STATISTICS WORKSHEET-6

Q1to Q15 are descriptive types. Answer in brief.

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

Answer: a) The outcome from the roll of a die

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

Answer: a) Discrete

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

Answer: a) pdf

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

Answer: c) mean

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

Answer: a) variance

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

Answer: b) standard deviation

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

Answer: c) 0 and 1

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

Answer: b) bootstrap

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

Answer: b) summarized

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

Answer: A histogram is a graphical representation of the distribution of continuous data by showing the frequency or proportion of observations in each bin, while a boxplot is a graphical summary of the distribution of any type of data by displaying the median, interquartile range, and outliers. The main difference is that histograms display the shape, center, and spread of the data, while boxplots provide a summary of the distribution, including the median, spread, and presence of outliers.

11. How to select metrics?

Answer: Selecting appropriate metrics in statistics depends on the specific research question or problem being investigated. Metrics should be selected based on their ability to accurately capture and quantify the phenomenon of interest. Here are some general steps to guide the process of selecting metrics:

1. Clearly define the research question or problem to be investigated.
2. Identify the variables that are relevant to the research question or problem.
3. Determine the appropriate scale of measurement for each variable (e.g. nominal, ordinal, interval, or ratio).
4. Consider what characteristics of the data are most important to measure in order to answer the research question or address the problem.
5. Choose metrics that are appropriate for the scale of measurement and characteristics of the data.
6. Evaluate the chosen metrics to ensure that they are meaningful, relevant, and interpretable.
7. Consider potential limitations or sources of bias in the selected metrics, and address these as appropriate.

Overall, selecting appropriate metrics in statistics requires careful consideration of the research question or problem, the relevant variables, and the characteristics of

the data. It is important to choose metrics that accurately capture the phenomenon of interest and are meaningful, relevant, and interpretable.

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

Answer:

1) state the research hypothesis

2) state the null hypothesis

3)select a probability of error level

4)select and compute the test for statistical significance

5)interpret the results

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

Answer: There are many types of data that do not follow a Gaussian or log-normal distribution. Here are some examples:

1. **Skewed data: Data that are skewed to one side or the other, with a long tail in one direction or the other. For example, income data often have a long tail to the right, because there are a few very high earners that skew the distribution.**
2. **Bimodal data: Data that have two peaks or modes in the distribution. For example, if we measured the heights of people in a population, we might see two peaks in the distribution corresponding to men and women.**
3. **Categorical data: Data that are not numerical, but rather represent categories or labels. For example, the type of car a person drives (e.g. sedan, SUV, truck) would be categorical data.**
4. **Count data: Data that represent counts of discrete events, such as the number of cars that pass through a toll booth in a given hour.**
5. **Power-law distributed data: Data that follow a power-law distribution, which is characterized by a few very large values and many small values. For example, the frequency of words in a natural language text often follows a power-law distribution.**

6. **Exponential data: Data that follow an exponential distribution, which is characterized by a rapid drop-off in the frequency of events as the magnitude of the event increases. For example, the time between earthquakes follows an exponential distribution.**

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

Answer: The median is often a better measure of central tendency than the mean when dealing with skewed data or data with outliers. In such cases, the mean can be significantly affected by extreme values, and may not accurately represent the typical value of the data set.

For example, consider the following data set representing the annual salaries of employees in a company:

$30,000, $35,000, $40,000, $45,000, $50,000, $55,000, $60,000, $65,000, $70,000, $75,000, $80,000, $85,000, $90,000, $95,000, $100,000

The mean salary in this data set is calculated as:

(30,000 + 35,000 + 40,000 + 45,000 + 50,000 + 55,000 + 60,000 + 65,000 + 70,000 + 75,000 + 80,000 + 85,000 + 90,000 + 95,000 + 100,000) / 15 = $62,000

However, if one employee in the company is a CEO with a salary of $1,000,000, the mean salary would increase to $125,000, which would not accurately represent the typical salary in the company.

In contrast, the median salary in this data set is the middle value when the salaries are arranged in order: $50,000

The median is not affected by the extreme value of the CEO's salary, and provides a more accurate representation of the typical salary in the company.

15. What is the Likelihood?

Answer: In statistics, likelihood is the probability of observing a set of data given a specific parameter value or set of parameter values in a statistical model. It is often used to estimate the parameters of a model, compare different models, and make inferences about the parameters. The likelihood function is a function of the parameters, given the observed data, and is not the same as the probability distribution of the data.