EASY

Q. What is statistics?

A. Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting data. It helps in making informed decisions and drawing meaningful conclusions from data.

Q. How are statistics used in everyday life?

A. Statistics is used in various real-life scenarios, such as analyzing opinion polls during elections, determining the effectiveness of medical treatments, and evaluating the performance of sports teams.

Q. What is the difference between population and sample in statistics?

A. Population: A population is the entire group or set of individuals, items, or data that you are interested in studying. For example, if you want to study the heights of all students in a school, the population would be all the students in that school.

Sample: A sample is a subset of the population selected for analysis. Using the school example, if you randomly select 100 students from the school to measure their heights, this group of 100 students is the sample.

Q. Why do we need sample statistics?

A. Sampling in statistics is done when population parameters are not known, especially when the population size is too large.

Q. Types of Sampling technique?

- Random Sampling:
  Random sampling is a method where each member of the population has an equal chance of being selected for the sample. It involves selecting individuals or items from the population in a completely random and unbiased manner.

- Stratified Sampling:
  Stratified sampling is a method in which the population is divided into subgroups or strata based on certain characteristics that are important to the study. Then, a random sample is selected from each stratum. This method ensures that each subgroup is adequately represented in the final sample

Q. What are quantitative data and qualitative data?

A. Qualitative data is used to describe the characteristics of data and is also known as Categorical data. For example, how many types. Quantitative data is a measure of numerical values or counts. For example, how much or how often. It is also known as Numeric data.

Q. What is categorical data?

A. Categorical data represents non-numeric categories or labels. It cannot be subjected to mathematical operations. Examples include gender (male, female), colors (red, blue, green), and vehicle types (sedan, SUV, truck).

Q. Define types of categorical data?

A. Categorical data can further be divided into two subtypes:

i. Nominal Data: These categories have no inherent order or ranking. For example, the types of cars listed above are nominal because there's no meaningful order among them.

ii. Ordinal Data: Ordinal data categories have a clear order or ranking, but the differences between them are not necessarily uniform. For instance, education levels (High School, Bachelor's, Master's, PhD) have an order, but the difference in "educational level" between High School and Bachelor's is not the same as between Bachelor's and Master's.

Q. Define numerical data.

A. Numerical data consists of measurable quantities with meaningful arithmetic operations. It can be further categorized as discrete (e.g., the number of employees) or continuous (e.g., height, weight, temperature).

Q. Define types of Numerical data?

A.Numerical data can be further categorized into two subtypes:

i. Discrete Data: Discrete data can only take on specific, distinct values, often counted in whole numbers. For example, the number of cars in a parking lot is discrete because you can't have a fraction of a car.

ii. Continuous Data: Continuous data can take on an infinite number of values within a given range. For example, temperature can be measured with decimal precision (e.g., 23.5°C), making it continuous.

MEDIUM

Q. What are descriptive statistics?

A. Descriptive statistics are used to summarize the basic characteristics of a data set in a study or experiment. It has three main types –

- Distribution – refers to the frequencies of responses.
- Central Tendency – gives a measure or the average of each response.
- Variability – shows the dispersion of a data set.

Measures of Central Tendency:

Q. Explain the mean.

A. The mean, also known as the average, is calculated by summing all values in a dataset and dividing by the number of values. For example, to find the mean of test scores (85, 90, 92, 78), you add them (85 + 90 + 92 + 78 = 345) and divide by 4 (the number of scores), resulting in a mean of 86.25.

Q. What is the median?

A. The median is the middle value when the data is ordered. If there's an even number of data points, it's the average of the two middle values. For instance, in the dataset (10, 15, 20, 25, 30), the median is 20. In (10, 15, 20, 25, 30, 35), the median is (20 + 25) / 2 = 22.5.

Q. Define mode.

A. The mode is the value that appears most frequently in a dataset. If there is no value repeated, the dataset is said to have no mode. In the dataset (2, 3, 3, 5, 7), the mode is 3.

Measures of Dispersion:

Q. What is the range?

A. The range is the difference between the maximum and minimum values in a dataset, providing a measure of the spread of the data. For instance, in the dataset (15, 20, 30, 45), the range is 45 - 15 = 30.

Q. Explain variance?

A. Variance quantifies how data points deviate from the mean. It is the average of the squared differences between each data point and the mean. For example, the variance of test scores (85, 90, 92, 78) is calculated as [(85-86.25)^2 + (90-86.25)^2 + (92-86.25)^2 + (78-86.25)^2] / 4, yielding the variance.

