MET CS555 B1
Data Analysis and Visualization

Lecture 1

Kia Teymourian

Course Description

- Will introduce the **statistical** tools most commonly used to process, analyze, and visualize data

- Will cover topics such as describing data, **statistical inference**, simple linear regression, multiple regression, logistic regression, analysis of variance, and regression diagnostics.

- The **statistical package R** will be use throughout the course. The focus is on understanding how to use and interpret output from R and how to visualize results.

- In each topic area, the methodology, including underlying assumptions and the mechanics of how it all works along with appropriate interpretation of the results, are discussed.

What will you get out of this course?

- Appreciate the **science of statistics** and the scope of its potential applications
- Summarize and present data in meaningful ways
- Select the appropriate statistical analysis depending on the research question
- Form testable hypotheses that can be evaluated using common statistical analyses
- Understand and verify the underlying assumptions of a particular analysis
- Effectively and clearly communicate results from analyses performed to others
- Conduct, present, and interpret common statistical analyses using R

Syllabus

- Module 1 - Describing and Interpreting Data
- Module 2 - Statistical Inference and Tests for Comparisons of Means
- Module 3 - Correlation and Simple Linear Regression
- Module 4 - Regression Diagnostics and Multiple Linear Regression
- Module 5 - Analysis of Variance
- Module 6 - Tests for Comparisons of Proportions and Logistic Regression

Course website (BU Blackboard site)

1. Go to http://onlinecampus.bu.edu

2. On left hand-side, there is a click-able link named **“Printable Lectures”**

3. Click on it. You will see the content on the right hand-side window

4. There are 2 lectures in each module. We will cover a lecture per week, which means we will cover a module in 2 weeks’ time.

Recommend to read the module lecture before class.

Prerequisite Courses

- CS546 (Quantitative Methods for Information Systems)

and

- CS544 (Foundations of Analytics)
- or equivalent background

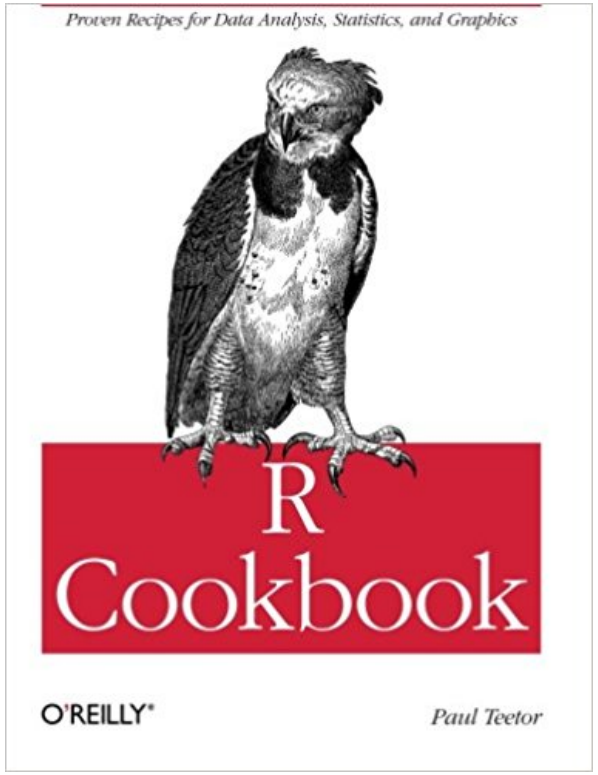
Other related Courses
(**Graduate Certificate in Data Analytics**)

- MET CS 544 Foundations of Analytics
- MET CS 688 Web Analytics and Mining
- MET CS 699 Data Mining

http://www.bu.edu/csmet/da/

Required textbook

- Teetor, P. (2011). R cookbook. Sebastopol, CA: O'Reilly. ISBN 9780596809157.

