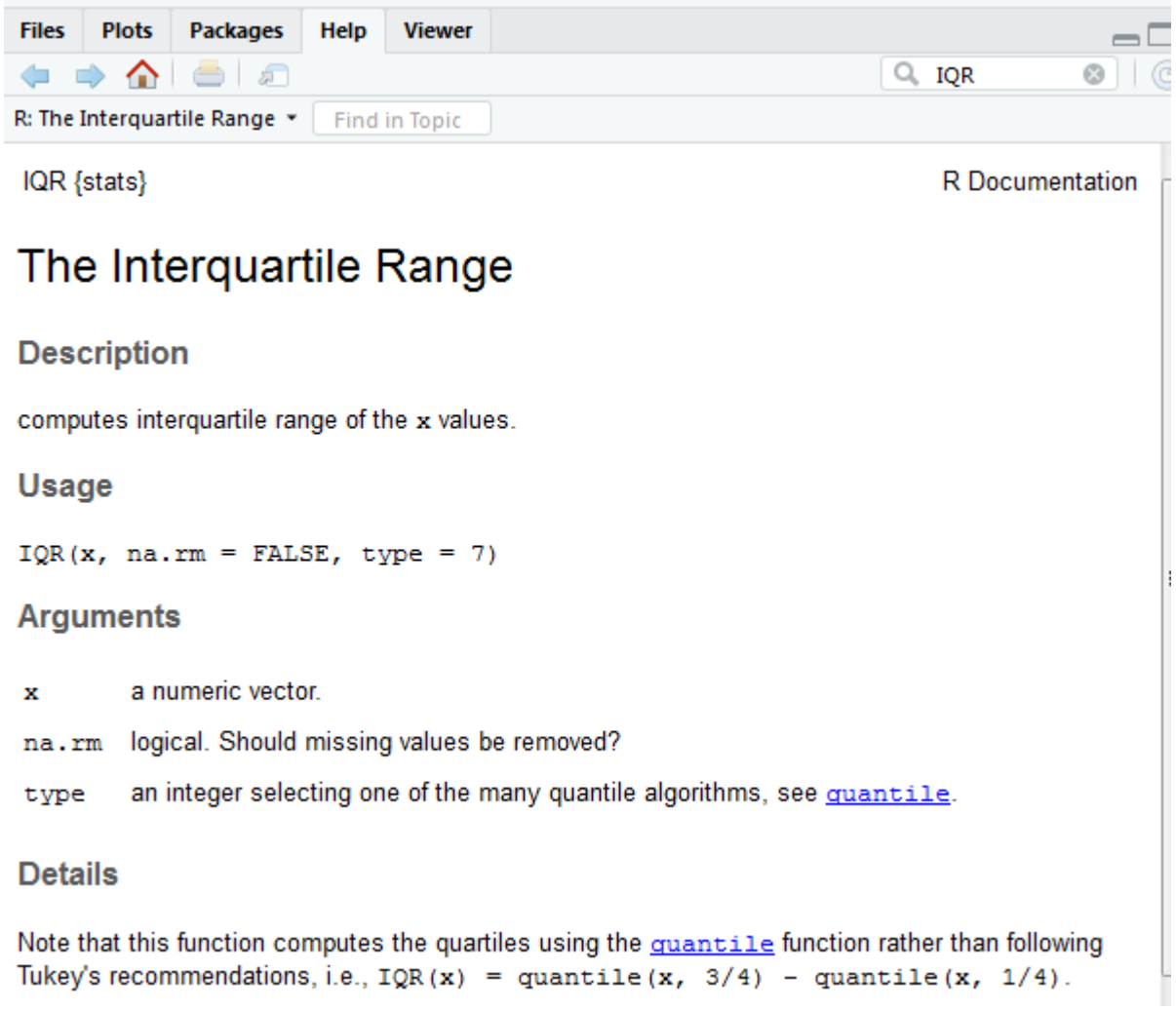
```
qdata( ~ height, c(0.025, 0.975), data = Galton)
```

```
##       quantile     p
## 2.5%        60 0.025
## 97.5%       73 0.975
```

The interquartile range is the width of the 50 percent coverage interval:

```
IQR( ~ height, data = Galton)
```

```
## [1] 5.7
```



Files | Plots | Packages | Help | Viewer

IQR

R: The Interquartile Range | Find in Topic

IQR {stats} R Documentation

The Interquartile Range

Description

computes interquartile range of the x values.

Usage

```
IQR(x, na.rm = FALSE, type = 7)
```

Arguments

x	a numeric vector.
na.rm	logical. Should missing values be removed?
type	an integer selecting one of the many quantile algorithms, see quantile.