Q. Define standard deviation?

A. The standard deviation is the square root of the variance. It measures the typical distance between each data point and the mean, providing a more interpretable measure of dispersion. In the test scores example, the standard deviation is the square root of the variance.

$$(\sigma) = \surd(\sum(x-\mu)2 / n)$$

Q. What is the interquartile range (IQR)?

A. The IQR is the range of the middle 50% of data points when data is sorted. To find the IQR, you calculate the difference between the third quartile (75th percentile) and the first quartile (25th percentile). It's useful for understanding the spread of data while ignoring extreme values. For example, consider the dataset (10, 15, 20, 25, 30). The first quartile is 15, the third quartile is 25, and the IQR is 25 - 15 = 10.

Measures of Relationship:

Q. What is covariance?

A. Covariance measures the degree to which two variables change together. A positive covariance suggests that when one variable increases, the other tends to increase, and vice versa. A negative covariance suggests an inverse relationship. For instance, in a study examining hours spent studying and exam scores, a positive covariance would indicate that more study hours are associated with higher scores.

Q. Explain correlation.

A. Correlation quantifies the strength and direction of a linear relationship between two variables, typically ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation. For example, a correlation coefficient of 0.8 between the amount of rainfall and crop yield suggests a strong positive correlation, indicating that increased rainfall is associated with higher crop yields.

Q. Describe the characteristics of a normal distribution.

A. A normal distribution is a symmetric, bell-shaped curve. It is characterized by a mean, median, and mode that are equal and located at the center of the distribution. The spread of the data is determined by the standard deviation, and it follows the 68-95-99.7 rule, indicating that approximately 68% of data falls within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations.

Q. What is the empirical rule?

A. The empirical rule, also known as the 68-95-99.7 rule, is a statistical guideline for a normal distribution. It states that:

Approximately 68% of the data falls within one standard deviation of the mean.

Approximately 95% of the data falls within two standard deviations of the mean.

Approximately 99.7% of the data falls within three standard deviations of the mean. This rule is used to understand the distribution of data in a standard normal curve.

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

A. In a normal distribution, the mean and the median are equal.

Q. What is skewness?

A. Skewness provides the measure of the symmetry of a distribution. If a distribution is not normal or asymmetrical, it is skewed. A distribution can exhibit positive skewness or negative skewness if the tail on the right is longer and the tail on the left side is longer, respectively.

Q. What is the left-skewed distribution and the right-skewed distribution?

A. In the left-skewed distribution, the left tail is longer than the right side.

Mean < median < mode

In the right-skewed distribution, the right tail is longer. It is also known as positive-skew distribution.

Mode < median < mean

Q. How to convert normal distribution to standard normal distribution?

A. Any point (x) from the normal distribution can be converted into standard normal distribution (Z) using this formula –

Z(standardized) = (x-μ) / σ

Here, Z for any particular x value indicates how many standard deviations x is away from the mean of all values of x.

Q. What is an outlier?

A. Outliers can be defined as the data points within a data set that varies largely in comparison to other observations. Depending on its cause, an outlier can decrease the accuracy as well as the efficiency of a model. Therefore, it is crucial to remove them from the data set.

Q. Define probability.

A. Probability is a numerical measure of the likelihood of an event occurring, expressed as a value between 0 and 1. A probability of 0 means the event is impossible, while a probability of 1 indicates certainty. For example, the probability of getting a heads in a fair coin toss is 0.5.

Q. What is the addition rule in probability?

A. The addition rule states that the probability of either of two mutually exclusive events occurring is the sum of their individual probabilities. Mutually exclusive events are events that cannot occur simultaneously. For instance, in a deck of cards, the probability of drawing either a red card or a face card is calculated as the sum of the probabilities of drawing a red card and drawing a face card.

Q. Differentiate between marginal and conditional probability.

A. Marginal probability refers to the probability of one event occurring independently of other events. It is based on the frequencies of single events. For example, in a two-dice roll, the marginal probability of getting a 6 on one die is 1/6, as it is unaffected by the outcome of the other die.

Conditional probability, on the other hand, takes into account prior information or events. It's the probability of one event occurring given that another event has already occurred. For example, in a deck of cards, the conditional probability of drawing a black card given that you've already drawn a spade is based on the reduced sample space.

