STAT 1291: Data Science

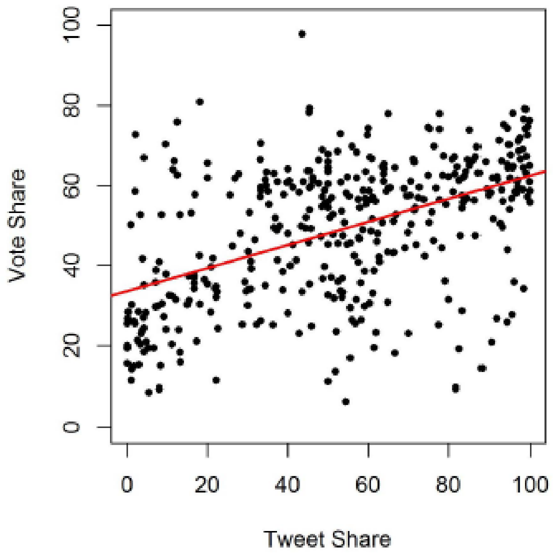
Lecture 2 - Doing Data Science

Sungkyu Jung

Last lecture

- What is Data Science?
- Course webpage: http://www.stat.pitt.edu/sungkyu/course/pds/

A case study "More Tweets, More Votes?"



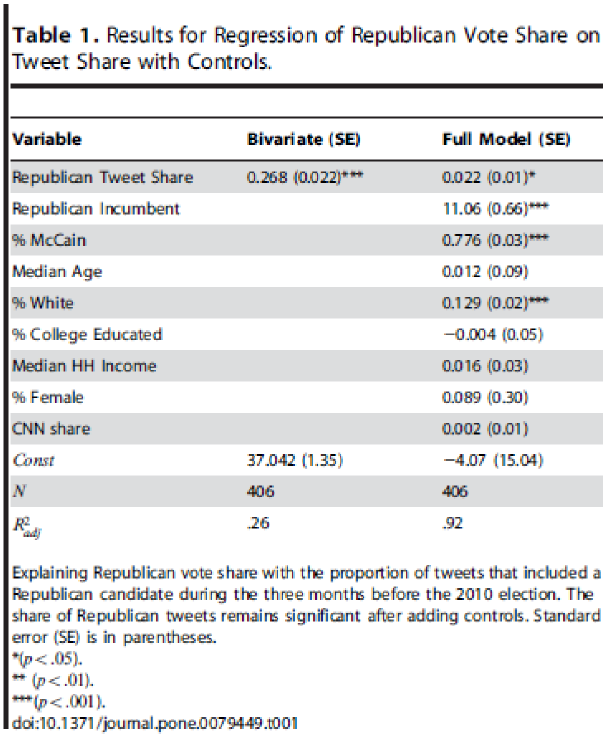


Table 1. Results for Regression of Republican Vote Share on Tweet Share with Controls.

Variable	Bivariate (SE)	Full Model (SE)
Republican Tweet Share	0.268 (0.022)***	0.022 (0.01)*
Republican Incumbent		11.06 (0.66)***
% McCain		0.776 (0.03)***
Median Age		0.012 (0.09)
% White		0.129 (0.02)***
% College Educated		−0.004 (0.05)
Median HH Income		0.016 (0.03)
% Female		0.089 (0.30)
CNN share		0.002 (0.01)
Const	37.042 (1.35)	−4.07 (15.04)
N	406	406
R^2_{adj}	.26	.92

Explaining Republican vote share with the proportion of tweets that included a Republican candidate during the three months before the 2010 election. The share of Republican tweets remains significant after adding controls. Standard error (SE) is in parentheses.
*($p<.05$).
** ($p<.01$).
***($p<.001$).
doi:10.1371/journal.pone.0079449.t001

Figure 1:

- How would you reproduce this study?
- Data can be found at Harvard DataVerse

Big Data and Data Science Hype

What is *big data* and what is *data science*?
Is data science the science of Big Data?
Is data science just an extension of statistics?

From wikipedia: Data Science is an interdisciplinary field about scientific **methods, processes, and systems** to **extract knowledge or insights** from data in various forms, either structured or unstructured, similar to **data mining**.

"Unstructured data" can include emails, videos, photos, social media, and other user-generated content.

Data science often requires sorting through a great amount of information and writing algorithms to extract insights from this data.

Today

- *What is Data Science?*
- How do we learn Data Science? (Course logistics)
- Data visualization

How do we learn?

