Introduction to Statistics

Agenda

1. Overview of Statistics
2. Statistical Thinking
3. Why Statistics is important for Data Science
4. Types of Statistics
5. Descriptive Statistics

Statistics - Overview

What is Statistics?

Statistics is the branch of mathematics dealing with the **collection, analysis, interpretation, and presentation of numerical data.** Statistics and data override intuition, inform decisions, and minimize risk and uncertainty.

In more common usage, statistics refers to numerical facts. The numbers that represent the income of a family, the age of a student, the percentage of passes completed by the quarterback of a football team, and the starting salary of a typical college graduate are examples of statistics in this sense of the word.

Educated guesses vs Pure guesses

Statistical methods help us make scientific and intelligent decisions in uncertain situations. Decisions made by using statistical methods can be called **educated guesses.** Decisions made without using statistical (or scientific) methods would be considered **pure guesses** and, hence, may prove to be unreliable.

Statistical Thinking

Let us try to answer a sample question: **Do first babies arrive late?**

Some people claim it's true, others say it's a myth, and some people say it's the other way around: first babies come early.

"My two friends that have given birth recently to their first babies, they BOTH went almost 2 weeks overdue before going into labour or being induced."

"I don't think that can be true because my sister was my mother's first and she was early, as with many of my cousins."

Reports like these are considered anecdotal evidence because they are based on data that is unpublished and usually personal.

Problems with this:
Small number of observations, selection bias, confirmation bias and inaccuracy

Statistical Thinking - The Approach

Data Collection	Collect data with the goal of generating statistically valid inferences about the population
Describe your Data	Generate statistics that summarize the data concisely, and evaluate different ways to visualize data
Explore your Data	Look for patterns, differences, and other features that address the question. Check for inconsistencies and limitations
Test your Hypothesis	Where we see apparent effects, we will evaluate whether the effect is real, or whether it might have happened by chance
Estimation	We will use data from a sample to estimate characteristics of the general population

Why is Statistics so important to Data Science?

- A popular quote says "***A data scientist is a person who is better at statistics than any programmer and better at programming than any statistician***." In other words, statistics is an inherently necessary component of Data Science.

- In Data Science, **Statistics is at the core of all sophisticated Machine Learning algorithms** - capturing and translating data patterns into actionable evidence. Data Scientists use statistics to gather, review, analyze, and draw conclusions from data, as well as apply quantified mathematical models to appropriate variables

- It is critical for Data Scientists to learn statistics because **it connects data to the questions businesses are asking across all disciplines.** Questions such as:
  - How can we create efficiencies?
  - How can we limit spending and increase revenue?
  - How can we maximize communication with our target audience?