Q. What is the formula for conditional probability?

A. Conditional Probability (A given B) = P(A and B) / P(B)

This formula calculates the probability of event A occurring given that event B has occurred. P(A and B) represents the joint probability of both events happening, and P(B) is the probability of event B occurring.

HARD

Q. Differentiate between descriptive and inferential statistics.

- Descriptive Statistics: Descriptive statistics involve methods for summarizing and presenting data in a meaningful way. It includes techniques like calculating means, medians, and standard deviations, as well as creating graphs and charts to provide an overview of data.
- Inferential Statistics: Inferential statistics is concerned with making inferences or predictions about a population based on sample data. It includes hypothesis testing, confidence intervals, and regression analysis, which help draw conclusions about a larger group based on a smaller subset of data. For example, estimating the average income of a city's entire population based on a sample survey is an application of inferential statistics.

Q. Give examples of descriptive and inferential statistics.

- Descriptive Statistics: Calculating the average age of students in a class, creating a histogram to show the distribution of test scores in a school, or computing the median income of a neighborhood.
- Inferential Statistics: Conducting a hypothesis test to determine if a new drug is effective in reducing cholesterol levels in a population, using regression analysis to predict housing prices based on factors like square footage and location, or calculating a confidence interval to estimate the proportion of defective products in a factory based on a random sample.

Q. What can you do with an outlier?

A. Outliers affect A/B testing and they can be either removed or kept according to what situation demands or the data set requirements.

Here are some ways to deal with outliers in data –

- Filter out outliers especially when we have loads of data.
- If a data point is wrong, it is best to remove the outliers.
- Alternatively, two options can be provided – one with outliers and one without.
- During post-test analysis, outliers can be removed or modified. The best way to modify them is to trim the data set.
- If there are a lot of outliers and results are critical, then it is best to change the value of the outliers to other variables. They can be changed to a value that is representative of the data set.
- When outliers have meaning, they can be considered, especially in the case of mild outliers.

Q. How to detect outliers?

A. The best way to detect outliers is through graphical means. Apart from that, outliers can also be detected through the use of statistical methods using tools such as Excel, Python, SAS, among others. The most popular graphical ways to detect outliers include box plot and scatter plot.

Q. What is the relationship between standard error and margin of error?

A. Margin of error = Critical value X Standard deviation for the population

and

Margin of error = Critical value X Standard error of the sample.

The margin of error will increase with the standard error.a

Q. What is Random Sampling? Give some examples of some random sampling techniques.

A. Random sampling is a sampling method in which each sample has an equal probability of being chosen as a sample. It is also known as probability sampling.

Let us check four main types of random sampling techniques –

- Simple Random Sampling technique – In this technique, a sample is chosen randomly using randomly generated numbers. A sampling frame with the list of members of a population is required, which is denoted by ‘n’. Using Excel, one can randomly generate a number for each element that is required.
- Systematic Random Sampling technique -This technique is very common and easy to use in statistics. In this technique, every k’th element is sampled. For instance, one element is taken from the sample and then the next while skipping the pre-defined amount or ‘n’.

In a sampling frame, divide the size of the frame N by the sample size (n) to get ‘k’, the index number. Then pick every k’th element to create your sample.

- Cluster Random Sampling technique -In this technique, the population is divided into clusters or groups in such a way that each cluster represents the population. After that, you can randomly select clusters to sample.
- Stratified Random Sampling technique – In this technique, the population is divided into groups that have similar characteristics. Then a random sample can be taken from each group to ensure that different segments are represented equally within a population.

Q. Describe Hypothesis Testing. How is the statistical significance of an insight assessed?

A. Hypothesis Testing in statistics is used to see if a certain experiment yields meaningful results. It essentially helps to assess the statistical significance of insight by determining the odds of the results occurring by chance. The first thing is to know the null hypothesis and then state it. Then the p-value is calculated, and if the null hypothesis is true, other values are also determined. The alpha value denotes the significance and is adjusted accordingly.

If the p-value is less than alpha, the null hypothesis is rejected, but if it is greater than alpha, the null hypothesis is accepted. The rejection of the null hypothesis indicates that the results obtained are statistically significant.

Q. What is the difference between the first quartile, the second quartile, and the third quartile?

A. The first quartile is denoted by Q1 and it is the median of the lower half of the data set.

The second quartile is denoted by Q2 and is the median of the data set.

