Introduction to Data Science With Probability and Statistics

Lecture 2: Exploratory Data Analysis & Summary Statistics

CSCI 3022 - Summer 2020
Sourav Chakraborty
Dept. of Computer Science
University of Colorado Boulder

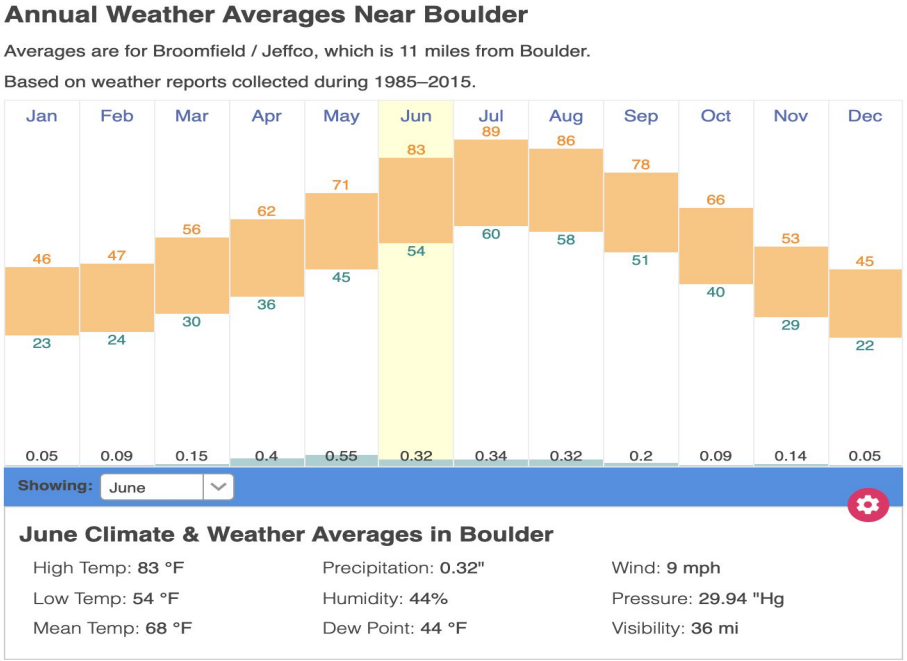
What will we learn today?

- Mean
- Median
- Mode
- Quartiles
- Sample Variance
- Sample Standard Deviation
- Interquartile Range (IQR)
- 5-Number Summary

❑ *A Modern Introduction to Probability and Statistics, section 16.1-16.3*

Exploratory Data Analysis

How do you summarize and describe a set of data



* Charts taken from TimeandData.com

Exploratory Data Analysis - Central Tendency

Center of a Dataset

Sample Mean: $\bar{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n}$

Sample Median: The *middle element* of the dataset when it is put in ascending order.
Note: When n is even, take the average of the middle two elements.

Example: Compute the mean and median of the following set of data.

10	11	14	19	23	11	13	16	17	18	9	13	11,375

Sample Mean: $\bar{x} = \dfrac{10+11+14+19+23+11+13+16+17+18+9+13+11{,}375}{13} = 888.384615$

Sample Median: 9 10 11 11 13 13 **14** 16 17 18 19 23 11,375

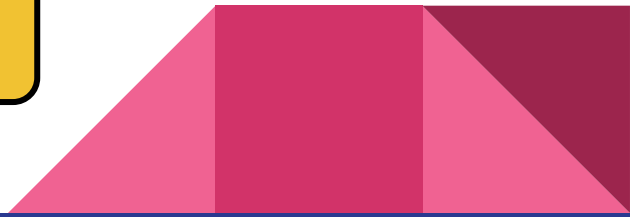
Exploratory Data Analysis - Central Tendency

Data scientists hope to learn about some characteristic of a population, typically we cannot study the entire population so we study a **sample**.

A **population** is a collection of units. (e.g. people, songs, tweets, incomes, daily temperatures ...)

A **sample** is a subset of the population

A **characteristic/variable of interest (VOI)** is something we want to measure for each unit



Exploratory Data Analysis

Example: Suppose that the city of Boulder wants to estimate its per-household income via a phone survey. They call every 50th number on a list of Boulder phone numbers between 6PM and 8PM.

What is the **population**?

