Statistics Questions for Interview

1. How is the statistical significance of an insight assessed?

Hypothesis testing is used to find out the statistical significance of the insight. To elaborate, the null hypothesis and the alternate hypothesis are stated, and the p-value is calculated.

After calculating the p-value, the null hypothesis is assumed true, and the values are determined. To fine-tune the result, the alpha value, which denotes the significance, is tweaked. If the p-value turns out to be less than the alpha, then the null hypothesis is rejected. This ensures that the result obtained is statistically significant.

2. Where are long-tailed distributions used?

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

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

3. What is the central limit theorem?

The central limit theorem states that the normal distribution is arrived at when the sample size varies without having an effect on the shape of the population distribution.

This central limit theorem is the key because it is widely used in performing hypothesis testing and also to calculate the confidence intervals accurately.

4. What is observational and experimental data in Statistics?

Observational data correlates to the data that is obtained from observational studies, where variables are observed to see if there is any correlation between them.

Experimental data is derived from experimental studies, where certain variables are held constant to see if any discrepancy is raised in the working.

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

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

It is considered a bad practice as it completely removes the accountability for feature correlation. This also means that the data will have low variance and increased bias, adding to the dip in the accuracy of the model, alongside narrower confidence intervals.

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

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

Outliers are determined by using two methods:

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

7. How is missing data handled in statistics?

There are many ways to handle missing data in Statistics:

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

8. What is exploratory data analysis?

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

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

9. What is the meaning of selection bias?

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

If correct randomization is not achieved, then the resulting sample will not accurately represent the population.

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

There are many types of selection bias as shown below:

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

11. What is the meaning of an inlier?

An inlier is a data point that lies at the same level as the rest of the dataset. Finding an inlier in the dataset is difficult when compared to an outlier as it requires external data to do so. Inliers, similar to outliers, reduce model accuracy. Hence, even they are removed when they're found in the data. This is done mainly to maintain model accuracy at all times.

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

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

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

Now, there are 7 ways of rolling an 8 (1+7, 7+1, 2+6, 6+2, 3+5, 5+3, 4+4)

P(Getting an 8) = 7/36 = 0.194

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

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

14. Can you give an example of root cause analysis?

Root cause analysis, as the name suggests, is a method used to solve problems by first identifying the root cause of the problem.

Example: If the higher crime rate in a city is directly associated with the higher sales in a red-colored shirt, it means that they are having a positive correlation. However, this does not mean that one causes the other.

Causation can always be tested using A/B testing or hypothesis testing.

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

Six sigma is a quality assurance methodology used widely in statistics to provide ways to improve processes and functionality when working with data.

A process is considered as six sigma when 99.99966% of the outcomes of the model are considered to be defect-free.

16. What is DOE?

DOE is an acronym for the Design of Experiments in statistics. It is considered as the design of a task that describes the information and the change of the same based on the changes to the independent input variables.

17. What is the meaning of KPI in statistics?

KPI stands for Key Performance Analysis in statistics. It is used as a reliable metric to measure the success of a company with respect to its achieving the required business objectives.

There are many good examples of KPIs:

- Profit margin percentage
- Operating profit margin
- Expense ratio

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

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

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

19. What is the Pareto principle?

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

A simple example of the Pareto principle is the observation that 80 percent of peas come from 20 percent of pea plants on a farm.

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

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

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

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

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

In short:

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

22. What are quantitative data and qualitative data?

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

23. What is Mean?

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

24. What is the meaning of standard deviation?

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

25. What is a bell-curve distribution?

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

26. What is skewness?

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

27. What is kurtosis?

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

28. What is correlation?

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

The -1 value represents a high negative correlation, i.e., if the value in one variable increases, then the value in the other variable will drastically decrease. Similarly, +1 means a positive correlation, and here, an increase in one variable will lead to an increase in the other. Whereas, 0 means there is no correlation.

If two variables are strongly correlated, then they may have a negative impact on the statistical model, and one of them must be dropped.

