Foundation of Data Science
Lecture 2, Module 1
Fall 2022

Rumi Chunara, PhD

Fine Print: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute without the instructor's permission.

Getting to Know the Data

- Statistical Descriptions of Data
- Measuring Data Similarity and Dissimilarity
- Summary

Descriptive vs. Inferential Statistics

Refresher!

- **Descriptive:**
  - Describes data you have but can't be generalized beyond that (e.g., median)
  - Scope: Exploratory Data Analysis
- **Inferential:**
  - Enables inference about a certain *population* beyond the data you have, based on *samples of this population* (e.g., t-test)
  - Adequate sampling techniques are key
  - Scope: Machine Learning and Prediction

Descriptive vs. Inferential Statistics

- **Descriptive:** **Focus today!**
  - Describes data you have but can't be generalized beyond that (e.g., median)
  - Scope: Exploratory Data Analysis
- **Inferential:**
  - Enables inference about a certain population beyond the data you have (e.g., t-test)
  - Scope: Machine Learning and Prediction

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Measures of central tendency
  - mean, median, mode
- Measures of dispersion (variation)
  - *max*, *min*, variance, standard deviation

Measuring Central Tendency

- Mean (algebraic measure) (sample vs. population):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

random sample population

Note: *n* is sample size and *N* is population size!

- Weighted arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Trimmed mean: chopping extreme values

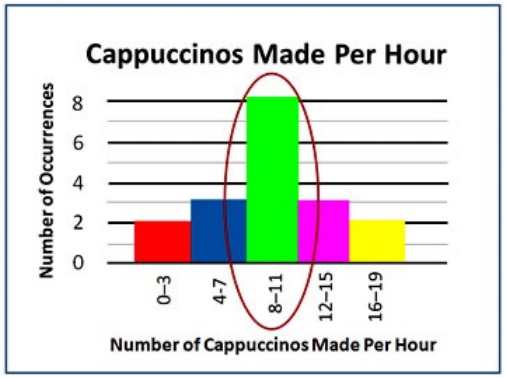
Measuring Central Tendency

- Median:
  - Middle value if odd number of values; average of the middle two values otherwise
  - **Estimated by interpolation (for *grouped data*)**

    Estimated Median class is halfway through groups

    Estimated Median = L + (((n/2) − B)/G) × w
    - **L** is the lower class boundary of the group containing the median
    - **n** is the total number of values
    - **B** is the cumulative frequency of the groups before the median group
    - **G** is the frequency of the median group
    - **w** is the group width

What is the estimated median here?



Measuring Central Tendency

- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal, multimodal
  - Empirical formula

$$mean - mode = 3 \times (mean - median)$$

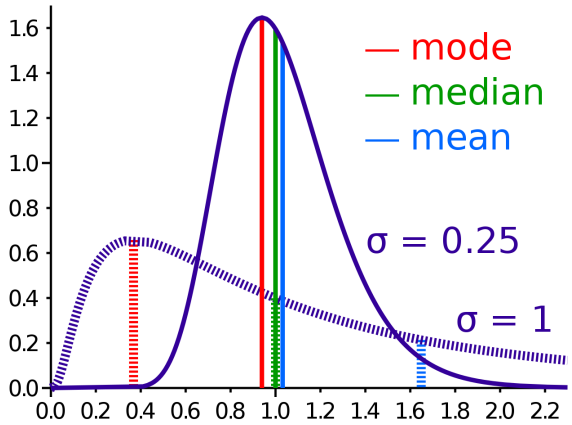
Approximation in the case of a normal distribution with a large sample size

Measuring Central Tendency

- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal, multimodal
  - Empirical formula

$$mean - mode = 3 \times (mean - median)$$

Approximation in the case of a normal distribution with a large sample size



How do mean and median compare?

- When should we use the mean? What are some of its drawbacks?
- And how about the median?

