100+ Statistics Interview Questions

Assignment

Submitted by : Abutalha D Maniyar

1. What are the most important topics in statistics?

Ans: Statistics is broadly divided into two categories

1. *Descriptive Statistics*
2. *Inferential Statistics*

Important topics in Descriptive Statistics are:

1. Types of Data, Population and Sample
2. Central Tendency(mean,median,mode)
3. Measures of Spread(Variance, Std, Range, IQR)
4. Covariance and Correlation
5. Quantiles (Percentiles, Quartiles)
6. 5 Number Summary and Box Plot
7. Univariate,Bivariate & Multivariate analysis
   (Histogram, Contingency Table, Scatter plot etc.,)
8. Normalization and Standardization

Important topics in Inferential Statistics are:

1. Distribution of Probability (PMF, PDF, CDF)
2. Hypothesis testing(Z-test, T-test, Chi Square, Anova)
   a. Null and Alternate hypothesis
   b. Steps in Hypothesis testing
   c. Performing Z-test
   d. Rejection Region Approach
   e. Type 1 and Type 2 error
   f. One sided and two sided test
   g. Statistical power
   h. P value and its interpretation
3. Central Limit Theorem and Standard Error

2. What is exploratory data analysis?

Ans: Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

3. What are quantitative data and qualitative data?

Ans: As the name suggests **Quantitative data** are those which are numerical. Example : Age in years, Number of students in a class etc., whereas **Qualitative data** refers to the categorical data or non-numerical data. Example: Size of shirt(S, M, L, XL) and Taste of Dish (Good, Bad)

4. What is the meaning of KPI in statistics?

Ans: KPI stands for key performance indicator, a quantifiable measure of performance over time for a specific objective. KPIs provide targets for teams to shoot for, milestones to gauge progress, and insights that help people across the organization make better decisions.

5. What Is the Difference Between Univariate, Bivariate, and Multivariate Analysis?

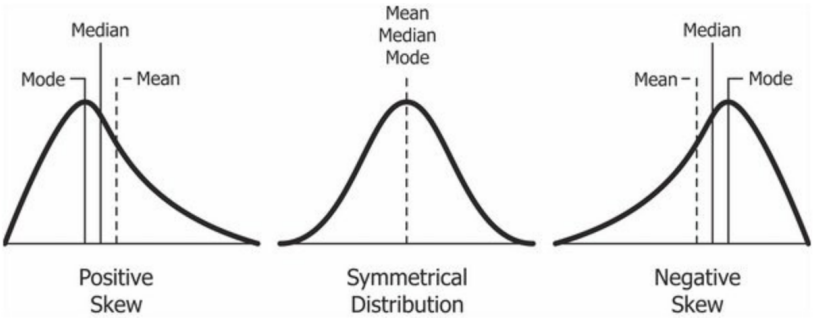
Ans: In Univariate analysis we consider only one feature (column) whereas in bivariate analysis we consider two features together for analysis and if we consider more than two features then that is called multivariate analysis.

6. How Would You Approach a Dataset That's Missing More Than 30 Percent of Its Values?

Ans: The approach will depend on the size of the dataset. If it is a large dataset, then the quickest method would be to simply remove the rows containing the missing values. Since the dataset is large, this won't affect the ability of the model to produce results.If the dataset is small, then it is not practical to simply eliminate the values. In that case, it is better to calculate the mean or mode of that particular feature and input that value where there are missing entries. Another approach would be to use a machine learning algorithm to predict the missing values. This can yield accurate results unless there are entries with a very high variance from the rest of the dataset.

7. Give an example where the median is a better measure than the mean?

Ans: When Data is **skewed** it is better to use median than mean. Example: In an exam 3 students perform really well but remaining 32 students have not performed that good hence in this case it is better to use median of marks than mean of marks. (from below fig we can see median does not get affected as much as the mean gets affected when data gets positively skewed or negatively)



8. What is the difference between Descriptive and Inferential Statistics?

Ans. Descriptive statistics describe some sample or population. **Inferential statistics** attempts to infer from some sample to the larger population.

9. What are descriptive statistics?

Ans: *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.

10. Can you state the method of dispersion of the data in statistics?

Ans: Standard Deviation is the method of dispersion and it can be stated as follows: It is a measure of spread of data about the mean. SD is the square root of the sum of squared deviation from the mean divided by the number of observations.

Sample

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Population

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

11. How can we calculate the range of the data?

Ans: Range = Maximum Value - Minimum Value

12. Is the range sensitive to outliers?

Ans: Yes, Range is very sensitive. If outliers are present in the data they adversely affect range.

13. What is the meaning of standard deviation?

Ans. Standard deviation is a statistic that measures the dispersion of a dataset relative to its mean. It is the average amount of variability in your dataset. It tells you, on average, how far each value lies from the mean.
A high standard deviation means that values are generally far from the mean, while a low standard deviation indicates that values are clustered close to the mean.

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

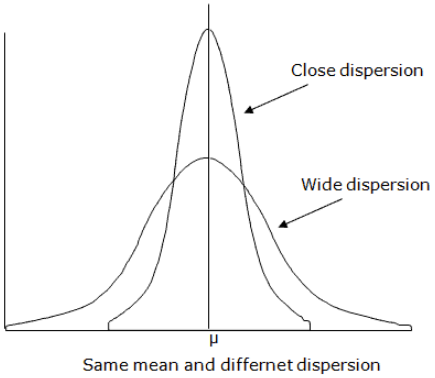
Ans: When the data has some special case of outliers they need to be interpreted accordingly. Suppose we have data of Blood Sugar Level of 100 students, if we remove the extreme data by calling it as an outlier we will lose that data point. And that data could be helpful in Assessing total health of the student or something like that. And in case of spam email detection we keep outliers. So outliers must be interpreted accordingly before completely removing them from the dataset.

15. What is Bessel's correction?

Ans: In statistics, Bessel's correction is the use of n-1 instead of n in several formulas, including the sample variance and standard deviation, where n is the number of observations in a sample. This method corrects the bias in the estimation of the population variance. It also partially corrects the bias in the estimation of the population standard deviation, thereby, providing more accurate results.

16. What do you understand about a spread out and concentrated curve?

Ans: Spread out curve has large deviation from center whereas concentrated curve has larger frequency of data around the center.



Same mean and differnet dispersion

17. Can you calculate the coefficient of variation?

Ans: If I know the S.D and Mean of the data I can find Coefficient of variation using the formula below, and is expressed in percentage by multiplying by 100

$$CV = \frac{\sigma}{\mu}$$

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

Ans: Case- Suppose In a City there are 100 people and The Richest Person of Asia is also in that city. If we calculate the mean of the wealth of 100 people. We would get a value which is more dragged towards the outlier(i.e, rich person's wealth) hence in this case we will go for median to avoid the misinterpretation of the mean and wealth of the city. Hence in this case median is a better measure than the mean.

19. How is missing data handled in statistics?

Ans: Depending upon the type of feature that we have we will try to handle the missing data. If the feature is categorical we will replace it with the mode of the variable. And if multimodal data is present we may randomly assign the missing values with the modes of the data.

If the data is numerical we may replace missing values with the median, mean or mode. Depending upon the variable we have and the distribution it follows and by domain knowledge we can handle the missing data.

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

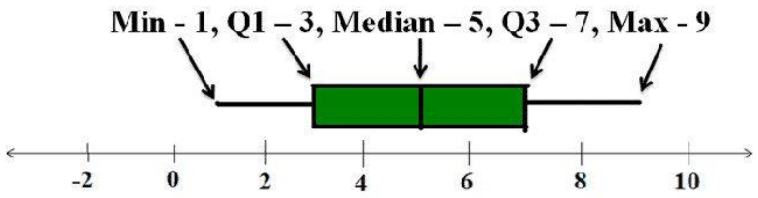
Ans: In statistics, imputation is the process of replacing missing data with substituted values. Mean imputation means replacing missing data with the mean of the available data. It is not completely a bad practice but when data is skewed, mean imputation will lead us to wrong conclusions and hence imputation must be done carefully.