What is the **sample**?

What is the **Variable of Interest**?

Exploratory Data Analysis

The **sample frame** is the source material or device from which the sample is drawn.

Simple random sample: randomly select people from a sample frame.

Systematic sample: Order the sample frame. Choose an integer k. Sample every k^{th} unit in the sample frame.

Census sample: sample everyone/everything in the population.

Stratified sample: if you have a heterogeneous population that can be broken up into homogeneous groups, randomly sample from each group proportionate to their prevalence in the population.

Exploratory Data Analysis - Central Tendency

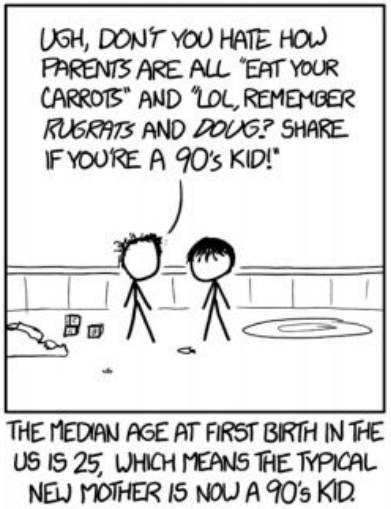
Summarizing the “center” of the sample is a popular way to characterize data.

Measures of Central Tendency: Mean, Median, and Mode

Sample Mean: $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{k=1}^{n} x_k$

Sample Median: The *middle element* of the dataset when it is put in ascending order.
Note: When n is even, take the average of the middle two elements

Sample Mode: Most commonly occurring value in a data set.



Exploratory Data Analysis - Central Tendency

Sample Mean: $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{k=1}^{n} x_k$

Calculation:

Add up all numbers in the data set.

Divide by number of numbers.

Example: Compute the mean of the following data set.

17 + 10 + 10 + 18 + 5 + 23

83/6 ≈ 13.

∴

Exploratory Data Analysis - Central Tendency

Sample Median: The *middle element* of the dataset when it is put in ascending order.
Note: When n is even, take the average of the middle two elements

Calculation:

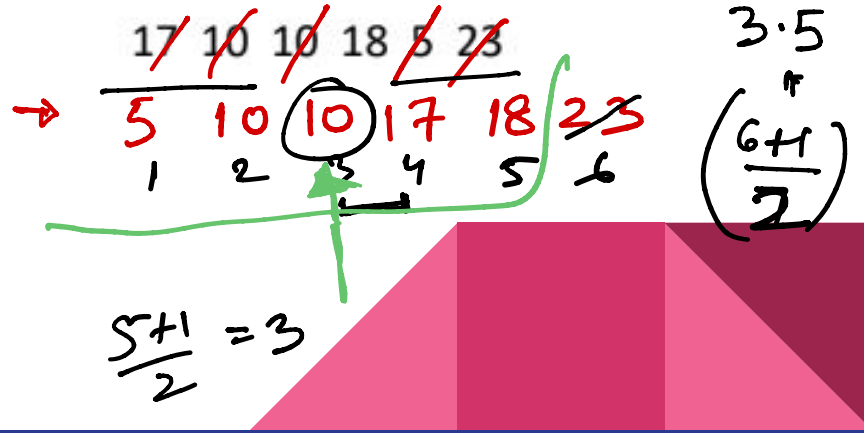
Order the n observations from smallest to largest
Include multiple instances of repeated values.

If n is odd, then $\tilde{x} = \left(\frac{n+1}{2}\right)^{th}$ ordered value.

If n is even, then $\tilde{x}$ = average of $\left(\frac{n}{2}\right)^{\text{th}}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$

$\bar{x}$, $\tilde{x}$

Example: Compute the median of the following data set.



Exploratory Data Analysis - Central Tendency

Sample Mode: Most commonly occurring value in a data set.

Calculation:

Count what number/data point occurs most often

Example: Compute the mode of the following data set.

17 10 10 18 5 23

Mode=10

1 2 3 4 5

Exploratory Data Analysis - Central Tendency

Example (from earlier): Compute the mean and median of the following set of data.

