STATISTICS

Ans 1: Central Limit theorem states that the distribution of a sample variable is approximately a normal distribution (bell shape curve) as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population’s actual distribution shape. Central Limit Theorem is a statistical premise that given a sufficiently large sample size from a population with a finite level of variance, the mean of all sampled variables from the same population will be approximately equal to the mean of the whole population. Further, these samples have approximately a normal distribution with their variances being approximately equal to the variance of the population as the sample size gets larger, according to the law of larger numbers

* 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.
* A key aspect of CLT is that the average of the sample means and standard deviations will equal the population mean and standard deviation.
* Sufficiently large sample size can predict the characteristics of a population more accurately.
* CLT is useful in finance when analyzing a large collection of securities to estimate portfolio distributions and traits for returns, risk, and correlation.

Ans 2: Sampling means selecting the group that you will actually collect data from in your research. For ex: If you are researching the opinions of students in your university, you could survey a sample of 100 students. In statistics, sampling allows you to test a hypothesis about the characteristics of a population.

These are two types of sampling methods:

Probability sampling involves random selection, allowing you to make strong statistical inferences about the whole group.

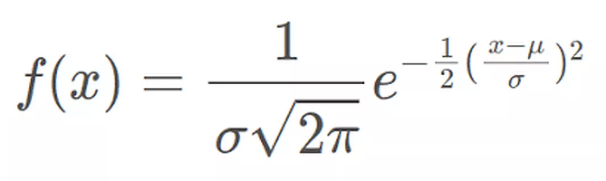
Non-probability sampling involves non-random selection based on convenience or other criteria, allowing you to easily collect data.

Ans 3: A type I error (false-positive) occurs if an investigator rejects a null hypothesis that is actually true in the population. A type II error (false-negative) occurs if the investigator fails to reject a null hypothesis that is actually false in the population. Although type I and type II errors can never be avoided entirely, the investigator can reduce their likelihood by increasing the sample size. False-positive and False-negative results can also occur because of Bias. Such errors are troublesome since they may be difficult to detect and cannot usually be qualified

Ans 4: Normal distribution also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing the data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a “Bell Curve”. Normal distribution is the most common type of distribution assumed in technical stock market analysis and in other types of statistical analyses. Normal distribution model is important in statistics and is key to the central limit theorem. Normal distribution is one type of Symmetrical distribution.

The normal distribution follows the following formula. Note that only the values of the mean (μ ) and standard deviation (σ) are necessary

Normal Distribution Formula.

  
where:

* *x* = value of the variable or data being examined and f(x) the probability function
* μ = the mean
* σ = the standard deviation

Ans 5: Correlation and Covariance are two mathematical concepts used in statistics. Both terms are used to describe how two variables relate to each other. Covariance is a measure of how two variables change together. Both terms describe the extent to which a random variable or a set of random variables can deviate from the expected value.

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables.

It not only shows the kind of relationship (in terms of direction) but also how strong the relationship is. Thus, we can say the correlation values have standardized notions, whereas the covariance values are not standardized and cannot be used to compare how strong or weak the relationship is because the magnitude has no direct significance. It can assume values from -1 to +1.

To determine whether the covariance of the two variables is large or small, we need to assess it relative to the standard deviations of the two variables.

Covariance signifies the direction of the linear relationship between the two variables. By direction we mean if the *variables* are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable). The values of covariance can be any number between the two opposite infinities. Also, it’s important to mention that covariance only measures how two variables change together, not the dependency of one variable on another one.

Ans 6:

Univariate data: This type of data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.

Bivariate data: This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship between the two variables.

Multivariate data: This type of data involves three or more variables, it is categorized under multivariate. It is similar to bivariate but contains more than one dependent variable. The ways to perform analysis on this data depending on the goals to be achieved. Some of the techniques are regression analysis, path analysis, and factor analysis.

Univariate statistics summarize only one [variable](https://www.scribbr.com/methodology/types-of-variables/) at a time.