Next up on this top Statistics Interview Questions and Answers blog, let us take a look at the intermediate set of questions.

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

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

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

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

Descriptive Statistics: Descriptive statistics is used to summarize a sample set of data like the standard deviation or the mean.

Inferential statistics: Inferential statistics is used to draw conclusions from the test data that are subjected to random variations.

31. What are the types of sampling in Statistics?

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

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

32. What is the meaning of covariance?

Covariance is the measure of indication when two items vary together in a cycle. The systematic relation is determined between a pair of random variables to see if the change in one will affect the other variable in the pair or not.

33. 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?

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

> $X = \mu + Z\sigma$
>
> Here:
>
> μ: Mean
>
> σ: Standard deviation
>
> X: Value to be calculated

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

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

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

35. What is Bessel's correction?

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

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

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

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

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

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

The significance level is the probability of obtaining a result that is extremely different from the condition where the null hypothesis is true. While the confidence level is used as a range of similar values in a population.

Both significance and confidence level are related by the following formula:

Significance level = 1 − Confidence level

39. A regression analysis between apples (y) and oranges (x) resulted in the following least-squares line: y = 100 + 2x. What is the implication if oranges are increased by 1?

If the oranges are increased by one, there will be an increase of 2 apples since the equation is:

> y = 100 + 2x.

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

Variables to be used for the Pearson's correlation coefficient must be either in a ratio or in an interval.

Note that there can exist a condition when one variable is a ratio, while the other is an interval score.

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

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

42. What are the examples of symmetric distribution?

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

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

- Uniform distribution
- Binomial distribution
- Normal distribution

43. Where is inferential statistics used?

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

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

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

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

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

That is,

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

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

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

47. What is one sample t-test?

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

48. What is an alternative hypothesis?

The alternative hypothesis (denoted by H1) is the statement that must be true if the null hypothesis is false. That is, it is a statement used to contradict the null hypothesis. It is the opposing point of view that gets proven right when the null hypothesis is proven wrong.

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

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

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

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

- The selection bias
- The survivorship bias
- The under coverage bias

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

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

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

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

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

- Define the confidence level (usually around 95%)
- Use sample sharks to measure
- Calculate the mean and standard deviation of the lengths
- Determine t-statistics values
- Determine the confidence interval in which the mean length lies

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

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

The following also apply:

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

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

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

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

55. How can you calculate the p-value using MS Excel?

Following steps are performed to calculate the p-value easily:

- Find the Data tab above
- Click on Data Analysis
- Select Descriptive Statistics
- Select the corresponding column

- Input the confidence level

56. What is the law of large numbers in statistics?

The law of large numbers in statistics is a theory that states that the increase in the number of trials performed will cause a positive proportional increase in the average of the results becoming the expected value.

Example: The probability of flipping a fair coin and landing heads is closer to 0.5 when it is flipped 100,000 times when compared to 100 flips.

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

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

Following are some of the important properties:

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

58. 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)?

The probability of not seeing a supercar in 20 minutes is:

> = 1 − P(Seeing one supercar)
>
> = 1 − 0.3
>
> = 0.7

Probability of not seeing any supercar in the period of 60 minutes is:

> = (0.7) ^ 3 = 0.343

Hence, the probability of seeing at least one supercar in 60 minutes is:

> = 1 − P(Not seeing any supercar)

$= 1 - 0.343 = 0.657$

59. What is the meaning of sensitivity in statistics?

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

The simple formula to calculate sensitivity is:

Sensitivity = Predicted True Events/Total number of Events

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

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

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

For reducing underfitting:

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

For reducing overfitting:

- Increase training data
- Stop early while training
- Lasso regularization
- Use random dropouts

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

Let's consider the population of men who have normally distributed weights, with a mean of 60 kg and a standard deviation of 10 kg, and the probability needs to be found out.

If one single man is selected, the weight is greater than 65 kg, but if 40 men are selected, then the mean weight is far more than 65 kg.

