Descriptive Statistics is that branch of statistics which is concerned with describing the population under study.
Descriptive Statistics refers to a discipline that quantitatively describes the important characteristics of the dataset.
Charts, Graphs and Tables
Inferential Statistics is a type of statistics, that focuses on drawing conclusions about the population, on the basis of sample analysis and observation.
Inferential Statistics is all about generalizing from the sample to the population, i.e. the results of the analysis of the sample can be deduced to the larger population
Probability

Where are long-tailed distributions used?

A long-tailed distribution is a type of distribution where the tail drops off gradually toward the end of the curve.

The Pareto principle and the product sales distribution are good examples to denote the use of long-tailed distributions. Also, it is widely used in classification and regression problems.

What is the central limit theorem?

The central limit theorem (CLT) **states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution**. Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.
The **central limit theorem** states that if we take repeated random samples from a population and calculate the mean value of each sample, then the distribution of the sample means will be approximately normally distributed, *even if the population the samples came from is not normal*.

What is observational and experimental data in Statistics?

we simply observe what is happening and record the observations. So, it would be correct to say that researchers do not impose any kind of treatment or restriction to the group nor do they randomly assign the subjects to a group.
Experimental data is derived from experimental studies, where certain variables are held constant to see if any discrepancy is raised in the working.

What is meant by mean imputation for missing data? Why is it bad?

Mean imputation is a rarely used practice where null values in a dataset are replaced directly with the corresponding mean of the data.

It is considered a bad practice as it **completely removes the accountability for feature correlation**.

What is an outlier? How can outliers be determined in a dataset?

Outliers are data points that vary in a large way when compared to other observations in the dataset. Depending on the learning process, an outlier can worsen the accuracy of a model and decrease its efficiency sharply.

Outliers are determined by using two methods:

- Standard deviation/z-score
- Interquartile range (IQR)

How is missing data handled in statistics?

There are many ways to handle missing data in Statistics:

- Prediction of the missing values
- Assignment of individual (unique) values
- Deletion of rows, which have the missing data
- Mean imputation or median imputation

What is exploratory data analysis?

Exploratory data analysis is the process of performing investigations on data to understand the data better.

In this, initial investigations are done to determine patterns, spot abnormalities, test hypotheses, and also check if the assumptions are right.

What is the meaning of selection bias?

Selection bias is a phenomenon that involves the selection of individual or grouped data in a way that is not considered to be random. Randomization plays a key role in performing analysis and understanding model functionality better.

What are the types of selection bias in statistics?

There are many types of selection bias as shown below:

- Observer selection
- Attrition
- Protopathic bias
- Time intervals
- Sampling bias

What is the meaning of an inlier?

- An inlier is a data value that lies in the interior of a statistical distribution and is in error.

An inlier is a data point that lies at the same level as the rest of the dataset. Finding an inlier in the dataset is difficult when compared to an outlier as it requires external data to do so. Inliers, similar to outliers reduce model accuracy.

What is the probability of throwing two fair dice when the sum is 5 and 8?

There are 4 ways of rolling a 5 (1+4, 4+1, 2+3, 3+2):

P(Getting a 5) = 4/36 = 1/9

Now, there are 5 ways of rolling an 8 (, 2+6, 6+2, 3+5, 5+3, 4+4)

P(Getting an 8) = 5/36

State the case where the median is a better measure when compared to the mean.

In the case where there are a lot of outliers that can positively or negatively skew data, the median is preferred as it provides an accurate measure in this case of determination.

What type of data does not have a log-normal distribution or a Gaussian distribution?

Exponential distributions do not have a log-normal distribution or a Gaussian distribution. In fact, any type of data that is categorical will not have these distributions as well.

Example: Duration of a phone car, time until the next earthquake, etc.

What is the Pareto principle?

The Pareto principle is also called the 80/20 rule, which means that 80 percent of the results are obtained from 20 percent of the causes in an experiment.

What is the meaning of the five-number summary in Statistics?

The five-number summary is a measure of five entities that cover the entire range of data as shown below:

- Low extreme (Min)
- First quartile (Q1)
- Median
- Upper quartile (Q3)
- High extreme (Max)

What are population and sample in Inferential Statistics, and how are they different?