Types of Statistics: Descriptive and Inferential Statistics

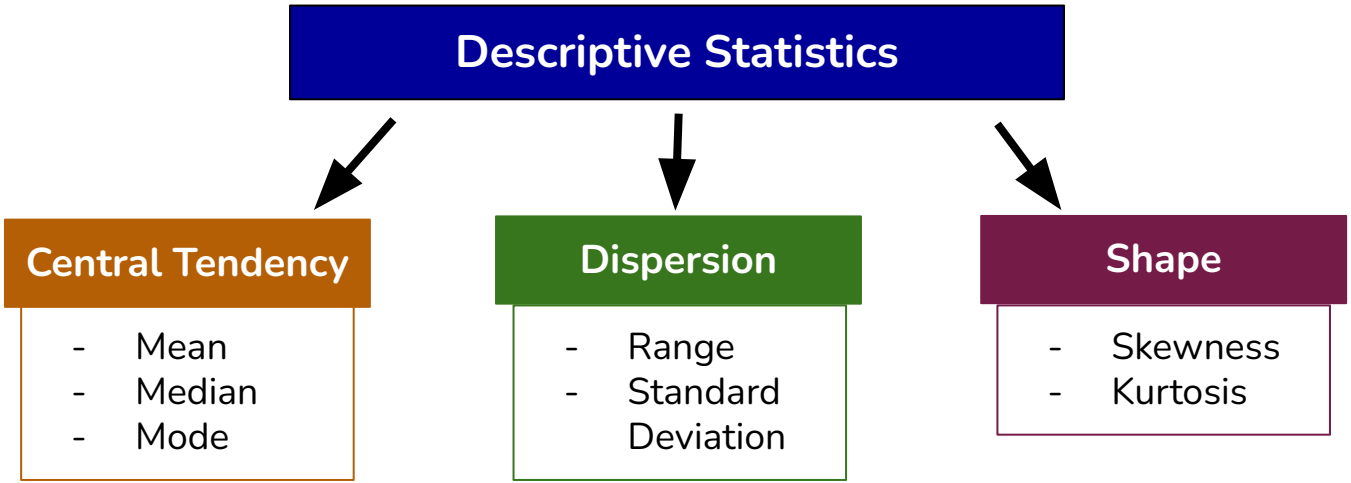
Descriptive Statistics - Overview

- Descriptive Statistics involves describing, summarizing and organizing the data so it can be easily understood.
- **Historical Note:** Statistics is considered to have originated during census activities carried out by the Babylonians and the Egyptians (4500 -3000BC). They used to maintain a record of the number of livestock each person owned and the crops each citizen harvested yearly.
- Descriptive Statistics is useful because it allows you to make sense of the data, and it helps in exploring and making conclusions about the data in order to make rational decisions.
- It also includes calculating things such as the average of the data, its spread and the shape its distribution produces.

How to describe your data?

The following measures are used to describe a data set:

- **Measures of position** (also referred to as **central tendency** or location measures).
- **Measures of spread** (also referred to as **variability** or **dispersion** measures).
- Measures of shape



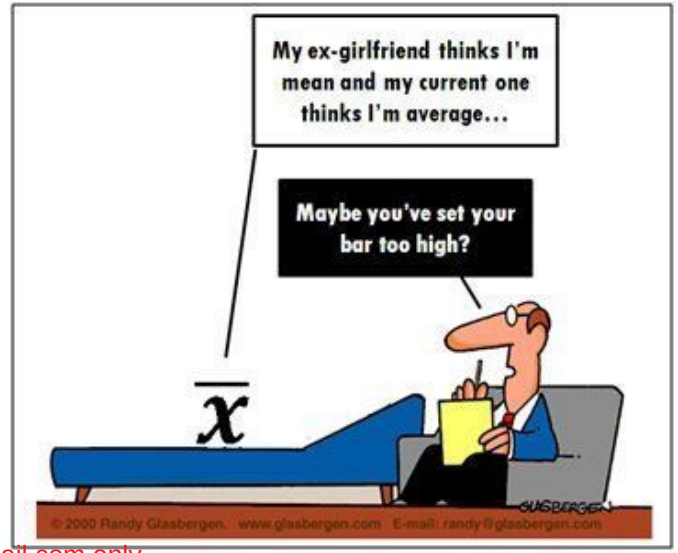
Descriptive Statistics - Measures of Central Tendency

Descriptive Statistics - Mean

- To calculate the arithmetic mean of a set of data we must first add up (sum) all of the data values (x) and then divide the result by the number of values (n). Since we obtain the following formula for the mean ($\bar{x}$).

