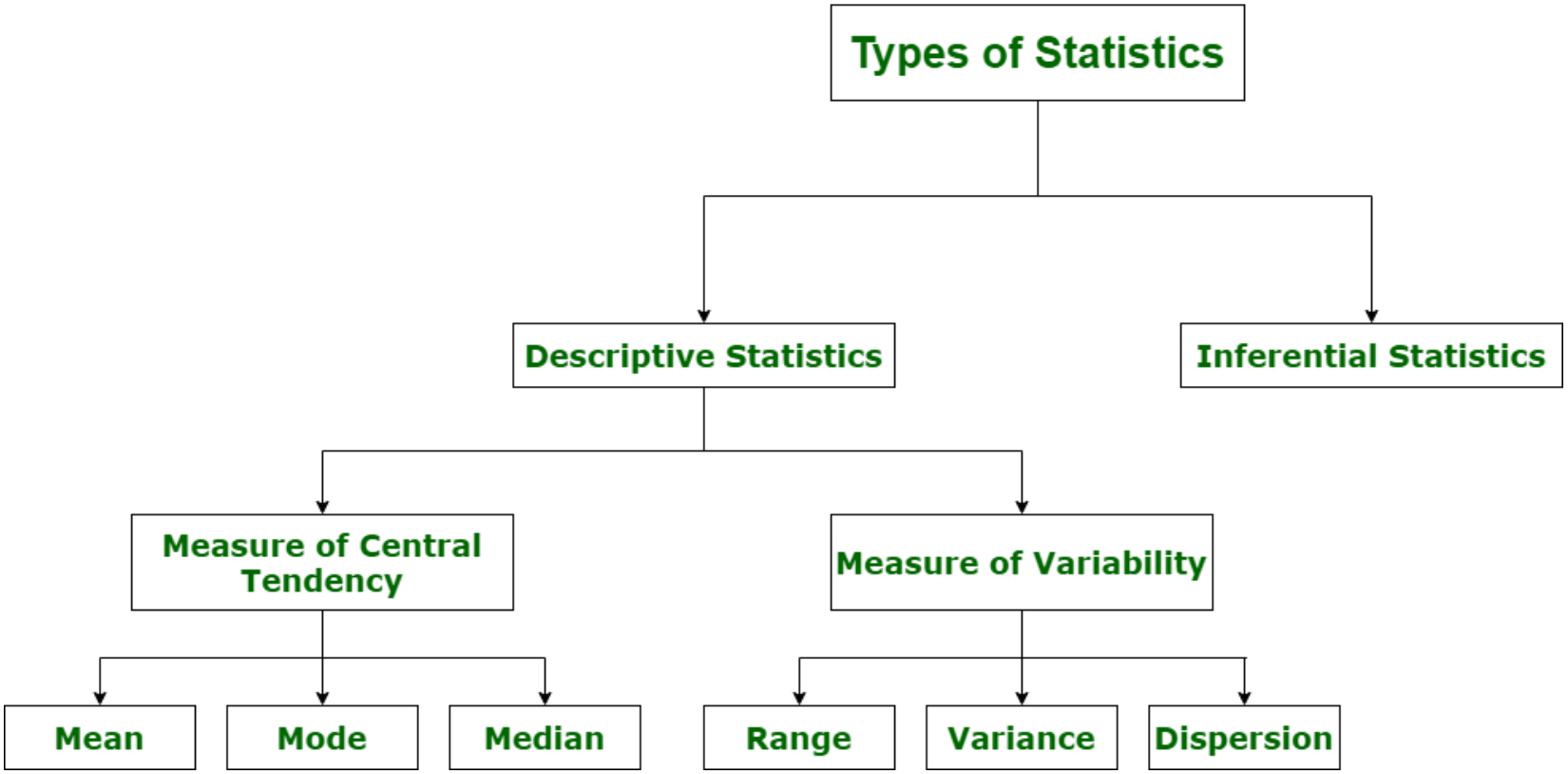
Statistical Descriptions of Data

Tushar B. Kute,
http://tusharkute.com

Types of statistics?



