CSCI 3022: Intro to Data Science

Lecture 2: Exploratory Data Analysis & Summary Statistics

Rachel Cox

Department of Computer Science

- quizlet 00 Due on Friday at 11:59pm
- office hours starting Friday

Course Logistics – Before Next Class

- ❑ Read the Syllabus. You are responsible for knowing the information contained in this document.
- ❑ quizlet 00
- ❑ Make sure you can access the Canvas page.
- ❑ Check out the Piazza page.
- ❑ Install Anaconda (or other reliable Jupyter notebook method). Python 3.
- ❑ Review and complete Numpy/Pandas tutorial
- ❑ Explore nb00 and nb01

What will we learn today?

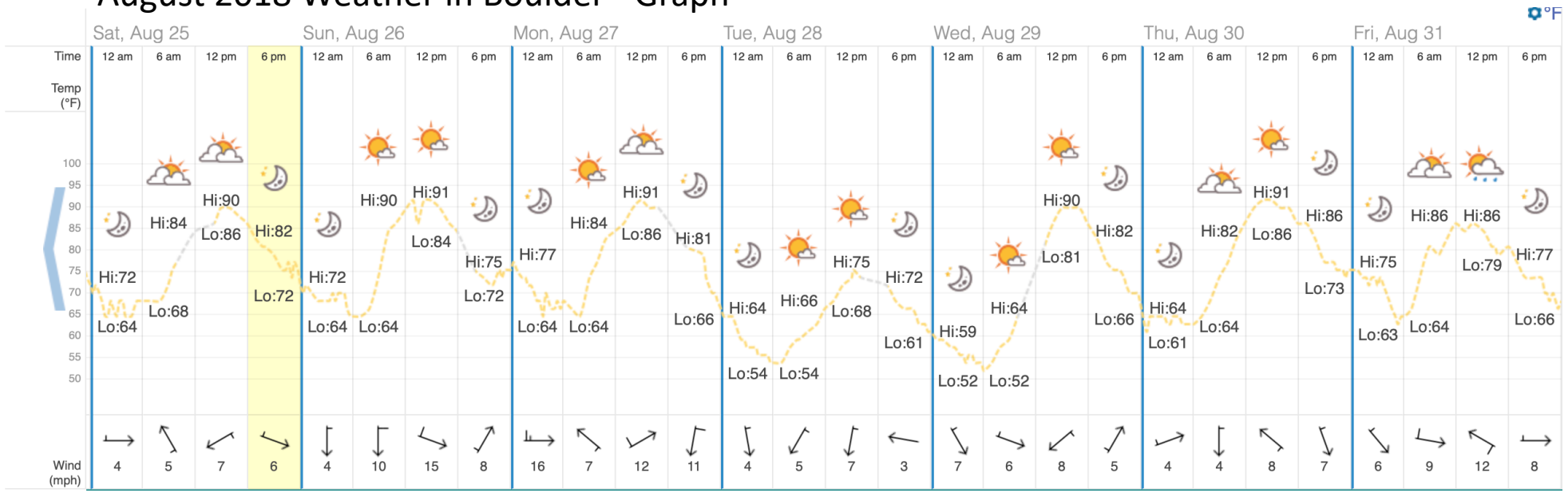
- ❑ Mean
- ❑ Median
- ❑ Mode
- ❑ Quartiles
- ❑ Sample Variance
- ❑ Sample Standard Deviation
- ❑ Interquartile Range (IQR)
- ❑ 5-Number Summary

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

Exploratory Data Analysis

How do you summarize and describe a set of data?

August 2018 Weather in Boulder - Graph



High & Low Weather Summary for August 2018

	Temperature	Humidity	Pressure
High	102 °F (Aug 5, 5:47 pm)	100% (Aug 14, 6:47 am)	30.35 "Hg (Aug 14, 6:47 am)
Low	52 °F (Aug 18, 10:52 pm)	9% (Aug 27, 3:45 pm)	29.84 "Hg (Aug 27, 3:45 pm)
Average	69 °F	46%	30.16 "Hg

* Reported Aug 1 12:51 am — Aug 31 11:53 pm, Boulder. Weather by CustomWeather, © 2019

* Charts taken from TimeandDate.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 6 PM and 8 PM.

- What is the **population**? Boulder residents
  sample frame
  Boulder residents that have a phone number
- What is the **sample**?
  every 50th person that actually picks up

What is the **Variable of Interest**?

household income

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.

