STATISTICS WORKSHEET- 6 ANSWER

1. Which of the following can be considered as random variable?

Ans. All of the mentioned

2. Which of the following random variable that take on only a countable number of possibilities?

Ans. Discrete

3. Which of the following function is associated with a continuous random variable?

Ans. pdf

4. The expected value or _____ of a random variable is the center of its distribution.

Ans. mean

5. Which of the following of a random variable is not a measure of spread?

Ans. variance

6. The _____ of the Chi-squared distribution is twice the degrees of freedom.

Ans. standard deviation

7. The beta distribution is the default prior for parameters between _____

Ans. 0 and 1

8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?

Ans. bootstrap

9. Data that summarize all observations in a category are called _____ data.

Ans. summarized

10. What is the difference between a boxplot and histogram?

Ans. A histogram is a graphical representation of the distribution of a set of numerical data, in which the data is divided into intervals or bins and the count or proportion of data points falling into each bin is shown using bars. It is used to show the underlying frequency distribution of a set of continuous or discrete data.

On the other hand, a box plot (or box and whisker plot) is a way of representing the distribution of a set of data through their quartiles. It shows the median of the data as a line inside a box, with the lower and upper quartiles of the data forming the bottom and top of the box, respectively. The whiskers extend from the box to the minimum and maximum values in the data set, excluding outliers.

In summary, a histogram is used to show the frequency distribution of data, while a box plot is used to show the range, central tendency, and variability of the data.

11. How to select metrics?

Ans. Selecting metrics in statistics involves identifying the relevant measures that will help in evaluating and understanding the performance or behavior of a system, process, or phenomenon. Here are some general steps to consider when selecting metrics in statistics:

- Identify the problem or question: Define the problem or question you want to address and the purpose of your analysis.
- Determine the data needed: Determine the type and amount of data required to answer the problem or question.
- Choose appropriate statistical methods: Choose appropriate statistical methods to analyze the data and answer the question.
- Identify metrics: Identify the metrics that will provide the most useful and relevant information to evaluate and understand the system or process.
- Evaluate the metrics: Evaluate the selected metrics to ensure they are appropriate and reliable.
- Refine the metrics: Refine the selected metrics as needed based on the results of the evaluation.
- Monitor the metrics: Monitor the metrics over time to identify trends and changes in the system or process being measured.

12. How do you assess the statistical significance of an insight?

Ans. To assess the statistical significance of insight, you can perform a hypothesis test. The hypothesis test involves making assumptions about the population parameter of interest and testing whether the observed data is consistent with those assumptions.

The general process for conducting a hypothesis test involves the following steps:

- Formulate the null hypothesis, which is the hypothesis that the parameter is equal to a specific value or falls within a specific range.
- Formulate the alternative hypothesis, which is the hypothesis that the parameter is not equal to the value specified in the null hypothesis.
- Select an appropriate test statistic and calculate its value using the observed data.
- Determine the p-value, which is the probability of obtaining a test statistic at least as extreme as the one observed, assuming the null hypothesis is true.
- Compare the p-value to the significance level (usually denoted by alpha), which is the maximum probability of rejecting the null hypothesis when it is true. If the p-value is less than or equal to the significance level, then the null hypothesis is rejected in favour of the alternative hypothesis. Otherwise, the null hypothesis is not rejected.

By assessing statistical significance, you can determine whether the observed data provide strong evidence in favour of the hypothesis being tested or whether the observed results could have occurred by chance.

13. Give examples of data that does not have a Gaussian distribution, nor log-normal.

Ans. There are many types of data that do not have a Gaussian (normal) distribution or a log-normal distribution. Here are a few examples:

- Poisson distributed data: This type of data arises when the number of occurrences of an event in a fixed interval of time or space follows a Poisson distribution. Examples include the

number of calls to a call center per hour, the number of cars passing a point on a road per minute, or the number of accidents per day.

- Binomial distributed data: This type of data arises when the outcome of an experiment can be one of two possible outcomes, and the probability of success is constant. Examples include the number of heads in a series of coin flips or the number of defective items in a production batch.
- Exponential distributed data: This type of data arises when the time between two successive events follows an exponential distribution. Examples include the time between arrivals at a queue, the time between calls to a call center, or the time between equipment failures.
- Power-law distributed data: This type of data arises when the frequency of an event is inversely proportional to its size raised to a power. Examples include the frequency of earthquakes, the distribution of income, or the number of citations of scientific papers.
- Beta distributed data: This type of data arises when the distribution of a proportion or a probability is described by a beta distribution. Examples include the proportion of successes in a series of Bernoulli trials or the probability of a user clicking on an ad.

14. Give an example where the median is a better measure than the mean.

Ans. The median is a better measure than the mean when dealing with skewed distributions or when there are extreme outliers. One example where the median is a better measure than the mean is household income. In many countries, the distribution of household income is highly skewed, with a small number of very high-income households and a large number of lower-income households. The mean household income is likely to be pulled upward by the high-income households, making it a less representative measure of central tendency. In this case, the median household income, which is the income level that separates the top 50% of households from the bottom 50%, is a better measure of central tendency because it is not affected by extreme values.

15. What is the Likelihood?

Ans. Likelihood is a concept in statistics that refers to the probability of obtaining a set of observations given a particular value of a parameter in a statistical model. The likelihood function is used to estimate the parameter values that are most likely to have produced the observed data. It is a function of the parameter of the model and is often used in maximum likelihood estimation. The likelihood function is proportional to the joint probability of the data given the parameter values, and it can be used to compare different models or to test hypotheses about the values of parameters.