10	11	14	19	23	11	13	16	17	18	9	13	11,375

Sample Mean: $\bar{x} = \frac{10+11+14+19+23+11+13+16+17+18+9+13+11,375}{13} = 888.384615$

Sample Median: 9 10 11 11 13 13 **14** 16 17 18 19 23 11,375

Sample Mean (aka Arithmetic Average)

Advantages: Simple to compute

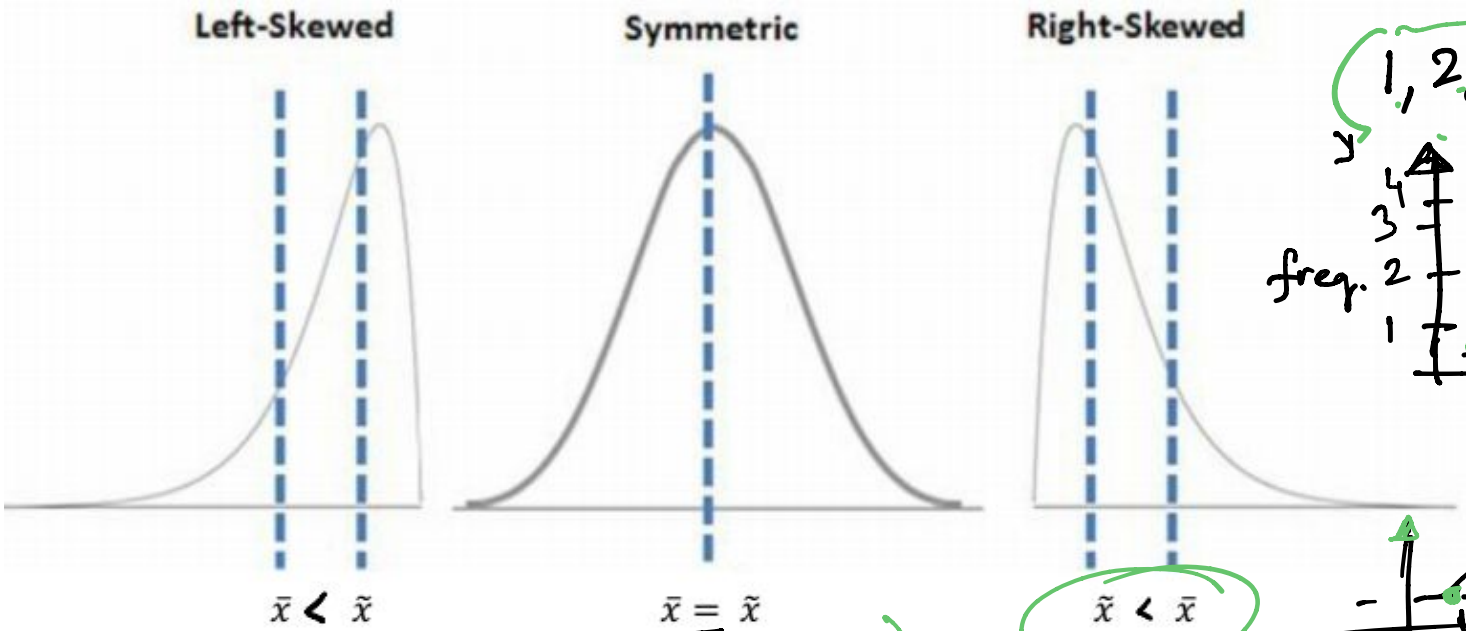
Disadvantages: Very sensitive to outliers

Sample Median

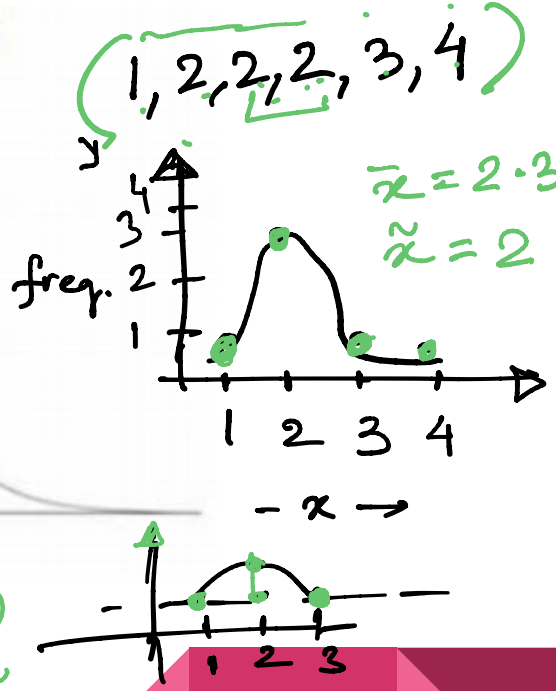
Advantages: Not as easily affected by a few outliers

Disadvantages: Doesn't depend on all entries in data set

Exploratory Data Analysis - Central Tendency



- The population mean and median will generally not be equal.