The solution to this can be as shown below:

> $Z = (x - \mu) / ? = (65 - 60) / 10 = 0.5$
>
> For a normal distribution $P(Z > 0.5) = 0.409$
>
> $Z = (65 - 60) / 5 = 1$
>
> $P(Z > 1) = 0.090$

62. What is linear regression?

In statistics, linear regression is an approach that models the relationship between one or more explanatory variables and one outcome variable. For example, linear regression can be used to quantify or model the relationship between various predictor variables such as age, gender, genetics, and diet on height, outcome variables.

63. What are the assumptions required for linear regression?

Four major assumptions for linear regression are as under –

- There's a linear relationship between the predictor (independent) variables and the outcome (dependent) variable. It means that the relationship between X and the mean of Y is linear.
- The errors are normally distributed with no correlation between them. This process is known as Autocorrelation.
- There is an absence of correlation between predictor variables. This phenomenon is called multicollinearity.
- The variation in the outcome or response variable is the same for all values of independent or predictor variables. This phenomenon of assumption of equal variance is known as homoscedasticity.

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

Some of the widely used low and high-bias Machine Learning algorithms are –

Low bias -Decision trees, Support Vector Machines, k-Nearest Neighbors, etc.

High bias -Linear Regression, Logistic Regression, Linear Discriminant Analysis, etc.

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

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.

66. What is the equation for confidence intervals for means vs for proportions?

To calculate the confidence intervals for mean, we use the following equation –

For n > 30

Use the Z table for the standard normal distribution.

For n<30

Use the t table with df=n-1

67. What is the empirical rule?

In statistics, the empirical rule states that every piece of data in a normal distribution lies within three standard deviations of the mean. It is also known as the 68–95–99.7 rule. According to the empirical rule, the percentage of values that lie in a normal distribution follow the 68%, 95%, and 99.7% rule. In other words, 68% of values will fall within one standard deviation of the mean, 95% will fall within two standard deviations, and 99.75 will fall within three standard deviations of the mean.

68. How are confidence tests and hypothesis tests similar? How are they different?

Confidence tests and hypothesis tests both form the foundation of statistics.

The confidence interval holds importance in research to offer a strong base for research estimations, especially in medical research. The confidence interval provides a range of values that helps in capturing the unknown parameter.

Hypothesis testing is used to test an experiment or observation and determine if the results did not occur purely by chance or luck using the below formula where ‘p’ is some parameter.

Confidence and hypothesis testing are inferential techniques used to either estimate a parameter or test the validity of a hypothesis using a sample of data from that data set. While the confidence interval provides a range of values for an accurate estimation of the precision of that parameter, hypothesis testing tells us how confident we are inaccurately drawing conclusions about a parameter from a sample. Both can be used to infer population parameters in tandem.

In case we include 0 in the confidence interval, it indicates that the sample and population have no difference. If we get a p-value that is higher than alpha from hypothesis testing, it means that we will fail to reject the null hypothesis.

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

Here are the conditions that must be satisfied for the central limit theorem to hold –

- The data must follow the randomization condition which means that it must be sampled randomly.
- The Independence Assumptions dictate that the sample values must be independent of each other.
- Sample sizes must be large. They must be equal to or greater than 30 to be able to hold CLT. Large sample size is required to hold the accuracy of CLT to be true.

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

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.

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

A population in inferential statistics refers to the entire group we take samples from and are used to draw conclusions. A sample, on the other hand, is a specific group we take data from and this data is used to calculate the statistics. Sample size is always less than that of the population.

Q72. What are the different methods to detect outliers in a dataset?

There are mainly 3 ways to detect outliers in a dataset:

- **Box-Plot**
    - Data points are divided into 4 different quartiles.
    - Box-plot marks Maximum, Minimum, lower quartile (Q1), median (Q2) and upper quartile (Q3).
    - Points outside the whisker are Outliers.
- **InterQuartile Range**
    - Arrange the data orderly (ascending)
    - Compute IQR = Q3 – Q1
    - Calculate bound (upper and lower) 1.5 IQR
    - Any point outside the upper and lower bound are the outlier.
- **Z-score**

In a normal distribution, any data point whose z-score is outside the 3rd standard deviation is an outlier.

