CONCEPTUAL QUESTIONS

For

BASIC STATISTICS

Website: www.analytixlabs.co.in

Email: info@analytixlabs.co.in

Q: How would you define statistics to a layman?

Ans: Statistics is a set of tools used to organize and analyze data. When referring to statistics, one usually means one or more of the following:

- A set of numerical data, such as the unemployment rate of a city, annual number of deaths due to bee stings, or the breakdown of the population of a city according to race in 2006 as compared to 1906
- Numbers or statistical values such as mean, which describes a sample of data, rather than the whole population
- Results of statistical procedures such as z-test or chi-square statistics
- A field of study using mathematical procedures to make inferences from data and decisions based on it

Q: How are statistical methods categorized?

Ans:There are two categories:

- **Descriptive statistics**: Descriptive statistics is the part of statistics that summarizes important characteristics of data sets. It is used to derive useful information from a set of numerical data.
- **Inferential statistics:**Inferential statistics consists of drawing conclusions, forecasts, judgments, or estimates about a larger data set, using the statistical characteristics of a sample of data.

Q: What are data frequency tables?

Ans: Data frequency tables are a method of representing a summary of data in a way in which identifying patterns or relationships becomes easier. By properly representing a data set in data frequency tables, we can unlock a lot of information about the data. It shows how many times something occurs in a given data set and how the data is distributed.

Q. What do you mean by Central tendency?

Ans: If all the homogeneous data plotted in graphically it is observed that data tends to cluster in one region which will have a midpoint which can denote the overall data. This characteristic is called as central tendency.

Q: How do you define location statistics?

Ans: Location statistics (the measure of a central tendency) is the central point of the data set. This statistic can be used to signify the representative (expected) value of a data set.

Mean (or average) is the numerical value of the center of a distribution and used when the data is concentrated)

$$\text{Mean} = \frac{X_1 + X_2 + X_3 + \ldots\ldots + X_n}{n}$$

Note: Mean is also referred to as $\overline{X}$.

The mean of a sample of data is typically denoted by $\overline{X}$, whereas the Greek letter μ (mu) is used to symbolize the mean of an entire population.

Mode is also average: Most occurring value (used for more discrete data OR categorical variables)

Median is also average: Exact middle point of the data distribution (used as a replacement for mean when the data has outliers).

Q: What are some of the key characteristics of a mean?

Ans: The following are key characteristics of a mean:

- Each data set, be it interval or ratio, has an arithmetic mean.
- Arithmetic mean computation includes all the data values.
- The arithmetic mean is unique.

Outliers (unusually large or small values) affect the calculated value of an arithmetic mean. The mean of 2, 5, 10, and 120 is 34.25, which does not provide a clear picture about the data. However, while calculating the arithmetic mean, all the observations are considered, which is a positive thing.

Q: What is a weighted mean?
Ans: When each observation disproportionately influences the mean, weighted mean is calculated. This can be caused either by outliers, or the data is structured in such a way that there are clusters within data groups that have varied characteristics.
For a set of numbers, the following equation is used to calculate weighted mean: where:

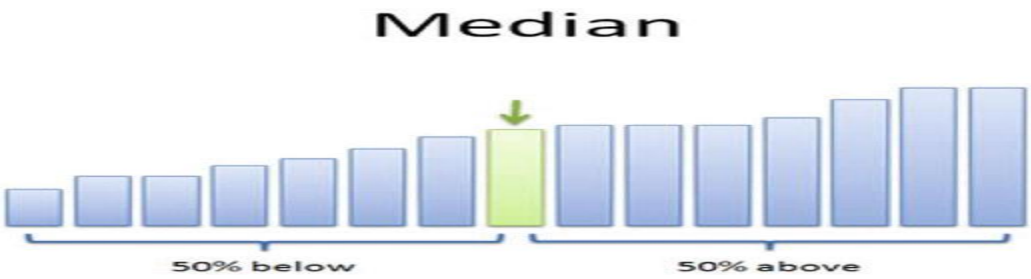
$$\overline{X}w = \sum_{i=1}^{n} w_i X_i = (w_1 X_1 + w_2 X_2 + \ldots\ldots\ldots + w_n X_n)$$

$$where:$$

$$X_1, X_2, \ldots\ldots\ldots X_n = observedvalues$$

$$w_1, w_2, \ldots\ldots\ldots w_n = correspondingweights$$

Q: What is a median? How do we determine the median of a data set?
Ans: Median (also known as the 50th percentile) is the middle observation in a data set. Median is calculated by sorting the data, followed by the selection of the middle value. The median of a data set having an odd number of observations is the observation number [N + 1] / 2. For data sets having an even number of observations, the median is midway between the observation numbers N / 2 and [N / 2] + 1. N is the number of observations.
As shown in Figure, the median is the middle observation point in a data set.

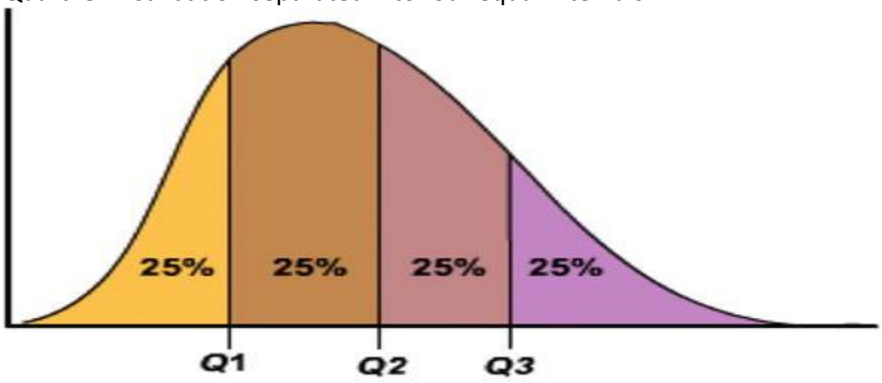


Q: When is a median used as opposed to a mean?
Ans: In the case of outliers (extreme values) or of a skewed data set (when one tail is significantly longer in a bell-shaped curve), the median is more applicable. If you want to represent the center of a distribution, such as in the case of the salaries of ten employees and one CEO, when the CEO has a significantly higher salary, using a median is more appropriate.

Q: What is meant by "mode"?
Ans: The value appearing most frequently in a data set is called the mode. A data set may have single or multiple modes, referred to as unimodal, bimodal, or trimodal, depending on the number of modes.

Q: Can you define quartiles, quintiles, deciles, and percentiles?
Ans: Quantile is the broad term for a value that divides a distribution into groups of equal size. For example:

- Quartile: Distribution separated into four equal intervals.



- Quintile: Distribution separated into five equal intervals
- Decile: Distribution separated into ten equal intervals
- Percentile: Distribution separated into 100 equal intervals (percent)

Q: Can you define standard deviation? Also, can you discuss briefly the notations used for it?
Ans: Standard deviation (StDev) is the measure of deviation of all observations in a distribution from the mean. It is also known as root-mean-square, for a sample data set calculated using the following formula.

$$S = \sqrt{\frac{\sum_{i=0}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}$$

S is used to signify the sample standard deviation, whereas the Greek letter s signifies the population standard deviation. The terms S or sˆ signify the estimated population standard deviation.

Q: What is the variance of a data set?
Ans: Variance, which is simply the standard deviation squared, represents dispersion of data. In other words, it is the average squared deviation from the mean.

$$S^2 = \sqrt{\frac{\sum_{i=0}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}$$

Note: Variance is additive in nature (standard deviations are not additive). The total variance is computed by adding individual variances.
Q. What is covariance?
Ans: Variance of one variable with respect to variance in other.

Q: What is unconditional probability?
Ans: Unconditional probability is the probability of an event irrespective of occurrences of other events in the past or future. For example, to find the probability of an economic recession, we calculate the unconditional probability of recession only, regardless of the changes in other factors, such as industrial output or consumer confidence.Unconditional probability is also known as marginal probability.