Details

Note that this function computes the quartiles using the quantile function rather than following Tukey's recommendations, i.e., IQR(x) = quantile(x, 3/4) - quantile(x, 1/4).

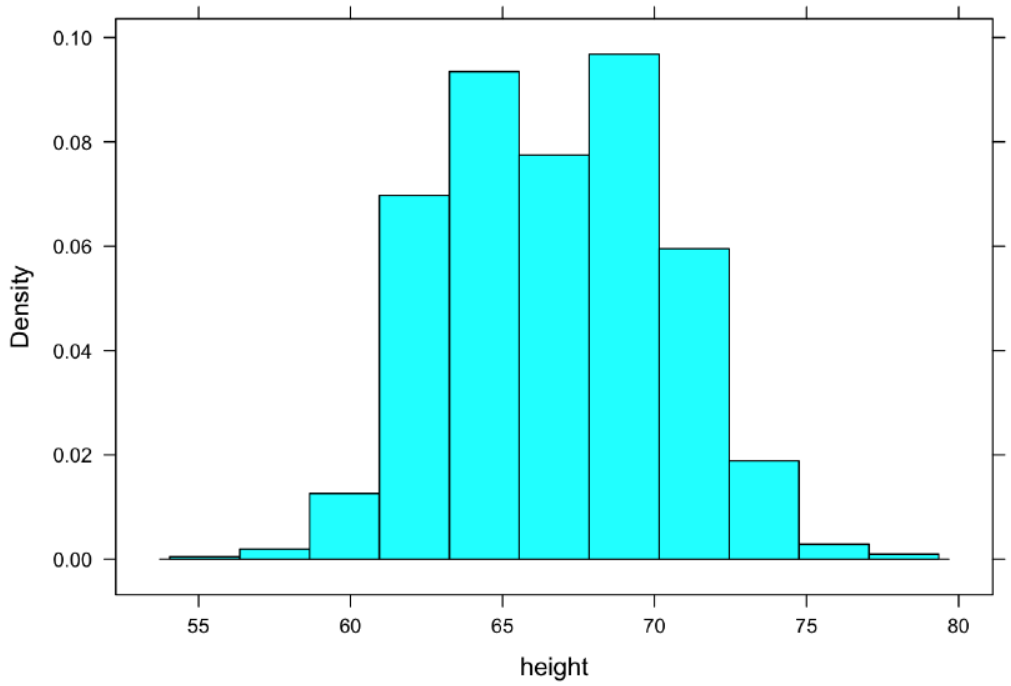
3.2 Simple Statistical Graphics

There are several basic types of statistical graphics to display the distribution of a variable: histograms, density plots, and boxplots. These are easily mastered by example.

3.2.1 Histograms and Distributions

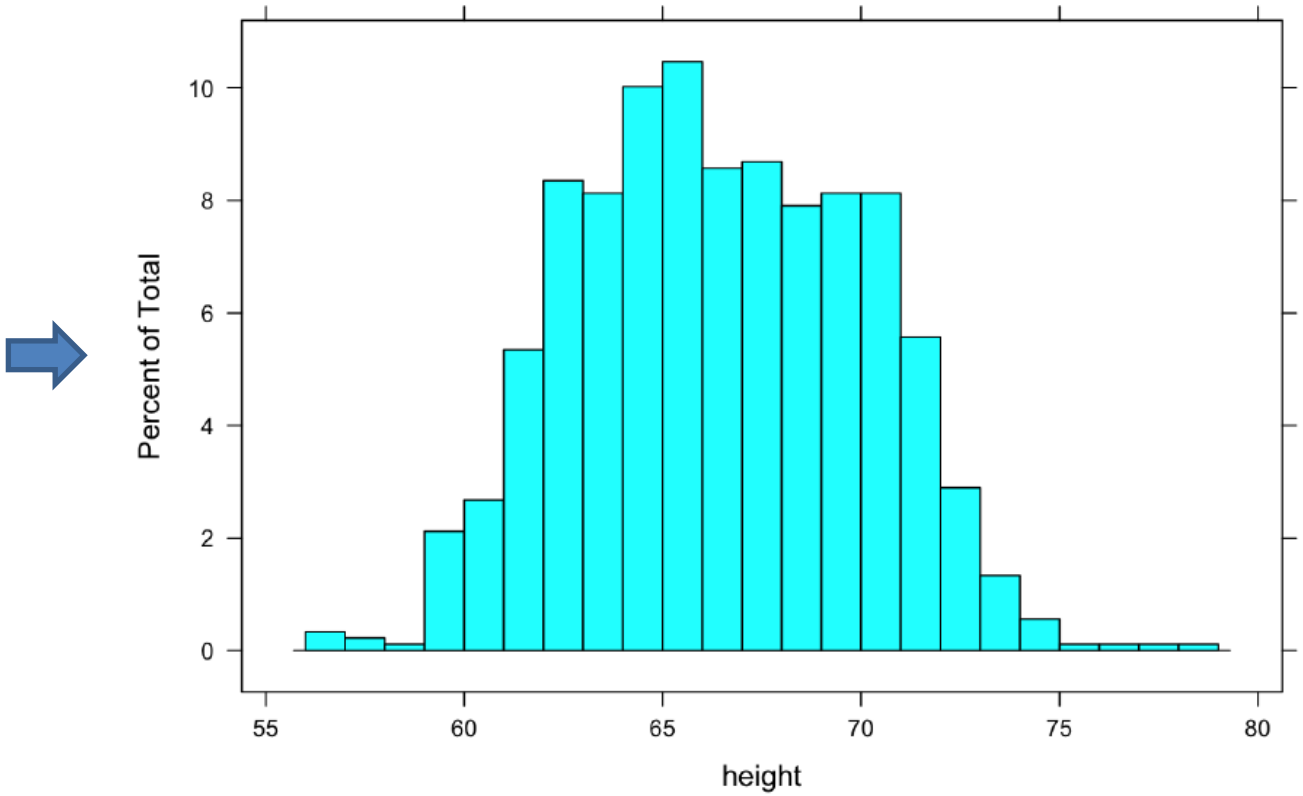
Constructing a histogram involves dividing the range of a variable up into bins and counting how many cases fall into each bin. This is done in an almost entirely automatic way:

```
histogram( ~ height, data = Galton)
```