73. What is the difference between data mining and data analysis?

Data Mining	Data Analysis
It refers to the process of identifying patterns in a pre-built database.	It is used to order and organize raw data in a meaningful manner.

Data mining is done on clean and well-documented data.	Data analysis involves cleaning the data hence it is not presented in a well-documented format.
The outcomes are not easy to interpret.	The outcomes are easy to interpret.
It is mostly used for Machine Learning where used to recognize the patterns with the help of algorithms.	It is used to gather insights from raw data, which has to be cleaned and organized before performing the analysis.

74. What is the Difference between Data Profiling and Data Mining?

Data Profiling: It refers to the process of analyzing individual attributes of data. It primarily focuses on providing valuable information on data attributes such as data type, frequency, length, occurrence of null values.
Data Mining: It refers to the analysis of data with respect to finding relations that have not been discovered earlier. It mainly focuses on the detection of unusual records, dependencies and cluster analysis.

75. What is the Process of Data Analysis?

Data analysis is the process of collecting, cleansing, interpreting, transforming, and modeling data to gather insights and generate reports to gain business profits. Refer to the image below to know the various steps involved in the process.

- Collect Data: The data is collected from various sources and stored to be cleaned and prepared. In this step, all the missing values and outliers are removed.
- Analyse Data: Once the data is ready, the next step is to analyze the data. A model is run repeatedly for improvements. Then, the model is validated to check whether it meets the business requirements.
- Create Reports: Finally, the model is implemented, and then reports thus generated are passed onto the stakeholders.

76. What is Data Wrangling or Data Cleansing/Cleaning?

Data Cleansing is the process of identifying and removing errors to enhance the quality of data. We must check for the following things and correct where needed:

- Are all variables as expected (variables names & variable types).
- Are there some variables that are unexpected?
- Are the data types and length across variables correct?
- For known variables, is the data type as expected (For example if age is in date format something is suspicious)
- Have labels been provided and are sensible?
- If anything is suspicious we can further investigate it and correct it accordingly.

77. What are Some of the Challenges You Have Faced during Data Analysis?

List out all the challenges you have had come across while analysing and cleaning the data. Here are some of the common challenges in a typical Data Analytics project:

- Poor quality of data, with lots of missing and erroneous values
- Lack of understanding of the data, variables, and availability data dictionary
- Unrealistic timelines and expectation from the business stakeholders
- Challenge in blending/ integrating the data from multiple sources, particular when there no consistent parameters and conventions
- Wrong selection of tools and data architecture to achieve analytics goals in a timely manner

78. What do you Understand by the Term Normal distribution?

- It is a continuous symmetric distribution for which mean, median, mode are all equal. It is a symmetric distribution which’s why it is normal.
- The distribution is symmetric on the y-axis and is bisected by the mean.

- The tails of the curve extend to infinity.
- Its mean and standard deviation differentiates the entire family of normal probability distributions.
- The highest point of the distribution is the mean which is also the median and mode of the distribution.

79. What is the Difference Between Uni-variate, Bi-variate, and Multivariate Analysis?

Univariate Analysis: It is the simplest form of analysis as this type of data consists of only one variable and hence the information deals with only one quantity that changes. It does not deal with the causes or relationships between the variables and the primary purpose of the Univariate analysis is to describe the data and find the patterns that exist within it. Example: univariate analysis of age or height. Either age or height – is only one variable and doesn't deal with cause or relationships.
The variable can be described using Measures Of Central Tendency (Mean, Median, and Mode), and the variation in the data or spread of the data can be checked by Measures of Dispersion (Range, Min, Max, Interquartile Range, Quartiles, Variance, Mod of Absolute Deviation and Standard Deviation). Frequency distribution tables can be made, and they can be visualized using: Histogram.