Q: What is conditional probability?
Ans: Conditional probability results when two events are related, such that one event occurs only when another event also occurs. Conditional probability is expressed as P(A|B); the vertical bar (|) indicates “given,” or “conditional upon.”

The probability of a recession when consumer confidence goes down is an example of conditional probability expressed as P (recession | decrease in consumer confidence). A conditional probability is also called its likelihood.

Q: What is the multiplication rule of probability?
Ans: Joint probability is the probability of two events happening together. The calculation of joint probability is based on the multiplication rule of probability and is expressed as
P(AB) = P(A|B) X P(B)

This is read as follows: the conditional probability of A given B, P(A|B), multiplied by the unconditional probability of B, P(B), is the joint probability of A and B.

Q: What is Bayes's theorem?
Ans: Bayes's theorem is one of the most common applications of conditional probability. A typical use of Bayes's theorem in the medical field is to calculate the probability of a person who tests positive on a screening test for a particular disease actually having the disease. Bayes's formula also uses several of the basic concepts of probability introduced previously and is, therefore, a good review for the entire chapter. Bayes's formula for any two events, A and B, is

$$P\left(A\middle|B\right) = \frac{P\left(A\&B\right)}{P(B)} = \frac{P\left(B\middle|A\right)P(A)}{P\left(B\middle|A\right)P(A) + P\left(B\middle|\sim A\right)P\left(\sim A\right)}$$

You would use this formula when you know P(B|A) but want to know P(A|B).

Q. Given two fair dices, what is the probability of getting scores that sum to 4 to 8?
Ans:

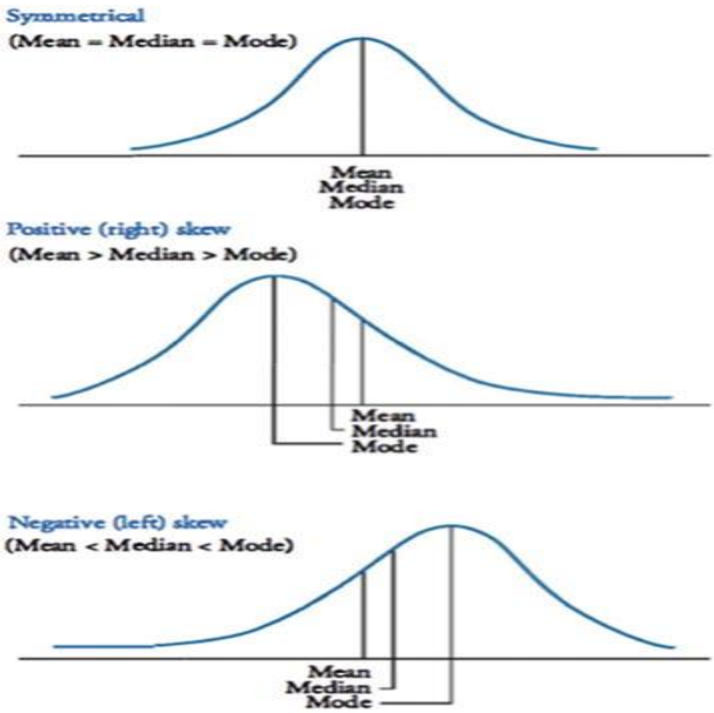
- Total: 36 combinations
- Of these, 3 involve a score of 4: (1,3), (3,1), (2,2)
- So: 3/36=1/12
- Considering a score of 8: (2,6), (3,5), (4,4), (6,2), (5,3)
- So: 5/36

Q. What is Normal distribution?
Ans: It is a continuous symmetric distribution for which Mean=Median=Mode. It is symmetric distribution that's why it is normal. Any distribution tends to normal if no. of observations tends to infinity.

Q. What is Poisson distribution?

Ans: In probability theory and statistics, the Poisson distribution (or Poisson law of small numbers) is a discrete probability distribution that expresses the probability of a number of event occurring in a fixed period of time if these events occur with a known average rate and independently of the timesince the last event.(The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.) Chancesofthefavorable outcome should be high (rare event).

Q: How would you define a symmetrical distribution?What is meant by the "skewness" of a distribution?
Ans: A symmetrical distribution has an identical shape on each side of its mean. This symmetry implies that on either side of the mean, the intervals will indicate the same frequency.
Skewness refers to the degree to which a distribution is not symmetrical. Skewness can be positive or negative, depending on the existence of outliers.



Q: What are outliers? How do they affect the skewness of a distribution?
Ans: Outliers are data points having extremely large positive or negative values, compared to the rest of the data points.
When the outliers lie in the upper region, or right tail, elongating this tail, they form a positively skewed distribution. When the outliers lie in the lower region, or left tail, elongating this tail, they form a negatively skewed distribution.

Q: How does skewness affect the location of the mean, median, and mode of a distribution?
Ans: For symmetrical distribution,

- Mode = Median = Mean
- For a positively skewed distribution,
- Mode < Median < Mean

In the case of a positively skewed distribution, the positive outliers pull the mean upward, or affect the mean positively. For example, a student scoring 0 on an exam, when all other students have scored above 50, pulls down the class average, making it skew negatively.

- For a negatively skewed distribution,
- Mean < Median < Mode

The negative outliers pull the median down, or to the left.
Q. Explain what a long-tailed distribution is and provide three examples of relevant phenomena that have long tails. Why are they important in classification and regression problems?
Ans:

- In long tailed distributions, a high frequency population is followed by a low frequency population, which gradually tails off asymptotically
- Rule of thumb: majority of occurrences (more than half, and when Pareto principles applies, 80%) are accounted for by the first 20% items in the distribution
- The least frequently occurring 80% of items are more important as a proportion of the total population
- Zipf's law, Pareto distribution, power laws

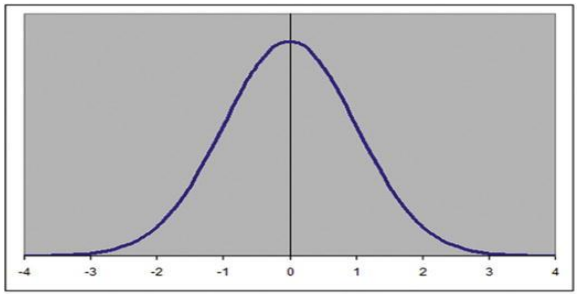
Examples:
1). Natural language
- Given some corpus of natural language - The frequency of any word is inversely proportional to its rank in the frequency table
- The most frequent word will occur twice as often as the second most frequent, three times as often as the third most frequent…
- "The" accounts for 7% of all word occurrences (70000 over 1 million)
- "of" accounts for 3.5%, followed by "and"…
- Only 135 vocabulary items are needed to account for half the English corpus!