Summary Stats: Limitations

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

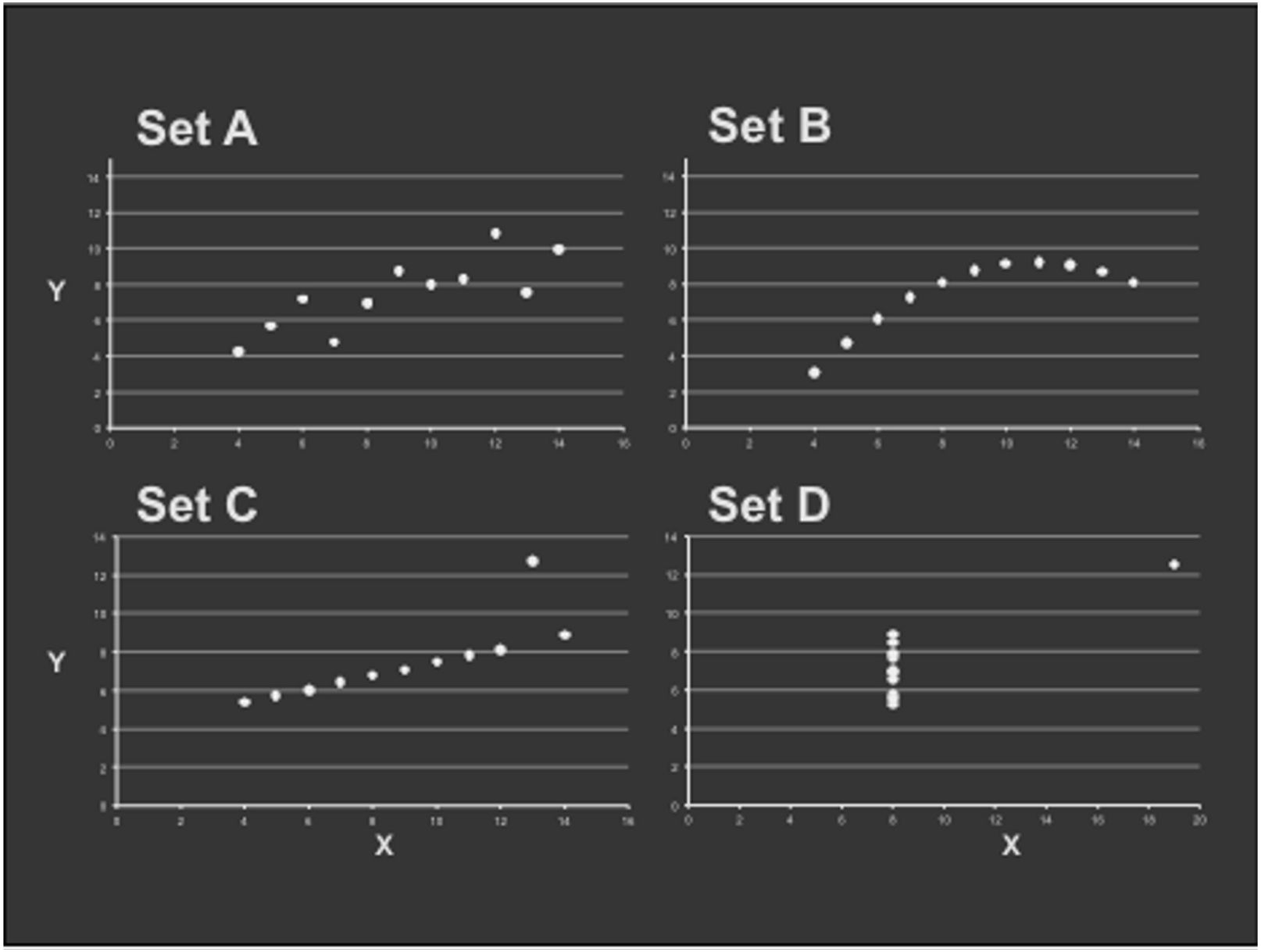
Summary Statistics Linear Regression

$u_X = 9.0$ $\sigma_X = 3.317$ $Y = 3 + 0.5\ X$

$u_Y = 7.5$ $\sigma_Y = 2.03$ $R^2 = 0.67$

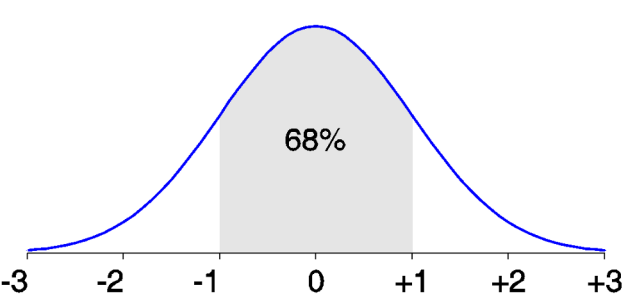
[Anscombe 73]

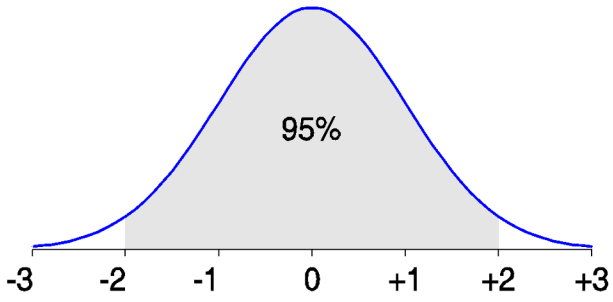
Looking at Data

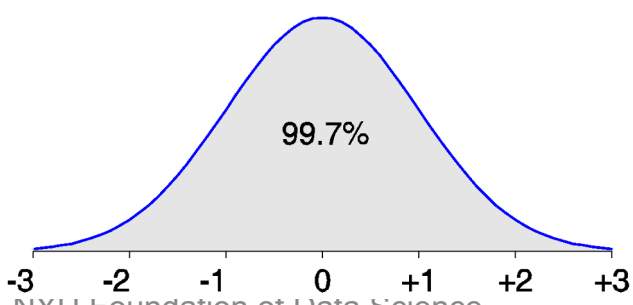


Properties of Normal Distribution Curve

- The normal (distribution) curve
- mean = median = mode
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it