Bivariate statistics compare two variables.

Multivariate statistics compare more than two variables

Ans 7: Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest, etc.).Sensitivity is nothing but “Predicted True events/ Total events”. True events here are the events that were true and the model also predicted them as true.

The calculation of seasonality is pretty straightforward

.**Seasonality**= ( **True Positives** ) / ( **Positives in Actual Dependent Variable** )

where true positives are positive events that are correctly classified as positives.

Ans 8: Hypothesis testing is an act in statistics whereby an analyst [tests](https://www.investopedia.com/terms/w/wilcoxon-test.asp) an assumption regarding a population parameter. In hypothesis testing, an [analyst](https://www.investopedia.com/terms/a/analyst.asp) tests a statistical sample, with the goal of providing evidence on the plausibility of the null hypothesis. Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed. All analysts use a random population sample to test two different hypotheses: the [null hypothesis](https://www.investopedia.com/terms/n/null_hypothesis.asp) and the alternative hypothesis. The null hypothesis is usually a hypothesis of equality between population parameters; e.g., a null hypothesis may state that the population means the return is equal to zero.

In hypothesis testing, there are two manually exclusive hypothesis

The Null Hypothesis (H0) and The Alternative Hypothesis(H1) One of these is the claim to be tested and based on the sampling results (which infers a similar measurement in the population) the claim will either be supported or not.

In two-tail test

Null hypothesis (H0): The null hypothesis here is what is currently stated to be true about the population.

Alternate Hypothesis (H1): The alternate hypothesis is always what is being claimed.

Ans 9: Qualitative data is the descriptive and conceptual findings collected through questionnaires, interviews, etc. Qualitative data is information that cannot be counted, measured, or easily expressed using numbers

Quantitative data are measures of values or counts and are expressed as numbers. It is data about numeric variables. Eg: how many, how much, how often.

Ans 10: The range is calculated by subtracting the lowest value from the highest value. While a large range means high variability, a small range means low variability in a distribution.

The formula to calculate the range is:

R = H – L

Where:

R = range

H = highest value

L = lowest value

To find the interquartile range (IQR), first, find the median (middle value) of the lower and upper half of the data. These values are quartile1 (Q1) and quartile (Q3). The IQR is the difference between Q3 and Q1.

The formula for finding the interquartile range takes the third quartile value and subtracts the first quartile value.

IQR = Q3 – Q1

The interquartile range is the region between the 75th and 25th percentile.

(75- 25 = 50 % of the data).

Ans 11: A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tends to have central, normal values, as peaks with low and high extremes tapering off relatively symmetrically on their side. Bell Curve is a visual representation of normal distribution, also called Gaussian Distribution. A normal distribution curve, when graphed out, typically follows a bell-shaped curve, hence the name. while the precise shape can vary according to the distribution of the population, the peak is always in the middle, and the curve is always symmetrical. Bell curves are useful for quickly visualizing a data set’s mean, median and mode are all the same. The long tail refers to the part of the bell curve that stretches out in either direction. If the diagram above represents a population under study, the fat area under the bell curve is where most of the population falls.

Ans 12: Outliers are data points that are far from other data points. In other words, they’re unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results.

There are various methods to detect outliers

## **Using Z-scores to Detect Outliers**

Z-scores can quantify the unusualness of observation when your data follow the normal distribution. Z-scores are the number of standard deviations above and below the mean that each value falls. For example, a Z-score of 2 indicates that observation is two standard deviations above the average while a Z-score of -2 signifies it is two standard deviations below the mean. A Z-score of zero represents a value that equals the mean.

To calculate the Z-score for observation, take the raw measurement, subtract the mean, and divide by the standard deviation. Mathematically, the formula for that process is the following:

z-score equation

The further away an observation’s Z-score is from zero, the more unusual it is. A standard cut-off value for finding outliers is a Z-score of +/-3 or further from zero. The probability distribution below displays the distribution of Z-scores in a standard normal distribution. Z-scores beyond +/- 3 are so extreme you can barely see the shading under the curve.

