MET CS 555 - Data Analysis and Visualization

Module-1: Introduction, Data Summarization, Normal Distribution
Lecture - 1

Mohammad Alaghemandi

Boston University

Table of contents

Basic R Programming

Why R?

- Freely available under the GNU General Public License
- Pre-compiled binary versions are provided for various operating systems
- Easy to install. Ready to use in a few minutes, frequent updates
- A few thousand supplemental packages
- Open source with a large support community: easy to find help!
- Many books, blogs, tutorials.
- More popular than major statistics packages (SAS, Stata, SPSS etc.)
- Getting more interest "Python" with packages Numpy, SciPy https://www.scipy.org/ Python Visualization package matplotlib https://matplotlib.org/
- As a data scientists you should learn python and R, and ...

R Package

- Open source programming language for statistical computing and graphical visualizations
- It is part of GNU project (`https://www.gnu.org/gnu/thegnuproject.en.html`)
- Written primarily in C and Fortran
- Available for various operating systems: Unix/Linux, Windows, Mac
- Can be downloaded and installed from the Comprehensive R Archive Network http://cran.r-project.org/

Resources to Learn R or ask Questions

- Textbooks
- R project website (http://www.r-project.org)
- R specific search engine (http://rseek.org)
- Search on the Web
- Ask questions on our **"Class Discussion Board"**
  You can ask questions anonymously.
  Useful:
  "How To Ask Questions The Smart Way" by Eric Steven Raymond
  `http://www.catb.org/esr/faqs/smart-questions.html`

Online Books

- An introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics, by W. N. Venables, et al.

  `http://onlinebooks.library.upenn.edu/webbin/book/lookupid?key=olbp44950`

- SimpleR - Using R for Introductory Statistics, by John Verzani.

  `https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf`

- R for Beginners, by Emmanuel Paradis.

  `https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf`

- The R Guide, by W. J. Owen.

  `https://cran.r-project.org/doc/contrib/Owen-TheRGuide.pdf`

- Institute for Digital Research and Education (UCLA)