2). Allocation of wealth among individuals: the larger portion of the wealth of any society is controlled by a smaller percentage of the people
3). File size distribution of Internet Traffic

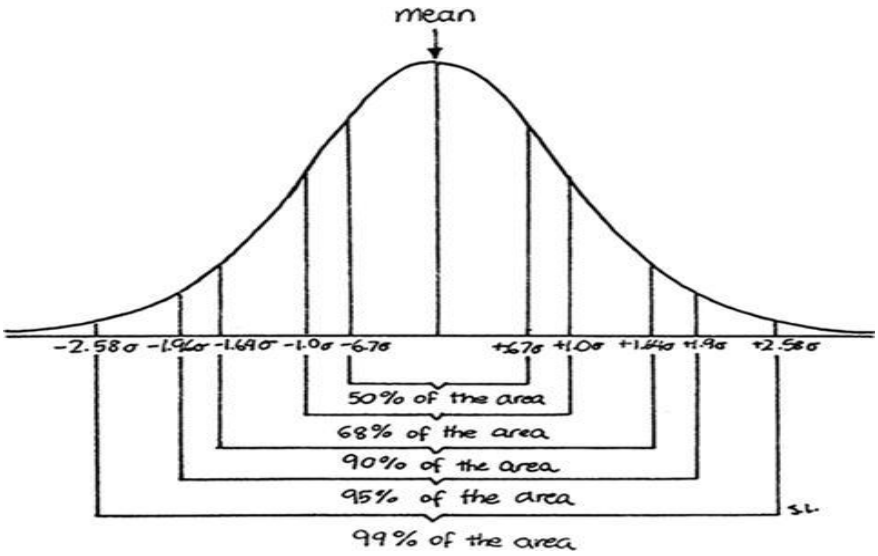
Additional: Hard disk error rates, values of oil reserves in a field (a few large fields, many small ones), sizes of sand particles, sizes of meteorites

Importance in classification and regression problems:
- Skewed distribution
- Which metrics to use? Accuracy paradox (classification), F-score, AUC
- Issue when using models that make assumptions on the linearity (linear regression): need to apply a monotone transformation on the data (logarithm, square root, sigmoid function…)
- Issue when sampling: your data becomes even more unbalanced! Using of stratified sampling of random sampling, SMOTE ("Synthetic Minority Over-sampling Technique") or anomaly detection approach

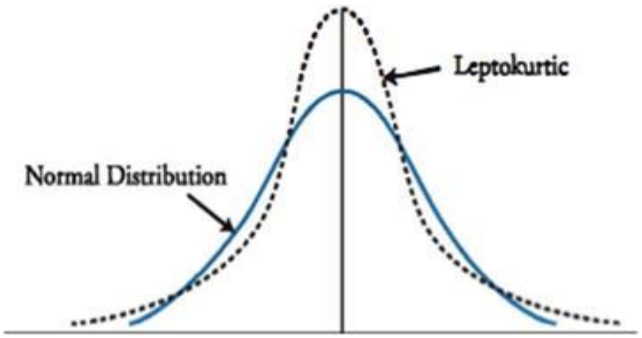
Q: What is a normal distribution?
Ans: This is a bell-shaped frequency distribution having the mean value in the middle of a demonstrable number of properties



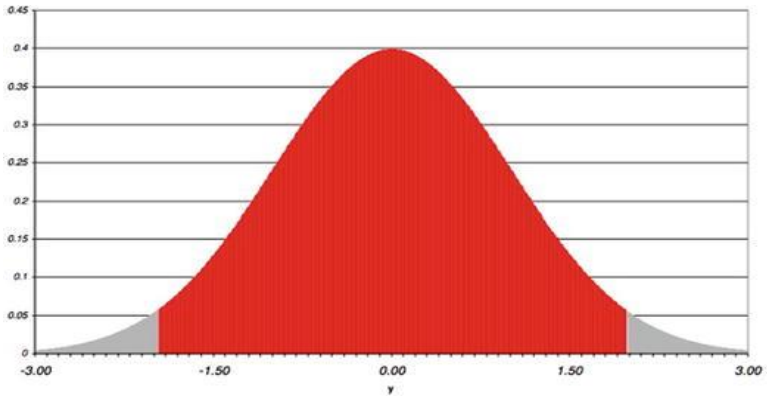
The total area of a normal curve is always 1, with the area on either side of the mean equal to 0.5. The area under the curve represents the probabilities.
An important attribute of the normal distribution curve is the way distribution is spread around the mean. We can see from Figure 3-5 that 68% of the population's values lie within $\mu \pm \sigma$, 95% of the population's values lie within $\mu \pm 1.96\sigma$, and 99% of population's values lie within $\mu \pm 2.58\sigma$.



Q: What is the kurtosis of a distribution? Also, what are the various types of kurtosis?
Ans: Kurtosis is a measure of how high or low a distribution is "peaked" compared to a normal distribution. A distribution that is peaked more than the normal curve is described as being leptokurtic, whereas a distribution flatter than the normal curve is called platykurtic. A distribution is mesokurtic if it has the same kurtosis as a normal distribution.



Q: What is a standard normal curve?
Ans: A standard normal curve is a normal curve with a mean of 0 and a standard deviation of 1. Practically, a standard normal curve is rare to come across, but we can transform or standardize a data set to make it a standard normal curve.



Q: What are some of the other continuous probability distributions?
Ans: Some commonly used continuous probability distributions are

- Normal Distribution
- Student's t-distribution
- Chi-square distribution
- F distribution
- Uniform distribution

Q: What is an F distribution?
Ans: An F distribution is used to test whether the ratios of two variances from normally distributed statistics are statistically different. In principle, this is the probability distribution associated with F statistics.

Q: What is a binomial probability distribution?
Ans: Experiments having only two outcomes—"success" or "failure"—such as flipping a coin to determine heads or tails, an out or not out in baseball or cricket, a person being dead or alive, a goal or no goal in soccer. These outcomes are labeled as "success" or "failure," or 1 or 0. Note that this carries no implication of "goodness."

Q. What is the difference between stratified and cluster sampling?
Ans:

a) Basically in a stratified sampling procedure, the population is first partitioned into disjoint classes (the strata) which together are exhaustive. Thus each population element should be within one and only one stratum. Then a simple random sample is taken from each stratum, the sampling effort may either be a proportional allocation (each simple random sample would contain an amount of variates from a stratum which is proportional to the size of that stratum) or according to optimal allocation, where the target is to have a final sample with the minimum variability possible.
b) The main difference between stratified and cluster sampling is that in stratified sampling all the strata need to be sampled. In cluster sampling one proceeds by firsts electing a number of clusters at random and then sampling each cluster or conduct a census of each cluster. But usually not all clusters would be included.

Q. Differences between observational and experimental data:
Ans:
Observational data: measures the characteristics of a population by studying individuals in a sample, but doesn't attempt to manipulate or influence the variables of interest
Example: find 100 women age 30 of which 50 have been smoking a pack a day for 10 years while the other have been smoke free for 10 years. Measure lung capacity for each of the 100 women. Analyze, interpret and draw conclusions from data.

Experimental data: applies a treatment to individuals and attempts to isolate the effects of the treatment on a response variable
Example: find 100 women age 20 who don't currently smoke. Randomly assign 50 of the 100 women to the smoking treatment and the other 50 to the no smoking treatment. Those in the smoking group smoke a pack a day for 10 years while those in the control group remain smoke free for 10 years. Measure lung capacity for each of the 100 women. Analyze, interpret and draw conclusions from data.

HYPOTHESIS TESTING

Hypothesis testing forms the next building block in learning statistical techniques. Now that you are familiar with probability distributions, the next step is to validate a data point or whether a sample falls into these distributions. Building a hypothesis is the first step in conducting an experiment or designing a survey. Don't just look on hypothesis testing as a statistical technique, but try to understand the core principles of this concept.

Q. What is sample and population?

Ans: Sample is a subset of the population which represents the population. It will have all characteristics of the population. All possible observations under study is called population.

Q. Specific characteristics of a sample?

Ans: It should be subset of population. It will have all characteristics of the population Different types of probability sampling.

- Simple random sampling
- Stratified sampling
- Cluster sampling
- Systematic sampling

Q. What is a cluster in sampling?

Ans:

a) While doing cluster sampling we divide the population into groups which are homogenous within the group and heterogeneous between the groups. These units are called cluster.
b) The technique works best when most of the variation in the population is within the groups, not between them.

Q. How do you determine the size of sample?

Ans:

- With experience
- Depending on requirement of the project
- Client request
- Time and cost budget
- Statistical formula for minimum sample size

Q. What is Central Limit theorem? Why is it important?

Ans: The CLT states that the arithmetic mean of a sufficiently large number of iterates of independent random variables will be approximately normally distributed regardless of the underlying distribution. i.e: the sampling distribution of the sample mean is normally distributed.