Bivariate Analysis: This type of data involves two different variables. The analysis of this type of data deals with the causes and relationships between the variables. The primary purpose of the Bivariate analysis is to find the relationship among the two variables. Example: Sale of AC/Cooler in the Summer season.
The visual depiction of the relationships among the two variables is done via Scatter Plot where these variables are plotted on the X and Y axis, and one of these variables (in our case, Sales of AC or Cooler, which is plotted on the Y-axis) is dependent on the other independent variables(e.g., Summer season which is plotted on X-axis)

Multivariate Analysis: When there are more than two variables, i.e., three or more, the analysis is categorized as Multivariate analysis. Example: A Telecom Service Provider wants to compare their 4 tariff plans, measure how it is used between female and male, and

Location-wise, and examine the relationships between these variables. The techniques used are Regression, ANOVA.

80. What is the Difference Between Mean, Median, and Mode? Which one do you prefer to use and why?

Mean (or average) is the numerical value of the centre of distribution and used when the data is concentrated)
Median (also known as the 50th percentile) is the middle observation in a data set. Median is calculated by sorting the data, followed by the selection of the middle value. The median of a data set has an odd number of observations is the observation number [N + 1] / 2
For data sets having an even number of observations, the median is midway between N / 2 and [N / 2] + 1. N is the number of observations.

A mode is a value that appears most frequently in a data set. A data set may have single or multiple modes, referred to as unimodal, bimodal, or trimodal, depending on the number of modes.
If outliers (extreme values) or of a skewed data set (when one tail is significantly longer in a bell-shaped curve), the median is more applicable and preferred over mean. Example: If you want to represent the centre of a distribution, such as in the case of the marks of the class and one student has a significantly lower mark, using a median is more appropriate as the mean will pull down the average of the class.

81. What is the Difference Between Covariance and Correlation?

Covariance measures the variance of the variable with itself, and correlation measures the strength and direction of a linear relationship between two or more variables. A correlation between two variables doesn’t imply that the change in one variable is the cause of the change in the other variable’s values. Correlation is the scaled version of covariance as it is unit-free and can be directly used in comparisons like corr(X1, Y1) > corr(X2, Y2). On the other hand, covariance cannot be directly compared this way.

82. What is Hypothesis Testing, Why is it Needed, and List Some of the Statistical Tests?

Hypothesis testing is the process in which statistical tests are used to check whether or not a hypothesis is true, using data. Based on hypothetical testing, we choose to accept or reject a hypothesis. It is needed to check whether the event is the result of a significant occurrence or merely of chance, hypothesis testing must be applied.

Some of the statistical tests are:
T-test: It is used to compare the means of two populations that is whether the given mean is significantly different from the sample mean or not. It can also be used to ascertain whether the regression line has a slope different from zero.
F-Test: It is used to determine the equality of the variances of the two normal populations. It can be also used to check if the data conforms to the regression model. In the Multiple Linear Regression model, examines the overall validity of the model or determines whether any of the independent variables is having a linear relationship with the dependent variable.
Chi-Square: It is used to check whether there is any statistically significant difference between the observed distribution and theoretical distribution.
ANOVA: It tests the equality of two or more population means by examining the variances of samples that are taken. ANOVA tests the general rather than specific differences among means.

83. What is A/B Testing?

A/B testing is dual-variable two-sample hypothesis testing performed on randomized experiments to determine which variation is better as compared to the other. In a user-experience design, we would want to identify changes to web pages that increased the clicks on a banner. The null hypothesis is there is no change or variation and the alternative hypothesis is there is variation, that is the clicks on the banner increased post the change in the design of the website.

84. Random number generator

There is an ideal random number generator, which given a positive integer M can generate any **real number** between 0 to M, and probability density function is uniform in [0, M].

Given two numbers A and B and we generate *x* and *y* using the random number generator with uniform probability density function [0, A] and [0, B] respectively, what's the probability that x + y is less than C? where C is a positive integer.

Input Format

The first line of the input is an integer N, the number of test cases.

N lines follow. Each line contains 3 positive integers A, B and C.

Constraints

All the integers are no larger than 10000.

Output Format

For each output, output a fraction that indicates the probability. The greatest common divisor of each pair of numerator and denominator should be 1.