Discriptive Statistics – Types

- There are 3 main types of descriptive statistics:
  - The distribution concerns the frequency of each value.
  - The central tendency concerns the averages of the values.
  - The variability or dispersion concerns how spread out the values are.
- You can apply these to assess only one variable at a time, in univariate analysis, or to compare two or more, in bivariate and multivariate analysis.

Discriptive Statistics – Example

- You want to study the popularity of different leisure activities by gender. You distribute a survey and ask participants how many times they did each of the following in the past year:
  - Go to a library
  - Watch a movie at a theater
  - Visit a national park
- Your data set is the collection of responses to the survey.
- Now you can use descriptive statistics to find out the overall frequency of each activity (distribution), the averages for each activity (central tendency), and the spread of responses for each activity (variability).

Reference: https://www.scribbr.com

Measure of Central Tendency

- A Measure of Central Tendency is a one number summary of the data that typically describes the center of the data.
- These one number summary is of three types.
  - Mean
  - Median
  - Mode

What is mean?

- Mean : Mean is defined as the ratio of the sum of all the observations in the data to the total number of observations.
- This is also known as Average.
- Thus mean is a number around which the entire data set is spread.

Example:

- Consider the following data points.

  17, 16, 21, 18, 15, 17, 21, 19, 11, 23

  Mean — Mean is calculated as

$$Mean = \frac{17+16+21+18+15+17+21+19+11+23}{10} = \frac{178}{10} = 17.8$$

What is median?

- Median is the point which divides the entire data into two equal halves. One-half of the data is less than the median, and the other half is greater than the same.
- Median is calculated by first arranging the data in either ascending or descending order.
  - If the number of observations are odd, median is given by the middle observation in the sorted form.
  - If the number of observations are even, median is given by the mean of the two middle observation in the sorted form.
- An important point to note that the order of the data (ascending or descending) does not effect the median.

What is median?

- To calculate Median, lets arrange the data in ascending order.

  11, 15, 16, 17, 17, 18, 19, 21, 21, 23

- Since the number of observations is even (10), median is given by the average of the two middle observations (5th and 6th here).

$$Median = \frac{5^{th}\ Obs + 6^{th}\ Obs}{2} = \frac{17 + 18}{2} = 17.5$$

What is mode?

- Mode is the number which has the maximum frequency in the entire data set, or in other words,mode is the number that appears the maximum number of times. A data can have one or more than one mode.
  - If there is only one number that appears maximum number of times, the data has one mode, and is called Uni-modal.
  - If there are two numbers that appear maximum number of times, the data has two modes, and is called Bi-modal.
  - If there are more than two numbers that appear maximum number of times, the data has more than two modes, and is called Multi-modal.

What is mode?

- The data:

  11, 15, 16, 17, 17, 18, 19, 21, 21, 23

- Mode is given by the number that occurs maximum number of times.
- Here, 17 and 21 both occur twice. Hence, this is a Bimodal data and the modes are 17 and 21.

Note:

- Since Median and Mode does not take all the data points for calculations, these are robust to outliers, i.e. these are not effected by outliers.
- At the same time, Mean shifts towards the outlier as it considers all the data points. This means if the outlier is big, mean overestimates the data and if it is small, the data is underestimated.
- If the distribution is symmetrical, Mean = Median = Mode. Normal distribution is an example.

Dispersion

- Dispersion refers to measures of how spread out our data is.
- Typically they’re statistics for which values near zero signify not spread out at all and for which large values (whatever that means) signify very spread out.