- Used in hypothesis testing
- Used for confidence intervals
- Random variables must be iid: independent and identically distributed
- Finite variance

Q: What is a hypothesis?

Ans: It is a supposition or assertion made about the world. A hypothesis is the starting point of an experiment, in which an assertion is made about some available data, and further investigation will be conducted to test if that assertion is correct or not.

a) A hypothesis is a proposed explanation for an observable phenomenon. For a hypothesis to be put forward as a scientific hypothesis, the scientific method requires that one can test it. Scientists generally

base scientific hypotheses on previous observations that cannot satisfactorily be explained with the available scientific theories. Even though the words "hypothesis" and “theory” are often used synonymously in common and informal usage, a scientific hypothesis is not the same as a scientific theory .A working hypothesis is a provisionally accepted hypothesis.

b) Ina related but distinguishable usage, the term hypothesis is used for the antecedent of a proposition; thus in proposition "If P, then Q ", P denotes the hypothesis (or antecedent); Q can be called a consequent. P is the assumption in a (possibly counterfactual) What If question.
c) The adjective hypothetical, meaning "having the nature of a hypothesis," or "being assumed to exist as an immediate consequence of a hypothesis," can refer to any of these meanings of the term "hypothesis."

Q: What is hypothesis testing?
Ans: It is the process in which statistical tests are used to check whether or not a hypothesis is true, using data. Based on hypothetical testing, we choose to accept or reject a hypothesis.
An example: Research is being conducted on the effect of TV viewing on obesity in children. A hypothesis for this would be that children viewing more than a certain amount of hours of television are obese. Data is then collected, and hypothesis testing is done to determine whether the hypothesis is correct or not.

Q: Why is hypothesis testing necessary?
Ans: When an event occurs, it can be the result of a trend, or it can occur by chance. To check whether the event is the result of a significant occurrence or merely of chance, hypothesis testing must be applied. In the preceding example of TV viewing and obesity, the hypothesis may be incorrect, and the data may show that it is merely chance that watching television makes some children obese.

Q: What are the criteria to consider when developing a good hypothesis?
Ans: A hypothesis is the initial part of a research study. If the hypothesis formed is incorrect, the research study is also likely to be incorrect; therefore, it should be properly considered and contemplated and should include the following criteria:

- The hypothesis should be logically consistent and make sense with regard to literature and language.
- The hypothesis should be testable. If a hypothesis cannot be tested, it has no use.
- It should be simple and clear, to avoid possible confusion.

Q: How is hypothesis testing performed?
Ans: There are several statistical tests available for hypothesis testing. The first step is to formulate a probability model based on the hypothesis. The probability model is also decided on the basis of the data available and the informed judgment of the researcher. Then, depending on the answers required, the appropriate statistical tests are selected.

Q: What are the various steps of hypothesis testing?
Ans: Hypothesis testing is conducted in four steps.

1. Identification of the hypothesis needed to be tested, for example, research to check the obesity in teenagers.
2. Selection of the criterion upon which a decision as to whether the hypothesis is correct or not is to be taken. For example, in the preceding problem, the criterion could be the body mass index (BMI) of the teenagers.
3. Determining from the random sample the statistics we are interested in. We select a random sample and calculate the mean. For example, a random sample of 1,000 teenagers is selected from a population, and their mean BMI is calculated.
4. Compare the result with the expected result, to check the validity. The discrepancy between the expected and real result helps to decide whether the claim is true or false.

Q. What is the Law of Large Numbers?

- A theorem that describes the result of performing the same experiment a large number of times
- Forms the basis of frequency-style thinking
- It says that the sample mean, the sample variance and the sample standard deviation converge to what they are trying to estimate
  Example: roll a dice, expected value is 3.5. For a large number of experiments, the average converges to 3.5

Q: What is the role of sample size in analytics?
Ans: Sample size for a statistical test is very important. Sample size is inversely proportional to standard error, i.e., the larger the sample size, the lesser the standard error and the greater the reliability. However, larger sample size means that a very small difference can become statistically significant, which may not be clinically or medically significant. The two main aspects of any study are generalizability (external validity) and validity (internal validity). Large samples have generalizability but not validity aspects.

Q. Why do we need weighting in market research?
Ans: To project characteristic with a sample for the population we have many constraints such as availability of proper sample. Like in population male female ratio is 50:50. but the sample shoes 80:20 then we need to put a weight to make it in same proportion.

Q. How do you calculate needed sample size?
Ans:Estimate a population mean:
- General formula is $ME = t \times \frac{S}{\sqrt{n}}$ or $ME = z \times \frac{s}{\sqrt{n}}$
- ME is the desired margin of error
- t is the t score or z score that we need to use to calculate our confidence interval
- s is the standard deviation
Example: we would like to start a study to estimate the average internet usage of households in one week for our business plan. How many households must we randomly select to be 95% sure that the sample mean is within 1minute from the true mean of the population? A previous survey of household usage has shown a standard deviation of 6.95 minutes.

- Z score corresponding to a 95% interval: 1.96 (97.5%, α/2=0.025)
- s=6.95
- n=(z×s/ME)^2=(1.96×6.95)^2=186

Estimate a proportion:
- Similar: $ME = z \times \sqrt{\frac{p(1-p)}{n}}$
Example: a professor in Harvard wants to determine the proportion of students who support gay marriage. She asks “how large a sample do I need?”
She wants a margin of error of less than 2.5%, she has found a previous survey which indicates a proportion of 30%.
n=(0.3×0.7)/0.025^2

Q: What is standard error?
Ans: The standard error (denoted by σ) is the standard deviation of a statistic. It reflects the variation caused by sampling. It is inversely proportional to sample size.

Q: What are null and alternate hypotheses?

Ans: A null hypothesis is the statement about a statistic in a population that is assumed to be true. It is the starting point of any research study. Based on statistical tests, a decision is taken as to whether the assumption is right or wrong.

An alternative hypothesis is the contradictory statement that states what is wrong with the null hypothesis.

We test the validity of a null hypothesis and not of an alternative hypothesis. An alternative hypothesis is accepted when the null hypothesis is proved to be wrong.

An example would be a study conducted to determine the mean height of a class of students. The researcher believes that the mean height is 170 centimeters (cms). In this case,

- H_0: $\mu_{height} = 170$ cms
- H_A: $\mu_{height} \neq 170$ cms

Q: Why are null and alternate hypotheses necessary?

Ans: Following are the reasons null and alternate hypotheses are necessary:

- The two hypotheses provide a rough explanation of the occurrences.
- They provide a statement to the researcher that acts as the base in a research study and is directly tested.
- They provide the outline for reporting the interpretations of the study.
- They behave as a working instrument of the theory.
- They verify whether or not the test is maintained and is detached from the investigator's individual standards and choices.

Q: How are the results of null/alternate hypotheses interpreted?

Ans: Statistical tests are conducted to check the validity of null hypotheses. When a null hypothesis is proved to be wrong, the alternate hypothesis is accepted. Consider, for example, a courtroom scenario. When a defendant is brought to trial, a null hypothesis is that he is innocent. The jury considers the evidence to decide whether or not the defendant is guilty.

In the preceding courtroom example, if there is insufficient evidence, the jury will free the defendant rather than convicting him or her. Similarly, in statistics, a null hypothesis is accepted if the research fails to prove otherwise, rather than endorsing an alternative hypothesis.

