STATISTICS WORKSHEET-4

STATISTICS ASSIGNMENT-4

Q1to Q15 are descriptive types. Answer in brief.

1. What is central limit theorem and why is it important?

ANS:

Statement of Central limit Theorem:

The central limit theorem states that if we have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample mean is asymptotically normal.

We can calculate the mean of the sample means for the random samples we choose from the population:

$\mu_{\bar{X}}=\mu$

As well as the standard deviation of sample means:

$\sigma_{\bar{X}}=\sigma n$

According to the central limit theorem, the form of the sampling distribution will approach normalcy as the sample size is sufficiently large (usually n>30). regardless of the population distribution.

Importance of Central Limit Theorem:

This is useful since the researcher never knows which mean in the sampling distribution corresponds to the population mean, but by taking numerous random samples from a population, the sample means will cluster together, allowing the researcher to obtain a very accurate estimate of the population mean.

2. What is sampling? How many sampling methods do you know?

ANS : Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger population. The method of sampling depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling.

Types of Audit Sampling

Random Sampling

With random sampling, every item within a population has an equal probability of being chosen. It is the furthest removed from any potential bias because there is no human judgement involved in selecting the sample.

For example, a random sample may include choosing the names of 25 employees out of a hat in a company of 250 employees. The population is all 250 employees, and the sample is random because each employee has an equal chance of being chosen.

Judgement Sampling

Auditor judgement may be used to select the sample from the full population. An auditor may only be concerned about transactions of a material nature. For example, assume the auditor sets the threshold for materiality for accounts payable transactions at $10,000. If the client provides a complete list of 15 transactions over $10,000, the auditor may just choose to review all transactions due to the small population size.

Alternatively, an auditor may identify all general ledger accounts with a variance greater than 10% from the prior period. In this case, the auditor is limiting the population from which the

sample selection is being derived. Unfortunately, human judgement used in sampling always comes with the potential for bias, whether explicit or implicit.

Block Sampling

Block sampling takes a consecutive series of items within the population to use as the sample. For example, a list of all sales transactions in an accounting period could be sorted in various ways, including by date or by dollar amount.

An auditor may request that the company's accountant provide the list in one format or the other in order to select a sample from a specific segment of the list. This method requires very little modification on the auditor's part, but it is likely that a block of transactions will not be representative of the full population.

Systematic Sampling

Systematic sampling begins at a random starting point within the population and uses a fixed, periodic interval to select items for a sample. The sampling interval is calculated as the population size divided by the sample size. Despite the sample population being selected in advance, systematic sampling is still considered random if the periodic interval is determined beforehand and the starting point is random.

Assume that an auditor is reviewing the internal controls related to a company's cash account and wants to test the company policy that stipulates that checks exceeding $10,000 must be signed by two people. The population consists of every company check exceeding $10,000 during the fiscal year, which, in this example, was 300. The auditor uses probability statistics and determines that the sample size should be 20% of the population or 60 checks. The sampling interval is 5 (300 checks / 60 sample checks).

Therefore, the auditor selects every fifth check for testing. Assuming no errors are found in the sampling test work, the statistical analysis gives the auditor a 95% confidence rate that the check procedure was performed correctly. The auditor tests the sample of 60 checks and finds no errors, so he concludes that the internal control over cash is working properly.

Example of Marketing Sampling

Businesses aim to sell their products and/or services to target markets. Before presenting products to the market, companies generally identify the needs and wants of their target audience. To do so, they may employ sampling of the target market population to gain a better understanding of those needs to later create a product and/or service that meets those needs. In this case, gathering the opinions of the sample helps to identify the needs of the whole.

3. What is the difference between type1 and typeII error?

ANS

Type I and Type II errors are subjected to the result of the null hypothesis. In case of type I or type-1 error, the null hypothesis is rejected though it is true whereas type II or type-2 error, the null hypothesis is not rejected even when the alternative hypothesis is true. Both the error type-i and type-ii are also known as "**false negative**". A lot of statistical theory rotates around the reduction of one or both of these errors, still, the total elimination of both is explained as a statistical impossibility.

Type I Error

A type I error appears when the null hypothesis (H_0) of an experiment is true, but still, it is rejected. It is stating something which is not present or a false hit. A type I error is often called a false positive (an event that shows that a given condition is present when it is absent). In words of community tales, a person may see the bear when there is none (raising a false alarm) where the null hypothesis (H_0) contains the statement: "There is no bear".

