STAT 350 Lecture 1: Exploratory Data Analysis

Graphical & Numerical Summaries of Data

(Chapter 1 of WMMY)

WMMY: Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying E. Ye (2016) *Probability and Statistics for Engineers and Scientists*, 9th edition.

Outline

1. Introduction
2. Data
3. Numerical Summaries
4. Graphical Summaries
5. The Empirical Rule
6. Linear Regression

Introduction to STAT 35000

Introduction to Statistics
Meghan Tooman
Office: LD 249
Email: mtooman@iupui.edu
Phone: (317) 274-6962
Office Hour: TR 3:00-4:00 PM (or by appointment)

Everything will be posted on **Canvas**!

Textbook: Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, Keying E. Ye (2016) *Probability and Statistics for Engineers and Scientists*, 9th edition.

Statistical Software: **R** is a free software environment on a wide variety of UNIX platforms, Windows and MacOS.
`http://www.r-project.org/index.html`

Calculator: **TI84**. **Required!**

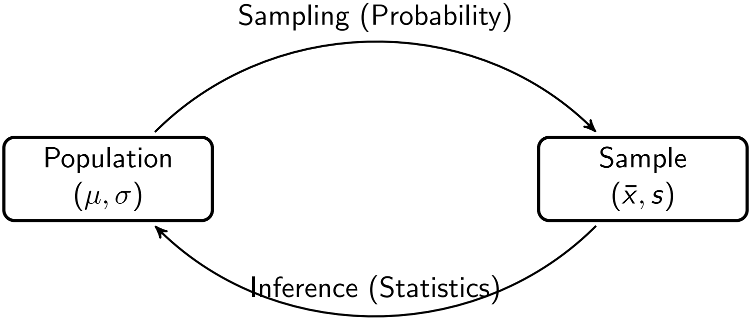
Key Issues on Syllabus

1. Please read the syllabus after class! [Required]
2. 8 closed book quizzes with the lowest score dropped.
3. 8 sets of homework with the lowest score dropped. All answers should be circled. Please do not use pencil for homework or exams. Otherwise, the mistakes made by the grader or the instructor won't be corrected! Staple your homework! No e-format. Do not slip your homework into the instructor's office.
4. 1 data analysis project using R. 1 midterm and 1 cumulative final.
5. No laptop/computer is allowed for exams and class meetings. Please practise calculation by hand or calculator when doing homework except for computer-related assignments. You are not allowed to use your cell phone or smart devices as calculator for exams.

Statistics in a nutshell

Statistics is **data science** pertaining collection, presentation, analysis and interpretation of data.

1. Population: a well-defined collection of objects.
2. Sample: a subset of the population.
3. Variable: characteristics of the objects.
4. Observation: an observed value of a variable.
5. Data: a collection of observations.



Introduction to Statistics

Read the following three articles and watch some videos from the YouTube channel "This is Statistics", and write a short report talking about your idea of Statistics within 800 words. Print it and submit it next Wednesday (1/20/2016) to earn 5 extra points.

1. For Today's Graduate, Just One Word: Statistics
2. Data, data everywhere
3. The Age of Big Data
4. This is Statistics

Outline

1 Introduction

2 Data

3 Numerical Summaries

4 Graphical Summaries

5 The Empirical Rule

6 Linear Regression

Terminology

Name	Gender	Age at death	Field of study
John	Male	45	Infectious Diseases
Janet	Female	89	Chronic Diseases
⋮	⋮	⋮	⋮

- *Individuals* (or *cases, observations*) are objects being described by a set of data, corresponding to the **rows in a data table**.
- *Variables* are characteristics of the individuals that can be measured, corresponding to the **columns in a data table**.
  - **Qualitative (categorical) variables** place an individual into one of several groups or categories (gender, eye color, letter grades, etc.).
  - **Quantitative variables** attach a numerical value to the individual such that adding or averaging these values makes sense (height, weight, age, income, etc.).

Qualitative variables

- Qualitative/categorical data can be either *nominal* or *ordinal*.
- Nominal variables put cases in to *unordered* groups.
  Example – Blood types of 16 children:

B	A	O	O	O	A	AB	O
AB	O	A	O	B	B	O	AB

- Ordinal variables put cases into *ordered* groups.
  Example – Class grades for 16 students:

B	D	A	B-	C	A	B	A-
F	C+	B	D	A	F	C-	C

- We won't deal with qualitative variables in STAT 35000!

