

Statistical Methods

(Central tendency, dispersion, standard error, coefficient of variation, Probability distributions and Confidence limits)

Course Code –BOTY 4204

Course Title- Techniques in plant sciences , biostatistics
and bioinformatics

By – Dr. Alok Kumar Shrivastava

Department of Botany

Mahatma Gandhi Central University, Motihari

Unit 4- Statistical Methods

Central tendency, dispersion, standard error, coefficient of variation; Probability distributions (normal, binomial of Poission) and Confidence limits. Test of statistical significance (t-test, Chi-square): Analysis of variance-Random Block Design and its application in plant breeding and genetics; Correlation and Regression.

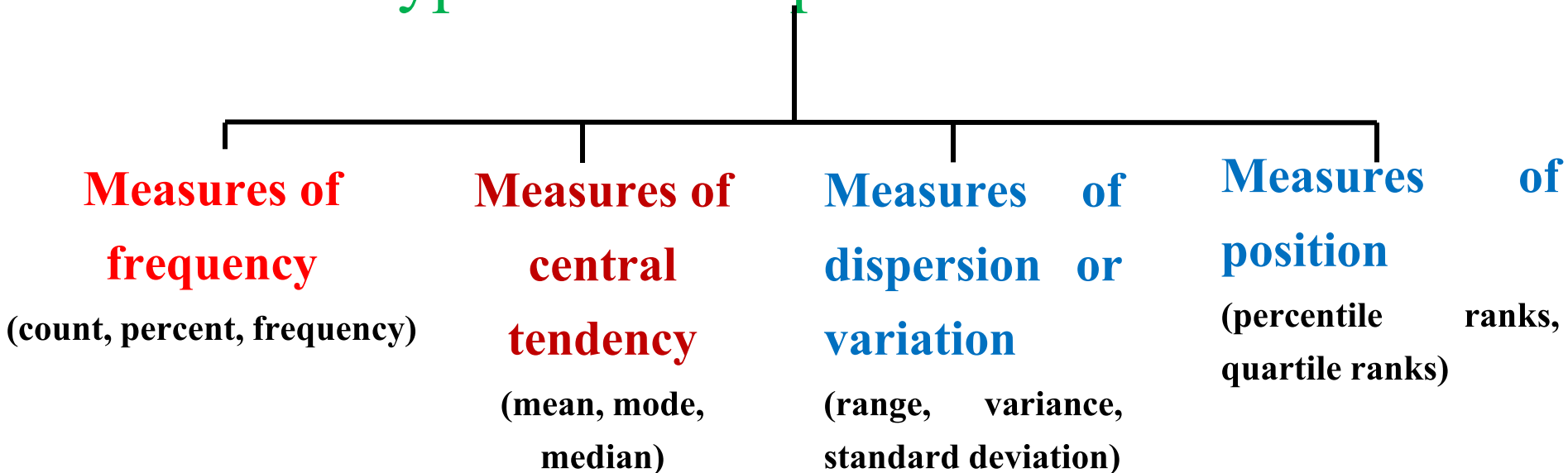
Statistics

- ✓ It is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data in such a way that meaningful conclusion can be drawn from them. Statistics teaches us to use a limited sample to make intelligent and accurate conclusions about a greater population. There are two types of statistical methods are used in analyzing data: descriptive and inferential statistics.
- ✓ Descriptive statistics are used to synopsise data from a sample exercising the mean or standard deviation.
- ✓ Inferential statistics are used when data is viewed as a subclass of a specific population.

Descriptive statistics

- Descriptive statistics uses the data to provide descriptions of the population, entire through numerical calculations or graphs or tables.
- Descriptive statistics enables us to present the data in a more meaningful way, which allows simpler interpretation of the data.