21. What is the benefit of using box plots?

Ans: The box plot is suitable for comparing range and distribution for groups of numerical data. Advantages: The box plot organizes large amounts of data, and visualizes outlier values. And Outliers can be detected.

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

Ans: Five Number summary describes given data in 5 numbers and they are:
1. Minimum
2. Q1
3. Q2 (or Median)
4. Q3
5. Maximum



Minimum : All the data lies above it. I.e., 100% data lies above this point.
Q1 : 75% of the data lies above this value.
Q2 : 50% of the data lies above this point.
Q3. 25% of the data lies above this point.
Maximum : All the data is below this point.

23. What is the difference between the First quartile, the IInd quartile, and the IIIrd quartile?

Ans: The difference between these quartiles is that they represent the value of the variable at a certain point. Point refers to a location in the sorted dataset where data gets divided into four equal parts.

24. What is the difference between percent and percentile?

Ans : The key difference between percentage and percentile is the percentage is a mathematical value presented out of 100 and percentile is the percent of values below a specific value. The percentage is a means of comparing quantities. A percentile is used to display position or rank.

25. What is an Outlier?

Ans: Outliers are extreme values that stand out greatly from the overall pattern of values in a dataset or graph.

26. What is the impact of outliers in a dataset?

Ans: Following are some impacts of outliers in the data set: It may cause a significant impact on the mean and the standard deviation. If the outliers are non-randomly distributed, they can decrease normality. They can bias or influence estimates that may be of substantive interest.

27. Mention methods to screen for outliers in a dataset.

Ans: We may use Box plot to screen outliers, in this case minimum and maximum are calculated based on IQR (generally we take min is 1.5 below IQR from Q1 and similarly maximum is 1.5 above IQR from Q3).
If data is skewed it has outliers. And we can find values that are above and below three S.D from mean.
In this way we can screen outliers.

28. How you can handle outliers in the datasets.

Ans : Depending on the specific characteristics of the data, there are several ways to handle outliers in a dataset. Let's review a few of the most common approaches to handle outliers below:

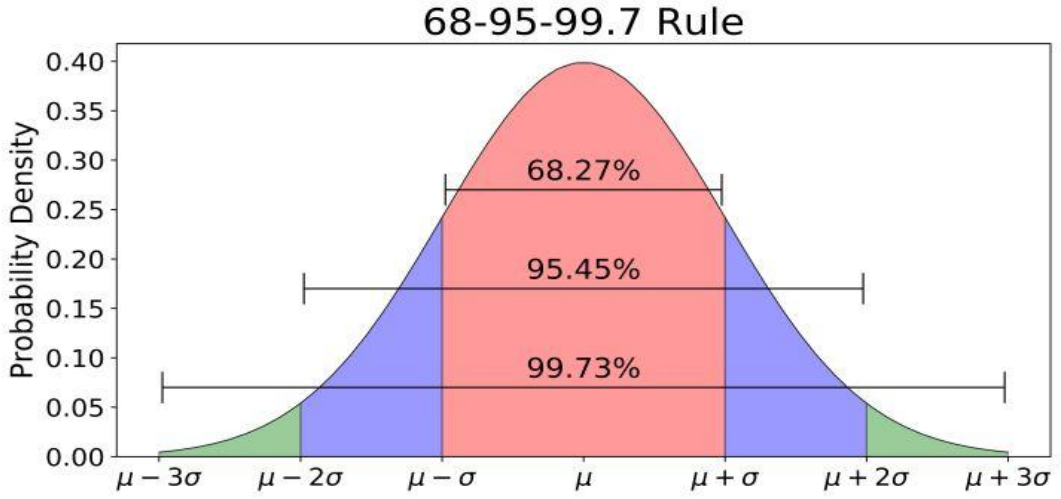
Remove outliers:
In some cases, it may be appropriate to simply remove the observations that contain outliers. This can be particularly useful if you have a large number of observations and the outliers are not true representatives of the underlying population.

Transform outliers:
The impact of outliers can be reduced or eliminated by transforming the feature. For example, a log transformation of a feature can reduce the skewness in the data, reducing the impact of outliers.

Impute outliers:
In this case, outliers are simply considered as missing values. You can employ various imputation techniques for missing values, such as mean, median, mode, nearest neighbor, etc., to impute the values for outliers.

29. What is the empirical rule?

Ans : The figure below shows the empirical rule



30. How to calculate range and interquartile range?

Ans: Range = Maximum value - Minimum value
IQR = Q3 - Q1

31. What is skewness?

Ans: Skewness is a measure of the asymmetry of a probability distribution. It is a statistical measure that describes the degree to which a dataset deviates from the normal distribution.

32. What are the different measures of Skewness?

Ans: The Three main measures of skewness are
1. Karl Pearson's coefficient of skewness
2. Bowley's coefficient of skewness
3. Kelly's coefficient of skewness

33. What is kurtosis?

Ans: Kurtosis is the fourth moment of statistic and it measure the tailedness of the distribution

There is 3 types of kurtosis

1) Leptokurtic
2) Mesokurtic
3) Platykurtic

34. Where are long-tailed distributions used?

Ans: In Finance, Pareto distribution is used to show the distribution of wealth. And In Video's comment section the length of comments follows a lognormal distribution. Wherever the data is skewed or is a non-gaussian long-tailed distribution comes into picture.

35. What is the central limit theorem?

Ans: The means of the samples (each of size n) drawn at random from population follows is normally distributed.
Let's say, we draw 100 samples each of size n=30, then the mean of these 100 samples follow a standard normal distribution.

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

Ans: Suppose we want to calculate the average milk produced by 1000 cows in a city. Let's say we sample 30 cows at random and calculate the mean and also we can calculate its standard deviation. By Central Limit theorem the mean is normally distributed. Hence we can say that the actual mean may lie around the calculated mean. And we generally define confidence interval to compensate for any chance error.

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

Ans: The data must be sampled randomly

The sample values must be independent of each other

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

38. What is the meaning of selection bias?

Ans: Selection bias is the bias introduced by the selection of individuals, groups, or data for analysis in such a way that proper randomization is not achieved, thereby failing to ensure that the sample obtained is representative of the population intended to be analyzed.

39. What are the types of selection bias in statistics?

Ans: **Sampling bias:** occurs when randomization is not properly achieved during data collection.

Convergence bias: occurs when data is not selected in a representative manner. e.g. when you collect data by only surveying customers who purchased your product and not another half, your dataset does not represent the group of people who did not purchase your product.

Participation bias: occurs when the data is unrepresentative due to participations gaps in the data collection process.

40. What is the probability of throwing two fair dice when the sum is 8?

Ans: There are 36 possible outcomes when rolling two dice because each die has 6 sides, and there are 6 possible outcomes for the first die and 6 possible outcomes for the second die (6 x 6 = 36).

Now, let's count the combinations that result in a sum of 8:

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

There are 5 combinations that result in a sum of 8.

So, the probability of getting a sum of 8 when throwing two fair dice is:

Probability = (Number of Favorable Outcomes) / (Total Number of Possible Outcomes)

Probability = 5 / 36

41. What are the different types of Probability Distribution used in Data Science?

Ans: Types of probability distribution used in data science are

1. Normal Distribution
2. LogNormal Distribution
3. Pareto Distribution
4. Bernoulli Distribution
5. Binomial Distribution
6. Poisson Distribution

42. What do you understand by the term Normal Distribution or What is a bell-curve distribution?

Ans: A normal distribution, often referred to as a bell curve, is a fundamental concept in statistics and probability theory. It describes a specific probability distribution of a continuous random variable where the data clusters around a central mean value, creating a symmetric, bell-shaped curve when plotted on a graph.