  `http://www.ats.ucla.edu/stat/`

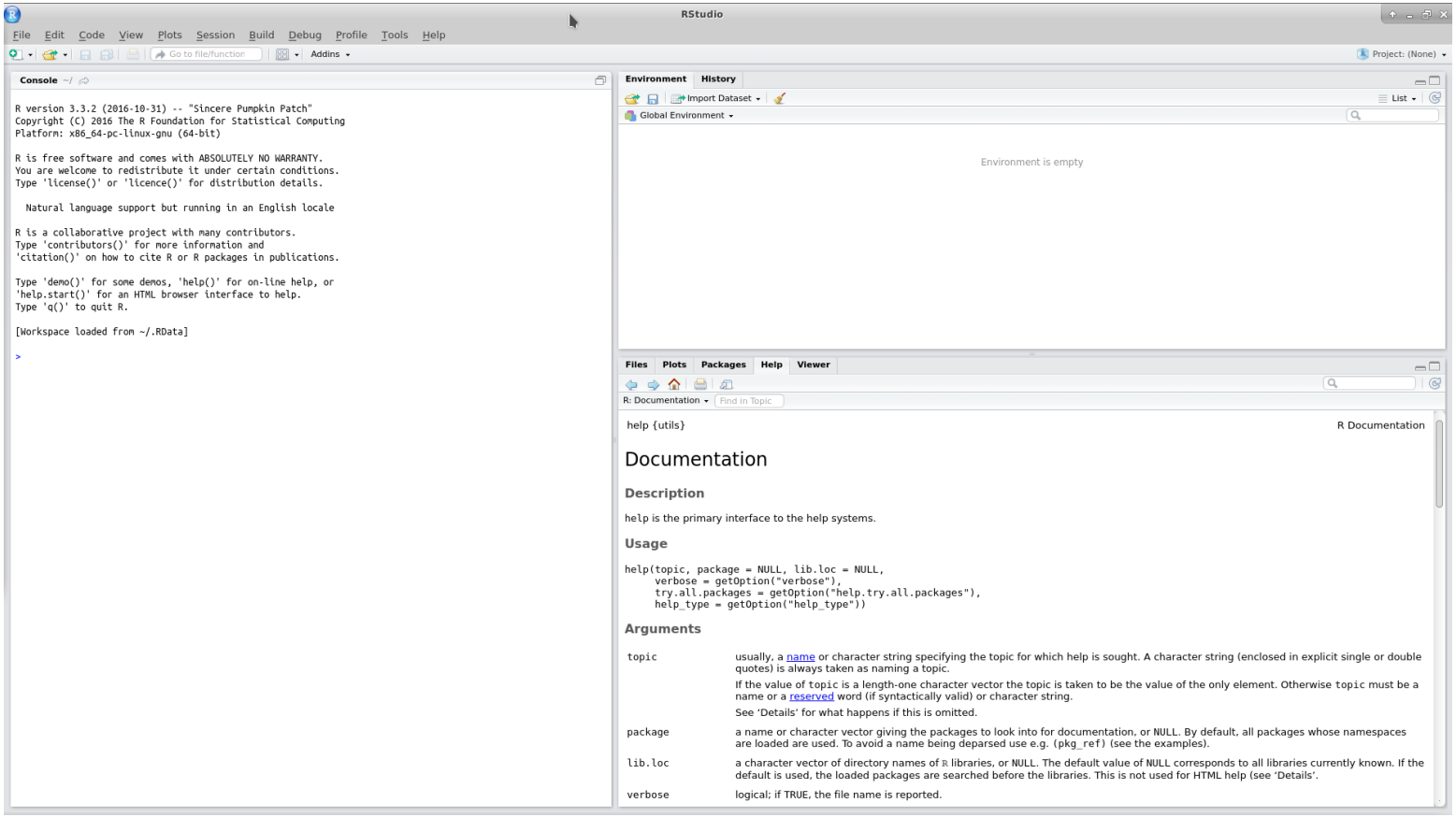
Installing R Software on Your Laptop

Go to R main website https://cran.r-project.org/ and download R based on your operating system.

- Install of R on Windows Operating System.
  Step by Step installation Video
  `https://www.youtube.com/watch?v=mfGFv-iB724`
- Install R on MacOS.
  Step by Step installation Video
  `https://www.youtube.com/watch?v=uxuuWXU-7UQ`

RStudio - IDE Recommended

RStudio `https://www.rstudio.com/` is a free and open-source Integrated Development Environment (IDE) for R Programming.



How to install RStudio. Step by step video

`https://www.youtube.com/watch?v=cX532N_XLIs`

Basic R Data Types

```
# numeric types: interger, double
> 348

# Characters
> "my string"

# logical
> TRUE
> FALSE

# Arithmetic operators as you'd expect
> 42 + 1 * 2^4

# So also logical operators/comparison
TRUE | FALSE
> 1 + 7 != 7
# Other logical operators:
# &, |, !
# <,>,<=,>=, ==, !=
```

Basic R Data Types

```
# Variables assignment is done with the <- operator
> mynumber <- 483

# typeof() tells use type
> typeof(mynumber)
[1] "double"

# we can convert between types
myint <- as.integer(mynumber)

> typeof(myint)
[1] "integer"
```

R Data Types Vector

```
# The vector is the most important data structure
# create it with c()
my.vec <- c(1, 2, 67, -8)

# get some properties
str(my.vec)

##
num [1:4] 1 2 67 -8
length(my.vec)

## [1] 4
# access elements with []
my.vec[3]

## [1] 67
my.vec[c(3,4)]

# can do assignment too
my.vec[5] <- 41.2
```

Working directory - Setting/Getting

It is the default location of all input and output files

```
# List all the objects in the current workspace
> getwd()

# Set working directory
> setwd("/YOUR-HOME-FOLDER/YOURFOLDER")
```

On Windows:

Remember to use double backslashes or use a single forward slash "/"

```
# List all the objects in the current workspace
> setwd("C:/Users/xyz/Documents/work/R")
```

You can use the RStudio menus to set your working directory.

Reading Data into R

Read a Comma-Separated Values (CSV) data file from a text file

```
> read.csv("filename")
```

First Line is the header, default value for header is True

```
> read.csv("filename", header=True)
```

It reads a **Dataframe** into R. Datatrame is an important data type in R.

It data type similar to an Excel Sheet or a Database Table like:

```
"age","job","marital","education","balance","housing","loan","contact"
30,"unemployed","married","primary",1787,"no","no","cellular"
33,"services","married","secondary",4789,"yes","yes","cellular"
35,"management","single","tertiary",1350,"yes","no","cellular"
30,"management","married","tertiary",1476,"yes","yes","unknown"
```

R Libraries and Packages

```
# Install a package (only need to do it once)
> install.packages("package name")
```

It will recognize dependencies between packages and install required sub packages

```
# Access the package
> library("package name")

# View a list of installed packages
> library()
```

Session Commands

```
> q()  # end R session

Save workshpace image? [y/n/c]:

# y  - yes
# n    no
# c    cancel

# Save content of the current workspace into .Rdata file
> save.image()
> save.image(file = "abc.Rdata")

# Save some objects of the current workspace into the file
> save.image(a, b, file = "abc.Rdata")
```

Load Stored Objects

```
> load("abc.Rdata")

# List all the objects in the current workspace
> ls()

OR

> objects()

# Remove objects from the current workspace
> rm(a, b)

# delete a file
> unlink("myFile.Rdata")
```

Learn R, in R. http://swirlstats.com/

You can learn R in R

Step by step Tutorial `http://swirlstats.com/students.html`

```
> install.packages("swirl")
> library("swirl")
> swirl()
```

Statistics

Probability and Statistics

- Statistics is the mathematical science behind the problem what can I know about a population if I'm unable to reach every member?