Quantitative variables

- Quantitative data can be either *discrete* or *continuous*.
- Discrete variables take values of a countable set.
  Example – Number of defective widgets produced per hour in an 8-hour shift:

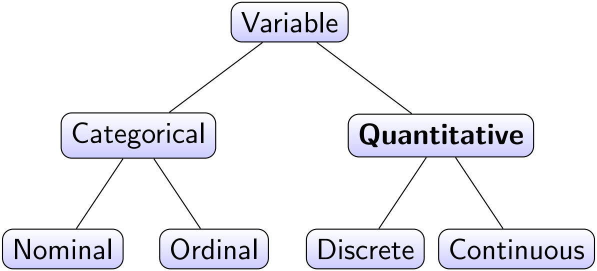
  4 2 4 5 10 5 3 6

  Gaps between values that “Number of defective widgets” can take. E.g. Integers.
- Continuous variables take values of an uncountable set.
  Example – Height (inches) of males:

72.0	69.9	66.8	66.0	70.3	67.8	70.1	65.2	70.4
69.8	67.5	64.9	75.3	69.6	66.0	68.9	68.3	73.9
71.8	72.6	65.9	65.1	69.6	69.9	70.2	66.3	70.1
71.0	69.2	69.8	66.0	71.0	70.7			

  Can be any number in an interval. E.g. real numbers
- Quantitative variables will be our focus and the distinction between discrete and continuous is important.

Types of Variables


Variable

Categorical

Quantitative

Nominal

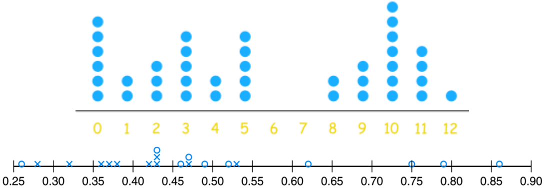
Ordinal

Discrete

Continuous

Distributions

- The different values in a data set, along with the frequency of (groups of) those values, makes up the **distribution**.
  - The simplest way to see the distribution of a variable is the **dot plot**.



- The goal of exploratory data analysis, or modeling, is to uncover patterns in the data that cannot be seen by simply looking at the list of numbers.
- Both graphical and numerical techniques can be used:
  - **Numerical methods** help when describing a data set to someone else (e.g., students generally want to know the average exam score, not the full list of scores).
  - **Graphical methods** show the shape of the distribution, which is particularly helpful for choosing a theoretical model.

Outline

1. Introduction
2. Data
3. Numerical Summaries
4. Graphical Summaries
5. The Empirical Rule
6. Linear Regression

Measures of center

One summary of the data is the location or center of the distribution. We will consider two such measures:

- **Sample mean** $\overline{x}$ of $\{x_1, \ldots, x_n\}$ is

  $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

  That is the average of the set of numbers.

- **Sample median** $\tilde{x}$ of $x_1, x_2, \ldots, x_n$ is the "middle value". How we calculate the median depends on if n is even or odd.

  1. Sort the data in ascending (increasing) order.
  2. If n is **odd**, then $\tilde{x}$ is the middle number.
     If n is **even**, $\tilde{x}$ is the average of the two "middle" numbers.

Examples

A study was conducted on the development of a relationship between the roots of trees and the action of a fungus. Two samples of 10 northern red oak seedlings were planted in a greenhouse, one containing seedlings treated with nitrogen and the other containing seedlings with no nitrogen. All seedlings contained the fungus *Pisolithus tinctorus*. The stem weights in grams were recorded after the end of 140 days. The data are given in the following table.

No Nitrogen	**Nitrogen**
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

- $\bar{x}$(no nitrogen)$=(0.32+0.53+\cdots+0.43)/10=0.399$.
- $\tilde{x}$(no nitrogen)$=(0.38+0.42)/2=0.400$.
- $\bar{x}$(nitrogen)$=(0.26+0.43+\cdots+0.46)/10=0.565$.
- $\tilde{x}$(nitrogen)$=(0.49+0.52)/2=0.505$.

Discussion: mean and median

Discussion: mean and median

- Are mean and median always approximately equal?

Discussion: mean and median

- Are mean and median always approximately equal?
  1. If the distribution is symmetric, the mean and median will be approximately equal.

Discussion: mean and median

- Are mean and median always approximately equal?
  1. If the distribution is symmetric, the mean and median will be approximately equal.
  2. The mean is sensitive to extreme values in the data set, whereas the median is not so much; for example

Data	Mean	Median
$\{1,2,3,4\}$	2.5	2.5
$\{1,2,3,14\}$	5	2.5

Discussion: mean and median

- Are mean and median always approximately equal?
  1. If the distribution is symmetric, the mean and median will be approximately equal.
  2. The mean is sensitive to extreme values in the data set, whereas the median is not so much; for example

Data	Mean	Median
$\{1, 2, 3, 4\}$	2.5	2.5
$\{1, 2, 3, 14\}$	5	2.5

- Could you come up with other robust measure of center?

Discussion: mean and median

- Are mean and median always approximately equal?
  1. If the distribution is symmetric, the mean and median will be approximately equal.
  2. The mean is sensitive to extreme values in the data set, whereas the median is not so much; for example

Data	Mean	Median
$\{1, 2, 3, 4\}$	2.5	2.5
$\{1, 2, 3, 14\}$	5	2.5

- Could you come up with other robust measure of center?
  1. **Trimmed mean**: is computed by "trimming away" a certain percent of both the largest and the smallest set of values. For example, the 10% trimmed mean is found by eliminating the largest 10% and smallest 10% and computing the average of the remaining values.