- Learn data science by doing data science
- use R and RStudio
- Two lectures and one recitation (lab) in a week

R

- R is a free software environment for statistical computing and graphics, and is the best data science language. URL https://www.r-project.org/
- (https://www.r-bloggers.com/why-you-should-learn-r-first-for-data-science/)
- (http://sharpsightlabs.com/blog/r-recommend-data-science/)

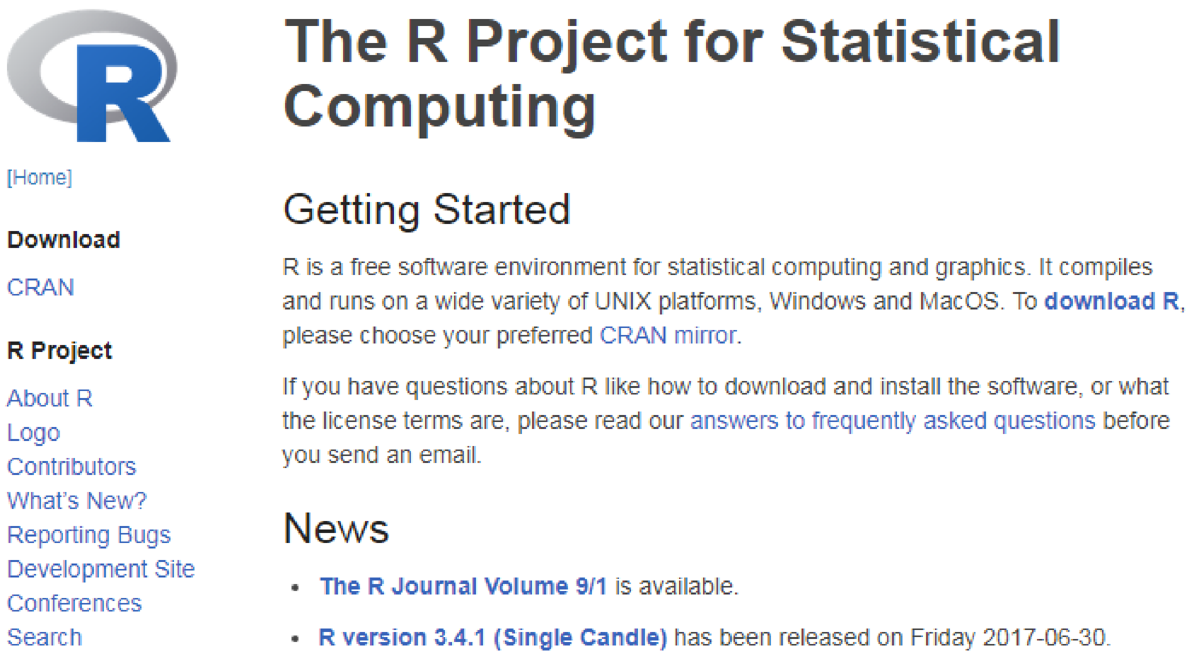
RStudio

RStudio is an open source and enterprise-ready professional software for R. URL https://www.rstudio.com/

Rstudio screen

How to learn R and RStudio

- R is a language for data science.
- This entire course is about doing data science using R
- Fridays classes (11 or 12 AM) will meet at STAT LAB (Posvar 1201) whenever possible
- We will begin using R on this Friday.
- If you've got a personal computer, install R and RStudio.
  - Visit https://cran.r-project.org/ to install R, then visit https://www.rstudio.com/ to install RStudio.
  - Take a look at "INSTALLING R and R Studio" document (at the course webpage)



[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Development Site
Conferences
Search

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

News

- **The R Journal Volume 9/1** is available.
- **R version 3.4.1 (Single Candle)** has been released on Friday 2017-06-30.

Figure 2:

  - Watch Lynda.com video at https://www.lynda.com/R-tutorials/Up-Running-R/120612-2.html (Use your Pitt ID to log-in)

- Computers in STAT LAB have R and RStudio. You can bring your laptop to the lab.

