ASSIGNMENT 3
STATISTICS WORKSHEET-3

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is the correct formula for total variation?
a) Total Variation = Residual Variation – Regression Variation
b) Total Variation = Residual Variation + Regression Variation
c) Total Variation = Residual Variation * Regression Variation
d) All of the mentioned
ANS: A

2. Collection of exchangeable binary outcomes for the same covariate data are called outcomes.
a) random
b) direct
c) binomial
d) none of the mentioned
ANS: A

3. How many outcomes are possible with Bernoulli trial?
a) 2
b) 3
c) 4
d) None of the mentioned
ANS: A

4. If Ho is true and we reject it is called
a) Type-I error
b) Type-II error
c) Standard error
d) Sampling error
ANS: A

5. Level of significance is also called:
a) Power of the test
b) Size of the test
c) Level of confidence
d) Confidence coefficient
ANS: B

6. The chance of rejecting a true hypothesis decreases when sample size is:
a) Decrease
b) Increase
c) Both of them

d) None
ANS: B

7. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned
ANS: B

8. What is the purpose of multiple testing in statistical inference?
a) Minimize errors
b) Minimize false positives
c) Minimize false negatives
d) All of the mentioned
ANS: D

9. Normalized data are centred at and have units equal to standard deviations of the original data
a) 0
b) 5
c) 1
d) 10
ANS: A

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What Is Bayes' Theorem?

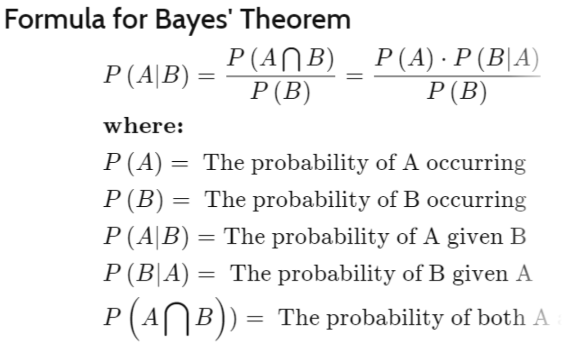
ANS: Bayes' Theorem states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event given the first event multiplied by the probability of the first event.

It is a mathematical formula used to calculate conditional probability. The likelihood of an outcome occurring based on a previous outcome occurring in similar circumstances is known as conditional probability. Given new or additional evidence, Bayes' theorem allows you to revise existing predictions or theories (update probabilities).

In finance, Bayes' Theorem can be used to rate the risk of lending money to potential borrowers. The theorem is also called Bayes' Rule or Bayes' Law and is the foundation of the field of Bayesian statistics.

Bayes' Theorem calculates the conditional probability of an event, based on the values of specific related known probabilities.

A Bayes’ Theorem Calculator figures the probability of an event A conditional on another event B, given the prior probabilities of A and B, and the probability of B conditional on A. It calculates conditional probabilities based on known probabilities.



Formula for Bayes' Theorem

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)}$$

where:

$P(A) =$ The probability of A occurring

$P(B) =$ The probability of B occurring

$P(A|B) =$ The probability of A given B

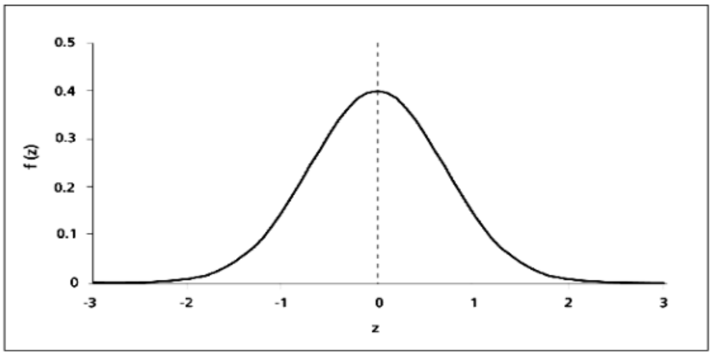
$P(B|A) =$ The probability of B given A

$P\left(A \bigcap B\right)) =$ The probability of both A

11. What is z-score?

ANS: A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units. The z-score is positive if the value lies above the mean, and negative if it lies below the mean.

It is also known as a standard score, because it allows comparison of scores on different kinds of variables by standardizing the distribution. A standard normal distribution (SND) is a normally shaped distribution with a mean of 0 and a standard deviation (SD) of 1.



It is useful to standardize the values (raw scores) of a normal distribution by converting them into z-scores because:

(a) It allows researchers to calculate the probability of a score occurring within a standard normal distribution;

(b) And enables us to compare two scores that are from different samples (which may have different means and standard deviations).

The formula for calculating a z-score is is z = (x-μ)/σ, where x is the raw score, μ is the population mean, and σ is the population standard deviation.

As the formula shows, the z-score is simply the raw score minus the population mean, divided by the population standard deviation.

12. What is t-test?

ANS: A T-test is the final statistical measure for determining differences between two means that may or may not be related. The testing uses randomly selected samples from the two categories or groups. It is a statistical method in which samples are chosen randomly, and there is no perfect normal distribution.

The type of T-test to be conducted is decided by whether the samples to be analysed are from the same category or distinct categories. The inference obtained in the process indicates the probability of the mean differences to have happened by chance. The test is useful when comparing population age, length of crops from two different species, student grades, etc.

A T-test studies a set of data gathered from two similar or different groups to determine the probability of the difference in the result than what is usually obtained. The accuracy of the test depends on various factors, including the distribution patterns used and the variants influencing the collected samples. Depending on the parameters, the test is conducted, and a T-value is obtained as the statistical inference of the probability of the usual resultant being driven by chance.

For example, if one wishes to figure out if the mean of the length of petals of a flower belonging to two different species is the same, a T-test can be done. The user can select petals randomly from two other species of that flower and come to a standard conclusion. The final T-test interpretation could be obtained in either of the two ways:

- A null hypothesis signifies that the difference between the means is zero and where both the means are shown as equal.
- An alternate hypothesis implies the difference between the means is different from zero. This hypothesis rejects the null hypothesis, indicating that the data set is quite accurate and not by chance.

This T-test, however, is only valid and should be done when the mean or average of only two categories or groups needs to be compared. As soon as the number of comparisons to be made is more than two, conducting this is not recommended.

13. What is percentile?

ANS: Percentiles are used in statistics to give you a number that describes the value that a given percent of the values are lower than.

Example: Let's say we have an array of the ages of all the people that lives in a street.

ages = [5,31,43,48,50,41,7,11,15,39,80,82,32,2,8,6,25,36,27,61,31]

What is 75. Percentile? The answer is 48, meaning that 75% of the people are 48 or younger.

Example to use the NumPy percentile () method to find the percentiles:

```
import numpy
ages = [25,31,43,48,50,41,39,60,52,32,27,46,47,55]
x = numpy.percentile(ages, 75)
print(x)
```

14. What is ANOVA?

ANS: Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.

- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.

15. How can ANOVA help?

ANS: The ANOVA test is the initial step in analyzing factors that affect a given data set. Once the test is finished, an analyst performs additional testing on the methodical factors that measurably contribute to the data set's inconsistency. The analyst utilizes the ANOVA test results in an f-test to generate additional data that aligns with the proposed regression models.

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them. The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

If no real difference exists between the tested groups, which is called the null hypothesis, the result of the ANOVA's F-ratio statistic will be close to 1. The distribution of all possible values of the F statistic is the F-distribution. This is actually a group of distribution functions, with two characteristic numbers, called the numerator degrees of freedom and the denominator degrees of freedom.