# STATISTICS WORKSHEET- 6

**Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.**

1. Which of the following can be considered as random variable?
   a) The outcome from the roll of a die
   b) The outcome of flip of a coin
   c) The outcome of exam
   d) All of the mentioned
   ANS: d
2. Which of the following random variable that take on only a countable number of possibilities?
   a) Discrete
   b) Non Discrete
   c) Continuous
   d) All of the mentioned

   ANS: a
3. Which of the following function is associated with a continuous random variable?
   a) pdf
   b) pmv
   c) pmf
   d) all of the mentioned

   ANS: a
4. The expected value or________of a random variable is the center of its distribution.
   a) mode
   b) median
   c) mean
   d) bayesian inference

   ANS: c
5. Which of the following of a random variable is not a measure of spread?
   a) variance
   b) standard deviation
   c) empirical mean
   d) all of the mentioned
   ANS: c

6. The__________of the Chi-squared distribution is twice the degrees of freedom.
   a) variance
   b) standard deviation
   c) mode
   d) none of the mentioned

   ANS: a

7. The beta distribution is the default prior for parameters between _____________
   a) 0 and 10
   b) 1 and 2
   c) 0 and 1
   d) None of the mentioned
   ANS: c
8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?

a) baggyer
b) bootstrap
c) jacknife
d) none of the mentioned

ANS: b

9. Data that summarize all observations in a category are called____________data.

a) frequency
b) summarized
c) raw
d) none of the mentioned

ANS: a

**Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.**

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

A histogram is a graph that displays the distribution of a numerical variable. It consists of a series of bars, where each bar represents a range of values and the height of the bar represents the frequency or proportion of observations that fall within that range. Histograms are useful for showing the shape of the distribution, the center of the distribution, and the spread of the distribution.

A boxplot, on the other hand, is a graph that displays the distribution of a numerical variable using five summary statistics: the minimum value, the first quartile (25th percentile), the median (50th percentile), the third quartile (75th percentile), and the maximum value. The boxplot consists of a rectangle (the box) that spans from the first to the third quartile, a line (the median) inside the box, and two whiskers that extend from the box to the minimum and maximum values. Boxplots are useful for showing the center and spread of the distribution, as well as identifying outliers.

In summary, while a histogram shows the distribution of a variable using bars, a boxplot displays the distribution using summary statistics and visualizes potential outliers.

11. How to select metrics?

Selecting appropriate metrics is an important step in analyzing and interpreting data. Here are some steps to follow when selecting metrics:

Define the objective: Start by defining the objective of your analysis. What are you trying to achieve? What questions are you trying to answer? This will help you identify the relevant metrics that will help you achieve your objectives.

Identify key performance indicators (KPIs): KPIs are metrics that measure the performance of a specific aspect of your business or project. They help you track progress towards your goals and identify areas for improvement. Identify the KPIs that are most relevant to your objective.

Choose measurable and relevant metrics: Choose metrics that are measurable and relevant to your objective. Make sure that the metrics you choose are easy to collect, analyze, and report. Avoid using metrics that are difficult to measure or not relevant to your objective.

Consider context: Consider the context in which the metrics will be used. Will the metrics be used to compare performance over time, across different departments, or against competitors? Make sure that the metrics you choose are appropriate for the context in which they will be used.

Balance metrics: Select a balance of metrics that provide a holistic view of performance. Avoid over-reliance on a single metric, as this can lead to a narrow focus and potential blind spots. Instead, choose a variety of metrics that provide a comprehensive picture of performance.

Test and refine: Test the metrics you have chosen and refine them as necessary. Monitor their performance over time and adjust them as needed to ensure that they continue to provide valuable insights.

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

Assessing the statistical significance of an insight involves determining the probability that the observed results occurred by chance or are actually representative of a true relationship or effect. Here are some general steps to assess statistical significance:

Define the null hypothesis: The null hypothesis is the assumption that there is no relationship or effect between the variables being studied.