When constructing a histogram, R makes an automatic but sensible choice of the number of bins. If you like, you can control this yourself. For instance:

```
histogram( ~ height, data = Galton, breaks = 25 )
```



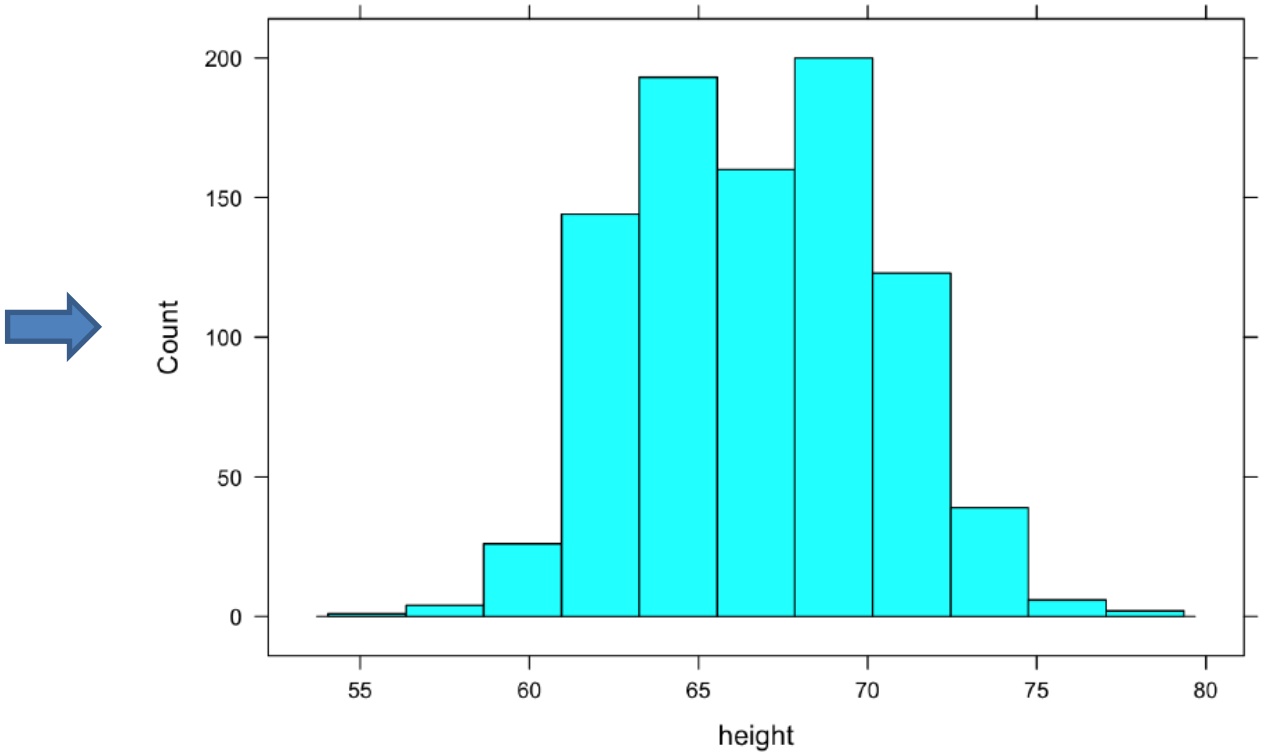
The horizontal axis of the histogram is always in the units of the variable. For the histograms above, the horizontal axis is in "inches" because that is the unit of the `height` variable.

