Base concept

Location Statistic

A **location statistic** (or **measure of central tendency**) is a key concept in descriptive statistics, and it is used to summarize a dataset by identifying a central or typical value around which other values tend to cluster. These statistics are often called **measures of central tendency** because they help to describe the "center" of a data distribution. They are critical in data analysis because they provide a simple summary that helps to understand and compare different datasets.

There are several common location statistics, each of which can give a slightly different perspective on the data:

Definition of Mean (Arithmetic Mean)

The **mean** is the average of a set of numbers. To find it, you add all the numbers together and then divide by the number of values in the set. It is commonly used to represent the central value of a dataset.

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Where $x_1, x_2, \cdots, x_n$ are the values in the dataset, and n is the number of values.

Definition of Mode

The **mode** is the value that appears most frequently in a dataset.
Basically, a dataset can have:

- No mode (if no number repeats)
- One mode (unimodal)
- Two modes (bimodal)
- More than two modes (multimodal)

Definition of Median

The **median** is the middle value of a dataset when it is ordered from least to greatest. If the dataset has an odd number of values, the median is the middle one. If the dataset has an even number of values, the median is the average of the two middle numbers.

Steps to find the median:

1. Order the dataset.
2. If n (number of values) is off, the median is the middle number.
3. If n is even, the median is the average of the two middle numbers.

There are several point that explain the importance of Location statistic:

- **Data Summary**: Location statistics allow you to summarize large datasets with a single number, giving a quick overview of the "center" or "typical" value in the data.
- **Comparison**: Location statistics provide a way to compare different datasets.
- **Data Distribution Insights**: The differences between the mean, median, and mode can give insights into the shape of the data distribution. For example:
  - If the **mean** is higher than the **median**, the distribution is **positively skewed** (tail to the right).
  - If the **mean** is lower than the **median**, the distribution is **negatively skewed** (tail to the left).
  - If the **mean** and **median** are approximately the same, the distribution is likely **symmetric**.

Different ways to define the Average

Arithmetic Mean

The arithmetic mean is the most commonly used average. It is calculated by summing all values in the dataset and dividing by the number of values.

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Weighted Mean

The weighted mean takes into account different weights for each data point, allowing for greater influence of some values over others.

$$\text{Weighted Mean} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

Where:

- w_i is the weight of the $i-th$ value x_i.
- $\sum w_i$ is the sum of all weights.

Geometric Mean

The geometric mean is used primarily for data involving growth rates (e.g., investment returns, population growth). It is calculated as the nth root of the product of the data values.

$$\text{Geometric Mean} = (\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$$

where:

- $\prod_{i=1}^{n} x_i$ is the product of all data values

Harmonic Mean

The harmonic mean is useful for rates, such as speeds or densities. It is calculated as the reciprocal of the arithmetic mean of the reciprocals of the data values.

$$\text{Harmonic Mean} = \frac{n}{\sum_{i=1}^{n}} \frac{1}{x_i}$$

where:

- $x_1, x_2, \cdots, x_n$ are the data values.

Proof that the mean minimizes the sum of the square distances

Question

Finde simple proof of:

$$\sum_{i=1}^{n} |x_i - c|$$

To prove that the mean minimizes the sum of squared distances, we'll go through a formal derivation. The key idea is that, given a set of numbers, the arithmetic mean (average) is the value that minimizes the sum of the squared differences between each number in the set and some constant c.

Problem Statement

Given a set of numbers $x_1, x_2, \ldots, x_n$, we want to find a value c that minimizes the sum of squared distances to each x_i. The objective is to minimize:

$$S(c) = \sum_{i=1}^{n} (x_i - c)^2$$

We will show that the value of c that minimizes this function is the arithmetic mean:

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

Proof

1. **Define the objective function**:
   The function we wish to minimize is the sum of squared distances:

   $$S(c) = \sum_{i=1}^{n} (x_i - c)^2$$

2. **Differentiate the function with respect to c**:
   To find the value of c that minimizes $S(c)$, take the derivative of $S(c)$ with respect to c:
   (Remember that we use partial derivate because with it we can find the minimum value of c)

   $$\frac{dS(c)}{dc} = \frac{d}{dc} \left(\sum_{i=1}^{n} (x_i - c)^2 \right)$$

To solve it, we need to apply:

$$\frac{dS(c)}{dc} = \sum_{i=1}^{n} 2(x_i - c)(-1)$$

Simplifying this expression:

$$\frac{dS(c)}{dc} = -2\sum_{i=1}^{n}(x_i - c)$$

\frac{dS(c)}{dc} = -2 \left(\sum_{i=1}^{n} x_i - nc \right)
$$

3. **Set the derivative equal to zero**:
   To find the critical points, set the derivative equal to zero:

$$-2\left(\sum_{i=1}^{n} x_i - nc\right) = 0$$

   This simplifies to:

$$\sum_{i=1}^{n} x_i = nc$$

4. **Solve for c**:
   Solving for c gives:

$$c = \frac{1}{n}\sum_{i=1}^{n} x_i$$

   Thus, the value of c that minimizes the sum of squared distances is the arithmetic mean:

$$c = \bar{x}$$

Conclusion

We have shown that the value of c that minimizes the sum of squared distances $\sum_{i=1}^{n}(x_i - c)^2$ is the arithmetic mean of the set:

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

Thus, the mean minimizes the sum of squared distances to the points in the set.