$$\bar{x} = \frac{\sum x}{n}$$

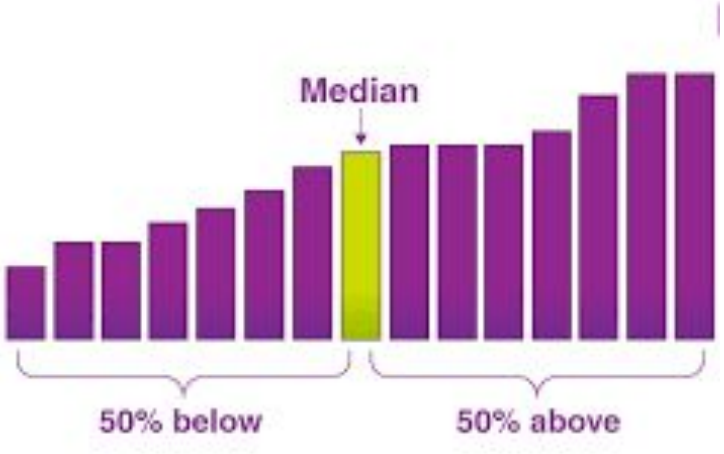
- Why do we need mean?
  - The mean gives us an idea of where the “center” of a dataset is located
  - The mean is also important because it carries a piece of information from *every* observation in a dataset.
  - Going forward you will notice how mean is used in Machine learning to minimize the prediction errors



Descriptive Statistics - Median

- Median is the middle number. It is found by putting the numbers in order and taking the actual middle number if there is one, or the average of the two middle numbers if not. Median is robust and is less affected by the outliers.

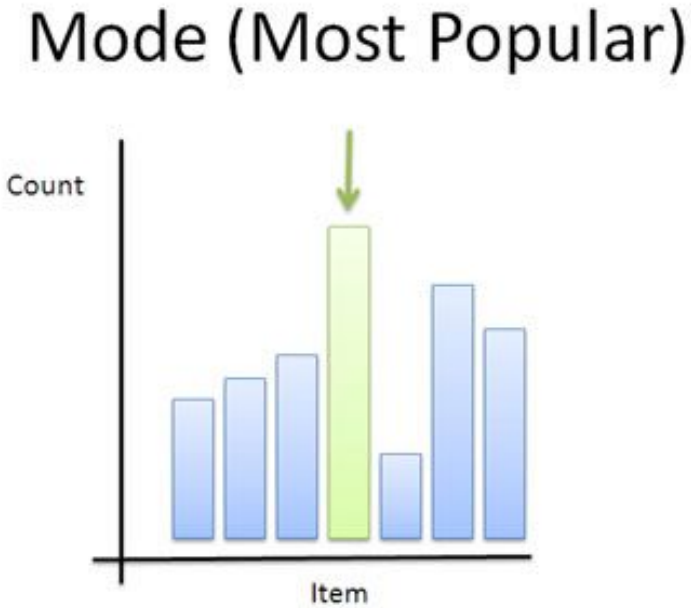
- Why do we need median?
  - The median also represents the 50th percentile of a dataset. That is, exactly half of the values in the dataset are larger than the median and half of the values are lower.
  - the median is a more useful metric than mean in the following circumstances:
    - When the distribution is skewed.
    - When the distribution contains outliers



Descriptive Statistics - Mode

- The value that occurs the most often in a data set. The mode is best used when you want to indicate the most common response or item in a data set.

- Why do we need mode?
  - *It can be used when the data is not numerical.*
  - It is more useful to distinguish between unimodal and multimodal distributions
    - When data has more than one peak



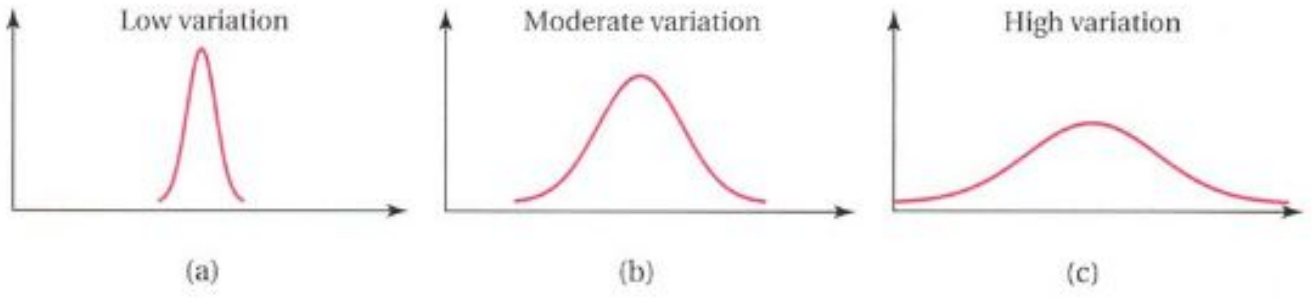
Descriptive Statistics - Measures of Spread