Textbooks

Required Textbook

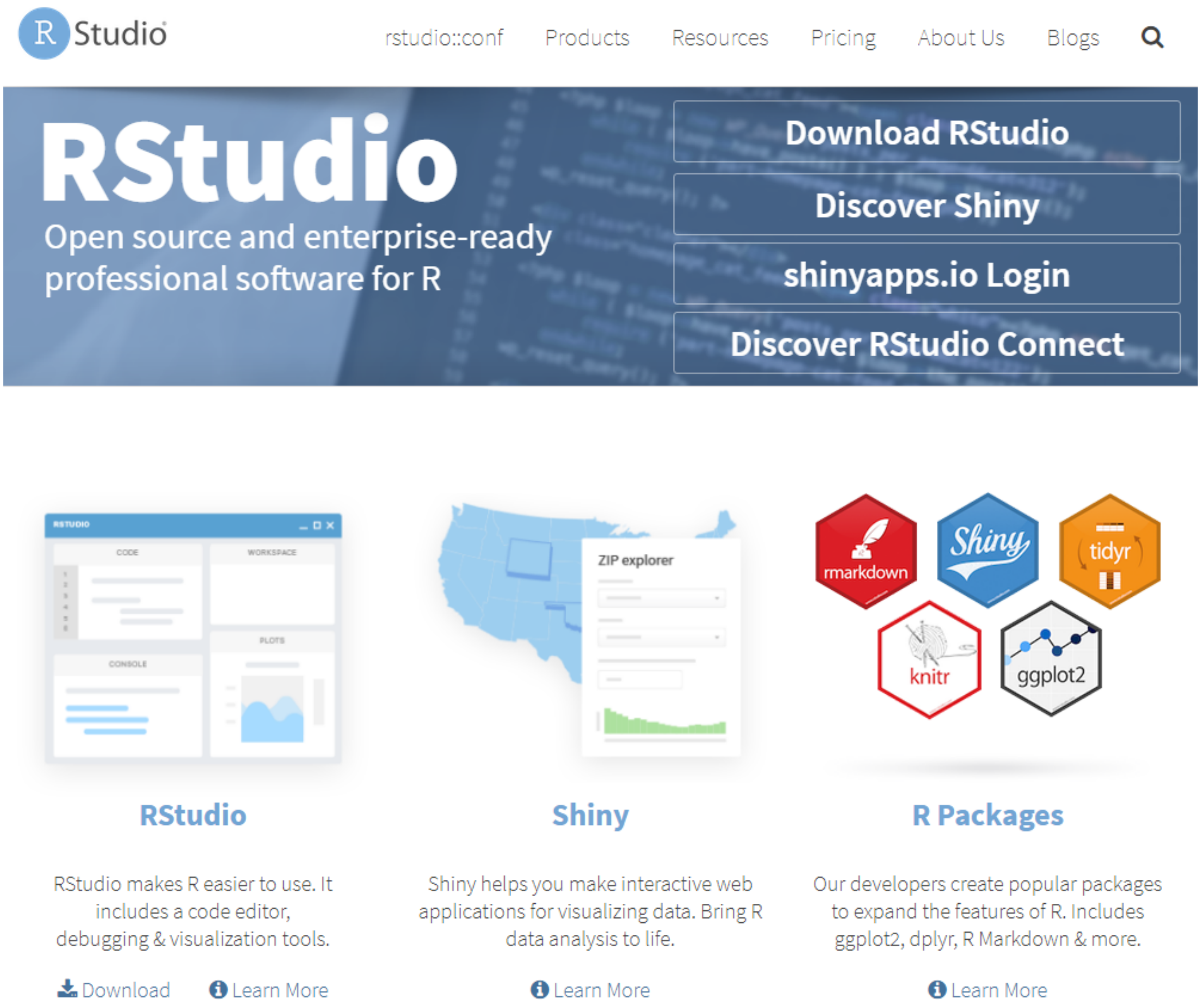
- Baumer et al., Modern Data Science with R. CRC Press. [Textbook webpage: https://mdsr-book.github.io/index.html]

Other Resources

- Grolemund and Wickham, R for Data Science. O'Reilly. [http://r4ds.had.co.nz/]

Topics

1. Introduction to Data Science
2. Introduction to Data Science tools: R and RStudio
3. Data Visualization
4. Data Wrangling
5. Ethics in Data Science
6. Statistical thinking in Data Science
7. Regression modeling
8. Machine Learning, dimension reduction, clustering, classification 9 A case study
9. Professional Reporting and reproducible analysis



RStudio

Open source and enterprise-ready professional software for R

RStudio

RStudio makes R easier to use. It includes a code editor, debugging & visualization tools.