The type I error significance level or rate level is the probability of refusing the null hypothesis given that it is true. It is represented by Greek letter α (alpha) and is also known as alpha level. Usually, the significance level or the probability of type i error is set to 0.05 (5%), assuming that it is satisfactory to have a 5% probability of inaccurately rejecting the null hypothesis.

Type II Error

A type II error appears when the null hypothesis is false but mistakenly fails to be refused. It is losing to state what is present and a miss. A type II error is also known as false negative (where a real hit was rejected by the test and is observed as a miss), in an experiment checking for a condition with a final outcome of true or false.

A type II error is assigned when a true alternative hypothesis is not acknowledged. In other words, an examiner may miss discovering the bear when in fact a bear is present (hence fails in raising the alarm). Again, H0, the null hypothesis, consists of the statement that, "There is no bear", wherein, if a wolf is indeed present, is a type II error on the part of the investigator. Here, the bear either exists or does not exist within given circumstances, the question arises here is if it is correctly identified or not, either missing detecting it when it is present, or identifying it when it is not present.

The rate level of the type II error is represented by the Greek letter β (beta) and linked to the power of a test (which equals 1−β).

4. covariance in statistics?

ANS:

Covariance is **a statistical tool that is used to determine the relationship between the movements of two random variables**. When two stocks tend to move together, they are seen as having a positive covariance; when they move inversely, the covariance is negative

$$\text{Covariance} = \frac{\sum (\text{Return}_{ABC} - \text{Average}_{ABC}) * (\text{Return}_{XYZ} - \text{Average}_{XYZ})}{(\text{Sample Size}) - 1}$$

Covariance Formula.

Where:

- x_i = a given x value in the data set
- x_m = the mean, or average, of the x values
- y_i = the y value in the data set that corresponds with x_i
- y_m = the mean, or average, of the y values

5. Differentiate between univariate ,Biavariate,and multivariate analysis.

ANS :

The term **univariate analysis** refers to the analysis of one variable. You can remember this because the prefix “uni” means “one.”

The term **multivariate analysis** refers to the analysis of more than one variable. You can remember this because the prefix “multi” means “more than one.”

There are three common ways to perform **univariate analysis**:

1. Summary Statistics

- We can calculate measures of central tendency like the mean or median for one variable.
- We can also calculate measures of dispersion such as the standard deviation for one variable.

2. Frequency Distributions

- We can create a frequency distribution, which describes how often each value occurs for one variable.

3. Charts

- We can create charts like boxplots, histograms, density curves, etc. to visualize the distribution of values for one variable.

There are two common ways to perform **multivariate analysis**:

1. Scatterplot Matrix

- We can create a scatterplot matrix, which allows us to visualize the relationship between each pairwise combination of variables in a dataset.

2. Machine Learning Algorithms

- We can use a supervised learning algorithm to fit a model like multiple linear regression that quantifies the relationship between multiple predictor variables and a response variable.
- We can also use an unsupervised learning algorithm like principal components analysis to find structure and relationships between multiple variables in a dataset at once.

The following examples show how to perform both univariate and multivariate analysis with the following dataset:

Household ID	Household Size	Annual Income	Number of Pets
1	2	$37,000	0
2	4	$49,000	0
3	4	$58,000	1
4	1	$68,000	3
5	3	$61,000	2
6	5	$64,000	2
7	6	$79,000	1
8	4	$89,000	1
9	7	$104,000	1
10	2	$95,000	0

Note: When you analyze exactly two variables, this is referred to as bivariate analysis.

Example: How to Perform Univariate Analysis

We could choose to perform univariate analysis on any of the individual variables in the dataset.

For example, we may choose to perform univariate analysis on the variable **Household Size**:

Household ID	Household Size	Annual Income	Number of Pets
1	2	$37,000	0
2	4	$49,000	0
3	4	$58,000	1
4	1	$68,000	3
5	3	$61,000	2
6	5	$64,000	2
7	6	$79,000	1
8	4	$89,000	1
9	7	$104,000	1
10	2	$95,000	0

We can calculate the following measures of central tendency for Household Size:

- Mean (the average value): 3.8
- Median (the middle value): 4

These values give us an idea of where the “center” value is located.