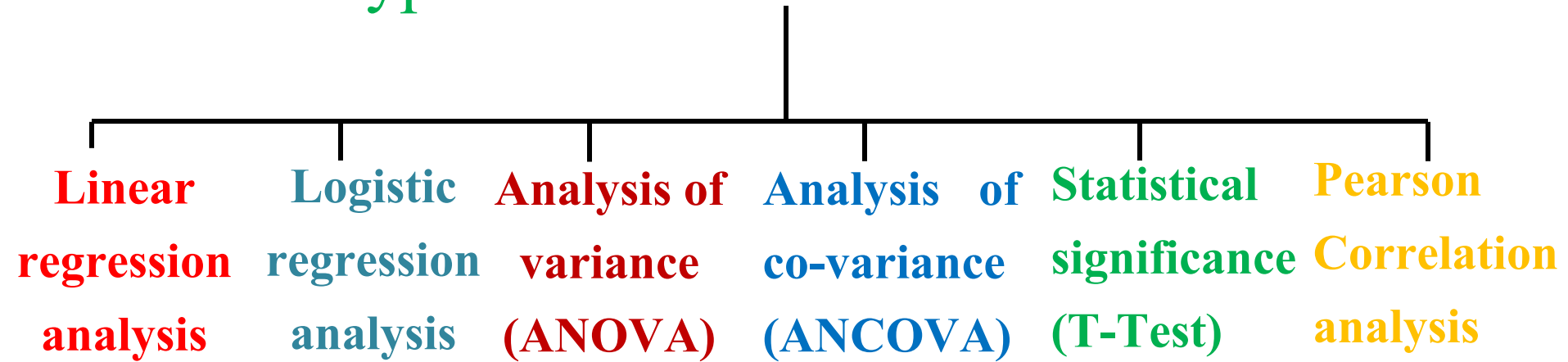
Types of descriptive statistics



Interferential statistics

- Interferential statistics makes interferences and predictions about a population based on a sample of data taken from the population in question.
- Through interferential statistics, one can take data from samples and make generalizations about a population.
- The most common methodologies in inferential statistics are hypothesis tests, confidence intervals, and regression analysis.