Measures of Spread

Measures of spread describe how similar or varied the set of observed values are for a particular variable. It gives an indication of the amount of variation in the process. It is an important indicator of quality of the obtained data.

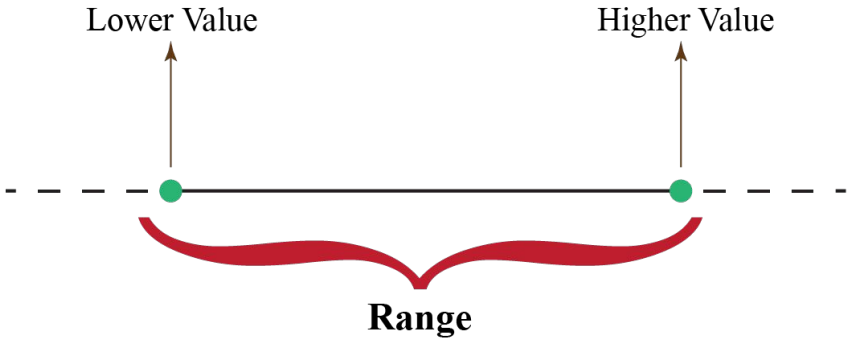
There are different statistics by which we can describe the spread of a data set:

- Range
- Standard deviation
- Variance



Measures of Spread - Range

- The difference between the highest and the lowest values.
- It is the simplest measure of variability. Range is often denoted by ‘R’. It does not make full use of the available data but still gives us a fair idea of spread or dispersion of the data.



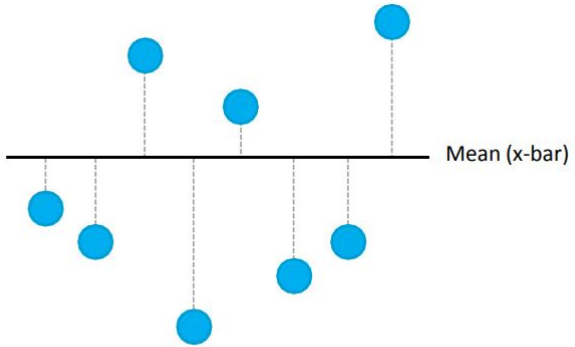
- It can be misleading when the data is skewed or in the presence of outliers.
- Just one outlier will increase the range dramatically.

Measures of Spread - Standard Deviation

- The average distance of the data points from their own mean. A low standard deviation indicates that the data points are clustered around the mean. A large standard deviation indicates that they are widely scattered around the mean.It is however a more robust measure of variability.

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

s = standard deviation
$\bar{x}$ = mean
x = values of the data set
n = size of the data set