- If we could measure the height of every resident of Australia, then we could make a statement about the average height of Australians at the time we took our measurement.
- This is where random sampling comes in.

Probability and Statistics

- If we take a reasonably sized random sample of Australians and measure their heights, we can form a statistical inference about the population of Australia.
- Probability helps us know how sure we are of our conclusions!

What is Data?

- **Data** = the collected observations we have about something.
- Data can be **continuous**: "What is the stock price?"
- or **categorical**: "What car has the best repair history?"

Data

Nominal

- Predetermined categories
- Can't be sorted

  Animal classification (mammal, fish, reptile)

  Political party (republican, democrat, independent)

Ordinal

- Can be sorted
- Lacks scale

  Survey responses

Interval

- Provides scale
- Lacks a "zero" point

  Temperature

Ratio

- Values have a true zero point

  Age, weight, salary

Why Data Matters?

- Helps us **understand things as they are**:
  "What relationships if any exist between two events?"
  "Do people who eat an apple a day enjoy fewer doctor's visits than those who don't?"
- Helps us **predict future behavior** to guide business decisions:
  "Based on a user's click history which ad is more likely to bring them to our site?"

Statistics

- A science that deals with the **collection, classification, analysis, and interpretation** of data.
- Deals with data collection, evaluation and interpretation.
- Statisticians use data to find patterns, answer important scientific questions and draw conclusions.

Two main areas of statistics:

- **Describing data** (including numerical and graphical summaries)
- **Drawing conclusions** about data (making estimates, predictions, and decisions) from data collected via sampling

Fundamental Elements of Statistics

- **Experimental unit** (or observational unit) = an object (for example, a person, thing, or event) about which we collect data about.
- **Population** = every member of a group
- **Sample** = a subset of members that time and resources allow you to measure
- When studying a population, we focus on one or more characteristics of the units of the population. We call these characteristics **variables**.

Variables can be classified into one of two general types:

- **Quantitative**
- **Qualitative**

Quantitative Variables

Quantitative

- Contain numeric data, (how many?; how much?; or how often?)
- **Examples: height, weight, number of houses sold**
- Numerical or Quantitative variables can further be categorized as **continuous** (like height) or **discrete** (number of pets in a household)

Qualitative Variables

Qualitative

- Place experimental units into categories
- Qualitative data are data about **categorical variables** (what type?)
- **Examples: hair color, religion, political party**
- Categorical variables can be **"ordered levels"** are called **"ordinal"**. For example quality of a product can be answered with: very unsatisfied, unsatisfied, neutral, satisfied and very satisfied.

Data Summarization

Quantitative data Summaries

Numerical Summaries focus on measures that describe the center and the spread.

- Mean
- Median
- Variance
- Standard Deviation
- Quartiles

Graphical Summaries

- **Histograms** perhaps the most popular graphical summary of quantitative variables; Data are first categorized into classes of equal width and then frequencies and relative frequencies are calculated.
- **Box plots** - the median, minimum, maximum, 1st and 3rd quartiles are used to create box plots

Qualitative Data Summaries

Numerically, we can summarize qualitative data in two ways:

1. by computing the **class frequency**
2. by computing the **class relative frequency.**

- ▷ The **class frequency** is the number of observations in the data set that fall into a particular class.
- ▷ The **class relative frequency** is the proportion of the number of observations in the data set that fall into a particular class to the total number of observations in the data set.

Graphically, we can often use **Pie Charts and Bar Graphs** to summarize qualitative data