Key characteristics of a normal distribution (bell curve) include:

1. Symmetry: The curve is perfectly symmetric, with the mean, median, and mode all coinciding at the center of the distribution.

2. Bell-Shaped: The probability density function (PDF) graph forms a distinctive bell shape, where data is most concentrated around the mean, and it gradually tapers off towards the tails.

3. Mean and Standard Deviation: The shape and spread of the curve are determined by the mean (μ) and standard deviation (σ) of the data. The mean represents the central value, while the standard deviation measures the dispersion or spread of data points around the mean.

4. Empirical Rule: A significant property of the normal distribution is the empirical rule (also known as the 68-95-99.7 rule), which 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.

5. Continuous Distribution: The normal distribution is continuous, meaning that it represents a continuous range of possible values rather than a discrete set of values.

43. Can you state the formula for normal distribution?

Ans: The probability density function (PDF) of a normal distribution, often denoted as N(μ, σ), where μ is the mean and σ is the standard deviation, is given by the following formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

In this formula:

- f(x) represents the probability density at a specific value x.
- μ is the mean or average of the distribution.
- σ (sigma) is the standard deviation, which measures the spread or dispersion of the data.
- e is the base of the natural logarithm, approximately equal to 2.71828.
- π is a mathematical constant approximately equal to 3.14159.

44. What type of data does not have a normal distribution or a Gaussian distribution?

Ans: Where the distribution is asymmetric like in Finance, Pareto distribution is used to show the distribution of wealth. And In Video's comment section the

length of comments follows a lognormal distribution. Wherever the data is skewed or is a non-gaussian distribution comes into picture.

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

Ans: In a normal distribution mean = median

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

Ans: It is symmetric around the mean and it follows the empirical rule
Empirical Rule: A significant property of the normal distribution is the empirical rule (also known as the 68-95-99.7 rule), which 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.

47. What is the assumption of normality?

Ans: The key assumptions of normality include:

1. Data Distribution: The data is assumed to be normally distributed, meaning that when you plot the data on a histogram or a probability density curve, it should resemble a bell-shaped curve with the characteristic symmetric shape.

2. Symmetry: The data distribution is symmetric around its mean. This means that the mean, median, and mode are all equal and located at the center of the distribution.

3. No Significant Skewness or Kurtosis: The data distribution does not exhibit substantial skewness (asymmetry) or kurtosis (tailedness) that deviates significantly from a normal distribution.

4. Constant Variance: The variance of the data is assumed to be constant across all levels of the independent variable(s). This is also known as homoscedasticity.

5. Independence: Observations or data points are assumed to be independent of each other.

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

Ans: Steps to convert a normal distribution to standard normal distribution
Step 1: Calculate mean and sd of the data
Step 2: Subtract mean from each data point and then divide by sd

$$z = \frac{x - \mu}{\sigma}$$

Step 3: Now the mean of the z values obtained is 0 and sd is 1

49. Can you tell me the range of the values in standard normal distribution?

Ans: In a standard normal distribution, which is also known as a Z-distribution, the range of values typically extends from negative infinity to positive infinity. This means that theoretically, a standard normal distribution can take on any real number as a value.

However, in practice, the values within a standard normal distribution are typically within a few standard deviations of the mean (which is 0 in the standard normal distribution). Most of the values are concentrated near the center of the distribution, and as you move away from the mean, the probability density decreases rapidly.

50. What is the Pareto principle?

Ans: The Pareto Principle, also known as the 80/20 Rule, is a concept that suggests a significant imbalance between inputs and outputs, effort and results, or causes and effects.
The Pareto Principle can be summarized as follows:

- Roughly 80% of the effects come from 20% of the causes.
- In various contexts, a small proportion (typically around 20%) of inputs, efforts, or factors often generates a large proportion (typically around 80%) of outputs, results, or outcomes.

51. What are left-skewed and right-skewed distributions?

Ans: Skewness is a way to describe the symmetry of a distribution. A left-skewed (Negative Skew) distribution is one in which the left tail is longer than that of the right tail. For this distribution,

mean < median < mode.

Similarly, right-skewed (Positively Skew) distribution is one in which the right tail is longer than the left one. For this distribution,

mean > median > mode.

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

Ans: In this case the mean will be greater than 20 (median).

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

Ans: In this case mean is less than 60 and mode is greater than 60.

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

Ans: Let the score be x, hence

1.20 = (x-160) / 15

Therefore, x = 178

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

Ans: True

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

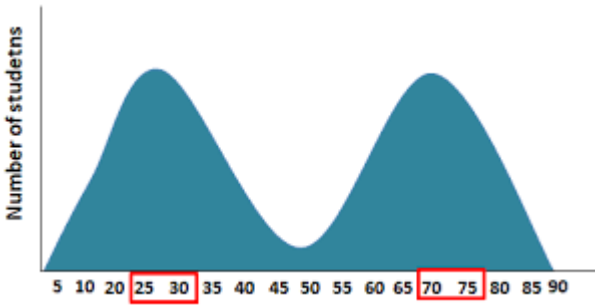
Ans: The population of the shark is large hence we randomly select 30 sharks and by some method we calculate the length of all 30 sharks. We estimate the parameter, and to do so. We calculate the mean and standard deviation of the data obtained. With a certain confidence interval we express the mean length of all sharks in the world.

57. Can you tell me the difference between unimodal bimodal and bell-shaped curves?

Ans: Unimodal has only one mode whereas bimodal has two modes. The bell-shaped curves have only one mode.

58. Does symmetric distribution need to be unimodal?

Ans: No, even bimodal distribution could be symmetric.



59. What are some examples of data sets with non-Gaussian distributions?

Ans. When data follows a non-normal distribution, it is frequently non-Gaussian. A non-Gaussian distribution is often seen in many statistics processes. This occurs when data is naturally clustered on one side or the other on a graph. For instance, bacterial growth follows an exponential or non-Gaussian distribution, which is non-normal.

60. What is the Binomial Distribution Formula?

Ans: $$P(X) = {}_nC_x\, p^x (1-p)^{n-x}$$

61. What are the criteria that Binomial distributions must meet?

Ans: The binomial distribution is a probability distribution that describes the number of successes in a fixed number of independent Bernoulli trials (experiments), where each trial has only two possible outcomes: success or failure. To use the binomial distribution, certain criteria must be met:

1. Two Outcomes: Each trial must result in one of two possible outcomes, often denoted as "success" and "failure." These outcomes are mutually exclusive.

2. Fixed Number of Trials: There must be a fixed number of trials (experiments), denoted as "n." You know in advance how many trials will be conducted.

3. Independence: The trials must be independent of each other. The outcome of one trial should not affect the outcome of any other trial.

4. Constant Probability of Success: The probability of success (denoted as "p") must remain constant from trial to trial. In other words, the probability of success does not change between trials.

With these criteria met, you can use the binomial probability formula to calculate the probability of obtaining a specific number of successes (k) in the fixed number of trials (n). The formula is:

P(X=k) = nCx * p^x * q^(n-x)

62. What are the examples of symmetric distribution?

Ans: Symmetric distributions are probability distributions where the data is evenly distributed on both sides of a central point or axis. In other words, the left and right sides of the distribution mirror each other. Here are some examples of symmetric distributions:
1. Normal Distribution (Gaussian Distribution)
2. Uniform Distribution
3. T-Distribution
4. Logistic Distribution
5. Triangular Distribution
6. Cauchy Distribution

63. How to find the mean length of all fishes in the sea?

Ans: Define the confidence level (most common is 95%). Take a sample of fishes from the sea (to get better results the number of fishes > 30). Calculate the mean length and standard deviation of the lengths. Calculate t-statistics

Get the confidence interval in which the mean length of all the fishes should be.