Middle school with 6th, 7th, & 8th graders

	num students	sample size
6	100	10
7	60	6
8	70	7

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

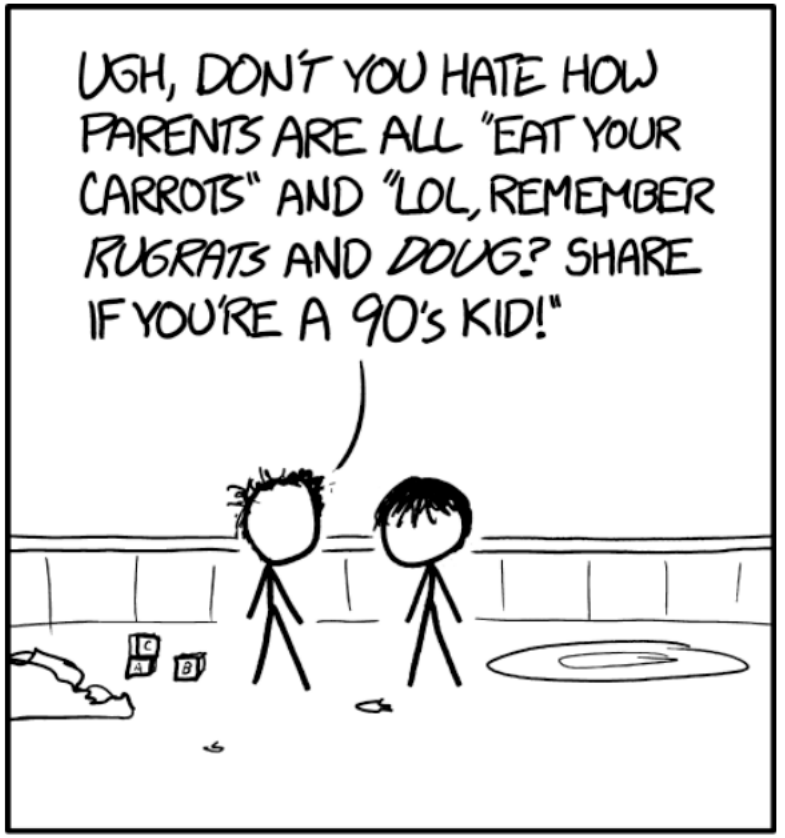
Summarizing the "center" of the sample is a popular way to characterize data.

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

$$\frac{17 + 10 + 10 + 18 + 5 + 23}{6}$$

$$= 13.8\overline{3}$$

Exploratory Data Analysis – Central Tendency

$\bar{x}$ \bar{x}

$\tilde{x}$ \tilde{x}

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 $\left(\frac{n}{2}+1\right)^{\text{th}}$ ordered values.

Example: Compute the median of the following data set.

17 10 10 18 5 23

5 10 10 17 18 23

average to find median

$\tilde{x} = \frac{10+17}{2} = 13.5$

Exploratory Data Analysis – Central Tendency

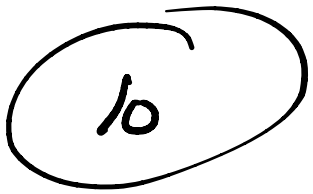
Sample Mode: Most commonly occurring value in a data set.

Calculation:

Count what number/data point occurs most often.

Example: Compute the mode of the following data set.