1 2 2 3

Mean = 2, Median = 2, mode = 2

Data Skewness

Right

~~Left~~ Skewness

$\frac{14+1}{2}$

1 2 3 4 5 6 7 8 9 10 11 12 13 14

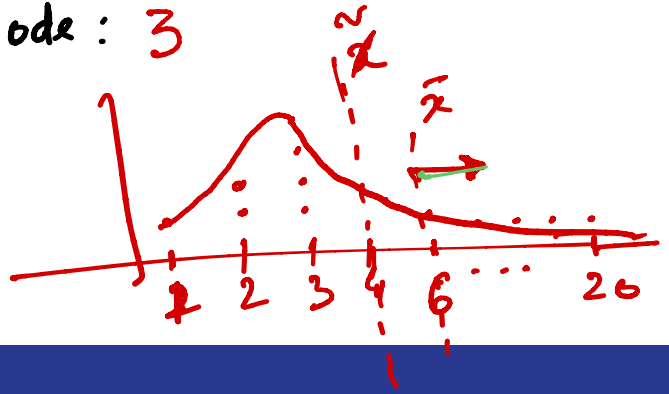
1, 2, 2, 3, 3, 3, 3, 5, 6, 7, 8, 11, 13, 20

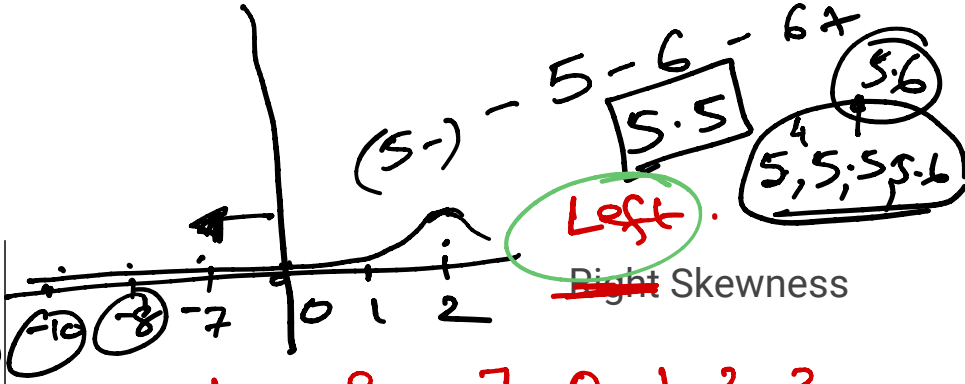
$\frac{70+20}{14}$

Mean: ≈ 6 $\bar{x}$

Median: 4 $\tilde{x}$

Mode: 3





~~Right~~ Skewness

-10 -8 -7 0 1 2 2

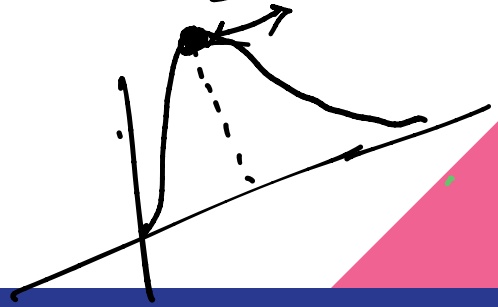
$\frac{-20}{7}$

Mean: $-20/7$

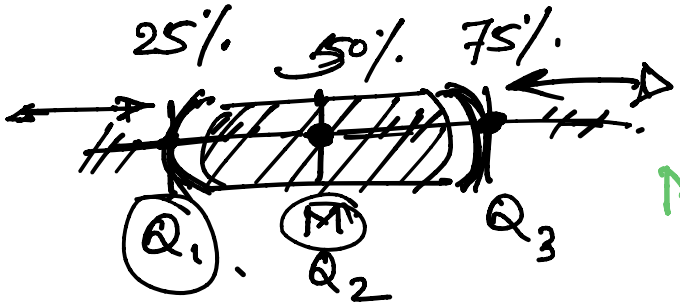
Median: 0

Mode: 2

$\bar{x} < \tilde{x}$



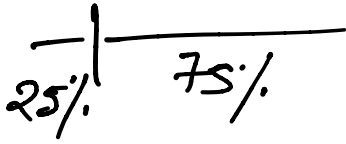
Exploratory Data Analysis



Other sample measures

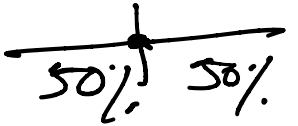
Quartiles: Divide the data into 4 equal parts

Lower quartile (Q_1 or P_{25}) splits the lowest 25% of the data from the other 75%



Middle quartile (Q_2 or P_{50}) splits the data in half (i.e. the median) Median