We can also calculate the following measures of dispersion:

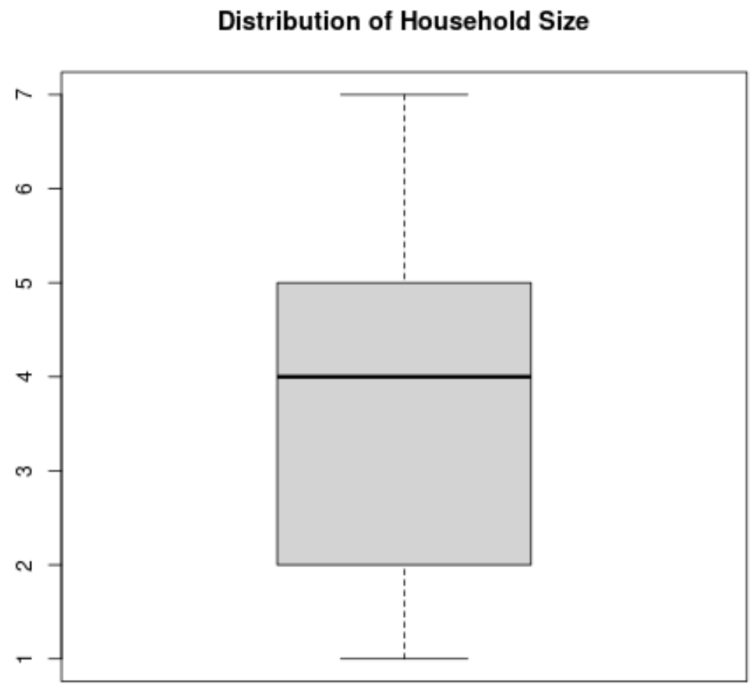
- Range (the difference between the max and min): 6
- Interquartile Range (the spread of the middle 50% of values): 2.5
- Standard Deviation (an average measure of spread): 1.87

These values give us an idea of how spread out the values are for this variable.

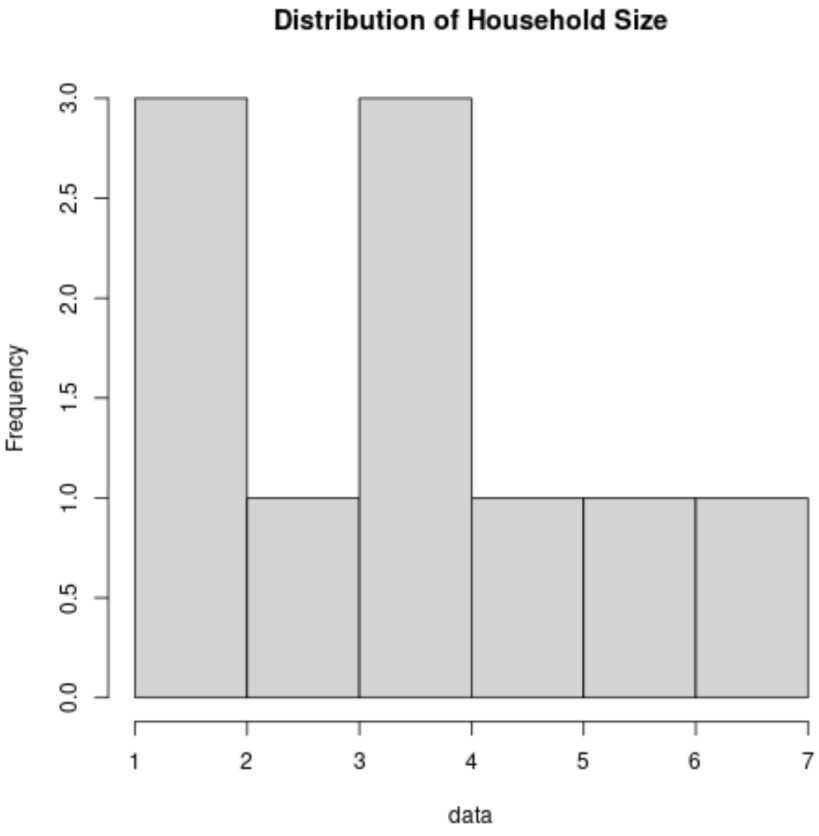
We can also create the following frequency distribution table to summarize how often different values occur:

Household Size	Frequency
1	1
2	2
3	1
4	3
5	1
6	1
7	1

We can also create a boxplot to visualize the distribution of values for household size:



Alternatively, we could create a histogram to visualize the distribution of values:



By calculating these metrics and creating these charts, we can gain a strong understanding of how the values are distributed for the variable Household Size.