64. What are the types of sampling in Statistics?

Ans: There are several types of sampling methods, each with its own advantages and disadvantages. Here are some common types of sampling methods:

1. Simple Random Sampling (SRS): In simple random sampling, every member of the population has an equal chance of being selected. This is typically done using random number generators or random sampling techniques.

2. Stratified Sampling: In stratified sampling, the population is divided into subgroups or strata based on certain characteristics (e.g., age, gender, location), and then random samples are drawn from each stratum. This ensures representation from each subgroup.

3. Systematic Sampling: Systematic sampling involves selecting every nth individual from a list or sequence. The starting point is often randomly determined. For example, you might select every 10th person from a list of employees.

4. Cluster Sampling: In cluster sampling, the population is divided into clusters or groups, and a random sample of clusters is selected. Then, all individuals within the selected clusters are surveyed. It's particularly useful when it's difficult to obtain a complete list of the population.

5. Convenience Sampling: Convenience sampling involves selecting individuals who are easiest to reach or survey. This method is quick and convenient but may not be representative of the entire population.

6. Purposive Sampling: Purposive sampling involves selecting individuals or items based on a specific purpose or criteria. It's often used in qualitative research or when certain characteristics are of interest.

7. Snowball Sampling: Snowball sampling is commonly used in situations where it's difficult to identify and access individuals in the population. An initial participant is selected, and then that participant helps identify and recruit additional participants.

65. Why is sampling required?

Ans: Sampling in statistics refers to the process of selecting a subset of individuals or items from a larger population for the purpose of making inferences or drawing conclusions about the entire population.

66. How do you calculate the needed sample size?

Ans: The sample size depends on several factors, including the desired level of confidence, the margin of error, the variability in the population, and the specific research objectives. Here's a general approach to calculate the needed sample size:

1. Determine Your Desired Confidence Level (CL):
The confidence level represents the degree of confidence you want to have in the results. It is typically expressed as a percentage (e.g., 95% confidence level).

2. Choose Your Margin of Error (MOE):
The margin of error is the maximum acceptable difference between the sample estimate and the true population parameter. It is often expressed as a percentage or a fixed value (e.g., ±3%).

3. Estimate Population Variability (if available):
If you have prior knowledge or data about the population's variability (standard deviation), use it to estimate the sample size. If not, you can use a conservative estimate (e.g., 0.5 for a binary response) or conduct a pilot study to estimate variability.

4. Minimum size of the sample is 30. The Standard error follows the law of Large Numbers. And sample size is chosen accordingly by plugging in the above values in respective formulae.

67. Can you give the difference between stratified sampling and clustering sampling?

Ans: *Stratified Sampling*: In stratified sampling, the population is divided into subgroups or strata based on certain characteristics (e.g., age, gender, location), and then random samples are drawn from each stratum. This ensures representation from each subgroup.
Selection Process: The sampling process involves selecting a separate random sample from each stratum. The samples from each stratum are often combined to form the final sample.

Cluster Sampling: In cluster sampling, the population is divided into clusters or groups, and a random sample of clusters is selected. Then, all individuals within the selected clusters are surveyed.
Selection Process: The sampling process involves two stages: first, random selection of clusters, and second, the inclusion of all units within the selected clusters.

68. Where is inferential statistics used?

Ans: Inferential statistics is a branch of statistics that involves drawing conclusions, making predictions, or testing hypotheses about a population based on a sample of data from that population. It is used in various fields and applications where researchers and analysts seek to make inferences or generalizations beyond the data they have collected. Some common areas where inferential statistics is used are:

1. *Scientific Research:* Inferential statistics is widely used in scientific research to make generalizations about entire populations based on sample data. Researchers use inferential statistics to test hypotheses, analyze experimental results, and draw conclusions about the broader scientific phenomena they are investigating.

2. *Business and Economics:* In the business world, inferential statistics is used to make decisions about product development, market research, financial forecasting, and quality control. Analysts use inferential statistics to draw conclusions about consumer behavior, economic trends, and market conditions.

3. *Medical and Health Sciences:* Inferential statistics plays a crucial role in medical research, clinical trials, and epidemiology. It is used to analyze clinical data, test the effectiveness of treatments, and make inferences about the health outcomes of larger populations.

4. *Social Sciences:* Researchers in fields such as psychology, sociology, and political science use inferential statistics to study human behavior, attitudes, and societal trends. It helps researchers test hypotheses about the factors influencing behavior and draw conclusions about larger populations.

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

Ans: In Inferential statistics, Population is larger and is the entire dataset. But Sample is a minimal amount of data from the population. The sample is representative of the population. Mean, SD, Median, anything of Population is called Parameter where as of the sample it is called statistic. And they may slightly differ from each other.

70. What is the relationship between the confidence level and the significance level in statistics?

Ans: The relationship between the confidence level and the significance level can be understood as follows:

The confidence level (1 - α) is complementary to the significance level (α). In other words, if you have a 95% confidence level, the significance level (α) is 5%.

A higher confidence level (e.g., 95%) corresponds to a lower significance level (e.g., α = 0.05), and vice versa.

The choice of significance level (α) in hypothesis testing is a decision made by the researcher and is often set based on the desired balance between the risk of Type I errors and the power of the test.

While a higher confidence level in estimating a parameter provides greater assurance that the interval contains the true parameter, it comes at the cost of wider confidence intervals, making them less precise.

71. What is the difference between Point Estimate and Confidence Interval Estimate?

Ans: Point Estimate and Confidence Interval Estimate are two different ways of summarizing and conveying information about population parameters based on sample data in statistics. And difference between them is
1. Point Estimate provides a single value, whereas Confidence Interval Estimate provides a range of values.
2. Point estimates are precise but do not convey the uncertainty associated with the estimate. Confidence intervals provide a measure of that uncertainty.
3. Confidence intervals are often used when researchers want to communicate the range of values that are plausible for the population parameter, taking into account the variability in the data.

72. What do you understand about biased and unbiased terms?

Ans: In statistics, the terms "biased" and "unbiased" refer to whether an estimator or a statistical method systematically overestimates or underestimates a population parameter on average.

Biased:

- If an estimator or method is "biased," it means that, on average, it tends to produce estimates that are systematically different from the true population parameter.
- In other words, if you were to use the biased estimator or method many times with different samples from the same population, the average of those estimates would not equal the true population parameter.
- Biased estimators/methods can introduce systematic errors and are less desirable because they do not provide accurate and unbiased estimates of the population parameter.

Unbiased:

- If an estimator or method is "unbiased," it means that, on average, it produces estimates that are equal to the true population parameter.
- In the long run, when using the unbiased estimator or method repeatedly with different samples from the same population, the average of those estimates will converge to the true population parameter.
- Unbiased estimators/methods are preferred because they provide accurate and reliable estimates that are not systematically too high or too low.

It's important to note that an estimator can still produce an estimate that is different from the true parameter for a specific sample (due to random sampling variation), even if it is unbiased. However, the key characteristic of

an unbiased estimator is that, on average, it provides accurate estimates when used repeatedly.

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

Ans: In statistics, the width of a confidence interval is inversely related to its precision. Specifically, as the desired level of confidence increases (e.g., from a 90% confidence interval to a 95% confidence interval), the width of the confidence interval also tends to increase. Conversely, as you decrease the level of confidence (e.g., from a 95% confidence interval to a 90% confidence interval), the width of the confidence interval tends to decrease.