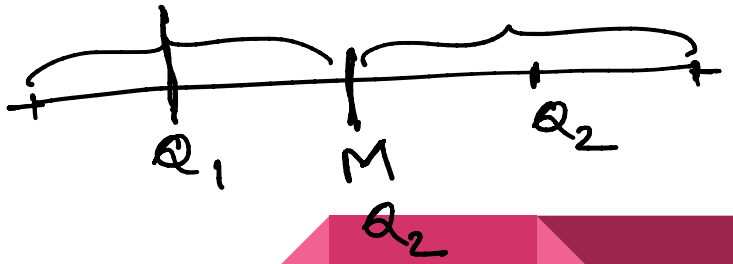
Upper quartile (Q_3 or P_{75}) splits the highest 25% of the data from the lowest 75%



Computation:

1) Use the median to divide the ordered data set into 2 halves

- **If n is odd, include the median in both halves.**
- **If n is even, split the data exactly in half.**



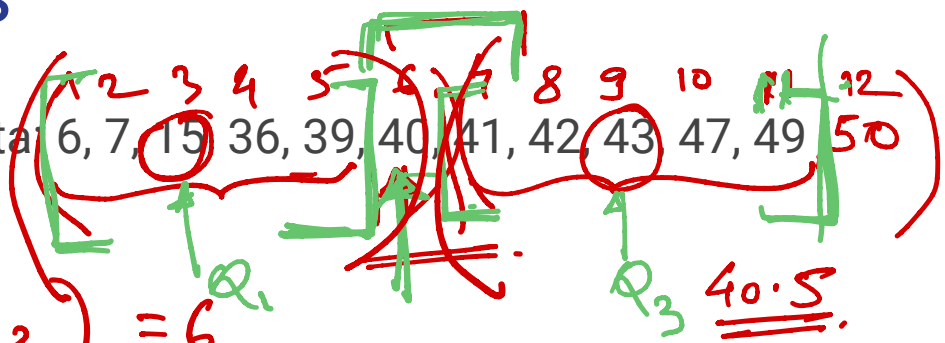
2) The lower quartile is the median of the lower half.

Q_1 : — 0 — 1

3) The upper quartile is the median of the upper half.

Exploratory Data Analysis

Example: Compute the quartiles of the data: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49, 50



$((11+1)/2) = 6$

Q_2 / Median: 40

Q_1: 15

Q_3: 43

n is odd, median is a single value, Q_1, Q_2

n is even, median is an avg

Exploratory Data Analysis

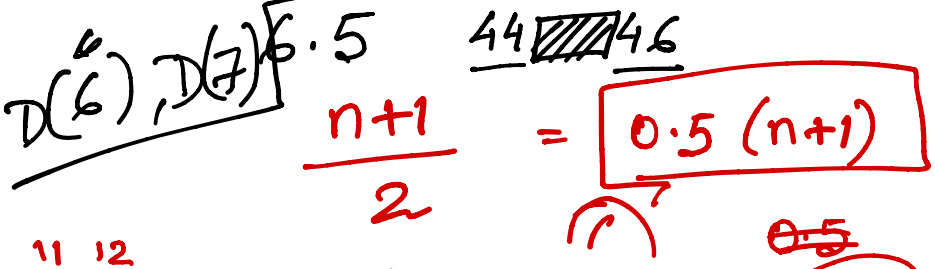
Quartiles: Divide the data into 4 equal parts