A population is a large volume of observations (data). The sample is a small portion of that population. Because of the large volume of data in the population, it raises the computational cost. The availability of all data points in the population is also an issue.

In short:

- We calculate the statistics using the sample.
- Using these sample statistics, we make conclusions about the population.

What are quantitative data and qualitative data?

- Quantitative data is also known as numeric data.
- Qualitative data is also known as categorical data.

What is Mean?

Mean is the average of a collection of values. We can calculate the mean by dividing the sum of all observations by the number of observations.

What is the meaning of standard deviation?

Standard deviation represents the magnitude of how far the data points are from the mean. A low value of standard deviation is an indication of the data being close to the mean, and a high value indicates that the data is spread to extreme ends, far away from the mean.

What is a bell-curve distribution?

A normal distribution can be called a bell-curve distribution. It gets its name from the bell curve shape that we get when we visualize the distribution.

What is skewness?

Skewness measures the lack of symmetry in a data distribution. It indicates that there are significant differences between the mean, the mode, and the median of data. Skewed data cannot be used to create a normal distribution.

What is kurtosis?

Kurtosis is used to describe the extreme values present in one tail of distribution versus the other. It is actually the measure of outliers present in the distribution. A high value of kurtosis represents large amounts of outliers being present in data. To overcome this, we have to either add more data into the dataset or remove the outliers.

What is correlation?

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It’s a common

tool for describing simple relationships without making a statement about cause and effect.

Correlation is used to test relationships between quantitative variables and categorical variables. Unlike covariance, correlation tells us how strong the relationship is between two variables. The value of correlation between two variables ranges from -1 to +1.

How is correlation measured?

The sample correlation coefficient, *r*, quantifies the strength of the relationship. Correlations are also tested for statistical significance.

What are some limitations of correlation analysis?

Correlation can't look at the presence or effect of other variables outside of the two being explored. Importantly, correlation doesn't tell us about cause and effect. Correlation also cannot accurately describe curvilinear relationships.

What do correlation numbers mean?

We describe correlations with a unit-free measure called the correlation coefficient which ranges from -1 to +1 and is denoted by *r*. Statistical significance is indicated with a p-value. Therefore, correlations are typically written with two key numbers: ***r*** = and ***p*** = .

- The closer *r* is to zero, the weaker the linear relationship.
- Positive *r* values indicate a positive correlation, where the values of both variables tend to increase together.
- Negative *r* values indicate a negative correlation, where the values of one variable tend to increase when the values of the other variable decrease.
- The p-value gives us evidence that we can meaningfully conclude that the population correlation coefficient is likely different from zero, based on what we observe from the sample.

What is a p-value?

A p-value is a measure of probability used for hypothesis testing.

What are left-skewed and right-skewed distributions?

A left-skewed distribution is one where the left tail is longer than that of the right tail. Here, it is important to note that the mean < median < mode.

Similarly, a right-skewed distribution is one where the right tail is longer than the left one. But, here mean > median > mode.

What are the types of sampling in Statistics?

There are four main types of data sampling as shown below:

- **Simple random**: Pure random division
- **Cluster**: Population divided into clusters
- **Stratified**: Data divided into unique groups
- **Systematical**: Picks up every 'n' member in the data

What is the meaning of covariance?

- Covariance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells you how a *single* variable varies, **co** variance tells you how **two** variables vary together.

Imagine that Jeremy took part in an examination. The test is having a mean score of 160, and it has a standard deviation of 15. If Jeremy's z-score is 1.20, what would be his score on the test?

To determine the solution to the problem, the following formula is used:

$X = \mu + Z\sigma$

Here:

μ: Mean

σ: Standard deviation

X: Value to be calculated

Therefore, X = 160 + (15*1.2) = 173.8 (Approximated to 174)

If a distribution is skewed to the right and has a median of 20, will the mean be greater than or less than 20?

If the given distribution is a right-skewed distribution, then the mean should be greater than 20, while the mode remains to be less than 20.

What is Bessel's correction?

Bessel's correction is a factor that is used to estimate a populations' standard deviation from its sample. It causes the standard deviation to be less biased, thereby, providing more accurate results.