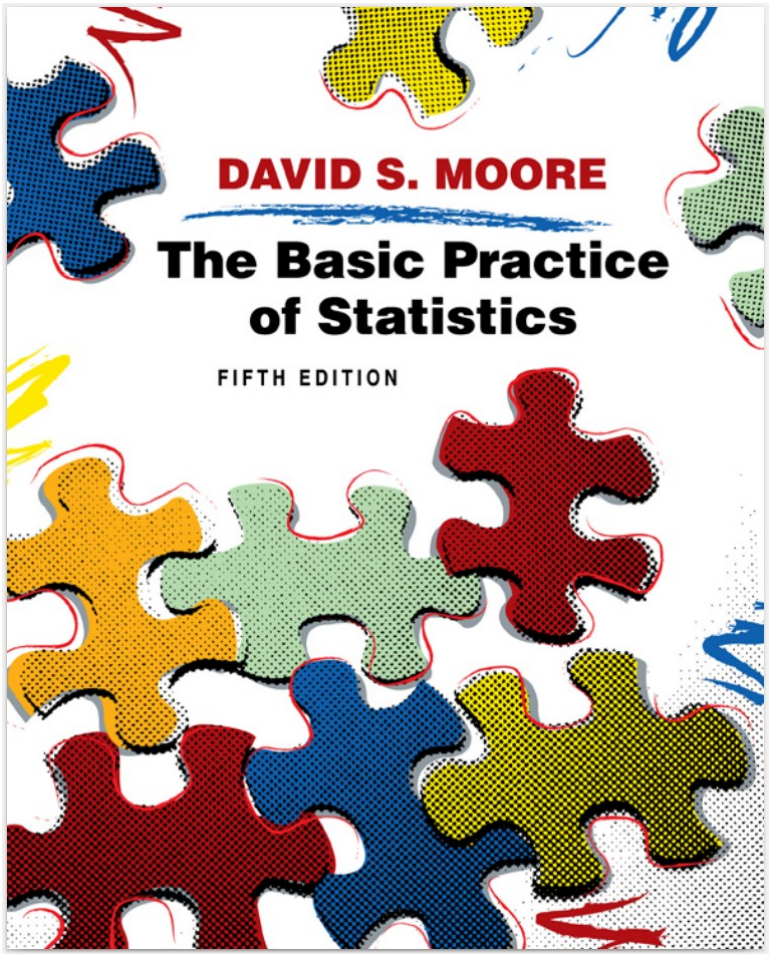


Free Online Access to the Book through BU Library

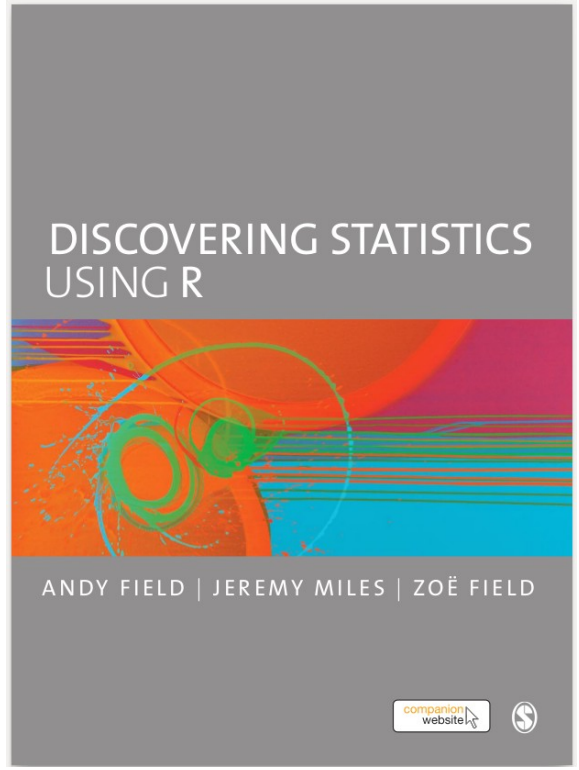
http://proquestcombo.safaribooksonline.com.ezproxy.bu.edu/9780596809287/id3381294

Statistic Book

- The Basic Practice of Statistics. DAVID S. MOORE



Additional Textbooks and Online Links



- Andy Field, Jeremy Miles and Zoe Field. (2012) ***Discovering Statistics Using R***. Publisher: SAGE Publications Ltd. ISBN-13: 978-1446200469

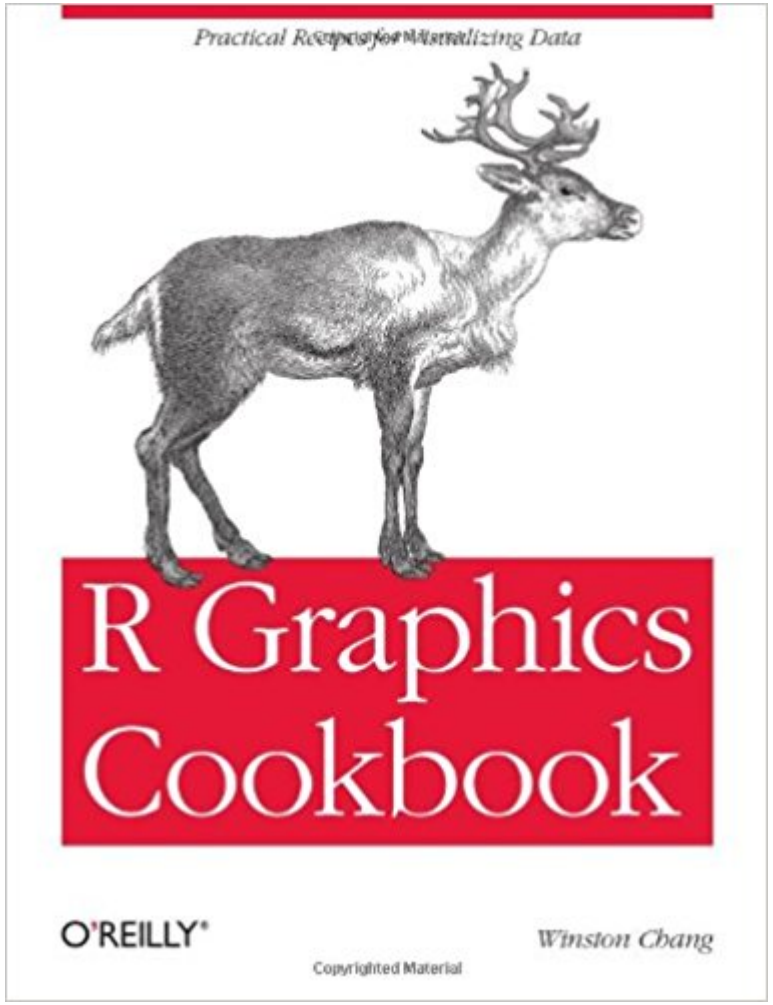
- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. (2013) ***An Introduction to Statistical Learning with Applications in R***. Springer.

- **http://onlinestatbook.com/ with some demos**

- **https://www.openintro.org/stat/** Free PDF for download & R tutorials and codes.

Recommended textbook

- Chang, W. (2013). R graphics cookbook. Sebastopol, CA: O'Reilly. ISBN 9781449316952.



Laptop Requirements

- You need to have a laptop for this class.
- You need a laptop for Midterm exam and final exam.
- Recommend to have your laptop with you in each class session

Grading Structure

- The **6 assignments** are focused on applying theory learned in the class to a set of data and analyzing that data in R.

- The **6 quizzes** will evaluate your understanding of concepts presented in the corresponding week's lecture.

- **Midterm exam** will cover material up to the end of Module 3.

- The **final exam** will be comprehensive and will cover material from the entire course. It will be an **open-book** proctored exam consisting of questions similar to the ones in the quizzes but longer in length.

Assignment Completion & Late Work

- All assignments should be submitted on time. If there is a delay, the student must be in touch with the instructor.

- Late submissions without reasons will result in grade deduction. You can turn in an assignment up to 24 hours late, in which case you receive a ***10% penalty*** (that is, 10 points are subtracted from an assignment that is worth 100 points), or up to 48 hours late, in which case you receive a ***20% penalty***. **Assignments turned in after that are not accepted**.