Dispersion - Types

- Absolute Deviation from Mean
- Variance
- Standard Deviation
- Range
- Quartiles
- Skewness
- Kurtosis

Reference: https://towardsdatascience.com

Mean Absolute Deviation

- The Absolute Deviation from Mean, also called Mean Absolute Deviation (MAD), describe the variation in the data set, in sense that it tells the average absolute distance of each data point in the set.
- It is calculated as,

$$Mean\ \ Absolute\ \ Deviation\ \ = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \overline{X}\right|$$

Variance

- In statistics, the variance is a measure of how far individual (numeric) values in a dataset are from the mean or average value.
- The variance is often used to quantify spread or dispersion. Spread is a characteristic of a sample or population that describes how much variability there is in it.
- A high variance tells us that the values in our dataset are far from their mean. So, our data will have high levels of variability.
- On the other hand, a low variance tells us that the values are quite close to the mean. In this case, the data will have low levels of variability.

Variance

- To calculate the variance in a dataset, we first need to find the difference between each individual value and the mean. The variance is the average of the squares of those differences. We can express the variance with the following math expression:

$$\sigma^2 = \frac{1}{n}\sum_{i=0}^{n-1}(x_i - \mu)^2$$

- In this equation, xi stands for individual values or observations in a dataset. μ stands for the mean or average of those values. n is the number of values in the dataset.
- The term xi - μ is called the deviation from the mean. So, the variance is the mean of square deviations. That's why we denoted it as σ^2.

Variance

Say we have a dataset [3, 5, 2, 7, 1, 3]. To find its variance, we need to calculate the mean which is:

$$(3+5+2+7+1+3)/6 = 3.5$$

Then, we need to calculate the sum of the square deviation from the mean of all the observations. Here's how:

$$(3-3.5)^2 + (5-3.5)^2 + (2-3.5)^2 + (7-3.5)^2 + (1-3.5)^2 + (3-3.5)^2 = 23.5$$

To find the variance, we just need to divide this result by the number of observations like this:

$$23.5/6 = 3.916666667$$

Variance

- That's all. The variance of our data is 3.916666667. The variance is difficult to understand and interpret, particularly how strange its units are.
- For example, if the observations in our dataset are measured in pounds, then the variance will be measured in square pounds.
- So, we can say that the observations are, on average, 3.916666667 square pounds far from the mean 3.5.
- Fortunately, the standard deviation comes to fix this problem

Standard Deviation

- The standard deviation measures the amount of variation or dispersion of a set of numeric values.
- Standard deviation is the square root of variance σ2 and is denoted as σ.
- So, if we want to calculate the standard deviation, then all we just have to do is to take the square root of the variance as follows:

$$Std\ \ Deviation\ \ = \sqrt{Variance\ \ } = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}$$

Standard Deviation

- Again, we need to distinguish between the population standard deviation, which is the square root of the population variance (σ2) and the sample standard deviation, which is the square root of the sample variance (S2).
- We'll denote the sample standard deviation as S:

$$S=\sqrt{S2}$$

Standard Deviation

- There are six steps for finding the standard deviation:
  - List each score and find their mean.
  - Subtract the mean from each score to get the deviation from the mean.
  - Square each of these deviations.
  - Add up all of the squared deviations.
  - Divide the sum of the squared deviations by N – 1.
  - Find the square root of the number you found.

Standard Deviation

- Low values of standard deviation tell us that individual values are closer to the mean.
- High values, on the other hand, tell us that individual observations are far away from the mean of the data.
- Values that are within one standard deviation of the mean can be thought of as fairly typical, whereas values that are three or more standard deviations away from the mean can be considered much more atypical. They're also known as outliers.

Standard Deviation

If we're trying to estimate the standard deviation of the population using a sample of data, then we'll be better served using **n - 1** degrees of freedom. Here's a math expression that we typically use to estimate the population variance:

$$\sigma_x = \sqrt{\frac{\sum_{i=0}^{n-1} (x_i - \mu_x)^2}{n-1}}$$

Note that this is the square root of the sample variance with **n - 1** degrees of freedom. This is equivalent to say:

$$S_{n-1} = \sqrt{S_{n-1}^2}$$

Range

- Range is the difference between the Maximum value and the Minimum value in the data set.
- It is given as,

$$\text{range} = \text{maximum} - \text{minimum}$$

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com