Summary Statistics

#CMSC320 #M1 #probability_statistics

Purpose

- Summary statistics provide general overview of what data is telling us
- Summary statistics are useful for comparing datasets with similar distributions
  - Example: can use summary statistics to compare the average weight of New York City rats to Baltimore rats
- Summary statistics can be misleading

Types of Summary Statistics

- Central Tendency
- Spread
- Skew

Note: The kind of distribution determines what summary statistics are meaningful and which measures to use.

Central Tendency

> **Summary**
>
> Central tendency shows where most of the data is centered. There are three types of central tendency: **mean**, **median**, and **mode**.

Note on Distributions

Certain measures of central tendency are not meaningful for certain distributions:

- *Bimodal*
  - Mean, median not useful
  - Mode may be useful
- *Power law*
- *Uniform*
  - Mode not useful

Measures of Central Tendency

- **Mean**: average of all numbers in dataset
- **Median**: middle element
- **Mode**: most common element in dataset

	Mean	Median	Mode
Good	• Most cases, default measure	• Data has significant outliers • Ordinal data • Skewed distribution	• Multimodal distributions • Categorical data • Few outliers
Bad	• Outliers in dataset • Skewed distribution	• Highly Skewed dataset • Distribution is not normal	• Non-normal distribution

Depending on the properties of the dataset being evaluated, certain measures of central tendency are better suited for determining the center of the data.

Types of Mean

Arithmetic mean

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

- What people usually think about when they hear "average"
- **When to Use**
  - Good for distributions that have similar tails on either side (ex. normal distribution)

Geometric mean

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

- Cannot be used when datasets have 0 or negatives (in the case the data does, shift values with zero or negative to apply geometric mean)
- **Properties**
  - Less vulnerable to outliers than arithmetic mean
  - Large outliers exacerbated by geometric mean
- **Sample Application**
  - Length of stay in a hospital (most patients are there for a short period of time, some patients are there for orders of magnitude longer)

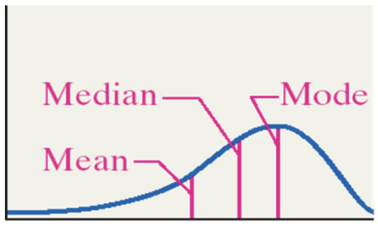
Harmonic mean

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n} \right)^{-1}.$$

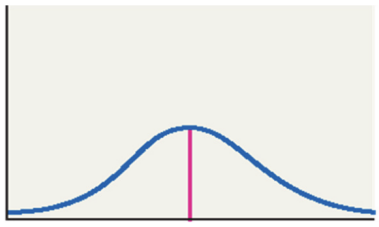
- Rarely used
- **When to Use**
  - Used in situations involving *rates and ratios*
- **Sample Application**
  - For instance, if a vehicle travels a certain distance d outbound at a speed x (e.g. 60 km/h) and returns the same distance at a speed y (e.g. 20 km/h), then its average speed is the harmonic mean of x and y (30 km/h) – not the arithmetic mean (40 km/h). The total travel time is the same as if it had traveled the whole distance at that average speed.

Mean and Median as Indicators

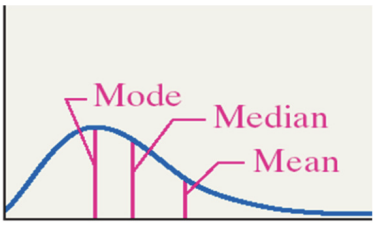
- When median and mean are different the distribution may have a skew or have many of outliers