The vertical axis is conventionally drawn in one of three ways: controlled by an optional argument named `type` .

1. **Absolute Frequency or Counts**

A simple count of the number of cases that falls into each bin. This mode is set with `type="count"` as in

```
histogram( ~ height, data = Galton, type = "count")
```



1. **Relative Frequency**

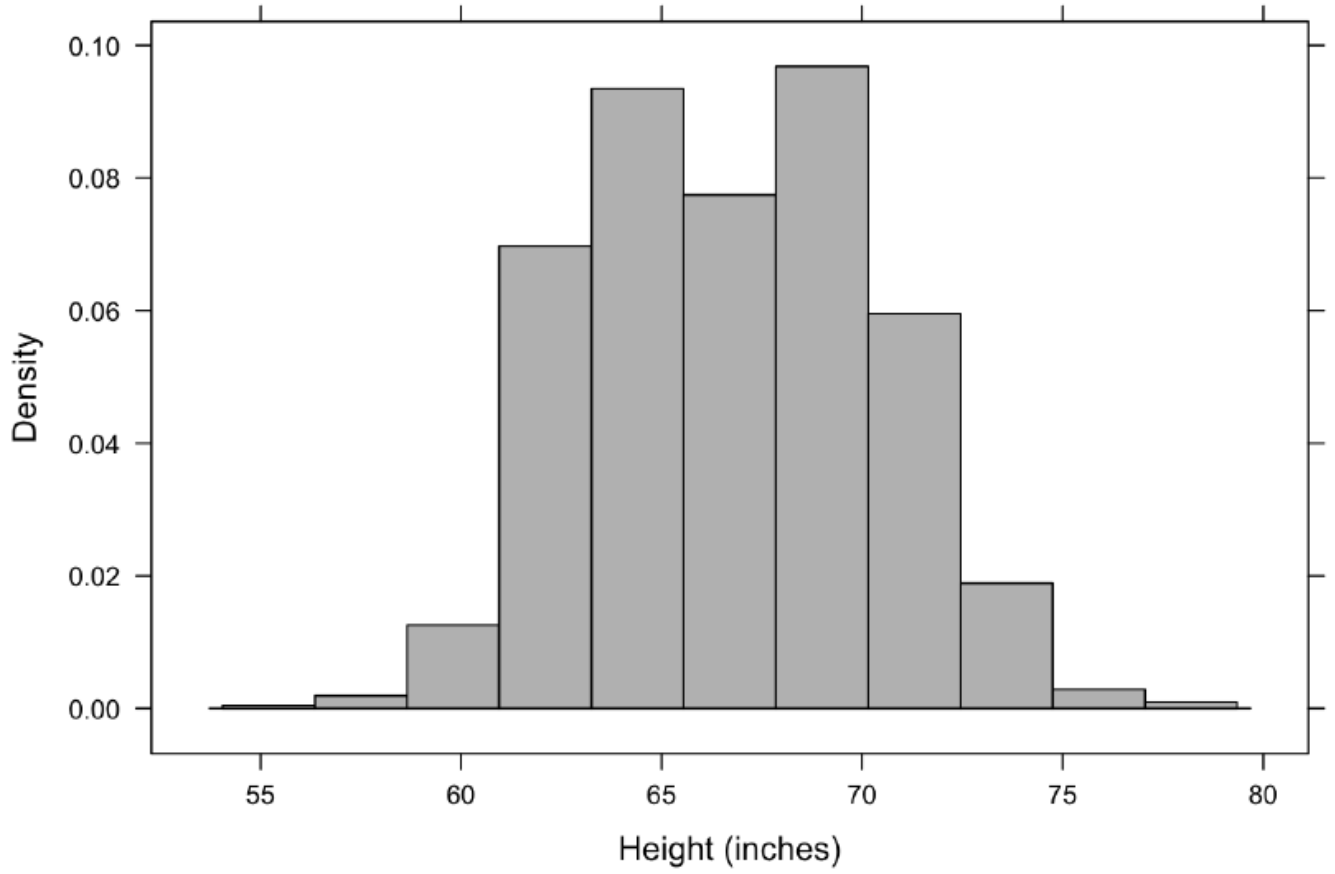
The vertical axis is scaled so that the height of the bar give the proportion of cases that fall into the bin. This is the default, that is, this is the `type` if you don't specify `type` in the command.

1. **Density**

The vertical axis *area* of the bar gives the relative proportion of cases that fall into the bin. Set `type = "density"` . In a density plot, areas can be interpreted as probabilities and the area under the entire histogram is equal to 1.