Q. What is level of confidence /Level of significance (i.e. what does 95% level significance stand for)?

Ans:

Level of significance is the criteria by which a decision is reached. In the courtroom example, the level of significance can be stated as the minimum level of evidence required by the jury to reach a verdict regarding the guilt or innocence of the defendant. Similarly, in statistics, it is the criterion by which a null hypothesis is rejected.

a) Statistics "significant" means probably true (not due to chance). A research finding may be true without being important. When statisticians say a result is "highly significant" they mean it is very probably true. They do not (necessarily) mean it is highly important.
b) The amount of evidence required to accept that an event is unlikely to have arisen by chance is known as the significance level or critical p-value: in traditional Fisherian statistical hypothesis testing, the p-value is the probability conditional on the null hypothesis of the observed data or more extreme data. If the obtained p-value is small then it can be said either the null hypothesis is false or an unusual event has occurred. It is worth stressing that p-values do not have any repeat sampling interpretation.
c) Where to establish the level of significance is determined by the alternative hypothesis. If the null hypothesis is true, the sample mean is equal to the mean population on average. If a = 5%, it means 95% of all the sample means lie within the range of $\mu \pm s$.

Let us consider an example in which the null hypothesis is that in the United States children watch three hours of TV. The level of significance is set at 95%. The other 5% value lies outside the range of μ±s.

The alternative hypothesis states that the children do not watch three hours of TV (either more or less).
If a sample has a mean of four hours, we will calculate the outcome by determining its likelihood.
We can see, then, how far the number of standard deviations for this result is from the mean. If the significance level is decided at 95%, and the distance from mean is more than one standard deviation from the mean, it implies that the null hypothesis is true.

Q. How do you assess the statistical significance of an insight?Is this insight just observed by chance or is it a real insight?
Ans: Statistical significance can be accessed using hypothesis testing:

- Stating a null hypothesis which is usually the opposite of what we wish to test (classifiers A and B perform equivalently, Treatment A is equal of treatment.
- Then, we choose a suitable statistical test and statistics used to reject the null hypothesis. Also, we choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)
- We calculate the observed test statistics from the data and check whether it lies in the critical region

Common tests:

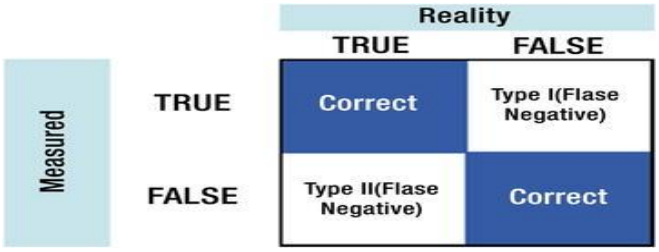
a). One sample Z test
b). Two-sample Z test
c). One sample t-test
d). Paired t-test
e). Two sample pooled equal variances t-test
f). Two sample unpooled unequal variances t-test and unequal sample sizes (Welch's t-test)
g). Chi-squared test for variances
h). Chi-squared test for goodness of fit
i). Anova (for instance: are the two regression models equal? F-test)
j). Regression F-test (i.e: is at least one of the predictor useful in predicting the response?)

Q. What is p-value?
Ans: The **P value**, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H_0) of a study question is true – the definition of 'extreme' depends on how the hypothesis is being tested. It is the least significant level at which you can accept the null hypothesis. Moreover, if p-value is high, you reject null hypothesis (based on p-value)

Q: What is test statistics?
Ans: Test statistics refers to a mathematical formula determining the likelihood of finding sample outcomes, if the null hypothesis is true, to make a decision regarding the null hypothesis.
If the level of significance is set at 95% and the test statistic value is less than 0.05, this would mean that the null hypothesis is wrong and should be rejected. Therefore, the researcher can take either of the following two decisions:

- Reject the null hypothesis, when the test statistic is less than a.
- Retain the null hypothesis, when the test statistic is greater than a.

Q: What are the different types of errors in hypothesis testing?
Ans: When we perform hypothesis testing, there can be errors such as falsely accepting or rejecting the null hypothesis. There are two types of distinguished errors: type I errors and type II errors.



A type I error occurs when a null hypothesis is incorrectly rejected and an alternate hypothesis is accepted. The type I error rate or significance level is denoted by a. It is generally set at 5%. In the courtroom example, if the judge convicts an innocent defendant, he/she is committing a type I error.

A type II error, or error of second kind, occurs when a null hypothesis is incorrectly accepted when the alternate hypothesis is true. If a type I error is a case of a false positive, a type II error is a case or a false negative. It is denoted by b and is related to the power of a test. In the example of the courtroom trial, if a judge lets a guilty defendant free, he is committing a type II error.

Q. What are one-tailed and two-tailed tests?

In a one-tailed test, the alternative hypothesis focuses on a specific direction (e.g., greater than or less than). In a two-tailed test, the alternative hypothesis is two-sided and considers deviations in both directions.

Q. What is the difference between a t-test and a z-test?

A t-test is used when the population standard deviation is unknown or when the sample size is small, while a z-test is used when the population standard deviation is known or when the sample size is large.

Q. How do you determine the sample size for a hypothesis test?

The sample size for a hypothesis test is determined based on factors such as the desired power of the test, significance level, effect size, and variability of the data.

Q. What is a confidence interval?

A confidence interval is a range of values calculated from sample data that is likely to contain the population parameter with a certain level of confidence. It provides a measure of the uncertainty associated with an estimate.

Q: What is meant by the statement "A result was said to be statistically significant at the 5% level."?
Ans: The result would be unexpected if the null hypothesis were true. In other words, we reject the null hypothesis.

Q: What are parametric and non-parametric tests?
Ans: In a parametric statistical test, assumptions such as that a population is normally distributed or has an equal-interval scale are made about the parameters (defining properties) of the population distribution. A non-parametric test is one that makes no such assumptions.

Q. When do you use t-test and what are assumptions there?
Ans: To test the small sample mean… whether it is equal to population mean or not

Q. When can you do T-test instead of Z-test?
Ans: When sample size is less than 30, we do t-test.

Q: What differentiates a paired vs. an unpaired test?
Ans: When we are comparing two groups, we have to decide whether to perform a paired test . A repeated-measures test, as it is called, is used when comparing three or more groups.
When the individual values are unpaired or matched or related among one another between groups, we use an unpaired test. In cases in which before and after effects of a study are required, a paired or repeated-measures test is used. In the case of measurements on matched/paired subjects, or in one of repeated lab experiments at dissimilar times, each with its own control, paired or repeated-measures tests are also used.
Paired tests are selected for closely correlated groups. The pairing can't be based on the data being analyzed, but before the data were collected, when the subjects were matched or paired.

Q. When do you use paired t-test?
Ans:

a) To test the difference between two dependent samples (basically one sample).
   Eg: before treatment and after treatment of a particular disease.

Q. What is Chi-square test?
Ans: Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. Chi-square tests are mainly of two types:

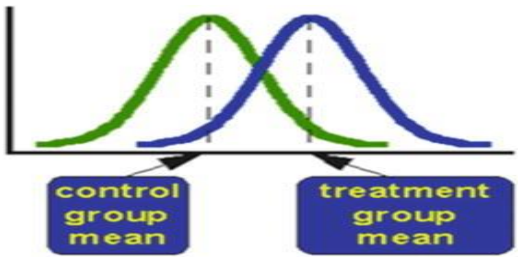
i. Goodness of fit.
ii. Independence of attributes.

A chi-square (χ^2) test is used to examine if two distributions of categorical variables are significantly different from each other. Categorical variables are the variables in which the value is in a category and not continuous, such as yes and no or high, low, and medium or red, green, yellow, and blue. Variables such as age and grade-point average (GPA) are numerical, meaning they can be continuous or discrete. The hypothesis for a χ^2 test follows:

- H_0: There is no association between the variables.
- H_A: There is association between them.

The type of association is not specified by the alternative hypothesis though. So interpretation of the test requires closer attention to the data.

Q: What is a t-test?
Ans: A t-test is a popular statistical test to draw inferences about single means or about two means or variances, to check if the two groups' means are statistically different from each other, where $n < 30$ and the standard deviation is unknown.
The means of the control and treatment group will most likely be located at different positions. The t-test checks if the means are statistically different for the two groups.