Discussion: mean and median

- Are mean and median always approximately equal?
  1. If the distribution is symmetric, the mean and median will be approximately equal.
  2. The mean is sensitive to extreme values in the data set, whereas the median is not so much; for example

Data	Mean	Median
$\{1, 2, 3, 4\}$	2.5	2.5
$\{1, 2, 3, 14\}$	5	2.5

- Could you come up with other robust measure of center?
  1. **Trimmed mean**: is computed by "trimming away" a certain percent of both the largest and the smallest set of values. For example, the 10% trimmed mean is found by eliminating the largest 10% and smallest 10% and computing the average of the remaining values.
  2. The trimmed mean is, of course, more insensitive to outliers than the sample mean but not as insensitive as the median. On the other hand, the trimmed mean approach makes use of more information than the sample median.

Measures of variability

- Student A Grades: 81, 83, 82, 82, 81, 81, 83
- Student B Grades: 60, 72, 85, 88, 81, 93, 94
- Student C Grades: 53, 69, 94, 68, 97, 94, 95

Which one you will choose to be your performance for our class? Why?

Measures of variability

- Measures of center alone are not sufficient for describing a distribution.
- For example, these two data sets have the same mean but are drastically different:

```
Data Set 1:  48 49 50 50 51 52
Data Set 2:   0 10 50 50 90 100
```

- The second dataset is much more *spread* out than the first.
- We'd like some way to measure how spread out a dataset (or distribution) is.
- We will consider only three measures of spread.
  - Range=$x_{\text{max}} - x_{\text{min}}$.
  - Variance/standard deviation.
  - Inter-quartile range (IQR).

Variance and standard deviation

- Variance and standard deviation (the two are equivalent) are the most common measures of variability.
- Roughly, the variance s^2 is the average (squared) distance of each observation x_i from the overall mean $\overline{x}$; in formula,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

- The use of $n - 1$ in the denominator instead of n is a bit mysterious, but there are good reasons for this!
- The standard deviation s is just the square root of the variance: $s = \sqrt{s^2}$.
- The standard deviation s is measured in the same units as the data

Example

An engineer is interested in testing the “bias” in a pH meter. Data are collected on the meter by measuring the pH of a neutral substance (pH = 7.0). A sample of size 10 is taken, with results given by

```
7.07 7.00 7.10 6.97 7.00 7.03 7.01 7.01 6.98 7.08.
```

1. Sample mean $\bar{x} = (7.07 + 7.00 + \cdots + 7.08)/10 = 7.025$;
2. Sample variance $s^2 = \frac{1}{9}[(7.07 - 7.025)^2 + (7.007.025)^2 + (7.107.025)^2 + \cdots + (7.087.025)^2] = 0.001939$.
3. Sample standard deviation $s = \sqrt{0.001939} = 0.044$.

Quartiles

- Quartiles are just certain percentiles:

$$
\begin{aligned}
Q_1 &= \text{1st quartile} = \text{25th percentile} \\
Q_2 &= \text{2nd quartile} = \text{50th percentile} = \tilde{x} \\
Q_3 &= \text{3rd quartile} = \text{75th percentile}
\end{aligned}
$$

- To find the quartiles:
  1. Split the data set at the median (**if *n* odd, count the median in both halves**)
  2. Q_1 is the median of the first half
  3. Q_3 is the median of the second half

Inter-quartile range (IQR)

- $\mathrm{IQR} = Q_3 - Q_1$
- Roughly, IQR tells us how wide the middle portion of the data set is.
- Not too sensitive to extreme values.
- Useful as a diagnostic for identifying *outliers*.

Five-number summary

- The five-number summary is just a list of five numbers that summarizes the center and spread of a distribution.
- Definition: $(\text{Min}, Q_1, Q_2(\tilde{x}), Q_3, \text{Max})$
- Can calulate the Range or IQR immediately from the five-number summary.

Find the five-number-summary statistics for the IUPUI male students' height data.

```
64.9 65.1 65.2 65.9 66.0 66.0 66.0 66.3 66.8 67.5 67.8
68.3 68.9 69.2 69.6 69.6 69.8 69.8 69.9 69.9 70.1 70.1
70.2 70.3 70.4 70.7 71.0 71.0 71.8 72.0 72.6 73.9 75.9
```

Five-number summary

- The five-number summary is just a list of five numbers that summarizes the center and spread of a distribution.
- Definition: $(\text{Min}, Q_1, Q_2(\tilde{x}), Q_3, \text{Max})$
- Can calulate the Range or IQR immediately from the five-number summary.