74. What is the meaning of standard error?

Ans: The standard error (SE) is a statistical measure of the precision or variability of a sample statistic, such as the sample mean or sample proportion, when estimating a population parameter. It quantifies the spread or dispersion of sample statistics around their expected values. In other words, it tells you how much you can expect the sample statistic to vary from one random sample to another.

The standard error is closely related to the standard deviation (SD), but it is specific to sample statistics, whereas the standard deviation is used to measure the dispersion of individual data points in a dataset.

75. What is a Sampling Error and how can it be reduced?

Ans: Sampling error is a type of error that occurs when the characteristics of a sample, which is a subset of a larger population, do not perfectly represent the characteristics of the entire population. In other words, it is the discrepancy between the sample statistic (e.g., sample mean or sample proportion) and the true population parameter.

Some ways to reduce sampling error are:

1. Increase the Sample Size: One of the most effective ways to reduce sampling error is to increase the sample size. As the sample size grows, the sample statistic becomes a more accurate estimate of the population parameter, and the sampling error decreases. Larger samples tend to provide more precise and reliable estimates.

2. Use Random Sampling: Ensure that the sampling process is truly random. Random sampling helps minimize bias and ensures that every element or individual in the population has an equal chance of being included in the sample. Common methods for random sampling include simple random sampling, stratified sampling, and cluster sampling.

3. Reduce Non-Response Bias: If survey or questionnaire data collection is involved, make efforts to reduce non-response bias by maximizing response rates. High non-response rates can introduce bias and increase sampling error.

4. Use Appropriate Sampling Methods: Choose the appropriate sampling method for the research question and population of interest. Different situations may require different sampling methods, such as stratified sampling for populations with diverse subgroups or cluster sampling for geographically dispersed populations.

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

Ans: The margin of error (MOE) is directly related to the standard error (SE). In fact, the MOE is typically calculated as a function of the SE and the chosen critical value from the probability distribution.

The margin of error defines the range within which the true population parameter is likely to fall with a specified level of confidence. It provides the "plus or minus" part of a confidence interval.

A wider confidence interval (larger MOE) implies lower precision and less certainty about the true parameter value, while a narrower confidence interval (smaller MOE) implies higher precision and greater certainty.

77. What is hypothesis testing?

Ans: Hypothesis testing is a statistical method used to make inferences and draw conclusions about populations based on sample data. It involves the formulation and testing of hypotheses, where a hypothesis is a statement or claim about a population parameter. The main goal of hypothesis testing is to determine whether there is enough evidence in the sample data to support or reject a specific hypothesis about the population.

78. What is an alternative hypothesis?

Ans: In hypothesis testing, the alternative hypothesis (often denoted as H1 or Ha) is a statement that represents a specific claim or assertion about a population parameter. It serves as the alternative to the null hypothesis (Ho), which represents a default or status quo assumption about the population parameter. The alternative hypothesis is what we aim to test or provide evidence for through the hypothesis testing process.

79. What is the difference between one-tailed and two-tail hypothesis testing?

Ans: *One-Tailed Hypothesis Testing:*
1. Research Question: One-tailed hypothesis testing is typically used when the researcher has a specific directional expectation about the difference or effect they are investigating. In other words, the research question involves determining whether a population parameter is significantly greater than or less than a specific value, but not both.

2. Alternative Hypothesis (H1): In one-tailed hypothesis testing, the alternative hypothesis (H1) specifies a specific direction of the effect or difference. It is expressed in the form of inequalities, such as:
H1: $\mu > \mu o$ (indicating that the population mean is greater than a specified value)
H1: $p < po$ (indicating that the population proportion is less than a specified value)

3. Critical Region: The critical region, where the rejection of the null hypothesis occurs, is located entirely in one tail of the probability distribution of the test statistic. This tail corresponds to the direction specified in the alternative hypothesis.

Two-Tailed Hypothesis Testing:

1. Research Question: Two-tailed hypothesis testing is used when the researcher is interested in detecting any significant difference or effect, regardless of direction. The research question involves determining whether a population parameter is significantly different from a specific value.

2. Alternative Hypothesis (H1): In two-tailed hypothesis testing, the alternative hypothesis (H1) does not specify a direction. Instead, it indicates that the population parameter differs from the null hypothesis value, but it does not specify whether it is greater or less than that value. It is expressed in the form of inequalities, such as:
H1: $\mu \neq \mu o$ (indicating that the population mean is not equal to a specified value)

H1: p ≠ po (indicating that the population proportion is not equal to a specified value)

3. Critical Region: The critical region for two-tailed hypothesis testing is divided into two tails of the probability distribution of the test statistic, representing both directions away from the null hypothesis value.

80. What is one sample t-test?

Ans: A one-sample t-test is a statistical hypothesis test used to determine whether the mean of a single sample of data is significantly different from a hypothesized population mean. It is particularly useful when you have collected data from a sample and want to assess whether the sample mean is statistically different from a known or assumed population mean.

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

Ans: In statistics, degrees of freedom (DF) refer to the number of values in the final calculation of a statistic that are free to vary. The concept of degrees of freedom is used in various statistical tests and estimations, such as hypothesis testing, regression analysis, and the calculation of sample statistics like the variance and t-statistic

Here's a more detailed explanation of degrees of freedom:
Variance and Standard Deviation:

- When calculating the sample variance or standard deviation, degrees of freedom are associated with the number of independent pieces of information in the sample.
- In a sample of size n, there are n data points, but not all of them are independent because they are subject to certain constraints.
- Degrees of freedom in this context are typically n-1, denoted as DF = n-1.
- The reason for n-1 degrees of freedom is that one value is fixed by the requirement that the sample mean (or sum) must equal the population mean (or sum). Once n-1 values are known, the n-th value is determined by this constraint.

82. What is the p-value in hypothesis testing?

Ans: A p-value is a number that describes the probability of finding the observed or more extreme results when the null hypothesis (H0) is True.

P-values are used in hypothesis testing to help decide whether to reject the null hypothesis or not. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

83. How can you calculate the p-value?

Ans: First we calculate test statistic considering Ho as true.
For a one-tailed test, by calculating the probability of observing a test statistic as extreme as or more extreme than the observed value in the direction specified by the alternative hypothesis.
For a two-tailed test, calculate the probability of observing a test statistic as extreme as or more extreme than the observed value in both tails.
You can do this by finding the area under the probability distribution curve associated with the test statistic in the tails beyond the critical value(s).

84. If there is a 30 percent probability that you will see a supercar in any 20-minute time interval, what is the probability that you see at least one supercar in the period of an hour (60 minutes)?

Ans: Hypothesis testing is a type of statistical inference that uses data from a sample to conclude about the population data.

Before performing the testing, an assumption is made about the population parameter. This assumption is called the null hypothesis and is denoted by H0. An alternative hypothesis (denoted Ha), which is the logical opposite of the null hypothesis, is then defined.

The hypothesis testing procedure involves using sample data to determine whether or not H0 should be rejected. The acceptance of the alternative hypothesis (Ha) follows the rejection of the null hypothesis (H0).

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

Ans: The p-value represents the probability of observing test results as extreme as or more extreme than the ones obtained from your sample data, assuming that the null hypothesis is true. In other words, it answers the question, "If the null hypothesis were true, how likely is it to observe the results we obtained?"

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

Ans: Type I and Type II errors are two different types of errors that can occur in hypothesis testing and statistical decision-making. They are associated with the acceptance or rejection of null and alternative hypotheses. Here's the difference between Type I and Type II errors:

Type I Error (False Positive):

- Definition: A Type I error occurs when you incorrectly reject a true null hypothesis. In other words, it is a false positive result, indicating that you conclude there is a significant effect or difference when there is none.
- Symbol: Often denoted as \(\alpha\) (alpha), the significance level or the probability of a Type I error.
- Consequence: Making a Type I error means you believe there is an effect or difference when there is not. It can lead to unwarranted conclusions and potentially costly actions.
- Example: In a medical test, a Type I error would occur if the test incorrectly diagnoses a healthy person as having a disease.