- **We kept on saying no exceptions**,
  but there are exceptions in very extreme circumstances, with proper documentation. For example, if you **obtain a doctor/dentist note stating that you were so ill** at the due date/time that you could not reasonably be expected to meet the deadline, it is possible to get an extension.

Grading Structure (cont’d)

6 Assignments	30%
6 Quizzes	10%
Midterm Exam	30%
Final Exam	30%

A	94–100
A−	90–93
B+	86–89
B	81–85
B−	76–80
C+	71–75
C	66–70
C−	61–65
D	56–60
F	0–55

Final Exam can improve your Final Grade

- If you write a very good final exam.

  If your final exam score is over 90% and you have from all other deliverables (Assignments+Quizzes+Midterm) over **70%**, then you will get an **“A”** as final grade.

  For example: someone has

Final exam	**28/30**
Assignments	23/30
Quizzes	6/10
Midterm	23/30

Time Plan

- Class will be **6:00 pm – 8:45 pm**

  15 min break

  First part: 6:00 - 7:15
  Second Part: 7:30 – 8:45

- **Calendar** on https://onlinecampus.bu.edu

- Two important appointments, **Midterm exam and Final exam**

About me

- **Kia Teymourian**
- Email: kiat@bu.edu
- Office hour: Thursday 3-5 PM
- Office address: 808 Commonwealth Ave, Room 257
- Academic Website: www.teymourian.info

R Programming

R package

- Open source programming language for statistical computing and graphical visualizations
- Part of GNU project
- 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/

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/

Resources

- Our textbooks (use Table of Contents and Index)
- R project website (http://www.r-project.org)
- R specific search engine (http://rseek.org)
- Search on the Web
- Email me
- Ask questions in Blackboard site -> Class Discussion -> Share what you've learned in R; And share what you've learned
- "How To Ask Questions The Smart Way" by Eric Steven Raymond http://www.catb.org/esr/faqs/smart-questions.html

Resources (continued)

Online Books:

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

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

- R for Beginners, by Emmanuel Paradis.

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

- Using R for Data Analysis and Graphics. Introduction, Code and Commentary, by J. H. Maindonald.

- CRAN R language manuals: http://cran.r-project.org/manuals.html

Others:

- https://www.codeschool.com/courses/try-r by Code School

- http://www.ats.ucla.edu/stat/ Institute for Digital Research and Education

Installing R

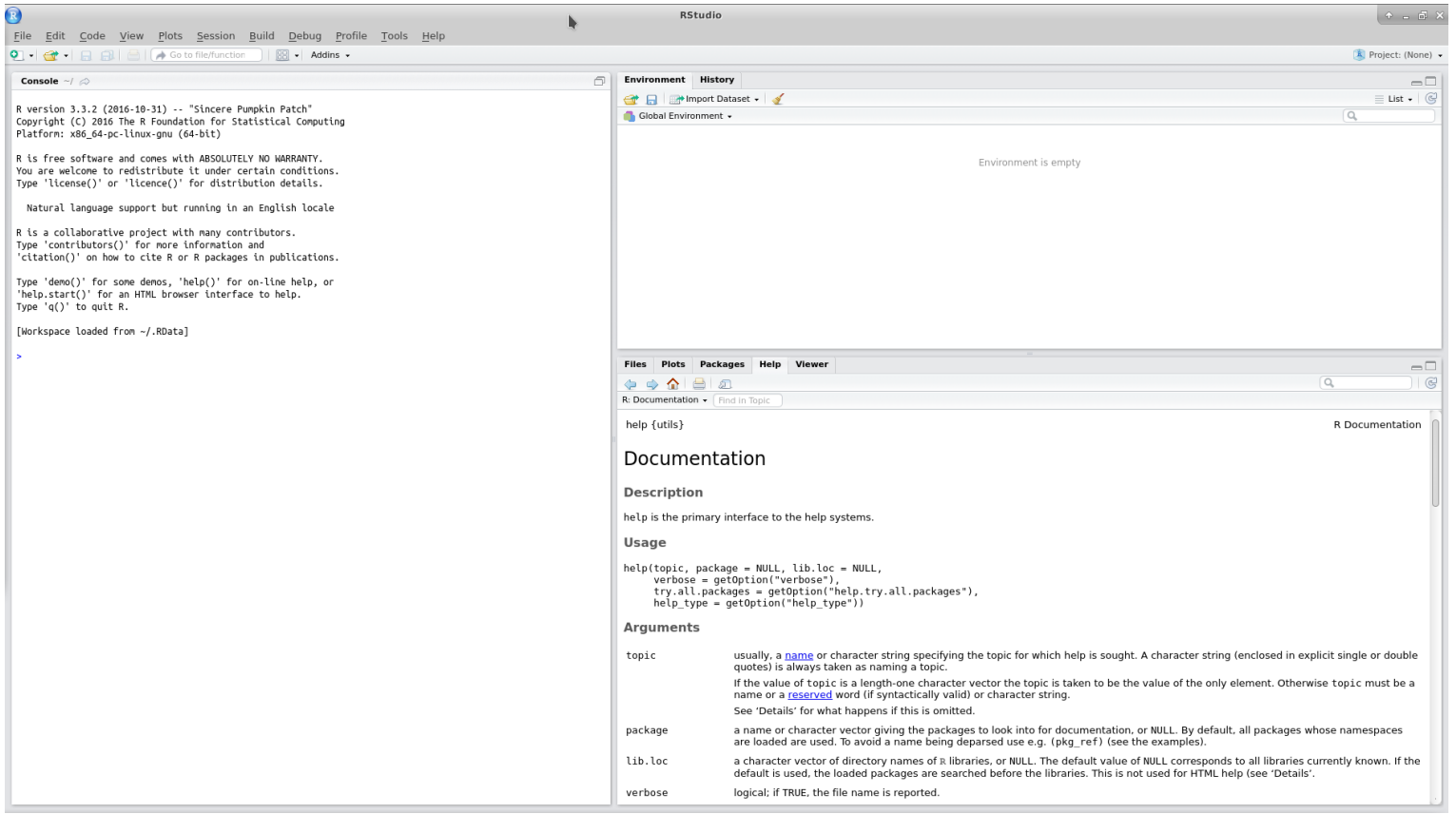
- Go to R main website https://cran.r-project.org/ and download R based on your operating system.