Find the five-number-summary statistics for the IUPUI male students' height data.

```
64.9 65.1 65.2 65.9 66.0 66.0 66.0 66.3 66.8 67.5 67.8
68.3 68.9 69.2 69.6 69.6 69.8 69.8 69.9 69.9 70.1 70.1
70.2 70.3 70.4 70.7 71.0 71.0 71.8 72.0 72.6 73.9 75.9
```

Five-number summary: $(64.9, 66.8, 69.8, 70.4, 75.9)$

Outline

1. Introduction
2. Data
3. Numerical Summaries
4. Graphical Summaries
5. The Empirical Rule
6. Linear Regression

Frequency Table – Discrete Data

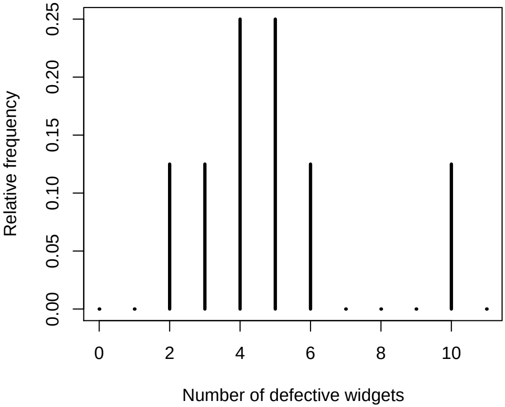
- A frequency table requires: all possible values in data set, corresponding frequencies / relative frequencies
- For discrete data, directly report separate values from data set, no groupings is necessary
- E.g. There are 8 production lines, each line could generate 2-10 defective items, lines are naturally grouped by # of defectives generated, we don't need to group

# defective items	frequency f	Relative frequency f/n
2	1	0.125
3	1	0.125
4	2	0.250
5	2	0.250
6	1	0.125
7	0	0.000
8	0	0.000
9	0	0.000
10	1	0.125
	$n = 8$	1.000

Note: $\sum f/n = 1$

Line graphs

- *Only for discrete variables!*
- y-axis could be frequency or relative frequency
  E.g. Number of defective widgets produced per hour in an 8-hour shift: 4 2 4 5 10 5 3 6



Frequency Table – Continuous Data

33 IUPUI male students' height in inches:

64.9 65.1 65.2 65.9 66.0 66.0 66.0 66.3 66.8 67.5 67.8

68.3 68.9 69.2 69.6 69.6 69.8 69.8 69.9 69.9 70.1 70.1

70.2 70.3 70.4 70.7 71.0 71.0 71.8 72.0 72.6 73.9 75.9

Frequency Table – Continuous Data

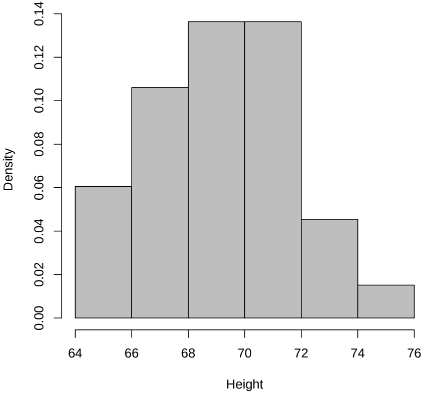
- For continuous data, we need to "group" the data ourselves!
- y-axis could be frequency, relative frequency, or density
- E.g. IUPUI male students' height in inches, need to group all possible heights

Height range	x	w	f	f/n	$d = f/wn$
$[64, 66)$	65	2	4	0.121	0.061
$[66, 68)$	67	2	7	0.212	0.106
$[68, 70)$	69	2	9	0.273	0.136
$[70, 72)$	71	2	9	0.273	0.136
$[72, 74)$	73	2	3	0.091	0.045
$[74, 76)$	75	2	1	0.030	0.015
			$n = 33$	1.000	

Note: $\sum wd = \sum f/n = 1$.
When y-axis is density total area of histogram is 1.

Histograms

- Histogram for IUPUI male students' height:



- *Only for continuous variables!*

Histograms, cont.

- When the groupings or bins are provided (as in height example) histograms are easy.
- But, in “real life,” no one will tell you the bins and, unfortunately, the choice of bins makes a difference.
- One way to choose the number of bins b is

Sturge’s Rule: Choose b such that $2^{b-1} \approx n$.

- In heights example, $n = 33$ and $b = 6$:

$$2^{6-1} = 32 \approx 33.$$

- When using software (e.g., R), the bins are conveniently chosen for you!

Stem-and-leaf plot

- A stem-and-leaf plot is another way to visualize the distribution.
- Has *stems* that act like the horizontal axis on the histogram, and *leaves* that act as the vertical axis.
- Difference between histogram and stem-leaf plot:
    - Individual data values cannot be recovered from a histogram picture alone
    - The complete data set can be recovered from a stem-and-leaf plot
    - Seems a stem-and-leaf plot is more informative than a histogram. But stem-and-leaf plot is not suitable for large data sets.

How to make stem-and-leaf plot

1. Select one or more leading digits for the stem values (any value appropriate). The trailing digits become the leaves.
2. List possible stem values in a vertical column.
3. Put the leaf for each observation besides the corresponding stem.
4. Indicate the units for stems and leaves.

