WORKSHEET

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: b) Total Variation = Residual Variation + Regression Variation

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: c) binomial

3. How many outcomes are possible with Bernoulli trial?

a) 2

b) 3

c) 4

d) None of the mentioned

Ans: a) 2

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) Type-I error

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: c) Level of confidence

6. The chance of rejecting a true hypothesis decreases when sample size is:
a) Decrease
b) Increase
c) Both of them
d) None
Ans: a) Decrease

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) Hypothesis

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) All of the mentioned

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) 0

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

10. What Is Bayes' Theorem?

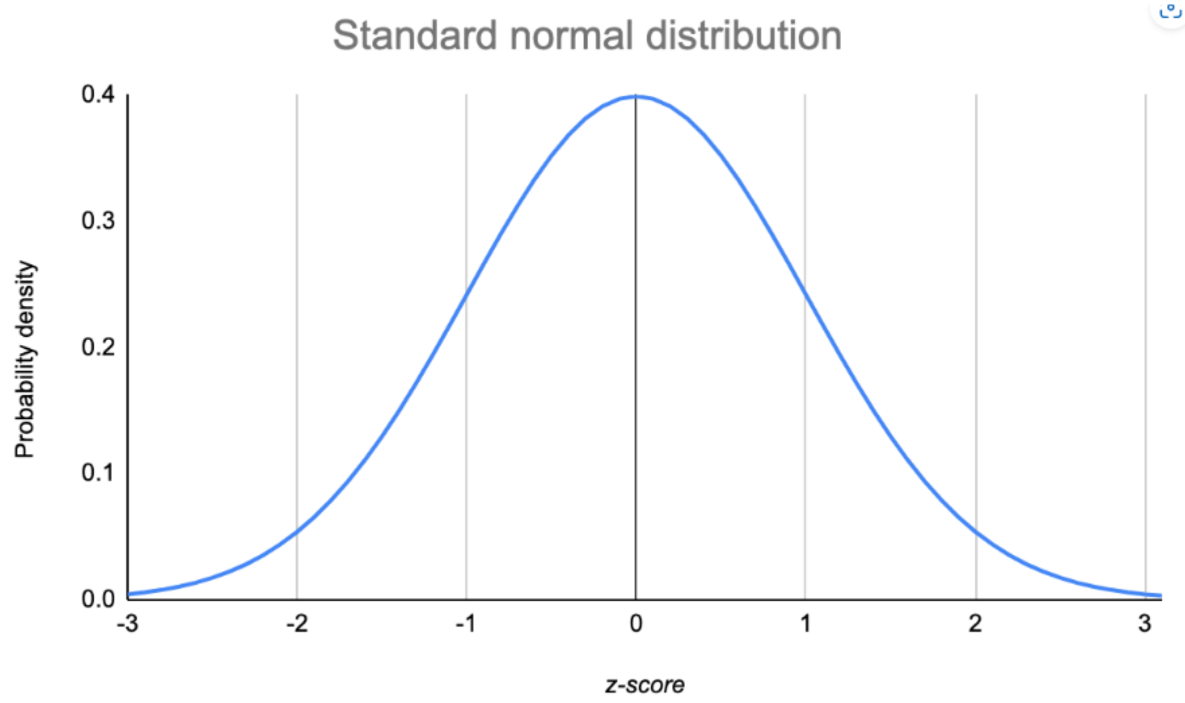
- It describes the probability of events that based on the condition that might related to the event. Bayes Theorem also known as Bayes rule or Bayes Law.
- In Probability, Bayes theorem is a mathematical formula, which is used to determine the conditional probability of the given event. Conditional probability is defined as the likelihood that an event will occur, based on the occurrence of a previous outcome.
- Bayes' Theorem allows you to update the predicted probabilities of an event by incorporating new information.
- Bayes' theorem relies on incorporating prior probability distributions in order to generate posterior probabilities.