- **Install 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 – Recommended

RStudio is a free and open-source integrated development environment for R



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

# character
"my string"

# logical
TRUE
FALSE

# arithmetic operators as you'd expect
42 + 1 * 2^4
# so too logical operators/comparison
TRUE | FALSE
1 + 7 != 7
# Other logical operators:
# &, |, !
# <,>,<=,>=, ==, !=
```

R Data Types – Cont.

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 – Cont.

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)]
## [1]
67 -8
# can do assignment too
my.vec[5] <- 41.2
```

Working directory

It is the default location of all input and output files

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

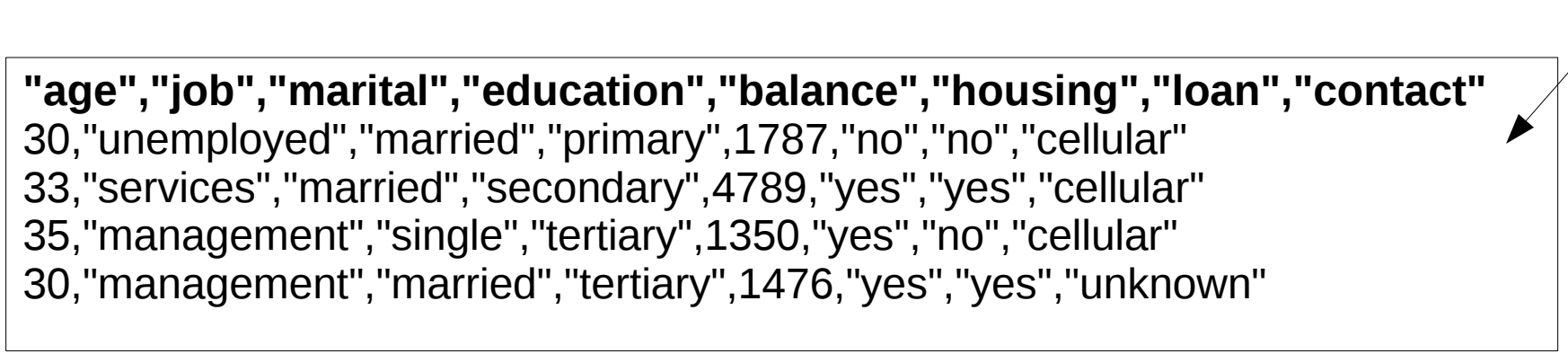
On Windows

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")
```

Read in data

Read a Comma-Separated Values data file from a text file
> **read.csv**(“filename”)



This is saved in the plain text file.

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.

Read in data

Read in data from a text file
> read.csv(“filename”)

Read in data from an excel file (.xlsx)
- Need to install and load package “xlsx”

> read.xlsx(“filename”, 1)

To make your life easier
- Save files in your working directory
- Have column headers in your data files
- Check the data after loading to make sure it is right

R 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()

Some commands to help you work efficiently

Command line window

- Use for one time or one-off commands
- User up/down arrows or Ctrl-P to view your recent inputs

```
# View the history of what you typed
>  history()
```

Use R Editor

- Go to “File” -> “New Script”
- Write your commands in R Editor; Save the script for future reference
- Use “#” to make comments
- To run part of the script, select the commands of interest, use Ctrl-R or toolbar icon

Get help in R

```
> # Get help on a function if you remember the function name
> ?cos
> # use ?? when don’t remember exact function name
> ??trigonometric
```

OR

```
> help(cos)

> help.start()  # help in HTML format

> # find all functions related to a subject of interest

> help.search(“trigonometric”)
```

Get help in R (continued)

```
> # List of demos
> demo()
> demo(lm.glm)  #  linear modelling examples
> demo(graphics)  # multiple graphics examples