Stem-and-leaf plot – men's height example

- Use tens & ones as stem, tenths as leaves.

```
64 | 9
65 | 1 2 9
66 | 0 0 0 3 8
67 | 5 8
68 | 3 9
69 | 2 6 6 8 8 9 9
70 | 1 1 2 3 4 7
71 | 0 0 8
72 | 0 6
73 | 9
74 |
75 | 9
```

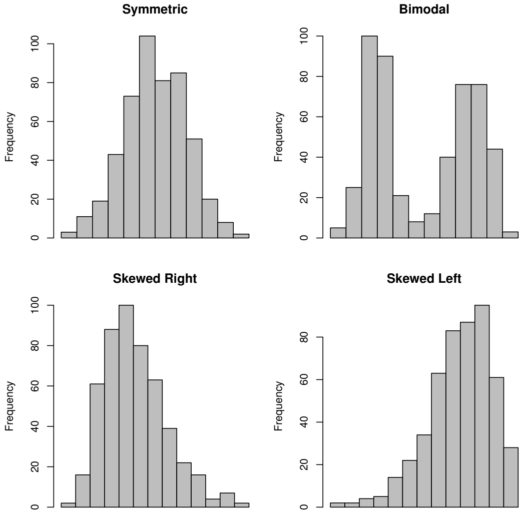
- Can you recover all data points by reading the stem-leaf plot?

Shapes of distributions

Three aspects:

1. *Does the histogram have a single, central hump or several separated humps?* These humps are called **modes**. A histogram with one peak, such as the earthquake magnitudes, is called **unimodal**; histograms with two peaks are **bimodal**, and those with three or more are called **multimodal**. A histogram that doesn't appear to have any mode and in which all the bars are approximately the same height is called **uniform**.
2. *Is the histogram* **symmetric***?* The usually thinner ends of a distribution are called the **tails**. If one tail stretches out farther than the other, the histogram is said to be **skewed** to the side of the longer tail.
3. *Do any unusual features stick out?* You should always mentions any stragglers, or **outliers**, that stand off away from the body of the distribution.

Shapes of distributions



Boxplots

- A graphical representation of the five-number summary.
- Also called Box-Whisker plot
- Reveals skewness: focus on the central box to tell the skewness based on the relative quartile differences $(Q_3 - Q_2)$ vs $(Q_2 - Q_1)$.
- Is particularly useful for comparing two distributions by using the so called side-by-side boxplot.

Boxplots

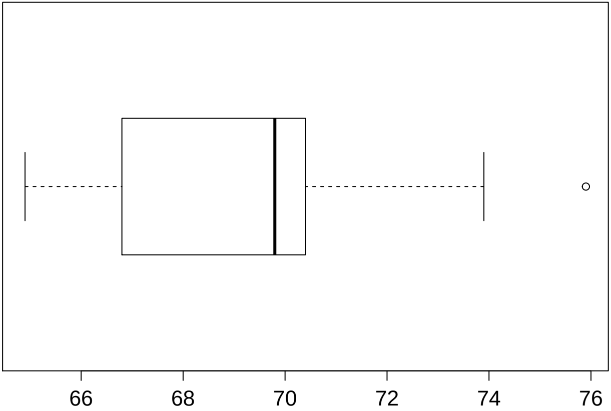
How to use 5-number-summary to draw boxplot:

1. Draw the box: Q_1, Median, Q_3;
2. Calculate $IQR = Q_3 - Q_1$;
3. Get upper fence and lower fence: uf=Q_3+1.5×IQR, lf=Q_1-1.5×IQR;
4. By fences, identify the outliers and the whiskers:
   - lower whisker is the smallest data point which is above or equal to the lower fence;
   - upper whisker is the biggest data point which is below or equal to the upper fence.
   - data points outside of the fences are classified to be outliers

The textbook adopts a simple version: no outliers, no whiskers

IUPUI male students' height example

Five-number summary: $(64.9, 66.8, 69.8, 70.4, 75.9)$



Side-by-side boxplots

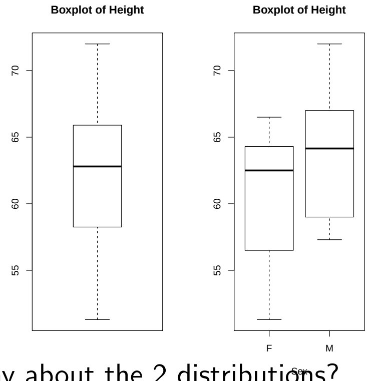
- When there are two related data sets in consideration, it is often important to compare the two.
- We could draw two histograms, but this might be hard to read.
- It would be nice to see both distributions on a single plot!
- We can put two (or more) boxplots on the same picture and easily compare center and spread.

Example: Middle School Class Data

```
       name sex age height weight
1     Alice   F  13   56.5   84.0
2     Becka   F  13   65.3   98.0
3      Gail   F  14   64.3   90.0
4     Karen   F  12   56.3   77.0
5     Kathy   F  12   59.8   84.5
6      Mary   F  15   66.5  112.0
...    ...     ... ...   ...    ...
12    Guido   M  15   67.0  133.0
13    James   M  12   57.3   83.0
14  Jeffrey   M  13   62.5   84.0
15     John   M  12   59.0   99.5
16   Philip   M  16   72.0  150.0
17   Robert   M  12   64.8  128.0
18   Thomas   M  11   57.5   85.0
19  William   M  15   66.5  112.0
```

Side-by-side boxplot – Example

- Boy's and girl's height.



- What can we say about the 2 distributions?
  Compare central tendency, range, spread, skewness.