Example: How to Perform Multivariate Analysis

Once again suppose we have the same dataset:

Household ID	Household Size	Annual Income	Number of Pets
1	2	$37,000	0
2	4	$49,000	0
3	4	$58,000	1
4	1	$68,000	3
5	3	$61,000	2
6	5	$64,000	2
7	6	$79,000	1
8	4	$89,000	1
9	7	$104,000	1
10	2	$95,000	0

One simple form of multivariate analysis we could perform on this dataset is to create a **scatterplot matrix**, which is a matrix that shows a scatterplot for each pairwise combination of numeric variables in the dataset.

We could create this type of matrix to visualize the relationship between household size, annual income, and number of pets all at once.

Resource: Check out this tutorial to see how to create a scatterplot matrix in R.

Another way to perform multivariate analysis on this dataset would be to fit a **multiple linear regression model**. For example, we could create a regression model that uses household size and number of pets to predict annual income.

Resource: Check out this tutorial to see how to perform multiple linear regression in R.

Yet another way to perform multivariate analysis on this dataset would be to perform **principal components analysis**, which allows us to find an underlying structure in the dataset.

Resource: Check out this tutorial to see how to perform principal components analysis in R.

Conclusion

Here's a quick summary of this article:

- Univariate analysis is the analysis of one variable.
- Multivariate analysis is the analysis of more than one variable.
- There are various ways to perform each type of analysis depending on your end goal.
- In the real world, we often perform both types of analysis on a single dataset.
- Univariate analysis allows us to understand the distribution of values for one variable while multivariate analysis allows us to understand the relationship between several variables.

6. What do you understand by sensitivity and how would you calculate it?

ANS:

Sensitivity analysis determines how different values of an independent variable affect a particular dependent variable under a given set of assumptions. In other words, sensitivity analyses study how various sources of uncertainty in a mathematical model contribute to the model's overall uncertainty. This technique is used within specific boundaries that depend on one or more input variables.

Sensitivity analysis is used in the business world and in the field of economics. It is commonly used by financial analysts and economists and is also known as a what-if analysis.

Sensitivity analysis is often performed in analysis software, and Excel has built in functions to help perform the analysis. In general, sensitivity analysis is calculated by leveraging formulas that reference different input cells. For example, a company may perform NPV analysis using a discount rate of 6%. Sensitivity analysis can be performed by analyzing scenarios of 5%, 8%, and 10% discount rates as well by simply maintaining the formula but referencing the different variable values.

7. What is hypothesis testing? What is H0 and H1? What is H0 and H1 for two-tail test?

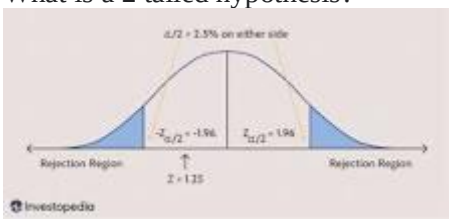
ANS

A statistical hypothesis is an assertion or conjecture concerning one or more populations. To prove that a hypothesis is true, or false, with absolute certainty, we would need absolute knowledge. That is, we would have to examine the entire population. Instead, hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis.

Hypothesis testing is formulated in terms of two hypotheses: • H0: the null hypothesis; • H1: the alternate hypothesis.

The hypothesis we want to test is if H1 is “likely” true. So, there are two possible outcomes: • Reject H0 and accept H1 because of sufficient evidence in the sample in favor or H1; • Do not reject H0 because of insufficient evidence to support H1.

What is a 2 tailed hypothesis?



A two-tailed hypothesis test is **designed to show whether the sample mean is significantly greater than and significantly less than the mean of a population**. The two-tailed test gets its name from testing the area under both tails (sides) of a normal distribution.

8. What is quantitative data and qualitative data?

ANS:

Quantitative data refers to any information that can be quantified. If it can be counted or measured, and given a numerical value, it’s quantitative data. Quantitative data can tell you “how many,” “how much,” or “how often”—for example, how many people attended last week’s webinar? How much revenue did the company make in 2019? How often does a certain customer group use online banking?

To analyze and make sense of quantitative data, you'll conduct statistical analyses.

Unlike quantitative data, qualitative data cannot be measured or counted. It's descriptive, expressed in terms of language rather than numerical values.

Researchers will often turn to qualitative data to answer "Why?" or "How?" questions. For example, if your quantitative data tells you that a certain website visitor abandoned their shopping cart three times in one week, you'd probably want to investigate why—and this might involve collecting some form of qualitative data from the user. Perhaps you want to know how a user feels about a particular product; again, qualitative data can provide such insights. In this case, you're not just looking at numbers; you're asking the user to tell you, using language, why they did something or how they feel.