The standard normal curve has a total area to be under one, and it is symmetric around zero. True or False?

True, a normal curve will have the area under unity and the symmetry around zero in any distribution. Here, all of the measures of central tendencies are equal to zero due to the symmetric nature of the standard normal curve.

In an observation, there is a high correlation between the time a person sleeps and the amount of productive work he does. What can be inferred from this?

First, correlation does not imply causation here. Correlation is only used to measure the relationship, which is linear between rest and productive work. If both vary rapidly, then it means that there is a high amount of correlation between them.

Confidence level vs Significance Level

Above, I defined a confidence level as answering the question: "…**if the poll/test/experiment was repeated (over and over), would the results be the same?**" In essence, confidence levels deal with repeatability. Significance levels on the other hand, have nothing at all to do with repeatability. They are set in the beginning of a specific type of experiment (a "hypothesis test"), and controlled by you, the researcher.

The significance level (also called the alpha level) is a term used to test a hypothesis. More specifically, it's the probability of making the wrong decision when the null hypothesis is true. In statistical speak, another way of saying this is that it's your probability of making a Type I error.

1. **Significance level:** In a hypothesis test, the significance level, alpha, is the probability of making the wrong decision when the null hypothesis is true.
2. **Confidence level:** The probability that if a poll/test/survey were repeated over and over again, the results obtained would be the same. A confidence level = 1 – alpha.
3. **Confidence interval:** A range of results from a poll, experiment, or survey that would be expected to contain the population parameter of interest. For

example, an average response. Confidence intervals are constructed using significance levels / confidence levels.

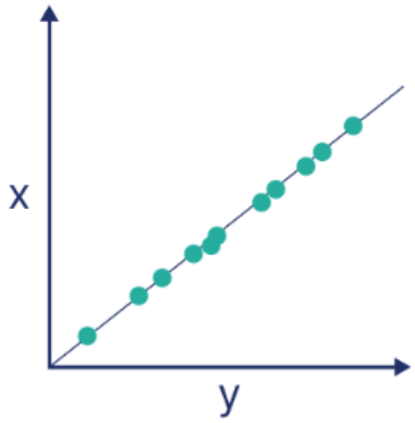
What types of variables are used for Pearson's correlation coefficient?