Measures of Central Tendency

Mean, Median, Mode

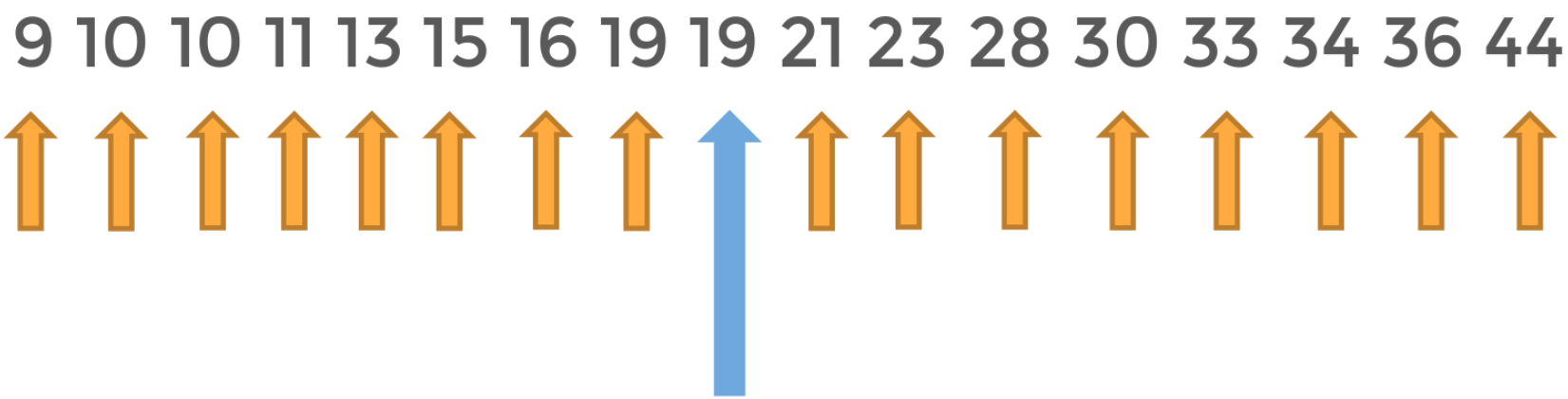
- Describe the "location" of the data
- Fail to describe the "shape" of the data

mean = "calculated average"
median = "middle value"
mode = "most occurring value"

Mean:

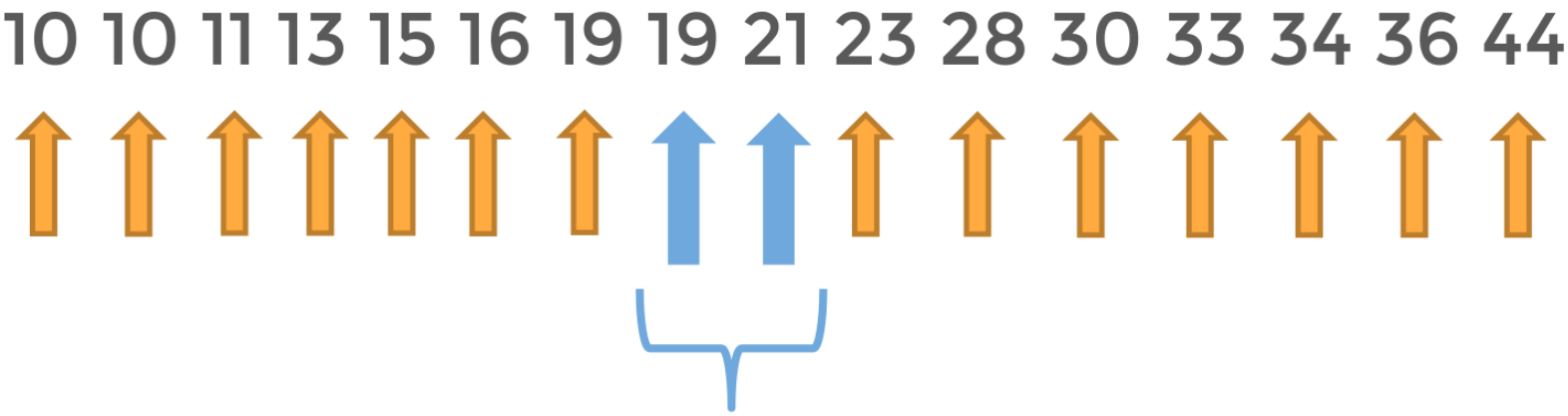
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median – odd number of values



= 19

Median – even number of values



$$\frac{19 + 21}{2} = 20$$

Measures of Central Tendency

Mean vs. Median

- The mean can be influenced by outliers.
  The mean of {2,3,2,3,2,12} is 4
  The median is 2.5
- The median is much closer to most of the values in the series!

Mode:

$= 16$

Measurement of Dispersion

Range, Variance, Standard Deviation

- **Range** = maximum value - minimum value
- **Variance** = calculated as the sum of square distances from each point to the mean
- **Standard Deviation** = square root of the variance (same units as the sample)

Measurement of Dispersion

Variance:

- POPULATION VARIANCE::

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- SAMPLE VARIANCE:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Measurement of Quartiles

- Another way to describe data is through **quartiles** and the **interquartile range** (IQR)
- Has the advantage that every data point is considered, not aggregated!