Sample Input

```
3
1 1 1
1 1 2
1 1 3
```

Sample Output

```
1/2
1/1
1/1
```

Change Theme

Code:

```
def solve(a, b, c):
    M = max(a, b)
    m = min(a, b)

    A = max(0, min(M, c - m))
    B = max(0, min(c, M))
    H = max(0, min(c, m))

    nom = (A + B) * H
    denom = 2 * M * m
    gcd = math.gcd(nom, denom)
    return '%s/%s' % (nom // gcd, denom // gcd)
```

85. Write a python code for calculating the mean of observations.

In Python, you can use the mean() method of the numpy library to calculate the mean of a list of observations:

```
import numpy as np

observations = [1, 2, 3, 4, 5]
mean = np.mean(observations)
print(mean)
```

or

```
observations = [1, 2, 3, 4, 5]
mean = sum(observations) / len(observations)
print(mean)
```

86. How to calculate median of observations in python?

In Python, you can use the **median()** method of the **numpy** library to calculate the median of a list of observations:

```
import numpy as np

observations = [1, 2, 3, 4, 5]
median = np.median(observations)
print(median)
```

or

```
observations = [1, 2, 3, 4, 5]
observations.sort()

if len(observations) % 2 == 0:
    median = (observations[len(observations)//2 - 1] + observations[len(observations)//2]) / 2
else:
    median = observations[len(observations)//2]

print(median)
```

87. How to calculate correlation in python?

In Python, you can use the corrcoef() method of the numpy library to calculate the Pearson correlation coefficient between two sets of observations:

```
import numpy as np

x = [1, 2, 3, 4, 5]
y = [5, 4, 3, 2, 1]
correlation = np.corrcoef(x, y)[0, 1]
print(correlation)
```

or

```
from scipy.stats import pearsonr

x = [1, 2, 3, 4, 5]
y = [5, 4, 3, 2, 1]
correlation, p_value = pearsonr(x, y)
print("Correlation:", correlation)
print("P-value:", p_value)
```

88. How to calculate statistical values of a data set in pandas?

In Pandas, you can use the following methods to calculate various statistical values of a data set:

- Mean: mean() method of the DataFrame or Series.
- Median: median() method of the DataFrame or Series.
- Mode: mode() method of the DataFrame or Series.
- Standard deviation: std() method of the DataFrame or Series.
- Variance: var() method of the DataFrame or Series.
- Correlation: corr() method of the DataFrame.
- Covariance: cov() method of the DataFrame.

Here is an example of calculating mean, median, and standard deviation of a DataFrame:

```
import pandas as pd
import numpy as np

data = {'A': [1, 2, 3, 4, 5], 'B': [5, 4, 3, 2, 1]}
df = pd.DataFrame(data)

mean = df.mean()
median = df.median()
std = df.std()

print("Mean:\n", mean)
print("\nMedian:\n", median)
print("\nStandard deviation:\n", std)
```

89. How to calculate the mean of the last 30 observations?

```
import pandas as pd

df = pd.read_csv("observations.csv") # replace with your file path

mean = df.tail(30).mean()
print(mean)
```

Assuming the observations are stored in a CSV file named "observations.csv", this code will load the data into a Pandas DataFrame, select the last 30 rows using the **tail()** method, and calculate the mean using the **mean()** method.

90. Write a code to apply t-test on a data set?

In Python, you can use the ttest_ind() method of the scipy.stats library to apply a two-sample t-test on a data set:

```
from scipy.stats import ttest_ind

sample1 = [1, 2, 3, 4, 5]
```

```
sample2 = [5, 4, 3, 2, 1]

t_statistic, p_value = ttest_ind(sample1, sample2)

print("T-statistic:", t_statistic)
print("P-value:", p_value)
```

This will perform a two-sample t-test to determine if the means of the two samples are equal. The t-statistic measures the difference between the means of the two samples and the p-value gives the probability of observing a t-statistic as extreme as the one calculated, assuming the null hypothesis (the means of the two samples are equal) is true. If the p-value is less than a significance level (e.g. 0.05), you can reject the null hypothesis and conclude that the means of the two samples are significantly different.