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

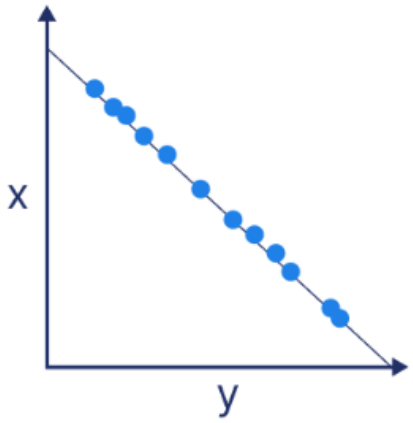
Perfect positive correlation

$r = 1$



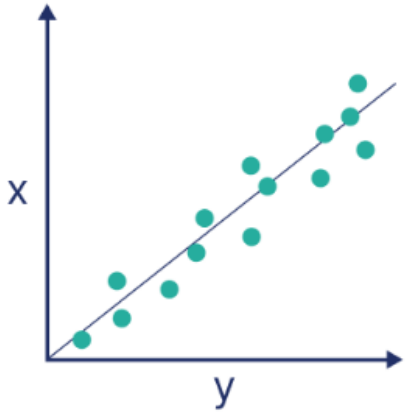
Perfect negative correlation

$r = 0$



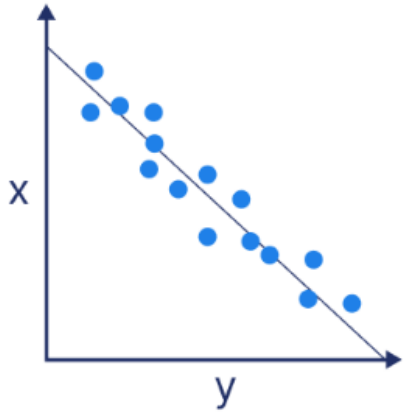
Strong positive correlation

$r > .5$



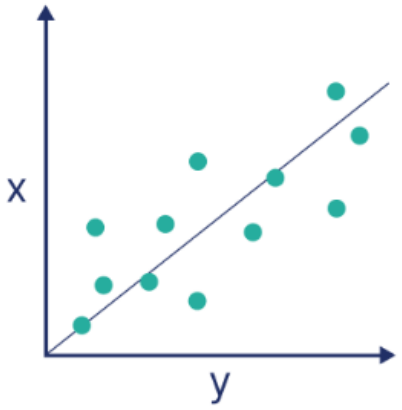
Strong negative correlation

$r < -.5$



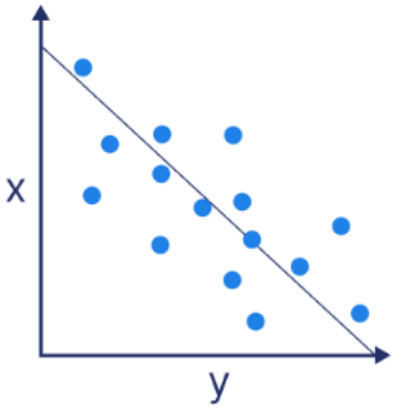
Weak positive correlation

$.3 > r > 0$



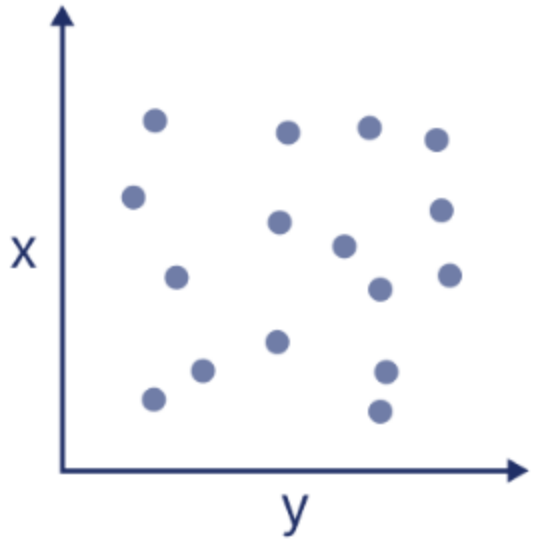
Weak negative correlation

$0 > r > -.3$



No correlation

$r = 0$



In a scatter diagram, what is the line that is drawn above or below the regression line called?

The line that is drawn above or below the regression line in a scatter diagram is called the residual or also the prediction error.

What are the examples of symmetric distribution?

Symmetric distribution means that the data on the left side of the median is the same as the one present on the right side of the median.

There are many examples of symmetric distribution, but the following three are the most widely used ones:

- Uniform distribution
- Binomial distribution
- Normal distribution

Where is inferential statistics used?

Inferential statistics is used for several purposes, such as research, in which we wish to draw conclusions about a population using some sample data. This is performed in a variety of fields, ranging from government operations to quality control and quality assurance teams in multinational corporations.

What is the relationship between mean and median in a normal distribution?

In a normal distribution, the mean is equal to the median. To know if the distribution of a dataset is normal, we can just check the dataset's mean and median.

What is the difference between the Ist quartile, the IInd quartile, and the IIIrd quartile?

Quartiles are used to describe the distribution of data by splitting data into three equal portions, and the boundary or edge of these portions are called quartiles.

That is,

- **The lower quartile (Q1)** is the 25th percentile.
- **The middle quartile (Q2)**, also called the median, is the 50th percentile.
- **The upper quartile (Q3)** is the 75th percentile.

How do the standard error and the margin of error relate?

The standard error and the margin of error are quite closely related to each other. In fact, the margin of error is calculated using the standard error. As the standard error increases, the margin of error also increases.

The **standard error** measures the preciseness of an estimate of a population mean. It is calculated as:

Standard Error = s / √n

where:

- **s:** Sample standard deviation
- **n:** Sample size

The **margin of error** measures the half-width of a confidence interval for a population mean. It is calculated as:

Margin of Error = z*(s/√n)

where:

- **z:** Z value that corresponds to a given confidence level
- **s:** Sample standard deviation
- **n:** Sample size

Let's check out an example to illustrate this idea.