Measurement of Quartiles

Consider the following series of 20 values:

1. Divide the series
2. Divide each subseries
3. These become quartiles

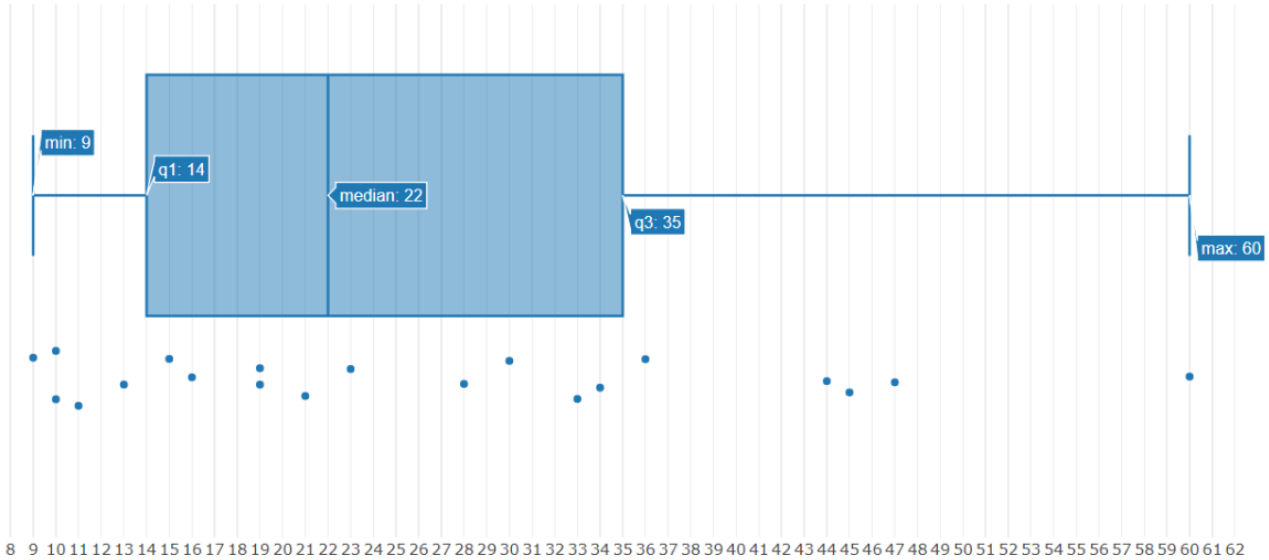
Measurement of Quartiles

Consider the following series of 20 values:

1^st quartile $= 14$
2^nd quartile $= 22$
3^nd quartile $= 35$

Measurement of Quartiles

9 10 10 11 13	15 16 19 19 21	23 28 30 33 34	36 44 45 47 60



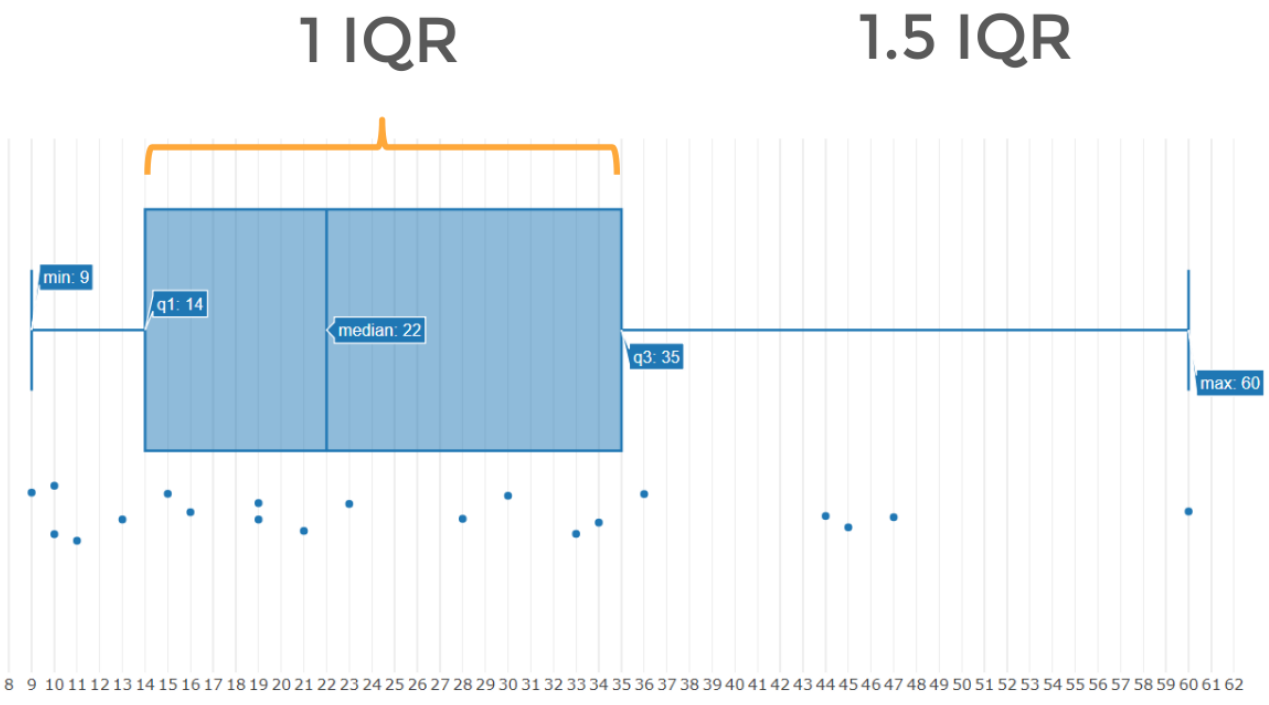
Quartile ranges are seldom the same size!