Types of inferential statistics



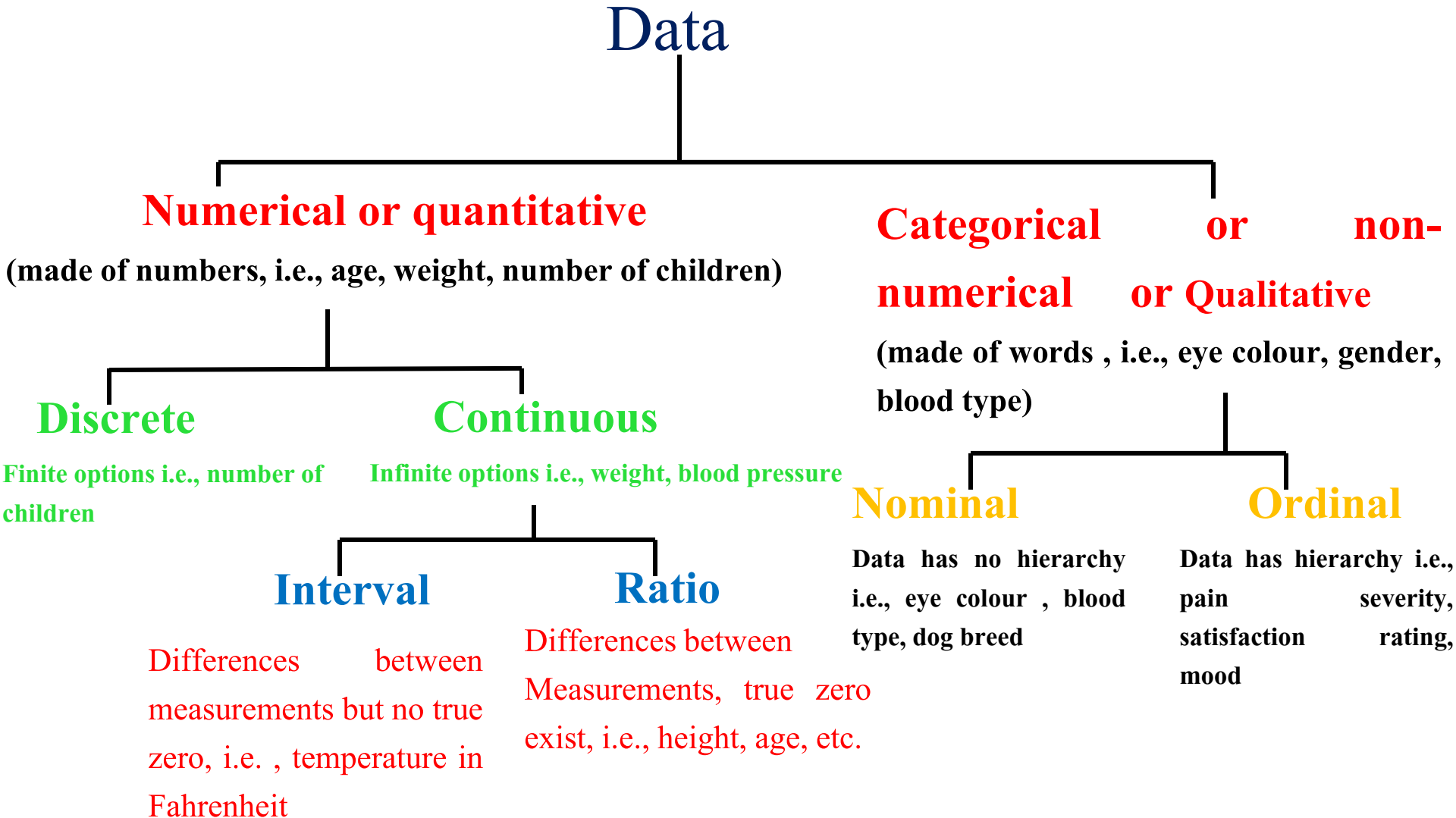
Characteristics of statistics

- ✓ Statistics are aggregates of facts.
- ✓ Statistics are numerically expressed.
- ✓ Statistics are affected to a marked extent by multiplicity of causes.
- ✓ Statistics are enumerated or estimated according to a reasonable standard of accuracy.

Data

- ✓ A piece of information which you collect through your study called data.
- ✓ In other words it is a set of facts and figures which are useful in a particular purpose like a survey or an analysis. When one arrange data in an organized form, they refers to as information.
- ✓ Data can be classified as qualitative and quantitative.

Types of statistical data



Categorical vs. numerical

- ✓ Categorical values or observations that can be sorted into groups or categories. Bar charts and pie graphs are used to graph categorical data.
- ✓ Numerical values or observations that can be measured as these number can be placed in ascending or descending order. Scatter plots and line graphs are used to graph numerical data.

Discontinuous (Discrete) vs. Continuous data

- ✓ A set of data is said to be discrete if the values belonging to the set are distinct and separate. This data obtained through counting. Discrete data is a whole number (integer) and it can not be subdivided into smaller and smaller parts, i.e., number of people in a classroom or numbers of brothers in a family, number of leaves in plant, etc. you cant have 30.5 people in the classroom. Discrete data can be numeric like numbers of apples but it can also be categorical like red or blue , or male or female, or good or bad.
- ✓ A set of data is said to be continuous if the values belonging to the set can take on any value within a finite or infinite interval. This data gathered through measurement like heights of plants, weights of plants, flowering time, etc. Continuous data continues on and on and on , at least that's how I think about it. These are data that can be broken into smaller and smaller units weight can be infinitely measured using precision equipment and does not have to stop at three meters.

Nominal vs. Ordinal data

- ✓ Nominal values or observations can be assigned a code in the form of a number where the numbers are simply labels. You can count but not order or measure nominal data. Example-sex, and eye colour.
- ✓ Ordinal values or observations can be ranked or put in order or have a rating scale attached. You can count and order, but not measure, ordinal data. Example- house number and swimming level.

Data measurement scale

The frame into which we wish to make everything fit is called scale. Every scale requires some unit and characteristics to classify measurements. The greater the refinement in the unit of measurement, the greater confidence in the measurement of findings. There are four measurement scales discovered by a psychologist researcher named Stanley Stevens.

Data measurement scale

Nominal

or

Labels or

classificatory scale

(use for classifying things without ranking)

Ordinal or

ranking scale

(use for classifying things with ranking)

Interval

(assume that intervals between each of the number are equal)

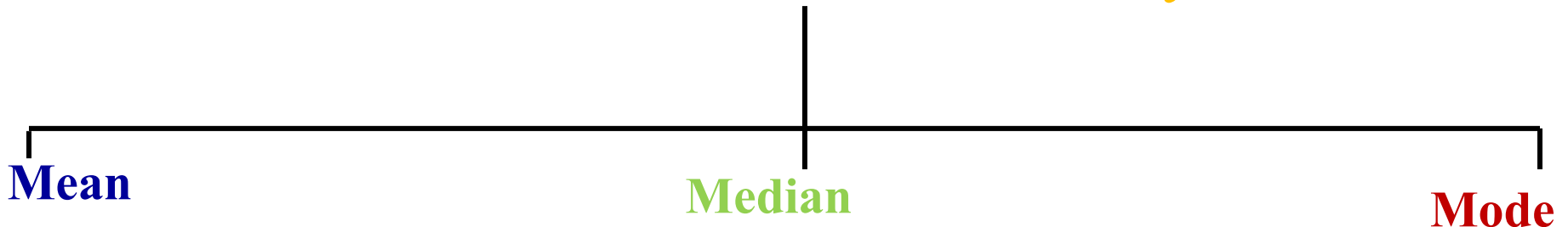
Ratio

(same as interval scales, but also contain absolute zero)