17 10 10 18 5 23



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

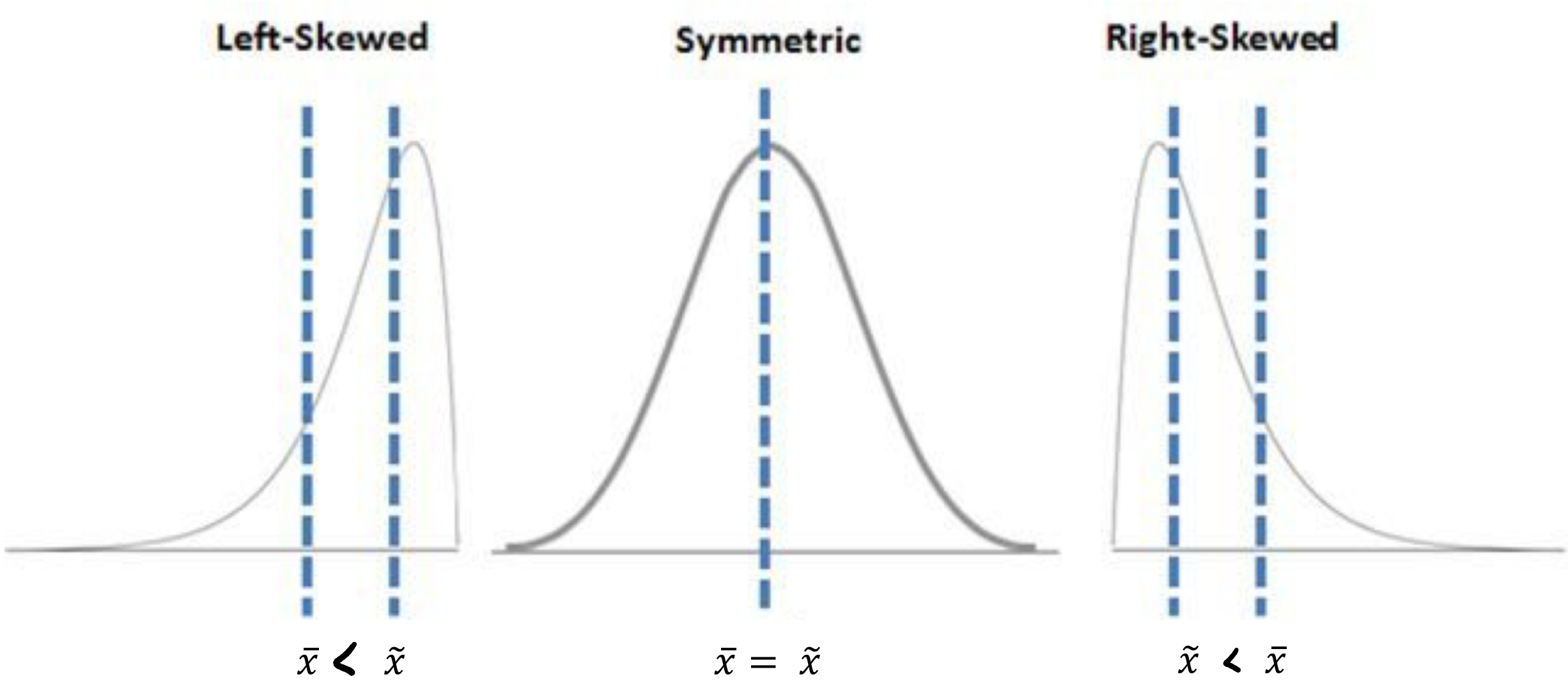
Disadvantages: Very sensitive to outliers

Sample Median

Advantages: Not as easily affected by a few outliers

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

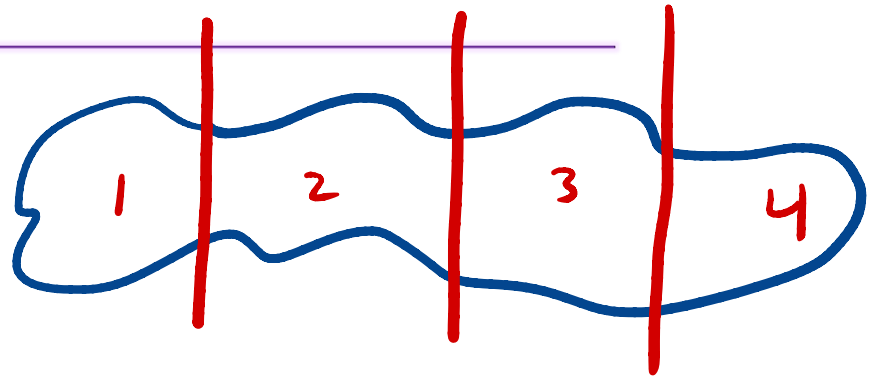
Exploratory Data Analysis – Central Tendency



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

Exploratory Data Analysis

Other sample measures



Quartiles: Divide the data into 4 equal parts

four

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

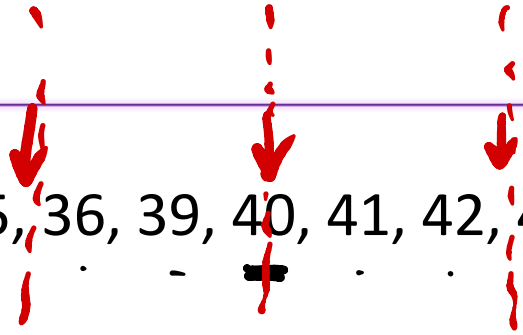
Middle quartile (Q_2 or P_{50}) splits the data in half (i.e. the median)

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

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.**
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

Q_2 (aka P_{50} aka median) = 40

use the median to split our list into two halves

lower half:

6, 7, 15, 36, 39, 40

- Q_1 is median of lower half

$$Q_1 = \frac{15+36}{2}$$

$$Q_1 = 25.5$$

upper half:

40, 41, 42, 43, 47, 49

- Q_3 is median of upper half

$$Q_3 = \frac{42+43}{2}$$

$$Q_3 = 42.5$$

quartiles

$Q_1 = 25.5$

$Q_2 = 40$

$Q_3 = 42.5$

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.

