SOLUTIONS STATISTICS WORKSHEET-6

1) D
2) A
3) A
4) C
5) C
6) A
7) C
8) B
9) B

10) A **boxplo**t and a **histogram** are two different types of visualizations that are used to represent and summarize data.

A boxplot, also known as a box-and-whisker plot, is a type of graph that displays the distribution of a set of numerical data by showing the median, quartiles, and outliers of the data. It provides a quick summary of the shape of the distribution, including skewness, outliers, and whether the data is symmetrical or not.

On the other hand, **a histogram** is a bar graph that represents the distribution of a set of continuous or discrete data by dividing the entire range of values into a set of intervals (also known as "bins") and counting how many values fall into each bin. The height of each bar represents the frequency of the data within each bin. A histogram provides a good sense of the shape of the data distribution, including whether the data is symmetrical, positively or negatively skewed, and whether it has multiple modes.

11) Selecting the appropriate metrics in statistics depends on the research question you're trying to answer and the type of data you have. Here are some general guidelines for selecting metrics:
    a) **Define the research question**: Start by clearly defining what you want to know. This will help you determine which metrics are relevant to your analysis.
    b) **Choose appropriate scale of measurement**: Depending on the scale of measurement of your data (nominal, ordinal, interval, or ratio), you can choose different types of metrics. For example, if you have nominal data, you would use counts or proportions; if you have interval data, you would use means or medians.
    c) **Consider the distribution of your data**: Depending on whether your data is normally distributed or not, you might choose different metrics. For example, if your data is normally distributed, you could use the mean and standard deviation

as measures of central tendency and variability. If your data is not normally distributed, you might consider using the median and interquartile range instead.

d) **Decide between parametric and non-parametric methods**: Depending on the distribution of your data and the type of research question you're trying to answer, you might choose to use either parametric or non-parametric statistical methods. Parametric methods assume that your data follows a normal distribution, while non-parametric methods do not make this assumption.

e) **Think about the interpretation of the results**: Make sure the metrics you choose are easy to interpret and communicate to your audience. Avoid using metrics that are difficult to understand or that require specialized knowledge.

12) To assess the statistical significance of an insight, we need to determine whether the result obtained is likely due to chance or if it is a real effect. There are several methods to do this, including **hypothesis testing**, **p-values**, and **confidence intervals**.

Hypothesis testing is a formal process where you formulate a null hypothesis (e.g., there is no effect) and an alternative hypothesis (e.g., there is an effect), and then use statistical methods to determine which hypothesis is more supported by the data. If the **p-value**, which is the probability of obtaining the observed results if the null hypothesis is true, is less than a certain level (often 0.05), the null hypothesis is rejected and the alternative hypothesis is accepted.

Confidence intervals provide a range of values for a population parameter (e.g., the mean) that is likely to include the true value with a certain level of confidence (e.g., 95%). If the confidence interval does not include zero, it suggests that the effect is statistically significant.

13) There are many types of data that do not follow a Gaussian (normal) distribution or a log-normal distribution. Some examples include:

a) **Categorical data**: This type of data represents a limited number of categories or labels. For example, the color of a fruit (red, green, yellow, etc.). Categorical data can be represented using a frequency table, but it does not have a Gaussian or log-normal distribution.

b) **Exponential data**: This type of data follows an exponential distribution, which is characterized by a long tail on one side. For example, the time between failures of a machine component might be modeled using an exponential distribution.

c) **Poisson data**: This type of data models the number of events that occur in a fixed interval of time or space. For example, the number of calls received by a call center in a given hour might follow a Poisson distribution.

d) **Weibull data**: This type of data models data that has a "bathtub" shaped distribution, with a decreasing hazard rate at the beginning and an increasing hazard rate at the end. This distribution is often used in reliability engineering to model the lifetime of a product.

e) **Pareto data**: This type of data models data that follows a power law distribution, with a few large values and many small values. For example, the size of cities or the wealth of individuals might be modeled using a Pareto distribution.

14) The median is often considered a better measure than the mean in the presence of outliers or extreme values. This is because the median represents the middle value in a dataset, whereas the mean is calculated by adding all values and dividing by the number of values, which can be heavily influenced by outliers.
For example, consider the following dataset of salaries of employees in a company:

₹10,000, ₹15,000, ₹20,000, ₹100,000

The mean salary is ₹31,250, which is significantly higher than the majority of the salaries in the dataset. However, the median salary is ₹20,000, which is a better representation of the "typical" salary in the company, since half of the employees earn less than ₹20,000 and half earn more.

In this case, the median provides a more representative measure of the central tendency of the data, since it is not influenced by the extreme value (the salary of ₹100,000). This is why the median is often preferred in cases where the data is not symmetrical or if there are outliers.

15) The likelihood is a mathematical concept used in statistics to describe the probability of obtaining a certain set of observations given a specific set of parameters. It's a measure of how well a particular model or hypothesis fits the observed data. The likelihood is used to estimate the parameters of a statistical model and to perform hypothesis testing. The likelihood is a function that maps the parameters of a model to the probability of observing the data, and it is usually maximized in order to find the best-fit parameters. The concept of likelihood is used in many different statistical techniques, including regression analysis, Bayesian statistics, and maximum likelihood estimation.