Type II Error (False Negative):

- Definition: A Type II error occurs when you incorrectly fail to reject a false null hypothesis. It is a false negative result, indicating that you conclude there is no significant effect or difference when there actually is one.
- Symbol: Often denoted as \(\beta\) (beta), the probability of a Type II error.
- Consequence: Making a Type II error means you miss a real effect or difference, potentially leading to missed opportunities for further investigation or intervention.
- Example: In a medical test, a Type II error would occur if the test fails to diagnose a person with a disease when they actually have it.

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

Ans: The choice between using a t-test and a z-test depends on several factors, including the characteristics of your data, the sample size, and your knowledge of the population standard deviation. Here are guidelines on when to use each test:

Use a t-test when:
1. Population Standard Deviation is Unknown: If you do not know the population standard deviation , or you are working with sample data and estimating the standard deviation from the sample (s), you should use a t-test. The t-test accounts for the uncertainty introduced by estimating the standard deviation.

2 Small Sample Size: For small sample sizes (typically when n < 30), a t-test is preferred. As the sample size increases, the t-distribution approaches the normal distribution (z-distribution), so the distinction becomes less important with larger samples.

3. Random Sampling: The data should be collected through random sampling or a process that approximates random sampling.

4. Approximately Normally Distributed Data: While t-tests are robust to deviations from normality, the data should be approximately normally distributed, especially for small sample sizes. For large sample sizes n > 30, the central limit theorem ensures that the t-distribution approximates the normal distribution even for non-normally distributed data.

Use a z-test when:

1. Population Standard Deviation is Known: If you know the population standard deviation, you can use a z-test. This is often the case in situations where you have complete information about the population.

2. Large Sample Size: For large sample sizes (typically when n >=30), the t-distribution closely approximates the normal distribution. In such cases, using a z-test is appropriate because the distinction between the t-distribution and the normal distribution becomes negligible with larger samples.

3. Comparison with Known Values: When comparing sample data to known population parameters (e.g., comparing a sample mean to a known population mean), a z-test can be used.

88. What is the difference between the f test and anova test?

Ans: The F-test and ANOVA (Analysis of Variance) test are related statistical tests, but they have different applications and purposes. Here are the key differences between the two:

F-Test:

1. Comparison of Two Variances: The F-test is primarily used to compare the variances of two or more populations or samples. Specifically, it assesses whether the variances are significantly different from each other.

2. Two Variance Groups: In its simplest form, the F-test compares the variances of two groups (two samples). This is known as a two-sample F-test. It is often used to check the homogeneity of variances assumption before conducting other statistical tests, such as t-tests or ANOVA.

3. Single Factor: The F-test is univariate and typically focuses on a single factor or variable. It assesses whether the variability in the data (variance) is significantly different across groups.

4. Statistical Output: The result of an F-test is an F-statistic and a p-value. The F-statistic measures the ratio of variances between groups to variances within groups. The p-value helps determine whether this ratio is statistically significant.

ANOVA (Analysis of Variance) Test:

1. Comparison of Multiple Groups: ANOVA is used to compare the means of three or more groups or treatments to determine if there are statistically significant differences among the group means.

2. Multiple Variance Groups: ANOVA assesses whether the variability between group means is greater than the variability within each group. It does involve the comparison of variances but as part of a larger analysis of means.

3. Multiple Factors: ANOVA can be used to analyze the impact of one or more factors (independent variables) on a dependent variable. For example, one-way ANOVA analyzes the effect of a single factor, while two-way ANOVA considers the effects of two factors.

4. Statistical Output: The result of an ANOVA test includes an F-statistic and a p-value. If the p-value is less than the chosen significance level (alpha), it indicates that at least one group mean is significantly different from the others. Post-hoc tests (e.g., Tukey's HSD or Bonferroni) are often used to identify which group means differ from each other.

89. What is Resampling and what are the common methods of resampling?

Ans: Common methods for resampling are

- K-fold cross-validation
- Bootstrapping

90. What is the proportion of confidence intervals that will not contain the population parameter?

Ans: The proportion of confidence intervals that will not contain the population parameter is equal to the chosen significance level (alpha). In other words, if you construct many confidence intervals from different samples and calculate their average, approximately (100% times alpha) of those intervals will not contain the true population parameter.

91. What is a confounding variable?

Ans: A confounding variable in statistics is an ‘extra’ or ‘third’ variable that is associated with both the dependent variable and the independent variable, and it can give a wrong estimate that provides useless results.

For example, if we are studying the effect of weight gain, then lack of workout will be the independent variable, and weight gain will be the dependent variable. In this case, the amount of food consumption can be the confounding variable as it will mask or distort the effect of other variables in the study. The effect of weather can be another confounding variable that may later the experiment design.

92. What are the steps we should take in hypothesis testing?

Ans.

1. State the null hypothesis

2.State the alternate hypothesis

3. Which test and test statistic to be performed

4. Collect Data

5. Calculate the test statistic

6. Construct Acceptance / Rejection regions

7. Based on steps 5 and 6, draw a conclusion about H0

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

Ans: The margin of error (MOE) is directly related to the standard error (SE). In fact, the MOE is typically calculated as a function of the SE and the chosen critical value from the probability distribution.

The margin of error defines the range within which the true population parameter is likely to fall with a specified level of confidence. It provides the "plus or minus" part of a confidence interval.

A wider confidence interval (larger MOE) implies lower precision and less certainty about the true parameter value, while a narrower confidence interval (smaller MOE) implies higher precision and greater certainty.

94. How would you describe what a 'p-value' is to a non-technical person or in a layman term?

Ans: The best way to describe the p-value in simple terms is with an example. In practice, if the p-value is less than the alpha, say of 0.05, then we're saying that there's a probability of less than 5% that the result could have happened by chance. Similarly, a p-value of 0.05 is the same as saying **"Only 5% of the time, we would see this by chance."**

95. What does interpolation and extrapolation mean? Which is generally more accurate?

Ans: Interpolation is a prediction made using inputs that lie within the set of observed values. Extrapolation is when a prediction is made using an input that's outside the set of observed values.

Generally, interpolations are more accurate.

96. What is an inlier?

Ans: An inlier is a data observation that lies within the rest of the dataset and is unusual or an error. Since it lies in the dataset, it is typically harder to identify than an outlier and requires external data to identify them. Once we identify them, we can simply remove them from the dataset to address them.

97. You roll a biased coin (p(head)=0.8) five times. What's the probability of getting three or more heads?

Ans: We can use the binomial probability formula. In this case, we want to find the probability of getting 3, 4, or 5 heads. The formula is:

$$P(X) = {}_nC_x\, p^x (1-p)^{n-x}$$

P(x = 3) = 10 * 0.8^3 * 0.2^2 = 0.2048
P(x = 4) = 5 * 0.8^4 * 0.2^1 = 0.4096
P(x = 5) = 1 * 0.8^5 * 0.2^0 = 0.3277

Therefore, P(x>=3) = P(x=3) + P(x=4) + P(x=5) = 0.9421 = 94.21 %

98. Infection rates at a hospital above a 1 infection per 100 person-days at risk are considered high. A hospital had 10 infections over the last 1787

person-days at risk. Give the p-value of the correct one-sided test of whether the hospital is below the standard.

Ans: To find the p-value for the one-sided test of whether the hospital infection rate is below the standard of 1 infection per 100 person-days at risk, you can use the Poisson distribution. The Poisson distribution is appropriate for modeling the number of rare events, such as infections in a hospital, over a known interval of time.