> example(matrix) # examples of a function use

> # list all function names that include the text matrix
>  apropos("matrix")
```

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

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

- An **experimental unit** (or observational unit) is an object (for example, a person, thing, or event) about which we collect data about.

- A **population** is a set of units (for example, people, objects, or events) that we are interested in studying.

- When studying a population, we focus on one or more characteristics of the units of the population. We call these characteristics **variables.**

Fundamental Elements of Statistics

Variables can be classified into one of two general types:

- **Quantitative** - contain numeric data, also referred to as continuous variables
  (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** - 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*.

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

Graphically, we can often use pie charts and bar graphs to summarize qualitative data

Quantitative data summaries

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

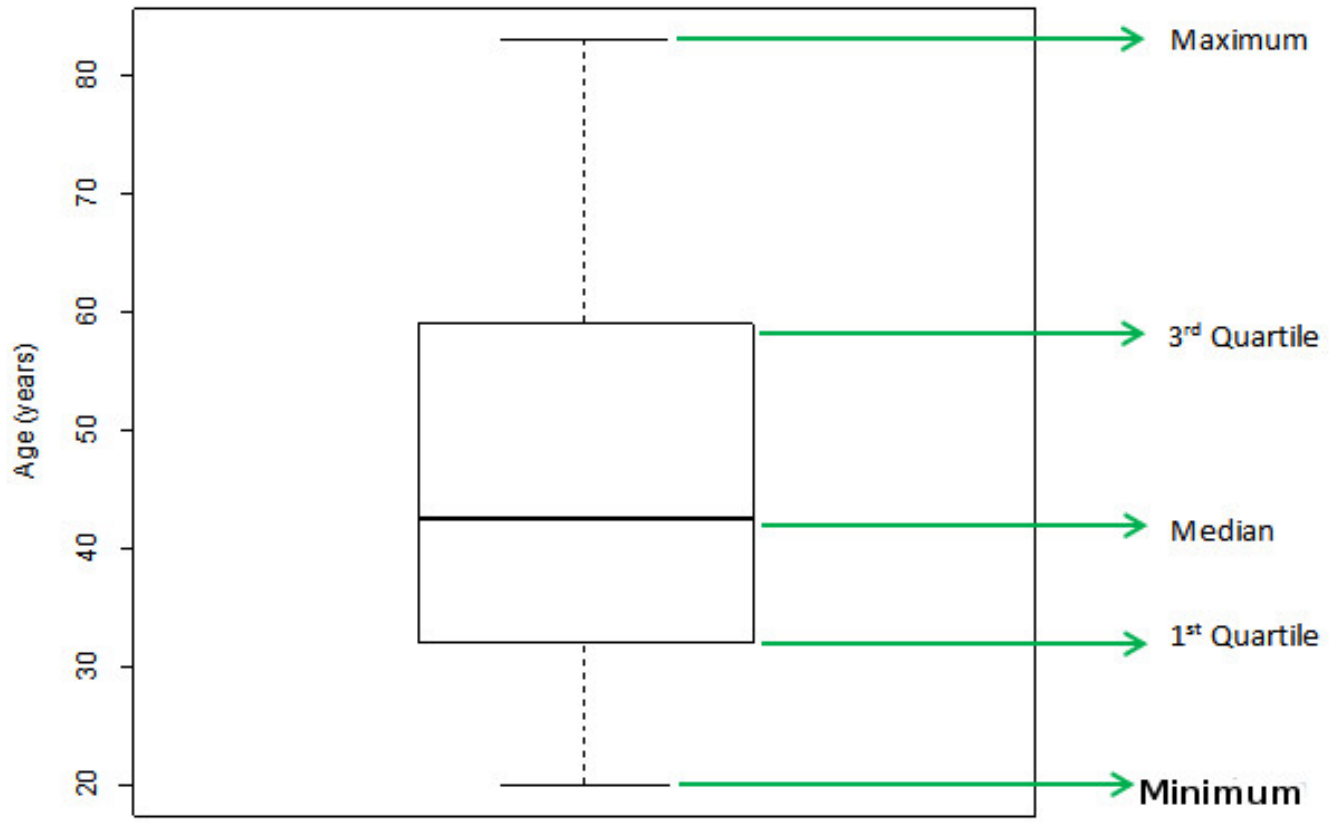
