**Statistical Quantities**

**Introduction:**

There are five basic statistical quantities Mean, Median, Mode, Variance and standard deviation. Mean, Median, Mode are also called central Tendency. Central tendency (or measure of central tendency) is a central or typical value for a [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution). It may also be called a center or location of the distribution. A central tendency can be calculated for either a finite set of values or for a theoretical distribution, such as the [normal distribution](https://en.wikipedia.org/wiki/Normal_distribution).

1. **Mean:**

In simple terms, Arithmetic mean (or simply, mean) is the sum of all measurements divided by the number of observations in the data set.

* The mean is:
  + the arithmetic average of all the scores (ΣX)/N
  + the number, m, that makes Σ(X - m) equal to 0
  + the number, m, that makes Σ(X - m)2 a minimum
* The mean of a population is represented by the Greek letter μ; the mean of a sample is represented by X
* You should use the mean when
  + the data are interval or ratio scaled
    - Many people will use the mean with ordinally scaled data too
* The mean is preferred because it is sensitive to every score
  + If you change one score in the data set, the mean will change

**Care to be taken**

* Mean is very sensitive to outliers (An abnormal value in the data). Do not uses mean if data contains outliers else lead to false assumptions.
* Mean is correct metric if data is normally distributed.

1. **Median**

The middle value that separates the higher half from the lower half of the data set. The median and the mode are the only measures of central tendency that can be used for ordinal data, in which values are ranked relative to each other but are not measured absolutely.

The median is simply another name for the 50th percentile

* It is the score in the middle; half of the scores are larger than the median and half of the scores are smaller than the median
* The median is often used when the distribution of scores is either positively or negatively skewed
  + The few really large scores (positively skewed) or really small scores (negatively skewed) will not overly influence the median

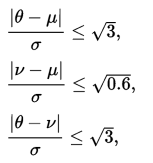
1. **Mode**

The most frequent value in the data set, one that occurs maximum number of times. This is the only central tendency measure that can be used with nominal data, which have purely qualitative category assignments.



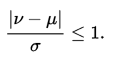
## Relationships between the mean, median and mode

For [unimodal distributions](https://en.wikipedia.org/wiki/Unimodal_distribution) the following bounds are known and are sharp

{\displaystyle {\frac {|\theta -\mu |}{\sigma }}\leq {\sqrt {3}},}{\displaystyle {\frac {|\nu -\mu |}{\sigma }}\leq {\sqrt {0.6}},}

Where μ is the mean, ν is the median, θ is the mode, and σ is the standard deviation.

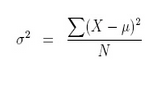
For every distribution,



The mean, mode and median do a nice job in telling where the center of the data set is, but often we are interested in more. For example, a pharmaceutical engineer develops a new drug that regulates iron in the blood. Suppose she finds out that the average sugar content after taking the medication is the optimal level. This does not mean that the drug is effective. There is a possibility that half of the patients have dangerously low sugar content while the other half has dangerously high content. Instead of the drug being an effective regulator, it is a deadly poison. What the pharmacist needs is a measure of how far the data is spread apart. This is what the variance and standard deviation do.

**Variance:**

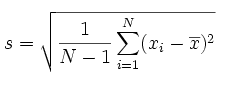
Variance measures how far a set of numbers are spread out from their mean. Variance is calculated by taking the differences between each number in the set and the mean, squaring the differences (to make them positive) and dividing the sum of the squares by the number of values in the set.



It is used in descriptive statistics, statistical inference, hypothesis testing, goodness of fit, Monte Carlo sampling, amongst many others.

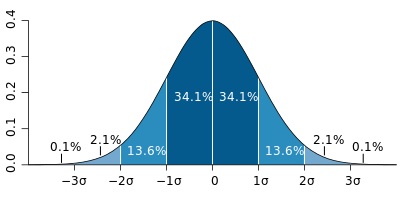
**Standard Deviation:**

Standard deviation tells you how much data deviates from the actual mean. It is the square root of the Variance

****

A low standard deviation indicates that the data points tend to be close to the mean, while a high standard deviation indicates that the data points are spread out over a wider range of values. A useful property of the standard deviation is that, unlike the variance, it is expressed in the same units as the data.

Standard Deviation can be used to check the data distribution (Ideal assumption is normally distributed data). If a data distribution is approximately normal then about 68 percent of the data values are within one standard deviation of the mean (mathematically, μ ± σ, where μ is the arithmetic mean), about 95 percent are within two standard deviations (μ ± 2σ), and about 99.7 percent lie within three standard deviations (μ ± 3σ). This is known as the [68-95-99.7 rule](https://en.wikipedia.org/wiki/68-95-99.7_rule), or the empirical rule.



**Additional information:**

Normal distribution is also called standard normal distribution in which μ=0 and σ=1.

**Applications**:

1. Mean, Median, Mode can be used for filling missing values
2. Standard deviation is used as a measure of the [risk](https://en.wikipedia.org/wiki/Risk#Finance) of a portfolio of assets
3. In the industrial applications, Standard deviation is often used to compare real-world data against a model to test the model.
4. Used to detect outliers
5. Used to understand data variability.

**Interview Questions:**

1. What is the Central Limit Theorem? Explain it. Why is it important?
2. What is [68-95-99.7 rule](https://en.wikipedia.org/wiki/68-95-99.7_rule)?
3. How will you compare two distributions?
4. What is the difference between parametric and non-parametric test?