- **Lower quartile** (Q_1 or P_{25}) splits the lowest 25% of the data from the other 75%
- **Middle quartile** (Q_2 or P_{50}) splits the data in half (i.e. the median)
- **Upper quartile** (Q_3 or P_{75}) splits the highest 25% of the data from the lowest 75%

Percentiles:

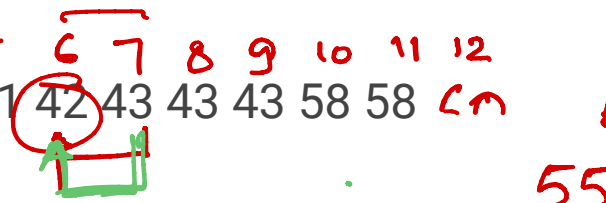
- Generalization of quartiles.
- Q_1 is the 25th percentile; P_{25}
- Can also calculate general percentiles: e.g. the 16th percentile, P_{16}, splits off the lower 16% of the data.

Quantiles Generalization



Example: 41 41 41 41 41 42 43 43 43 58 58 (n)

1 2 3 4 5 6 7 8 9 10 11 12



$\tilde{x} = D\left(\frac{11+1}{2}\right)$

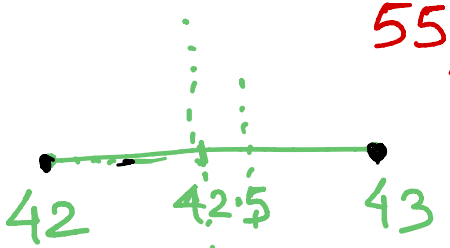
55%

0.55 = 6 0.5

= 50%

0.55 * 12

= [6.6] = 6



(42, 43)

42 42.5 43

$idx = \left\lfloor \frac{p}{100} \cdot (n+1) \right\rfloor$ | 0.6 * (43 − 42) = 0.6

42 + 0.6 = 42.6

$\alpha = \frac{p}{100}(n+1) - idx = 0.6$

42,50 6.6

46

$[D(idx) + \alpha \cdot (D(idx+1) - D(idx))]$

Exploratory Data Analysis

$D_1: 1, 2, 3$

$D_2: -1, 2, 5$

$\bar{x}_1 = \bar{x}_2 = 2$

Other ways to analyze data other than central tendency?

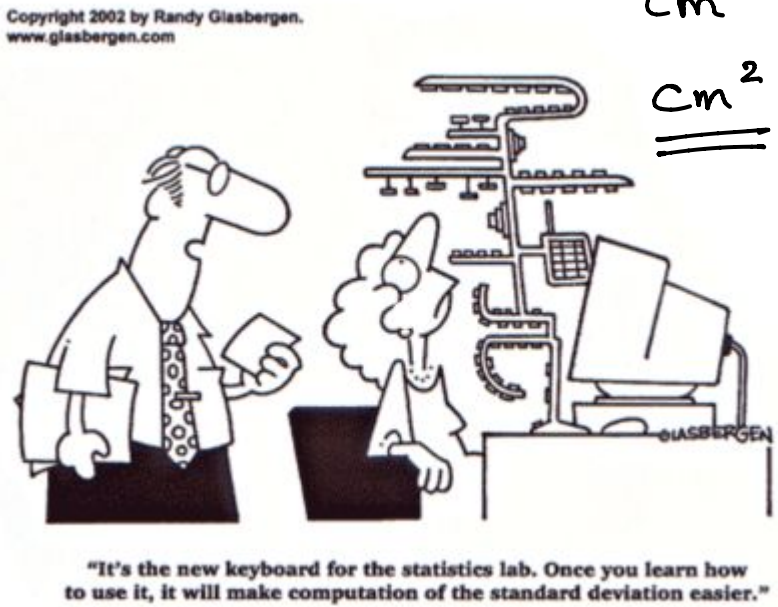
How about measuring the spread or **variability** of the data.

The **range** of a sample is the difference between the maximum and minimum values.

$R_1 = 3 - 1 = 2$

$R_2 = 5 - (-1) = 6$

Exploratory Data Analysis



"It's the new keyboard for the statistics lab. Once you learn how to use it, it will make computation of the standard deviation easier."