- Mean
- Median
- Variance
- Standard Deviation
- Quartiles

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.

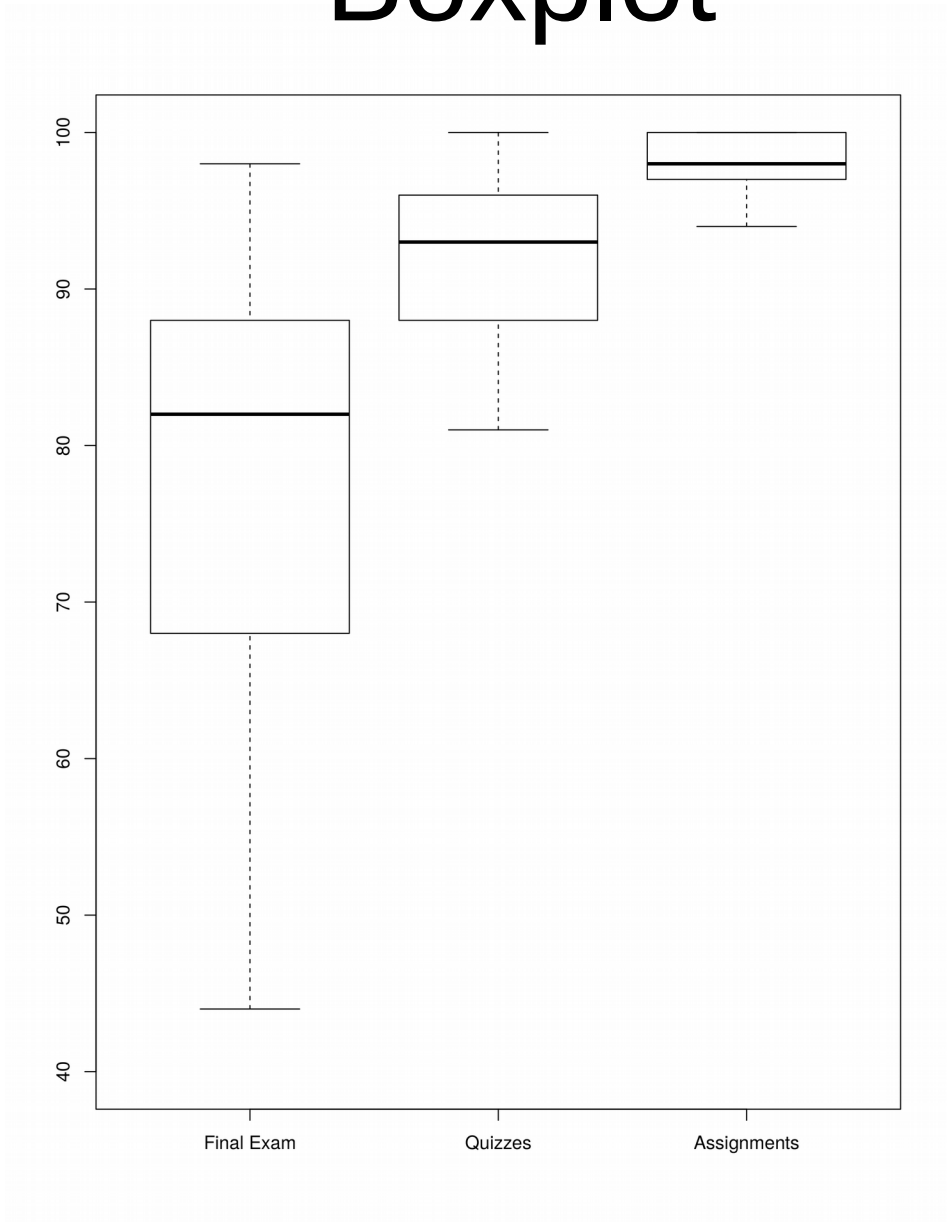
- Box plots - the median, minimum, maximum, 1st and 3rd quartiles are used to create box plots

Boxplots

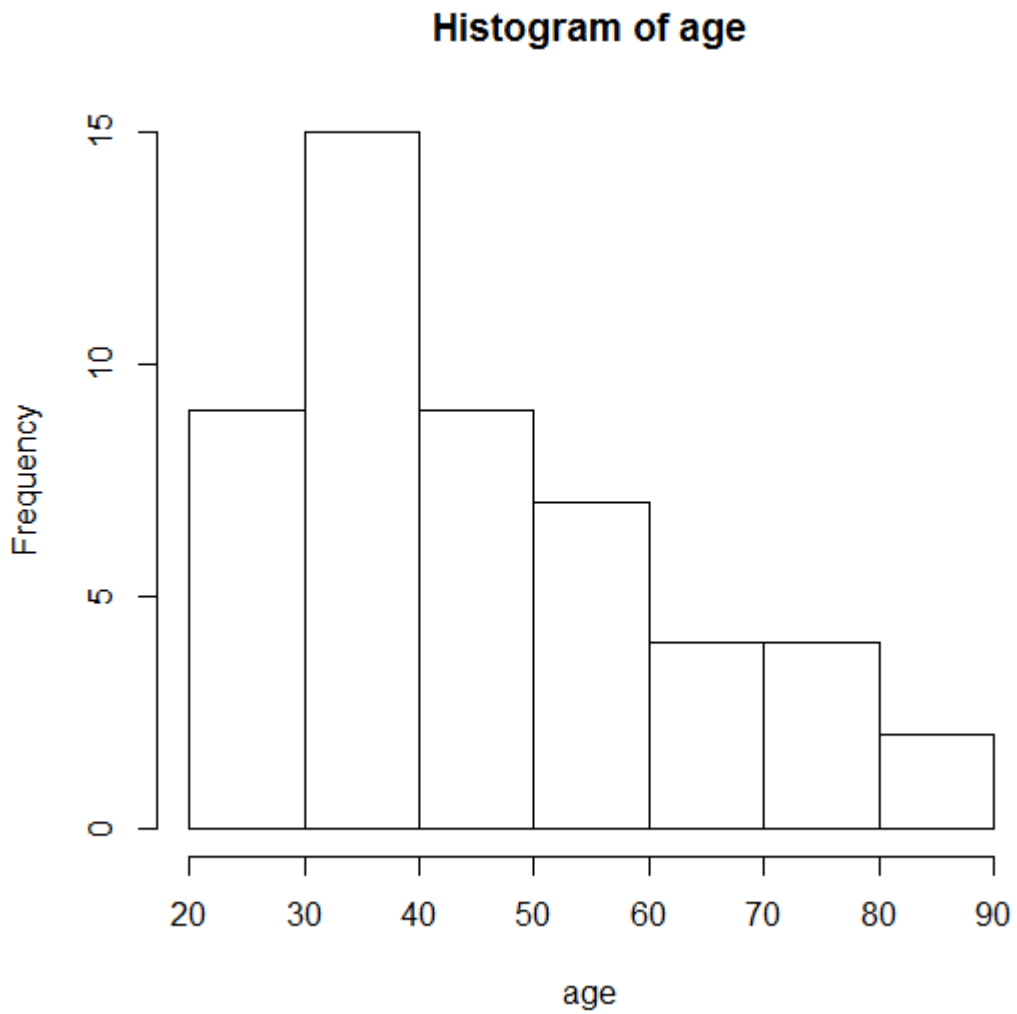

Age (years)
20
30
40
50
60
70
80
Maximum
3rd Quartile
Median
1st Quartile
Minimum

Boxplot

Last Semester Scores



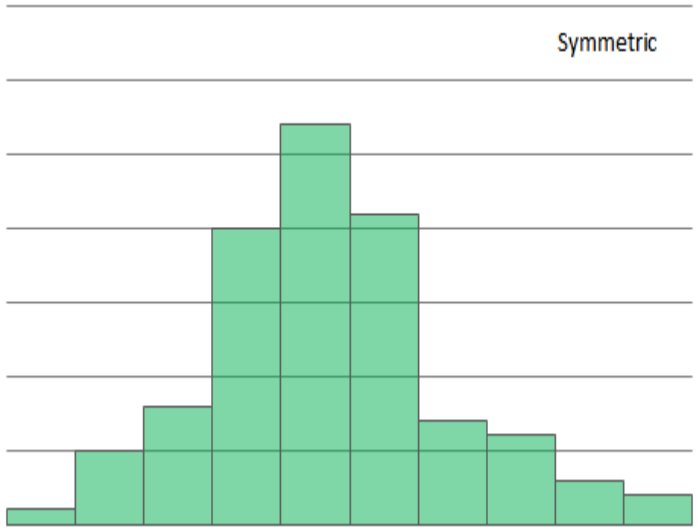
Histogram

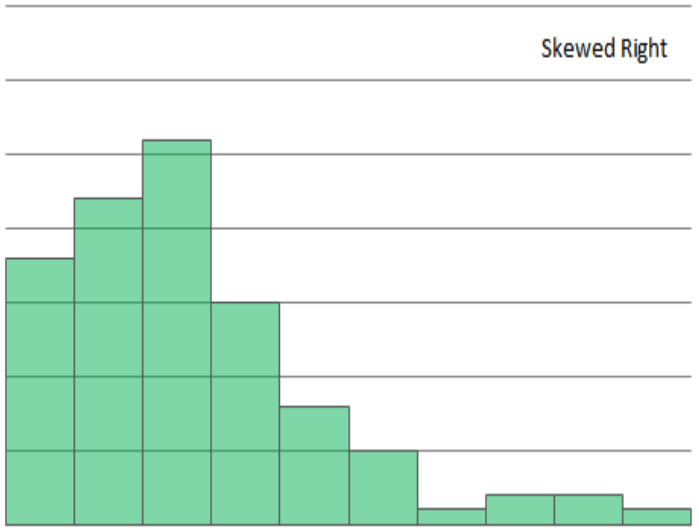


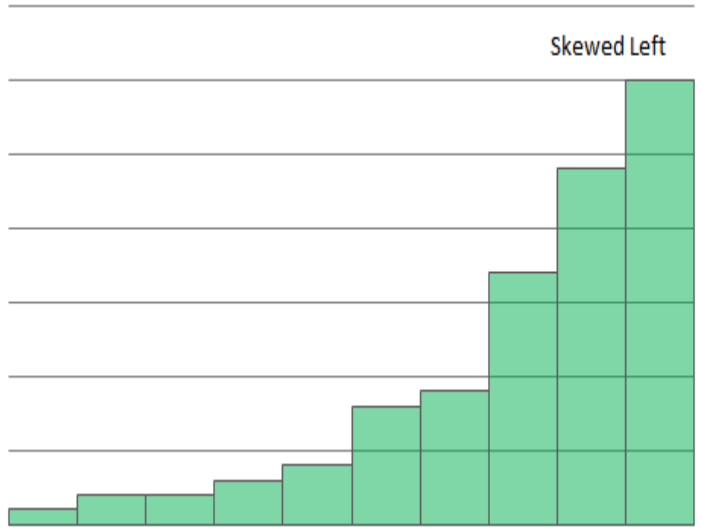
Histograms

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.







Quantitative data summaries

- The mean and median are numerical measures of the center of a distribution
- The population mean is the mean of all observations for the entire population: $\mu = \frac{\sum x_i}{N}$
- The sample mean for a sample of size n is $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- The variance and the standard deviation measure how far each observation is from the mean.

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

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

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)
```