The t-test judges the difference between the means relative to the spread or variability of the scores of the two groups.

Q: What is a one-sample t-test?
Ans: A one-sample t-test compares the mean of a sample to a given value, usually the population mean or a standard value. Basically, it compares the observed average (sample average) with the expected average (population average or standard value), adjusting the value of the number of cases and the standard deviation.

Q: What is a two-sample t-test?
Ans: The purpose of the two-sample t-test is to determine if two population means are significantly different. The test is also known as the independent samples t-test, since the two samples are not related to each other and can therefore be used to implement a between-subjects design. In addition to the assumption of independence, both distributions must be normal, and the population variances must be equal (i.e., homogeneous).

Q: What is a paired-sample t-test?
Ans: The purpose of the repeated-measures t-test (or paired-sample t-test) is to test the same experimental units under different treatment conditions—usually experimental and control—to determine the treatment effect, allowing units to act as their own controls. This is also known as the dependent samples t-test, because the two samples are related to each other, thus implementing a within-subjects design. The other requirement is that sample sizes be equal, which is not the case for a two-sample t-test.
Q: Briefly, what are some issues related to t-tests?
Ans: The biggest issue with t-tests results from the confusion of its application as opposed to the z-test. Both statistical tests are used for almost the same purpose, except for a slight difference; the difference being when to use which test. When a sample is large ($n \geq 30$), and whether the population standard deviation is known or not, a z-test is used. For a limited sample ($n < 30$), when the standard deviation of the population is unknown, a t-test is chosen.

Q. What is ANOVA?
Ans:

a) Analysis of variance between two or more variables. There are different types of ANOVA like One-way, two-way, multiple. The ANOVA is based on the fact that two independent estimates of the population variance can be obtained from the sample data. A ratio is formed for the two estimates, where:
b) One is sensitive to® treatment effect & error between groups estimate and the other to ®error within groups estimate
c) Given the null hypothesis (in this case $H_O: m_1=m_2=m_3$), the two variance estimates should be equal. That is, since the null assumes no treatment effect, both variance estimates reflect error and their ratio will equal1. To the extent that this ratio is larger than1, it suggests a treatment effect (i.e., differences between the groups).

Q: What is understood by one-way analysis of variance?

Ans: The one-way analysis of variance (ANOVA) test is used to determine whether the mean of more than two groups of a data set are significantly different from each other.
Imagine, for example, that we are conducting a BOGO (buy one get one) campaign involving five groups of a hundred customers each. Each group is different in terms of its demographic attributes. We would like to determine whether these five groups respond differently to the campaign. This would help us to optimize the right campaign for the right demographic group, increase the response rate, and reduce the cost of campaign.

Q: In a nutshell, how does the ANOVA technique work?
Ans: The "analysis of variance" works by comparing the variance between the groups to that within the group variance. The core of this technique lies in assessing whether all the groups are, in fact, part of one larger population or a completely different population with different characteristics.

Q: What is the null hypothesis that ANOVA tests?
Ans: The null hypothesis is $H_o = \mu_o = \mu_1 = \mu_2 = \mu_3 = \ldots\ldots\ldots = \mu_k$
Where μ = the group mean and $_k$ = number of groups.
In a null hypothesis, the means of all groups are equal to one another. If there are even two groups with significantly different means, then we accept an alternate hypothesis.

Q. What is statistical power?

- sensitivity of a binary hypothesis test
- Probability that the test correctly rejects the null hypothesis H0 when the alternative is true H1
- Ability of a test to detect an effect, if the effect actually exists
- Power=P(rejectH0|H1istrue)
- As power increases, chances of Type II error (false negative) decrease
- Used in the design of experiments, to calculate the minimum sample size required so that one can reasonably detects an effect. i.e: "how many times do I need to flip a coin to conclude it is biased?"

Q. Explain selection bias (with regard to a dataset, not variable selection). Why is it important? How can data management procedures such as missing data handling make it worse?
Ans: Selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved.

Types:
Sampling bias: systematic error due to a non-random sample of a population causing some members to be less likely to be included than others

Time interval: a trial may terminated early at an extreme value (ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all the variables have similar means

Data: "cherry picking", when specific subsets of the data are chosen to support a conclusion (citing examples of plane crashes as evidence of airline flight being unsafe, while the far more common examples of flights that are Complete safely)

Studies: performing experiments and reporting only the most favourable results
can lead to inaccurate or even erroneous conclusions
Statistical methods can generally not overcome it

Why data handling makes it worse?
Example: individuals who know or suspect that they are HIV positive are less likely to participate in HIV surveys
Missing data handling will increase this effect as it's based on most HIV negative
Prevalence estimates will be inaccurate

Q. Is mean imputation of missing data acceptable practice? Why or why not?
Ans:

- Bad practice in general
- If just estimating means: mean imputation preserves the mean of the observed data
- Leads to an underestimate of the standard deviation
- Distorts relationships between variables by "pulling" estimates of the correlation toward zero

Q. Provide a simple example of how an experimental design can help answer a question about behaviour. How does experimental data contrast with observational data?
Ans:

- You are researching the effect of music-listening on studying efficiency
- You might divide your subjects into two groups: one would listen to music and the other (control group) wouldn't listen anything!
- You give them a test
- Then, you compare grades between the two groups

8. What is an outlier? Explain how you might screen for outliers and what would you do if you found them in your dataset. Also, explain what an inlier is and how you might screen for them and what would you do if you found them in your dataset
Ans:
Outliers: An observation point that is distant from other observations
Can occur by chance in any distribution
Often, they indicate measurement error or a heavy-tailed distribution
Measurement error: discard them or use robust statistics
Heavy-tailed distribution: high skewness, can't use tools assuming a normal distribution
Three-sigma rules (normally distributed data): 1 in 22 observations will differ by twice the standard deviation from the mean
Three-sigma rules: 1 in 370 observations will differ by three times the standard deviation from the mean
Three-sigma rules example: in a sample of 1000 observations, the presence of up to 5 observations deviating from the mean by more than three times the standard deviation is within the range of what can be expected, being less than twice the expected number and hence within 1 standard deviation of the expected number (Poisson distribution).

If the nature of the distribution is known a priori, it is possible to see if the number of outliers deviates significantly from what can be expected. For a given cutoff (samples fall beyond the cutoff with probability p), the number of outliers can be approximated with a Poisson distribution with lambda=pn. Example: if one takes a normal distribution with a cutoff 3 standard deviations from the mean, p=0.3% and thus we can approximate the number of samples whose deviation exceed 3 sigma's by a Poisson with lambda=3

Identifying outliers:
No rigid mathematical method
Subjective exercise: be careful
Boxplots
QQ plots (sample quantiles Vs theoretical quantiles)

Handling outliers:
Depends on the cause
Retention: when the underlying model is confidently known

Regression problems: only exclude points which exhibit a large degree of influence on the estimated coefficients (Cook's distance)

Inlier:
Observation lying within the general distribution of other observed values
Doesn't perturb the results but are non-conforming and unusual
Simple example: observation recorded in the wrong unit (°F instead of °C)

Identifying inliers:
Mahalanobi's distance
Used to calculate the distance between two random vectors
Difference with Euclidean distance: accounts for correlations
Discard them

Q. How do you handle missing data? What imputation techniques do you recommend?
Ans:

- If data missing at random: deletion has no bias effect, but decreases the power of the analysis by decreasing the effective sample size
- Recommended: Knn imputation, Gaussian mixture imputation

Q. You have data on the durations of calls to a call centre. Generate a plan for how you would code and analyze these data. Explain a plausible scenario for what the distribution of these durations might look like. How could you test, even graphically, whether your expectations are borne out?
Ans:

1. Exploratory data analysis
    - Histogram of durations
    - Histogram of durations per service type, per day of week, per hours of day (durations can be systematically longer from 10am to 1pm for instance), per employee
2. Distribution: lognormal?
3. Test graphically with QQ plot: sample quantiles of log(durations) Vs normal quantiles

