Handling Missing Data: Business Problems

Statistical Function

Mean

In statistics, the mean, also known as the arithmetic mean or average, is a measure of central tendency of a set of numerical values. It is calculated by adding up all the values in the data set and then dividing the sum by the total number of values. The formula for calculating the mean is:

$$m = \frac{\text{sum of the terms}}{\text{number of terms}}$$

Here, m = mean. The mean is a useful tool for summarizing a set of data and understanding its general properties. It can be affected by extreme values or outliers, so it is important to consider other measures of central tendency, such as the median and mode, as well as the variability of the data when analyzing a dataset.

In this course, you will use the **AVERAGE** function in Google spreadsheets to compute the mean.

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

Here, $\bar{x}$ denotes the mean value and x_i denotes the sum of n values where n is the number of values in the sample.

Median

The median is a statistical measure that represents the middle value of a dataset when it is arranged in order of magnitude. It is the value that divides the data set into two equal halves, such that half of the values are above the median and half are below it.

To find the median, the data set is first arranged in ascending or descending order.

1. If the data set contains an odd number of values, then the median is the middle value. For example, the median for a sorted list of 15 observations is the 8th value.
2. If the data set contains an even number of values, then the median is the average of the two middle values. For example, the median for a sorted list of 16 observations is the average of the 8th and 9th values.

$$\mathrm{Med}(X) = \begin{cases} X[\frac{n+1}{2}] & \text{if n is odd} \\ \frac{X[\frac{n}{2}]+X[\frac{n}{2}+1]}{2} & \text{if n is even} \end{cases}$$

X = ordered list of values in data set

n = number of values in data set

Unlike the mean, the median is not influenced by extreme values or outliers in the data set, making it a useful measure of central tendency in skewed or asymmetric distributions. A skewed or asymmetric distribution is a type of data distribution where the values are not evenly spread out around the average or middle of the data. In this type of distribution, the data tends to be concentrated on one side of the center, and the other side has fewer values that are more spread out.

In this course, you will use the **MEDIAN** function in Google spreadsheets to compute the mean.

Mode

In statistics, the mode is a measure of central tendency that represents the most frequent value in a dataset. More specifically, the mode is the value that occurs with the highest frequency in a set of observations or data points. It is one of the three main measures of central tendency, along with the mean and median.

The mode is particularly useful when dealing with categorical or discrete data, such as the number of times a certain event occurs, or the most common color of cars in a parking lot. It is also useful when dealing with continuous data that can be grouped into categories or bins.

Unlike the mean and median, the mode does not take into account the actual values of the data, only their frequency of occurrence. This makes it less sensitive to outliers or extreme values that may affect the mean or median. However, it may not be a representative measure of central tendency if there are multiple modes in the dataset or if the frequency of the modes is close to each other. You can use the **MODE** function to compute the mode.

Variance

In statistics, variance is a measure of how spread out a dataset is. More specifically, it measures the average squared difference between each data point and the mean of the dataset. Variance is represented by the symbol σ^2 for a population and s^2 for a sample.
Variance is commonly used in statistics to describe the variability or spread of a dataset. It is a useful tool for comparing the spread of two or more datasets, as well as for identifying outliers or extreme values in a dataset.

The formula for variance depends on whether you are calculating the variance of a population or a sample.

• Population Variance (σ^2)

The population variance is calculated using the following formula:

$$\sigma^2 = \Sigma(x - \mu)^2 / N$$

- σ^2 represents the population variance
- x represents each data point in the dataset
- μ represents the population mean
- N represents the total number of data points in the dataset

This formula calculates the average squared difference between each data point and the population mean.

• Sample Variance (s^2)

The sample variance is calculated using a similar formula, but with n-1 in the denominator instead of N to account for the fact that the sample mean is an estimate of the population mean:

$$s^2 = \Sigma(x - \bar{x})^2 / (n-1)$$

- s^2 represents the sample variance
- x represents each data point in the dataset
- $\bar{x}$ represents the sample mean
- n represents the sample size

This formula calculates the average squared difference between each data point and the sample mean.

Both formulas involve squaring the differences between each data point and the mean, which gives more weight to larger differences and emphasizes the spread of the dataset. The result is a measure of how much the data points in a dataset vary from the mean. Google Spreadsheets use the sample variance formula. You can use the **VAR** function to compute the variance.

Standard Deviation

Standard deviation is a statistical measure that is used to quantify the amount of variability or dispersion in a set of data. It is defined as the square root of the variance and is typically denoted by the symbol σ (sigma).

The standard deviation tells us how spread out the data is from the mean, or average, value. A low standard deviation means that the data points tend to be close to the mean, while a high standard deviation means that the data points are spread out over a wider range.