Other useful optional arguments set the labels for the axes and the graph as a whole and color the bars. For example,

```
histogram( ~ height, data = Galton, type = "density",
     xlab="Height (inches)",
     main="Distribution of Heights",
     col="gray")
```

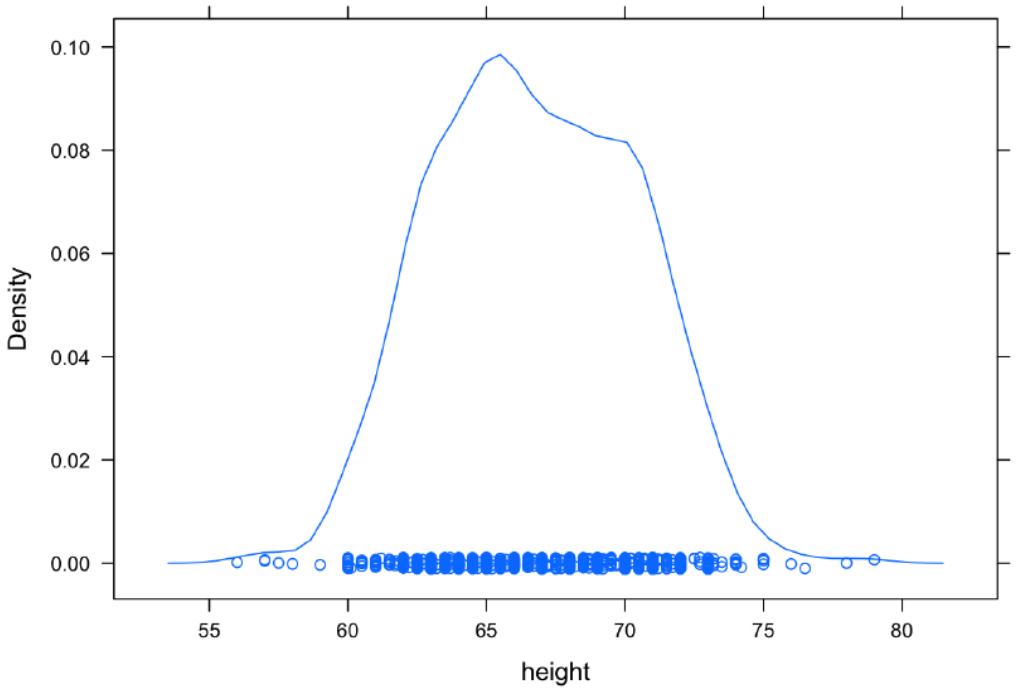


The above command is so long that it has been broken into several lines for display purposes. R ignores the line breaks, holding off on executing the command until it sees the final closing parentheses. Notice the use of quotation marks to delimit the labels and names like `"blue"`. Also note the `+` signs that appear in place of the prompt (`>`) on new lines when the command is still incomplete. If a plus sign appears when you are expecting a prompt, R is telling you that the previous command is incomplete. You need to complete it before you get back to the prompt.

3.2.2 Density Plots

A ***density plot*** avoids the need to create bins and plots out the distribution as a continuous curve. Making a density plot involves two operators. The ***density*** operator performs the basic computation which is then displayed using either the `plot` or the `lines` operator. For example:

```
densityplot( ~ height, data = Galton)
```



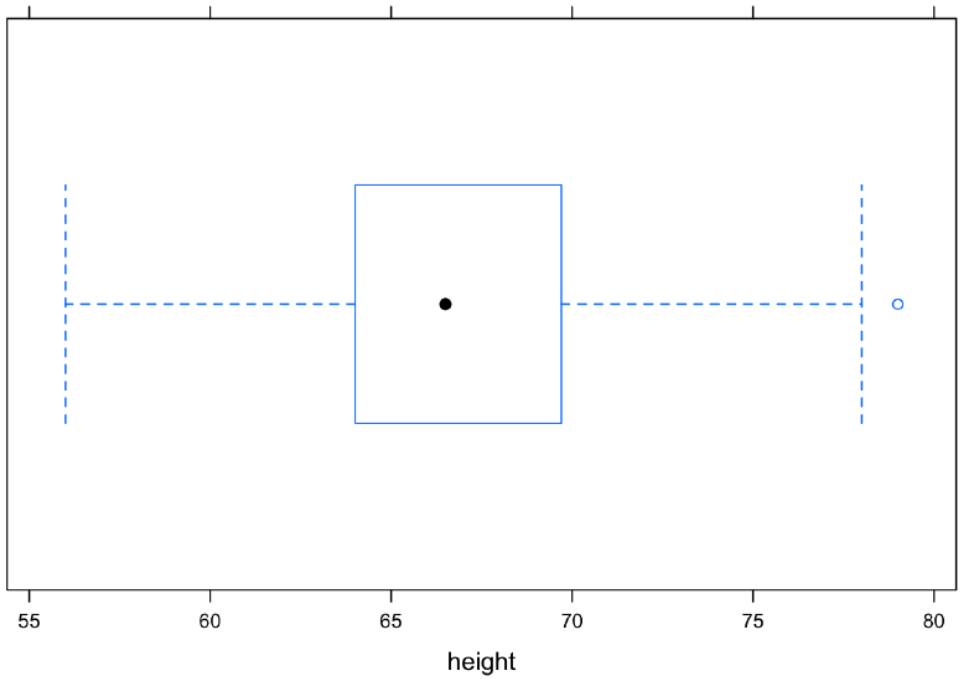
If you want to suppress the rug-like plotting of points at the bottom of the graph, use

```
densityplot( ~ height, data = Galton, plot.points = FALSE) .
```