Measurement of Quartiles

Fences & Outliers

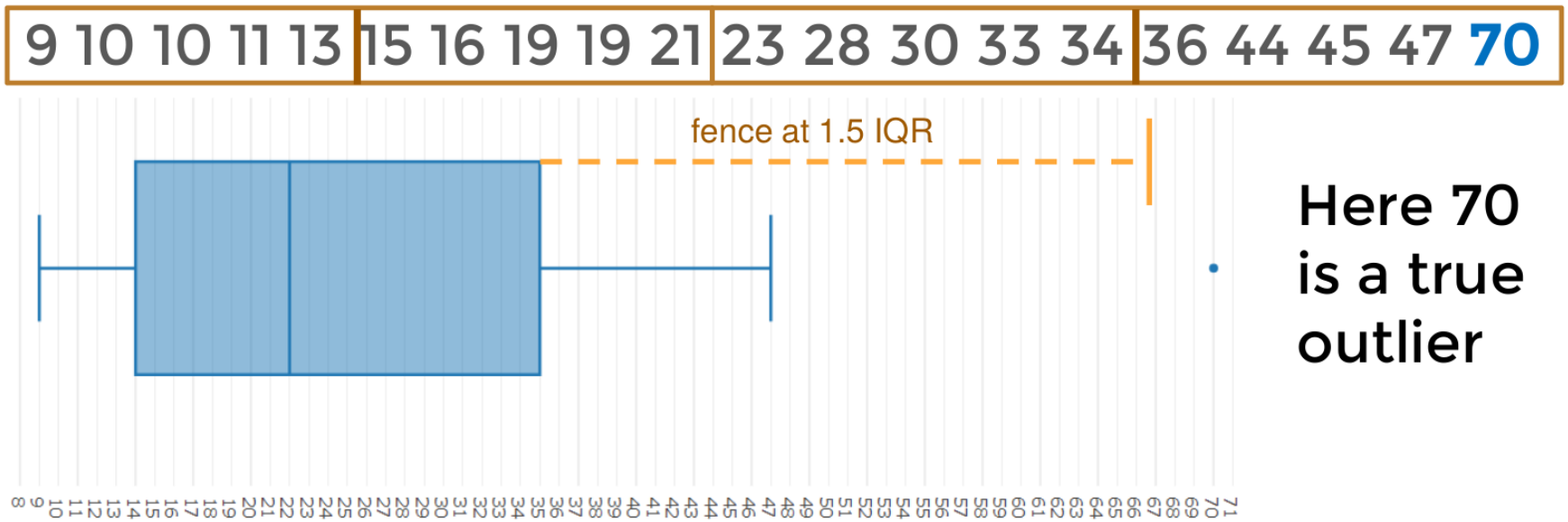
- What is considered an **"outlier"**?
- A common practice is to set a **"fence"** that is **1.5** times the width of the IQR
- Anything outside the fence is an outlier
- This is determined by the *data*, not an arbitrary percentage!