Determine the appropriate statistical test: The choice of statistical test depends on the nature of the data, the research question, and the study design. Common statistical tests include t-tests, ANOVA, chi-square tests, and regression analysis.

Calculate the p-value: The p-value is the probability of obtaining the observed results or more extreme results under the null hypothesis. A p-value of 0.05 or less is typically considered statistically significant, indicating that the observed results are unlikely to have occurred by chance.

Interpret the results: If the p-value is less than the predetermined level of significance, the null hypothesis can be rejected, and the results are considered statistically significant. However, it's important to also consider effect size, confidence intervals, and other factors when interpreting statistical significance.

Consider the limitations: Statistical significance does not necessarily imply practical or clinical significance, and it's important to consider the limitations of the study design, sample size, and potential confounding variables when interpreting the results.

13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal.

Here are some examples of data that do not have a Gaussian distribution or a log-normal distribution:

Power law distribution: Power law distributions are characterized by a small number of values occurring frequently and a large number of values occurring rarely. Examples of power law distributions include the distribution of city populations, the distribution of the frequency of words in a language, and the distribution of the sizes of earthquakes.

Pareto distribution: The Pareto distribution is another example of a power law distribution that is often used to describe the distribution of wealth, income, or other resources in a population.

Exponential distribution: The exponential distribution is a probability distribution that describes the time between events in a Poisson process. It is often used in reliability engineering to model the failure time of mechanical or electronic systems.

Gamma distribution: The gamma distribution is a two-parameter family of continuous probability distributions that is often used to model the waiting time between Poisson distributed events. It can also be used to model the distribution of sizes of particles or the time taken to complete a task.

Weibull distribution: The Weibull distribution is another widely used distribution in reliability engineering, as it is used to model the distribution of failure times of mechanical or electronic systems. It can also be used to model the lifetime of certain biological organisms or the time to failure of a product.

These are just a few examples of non-Gaussian and non-log-normal distributions that are commonly used in different fields. There are many other types of distributions as well that can be used to model different types of data.

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

The median and the mean are both measures of central tendency that describe the typical value of a dataset. However, the median is a better measure than the mean in certain situations, such as when the dataset contains outliers or when the distribution is skewed. Here's an example to illustrate this:

Suppose a small company has four employees and their monthly salaries are as follows:

Employee A: $50,000
Employee B: $55,000
Employee C: $60,000
Employee D: $1,000,000

The mean salary of these employees is ($50,000 + $55,000 + $60,000 + $1,000,000) / 4 = $291,250, which is heavily influenced by the outlier salary of $1,000,000. This mean value may not accurately reflect the typical salary of an employee in this company.

On the other hand, the median salary of these employees is the middle value when the salaries are arranged in order, which is $57,500. The median is not affected by the outlier salary of $1,000,000 and provides a more accurate representation of the typical salary in the company.

In this example, the median is a better measure than the mean because it is less sensitive to outliers and gives a better representation of the central tendency of the dataset.

15. What is the Likelihood?

In statistics, likelihood refers to the probability of observing a set of data, given a particular value of a parameter or set of parameters in a statistical model. It is commonly denoted by the symbol L(theta | x), where theta represents the parameter(s) of the model and x represents the observed data.

The likelihood function is the probability density function (PDF) or probability mass function (PMF) of the data, treated as a function of the unknown parameters. The likelihood function is a key component of maximum likelihood estimation, a method for estimating the parameters of a statistical model based on observed data.

In simple terms, likelihood is a measure of how well a statistical model fits the observed data. The goal of maximum likelihood estimation is to find the values of the parameters that maximize the likelihood function, which are the values that make the observed data most probable under the model.

A likelihood function is an important tool for comparing different statistical models and selecting the model that best fits the data. The model with the highest likelihood is often considered the most plausible or likely explanation for the observed data.