Central tendency

A measure of central tendency in a single value that describes the way in which a group of data cluster around a central value. In other words, it is a way to describe the center of a data set. It may also be called a center or location of the distribution. There are three measures of central tendency: the mean, the median and the mode. Mean is the most frequently used measure of central tendency and generally considered the best measure of it. However , there are some situations where either median or mode are preferred. Median is the preferred measure of central tendency when: there are a few extreme scores in the distribution of the data.

Measure of central tendency



Mean

- The mean is the average of a data set. That is used to derive the central tendency of the data in question.
- It is found by adding all data points and dividing by the number of data points. The resulting number is known as mean or average.
- The mean is essentially a model of your dataset. It is the value that is most common.
- By using the mean one can describe the sample with a single value that represents the center of the data.
- $\text{mean} = \text{sum of data} / \text{number of data points}$

Median

- It is a simple measure of central tendency. It is the most suitable measure of average for data classified on an ordinal scale.
- The middle number, found by ordering all data points and picking out the one in the middle.
- It is the value which separating the lower half from higher half of a data sample.
- By comparing the median to the mean, one can get an idea of the distribution of a data set. When the mean and the median are same, the data set is more or less evenly distributed from lowest to highest values.

- If the number of observations is odd then the median is the observations that are ranked at position $N+1/2$ (It is the middle value).
- If the number of observations is even then the median is the average value of the observations that are ranked at numbers $N/2$ and $(N/2)+1$ (It is the average or mean of the two middle most values).
- Outliers and skewed data have a smaller effect on the median consequently , when some of the value s are more extreme , the effect on the median is smaller. When data distribution is skewed the median is better measure of the central tendency than the mean.

Mode

- The most frequent number that is, the number that occurs the highest number of times in any data set. In other words one can say the mode of a set of data values is the value that appears most often in data set. The mode is the measure of average that can be used with nominal data. A data set may be bimodal (data set have two modes), trimodal (data set have three modes), multimodal (data set have more than three modes) or no-model (If all the number appear the same number of times in any data). The term mode originates with Karl Pearson in 1895.
- Like mean and median, the mode is a way of expressing , in a single number, important information about a random variable or a population.
- The numerical value of the mode is the same as that of the mean and median in a normal distribution, and it may be very different in highly skewed distributions.

Dispersion

Dispersion is the extent to which a distribution is stretched or squeezed. It is also called variability, scatter or spread. Common example of measures of dispersion in statistic are the variance, standard deviation and interquartile range. Measure of statistical dispersion is a non negative that is zero if all the data are the same and increases as the data become more diverse.

Variance

- Variance is the expectation of the square deviation of a random variable from its mean. It measures how far a set of numbers or data are spread out from their average value. A high variance indicates the data points are very spread out from the mean, and from one another. Variance is the average of the squared distances from each point to the mean.
- It is calculated as the average squared deviation of each number from the mean of a data set.

Standard deviation

- Standard deviation is a statistic that measures the dispersion or variability of a dataset relative to its mean and is calculated as the square root of the variance
- Standard deviation is a number used to tell how measurement for a group are spread out from the average (mean), or expected value. A low standard deviation means that most of the numbers are close to the average. A high standard deviation means that the numbers are more spread out.
- It is the best measure of variation.
- If standard deviation is zero it means that all observations are identical.

Standard deviation vs. variance

- Standard deviation and variance are the most commonly used measures of spread.
- Standard deviation looks at how spread out a group of numbers is from the means, by looking at the square root of the variance . The variance measures the average degree to which each point differs from the mean(the average of all data point).
- Standard deviation symbolized by ' S^2 ' and variance symbolized by 'S'

Interquartile range

- ✓ In descriptive statistics , the interquartile range also called the midspread, middle 50%, or H-spread.
- ✓ It is a measure of statistical dispersion based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts the values that divide each part are called the first (Q1), second (Q2), and third (Q3) quartiles.
- ✓ It is equal to the difference between 75th and 25th percentiles, or between third quartile (upper quartile) and first quartile (lower quartiles).

$$\text{IQR} = \text{Q3} - \text{Q1}$$

- ✓ It is less sensitive measure of the spread of a data set and due to its resistance to outliers, it is useful in identifying when a value is an outlier.