Here's how to calculate the p-value for this test:

1. Calculate the expected number of infections under the standard rate:
   Standard infection rate = 1 infection per 100 person-days.
   Expected infections = (1787) * (1/100) = 17.87

2. Use the Poisson distribution to find the probability of observing 10 or fewer infections when the expected number is 17.87. The Poisson probability mass function is:

$$P(X = x) = \frac{e^{-\lambda} * \lambda^x}{x!}$$

3. Calculate the cumulative probability of observing 10 or fewer infections:

$$P(X \leq 10) = \sum_{x=0}^{10} \frac{e^{-17.87} * 17.87^x}{x!}$$

4. Find the p-value, which is the probability of observing 10 or fewer infections:

P(X ≤ 10) = 0.033

So, the p-value for the one-sided test of whether the hospital is below the standard infection rate of 1 infection per 100 person-days at risk is approximately 0.033. This p-value indicates strong evidence that the hospital's infection rate is below the standard, as it is smaller than a typical significance level α such as 0.05.

99. *In a population of interest, a sample of 9 men yielded a sample average brain volume of 1,100cc and a standard deviation of 30cc. What is a 95% Student's T confidence interval for the mean brain volume in this new population?*

Ans: To calculate a 95% Student's T confidence interval for the mean brain volume in the new population based on a sample of 9 men with a sample average of 1,100cc and a standard deviation of 30cc, you can use the formula for the confidence interval:

$$CI = \bar{x} \pm t\left(\frac{s}{\sqrt{n}}\right)$$

Where:

- x̄ is the sample mean (1,100cc).
- s is the sample standard deviation (30cc).
- n is the sample size 9.
- t is the critical value from the Student's T-distribution for the desired confidence level.

For a 95% confidence interval and 8 degrees of freedom (n - 1 = 9 - 1 = 8), you can find the critical value, t, from the T-distribution table or using a statistical calculator. For a 95% confidence level and 8 degrees of freedom, t is approximately 2.306.

Therefore, CI = 1,100 + 2.306*10

Lower Limit: 1,100 - 23.06 = 1,076.94cc

Upper Limit: 1,100 + 23.06 = 1,123.06cc

So, the 95% Student's T confidence interval for the mean brain volume in the new population is approximately 1,076.94cc to 1,123.06cc. This means we are 95% confident that the true population mean brain volume falls within this interval.

100. What Chi-square test?

Ans: A statistical method is used to find the difference or correlation between the observed and expected categorical variables in the dataset.

Example: A food delivery company wants to find the relationship between gender, location and food choices of people in India.

It is used to determine whether the difference between 2 categorical variables is:

- Due to chance or
- Due to relationship

101. What is the ANOVA test?

Ans: ANOVA, or Analysis of Variance, is a statistical technique used to analyze the variation between multiple groups or treatments to determine if there are statistically significant differences among the group means. ANOVA is particularly useful when you have more than two groups to compare.

102. How to calculate p-value using a manual method?

Ans: The p-value is the probability of observing a test statistic as extreme as or more extreme than the one you calculated, assuming the null hypothesis is true.
To calculate it:
We can make use of the Cumulative Distribution Function to calculate the required probabilities.
For one-sided tests:

- If your test statistic is in the right tail of the distribution, find the probability of values as extreme or greater than your statistic.
- If your test statistic is in the left tail, find the probability of values as extreme or smaller than your statistic.

For two-sided tests:

- Find the probability of values as extreme as or more extreme than your statistic in both tails.

103. What do we mean by – making a decision based on comparing p-value with significance level? What is the goal of A/B testing?

Ans: If p-value ≤ α: You reject the null hypothesis. This means that the observed data provides strong evidence against the null hypothesis, and you conclude that there is a statistically significant effect or difference in the data.

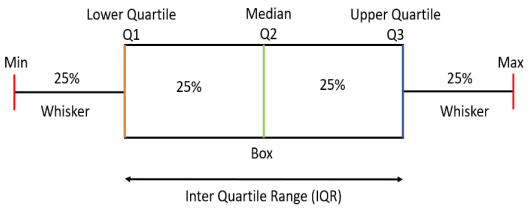
If p-value > α:
You fail to reject the null hypothesis. This means that the observed data does not provide strong enough evidence to conclude that there is a statistically significant effect or difference. It does not prove that the null hypothesis is true; it simply means you don't have enough evidence to reject it.

Goal of A/B Testing: The primary goal of A/B testing (also known as split testing) is to compare two or more versions of a web page, app, or marketing campaign to determine which one performs better in terms of a specific outcome or metric. Common objectives include improving conversion rates, click-through rates, user engagement, or other key performance indicators (KPIs).

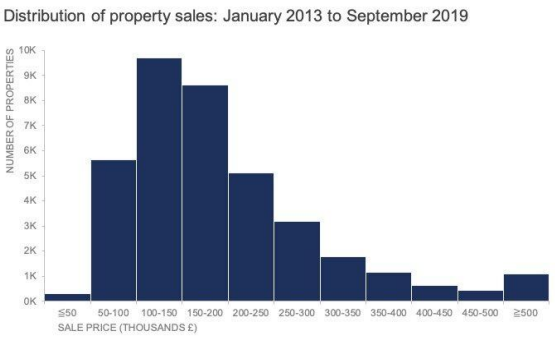
104. What is the difference between a box plot and a histogram?

Ans: Box plot summarizes the data in 5 numbers. Whereas histogram shows the complete distribution of data when appropriate bin size is used. Box Plot is more generally used to compare the distributions. Both the plots help in identifying the outliers.

Box Plot



Histogram



105. A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random, and toss it 10 times. Given that you see 10 heads, what is the probability that the next toss of that coin is also a head?

Ans: We use Bayes Theorem to find the answer. Let's split problem into two parts:
1) What is the probability you picked the double-headed coin (referred as D)
2) What is the probability of getting a head on the next toss?
Question 2 follows very naturally after question 1, so let's tackle question 1.

We are trying to find the probability of having a double-headed coin. We know that the same coin has been flipped 10 times, and we've gotten 10 heads (intuitively, you're probably thinking that there is a significant chance we have the double-headed coin). Formally, we're trying to find P(D | 10 heads).

Using Bayes rule:

$$P(D \mid 10\,H) = \frac{P(10\,H \mid D) * P(D)}{P(10H)}$$

Tackling the numerator, the prior probability, P(D) = 1/1000. If we used the double headed coin, the chance of getting 10 heads, P(10 H | D) = 1 (we always flip heads). So the numerator = 1 / 1000 * 1 = 1 / 1000.

The denominator, P(10H) is just P(10 H | D) * P(D) + P(10 H | Fair) * P(Fair). This makes sense because we are simply enumerating over the two possible coins. The first part of P(10H) is the exact same as the numerator (1 / 1000). Then the second part: P(Fair) = 999/1000. P(10 H | Fair) = (1/2) ^ 10 = 1/1024. Thus P(10 H | Fair) * P(Fair) = .0009756. The denominator then equals .001 + .0009756.

Since we have all the components of P(D | 10 H), compute and you'll find the probability of having a double headed coin is .506. We have finished the first question.

The second question is then easily answered: we just compute the two individual possibilities and add.