Ans 13: p-value is a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true. P-values are used in hypothesis testing to help decide whether to reject the null hypothesis. The smaller the p-value, the more likely you are to reject the null hypothesis. The p-value, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis. It does this by calculating the likelihood of your test statistic, which is the number calculated by a statistical test using your data. The p-value tells you how often you would expect to see a test statistic as extreme or more extreme than the one calculated by your statistical test if the null hypothesis of that test was true. The p-value gets smaller as the test statistic calculated from your data gets further away from the range of test statistics predicted by the null hypothesis.

The p-value is a proportion : If your p-value is 0.05, that means that 5% of the time you would see a test statistic at least as extreme as the one you found if the null hypothesis was true

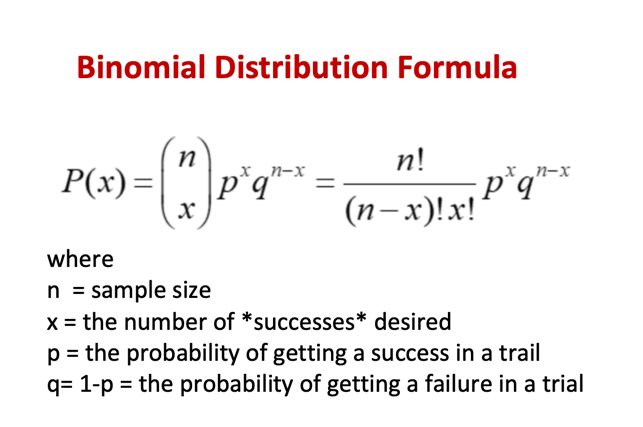
Ans 14: The binomial distribution forms the base for the famous binomial test of statistical importance. A test that has a single outcome such as success/failure is also called a Bernoulli trial or Bernoulli experiment, and a series of outcomes is called a Bernoulli process. Consider an experiment where each time a question is asked for a yes/no with a series of n experiments. Then in the binomial probability distribution, the boolean-valued outcome the success/yes/true/one is represented with probability p and the failure/no/false/zero with probability q (q = 1 − p). In a single experiment when n = 1, the binomial distribution is called a Bernoulli distribution.

The binomial distribution formula is for any random variable X, given by;  P(x:n,p) = nCxx px(1-p)n-x **Or** P(x:n,p) = nCx px (q)n-x

where,

* n = the number of experiments
* x = 0, 1, 2, 3, 4, …
* p = Probability of success in a single experiment
* q = Probability of failure in a single experiment (= 1 – p)

The binomial distribution formula is also written in the form of n-Bernoulli trials, where nCx = n!/x!(n-x)!. Hence, P(x:n,p) = n!/[x!(n-x)!].px.(q)n-x



Ans 15: Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts :

Systematic factors and random factors. The systematic factors do not. Analysts use the ANOVA test to determine the influence of the given dataset, while the random factors do not. Analysts use the ANOVA test to determine the influence the independent variables have on the dependent variable in a regression study. ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups difference by comparing the means of each group and includes spreading out the variance into diverse sources. It is employed with subjects, test groups, between groups, and within groups.

**One-Way ANOVA Versus Two- Way ANOVA**

There are two main types of ANOVA: one-way (or unidirectional) and two-way. There are also variations of ANOVA. For example, MANOVA (multivariate ANOVA) differs from ANOVA as the former tests for multiple dependent variables simultaneously while the latter assesses only one dependent variable at a time. One-way or two-way refers to the number of independent variables in your analysis of the variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable. It determines whether all the samples are the same. The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.

A two-way ANOVA is an extension of one-way ANOVA. With a one-way, you have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents. For example, a two-way ANOVA allows a company to compare worker productivity based on two independent variables, such as salary and skill set. It is utilized to observe the interaction between the two factors and test the effect of two factors at the same time.