Example: Margin of Error vs. Standard Error

Suppose we collect a random sample of turtles with the following information:

- Sample size **n = 25**
- Sample mean weight **x = 300**
- Sample standard deviation **s = 18.5**

Now suppose we'd like to create a 95% confidence interval for the true population mean weight of turtles. The formula to calculate this confidence interval is as follows:

Confidence Interval = x +/- z*(s/√n)

where:

- **x:** Sample mean
- **s:** Sample standard deviation
- **n:** Sample size
- **z:** Z value that corresponds to a given confidence level

The z-value that you will use is dependent on the confidence level that you choose. The following table shows the z-value that corresponds to popular confidence level choices:

Confidence Level	z-value
0.90	1.645
0.95	1.96
0.99	2.58

Notice that higher confidence levels correspond to larger z-values, which leads to wider confidence intervals. This means that, for example, a 99% confidence interval will be wider than a 95% confidence interval for the same set of data.

What is one sample t-test?

This T-test is a statistical hypothesis test in which we check if the mean of the sample data is statistically or significantly different from the population's mean.

What is an alternative hypothesis?

The alternative hypothesis (denoted by H1) is the statement that must be true if the null hypothesis is false. That is, it is a statement used to contradict the null hypothesis

Given a left-skewed distribution that has a median of 60, what conclusions can we draw about the mean and the mode of the data?

Given that it is a left-skewed distribution, the mean will be less than the median, i.e., less than 60, and the mode will be greater than 60.

What are the types of biases that we encounter while sampling?

Sampling biases are errors that occur when taking a small sample of data from a large population as the representation in statistical analysis. There are three types of biases:

- The selection bias
- The survivorship bias
- The undercoverage bias

What are the scenarios where outliers are kept in the data?

There are not many scenarios where outliers are kept in the data, but there are some important situations when they are kept. They are kept in the data for analysis if:

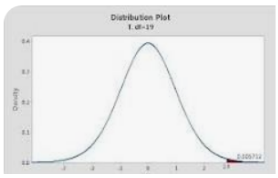
- Results are critical
- Outliers add meaning to the data
- The data is highly skewed

Briefly explain the procedure to measure the length of all sharks in the world.

Following steps can be used to determine the length of sharks:

- Define the confidence level (usually around 95%)
- Use sample sharks to measure
- Calculate the mean and standard deviation of the lengths
- Determine t-statistics values
- Determine the confidence interval in which the mean length lies
- The t-value **measures the size of the difference relative to the variation in your sample data**. Put another way, T is simply the calculated difference represented in units of standard error. The greater the magnitude of T, the greater the evidence against the null hypothesis. 04-Nov-2016

  https://blog.minitab.com › what-are-t-values-and-p-values...



How does the width of the confidence interval change with length?

The width of the confidence interval is used to determine the decision-making steps. As the confidence level increases, the width also increases.

The following also apply:

- Wide confidence interval: Useless information
- Narrow confidence interval: High-risk factor

What is the meaning of degrees of freedom (DF) in statistics?

Degrees of freedom or DF is used to define the number of options at hand when performing an analysis. It is mostly used with t-distribution and not with the z-distribution.

If there is an increase in DF, the t-distribution will reach closer to the normal distribution. If DF > 30, this means that the t-distribution at hand is having all of the characteristics of a normal distribution.

What are some of the properties of a normal distribution?

A normal distribution, regardless of its size, will have a bell-shaped curve that is symmetric along the axes.

Following are some of the important properties:

- Unimodal: It has only one mode.
- Symmetrical: Left and right halves of the curve are mirrored.
- Central tendency: The mean, median, and mode are at the midpoint.

What is the meaning of sensitivity in statistics?

Sensitivity, as the name suggests, is used to determine the accuracy of a classifier (logistic, random forest, etc.):

The simple formula to calculate sensitivity is:

Sensitivity = Predicted True Events/Total number of Events

What are some of the low and high-bias Machine Learning algorithms?

There are many low and high-bias Machine Learning algorithms, and the following are some of the widely used ones:

- **Low bias**: **SVM**, decision trees, KNN algorithm, etc.
- **High bias**: Linear and logistic regression

What are some of the techniques to reduce underfitting and overfitting during model training?