The **sample variance** denoted by s^2 is given by

$$s^2 = \frac{1}{(n-1)} \sum_{k=1}^{n} (x_k - \bar{x})^2$$

The **sample standard deviation**, denoted by s, is given by the square root of the variance.

$$s = \sqrt{s^2}$$

Exploratory Data Analysis - Variability

Example: Compute the standard deviation (SD) of the data: 2, 4, 3, 5, 6, 4

$$\bar{x} = \frac{2+4+3+5+6+4}{6} = \frac{24}{6} = 4.$$

$$\sigma^2 = \frac{1}{6-1}\left((2-4)^2 + (4-4)^2 + (3-4)^2 + (5-4)^2 + (6-4)^2 + (4-4)^2\right)$$

$$= \frac{1}{5}(4+0+1+1+4+0) = \frac{10}{5} = 2$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2}.$$

Exploratory Data Analysis – Interquartile Range

The interquartile range is defined to be the difference between the upper and lower quartiles:

$$IQR = Q_3 - Q_1$$

Assumption : Excluding the median from the computation of Q_1, Q_3 if n is odd

- IQR gives the spread of 50% of the data

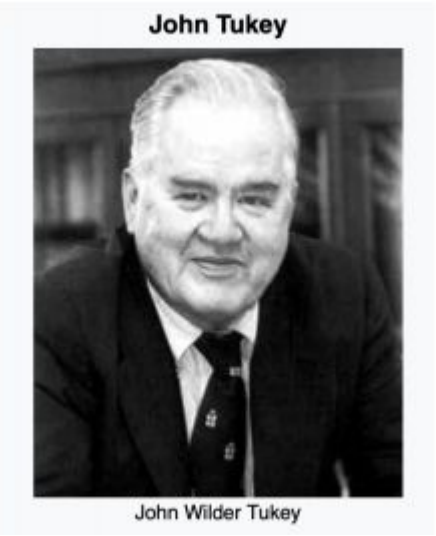
Example: Compute the IQR of the data: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

Q_1 Q_2 Q_3

Q_1 : 15
Q_2 : 40
Q_3 : 43

$IQR = Q_3 - Q_1 = 43 - 15$
$= \boxed{28}$

Exploratory Data Analysis – 5-Number Summary



John Wilder Tukey

Born	June 16, 1915 New Bedford, Massachusetts, U.S.
Died	July 26, 2000 (aged 85) New Brunswick, New Jersey, U.S.

Tukey advocated summarizing data sets with 5 values:

1) Minimum value
2) Lower Quartile
3) Median
4) Upper Quartile
5) Maximum value

Exploratory Data Analysis – 5-Number Summary

Example: Find the 5-number summary of the data: 6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49

Min ?

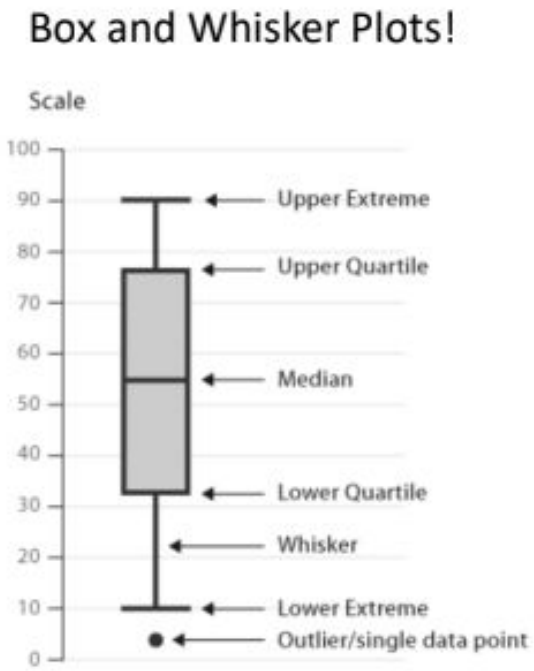
Q_1 ?

Q_2 ?

Q_3 ?

Max ?

Next Time:


Box and Whisker Plots!
Scale
100
90
80
70
60
50
40
30
20
10
0
Upper Extreme
Upper Quartile
Median
Lower Quartile
Whisker
Lower Extreme
Outlier/single data point