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
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

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))

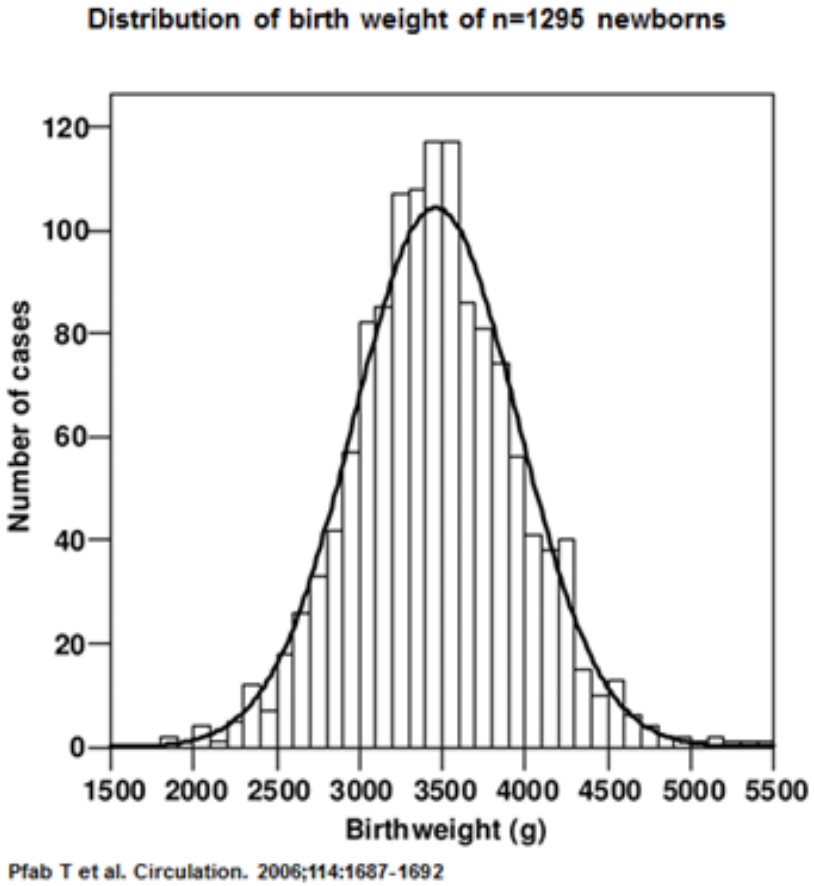
Qualitative data summary – An example

```
# 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)

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")
```

Normal Distribution

Many statistical inference procedures are based on the normal distribution.

The density curve (a mathematical model that represents the pattern of data) of a normal distribution has a very familiar, bell curve shape with a single peak. It is perfectly symmetrical.