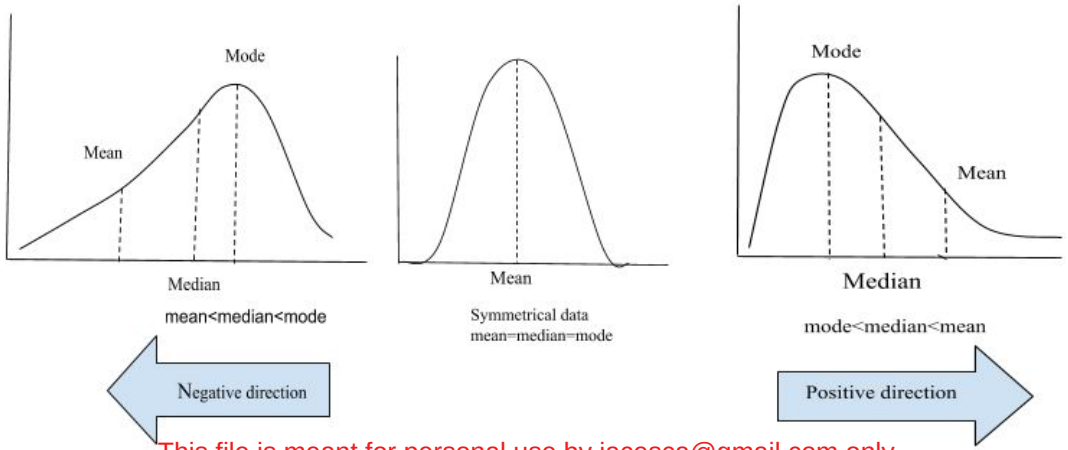


- What would be the standard deviation for the grades of 6 students who scored 40,40,40,40,40,40.

Descriptive Statistics - Measures of Shape

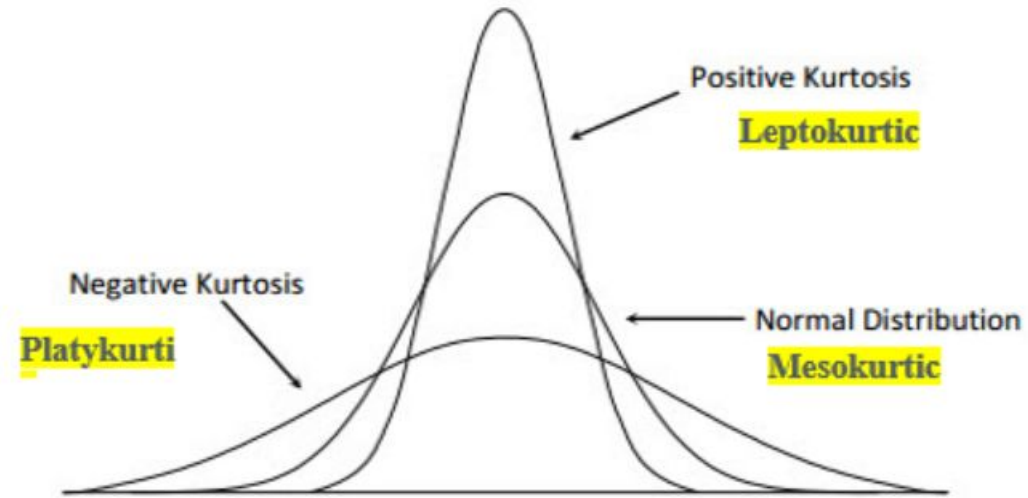
Measures of Shape - Skewness

- Skewness describes whether the data is distributed symmetrically around the mean.
- A skewness value of zero indicates perfect symmetry. Which means the mean= median =mode
- A negative value implies left-skewed data.
- A positive value implies right-skewed data.



Measures of Shape - Kurtosis

- Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a Symmetrical distribution. In other words, kurtosis identifies whether the tails of a given distribution contain extreme values.



Appendix

Statistics vs Machine Learning

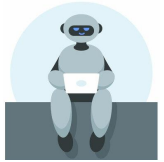
Statistical Modeling

Statistical models are subset of mathematics that are designed to make inference about the relationships between variables

Statistician: "The model is 85% accurate in predicting Y, given a, b and c; and I am 90% certain that you will obtain the same result."

statisticians must understand how the data was collected, statistical properties of the estimator (p-value, unbiased estimators), the underlying distribution of the population

Machine Learning



Machine learning models are subset of AI and CS which are designed to make the most accurate predictions possible.

ML : Professional :The model is 85% accurate in predicting Y, given a, b and c.

No prior assumptions about the underlying relationships between the variables.Machine learning treats an algorithm like a black box, as long it works.