91. Write a code to apply a Z-test on a data set?

In Python, you can use the z-score() method of the ***scipy.stats*** library to perform a one-sample z-test on a data set:

```
from scipy.stats import zscore
import numpy as np

sample = [1, 2, 3, 4, 5]
mean = 3

z_score = (np.mean(sample) - mean) / np.std(sample)

print("Z-score:", z_score)
```

This will calculate the z-score of the sample mean compared to a specified mean. The z-score is a standard normal deviate and measures the number of standard deviations the sample mean is from the specified mean.

Note that this is a one-sample z-test and assumes that the population standard deviation is known. If the population standard deviation is not known, you can use a t-test instead.

92. Write a Code for calculating conditional probability in python?

To calculate conditional probability on a data set in Python, you can use the Pandas library to count the number of events that meet specific conditions and divide by the total number of events:

```
import pandas as pd

df = pd.read_csv("data.csv") # replace with your file path

# Example: Calculate the conditional probability of event A given event B
event_a = df[df["B"] == 1]["A"].sum()
event_b = df[df["B"] == 1]["B"].count()

cond_prob = event_a / event_b
print("Conditional probability of A given B:", cond_prob)
```

Assuming the data set is stored in a CSV file named "data.csv", this code will load the data into a Pandas DataFrame, count the number of events where "B" is 1 using the **count()** method, count the number of events where "A" is 1 and "B" is 1 using the **sum()** method, and finally calculate the conditional probability by dividing the two values.

93. Write a code to calculate the mean by using a rolling function.

To calculate the mean using the rolling function in Python, you can use the Pandas library and the rolling().mean() method:

```
import pandas as pd

df = pd.read_csv("data.csv") # replace with your file path

# Example: Calculate the rolling mean of the "column_name" column with a window size of 30
rolling_mean = df["column_name"].rolling(window=30).mean()
print(rolling_mean)
```

Assuming the data set is stored in a CSV file named "data.csv", this code will load the data into a Pandas DataFrame, calculate the rolling mean of the "column_name" column with a window size of 30 using the **rolling().mean()** method, and print the result.

94. Write a code to apply a description command in pandas.

Here is a code in Python to apply the **describe()** command in Pandas:

```
import pandas as pd

df = pd.read_csv("data.csv") # replace with your file path

# Apply the describe() command to the DataFrame
desc = df.describe()
print(desc)
```

Assuming the data set is stored in a CSV file named "data.csv", this code will load the data into a Pandas DataFrame, apply the **describe()** command to the DataFrame, and print the result. The **describe()** command returns a summary of the central tendency, dispersion, and shape of the distribution of a set of continuous values, excluding **NaN** values by default.

95. Write a code to apply ANOVA test on a data set.

Here is a code in Python to perform an ANOVA test on a data set using the f_oneway() method from the scipy.stats library:

```
import pandas as pd
from scipy.stats import f_oneway
import numpy as np

df = pd.read_csv("data.csv") # replace with your file path

# Example: Perform an ANOVA test on the "column_name" column
group1 = df[df["group"] == 1]["column_name"]
group2 = df[df["group"] == 2]["column_name"]
group3 = df[df["group"] == 3]["column_name"]
```

```
f_value, p_value = f_oneway(group1, group2, group3)
print("F-value:", f_value)
print("P-value:", p_value)
```

Assuming the data set is stored in a CSV file named "data.csv", this code will load the data into a Pandas DataFrame, split the "column_name" column into three groups based on the values in the "group" column, perform an ANOVA test on the groups using the **f_oneway()** method, and print the F-value and P-value.

The F-value measures the ratio of the between-group variability to the within-group variability and the P-value tests the null hypothesis that the means of all groups are equal. A low P-value (e.g. less than 0.05) indicates that the means are significantly different, while a high P-value (e.g. greater than 0.05) indicates that the means are not significantly different.