WORKSHEET-4

STATISTICS

Q1to Q15 are descriptive types. Answer in brief.

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

Answer:---**Central limit theorem:-** The CLT is a statistical theory that states that - if you take a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from that population will be roughly equal to the population mean.

Importance:- 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?

Answer:-- **Sampling:--**Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger population. Sampling means selecting the group that you will actually collect data from in your research.

Methods:-- Sampling in market action research is of two types – probability sampling and non-probability sampling.

Probability sampling:- Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.

Non-probability sampling:-- In non-probability sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.

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

Answer:--

Type 1 error	Type II error
1. It is also known as a false-positive.	2. It is also known as a false negative.
2. It occurs if the researcher rejects a correct null hypothesis in the population. i.e., incorrect rejection of the null hypothesis.	2. It occurs if a researcher fails to reject a null hypothesis that is actually a false hypothesis.
3. Measured by alpha (significance level) . If the significance level is fixed at 5%, It means there are about five chances of type – 1 error out of 100.	3. Measured by beta (the power of test). The probability of committing a type -2 error is calculated by 1 – beta (the power of test).
4. Sample size is not considered	4. It is caused by a smaller sample size
5. It can be reduced by decreasing the level of significance.	• 5. It can be reduced by increasing the level of significance.

4. What do you understand by the term Normal distribution?

Answer:--

-**Normal distribution:-** Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "bell curve".

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena.

5. What is correlation and covariance in statistics?

Answer:--

Correlation:-- we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other. Correlation is a statistical measure that indicates how strongly two variables are related.

Covariance:-- Covariance evaluates how the mean values of two random variables move together. For example, if stock A's return moves higher whenever stock B's return moves higher, and the same relationship is found when each stock's return decreases, these stocks are said to have positive covariance.

Relation:-- Both correlation and covariance can be positive or negative, depending on the values of the variables. A positive covariance always leads to a positive correlation, and a negative covariance always outputs a negative correlation. This is due to the fact that correlation coefficient is a function of covariance.

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

Answer:-

Univarate Analysis

Univariate analysis is the simplest form of data analysis where the data being analyzed contains only one variable. Since it's a single variable it doesn't deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.
You can think of the variable as a category that your data falls into. One example of a variable in univariate analysis might be "age". Another might be "height". Univariate analysis would not look at these two variables at the same time, nor would it look at the relationship between them..

Bivarate Analysis

Bivariate analysis is used to find out if there is a relationship between two different variables. Something as simple as creating a scatterplot by plotting one variable against another on a Cartesian plane (think X and Y axis) can sometimes give you a picture of what the data is trying to tell you. If the data seems to fit a line or curve then there is a relationship or correlation between the two variables.

Multivariate Analysis

Multivariate analysis is the analysis of three or more variables. There are many ways to perform multivariate analysis depending on your goals. Some of these methods include:

- Additive Tree
- Canonical Correlation Analysis
- Cluster Analysis

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

Answer:---**sensitivity:--** The sensitivity is calculated by dividing the percentage change in output by the percentage change in input.

$$Z = X^2 + Y^2$$

Sensitivity analysis is an analysis technique that works on the basis of what-if analysis like how independent factors can affect the dependent factor and is used to predict the outcome when analysis is performed under certain conditions. It is commonly used by investors who takes into consideration the conditions that affect their potential investment to test, predict and evaluate result.

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

Answer:--

Hypothesis testing:-- Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution.

H0 and H1 testing;-- In hypothesis testing there are two mutually exclusive hypotheses; the Null Hypothesis (H0) and the Alternative Hypothesis (H1).

H0- A null hypothesis is a type of statistical hypothesis that proposes that no statistical significance exists in a set of given observations. Hypothesis testing is used to assess the credibility of a hypothesis by using sample data. Sometimes referred to simply as the "null," it is represented as H_0.