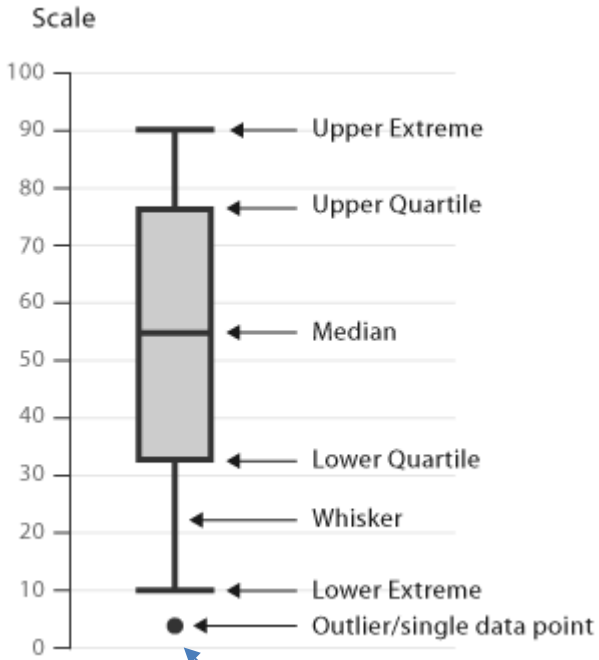
3.2.3 Box-and-Whisker Plots

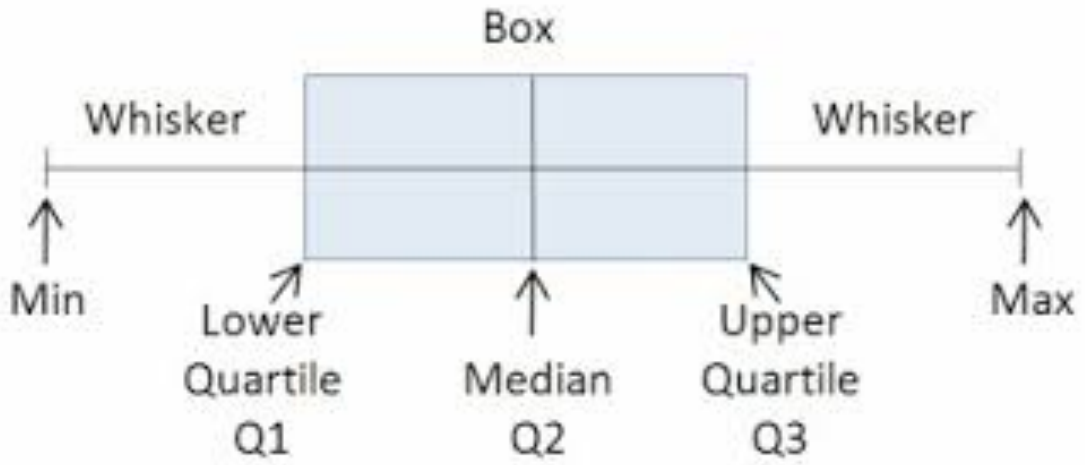
Box-and-whisker plots are made with the `bwplot` command:

```
bwplot( ~ height, data = Galton)
```



The median is represented by the heavy dot in the middle. Outliers, if any, are marked by dots outside the whiskers.