The third quartile is denoted by Q3 and is the median of the upper half of the data set.

About 25% of the data set lies above Q3, 75% lies below Q3 and 50% lies below Q2. The Q1, Q2, and Q3 are the 25th, 50th, and 75th percentile respectively.

Q. How to screen for outliers in a data set?

A. There are many ways to screen and identify potential outliers in a data set. Two key methods are described below –

Standard deviation/z-score – Z-score or standard score can be obtained in a normal distribution by calculating the size of one standard deviation and multiplying it by 3. The data points outside the range are then identified. The Z-score is measured from the mean. If the z-score is positive, it means the data point is above average.

If the z-score is negative, the data point is below average.

If the z-score is close to zero, the data point is close to average.

If the z-score is above or below 3, it is an outlier and the data point is considered unusual.

The formula for calculating a z-score is –

z= data point−mean/standard deviation OR $z=x-\mu/ \sigma$

Interquartile range (IQR) – IQR, also called midspread, is a method to identify outliers and can be described as the range of values that occur throughout the length of the middle of 50% of a data set. It is simply the difference between two extreme data points within the observation.

IQR=Q3 – Q1

Other methods to screen outliers include Isolation Forests, Robust Random Cut Forests, and DBScan clustering.

Q. What is the meaning of an inlier?

A. An Inliner is a data point within a data set that lies at the same level as the others. It is usually an error and is removed to improve the model accuracy. Unlike outliers, inlier is hard to find and often requires external data for accurate identification.

Q. What is the meaning of six sigma in statistics?

A. Six sigma in statistics is a quality control method to produce an error or defect-free data set. Standard deviation is known as Sigma or σ. The more the standard deviation, the less likely that process performs with accuracy and causes a defect. If a process outcome is 99.99966% error-free, it is considered six sigma. A six sigma model works better than 1σ, 2σ, 3σ, 4σ, 5σ processes and is reliable enough to produce defect-free work.

Q. What is the meaning of KPI in statistics?

A. KPI is an acronym for a key performance indicator. It can be defined as a quantifiable measure to understand whether the goal is being achieved or not. KPI is a reliable metric to measure the performance level of an organization or individual with respect to the objectives. An example of KPI in an organization is the expense ratio.

Q. What is the Pareto principle?

A. Also known as the 80/20 rule, the Pareto principle states that 80% of the effects or results in an experiment are obtained from 20% of the causes. A simple example is – 20% of sales come from 80% of customers.

Q. What are some of the properties of a normal distribution?

A. Also known as Gaussian distribution, Normal distribution refers to the data which is symmetric to the mean, and data far from the mean is less frequent in occurrence. It appears as a bell-shaped curve in graphical form, which is symmetrical along the axes.

The properties of a normal distribution are –

- Symmetrical – The shape changes with that of parameter values
- Unimodal – Has only one mode.
- Mean – the measure of central tendency
- Central tendency – the mean, median, and mode lie at the centre, which means that they are all equal, and the curve is perfectly symmetrical at the midpoint.

Q. How would you describe a ‘p-value’?

A. P-value in statistics is calculated during hypothesis testing, and it is a number that indicates the likelihood of data occurring by a random chance. If a p-value is 0.5 and is less than alpha, we can conclude that there is a probability of 5% that the experiment results occurred by chance, or you can say, 5% of the time, we can observe these results by chance.

Q. What is the difference between type I vs type II errors?

A. A type 1 error occurs when the null hypothesis is rejected even if it is true. It is also known as false positive.

A type 2 error occurs when the null hypothesis fails to get rejected, even if it is false. It is also known as a false negative.

Q. Give an example of a data set with a non-Gaussian distribution?

A. A non-Gaussian distribution is a common occurrence in many processes in statistics. This happens when the data naturally follows a non-normal distribution with data clumped on one side or the other on a graph. For example, the growth of bacteria follows a non-Gaussian or exponential distribution naturally and Weibull distribution.

Q. When should you use a t-test vs a z-test?

A. The z-test is used for hypothesis testing in statistics with a normal distribution. It is used to determine population variance in the case where a sample is large.

The t-test is used with a t-distribution and used to determine population variance when you have a small sample size.

In case the sample size is large or n>30, a z-test is used. T-tests are helpful when the sample size is small or n<30.

Q. What is selection bias and why is it important?

A. Selection bias is a term in statistics used to denote the situation when selected individuals or a group within a study differ in a manner from the population of interest that they give systematic error in the outcome.