H1-- The alternative hypothesis, H1 or Ha, is a statistical proposition stating that there is a significant difference between a hypothesized value of a population parameter and its estimated value. When the null hypothesis is tested, a decision is either correct or incorrect.

A two-tailed test is the statistical testing of whether a distribution is two-sided and if a sample is greater than or less than a range of values.

Null hypothesis (H0): The null hypothesis here is what currently stated to be true about the population. In our case it will be the average height of students in the batch is 100. Alternate hypothesis (H1): The alternate hypothesis is always what is being claimed

9. What is quantitative data and qualitative data?

Answer:--

Quantitative data:-- Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often). Quantitative data is data that can be counted or measured in numerical values. The two main types of quantitative data are discrete data and continuous data. Height in feet, age in years, and weight in pounds are examples of quantitative data.. There are four main types of Quantitative data: Descriptive, Correlational, Causal-Comparative/Quasi-Experimental, and Experimental Research. attempts to establish cause- effect relationships among the variables.

Qualitative data:- Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code. Qualitative data are data about categorical variables (e.g. what type). In statistics, qualitative data is known as categorical data. Qualitative data is data that can be felt or described. The three main types of qualitative data are binary, nominal, and ordinal. There are many different types of qualitative data, like data in research, work, and statistics.

10. How to calculate range and interquartile range?

Answer:**--**

Range: the difference between the highest and lowest values.

Interquartile range: the range of the middle half of a distribution.

The range is calculated by subtracting the lowest value from the highest value.. The IQR describes the middle 50% of values when ordered from lowest to highest. To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference between Q3 and Q1.

Calculation:--

The formula for quartiles is given by:

1. Lower Quartile (Q1) = (N+1) * 1 / 4.
2. Middle Quartile (Q2) = (N+1) * 2 / 4.
3. Upper Quartile (Q3)= (N+1) * 3 / 4.
4. Interquartile Range = Q3 – Q1.

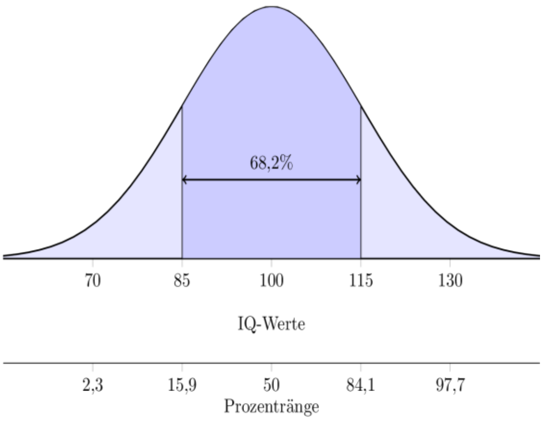
Where,
IQR=Inter-quartile range
Q_1 = First quartile
Q_3 = Third quartile

5. Q_1 can also be found by using the following formula
6. Q1=(n+14)thterm
7. Q_3 can also be found by using the following formula**:**
8. Q3=(3(n+1)4)thterm
9. In these cases, if the values are not whole number, we have to round them up to the nearest integer.
10. Q_2 can also be found by using the following formula:
11. $Q_2 = Q_3 – Q$

11. What do you understand by bell curve distribution ?

Answer:--

A bell curve is a type of graph that is used to visualize the distribution of a set of chosen values across a specified group that tend to have a central, normal values, as peak with low and high extremes tapering off relatively symmetrically on either side.



This Photo by Unknown Author is licensed under

12. Mention one method to find outliers.

Answer:-- **Outliers:--** Outliers can give helpful insights into the data you're studying, and they can have an effect on statistical results. This can potentially help you disover inconsistencies and detect any errors in your statistical processes.