Q. You are compiling a report for user content uploaded every month and notice a spike in uploads in October. In particular, a spike in picture uploads. What might you think is the cause of this, and how would you test it?
Ans

- Halloween pictures?
- Look at uploads in countries that don't observe Halloween as a sort of counter-factual analysis
- Compare uploads mean in October and uploads means with September: hypothesis testing

Q. You're about to get on a plane to Seattle. You want to know if you should bring an umbrella. You call 3 random friends of yours who live there and ask each independently if it's raining. Each of your friends has a 2/3 chance of telling you the truth and a 1/3 chance of messing with you by lying. All 3 friends tell you that "Yes" it is raining. What is the probability that it's actually raining in Seattle?
Ans:

- All say yes: all three lie or three say the truth
- $P(\text{" all say the truth "}) = (\frac{2}{3})^3 = \frac{8}{27}$
- $P(\text{" all lie "}) = (\frac{1}{3})^3 = \frac{1}{27}$
- $P(\text{" all yes "}) = \frac{1}{27} + \frac{8}{27} = \frac{1}{3}$
- Out of these numbers, there is $\frac{\frac{8}{27}}{\frac{1}{3}} = \frac{8}{9}$ chance it's actually raining

Q. There's one box - has 12 black and 12 red cards, 2nd box has 24 black and 24 red; if you want to draw 2 cards at random from one of the 2 boxes, which box has the higher probability of getting the same color? Can you tell intuitively why the 2nd box has a higher probability

Ans:

- First select: for both, then and ; compare them
- B/A=529/517

Q. When you sample, what bias are you inflicting?

Ans:

Selection bias: An online survey about computer use is likely to attract people more interested in technology than in typical

Under coverage bias: Sample too few observations from a segment of population

Survivorship bias: Observations at the end of the study are a non-random set of those present at the beginning of the investigation

In finance and economics: the tendency for failed companies to be excluded from performance studies because they no longer exist

Q. How do you control for biases?

Ans:

- Choose a representative sample, preferably by a random method
- Choose an adequate size of sample
- Identify all confounding factors if possible
- Identify sources of bias and include them as additional predictors in statistical analyses
- Use randomization: by randomly recruiting or assigning subjects in a study, all our experimental groups have an equal chance of being influenced by the same bias

Notes:

Randomization: in randomized control trials, research participants are assigned by chance, rather than by choice to either the experimental group or the control group.

Random sampling: obtaining data that is representative of the population of interest

Q. What are confounding variables?

Ans:

- Extraneous variable in a statistical model that correlates directly or inversely with both the dependent and the independent variable
- A spurious relationship is a perceived relationship between an independent variable and a dependent variable that has been estimated incorrectly
- The estimate fails to account for the confounding factor
- See Question 18 about root cause analysis

Q. What is A/B testing?

Ans:

- Two-sample hypothesis testing
- Randomized experiments with two variants: A and B

- A: control; B: variation
- User-experience design: identify changes to web pages that increase clicks on a banner
- Current website: control; NULL hypothesis
- New version: variation; alternative hypothesis

Q. An HIV test has a sensitivity of 99.7% and a specificity of 98.5%. A subject from a population of prevalence 0.1% receives a positive test result. What is the precision of the test (i.e the probability he is HIV positive)?
Ans:

Bayes rule: $P(Actu_+|Pred_+) = \frac{P(Pred_+|Actu_+)\times P(Actu_+)}{P(Pred_+|Actu_+)\times P(Actu_+)+P(Pred_+|Actu_-)P(Actu_-)}$

We have: $\frac{sensitivity\times prevalence}{sensitivity\times prevalence+(1-specificity)\times(1-prevalence)} = \frac{0.997\times 0.001}{0.997\times 0.001+0.15\times 0.999} = 0.62$

Q. What is difference between likelihood function and joint distribution?
Ans: Joint distribution is functiona function of parameters whereas likelihood function of sample observations.

Q. Explain likely differences between administrative datasets and datasets gathered from experimental studies. What are likely problems encountered with administrative data? How do experimental methods help alleviate these problems? What problem do they bring?
Ans:
Advantages:
- Cost
- Large coverage of population
- Captures individuals who may not respond to surveys
- Regularly updated, allow consistent time-series to be built-up

Disadvantages:
- Restricted to data collected for administrative purposes (limited to administrative definitions. For instance: incomes of a married couple, not individuals, which can be more useful)
- Lack of researcher control over content
- Missing or erroneous entries
- Quality issues (addresses may not be updated or a postal code is provided only)
- Data privacy issues
- Underdeveloped theories and methods (sampling methods…)

CORRELATION

This topic is concerned with measuring the relatedness between two variables. A simple measure, the correlation coefficient, is commonly used to quantify the degree of relationship between two variables

Q: What is correlation and what does it do?
Ans: In analytics, we try to find relationships and associations among various events. In the probabilistic context, we determine the relationships between variables. Correlation is a method by which to calculate the relationship between two variables. The coefficient of a correlation is a numerical measure of the relationship between paired observations (Xi,Yi), i = 1,…, n. For different coefficients of correlations, the relationship between variables and their interpretation varies.

There are a number of techniques that have been developed to quantify the association between variables of different scales (nominal, ordinal, interval, and ratio), including the following:

- Pearson product-moment correlation (both variables are measured on an interval or ratio scale)
- Spearman rank-order correlation (both variables are measured on an ordinal scale)
- Phi correlation (both variables are measured on a nominal/dichotomous scale)
- Point bi-serial (one variable is measured on a nominal/dichotomous scale, and one is measured on an interval or ratio scale)

Q: When should correlation be used or not used?

Ans: Correlation is a good indicator of how two variables are related. It is a good metric to look at during the early phases of research or analysis. Beyond a certain point, however, correlation is of little use.

A dip in a country's gross domestic product (GDP), for example, would lead to an increase in the unemployment rate. A casual look at the correlation between these two variables would indicate that there is a strong relationship between them.

Yet, the extent or measure of this relationship cannot be ascertained through a correlation. A correlation of 75% between two variables would not mean that the two variables are related to each other by a measure of .75 times.

This brings us to another issue that is often misunderstood by analysts. Correlation does not mean causation. A strong correlation between two variables does not necessarily imply that one variable causes another variable to occur.

In our GDP vs. unemployment rate example, this might be true, i.e., a lower GDP rate might increase unemployment. But we cannot and should not infer this from a correlation. It should be left to the sound judgment of a competent researcher.

Q: What is the Pearson product-moment correlation coefficient?

Ans: It is easy to determine whether two variables are correlated, simply by looking at the scatter plot (each variable on the 2 axis of the plot). Essentially, the values should be scattered across a straight line of the plot, in order for the two variables to have a strong correlation.

However, to quantify the correlation, we use the Pearson's product-moment correlation coefficient for samples, otherwise known as Pearson's r.

The correlation can be either positive or negative.

So, the value of ρ can be any number between -1 to 1 ([-1 <= ρ <= +1])

A correlation coefficient of less than zero would mean that the increase of one variable generally leads to the decrease of the other variable. A coefficient greater than zero implies that an increase of one variable leads to an increase in the other variable.

Higher values mean stronger relationships (positive or negative), and values closer to zero depict weak relationships. A correlation of 1.00 means that the two values are completely or perfectly positively correlated; -1.00 means that they are perfectly negatively correlated; and a correlation of 0.00 means that there is no relationship between the two variables.

Q: What is the formula for calculating the correlation coefficient?

Ans: The formula is as follows:

$$r = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$$

To compute r, the following algorithm, corresponding to the preceding formula, is used as follows:

- For each (x, y) set of coordinates, subtract the mean from each observation for x and y.
- Divide by the corresponding standard deviation.
- Multiply the two results together.
- The result is then added to a sum.
- The sum is divided by the degrees of freedom, n - 1.