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

- Bayes Theorem provides a useful method for thinking about the relationship between a data set and a probability. In other words, the theorem says that the probability of a given hypothesis being true based on specific observed data can be stated as finding the probability of observing the data given the hypothesis multiplied by the probability of the hypothesis being true regardless of the data, divided by the probability of observing the data regardless of the hypothesis.
- Besides statistics, the Bayes' theorem is also used in various disciplines, with medicine and pharmacology. The theorem is commonly employed in different fields of finance. Some of the applications include but are not limited to, modelling the risk of lending money to borrowers or forecasting the probability of the success of an investment.

11. What is z-score?

- A z score is score that designates how many standard deviations a particular score is above or below the mean.
- z scores can be positive or negative depending on whether the score in question is above or below the mean. Thus, the higher the magnitude of the score (meaning independent of whether it's positive or negative) the further the score is away from the mean.
- Z-test is a hypothesis test in which the z-statistic follows a normal distribution**.**



- A Z-score is a **numerical measurement that describes a value's relationship to the mean of a group of values**. Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score.
- 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

$$\textbf{Z Score} = (x - \bar{x})/\sigma$$

- x = Standardized random variable
- $\bar{x}$ = Mean
- σ = Standard deviation.

- Z-Score is measured in terms of standard deviation from the mean. For example, a standard deviation of 2 indicates the value is 2 standard deviations away from the mean. In order to use a z-score, we need to know the population mean (μ) and also the population standard deviation (σ).

12. What is t-test?

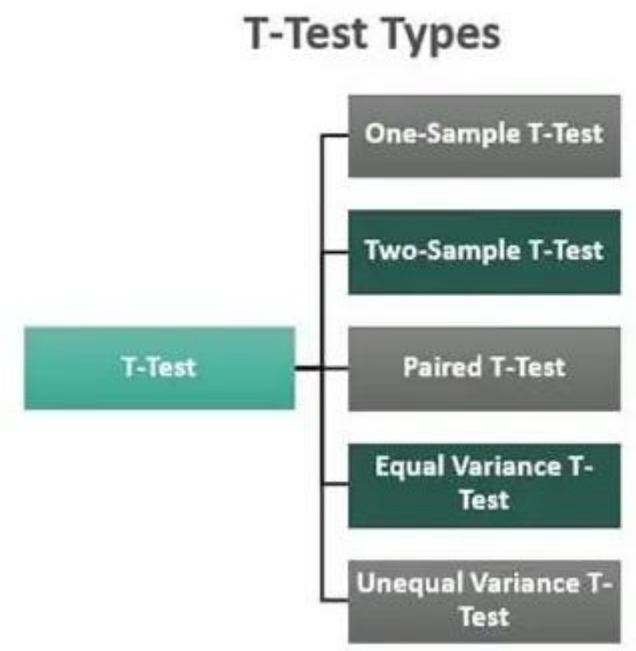
- A t-test is an inferential statistic used to determine if there is a significant difference between the means of two groups and how they are related.

- T-tests are used when the data sets follow a normal distribution and have unknown variances, like the data set recorded from flipping a coin 100 times.
- The t-test is a test used for hypothesis testing in statistics.
- Calculating a t-test requires three fundamental data values including the difference between the mean values from each data set, the standard deviation of each group, and the number of data values.
- T-tests can be dependent or independent.
- A t-test compares the average values of two data sets and determines if they came from the same population.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\Delta}}$$

where

$$s_{\Delta} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$



- The tests are completely based on random sampling. As no individuality is maintained in the samples, the reliability is often questioned.
- 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.
- The measurement scale used for such hypothesis testing follows a set of continuous or ordinal patterns. The accounted parameters and variants influencing the samples and surrounding the groups are based on the standard consideration.
- When the data is plotted with respect to the T-test distribution, it should follow a normal distribution and bring about a bell-curved graph.
- For a clearer **bell curve**, the **sample size** needs to be bigger.
- The variance should be such that the **standard deviations** of the samples are almost equal.