Outline

1. Introduction
2. Data
3. Numerical Summaries
4. Graphical Summaries
5. The Empirical Rule
6. Linear Regression

The Empirical Rule (68-95-99.7 Rule)

If a data set is approx. normal with sample mean $\overline{x}$, sample standard deviation s, then:

- Approximately 68% of the observations lie within $\overline{x} \pm s$
- Approximately 95% of the observations lie within $\overline{x} \pm 2s$
- Approximately 99.7% of the observations lie within $\overline{x} \pm 3s$

Note that $z = \frac{x-\bar{x}}{s}$ is very important. It is a measure of relative standing by taking off the effects of center and spread out.

Outline

1 Introduction

2 Data

3 Numerical Summaries

4 Graphical Summaries

5 The Empirical Rule

6 Linear Regression

Why regression?

If you want to buy your girlfriend a ring but you have no idea about the ring size of your girlfriend, could you use the shoe size of your girlfriend to predict it?

- Want to model a functional relationship between two continuous variables, one "**predictor** variable" (input, independent variable, etc.), shoe size; and one "**response** variable" (output, dependent variable, etc.), ring size.
- The roles of the two variables are different. We are more interested in the response variable than the predictor variable.
- Two distinct goals
  - Understanding the relationship between predictor and response—whether the relationship between ring size and shoe size is significant? (answered at the end of the semester)
  - Predicting the future response given the new observed predictor—predict the ring size for your girlfriend with her shoe size by the model learned from all of your girl friends

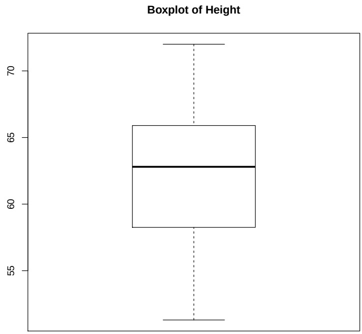
Middle School Class Data Example

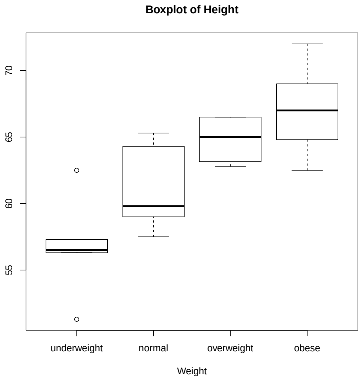
```
      name sex age height weight
1    Alice   F  13   56.5   84.0
2    Becka   F  13   65.3   98.0
3     Gail   F  14   64.3   90.0
4    Karen   F  12   56.3   77.0
5    Kathy   F  12   59.8   84.5
6     Mary   F  15   66.5  112.0
...  ...    ... ...  ...    ...
12   Guido   M  15   67.0  133.0
13   James   M  12   57.3   83.0
14 Jeffrey   M  13   62.5   84.0
15    John   M  12   59.0   99.5
16  Philip   M  16   72.0  150.0
17  Robert   M  12   64.8  128.0
18  Thomas   M  11   57.5   85.0
19 William   M  15   66.5  112.0
```

How to study the relationship

1. Recall that we can use the side-by-side boxplot to study the relationship between one categorical variable and one continuous variable.
2. How to study the relationship between two continuous variables, height and weight?
   - Grouping weight into different categories: underweight, normal, overweight, obese.
   - Using side-by-side boxplot to see the relationship.

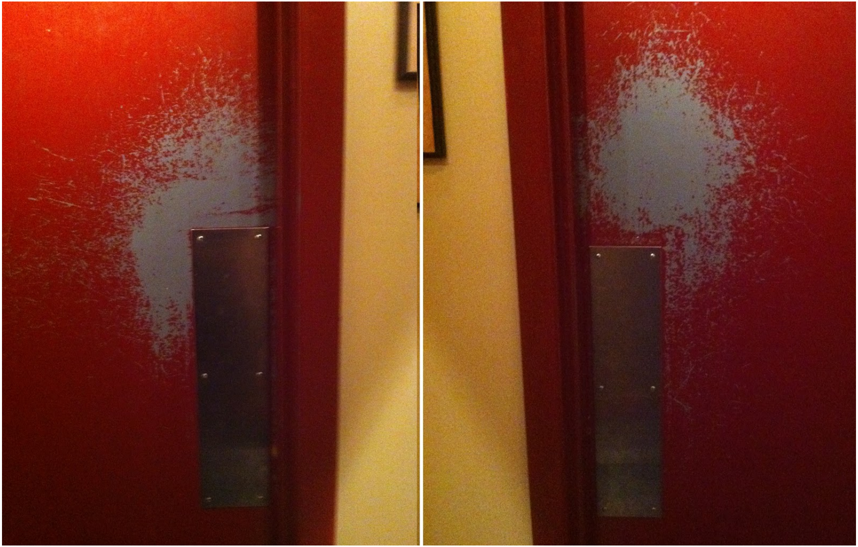
Side-by-side boxplot after grouping





1. Compare the variation of height for the whole group and that for those four subgroups.
2. The variation in height is partially explained by the weight.