(a) Skewed Left
Mean < Median



(b) Symmetric
Mean = Median



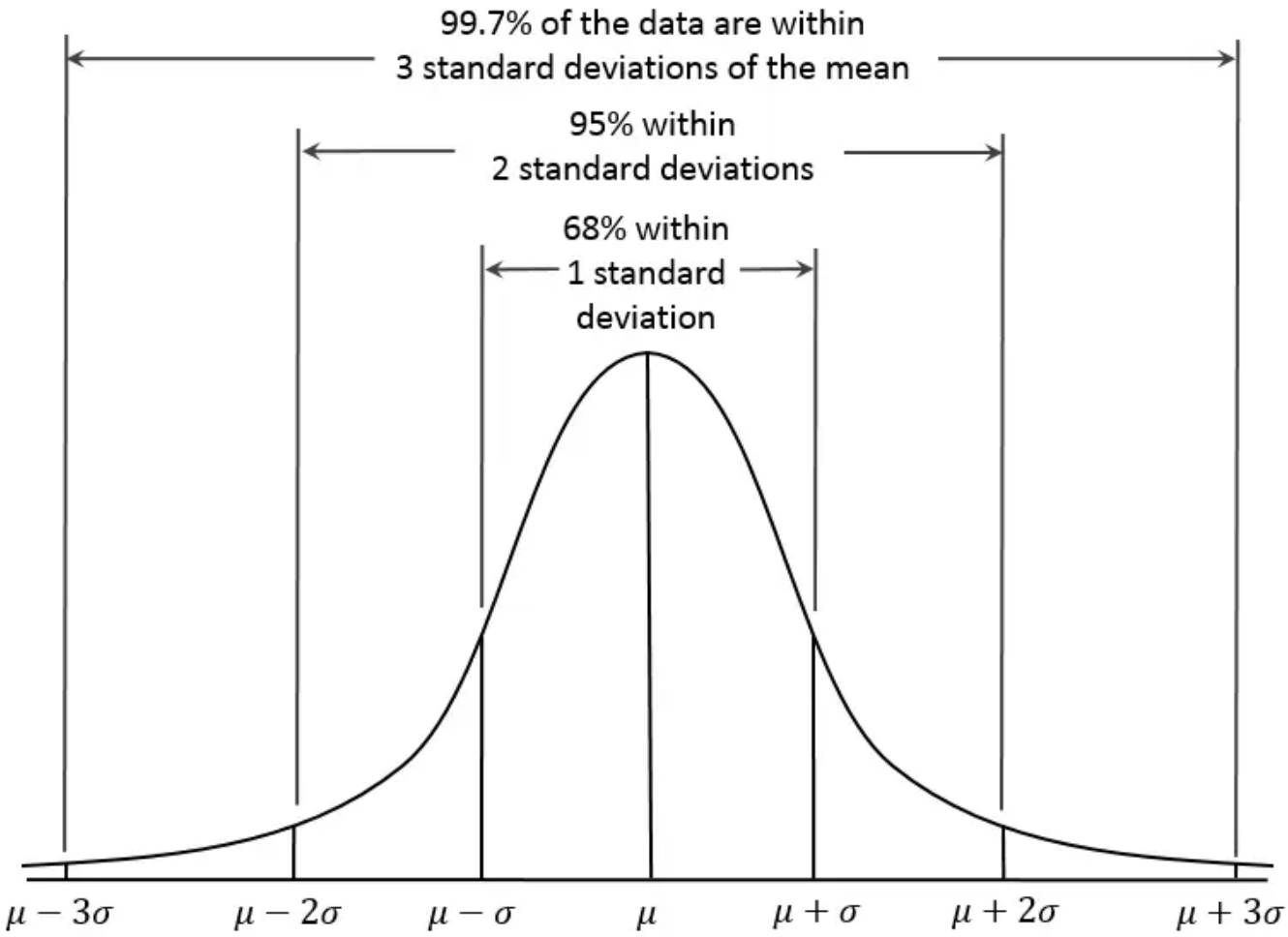
(c) Skewed Right
Mean > Median

Spread

Summary

Spread measures how spread out a distribution is. The most common measures of spread is **standard deviation** and **variance**.

Variance / Standard deviation



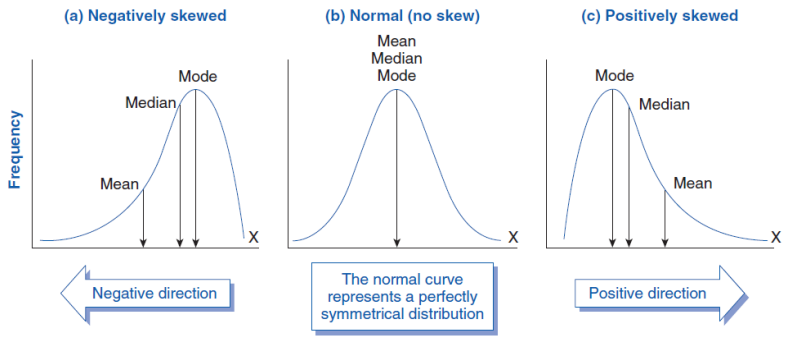
$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

- Standard deviation measures the average amount by which all the data points in a dataset deviate from the mean
- Large standard deviation means dataset is more spread out
- Standard deviation is in same units as actual data set
- Standard deviation and variance less useful with highly skewed distributions
- Standard deviation used when determining the confidence that a given data point came from the distribution in question

Skew

Summary

Skew measures how shifted a distribution is, or how many outliers it has.



Kurtosis

- A measure of how likely a distribution is to produce outliers

Skewness

- For a unimodal distribution,
  - **Negative skew**: indicates that the tail is on the left side of the distribution
  - **Positive skew**: indicates that the tail is on the right side of the distribution
- In cases where one tail is long but the other tail is fat, skewness does not obey a simple rule

Normalcy

- How close to the normal distribution a given dataset is

Miscellaneous

Min / max

- Often as a data scientist it's useful to look at a collection of the data points at the edges
  - Example: Determining the top and bottom 1%. This can help tell us if there is a fat tail, one or a few wild outliers, or some other pattern.

Quartiles

- Useful for seeing how tightly grouped the data is

Note: Quintiles used as well, though not as common

Outlier detection

- Z score:

$$Z = \frac{x - \mu}{\sigma}$$

- Z-score is useful when comparing individuals in different metrics
  - *Examples*
    - Are exceptionally tall women also exceptionally strong?
    - Are lakes with average amounts of pollutants also of average size?