Measurement of Quartiles



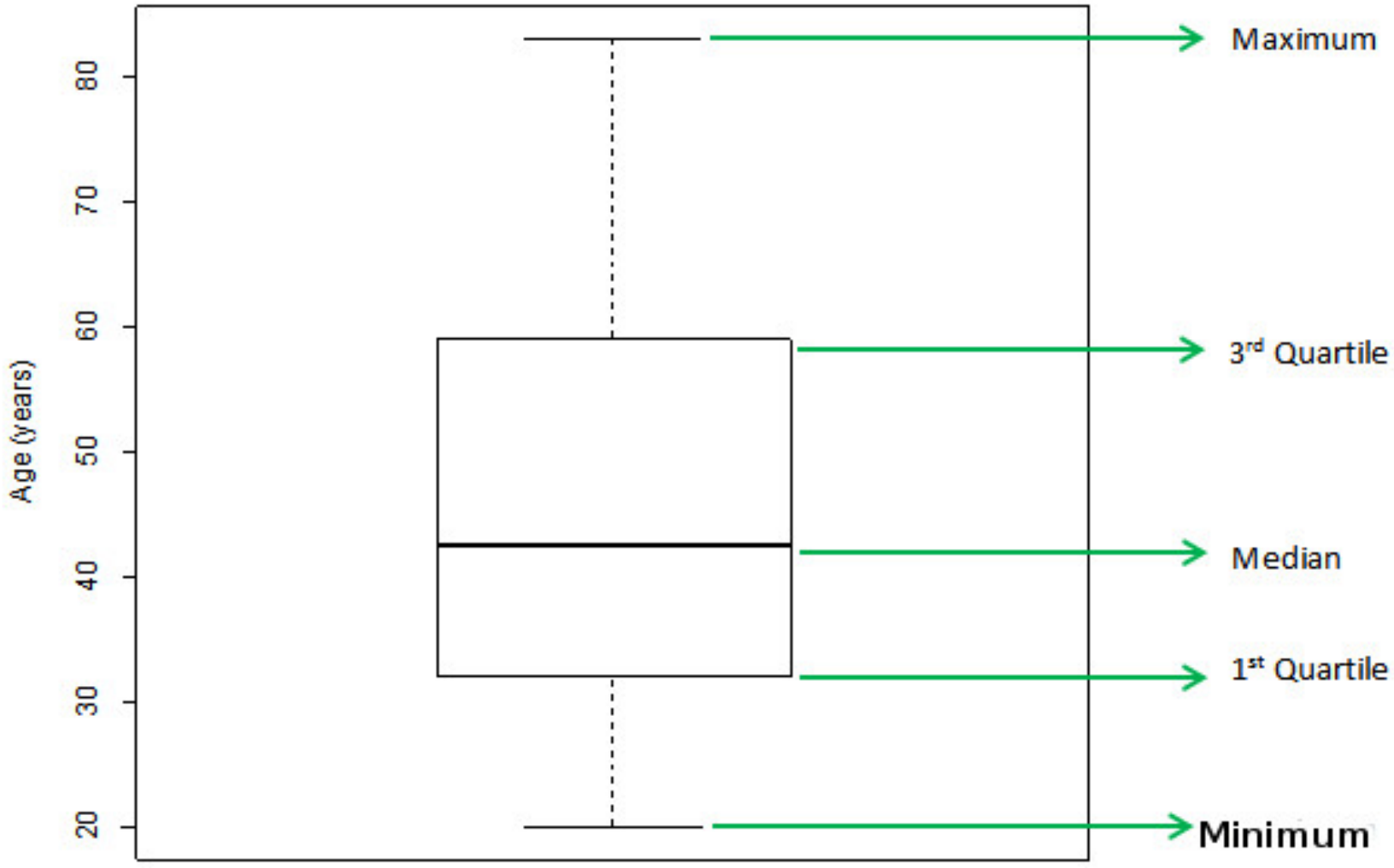
In this set, 60 is *not* an outlier, but 70 would be

Measurement of Quartiles



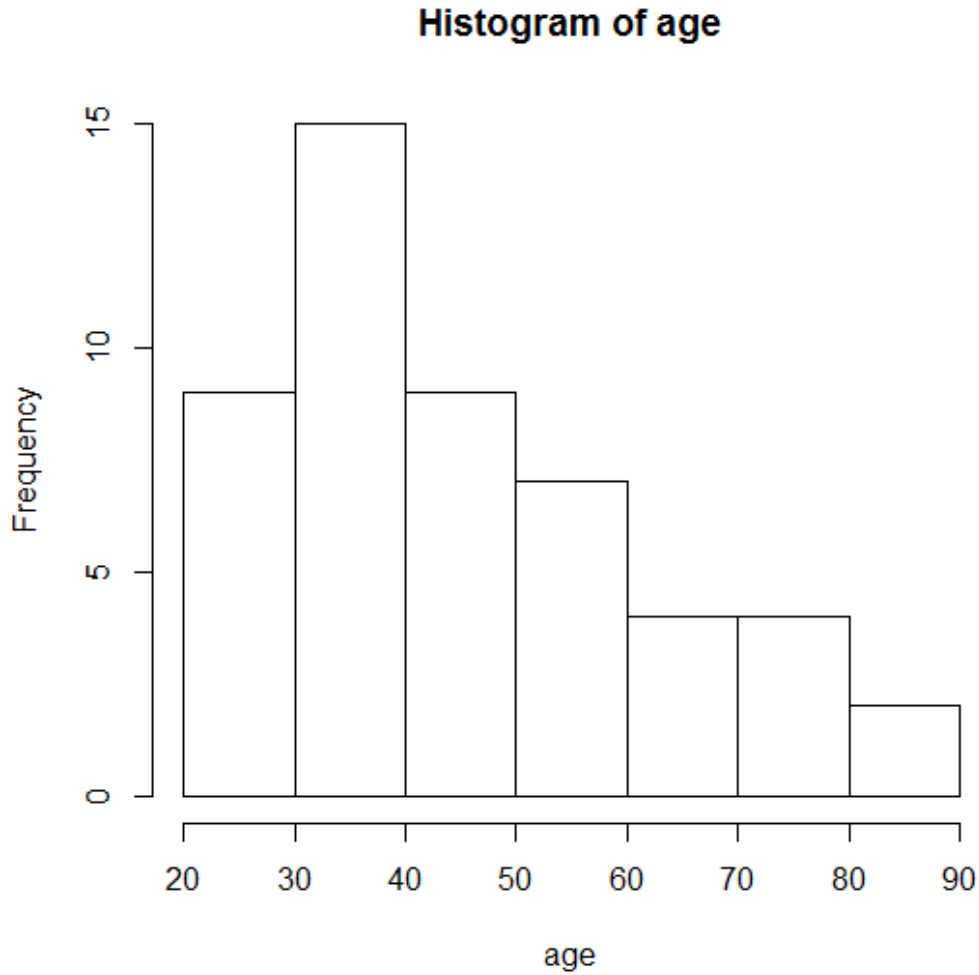
When drawing box plots, the whiskers are brought inward to the outermost values inside the fence.