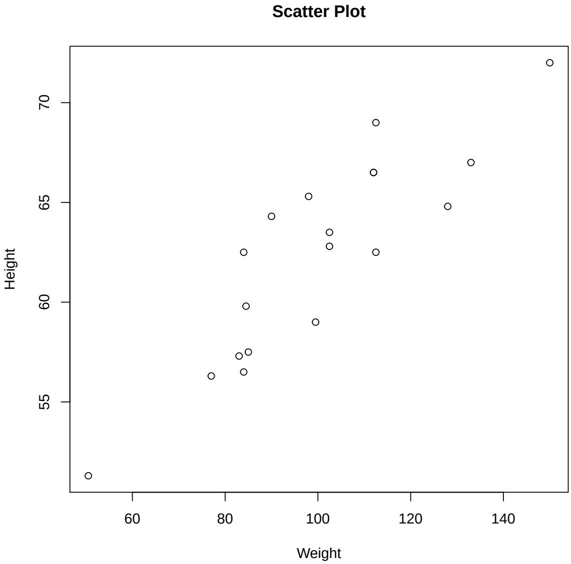
Explanation of Variation



Scatterplots

- A *scatterplot* is simply a plot of the observed data points $(x_1, y_1), \ldots, (x_n, y_n)$.
- It's just a picture of dots—no lines connecting them!
- Scatterplots can be used to identify the relationship or *association* between the two variables.
- In particular, look at
  - *Form.* Linear? Non-linear?
  - *Direction.* Positive? Negative?
  - *Strength.* Strong? Weak?

Scatter plot



Correlation coefficient

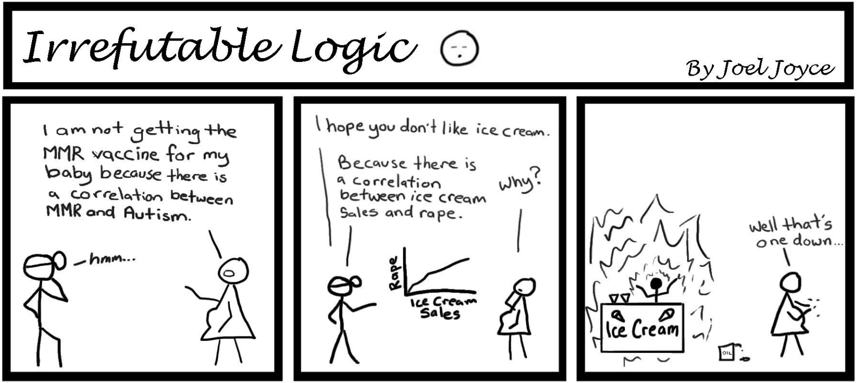
- The correlation coefficient r is a measure of the direction and strength of a **linear** relationship between two quantitative variables.
- Notation:
  - Sample data $(x_1, y_1), \ldots, (x_n, y_n)$.
  - Sample means $\overline{x}$ and $\overline{y}$
  - Sample std devs s_x and s_y.
- Then the correlation is defined as

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \underbrace{\left(\frac{x_i - \overline{x}}{s_x}\right)}_{\color{red}{z_{x_i}}} \underbrace{\left(\frac{y_i - \overline{y}}{s_y}\right)}_{\color{red}{z_{y_i}}} = \frac{s_{xy}^2}{s_x s_y}.$$

Properties of r

- r is independent of the units of x and y.
- $r > 0$ indicates a positive association.
- $r < 0$ indicates a negative association.
- r is always between ± 1.
- r values near 0 indicate a weak linear relationship, while values near ± 1 indicate a strong linear relationship. In particular
  - $r \in (0.8, 1]$: strong linear;
  - $r \in (0.3, 0.8]$: moderate linear;
  - $r \in [0, 0.3]$: weak linear.
- Remember, r measures only *linear* relationships!

Correlation is not Causation


Irrefutable Logic
By Joel Joyce
I am not getting the MMR vaccine for my baby because there is a correlation between MMR and Autism.
-hmm...
I hope you don't like ice cream.
Because there is a correlation between ice cream sales and rape.
Why?
Rape
Ice Cream Sales
Ice Cream
OIL
Well that's one down...

Least Squares Linear Regression

- Goal: make y_i and $a + bx_i$ close for all $i = 1, 2, \cdots, n$.
- Proposal: minimize

$$Q(a, b) := \sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

- Choose a, b as estimators for α, β, denoted as $\hat{\alpha}, \hat{\beta}$, i.e.