Underfitting refers to a situation where data has high bias and low variance, while overfitting is the situation where there are high variance and low bias.

Following are some of the techniques to reduce underfitting and overfitting:

For reducing underfitting:

- Increase model complexity
- Increase the number of features
- Remove noise from the data
- Increase the number of training epochs

For reducing overfitting:

- Increase training data
- Stop early while training
- Lasso regularization

Can you give an example to denote the working of the central limit theorem?

-
- For example, an economist may collect a simple random sample of 50 individuals in a town and use the average annual income of the individuals in the sample to estimate the average annual income of individuals in the entire town.
- If the economist finds that the average annual income of the individuals in the sample is $58,000, then her best guess for the true average annual income of individuals in the entire town will be $58,000.

What is the benefit of using box plots?

Box plots allow us to provide a graphical representation of the 5-number summary and can also be used to compare groups of histograms.

Does a symmetric distribution need to be unimodal?

A symmetric distribution does not need to be unimodal (having only one mode or one value that occurs most frequently). It can be bi-modal (having two values that

have the highest frequencies) or multi-modal (having multiple or more than two values that have the highest frequencies).

What is the impact of outliers in statistics?

Outliers in statistics have a very negative impact as they skew the result of any statistical query. For example, if we want to calculate the mean of a dataset that contains outliers, then the mean calculated will be different from the actual mean (i.e., the mean we will get once we remove the outliers).

When creating a statistical model, how do we detect overfitting?

Overfitting can be detected by cross-validation. In cross-validation, we divide the available data into multiple parts and iterate on the entire dataset. In each iteration, one part is used for testing, and others are used for training. This way, the entire dataset will be used for training and testing purposes, and we can detect if the data is being overfitted.

What is a survivorship bias?

The survivorship bias is the flaw of the sample selection that occurs when a dataset only considers the 'surviving' or existing observations and fails to consider those observations that have already ceased to exist.

What is an under coverage bias?

The under coverage bias is a bias that occurs when some members of the population are inadequately represented in the sample.

What is the relationship between standard deviation and standard variance?

Standard deviation is the square root of standard variance. Basically, standard deviation takes a look at how the data is spread out from the mean. On the other hand, standard variance is used to describe how much the data varies from the mean of the entire dataset.

What is expected value
For probability distribution,mean of distribution is called expected value.
What is fourth moment business decision?

The fourth moment business decision is measure of peakedness of distribution.Kurtosis.

Histogram is used for:
Select one:

- a. **Identify the Distribution of the Data**
- b. Identify the shape of the Distribution
- c. Identify the relationship between 2 variables
- d. Both 2 and 3

The number of Dimensions, Barplot is plotted on:

1 diamensional

b. 2 dimensional

- c. 3 dimensional
- d. 4 dimensional

The difference of first quartile and median is greater than the difference of median and third quartile then distribution is classified as
Select one:

- a. Symmetrical
-
- **c. Left Skewed**
- d. Not Skewed at All

b. Right Skewed

What percentage of data lies in IQR?
Select one:

-
- **c. 50%**
- d. 75%

b. 25%
a. 20%

What are the four main things we should know before studying data analysis?
Descriptive statistics
Inferential statistics

Distributions (normal distribution / sampling distribution)
Hypothesis testing

Most common characteristics used in descriptive statistics?

- Center – middle of the data. Mean / Median / Mode are the most commonly used as measures.
  - Mean – average of all the numbers
  - Median – the number in the middle
  - Mode – the number that occurs the most. The disadvantage of using Mode is that there may be more than one mode.
- Spread – How the data is dispersed. Range / IQR / Standard Deviation / Variance are the most commonly used as measures.
  - Range = Max – Min
  - Inter Quartile Range (IQR) = Q3 – Q1
  - Standard Deviation (σ) = $\sqrt{(\sum(x-\mu)^2 / n)}$
  - Variance = σ^2
- Shape – the shape of the data can be symmetric or skewed
  - Symmetric – the part of the distribution that is on the left side of the median is same as the part of the distribution that is on the right side of the median
  - Left skewed – the left tail is longer than the right side
  - Right skewed – the right tail is longer than the left side
- Outlier – An outlier is an abnormal value
-
  - Keep the outlier based on judgement
  - Remove the outlier based on judgement
- **How to calculate range and interquartile range?**
- IQR = Q3 – Q1
- Where, Q3 is the third quartile (75 percentile)
- Where, Q1 is the first quartile (25 percentile)

How to convert normal distribution to standard normal distribution?