Standard Error

- The standard error is the approximate standard deviation of sample population. Standard error is the statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation. In statistics, a sample mean deviates from the actual mean of a population-this deviation is the standard error of the mean.

- The standard error is the approximate standard deviation of a statistical sample population.
- The standard error can include the variation between the calculated mean of the population and one which is considered known or accepted as accurate.
- The more data points involved in the calculations of the mean, the smaller the standard error tends to be.
- The term standard error is used to refer to the standard deviation of various sample statistics, such as the mean or median. For example, the “standard error of the mean” refers to the standard deviation of the distribution of sample means taken from the population.
- The smaller the standard error , the more representative the sample will be of the overall population.

- In statistics, a sample mean deviates from the actual mean of a population-this deviation is the standard error of the mean which is inversely proportional to the sample size.

Standard deviation vs. standard error

- The standard deviation (SD) measures the amount of variability or dispersion, for a subject set of data from the average (mean) or expected value while standard error of the sample mean is an estimate of how far the sample mean of the data is likely to be from the true population mean.
- Standard error is always smaller than standard deviation.

Coefficient of variation

- The coefficient of variation also known as relative standard deviation, is a standardized measure of dispersion of a probability distribution or frequency distribution.
- It is useful static for comparing the degree of variation from one data series to another. The greater the SD value the less precise the data because it increases the acceptable range within the deviation. Value of CV between 2%-3% is good and acceptable.
- It is often expressed as a percentage, and is defined as the ratio of the standard deviation to the mean.

Probability distributions

- A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range. This range will be bounded between the minimum and maximum possible values, but precisely where the possible value is likely to be plotted on the probability distribution depends on number of factors. These factors include the distribution's mean, standard deviation, skewness and kurtosis.

- The most common probability distribution is the normal distribution, or ‘bell curve’, although several distributions exist that are commonly used. Typically, the data generating process of some phenomenon will dictate its probability distribution. This process is called the probability density function.
- There are many different classifications of probability distributions. Some of them include the normal distribution, chi square distribution, binomial distribution and Poisson distribution.
- The different probability distributions serve different purposes and represent different data generation processes.
- The binomial distribution evaluates the probability of an event occurring several times over a given number of trials and given the event's probability in each trial. A binomial distribution is discrete, as opposed to continuous, since only 1 or 0 is a valid response.

Normal distribution or Gaussian distribution

- It is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.
- Normal distributions are symmetrical, but not all symmetrical distribution are normal
- The normal distribution is fully characterized by its mean and standard deviation, meaning the distribution is not skewed and does exhibit kurtosis. This makes the distribution symmetric and it is depicted as a bell-shaped curve. A normal distribution is defined by a mean (average) of 0 and a standard deviation of 1, with skewness of 0 and kurtosis=3.
- In a normal distribution, approximately 68% of the data collected fall within +/- 1 standard deviation of the mean; approximately 95% of the data within +/- 2 standard deviation of the mean; and 99.7% within +/- 3 standard deviation.
- Unlike the binomial distribution the normal distribution is continuous, meaning that all possible values are represented (as opposed to 0 and 1 with nothing in between).

Skewness

- Skewness refers to distortion or asymmetry in a symmetrical bell curve, or normal distribution, in a set of data.
- Real life data rarely follow a perfect normal distribution. The skewness and kurtosis coefficients measure how different a given distribution is from a normal distribution. The skewness measures the symmetry of a distribution. The normal distribution is symmetric and has a skewness of 0. If the distribution of the data set has a skewness less than zero, or negative skewness, then the left tail of the distribution is longer than the right tail; positive skewness implies that the right tail of the distribution is longer than the left.

Types of Skewness



Positive

Curve has longer tail toward the right and Mean exceeds Mode and Median

Mode < Median < Mean

Normal

Mean, Mode and Median are equal

Mean = Median = Mode

Negative

Curve has longer tail toward the left And Mode exceeds Mean and Median

Mean < Median < Mode

Kurtosis

- Like skewness, kurtosis is a statistical measure that is used to describe the distribution. Whereas skewness differentiates extreme values in one versus other tail, kurtosis measures extreme values in either tail.
- The kurtosis statistic measures the thickness of the tail ends of the distribution in relation to the tails of the normal distribution.
- Distributions with large kurtosis exhibits tail data exceeding the tails of the normal distribution(e.g., 5 or more standard deviations from the mean).
- Distributions with low kurtosis exhibits tail data that is generally less extreme than the tails of the normal distribution.
- The normal distribution has a kurtosis of three, which indicates the distribution has neither fat nor thin tails.
- If an observed distribution has a kurtosis greater than 3, the distribution is said to have heavy tails when compared to the normal distribution. If the distribution has a kurtosis of less than 3, it is said to have thin tails when compared to the normal distribution. This phenomenon is known as kurtosis risk.