Measuring the Dispersion of Data

- Variance and standard deviation (*sample vs population)*
  - **Variance** (algebraic, scalable computation)
    - How far a set of numbers is spread out from their mean value

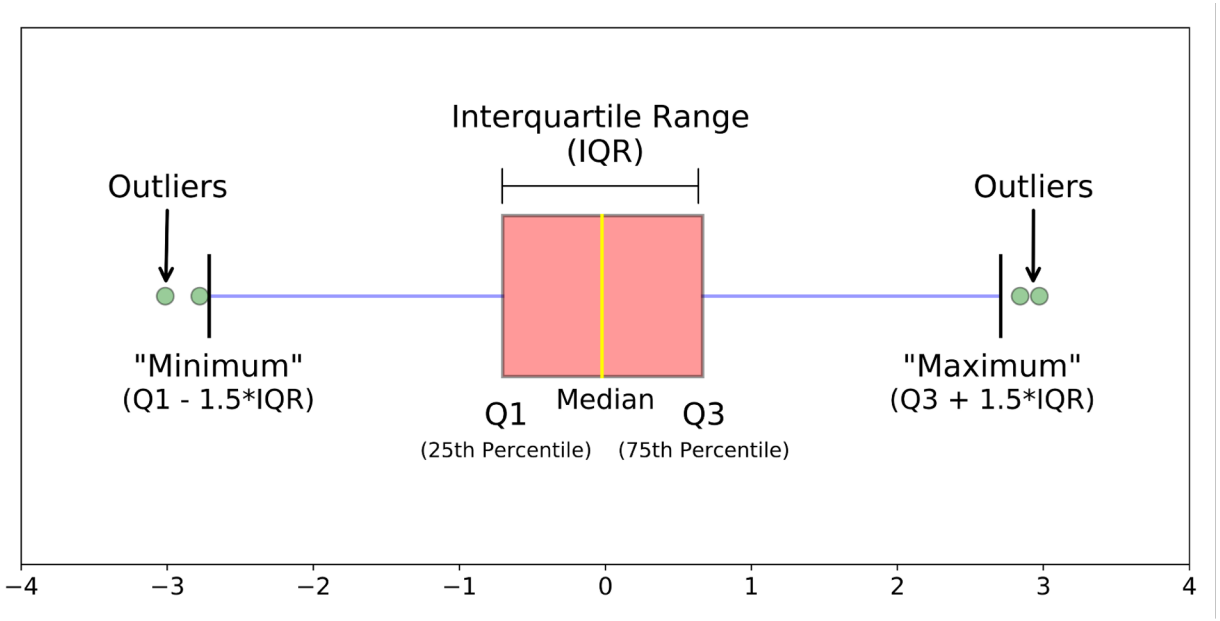
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$$

  - **Standard deviation** *s (or σ)* is the square root of variance
    - May give more clarity about the deviation of data from a mean

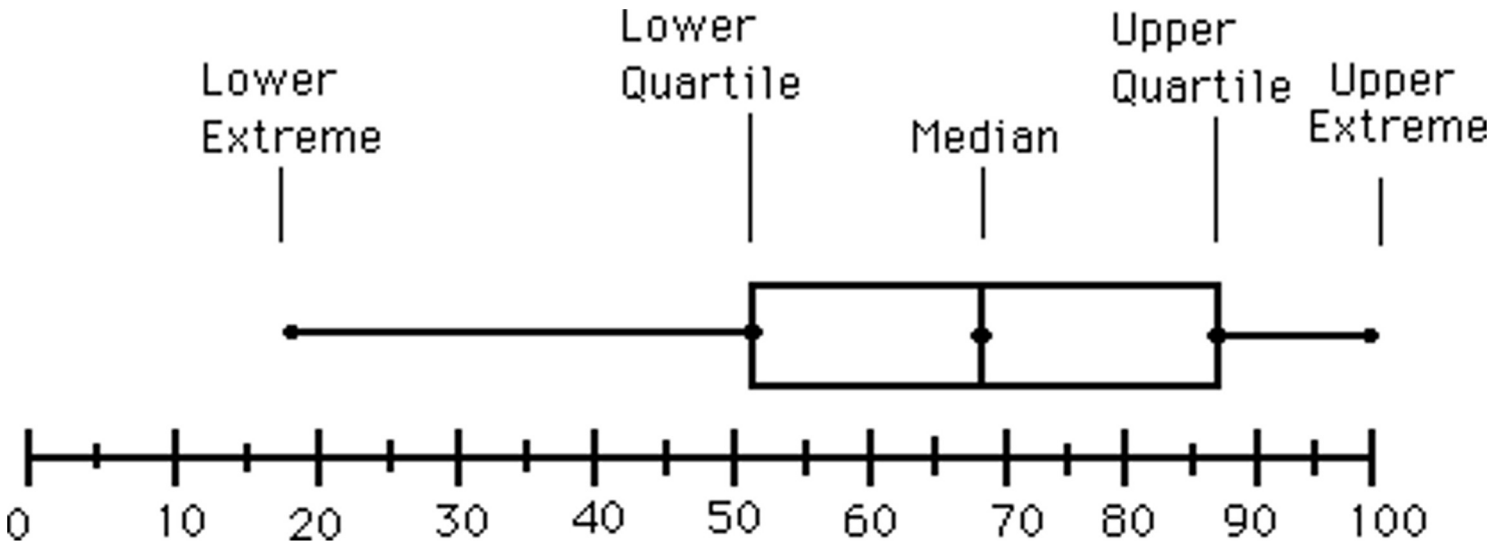
Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually



Is the data normally distributed?

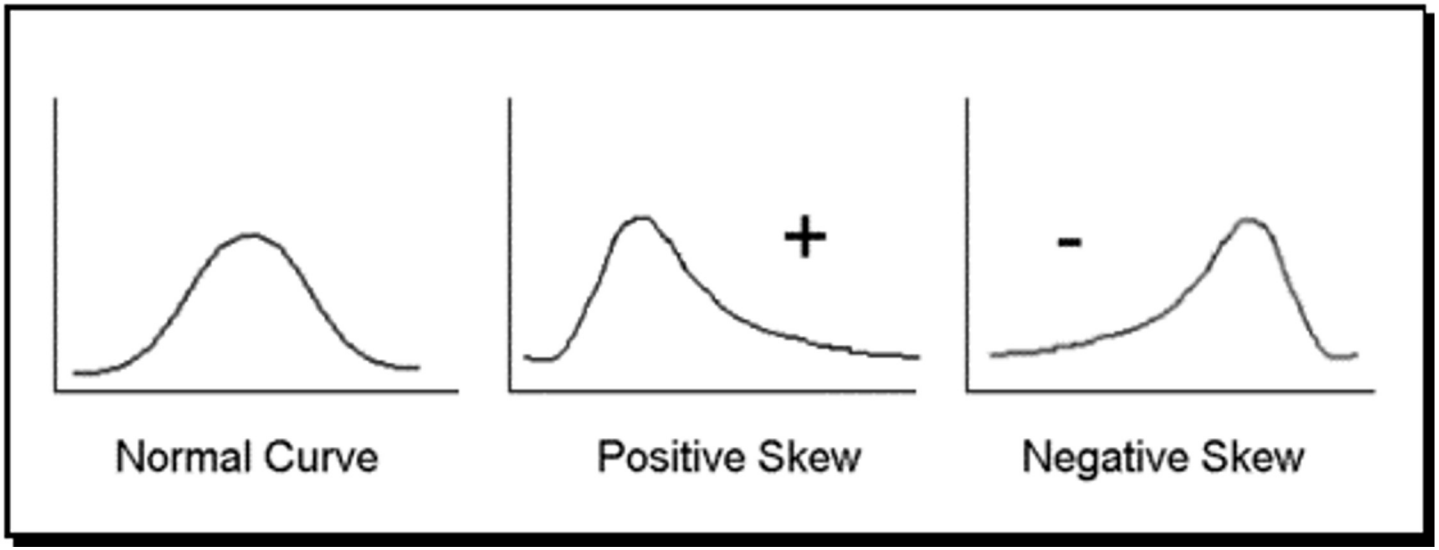
- Many statistical tools (mean, t-test, Pearson's correlation etc.) **assume data is normally distributed**.
  - Often not true!
  - Box-and-whisker plot is a good clue



Is the data normally distributed?

- Many statistical tools (mean, t-test, Pearson's correlation etc.) **assume data is normally distributed**.
  - Often not true!

- Whenever it is asymmetric, the data cannot be normal
  - The histogram gives even more information.



Distributions

Some other important distributions:

- **Poisson:** the distribution of counts that occur at a certain *rate.*
  - Number of visits to a website in a fixed time interval.
  - Number of website clicks in an hour.
- **Zipf/Pareto/Yule distributions:** govern the frequencies of different terms in a document, or website visits
  - Long-tails.
- **Binomial/Multinomial:** The number of counts of events out of n trials
  - Coin flips = 6, how many heads did I see?

You should understand the distribution of your data before applying any model!