- Standardized normal distribution has mean = 0 and standard deviation = 1
- To convert normal distribution to standard normal distribution we can use the formula
- X (standardized) = $(x-\mu) / \sigma$

Mention one method to find outliers?

- Shows the 5-number summary can be used to identify the outlier
- Widely used – Any data point that lies outside the 1.5 * IQR

What is the difference between population parameters and sample statistics?

- Population parameters are:
  - Mean = μ
  - Standard deviation = σ
- Sample statistics are:
  - Mean = x (bar)
  - Standard deviation = s
- **Why we need sample statistics?**
- Population parameters are usually unknown hence we need sample statistics.

What is p-value in hypothesis testing?

- If the p-value is more than then critical value, then we fail to reject the H0
  - If p-value = 0.015 (critical value = 0.05) – strong evidence
  - If p-value = 0.055 (critical value = 0.05) – weak evidence
- If the p-value is less than the critical value, then we reject the H0
  - If p-value = 0.055 (critical value = 0.05) – weak evidence
  - If p-value = 0.005 (critical value = 0.05) – strong evidence

What do you mean by Mode?

Ans:- The mode is defined as that element of the data sample, which appears most often in the collection.

Most common characteristics used in descriptive statistics?

Ans:- Center, spread, shape, and outlier are the most common characteristics used in descriptive statistics.

- The Center is in the middle of the data. Mean, Median and Mode are the most commonly used as measures.
- Spread how the data is dispersed. Rane, IQR, Variance, and Standard Deviation are the most commonly used as measures.
- Shape, the shape of the data can be symmetric or skewed.
- Outliner, an outlier is an abnormal value.

What do you mean by skewness?

Ans:- Skewness is described as data asymmetry, which is centered around a mean. If skewness is negative, the data is spread more on the left of the mean to the right. If skewness is seen as positive, then the data is moving more to the right.

What general conditions must be satisfied for the central limit theorem to hold?

Ans:- The data should be sampled randomly.
The sample values must be independent of each other.

The sample size should be sufficiently large, generally, it needs to be greater or equal than 30

Can you give some examples of where to use the mean or where to use the median?

Mean is the average value of a set of observations. It is used when data is normally distributed. Median is the middle value of a given set of observations. It is used when data is the skewed or long tail. For example, Income variable, income is highly skewed in the real world. If you use mean then the value of the mean will be dominated by the outliers.

What is kurtosis?

Kurtosis measures the thickness of the tail. The high value of kurtosis means heavy-tailed which indicates more outliers. The low value of kurtosis means less tailed which means less number of outliers in the observations.

What is the difference between sample distribution and sampling distribution?

Sample distribution is the distribution of all the values of the sample and sampling distribution displays all the values of possible samples from the population.

What is the advantage of using the standard normal distribution over the normal distribution?

Normal distribution fits into all kinds of real-life scenarios such as heights, exam scores, and blood pressure. The standard normal distribution is a specific distribution with a mean 0 and a standard deviation of 1. It is also known as the Gaussian distribution and the bell curve. Standardizing normal distribution makes it easier to compare with other metrics. It all boils down to the central limit theorem. The standard normal distribution uses Z values that can be easily compared and interpreted by a trained statistician.

Difference between parameter and static?

A Parameter is a number that describes the data from the *population* whereas, a *Statistic* is a number that describes the data from a *sample*.

Null Hypothesis: Null hypothesis is a statistical theory that suggests there is no statistical significance exists between the populations.

It is denoted by $\mathbf{H_0}$ and read as **H-naught**.

Alternative Hypothesis: An Alternative hypothesis suggests there is a significant difference between the population parameters.

Test Statistic: It is denoted by **t** and is dependent on the test that we run. It is deciding factor to reject or accept Null Hypothesis.

Type I error: Occurs when we reject a True Null Hypothesis and is denoted as α.

Type II error: Occurs when we accept a False Null Hypothesis and is denoted as β.