Typically selection bias can be identified using bivariate tests apart from using other methods of multiple regression such as logistic regression.

It is crucial to understand and identify selection bias to avoid skewing results in a study. Selection bias can lead to false insights about a particular population group in a study.

Different types of selection bias include –

- Sampling bias – It is often caused by non-random sampling. The best way to overcome this is by drawing from a sample that is not self-selecting.
- Participant attrition – The dropout rate of participants from a study constitutes participant attrition. It can be avoided by following up with the participants who dropped off to determine if the attrition is due to the presence of a common factor between participants or something else.
- Exposure – It occurs due to the incorrect assessment or the lack of internal validity between exposure and effect in a population.
- Data – It includes dredging of data and cherry-picking and occurs when a large number of variables are present in the data causing even bogus results to appear significant.
- Time-interval – It is a sampling error that occurs when observations are selected from a certain time period only. For example, analyzing sales during the Christmas season.
- Observer selection- It is a kind of discrepancy or detection bias that occurs during the observation of a process and dictates that for the data to be observable, it must be compatible with the life that observes it.

Q. What is the chance of rolling at least one five with two dice?

A. Let’s assume that event A is that we get 5 on 1st dice and B is that we get 5 on 2nd dice

Since the outcome of throwing the second die wouldn't be affected by the outcome of throwing the first dice, we can calculate the probability of independent events A and B both occurring as: $P(A \cap B)=P(A)*P(B)$

The probability of getting at least one 5 can be computed using the probability of the union of two events:

$P(A \cup B)=P(A)+P(B)-P(A \cap B)$The probability of getting any specific outcome from a die is ⅙.

Thus, $P(A \cup B)=1/6+1/6-1/(6*6)=1/3-1/36=11/36$

Thus the probability of rolling at least one five with two dice is 11/36.

Q. Given that a die is rolled twice and the sum of the numbers is noted to be 6, what is the conditional probability that the number 4 has occurred at least once?

A. If you roll the dice twice, you'll get the following sample space:

S = { (1,1)(1,2)(1,3)(1,4)(1,5)(1,6)

(2,1)(2,2)(2,3)(2,4)(2,5)(2,6)

(3,1)(3,2)(3,3)(3,4)(3,5)(3,6)

(4,1)(4,2)(4,3)(4,4)(4,5)(4,6)

(5,1)(5,2)(5,3)(5,4)(5,5)(5,6)

(6,1)(6,2)(6,3)(6,4)(6,5)(6,6)}

Provided the given data, calculate the probability that 4 has appeared at least once, given that the sum of the numbers is 6.

Assume that F: The total of two numbers is six.

Take E, for example, 4 has appeared at least once.d

As a result, we must locate P(E|F).

Obtaining P (E):

The chances of collecting four at least once are:

E = (1, 4), (2, 4), (3, 4), (4, 4), (5, 4), (6, 4), (4, 1), (4, 2), (4, 3), (4, 5), (4, 6), (4, 1), (4, 2), (4, 3), (4, 5), (4, 6)

As a result, $P(E) = 11/36$.

Identifying P (F):

The chance of getting the sum of two numbers is 6:

$F = \{(1, 5), (5, 1), (2, 4), (4, 2), (3, 3)\}$

As a result, $P(F) = 5/ 36$

In addition, $E \cap F = \{(2,4), (4,2)\}$

$P(E \cap F) = 2/36$

As a result, $P(E|F) = (P(E \cap F)) / (P (F))$

Now, Substitute the computed probability values= $(2/36)/ (5/36)$

Hence, 2/5   is the required probability.

Q. What is the likelihood of drawing two cards with the same suite (from the same deck)?

A. This is an illustration of a dependent event. According to this definition, the likelihood that two events will occur in the case of a dependent event is:

The probability of events A and B occurring simultaneously is equal to the chance of events A occurring multiplied by the likelihood of events B occurring given the outcome of events A.

This is expressed as $P(A \cap B)=P(A) \ast P(B|A)$

In this scenario, a deck of cards comprises four suites, each of which contains 13 cards.

Our chance of getting a card from a certain suite in the initial draw would be 13/52. Our chances of drawing a card from the same suite as the previous one in the subsequent draw would drop from 13/52 to 12/51.

Hence P(two cards same suite)

$=4 \ast 13/52 * 12/51$

$=4/17$