To calculate the standard deviation, first find the mean of the data set. Then, for each data point, subtract the mean and square the result. Next, find the average of these squared differences, which is the variance. Finally, take the square root of the variance to get the standard deviation. The formula for the standard deviation is:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

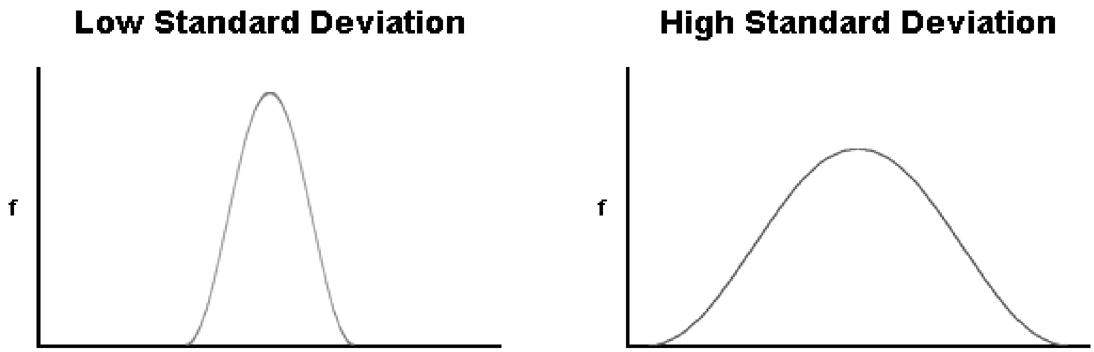
σ = population standard deviation

N = the size of the population

x_i = each value from the population

μ = the population mean

The higher the standard deviation the more variability or spread you have in your data. The larger your standard deviation, the more spread or variation in your data. Small standard deviations mean that most of your data is clustered around the mean.

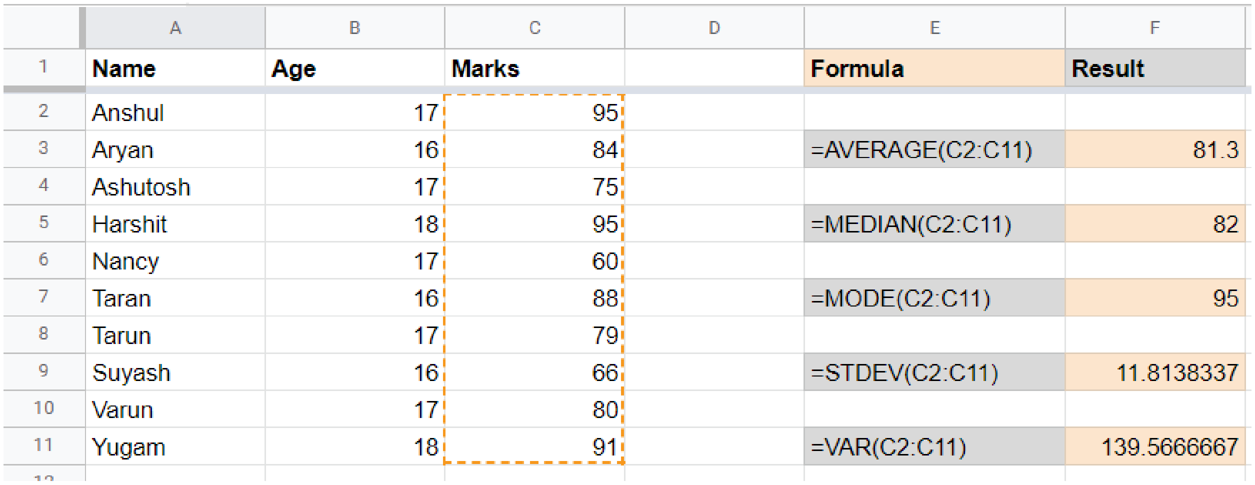


As you can see in the graph When Low standard deviation values are clustered around the mean but in another case, the values are spread. You can use **STDEV** function to compute the standard deviation.

How to impute missing values?

1. If the Standard deviation is similar/ near to Average or bigger value, then we replace the missing value with Median
2. If the Standard deviation is less than the average value or has a small value that means values are clustered near to Average, then we replace the missing value with the Average

Here is the implementation of all the mentioned statistical operations in a given Student dataset



	A	B	C	D	E	F
1	Name	Age	Marks		Formula	Result
2	Anshul	17	95			
3	Aryan	16	84		=AVERAGE(C2:C11)	81.3
4	Ashutosh	17	75			
5	Harshit	18	95		=MEDIAN(C2:C11)	82
6	Nancy	17	60			
7	Taran	16	88		=MODE(C2:C11)	95
8	Tarun	17	79			
9	Suyash	16	66		=STDEV(C2:C11)	11.8138337
10	Varun	17	80			
11	Yugam	18	91		=VAR(C2:C11)	139.5666667

Statistical operations were performed using the “Marks” column in the dataset

Read More

- https://support.google.com/docs/answer/3094063?hl=en