Types of kurtosis

```
graph TD; A[Types of kurtosis] --- B[Mesokurtosis or Normal]; A --- C[Leptokurtosis or Positive]; A --- D[Platykurtosis or Negative]; B --- B1["Distribution identical to the normal"]; B --- B2["=3"]; C --- C1["A distribution more peaked to the normal"]; C --- C2[">3"]; D --- D1["A distribution less peaked to the normal"]; D --- D2["<3"];
```

**Mesokurtosis or
Normal**

**Distribution identical to the
normal**

=3

**Leptokurtosis or
Positive**

**A distribution more
peaked to the normal**

>3

**Platykurtosis or
Negative**

**A distribution less peaked
to the normal**

<3

Poisson distribution

- ✓ It is discrete probability distribution of the number of events occurring in a given time period, given the average number of times the event occurs over that time period .
- ✓ It is a tool that helps to predict the probability of certain events from happening when you know how often the event has occurred. It gives us the probability of a given number of events happening in a fixed interval of time.
- ✓ Poisson is positively skewed and when the mean of Poisson distribution is large, it becomes similar to a normal distribution or symmetrical.
- ✓ The shape of a Poisson distribution changes for example , a Poisson distribution with low mean is highly skewed, with 0 as the mode. All the data pushed up against 0 , with the tail extending to the right.

Binomial distribution

- ✓ A binomial distribution can be thought of as simply the probability of a success or failure outcome in an experiment or survey that is repeated multiple times. The binomial is a type of distribution that has two possible outcomes.
- ✓ It is common way to test the distribution and it is frequently used in statistic.
- ✓ There are two most important variables in the binomial formula such as-
 1. 'N' it stands for the number of times the experiment is conducted
 2. ' π ' represents the possibility of one specific outcome

Criteria of binomial distribution

- ✓ For binomial distribution the number of trial must be fixed.
- ✓ Every trial must be independent.
- ✓ The probability always stays the same and equal.

Formula for binomial distribution

$$P_{(x)} = \frac{N!}{x!(N-x)!} * \pi^x (1-\pi)^{N-x}$$

or

$$P_{(x)} = \frac{N!}{x!(N-x)!} * \pi^x (q)^{N-x}$$

$P_{(x)}$ = probability of x successes out of N trials

N == number of trials

π = probability of success on a given trial

Mean of the distribution (μ) = $n\pi$

Variance (σ^2) = $n\pi q$ (it is always less than its mean in binomial variable. For maximum variance $\pi=q=0.5$ and $q_{\max=n/4}$)

$$q = \pi - 1$$

q = probability of failure on a given trial

Confidence limits

- ✓ A confidence interval is a range of values we are fairly sure our true value lies in.
- ✓ A confidence interval, in statistics, refers to the probability that a population parameter will fall between two set values for a certain proportion of times. Confidence intervals measure the degree of uncertainty or certainty in a sampling method. A confidence interval can take any number of probabilities, with the most common being a 95% or 99% confidence level.
- ✓ Confidence level and confidence interval are interrelated but are not exactly the same.

- ✓ A confidence interval is a range of values likely would contain an unknown population parameter. Confidence level refers to the percentage of probability, or certainty , that the confidence interval would contain the true population parameter when you draw a random sample many times. Or, in vernacular , “we are 99% certain (confidence level) that most of these datasets (confidence intervals) contain the true population parameter.”
- ✓ Confidence interval calculates the probability that a population parameter will fall between two set values.
- ✓ Confidence intervals measure the degree of uncertainty in a sampling method.
- ✓ Most often confidence intervals reflect confidence levels of 95% or 99%.

✓ For example a result might be reported as $50\% \pm 6\%$, with a 95% confidence. So the confidence interval : $50\% \pm 6\% = 44\%$ to 56% and the confidence level: 95%.

✓ Confidence level is expressed as percentage.

✓ A 0% confidence level means you have no faith at all that if you repeated the experiments that you would get the same results. In fact, you are sure the results would be completely different.

✓ A 100% confidence level means there is no doubt at all that if you repeated the experiments that you would get the same results. In fact, you are sure the results would be repeatable of the time.

Thank you