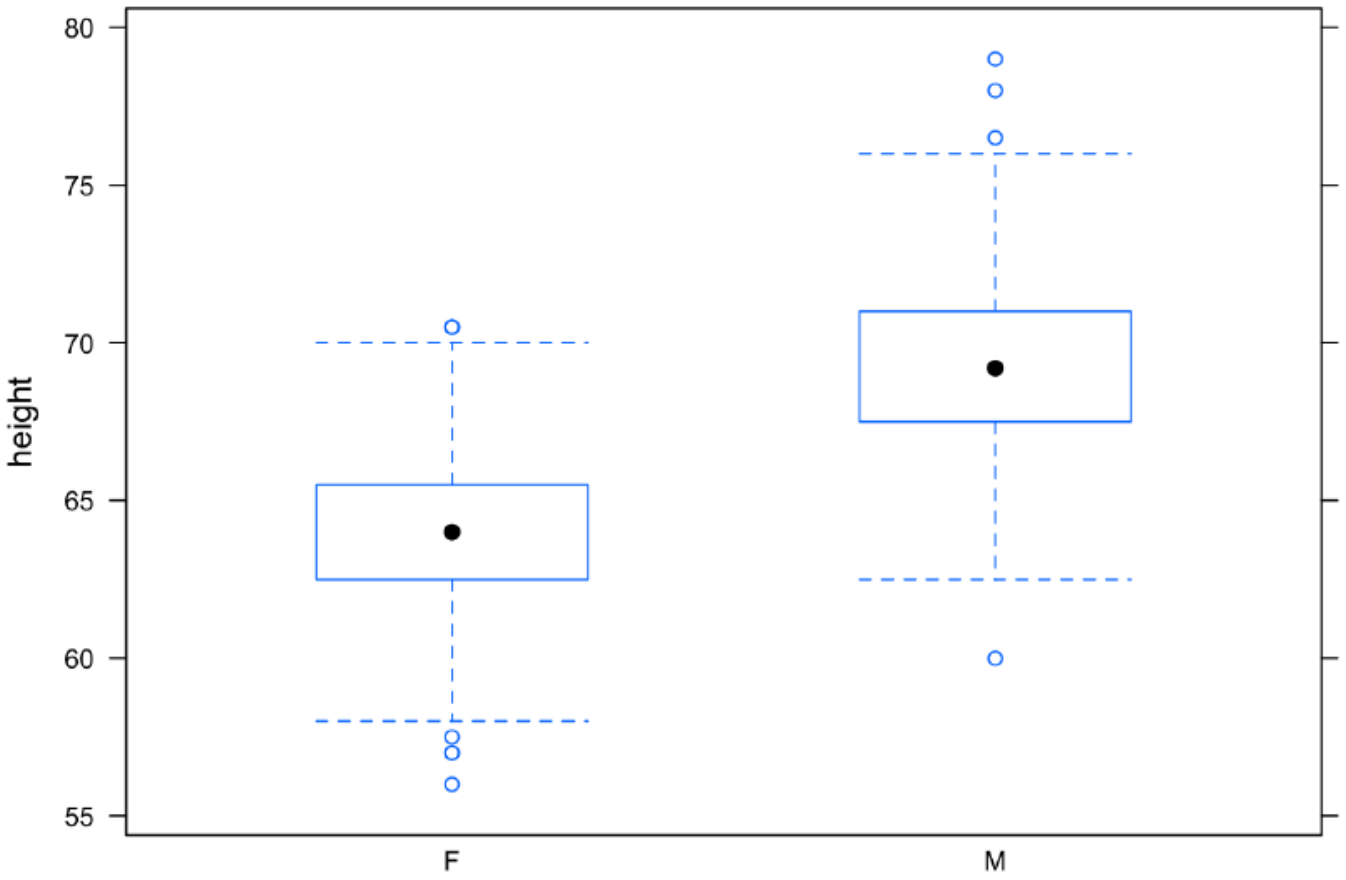




outliers are those observations that lie outside 1.5 * IQR

The real power of the box-and-whisker plot is for comparing distributions. This will be raised again more systematically in later chapters, but just to illustrate, here is how to compare the heights of males and females:

```
bwplot(height ~ sex, data = Galton)
```



Que conclusão poderíamos tirar desse gráfico ?

3.3 Displays of Categorical Variables

For categorical variables, it makes no sense to compute descriptive statistics such as the mean, standard deviation, or variance. Instead, look at the number of cases at each level of the variable.