13. What is percentile?

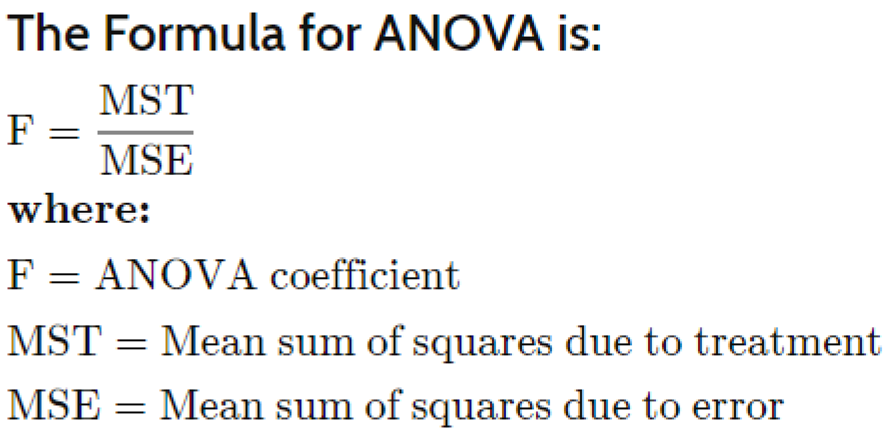
- Percentile is used to understand and interpret data. They indicate the values below which a certain percentage of the data in a data set is found.

$$\mathbf{n = (P/100) \times N}$$

- It is used to express fractions of a whole, while percentiles are the values below which a certain percentage of the data in a data set is found. In practical terms, there is a significant difference between the two. For example, a student taking a difficult exam might earn a score of 75 percent.
- This means that he correctly answered every three out of four questions. A student who scores in the 75th percentile, however, has obtained a different result. This percentile means that the student earned a higher score than 75 percent of the other students who took the exam.
- In other words, the percentage score reflects how well the student did on the exam itself; the percentile score reflects how well he did in comparison to other students.
- Percentile is nothing but, is a measure that is used in the statistics indicating the values below which the given percentage of an observation in a given data falls.
- The percentile rank of a score is the percentage of scores in its distribution that are less than it, an exclusive definition, and one that can be expressed with a single, simple formula. Percentile scores and percentile ranks are often used in the reporting of test scores from norm-referenced tests.

14. What is ANOVA?

- 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.
- 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.



The Formula for ANOVA is:

$$F = \frac{\text{MST}}{\text{MSE}}$$

where:

$\text{F} = \text{ANOVA coefficient}$

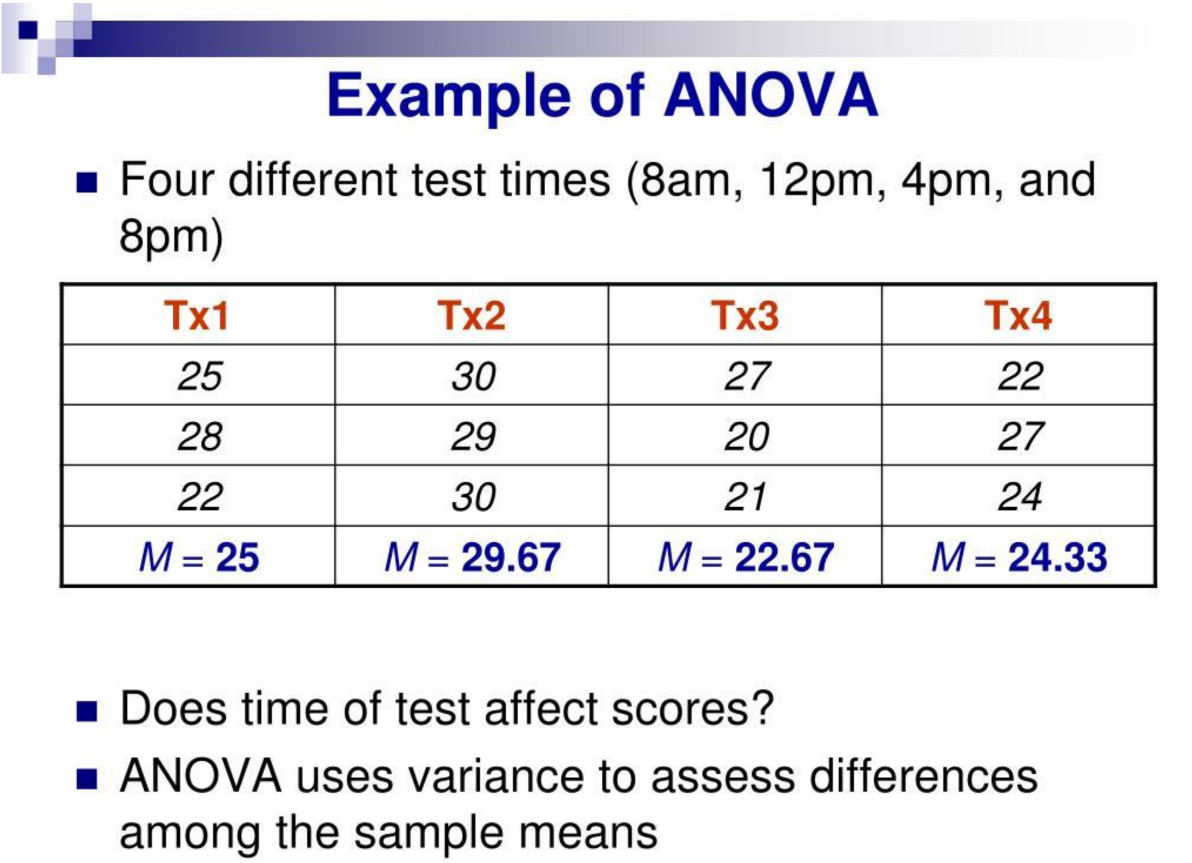
$\text{MST} = \text{Mean sum of squares due to treatment}$

$\text{MSE} = \text{Mean sum of squares due to error}$

- 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.
- There are two main types of ANOVA: **one-way** (or unidirectional) and **two-way**.
- ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests.
- ANOVA is used in regression to determine the influence of independent variable on dependent variables.
- The analyst utilizes the ANOVA test results in an f-test to generate additional data that aligns with the proposed regression models.

15. How can ANOVA help?

- The **ANOVA** is a technique that is actually used all the time in a variety of fields in real life.
- The two most common types of ANOVAs are the one-way ANOVA and two-way ANOVA.
- ANOVA is used in a wide variety of real-life situations, but the most common include:
- **Retail:** Store are often interested in understanding whether different types of promotions, store layouts, advertisement tactics, etc. lead to different sales. This is the exact type of analysis that ANOVA is built for.
- **Medical:** Researchers are often interested in whether or not different medications affect patients differently, which is why they often use one-way or two-way ANOVA's in these situations.
- **Environmental Sciences:** Researchers are often interested in understanding how different levels of factors affect plants and wildlife. Because of the nature of these types of analyses, ANOVA's are often used.



Example of ANOVA

- Four different test times (8am, 12pm, 4pm, and 8pm)

Tx1	Tx2	Tx3	Tx4
25	*30*	*27*	*22*
28	*29*	*20*	*27*
22	*30*	*21*	*24*
M = **25**	M = **29.67**	M = **22.67**	M = **24.33**

- Does time of test affect scores?
- ANOVA uses variance to assess differences among the sample means

- when we have more than two groups, t-test is not the optimal choice because a separate t-test needs to perform to compare each pair.
- Assume we are comparing three countries, A, B, and C. We need to apply a t-test to A-B, A-C and B-C pairs. As the number of groups increase, this becomes harder to manage. Thus, we choose to go with ANOVA.
- ANOVA is a method to determine if the mean of groups is different. In inferential statistics, we use samples to infer properties of populations. Statistical tests like ANOVA help us justify if sample results are applicable to populations.
- The difference between t-test and ANOVA is that t-test can only be used to compare two groups where ANOVA can be extended to three or more groups.
- ANOVA can also be used in feature selection process of machine learning. The features can be compared by performing an ANOVA test and similar ones can be eliminated from the feature set.