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{a,b} \sum_{i=1}^{n} [y_i - (a + bx_i)]^2.$$

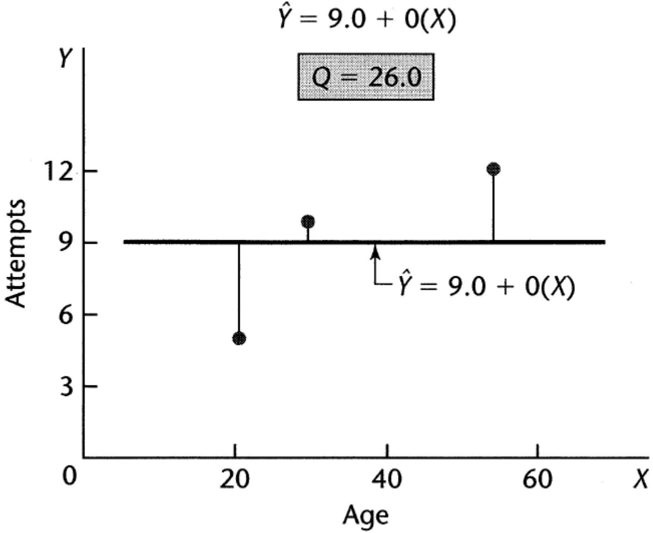
- The regression line is of the form $\hat{y} = \hat{\alpha} + \hat{\beta}x$ for a given new x. $\hat{y}$ is called the predicted response for X.
- Residual for observation i is defined to be: $e_i = y_i - \hat{y}_i$, where y_i is true value of response for observation i and $\hat{y}_i$ is estimated response for observation i.

Example

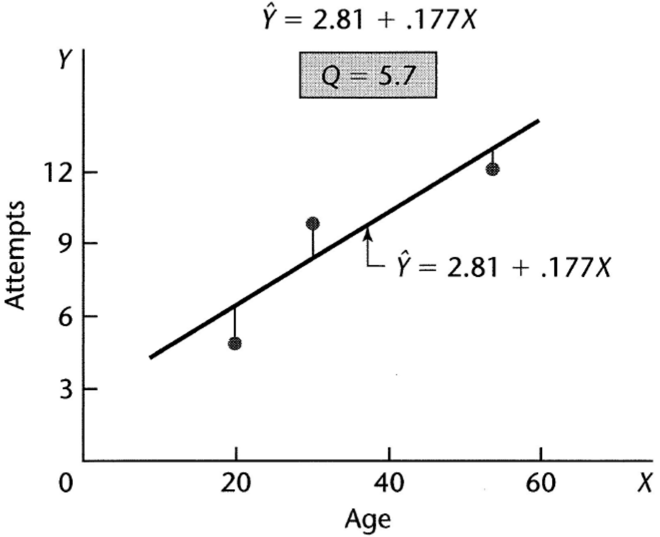
An experimenter gave three subjects a very difficult task. Data on the age of the subject (x) and on the number of attempts to accomplish the task before giving up (y) follow:

Subject i	1	2	3
Age x_i	20	55	30
# of attempts y_i	5	12	10

Guess 1



Guess 2


$\hat{Y} = 2.81 + .177X$
$Q = 5.7$
Y
12
9
6
3
0
Attempts
$\hat{Y} = 2.81 + .177X$
20
40
60
X
Age

LS estimates

- By solving the minimization problem, you'll find that the LS estimates of α, β are

$$\hat{\alpha} = \frac{\sum_{i=1}^n y_i - \hat{\beta}\sum_{i=1}^n x_i}{n} = \overline{y} - \hat{\beta}\overline{x},$$
$$\hat{\beta} = \frac{n\sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = r\frac{s_y}{s_x}.$$

- Then the estimated regression line is $\hat{y} = \hat{\alpha} + \hat{\beta}x$, i.e.

$$\frac{\hat{y} - \overline{y}}{s_y} = r\frac{x - \overline{x}}{s_x}$$

- If $-1 < r < 1$, then we say that the data points exhibit **regression** toward the mean. In other words, the predicted standardized value of y is closer to its mean than the standardized value of x is to its mean.

Parameter Interpretation

- Intercept α: α is **average response** when $x = 0$
- Slope β: β is change of **average response** when x increases by 1 unit.
- **Coefficient of determination** $R^2 = r^2$: R^2 is proportion of variation in response y that is explained by linear relationship with predictor x.

Middle School Class Data Example

```
summary statistics for height and weight:

    height           weight
 Min.   :51.30   Min.   : 50.50
 1st Qu.:58.25   1st Qu.: 84.25
 Median :62.80   Median : 99.50
 Mean   :62.34   Mean   :100.03
 3rd Qu.:65.90   3rd Qu.:112.25
 Max.   :72.00   Max.   :150.00

variance and covariance for height and weight:

         height   weight
height  26.2869 102.4934
weight 102.4934 518.6520
```

1. $\overline{x} = 100.03, \overline{y} = 62.34, s_x = \sqrt{518.6520}, s_y = \sqrt{26.2869}, s_{xy}^2 = 102.4934$
2. $r = \frac{s_{xy}^2}{s_x s_y} = 0.8777813$
3. $\hat{\beta} = r\frac{s_y}{s_x} = 0.19761$
4. $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = 42.57014$

Middle School Class Data Example

From the results shown above, we can see that the prediction line equation is

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = 42.57014 + 0.19761x.$$

And the coefficient of determination $R^2 = r^2$=0.7705, which means there are 77.05% of the variation in the response variable `height` explained by the predictor variable `weight`. And $r = 0.8777813$, which indicates that the linear relationship between `height` and `weight` is positive and strong.