Exploratory Data Analysis – Variability

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.

Exploratory Data Analysis – Variability

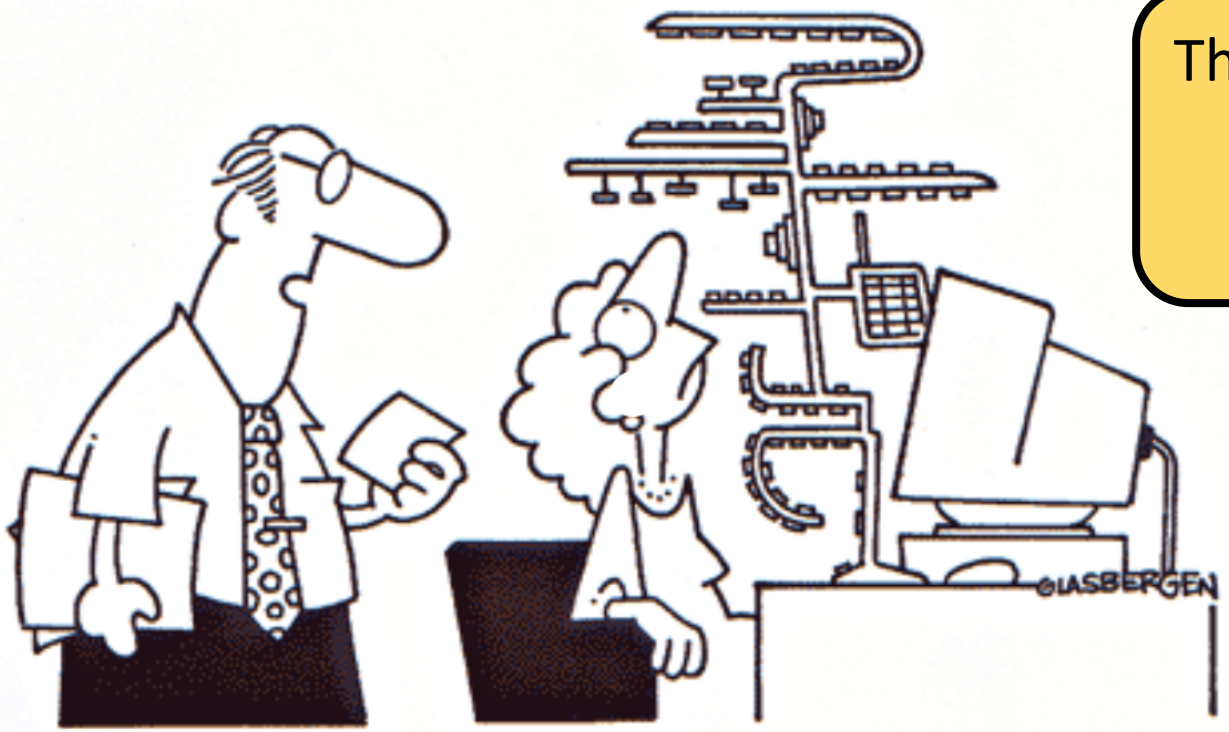
We will prove this later on.

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.

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



"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."

$x_1, x_2, x_3, \ldots, x_{10}$

mean $\bar{x}$

$x_1 - \bar{x}, x_2 - \bar{x}, \ldots x_{10} - \bar{x}$

Exploratory Data Analysis – Variability

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

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

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

$$= 2$$

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

Exploratory Data Analysis – Interquartile Range

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

$$IQR = Q_3 - Q_1$$

- 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

$Q_1 = 25.5$

$Q_2 = 40$

$Q_3 = 42.5$

$IQR = 42.5 - 25.5$

$= 17$

Exploratory Data Analysis – 5–Number Summary

John Tukey

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

6, 25.5, 40, 42.5, 49

Next Time:

Box and Whisker Plots!