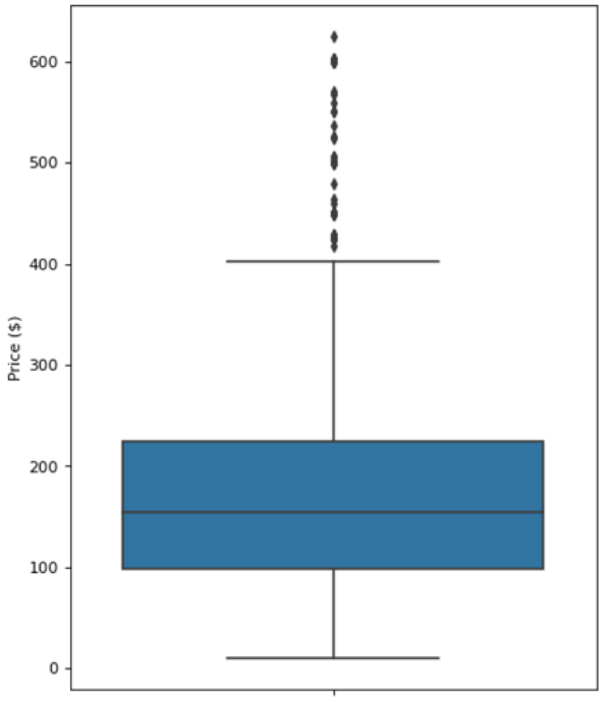
Method:--

Z-score is just another form of standard deviation procedure. Z-score is used to convert the data into another dataset with mean = 0.

$$Z = \frac{x_i - \bar{x}}{s}$$

Here, X-bar is the mean value and s is standard deviation. Once the data is converted, the center becomes 0 and the z-score corresponding to each data point represents the distance from the center in terms of standard deviation. For example, a z-score of 2.5 indicates that the data point is 2.5 standard deviation away from the mean. Usually z-score =3 is considered as a cut-off value to set the limit. Therefore, any z-score greater than +3 or less than -3 is considered as outlier which is pretty much similar to standard deviation method.

We found that the number of outliers is 21 before implementing this method and obtained 20 after removing those 21 outliers.



In fact, these 20 outliers are the same data point that we obtained from 3 times stdev method. Therefore, the user may proceed with either one.

13. What is p-value in hypothesis testing?

Answer:---

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.

the significance level is declared in advance to determine how small the P-value needs to be such that the null hypothesis is rejected. The levels of significance vary from one researcher to another; so it can get difficult for readers to compare results from two different tests. That is when P-value makes things easier.

14. What is the Binomial Probability Formula?

Answer:---

Binomial probability distribution:-- The binomial distribution is the discrete probability distribution that gives only two possible results in an experiment, either success or failure.

Formula:-- The formula for binomial distribution is:
P(x: n,p) = nCx px (q)n-x
Where p is the probability of success, q is the probability of failure,
n= number of trials

Example:- A financial analyst wants to evaluate the binomial distribution for the success rate of achieving more than a 15% return based on the historical data of past investment activities. If the analyst establishes each quarterly period as a single trial and evaluates 18 periods, they apply this value to the n variable in the probability formula. Assuming the financial analyst calculates seven successful return rates over 15%, they can calculate the binomial distribution for a 50% chance of earning over 15% in returns. Using the formula, these values result in the following binomial distribution:

Px = 18C7 · (0.5)7 · (1 - 0.5)18-7 =

Px = (31,824) · (0.0078125) · (0.000488281) = 0.12139892578

This gives the financial analyst an approximate success rate of 12% for achieving investment returns that are greater than 15%. Using this information, the analyst can then evaluate various investment instruments and select the most profitable options that can result in similar outcomes.

15. Explain ANOVA and it's applications.

Answer:--

ANOVA:-- Analysis of Variance (ANOVA) is **a** statistical formula used to compare variances across the means (or average) of different groups. A range of scenarios use it to determine if there is any difference between the means of different groups.

The formula for Analysis of Variance is: ANOVA coefficient, F= Mean sum of squares between the groups (MSB)/ Mean squares of errors (MSE). Therefore **F = MSB/MSE**.

Application:- ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.