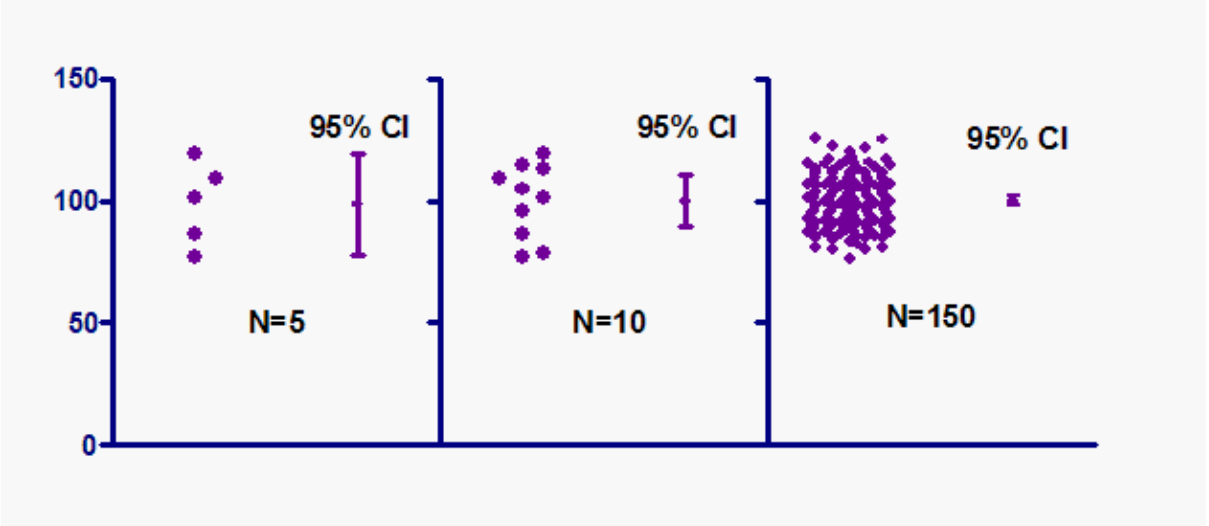
P(H) = P(D) * P(H | D) + P(Fair) * P(H | Fair) = .506 * 1 + (1 - .506) * (.5) = .753.

So there is a **75.3%** chance you will flip a heads.

106. What is a confidence interval and how do you interpret it?

Ans: A confidence interval gives a range where we think a certain number (like an average) lies for the whole population, based on our sample data.

The following graph shows three samples (of different size) all sampled from the same population.



With the small sample on the left, the 95% confidence interval is similar to the range of the data. But only a tiny fraction of the values in the large sample on the right lie within the confidence interval. This makes sense. The 95% confidence interval defines a range of values that you can be 95% certain contains the population mean. With large samples, you know that mean with much more precision than you do with a small sample, so the confidence interval is quite narrow when computed from a large sample.

107. How do you stay up-to-date with the new and upcoming concepts in statistics?

Ans: By following the experts and influencers who share their insights, opinions, and best practices on various platforms. We can follow them on social media, blogs, podcasts, newsletters, webinars, or online communities

108. What is correlation?

Ans: Correlation is a statistical measure that describes the degree to which two variables are related or move together in a linear fashion. It quantifies the strength and direction of the linear relationship between two continuous variables. Correlation does not imply causation; it simply indicates that there is a statistical association between the variables.
1. Strength of Relationship: Correlation coefficients range from -1 to 1. The absolute value of the correlation coefficient indicates the strength of the relationship:
1: Perfect positive correlation. When one variable increases, the other increases proportionally.
0: No correlation. The variables are not related.
-1: Perfect negative correlation. When one variable increases, the other decreases proportionally.

2. Direction of Relationship: The sign of the correlation coefficient (+ or -) indicates the direction of the relationship:
Positive correlation: As one variable increases, the other tends to increase.
Negative correlation: As one variable increases, the other tends to decrease.

3. Correlation Coefficient: The most common correlation coefficient is the Pearson correlation coefficient (often denoted as r. It measures the linear relationship between two variables. It's calculated as:

$$\frac{\Sigma[(X - X_m) * (Y - Y_m)]}{\sqrt{[\Sigma(X - X_m)^2 * \Sigma(Y - Y_m)^2]}}$$

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

Ans: To get Correlation of two variables they must be numerical variables and outliers must be avoided to get precise correlation between them.
Quantities used to evaluate coefficient of correlation are

1. Covariance 2. Standard Deviation

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

Ans: A high correlation between the time a person sleeps and the amount of productive work they do suggests a statistically significant relationship between these two variables.

1. Statistical Association: The high correlation indicates that there is a statistical association between the time spent sleeping and the amount of productive work. In other words, as one variable (sleep) changes, there tends to be a corresponding change in the other variable (productive work).

2. Direction of the Relationship: If the correlation is positive, it suggests that as the amount of sleep increases, productive work tends to increase as well. Conversely, if the correlation is negative, it implies that as the amount of sleep decreases, productive work tends to decrease.

3. Strength of the Relationship: A high correlation coefficient (close to +1 or -1) suggests a strong linear relationship between the variables. This means that changes in one variable are closely related to changes in the other variable. A lower correlation coefficient (closer to 0) would indicate a weaker relationship.

However, it's important to clarify the nature of the relationship and be cautious about making causal inferences.

111. What is the meaning of covariance?

Ans: Covariance is a statistical measure that assesses whether an increase in one variable is associated with an increase or decrease in another variable. In other words, covariance measures the joint variability of two variables. It ranges from $(-\infty,+\infty)$. We prefer to use coefficient of correlation instead of covariance, since covariance is sensitive to scale.

112. What does autocorrelation mean?

Ans: Autocorrelation, also known as serial correlation, refers to the correlation between the values of a single variable at different time points or lags. In other words, it measures the degree to which a variable is correlated with itself over time, with each observation being compared to previous observations in a time series data set.

113. How will you determine the test for the continuous data?

Ans: Determining the appropriate statistical test for continuous data depends on several factors, including the research question, the nature of the data, and the specific hypothesis you want to test.

1. Two-Sample T-Test:
Use case: Use a two-sample t-test when you want to compare the means of two independent groups or samples. It's suitable for testing whether there is a statistically significant difference between the means of two continuous variables.

2. Paired T-Test:
Use case: Use a paired t-test when you want to compare the means of two related or paired groups. It's suitable when the same subjects or items are measured before and after an intervention or treatment.

3. Analysis of Variance (ANOVA):
Use case: ANOVA is used when you want to compare the means of three or more independent groups or conditions. It helps determine whether there are statistically significant differences among the group means.
Always conduct exploratory data analysis (EDA) to understand your data's characteristics before choosing a statistical test. If in doubt, consulting with a statistician or data analyst can be valuable in selecting the most appropriate test for your analysis.

114. What can be the reason for non normality of the data?

Ans: Non-normality in data, where the distribution of data points deviates from a normal (Gaussian) distribution, can be caused by various factors and phenomena. Understanding the reasons for non-normality is important for appropriate data analysis and modeling. Some reasons for non-normality in data can be:

1. Skewness
2. Outliers
3. Bimodality
4. Heavy tails
5. Transformation

115. why is there no such thing like 3 samples t- test?? why t-test failed with 3 samples

Ans: There's no three-sample t-test because it would not address the complexities of comparing multiple groups and the increased risk of Type I errors associated with multiple pairwise comparisons. Instead, ANOVA is the appropriate statistical test for comparing means across three or more groups, and it provides a more comprehensive analysis of group differences.
The reason why t-test fails with 3 samples is
When you have three or more groups, using a series of t-tests to compare each pair of groups (i.e., pairwise t-tests) is possible, but it can lead to an increased risk of making a Type I error (false positives). This is because with multiple tests, the probability of finding a significant result by chance increases.

Lack of Overall Conclusion: Conducting multiple t-tests doesn't provide an overall conclusion about whether at least one group is different from the others. It only tells you which specific pairs of groups are different, but it doesn't address the broader question of group differences in general.

Resources :

1. Statistics in one shot :

Complete Statistics For Data Science In 6 hours By Krish Naik

2. Stats In English Detailed Playlist:

How to Learn Statistics for Data Science As A Self Starter- Follow My Way

3. Stats in Hindi Detailed Playlist:

Starter Roadmap For Learning Statistics For Data Analyst & Data Science In ...