Boxplots



Sample Boxplot

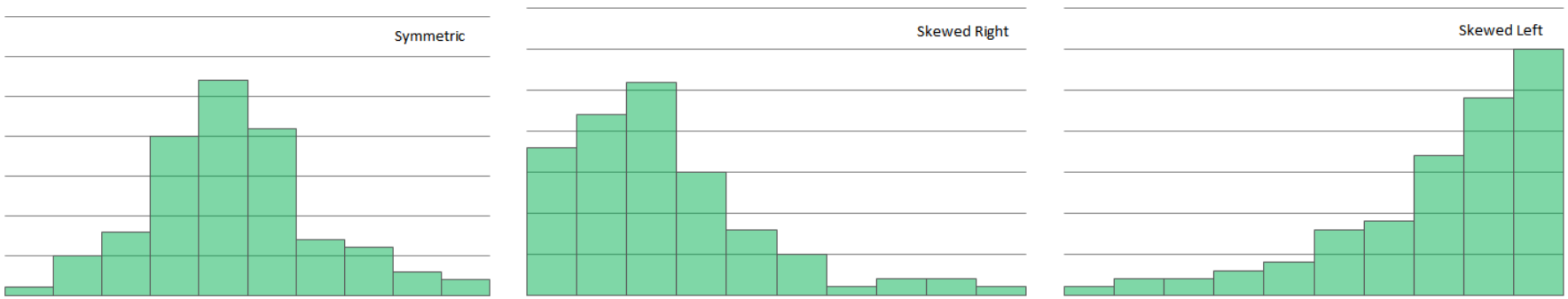
Histograms



Sample Histogram - Age Distribution

Histograms

- A distribution is **skewed to the right** if the right side (containing about half of the observations) of the histogram extends much further out than the left side.
- It is **skewed to the left** if the left side of the histogram extends much farther to the left than to the right side.



Symmetric, Skewed right, Skewed left

R Functions - Quantitative data summaries commands

```
> mean(data$variable)
> median(data$variable)
> min(data$variable)
> max(data$variable)
> quantile(data$variable)
> var(data$variable)
> sd(data$variable)
> summary(data$variable)
```

R Functions - Graphical data summaries

Histograms

```
> hist(data$variable)
> hist(data$variable, bins) # specify the number of bins
> hist(data$variable, breaks=c(x,y,z..)) # specify cutpoints
> hist(data$variable, breaks=seq(a,b,by=c)) # specify cutpoints
```

Boxplots

```
> boxplot(data$variable)
```

Make Your Graphs Look Better

Labeling

- Title: main="Histogram of xyz"
- X-axis label: xlab="Nile flow"
- Y-axis lable: ylab = "Frequency"

Colors

- color: col="dark red"

`http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf`

Controlling the window

- X-axis: xlim=c(min, max)
- Y-axis: ylim=c(min, max)

Combine multiple plots into one overall graph

```
> par(mfrow=c(2,2)) # 2 by 2 panels
> par(mfrow=c(1,1)) # Go back to single graph mode
```

R Functions - Qualitative data summaries

Numerical summary

- Class Frequencies

```
> table(data$variable)
Or
> summary(data$variable)
```

- Relative Class Frequencies Divide class frequencies by number of rows in the dataset using nrow(data)

Graphical summary

```
Pie(table(data$variable))
Barplot(table(data$variable))
```

Qualitative data summary - An example

Read in data

```
> read.csv("ceo.csv")
# Numerical summaries
Frequencies:

> table(data$Education)  or
> summary(data$Education)


Relative frequencies:
> table(data$Education)/nrow(data) or
> summary(data$Education)/nrow(data
```

R-Example - Graphical Summary

Graphical Summary

```
> pie(table(data$variable))

> barplot(summary(data$variable))/nrow(data))

> barplot(summary(data$Education), main="CEO Education Levels", xlab="
    Education level", ylab="Frequency")
```