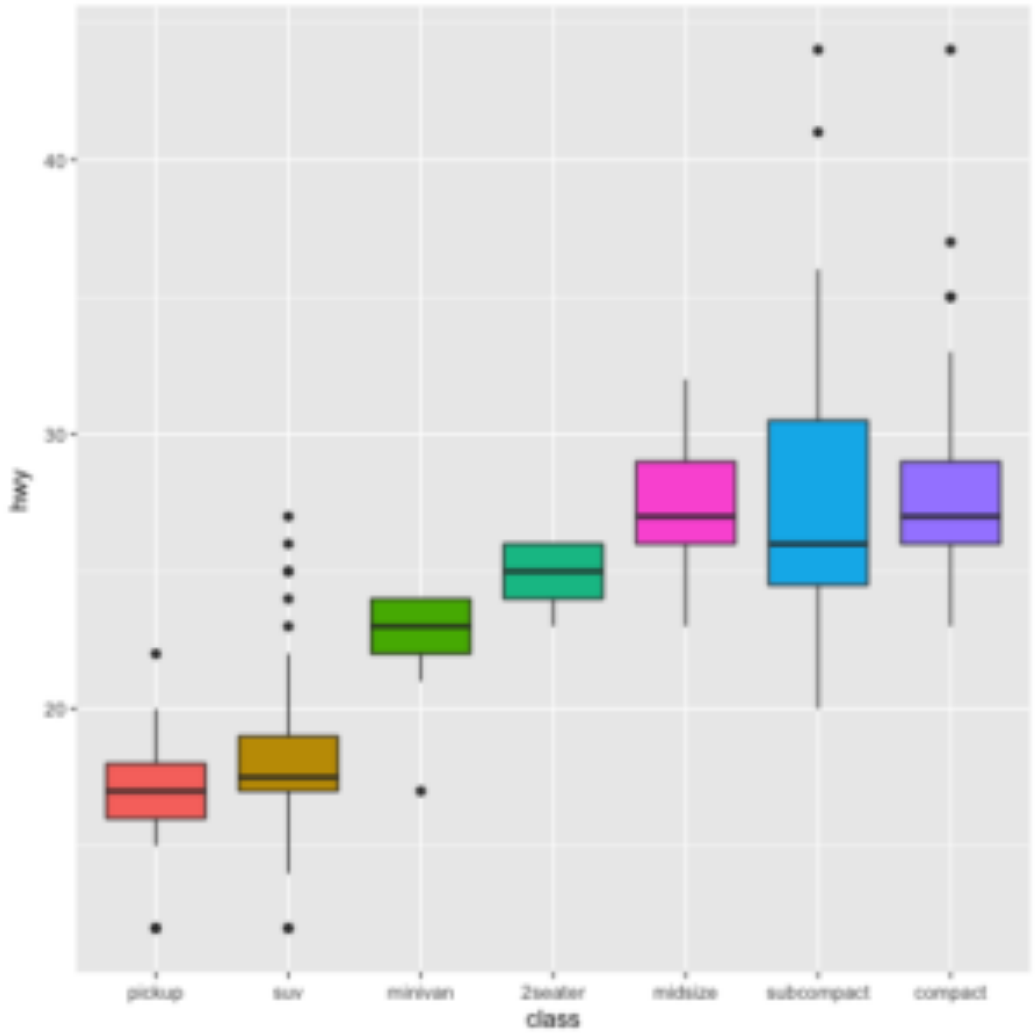
```
tally( ~ sex, data = Galton)
```

```
##
##   F   M
## 433 465
```

Proportions can be found by dividing the tallies by the total:

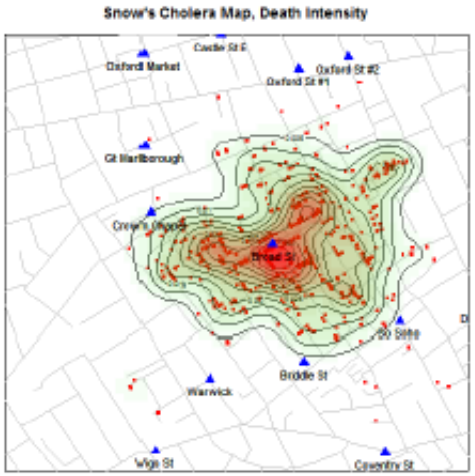
```
tally( ~ sex, data = Galton)/nrow(Galton)
```

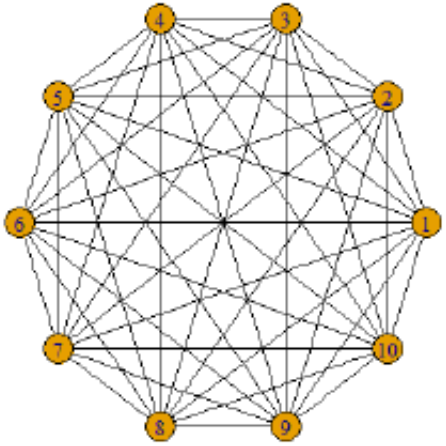
```
##
##         F         M
## 0.4821826 0.5178174
```


hwy
40
30
20
pickup
suv
minivan
2seater
midsize
subcompact
compact
class

Curso de Gráficos em R

http://www.datavis.ca/courses/RGraphics/





An introduction to R Graphics

Michael Friendly

SCS Short Course

March, 2018

http://datavis.ca/courses/RGraphics/

Statistical Consulting Service

YORK UNIVERSITÉ UNIVERSITY