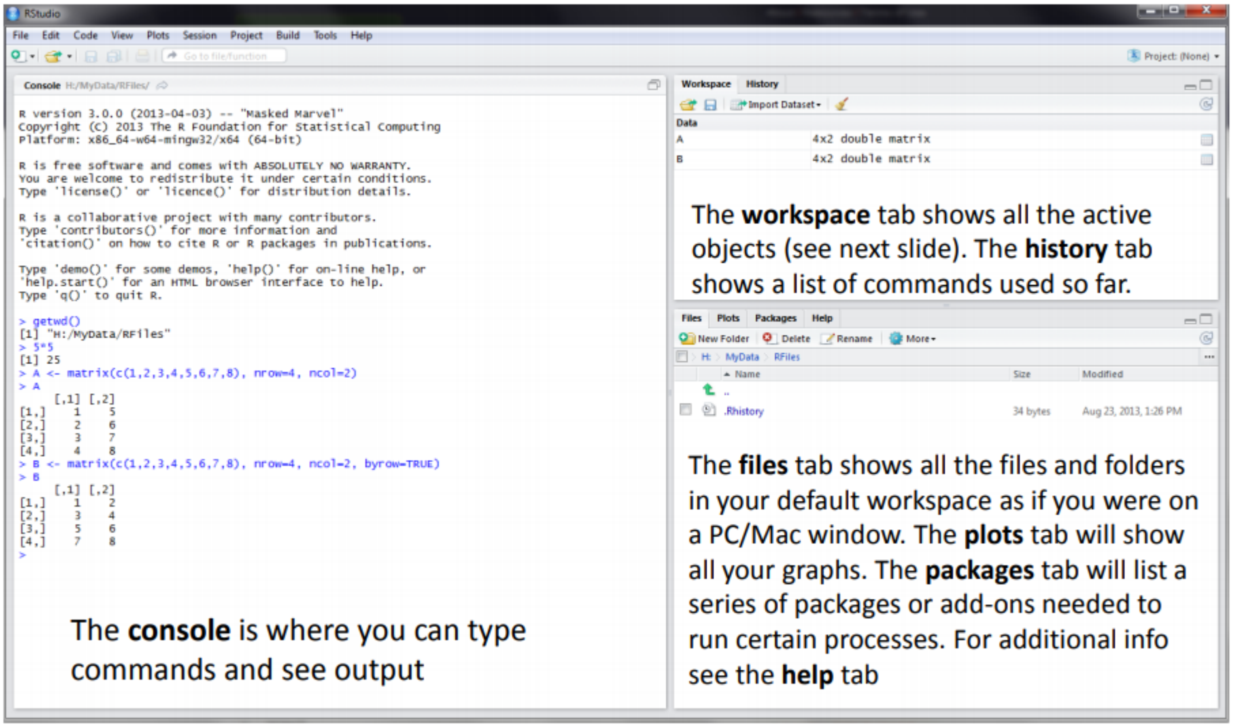
Shiny

Shiny helps you make interactive web applications for visualizing data. Bring R data analysis to life.

R Packages

Our developers create popular packages to expand the features of R. Includes ggplot2, dplyr, R Markdown & more.

Figure 3:


RStudio
File Edit Code View Plots Session Project Build Tools Help
Go to file/function
Project: (None)
Console H:/MyData/RFiles/
Workspace History
Import Dataset
Files Plots Packages Help
New Folder Delete Rename More
H: MyData RFiles
Name Size Modified
..
.Rhistory 34 bytes Aug 23, 2013, 1:26 PM

```
R version 3.0.0 (2013-04-03) -- "Masked Marvel"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> getwd()
[1] "H:/MyData/RFiles"
> 5*5
[1] 25
> A <- matrix(c(1,2,3,4,5,6,7,8), nrow=4, ncol=2)
> A
     [,1] [,2]
[1,]    1    5
[2,]    2    6
[3,]    3    7
[4,]    4    8
> B <- matrix(c(1,2,3,4,5,6,7,8), nrow=4, ncol=2, byrow=TRUE)
> B
     [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
[4,]    7    8
>
```

Data	
A	4x2 double matrix
B	4x2 double matrix

The **console** is where you can type commands and see output

The **workspace** tab shows all the active objects (see next slide). The **history** tab shows a list of commands used so far.

The **files** tab shows all the files and folders in your default workspace as if you were on a PC/Mac window. The **plots** tab will show all your graphs. The **packages** tab will list a series of packages or add-ons needed to run certain processes. For additional info see the **help** tab

Figure 4:

Syllabus

Visit Course webpage at http://www.stat.pitt.edu/sungkyu/course/pds/

Data Wrangling and Data Visualization

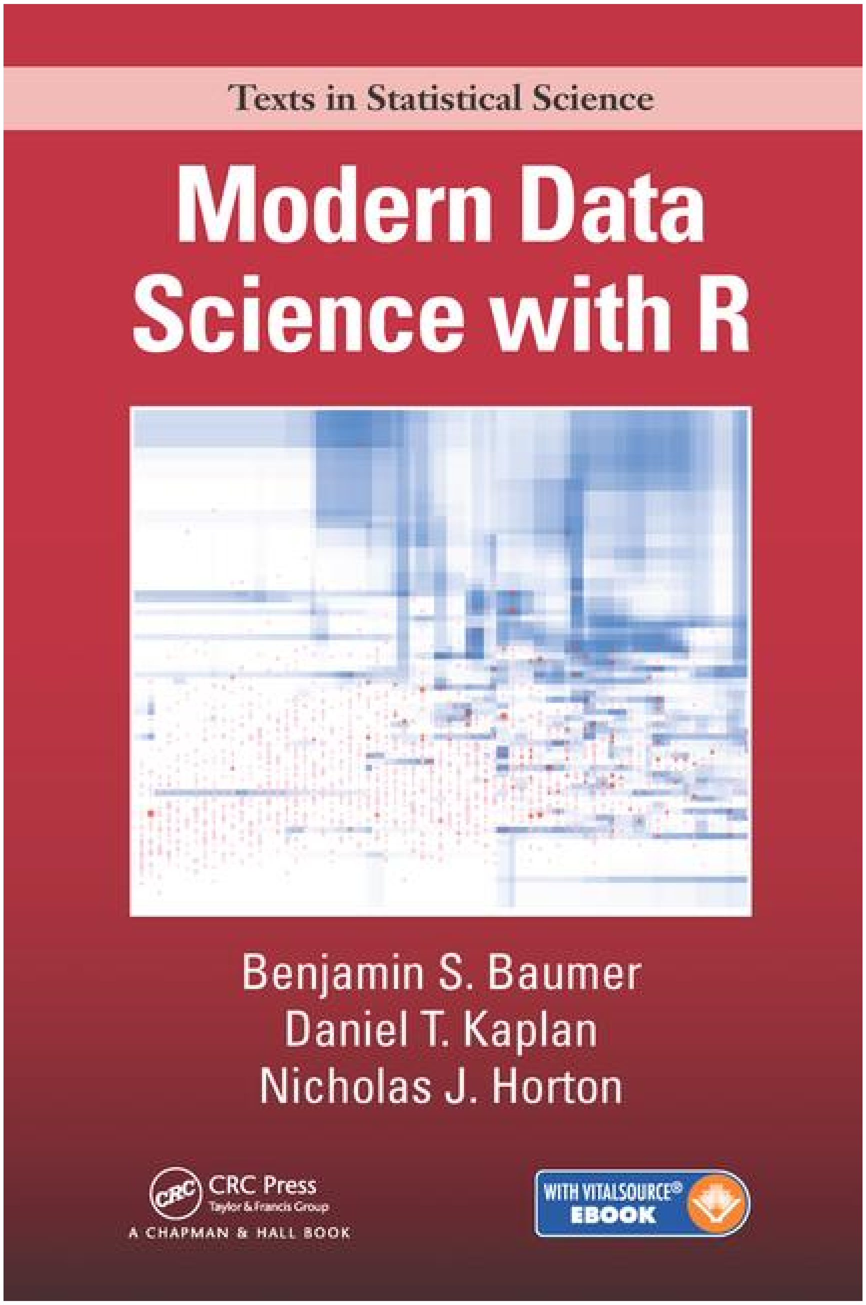
See an example Data, collecting sex and height from a group of people data

```
           Timestamp Height    Sex
1 9/2/2014 13:40:36     75   Male
2 9/2/2014 13:46:59     70   Male
3 9/2/2014 13:59:20     68   Male
4 9/2/2014 14:51:53     74   Male
5 9/2/2014 15:16:15     61   Male
6 9/2/2014 15:16:16     65 Female
```

