FLIP ROBO

WORKSHEET

STATISTICS WORKSHEET- 6

1. Which of the following can be considered as random variable?
d) All of the mentioned

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

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

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

5. Which of the following of a random variable is not a measure of spread?
c) empirical mean

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

7. The beta distribution is the default prior for parameters between _____
c) 0 and 1

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

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

10. What is the difference between a boxplot and histogram?
Ans. A histogram and a boxplot are both graphical methods for visualizing the distribution of a continuous variable, but they differ in their approach and the information they provide.
A histogram displays the distribution of the data by dividing the range of values into a set of intervals, called bins, and counting the number of observations that fall into each bin. The bins are typically represented by adjacent rectangles that touch each other, and the height of each rectangle represents the frequency or density of observations in that bin. Histograms provide information on the shape, center, and spread of the distribution, and can reveal patterns such as symmetry, skewness, or multimodality.
A boxplot, also called a box-and-whisker plot, displays the distribution of the data by showing the median, quartiles, and range of the variable. The box in the middle of the plot represents the interquartile range (IQR), which is the range of values between the 25th and 75th percentiles of the data. The whiskers extend from the box to the minimum and maximum values that are within 1.5 times the IQR from the edges of the box. Points outside the whiskers are considered outliers and are plotted individually.
Boxplots provide information on the median, range, and spread of the distribution, and can reveal patterns such as skewness, outliers, and variability.

In summary, histograms show the distribution of the data by dividing it into bins and counting the frequency in each bin, while boxplots summarize the distribution by showing the median, quartiles, and range of the variable. Both methods are useful for visualizing and summarizing the distribution of continuous data, and can provide complementary insights into the data.

11. How to select metrics?

Ans. Selecting the right metrics is an important step in any data analysis project or performance measurement initiative. The following are some general steps that can be followed to select metrics:

Define the goals and objectives: The first step is to clearly define the goals and objectives of the project or initiative. This will help to identify the key performance indicators (KPIs) that are most relevant to measuring progress towards those goals.

Identify the stakeholders: Identify the stakeholders who will be impacted by the project or initiative, and consider their needs and expectations. This will help to identify the metrics that are most relevant to those stakeholders.

Identify the data sources: Identify the data sources that are available or can be collected to measure the KPIs. Consider the quality and reliability of the data, as well as any limitations or biases that may affect the results.

Define the metrics: Based on the goals, objectives, stakeholders, and data sources, define the metrics that will be used to measure progress towards the KPIs. Metrics should be specific, measurable, and relevant to the goals and objectives.

Test the metrics: Test the metrics to ensure that they are reliable, valid, and actionable. This may involve collecting additional data, comparing results with other sources, or using statistical methods to test the validity and reliability of the metrics.

Communicate the metrics: Communicate the metrics and results to stakeholders in a clear and concise manner. This may involve using data visualization techniques, such as graphs, charts, and dashboards, to help stakeholders understand the results and take action based on the insights gained from the metrics.

Overall, the process of selecting metrics should be guided by the goals and objectives of the project or initiative, as well as the needs and expectations of stakeholders. By carefully selecting and testing metrics, and communicating the results effectively, data-driven decisions can be made to improve performance and achieve the desired outcomes.

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

Ans. To assess the statistical significance of an insight, one typically performs hypothesis testing using statistical methods. Hypothesis testing involves setting up a null hypothesis and an alternative hypothesis, and then calculating a test statistic and a p-value to determine the likelihood of observing the results if the null hypothesis were true.

The following are the general steps for assessing statistical significance:

Define the null and alternative hypotheses: The null hypothesis is the assumption that there is no relationship or difference between variables, while the alternative hypothesis is the opposite. For example, the null hypothesis may be that there is no difference in means between two groups, while the alternative hypothesis may be that there is a difference.

Select the appropriate statistical test: The statistical test selected depends on the type of data and research question being investigated. Common statistical tests include t-tests, ANOVA, chi-squared tests, and regression analysis.

Calculate the test statistic and p-value: The test statistic is calculated based on the data and the selected statistical test. The p-value is the probability of observing the results if the null hypothesis were true. A p-value less than the predetermined significance level (often 0.05) indicates that the results are statistically significant and the null hypothesis can be rejected.

Interpret the results: If the p-value is less than the significance level, then the results are statistically significant and the null hypothesis can be rejected. If the p-value is greater than the significance level, then the results are not statistically significant and the null hypothesis cannot be rejected.
It is important to note that statistical significance does not necessarily imply practical significance or importance. It is also important to consider the context and limitations of the study when interpreting the results.

13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal.
Ans. There are many types of data that do not have a Gaussian distribution or a log-normal distribution. Here are some examples:
Skewed data: Data that are heavily skewed to the left or right are not normally distributed. For example, income data often have a long tail to the right, indicating that a small proportion of people have very high incomes.
Bimodal data: Data that have two distinct peaks or modes are not normally distributed. For example, the heights of people may have two peaks due to the presence of two distinct sub-populations (e.g., men and women).
Categorical data: Data that are nominal or ordinal in nature, such as gender, race, or education level, are not normally distributed.
Count data: Data that represent counts or frequencies, such as the number of births per day or the number of phone calls received per hour, are often not normally distributed.
Power law distributions: Data that follow a power law distribution, which has a long tail, are not normally distributed. For example, the distribution of the number of followers for social media accounts often follows a power law distribution.
Exponential distributions: Data that follow an exponential distribution, which has a rapidly decreasing probability density function, are not normally distributed. For example, the time between customer arrivals in a queue may follow an exponential distribution.

14. Give an example where the median is a better measure than the mean.
Ans. The median is often a better measure than the mean in situations where the data are skewed or have extreme values (outliers). Here is an example:
Suppose we want to calculate the "average" income of people in a small town. We collect data from 10 people, and their incomes (in thousands of dollars per year) are:
$20, $25, $30, $30, $30, $30, $35, $40, $45, $100
The mean income of these people is (20+25+30+30+30+30+35+40+45+100)/10 = $37.5 thousand per year. However, we can see that the income distribution is heavily skewed to the right, with an outlier value of $100 thousand per year. The mean is heavily influenced by this outlier value, and does not accurately represent the "typical" income of people in the town.
In this situation, the median income may be a better measure of central tendency. The median is the middle value of the data when it is arranged in ascending or descending order. In this case, the median income is $30 thousand per year, which is a more representative value of the "typical" income in the town, since it is not affected by the outlier value.
Therefore, in situations where the data are skewed or have outliers, the median may be a better measure than the mean.

15. What is the Likelihood?
Ans. In statistics, the likelihood refers to the probability of observing a set of data given a particular model and set of parameters. It is often used in the context of maximum likelihood estimation, which is a method for estimating the parameters of a statistical model that maximize the likelihood of the observed data.

More formally, suppose we have a set of observed data X, and a statistical model with a set of parameters θ. The likelihood function $L(\theta|X)$ is defined as the probability of observing the data X given the parameters θ. Mathematically, the likelihood function is expressed as:

$L(\theta|X) = P(X|\theta)$

The maximum likelihood estimator (MLE) is the value of the parameter that maximizes the likelihood function, given the observed data. In other words, the MLE is the value of θ that makes the observed data most probable, according to the model.

The likelihood is an important concept in statistical inference, as it allows us to compare different models and estimate their parameters based on observed data. By maximizing the likelihood function, we can find the "best fit" parameters for the model, and use these parameters to make predictions or test hypotheses about the underlying population.