Qualitative data also refers to the words or labels used to describe certain characteristics or traits—for example, describing the sky as blue or labeling a particular ice cream flavor as vanilla.

9. How to calculate range and interquartile range?

The range is the difference between the largest and smallest values in a data set.

The interquartile range (IQR) is the difference between the upper and lower quartiles.

For example: {2, 3, 5, 5, 7, 8, 9, 11, 12, 12, 50}

The range is 50 - 2 = 48

To find IQR, first find the median (the middle value) which is 8.

Then find the median of the numbers less then 8 and the numbers greater than 8. These give the lower quartile (5) and the upper quartile (12). The IQR is 12 - 5 = 7.

The IQR is resistant to outliers, such as the 50 in this data set. While that value makes the range large, it makes no difference to the IQR if that number is 50 or 12.

10. What do you understand by bell curve distribution ?

ANS: A bell curve is a common type of distribution for a variable, also known as the normal distribution. The term "bell curve" originates from the fact that the graph used to depict a normal distribution consists of a symmetrical bell-shaped curve.

The highest point on the curve, or the top of the bell, represents the most probable event in a series of data (its mean, mode, and median in this case), while all other possible occurrences are symmetrically distributed around the mean, creating a downward-sloping curve on each side of the peak. The width of the bell curve is described by its standard deviation.

A bell curve is a symmetric curve centered around the mean, or average, of all the data points being measured. The width of a bell curve is determined by the standard deviation—68% of the data points are within one standard deviation of the mean, 95% of the data are within two standard deviations, and 99.7% of the data points are within three standard deviations of the mean.

11. Mention one method to find outliers.

- ANS:

Outlier Detection and Analysis Methods

Outlier detection is a key consideration within the development and deployment of machine learning algorithms. Models are often developed and leveraged to perform outlier detection for different organisations that rely on large datasets to function. Economic modelling, financial forecasting, scientific research, and ecommerce campaigns are some of the varied areas that machine learning-driven outlier detection is used.

Identifying and dealing with outliers is an integral part of working with data, and machine learning is no different. Algorithm development usually relies on huge arrays of training data to achieve a high level of accuracy. Once deployed, models will process huge amounts of data, providing insights into trends and patterns. In this data-rich environment, organisations can expect to have to deal with outlier data. Outliers can skew trends and have a serious impact on the accuracy of models. The presence of outliers can be a sign of concept drift, so ongoing outlier analysis in machine learning is needed.

Machine learning models learn from data to understand the trends and relationship between data points. Outliers can skew results, and anomalies in training data can impact overall model effectiveness. Outlier detection is a key tool in safeguarding data quality, as anomalous data and errors can be removed and analysed once identified.

Outlier detection is an important part of each stage of the machine learning process. Accurate data is integral during the development and training of algorithms, and outlier detection is performed after deployment to maintain the effectiveness of models. This guide explores the basics of outlier detection techniques in machine learning, and how they can be applied to identify different types of outlier.

12. What is p-value in hypothesis testing?

ANS:

The p-value is **a number, calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true**.

P-values are used in hypothesis testing to help decide whether to reject the null hypothesis

13. What is the Binomial Probability Formula?

ANS: :

In the binomial probability, the number of successes X in ‘n’ trials of a binomial experiment is called a binomial random variable. The probability distribution of the random variable X is called a binomial distribution, and is given by the formula as below:
P(X)=Cnxpxqn−xP(X)=Cxnpxqn−x
Where n is the number of trials, x is 0, 1, 2..., n, p is the probability of success in a single trial, q is the probability of failure in a single trial and the value of q is 1-p. P(X) gives the probability of successes in n binomial trials.
The combination formula is Cnx=n!x!(n−x)!Cxn=n!x!(n−x)!.

14. Explain ANOVA and it’s applications.

ANS:

Developed by Ronald Fisher, ANOVA stands for Analysis of Variance. One-Way Analysis of Variance tells you if there are any statistical differences between the means of three or more independent groups.

You might use Analysis of Variance (ANOVA) as a marketer, when you want to test a particular hypothesis. You would use ANOVA to help you understand how your different groups respond, with a null hypothesis for the test that the means of the different groups are equal. If there is a statistically significant result, then it means that the two populations are unequal (or different).

The one-way ANOVA can help you know whether or not there are significant differences between the means of your independent variables (such as the first example: age, sex, income). When you understand how each independent variable’s mean is different from the others, you can begin to understand which of them has a connection to your dependent variable (landing page clicks), and begin to learn what is driving that behavior.