Q: Briefly, what are the other techniques for calculating correlation?

Ans: Although the Pearson product-moment correlation is the most widely used correlation technique, other correlation techniques must be applied if the main tenet of the Pearson technique is violated, i.e., both variables should be on an interval or ratio scale.

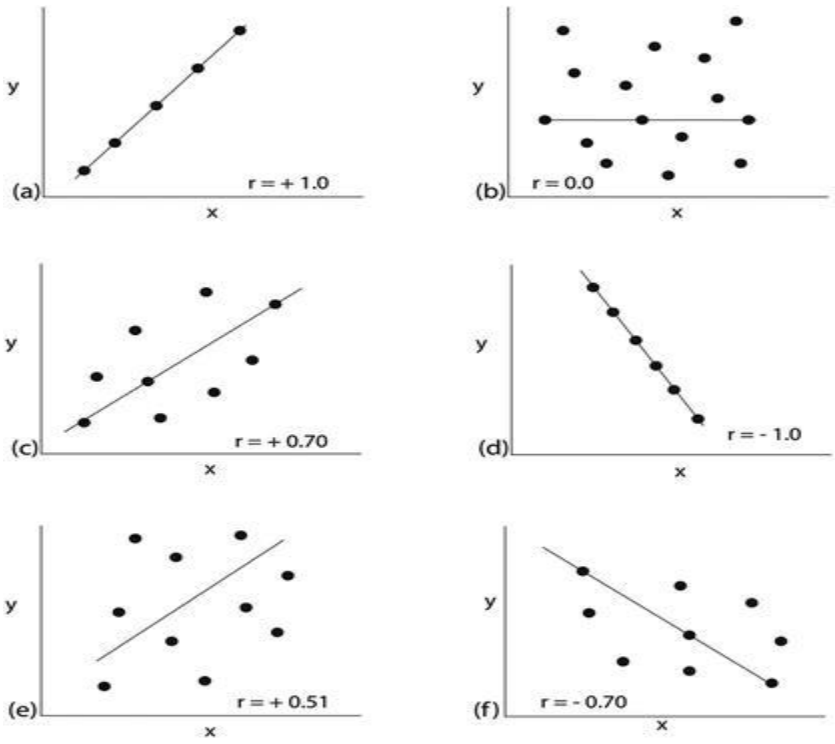
SPEARMAN RANK-ORDER CORRELATION: This technique is used when both variables are on an ordinal scale. Both variables are ranked manually between values of 1 to 10. The main tenet here is that the rank for a particular observation should be high for both variables, for any correlation to exist.

PHI CORRELATION: This is also used as an after test for a chi-square test. It is used when variables are nominal.

POINT BISERIAL: This method is used when one variable is on a nominal/dichotomous scale and one is measured on an interval or ratio scale.

Q: How would you use a graph to illustrate and interpret a correlation coefficient?

Ans: Visualization helps us to understand the relationships among variables better. In Figure, graphs have been plotted to illustrate and visualize relationships between the variables x and y that are being statistically analyzed.



Graphs plotting the statistical analysis of relationships between the variables x and y

Now let us interpret these graphs.

Figure (a): r = 1.0

This represents a perfect linear association. All the data points fall on the line.

Figure (b): r = 0

No linear relationship exists between the variables. The data points are scattered randomly and may approximate a circle. Changing the value of one variable has no effect on the value of the other.

Figure (c): r = 0.70

There is some positive linear relationship, although it is not perfect. Most of the data points fall on or closer to a straight line.

Figure (d): r = -1.0

This shows a perfect linear relationship between the variables, similar to figure (a), with the difference being that the variables are inversely related, i.e., increasing the value of one variable results in a decrease in the other variable.

Figure (e): r = 0.51

The relationship between the variables is not very strong, and the data points are a little scattered, although still closer to a straight line.

Figure (f): r = -0.70

This is similar to figure (c), with the difference being that the variables are negatively correlated.

We can see that as the value of r decreases, the data points are more scattered, whereas the data points are closer to a straight line when the value of r approaches -1.0 or +1.0.

Q: What are the various issues with correlation?

Ans: The various issues with correlation are

- Correlation analysis does not measure the strength of a nonlinear association between variables.
- Accidental or spurious relationships are not accounted for.
- Research problems, such as data contamination, sample bias, etc., hinder drawing reliable conclusions.
- Correlation analysis measures the relationship and does not provide an explanation or basis for it, which can result in false conclusions.

Q: How would you calculate a correlation coefficient in Excel?

Ans: A correlation coefficient between two arrays can be calculated in Excel using the CORREL function. The syntax for using the CORREL function is CORREL (array1, array2).

For eg: CORREL (A1:A20, B1:B20)

The CORREL function gives error in the case of the following:

- A #N/A error for unequal numbers of data points in the two arrays
- A #DIV/0! error when the standard deviation of array1 and array2 is zero

Q. What are the different kinds of multivariate analyses?

Ans:

- Factor analysis.
- Cluster analysis
- Conjoint analysis
- Discriminant analysis
- Logistic regression
- Multiple regressions

Q. Define: quality assurance, six sigma.

Ans:

Quality assurance:

- A way of preventing mistakes or defects in manufacturing products or when delivering services to customers
- In a machine learning context: anomaly detection

Six sigma:

- Set of techniques and tools for process improvement
- 99.99966% of products are defect-free products (3.4 per 1 million)
- 6 standard deviation from the process mean

Q. Give examples of data that does not have a Gaussian distribution, nor log-normal.

Ans:

- Allocation of wealth among individuals
- Values of oil reserves among oil fields (many small ones, a small number of large ones)

Q. What is root cause analysis? How to identify a cause vs. a correlation? Give examples

Ans:

Root cause analysis:

- Method of problem solving used for identifying the root causes or faults of a problem
- A factor is considered a root cause if removal of it prevents the final undesirable event from recurring

Identify a cause vs. a correlation:
- **Correlation:** statistical measure that describes the size and direction of a relationship between two or more variables. A correlation between two variables doesn't imply that the change in one variable is the cause of the change in the values of the other variable
- **Causation:** indicates that one event is the result of the occurrence of the other event; there is a causal relationship between the two events
- Differences between the two types of relationships are easy to identify, but establishing a cause and effect is difficult

Example: sleeping with one's shoes on is strongly correlated with waking up with a headache. Correlation-implies-causation fallacy: therefore, sleeping with one's shoes causes headache.
More plausible explanation: both are caused by a third factor: going to bed drunk.
Identify a cause Vs a correlation: use of a controlled study
- In medical research, one group may receive a placebo (control) while the other receives a treatment If the two groups have noticeably different outcomes, the different experiences may have caused the different outcomes

Q. Difference between correlation and regression?
Ans: One is a model(equation).one is a coefficient.

Q. What difference between covariance and correlation?
Ans: Both are one and the same, corris scaled version of cov(correlation is unit free, can be directly used in comparisons like corr(X1,Y1)>corr(X2,Y2), but it can't be done directly with covariance).

OTHER QUESTIONS (Without Answers)

1. What is the difference between mean/median/mode? Which one you prefer and why?
2. What is normal distribution and why it is so important?
3. What is the central limit theorem and its importance?
4. What is sampling? What are different sampling methods you used and difference between them? (Simple random sample/stratified sample/cluster sample)
5. What is hypothesis testing? And importance?
6. What are type-1 and type-2 errors? Which is most dangerous?
7. What is t-statistics?
8. Is there any relationship between p-value and t-statistics?
9. When do you use t-test?
10. What are different t-tests and what is the difference between them?
11. When do you use chi-square test?
12. Have you used ANOVA any time? What is ANOVA? What is the difference between t-test and ANOVA?
13. What are parametric and non-parametric tests?
14. How will you perform different statistical tests in SAS/R/Python? And interpretation?