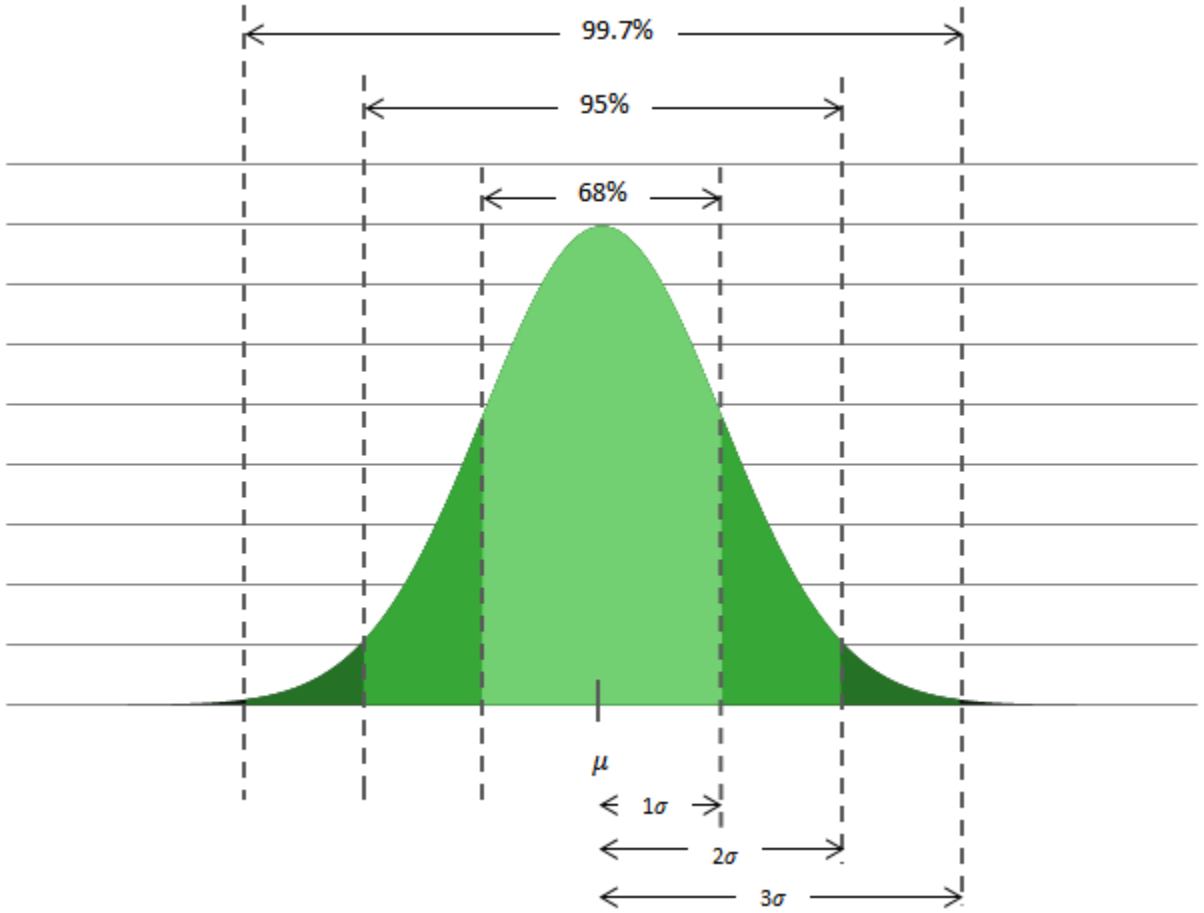


Pfab T et al. Circulation. 2006;114:1687-1692

68-95-99.7 Rule

For a normal distribution with mean, μ, and standard deviation, σ,

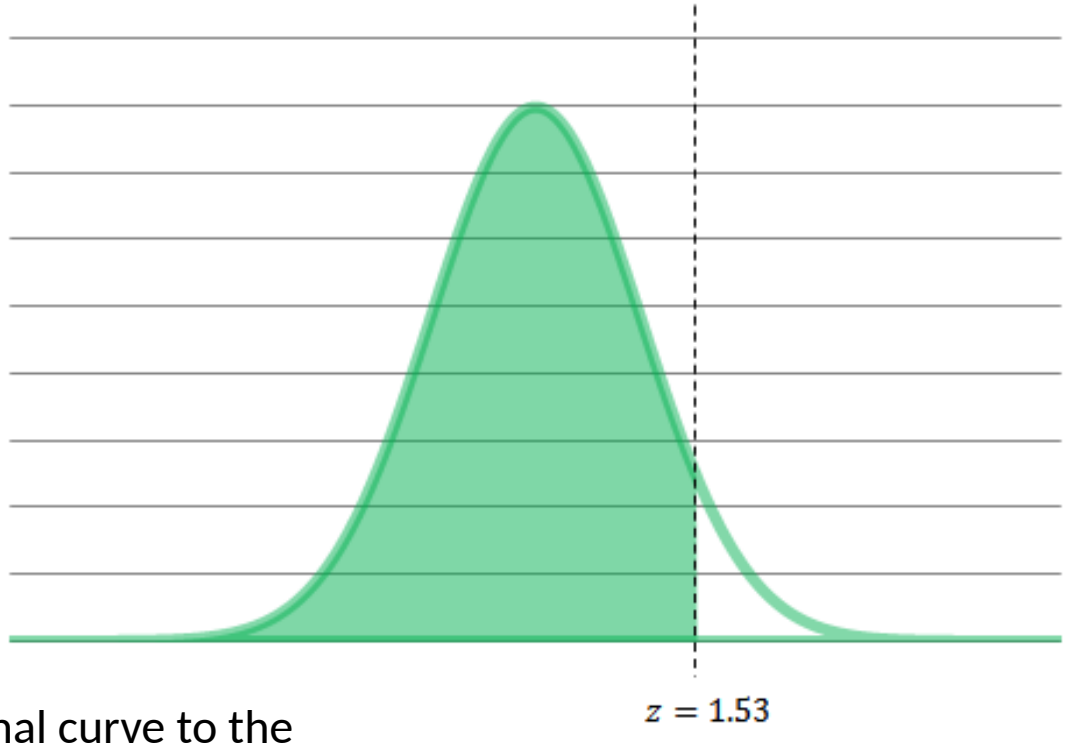
- 68% of the observations fall within one standard deviation of the mean
- 95% of the observations fall within two standard deviations of the mean
- 99.7% of the observations fall within three standard deviations of the mean



Standardized Normal Distribution

- Convert a value into standard deviation units by calculating its Z-score
- Z-score tells us how many standard deviation x is from the mean

$$z = \frac{x - \mu}{\sigma}$$



The area under the standard Normal curve to the left of z is 0.9370.

Normal Distribution – R commands

Areas
> pnorm(z) # calculate the area to the left of z

For not standardized normal distribution
> pnorm(x, mean=a, sd=b) # calculate the area to the left of x

Z-score associated with an area
> qnorm(a) # calculate z-score with a as the area to the left

For not standardized normal distribution
> qnorm(c, mean=a, sd=b)