Motivating Data Wrangling

Note that some entries are not in inches.

```
             Timestamp Height    Sex
127 9/2/2014 15:16:56   5'7"   Male
150 9/2/2014 15:17:09   5'3" Female
187 9/2/2014 15:18:00 5'8.11   Male
202 9/2/2014 15:19:48   5'11   Male
236  9/4/2014 0:46:45  5'9''   Male
55   9/2/2014 15:16:37  165cm Female
```



Texts in Statistical Science

Modern Data Science with R

Benjamin S. Baumer
Daniel T. Kaplan
Nicholas J. Horton

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

WITH VITALSOURCE®
EBOOK

Figure 5:

Fixing this is part of what we call data wrangling.

Data Wrangling

After fixing the above issue, there are still some problems:

```
             Timestamp   Height    Sex
12  9/2/2014 15:16:23     6.00   Male
40  9/2/2014 15:16:32     5.30 Female
66  9/2/2014 15:16:41   511.00   Male
84  9/2/2014 15:16:46     6.00   Male
99  9/2/2014 15:16:50     2.00 Female
126 9/2/2014 15:16:56  9000.00   Male
194 9/2/2014 15:18:14     5.25 Female
231 9/3/2014 21:43:00     5.50   Male
235 9/3/2014 23:55:37 11111.00   Male
241  9/4/2014 5:15:28     6.00 Female
242  9/4/2014 6:31:03     6.50   Male
244  9/4/2014 9:24:41   150.00 Female
```

We sometimes have to fix these “by hand”

Understanding Univariate Data

Look at the **distribution** of univariate data

$$F(a) = \text{Prob}(\text{Height} \leq a)$$

Distributions

Histograms show: $F(b) - F(a)$ for several intervals $(a, b]$

Easier to interpret than cumulative distribution functions

Normal Approximation

The distribution of many outcomes in nature are approximated by the normal distribution:

- μ is the average (also called the mean)
- σ is the standard deviation

Normal Approximation

If our data follows the normal distribution then μ and σ are a sufficient summary: they tell us everything!

All we need to know is μ and σ

```
       Average SD
Male        70  3
Female      65  3
```

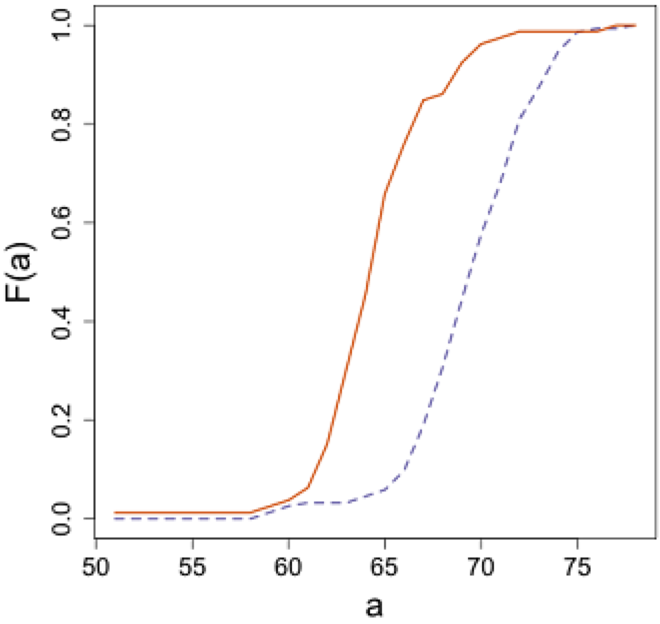


Figure 6:

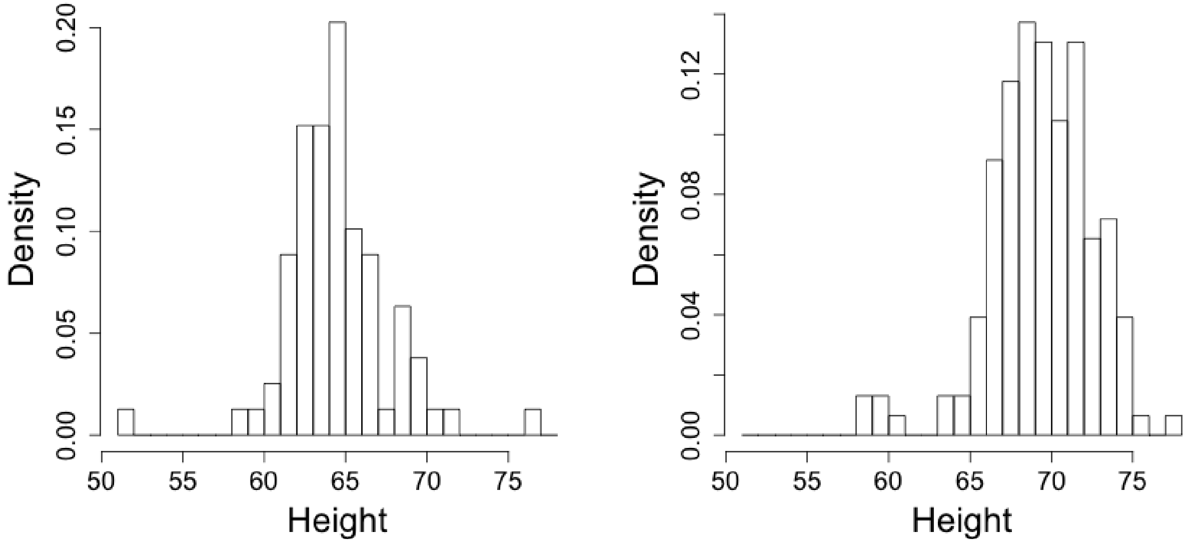


Figure 7:

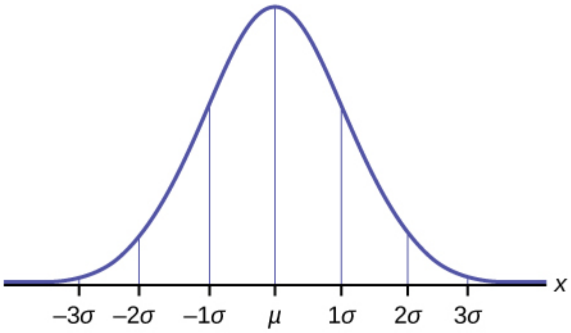


Figure 8:

How good is the normal approximation?

Here are the approximations for males

```
  Height Real Approx
1     63 0.02   0.03
2     65 0.07   0.06
3     67 0.16   0.10
4     68 0.31   0.31
5     70 0.50   0.44
6     71 0.69   0.68
7     73 0.84   0.88
8     75 0.93   0.95
9     76 0.98   0.99
```

QQ-plots

Observed versus normal approximation quantiles

Two variables

Normal approximation for two variables

Many pairs of data are bivariate normal

- The blue line is the average within each strata
- It is called the regression line

Regression line

The regression line is defined by this formula

$$\frac{Y - \mu_Y}{\sigma_Y} = \rho \frac{X - \mu_X}{\sigma_X}$$

- ρ is called the correlation coefficient
- For fathers and son heights it is 0.5

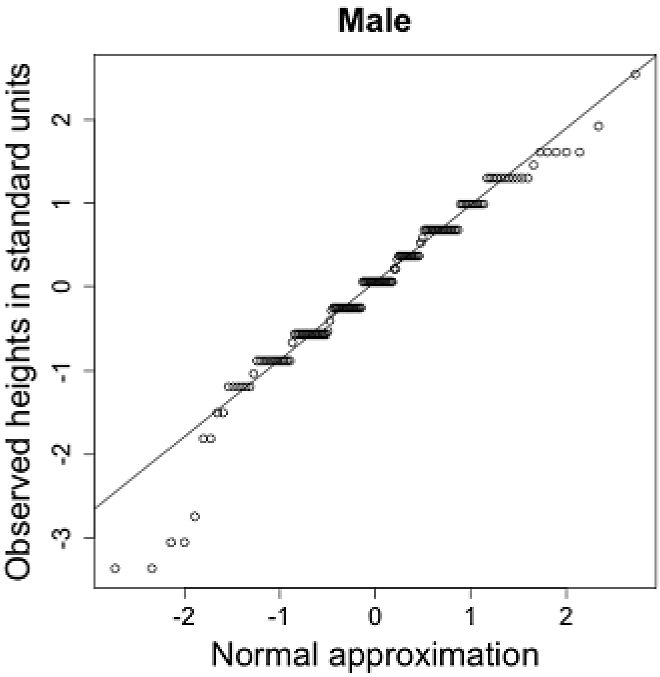


Figure 9:

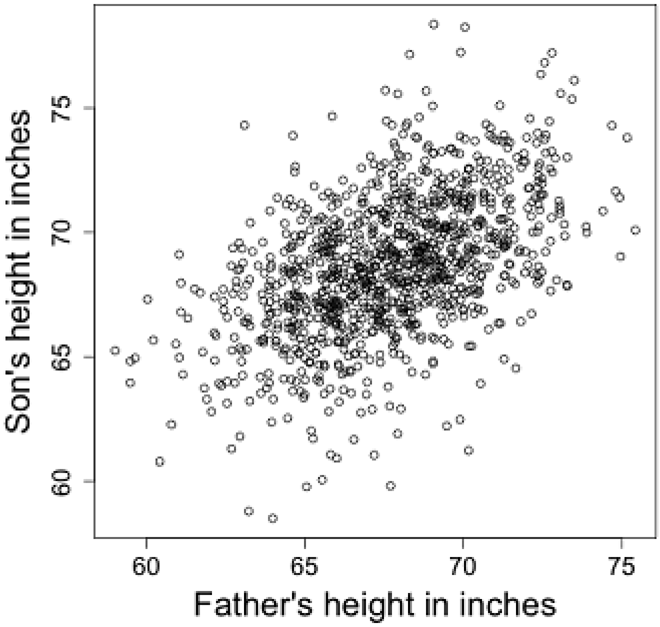


Figure 10:

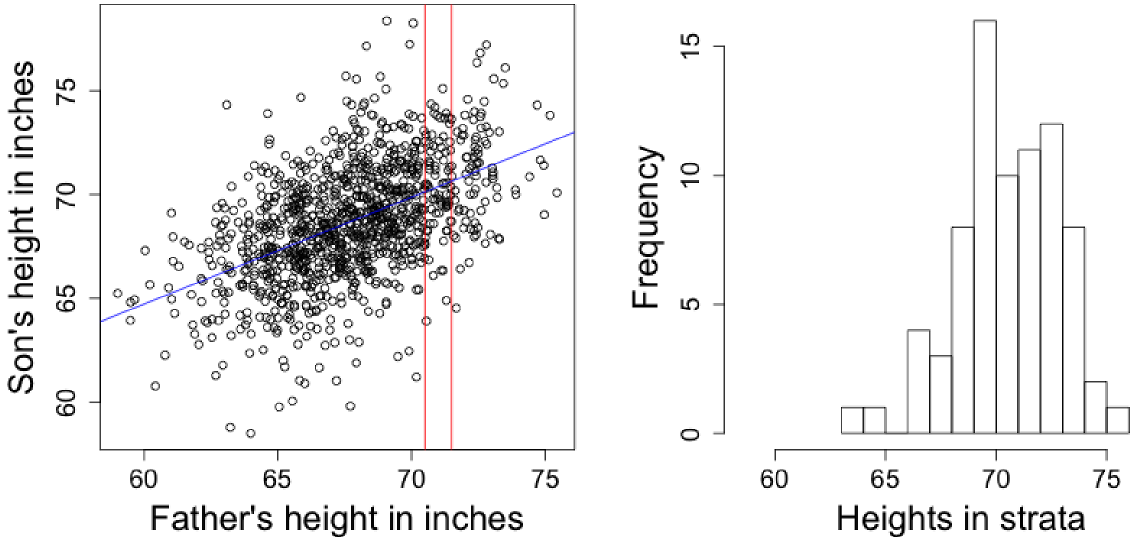


Figure 11:

- For bivariate normal pairs of data these five numbers provide a complete summary:

$$\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$$

Anscombe's quartet

Most data are not normal

For example, look at compensation for 199 US CEOs (2000)

Average is $600,000 but 84%, not 50%, make less.

The normal approximation is not useful here.

