ASSIGNMENT – 3 STATISTICS

1. B) Total Variation = Residual Variation + Regression Variation.
2. C) Binomial outcomes.
3. A) 2
4. A) Type I error
5. B) Size of the test
6. B) Increase
7. B) Hypothesis
8. D) All of the above.
9. A) 0
10. In statistics, Bayes' theorem is used to update the probability of an event occurring (e.g. a hypothesis being true) based on new evidence or data. It is often used in Bayesian statistics, which is a branch of statistics that uses Bayes' theorem to update probabilities based on new data. The formula for Bayes' theorem is: $P(A|B) = P(B|A) * P(A) / P(B)$

    Where: $P(A|B)$ is the probability of event A occurring given that event B has occurred (i.e. the posterior probability) $P(B|A)$ is the probability of event B occurring given that event A has occurred (i.e. the likelihood) $P(A)$ is the prior probability of event A occurring (i.e. the probability of A before taking B into account) $P(B)$ is the probability of event B occurring (i.e. the marginal probability of B)

    Bayes' theorem is used to update the probability of an event A occurring (the posterior probability) based on new information or data about event B. The prior probability of A is multiplied by the likelihood of B given A, and then divided by the marginal probability of B. This allows for the incorporation of new information into the probability estimation process. An example of the use of Bayes' theorem in statistics is in medical testing. A test for a certain disease may have a high rate of false positives, meaning that it may indicate the presence of the disease even when the person doesn't have it. By using Bayes' theorem, the probability of actually having the disease can be calculated based on the test results and the prior probability of having the disease.

11. A z-score, also known as a standard score, is a measure of how many standard deviations an observation or data point is from the mean of a distribution. It is calculated by subtracting the mean of the distribution from an individual data point, and then dividing the result by the standard deviation of the distribution. The formula for calculating a z-score is: $z = (x - \mu) / \sigma$ Where: x is the individual data point μ is the mean of the distribution σ is the standard deviation of the distribution

    The resulting z-score will be a number that tells you how many standard deviations away from the mean the data point is. A positive z-score indicates that the data point is above the mean, and a negative z-score indicates that the data point is below the mean. A z-score can be used to standardize the data, which can be useful in cases where you want to compare data from different distributions that have different units or ranges. It also can be used to calculate the probability of a data point being within a certain range of the mean, which is useful in statistical hypothesis testing and in understanding the spread of data.

In summary, z-score is a useful statistic because it standardizes data and provides a way to compare observations from different distributions. It also helps in identifying outliers, understanding the spread of data and measuring the probability of an observation being within a certain range of the mean.

12. A t-test is a statistical hypothesis test that is used to determine whether there is a significant difference between the means of two groups. There are different types of t-tests, but the most common is the independent samples t-test, which is used to compare the means of two groups that are independent of each other. The independent samples t-test is used to test the null hypothesis that the means of two groups are equal, against the alternative hypothesis that they are not equal. The test statistic is calculated as the difference between the means of the two groups, divided by the standard error of the difference. The test statistic follows a t-distribution, which is a probability distribution that is similar to the normal distribution but has heavier tails. The t-test is sensitive to the sample size and the variance of the groups being compared. When the sample size is small or the variances are not equal, the t-test should be used instead of the z-test.

    The t-test can be one-tailed or two-tailed, depending on the research question and the direction of the alternative hypothesis. A one-tailed test is used when the direction of the difference is predicted, while a two-tailed test is used when the direction of the difference is not predicted. In summary, t-test is a statistical hypothesis test used to compare the means of two independent groups. It is used to test the null hypothesis that the means are equal, against the alternative hypothesis that they are not equal. The test statistic follows a t-distribution, which is similar to the normal distribution but has heavier tails. T-test is sensitive to the sample size and the variances of the groups being compared, and can be one-tailed or two-tailed.

13. A percentile is a measure that indicates the value below which a certain percentage of observations in a data set fall. For example, the 50th percentile, also known as the median, is the value that separates the lowest 50% of observations from the highest 50% of observations. Percentiles are often used to describe the distribution of a data set and to compare different sets of data.
    The formula for calculating a percentile is: P(x) = (n x / 100) Where: P(x) is the percentile n is the total number of observations in the data set x is the percentage of observations that fall below the desired percentile For example, if you have a data set of 100 observations, and you want to find the 90th percentile, you would first sort the data set in ascending order. Then you would find the value that corresponds to 90% of the observations, which would be the 90th value in the sorted data set. It's important to note that percentiles are not the same as percentages, and should not be confused. Percentiles are used to describe the distribution of data, while percentages are used to describe the proportion of observations that fall into a certain category.
    In summary, percentile is a measure that indicates the value below which a certain percentage of observations in a data set fall. Percentiles are commonly used to describe

the distribution of a data set and to compare different sets of data. Percentile values can be calculated using the formula P(x) = (n x / 100) where n is the total number of observations in the data set and x is the percentage of observations that fall below the desired percentile.

14. ANOVA stands for "Analysis of Variance". It is a statistical method used to test for significant differences between the means of two or more groups. It is based on the idea that if there is no difference between the means of the groups, then the variation within the groups should be similar to the variation between the groups. There are different types of ANOVA, including:

- One-way ANOVA: This is used to compare the means of two or more independent groups.
- Two-way ANOVA: This is used to compare the means of two or more groups while taking into account the effect of two independent variables.
- Repeated measures ANOVA: This is used to compare the means of two or more groups while taking into account the effect of repeated measurements on the same individuals. The ANOVA test is based on the F-distribution, which is a probability distribution that is similar to the chi-squared distribution but with a different number of degrees of freedom. The test statistic, known as the F-value, is calculated as the ratio of the variation between the groups to the variation within the groups.
- The null hypothesis for ANOVA is that there is no significant difference between the means of the groups, while the alternative hypothesis is that there is a significant difference between at least two of the means. The p-value is used to determine the level of significance of the test.

15. ANOVA can help in many ways : Identifying significant differences: ANOVA can be used to determine whether there is a significant difference between the means of two or more groups. This can be useful in fields such as marketing, where ANOVA can be used to test whether different advertising campaigns have different effects on sales, or in biology, where ANOVA can be used to test whether different treatments have different effects on plant growth.

- Understanding relationships between variables: ANOVA can be used to understand the relationship between two or more independent variables and a dependent variable. For example, in an experiment to study the effect of a new drug on blood pressure, ANOVA can be used to determine the effect of dose and age on blood pressure.
- Multiple comparisons: ANOVA can be used to make multiple comparisons between groups, without increasing the risk of type I errors (false positives).
- Identifying outliers: ANOVA can be used to identify outliers in the data, which are data points that are significantly different from the other data points in the group.
- Exploratory data analysis: ANOVA can be used as a tool for exploratory data analysis, to identify patterns and relationships in the data that might not be immediately apparent.
- Design of experiments: ANOVA can be used to design experiments by identifying the key factors that are likely to affect the outcome of the experiment, and the appropriate levels of these factors.