Rahul Krishna Gunneri **2210264**

Section A

1. In your own words, define skewness? What are some real life examples of skewed distributions?

Ans. Skewness is a measure of the asymmetry of a distribution, indicating the extent to which the data is skewed towards one tail or the other. A perfectly symmetrical distribution will have a skewness of zero, meaning that the data is equally distributed around the mean. A positive skewness indicates that the distribution is skewed towards the right tail, with a longer tail on the right side of the distribution. Conversely, a negative skewness indicates that the distribution is skewed towards the left tail, with a longer tail on the left side of the distribution.

Real life examples of skewed distributions can be found in many different areas. For example, income distribution is often positively skewed, with a long tail towards the higher incomes. This means that most people have relatively low to moderate incomes, but there are a few people who have very high incomes. Another example is exam scores, which are often negatively skewed, with a long tail towards the lower scores. This means that most students perform relatively well on the exam, but there are a few students who perform very poorly. Additionally, stock prices can also exhibit skewness, with a long tail towards extreme values due to infrequent events such as market crashes or surges.

2. Define type 1 and type 2 errors and give a relevant example from the field of petroleum engineering. How do you reduce these errors?

Ans. Type 1 error and Type 2 error are two types of errors that can occur in hypothesis testing.

Type 1 error occurs when the null hypothesis is rejected even though it is true. In other words, it is a false positive result. The probability of making a type 1 error is denoted by the Greek letter alpha (α).

Type 2 error, on the other hand, occurs when the null hypothesis is not rejected even though it is false. In other words, it is a false negative result. The probability of making a type 2 error is denoted by the Greek letter beta (β).

In the field of petroleum engineering, an example of a Type 1 error would be to conclude that a particular reservoir contains significant amounts of oil, even though it does not. This can lead to expensive drilling and exploration activities that do not result in any significant oil discoveries.

An example of a Type 2 error in the field of petroleum engineering would be to conclude that a particular reservoir does not contain significant amounts of oil, even though it does. This can lead to missed opportunities for oil discovery and development.

To reduce these errors, it is essential to design experiments and data analysis methods that are statistically sound and have appropriate sample sizes. Proper planning and thorough analysis can help minimize the likelihood of Type 1 and Type 2 errors. Additionally, it is essential to establish appropriate decision thresholds and evaluate the costs associated with making each type of error to ensure that the trade-offs are acceptable.

3. What are the common statistical assumptions?

Ans. There are several common statistical assumptions that underlie many statistical techniques. Violations of these assumptions can affect the validity of statistical inferences and conclusions. The most common statistical assumptions include:

1) Normality assumption: The assumption that the distribution of the data follows a normal or Gaussian distribution. Many statistical techniques, such as t-tests and ANOVA, assume that the data are normally distributed.
2) Independence assumption: The assumption that the observations in a sample are independent of each other. In other words, the value of one observation does not influence the value of another observation. This assumption is necessary for many statistical techniques to be valid.
3) Homogeneity of variance assumption: The assumption that the variance of the data is the same across different groups or conditions. Violations of this assumption can lead to biased estimates of the standard error and incorrect statistical inferences.
4) Linearity assumption: The assumption that the relationship between two variables is linear. Many regression techniques assume a linear relationship between the independent and dependent variables.
5) Random sampling assumption: The assumption that the sample is selected randomly from the population of interest. This assumption is necessary for making inferences about the population based on the sample.
6) Equal group size assumption: The assumption that the sample sizes are equal across different groups or conditions. Violations of this assumption can lead to biased estimates of the standard error and incorrect statistical inferences.
7) Absence of outliers assumption: The assumption that the data do not contain extreme or influential observations that can bias the results of the analysis.

It is important to assess these assumptions before using statistical techniques to ensure that the results are valid and reliable. If these assumptions are violated, appropriate adjustments or alternative methods may be needed to analyze the data properly.

4. List the best practices for cleaning data. How do you remove duplicate observations from a data frame?

Ans. Best practices for cleaning data include:

1) Check for missing values and decide on the best way to handle them (impute or remove).
2) Identify and remove duplicate observations.
3) Correct or remove any inconsistent or erroneous data.
4) Check for outliers and decide on the best way to handle them (remove or transform).
5) Standardize data types and formats.
6) Ensure data consistency across different data sources.

To remove duplicate observations from a data frame in Python, drop_duplicates() function will be from the pandas library. This function removes all duplicate rows from the data frame.

5. What are outliers? Why do they matter? When should you remove a outlier from a data set?

Ans. Outliers are observations in a dataset that are significantly different from other observations. They can be caused by measurement errors, data entry errors, or legitimate extreme values in the data. Outliers can be identified using statistical methods such as the interquartile range (IQR) or standard deviation (SD).

Outliers matter because they can affect the accuracy and reliability of statistical analyses. They can skew the results of descriptive statistics, such as mean and standard deviation, and also affect the results of inferential statistics, such as hypothesis testing and regression analysis. Outliers can also affect the normality assumption of statistical models, leading to biased parameter estimates and incorrect inferences.

Whether to remove an outlier from a data set depends on the nature of the data and the research question. In some cases, outliers may be legitimate and informative data points that should not be removed. For example, in medical research, extreme values may be the result of a rare disease or an unexpected response to treatment. In other cases, outliers may be the result of measurement errors or data entry errors and should be removed to prevent bias in statistical analysis.

In general, outliers should be removed when they are the result of measurement or data entry errors, or when they significantly affect the results of statistical analysis. However, removing too many outliers can also lead to bias in the analysis, so it is important to use careful judgment when deciding which outliers to remove. It is also important to report the presence of outliers in the data and any decisions made about their treatment.

6. What are the types of missing data? How do you deal with missing data?

Ans. There are four types of missing value, they are:

1. Missing completely at random (MCAR) is a type of missing data mechanism in which the probability of a data point being missing is unrelated to both its observed and unobserved values, and is also unrelated to any other variables in the dataset.

2. Missing at random (MAR) is a type of missing data mechanism where the probability of a data point is missing depends only on observed data and can be predicted by other variables in the dataset. This means that the missingness in the dataset is related to other variables in the dataset, but not to the missing values themselves.

3. Structurally missing data is a type of missing data mechanism where the missingness in a dataset is related to the underlying structure or design of the study or data collection process. This means that some variables or observations are missing from the dataset because they were not included in the study or data collection process in the first place.

4. "Missing not at random" (MNAR) is a term used in statistics and data analysis to describe a situation where the probability of a data point being missing is dependent on the value of the missing data itself, in a way that is not explained by the observed data.

To deal with missing data, there are several strategies that can be used, including:

Complete Case Analysis: This involves only analyzing the data for observations that have complete information for all variables. However, this can result in loss of information and reduced statistical power.

Imputation: This involves filling in the missing data with estimated values based on observed data. There are several methods for imputation, including mean imputation, regression imputation, and multiple imputation.

Model-Based Methods: These methods involve fitting a statistical model that can handle missing data, such as maximum likelihood estimation or Bayesian methods.

Weighting Methods: These methods involve assigning weights to observations based on their likelihood of being missing, which can help account for the potential biases in the analysis due to missing data.

7. Define Arithmetic mean, Harmonic mean and Geometric mean. How are they different in terms of their power?

Ans. Arithmetic mean, harmonic mean, and geometric mean are all measures of central tendency that are used to describe a set of data. Here's how they are defined:

Arithmetic mean: The arithmetic mean is the sum of all the values in a dataset divided by the total number of values. It is also known as the average.

Arithmetic Mean = (Sum of all values) / (Number of values)

Harmonic mean: The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals of the values in a dataset. It is useful for averaging rates or ratios.

Harmonic Mean = (Number of values) / (Sum of the reciprocals of the values)

Geometric mean: The geometric mean is the nth root of the product of all the values in a dataset, where n is the total number of values. It is useful for calculating growth rates and compound interest.

Geometric Mean = (Product of all values)^(1/n).

The arithmetic mean, harmonic mean, and geometric mean all have different powers and are used in different contexts.

The arithmetic mean is the most commonly used measure of central tendency and is useful for summarizing data that is normally distributed or symmetrical. However, it is sensitive to outliers and can be affected by extreme values.

The harmonic mean is useful for averaging rates or ratios, such as speed or price-to-earnings ratio, and is less affected by outliers than the arithmetic mean. However, it can be undefined or distorted when there are zero or negative values in the dataset.

The geometric mean is useful for calculating growth rates and compound interest and is less affected by extreme values than the arithmetic mean. However, it can only be used with positive values and can be distorted by very small or very large values.

In summary, the choice of mean depends on the nature of the data and the purpose of the analysis. The arithmetic mean is the most commonly used measure of central tendency, but the harmonic mean and geometric mean are useful in specific contexts where rates or ratios are involved.

8. What are the tests required for Normality, Linearity and Homoskedasticity?

Ans. Normality, Linearity, and Homoscedasticity are assumptions that must be met in many statistical analyses. There are several tests available to check whether these assumptions are met, some of which are:

Normality:

The normality assumption states that the data should follow a normal distribution. The following are some tests to check for normality:

Shapiro-Wilk test: It tests the null hypothesis that the data are normally distributed.

Kolmogorov-Smirnov test: It tests the null hypothesis that the data come from a normal distribution.

Anderson-Darling test: It tests the null hypothesis that the data come from a specific distribution, such as the normal distribution.

Linearity:

The linearity assumption states that the relationship between two variables should be linear. The following are some tests to check for linearity:

Scatter plot: A scatter plot can be used to visually check for linearity.

Residual plot: A residual plot can be used to check for the linearity of the relationship between the predictor and the response variable.

Homoscedasticity:

The homoscedasticity assumption states that the variance of the errors should be constant across all levels of the predictor variable. The following are some tests to check for homoscedasticity:

Residual plot: A residual plot can be used to check for the homoscedasticity of the data. If the plot shows a funnel shape, it indicates heteroscedasticity.

Breusch-Pagan test: It tests the null hypothesis that the variance of the errors is constant across all levels of the predictor variable.

White test: It tests the null hypothesis that the errors are homoscedastic.

These tests can be performed using statistical software packages such as Python. It is important to check the assumptions of normality, linearity, and homoscedasticity before performing statistical analyses to ensure that the results are valid and reliable.

9. What is the statical test to be used to evaluate the association between production and a formation (qualitative variable) in a study that has 5 different formations?

Ans. To evaluate the association between production and a qualitative variable (formation in this case), we can use the chi-square test of independence. This test is used to determine whether there is a significant association between two categorical variables.

In this scenario, we have one categorical variable (formation) with 5 levels and another variable (production) which is likely to be continuous or quantitative. Therefore, we need to first categorize or group the production variable, for example, by dividing it into quartiles or quintiles.

After categorizing the production variable, we can create a contingency table that shows the number of observations in each combination of the two variables (formation and production category). We can then perform a chi-square test of independence on the contingency table to determine whether there is a significant association between the two variables.

The null hypothesis for the chi-square test is that there is no association between the two variables, while the alternative hypothesis is that there is an association between the two variables. If the p-value is less than the chosen significance level (e.g. 0.05), we reject the null hypothesis and conclude that there is a significant association between the two variables.

It is important to note that the chi-square test assumes that the observations are independent and that the expected frequency in each cell of the contingency table is at least 5. If these assumptions are violated, then alternative tests such as Fisher's exact test may be used.

10. What are the conclusions of a successful data wrangling process?

Ans. A successful data wrangling process should result in clean, organized, and useful data that can be used for further analysis or modeling. Some conclusions of a successful data wrangling process are:

Data completeness: All the required data should be present without any missing values.

Data consistency: The data should be uniform, without any duplications or inconsistencies, and in the format required for analysis.

Data accuracy: The data should be valid and accurate, without any errors or outliers that could bias the analysis.

Data relevance: The data should be relevant to the problem being addressed, and any irrelevant data should be removed.

Data accessibility: The data should be easily accessible, organized, and documented, so that other team members or stakeholders can use it for analysis.

Overall, a successful data wrangling process should result in high-quality data that can be trusted and used for making informed decisions. It should also save time and resources in the long run by reducing the likelihood of errors and inconsistencies in the analysis.