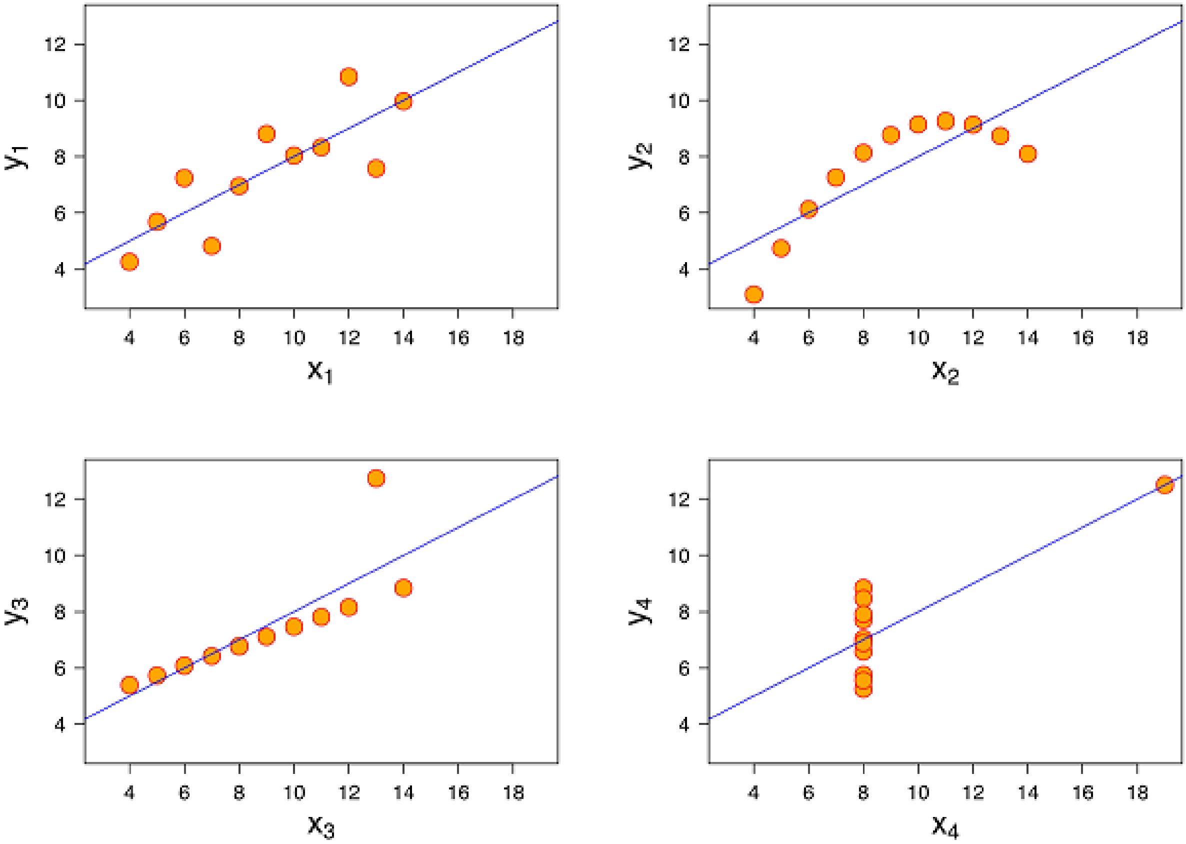

12
10
8
6
4
y_1
4
6
8
10
12
14
16
18
x_1
12
10
8
6
4
y_2
4
6
8
10
12
14
16
18
x_2
12
10
8
6
4
y_3
4
6
8
10
12
14
16
18
x_3
12
10
8
6
4
y_4
4
6
8
10
12
14
16
18
x_4

Figure 12:

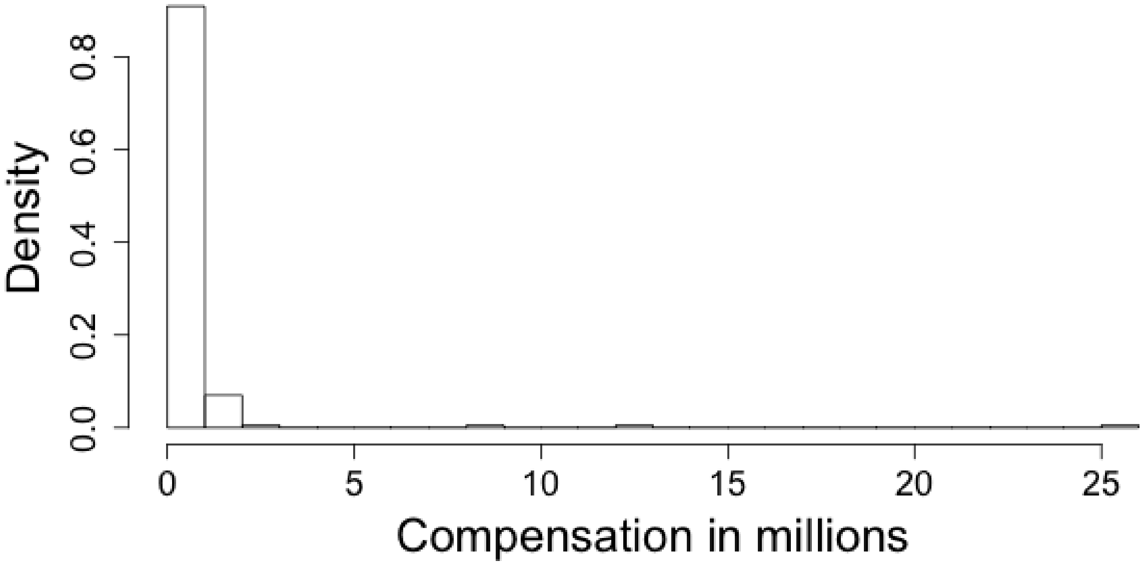


Figure 13: