# STATISTICS WORKSHEET- 6- Ans

| | |
|---|---|
| 1 | Which of the following can be considered as random variable?<br><br>a) The outcome from the roll of a die<br>b) The outcome of flip of a coin<br>c) The outcome of exam<br>d) All of the mentioned |
| Ans | **d) All of the mentioned** |
| 2 | Which of the following random variable that take on only a countable number of possibilities?<br>a) Discrete<br>b) Non Discrete<br>c) Continuous<br>d) All of the mentioned |
| | a) Discrete |
| 3 | Which of the following function is associated with a continuous random variable?<br>a) pdf<br>b) pmv<br>c) pmf<br>d) all of the mentioned |
| Ans | a) pdf |
| | |
| 4 | The expected value or _____ of a random variable is the center of its distribution.<br>a) mode<br>b) median<br>c) mean<br>d) bayesian inference |
| Ans | c) mean |
| | |
| 5 | Which of the following of a random variable is not a measure of spread?<br>a) variance<br>b) standard deviation<br>c) empirical mean<br>d) all of the mentioned |
| Ans | c) empirical mean |
| | |
| 6 | The _____ of the Chi-squared distribution is twice the degrees of freedom. |

| | |
|---|---|
| | a) variance<br>b) standard deviation<br>c) mode<br>d) none of the mentioned |
| Ans | a) variance |
| | |
| 7 | The beta distribution is the default prior for parameters between _____<br>a) 0 and 10<br>b) 1 and 2<br>c) 0 and 1<br>d) None of the mentioned |
| Ans | c) 0 and 1 |
| | |
| 8 | Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?<br>a) baggyer<br>b) bootstrap<br>c) jacknife<br>d) none of the mentioned |
| Ans | b) bootstrap |
| | |
| 9 | Data that summarize all observations in a category are called _____ data.<br>a) frequency<br>b) summarized<br>c) raw<br>d) none of the mentioned |
| Ans | **a) frequency** |
| 10 | What is the difference between a boxplot and histogram? |
| | A boxplot and histogram are both graphical tools used to summarize and display data, but they represent different aspects of the data.<br><br>A histogram is a graph that displays the distribution of a dataset by showing the frequency of observations that fall into different bins or intervals. The x-axis represents the range of values in the dataset, and the y-axis represents the frequency or proportion of observations that fall into each bin. Histograms are useful for identifying the shape of the distribution, including any patterns such as skewness or bimodality.<br><br>A boxplot, also known as a box-and-whisker plot, is a graph that displays the distribution of a dataset by showing the range, median, and quartiles of the data. The box represents the middle 50% of the data, with the median indicated by a line inside |

| | |
|---|---|
| | the box. The whiskers extend to the minimum and maximum values that fall within a certain range, typically 1.5 times the interquartile range (IQR) above and below the box. Boxplots are useful for identifying outliers and visualizing the spread and central tendency of the data.<br><br>In summary, a histogram provides a visual representation of the frequency or density distribution of the data, while a boxplot summarizes the spread, center, and outliers of the data. Both are useful tools for exploratory data analysis and can be used in combination to gain a more complete understanding of the data. |
| | |
| 11 | How to select metrics? |
| | Selecting appropriate metrics is a critical step in any data analysis or decision-making process. Here are some steps to follow when selecting metrics:<br><br>**Define the goal:** Start by clearly defining the goal or objective of the analysis or decision-making process. This will help identify the key factors that need to be measured and evaluated.<br><br>Identify key performance indicators (KPIs): Identify the KPIs that are relevant to the goal or objective. KPIs are specific metrics that are used to track progress towards a particular goal or objective. They should be relevant, measurable, and aligned with the overall goal.<br><br>Evaluate the data: Evaluate the available data to determine which metrics are feasible to collect and analyze. Consider the accuracy and completeness of the data, as well as any potential biases or limitations.<br><br>Consider trade-offs: Consider the trade-offs between different metrics, as well as any potential unintended consequences. For example, a metric that focuses on short-term gains may not be sustainable in the long run.<br><br>Test and refine: Test the selected metrics and evaluate their effectiveness in achieving the goal or objective. Refine the metrics as needed based on feedback and results.<br><br>Communicate: Communicate the selected metrics and their importance to stakeholders to ensure buy-in and alignment towards the goal or objective.<br><br>By following these steps, we can select metrics that are relevant, effective, and aligned with the overall goal or objective. |
| | |

| | |
|---|---|
| 12 | How do you assess the statistical significance of an insight? |
| | To assess the statistical significance of an insight, you typically need to perform hypothesis testing. Here are the general steps:<br>**Define the null hypothesis**: Start by defining the null hypothesis, which is the default assumption that there is no significant difference or relationship between two variables.<br><br>**Choose a significance level**: Choose a significance level, typically denoted by alpha (α), which is the probability of rejecting the null hypothesis when it is actually true. A commonly used significance level is 0.05, which means that there is a 5% chance of rejecting the null hypothesis when it is true.<br><br>**Collect and analyze the data**: Collect and analyze the data to calculate a test statistic, which is a measure of how far the observed data deviate from the null hypothesis. The choice of test statistic depends on the type of hypothesis test and the type of data being analyzed.<br><br>**Calculate the p-value:** Calculate the p-value, which is the probability of observing a test statistic as extreme or more extreme than the observed value, assuming the null hypothesis is true. A p-value less than the significance level indicates that the observed data are unlikely to have occurred by chance alone and provides evidence to reject the null hypothesis.<br><br>**Interpret the results:** Interpret the results based on the p-value and the chosen significance level. If the p-value is less than the significance level, reject the null hypothesis and conclude that the observed data provide evidence of a significant difference or relationship between the variables. |
| | |
| 13 | Give examples of data that does not have a Gaussian distribution, nor log-normal. |
| ANS | Here are some examples of data that do not have a Gaussian (normal) distribution:<br>**Binary data:** Data that can only take on two values, such as yes or no, true or false, or 0 or 1. This type of data follows a Bernoulli distribution.<br>**Count data**: Data that consists of whole numbers that count the frequency of an event, such as the number of emails received in a day or the number of customers who enter a store in an hour. This type of data follows a Poisson distribution.<br>**Skewed data:** Data that has a long tail on one side of the distribution, such as income or house prices. This type of data can follow a log-normal or a Weibull distribution.<br><br>**Bimodal data:** Data that has two distinct peaks in the distribution, such as the number of hours of sleep per night in a population. This type of data can follow a mixture of two normal distributions or a bimodal distribution. |

| | |
|---|---|
| | **Uniform data:** Data that is equally likely to fall within any range of values, such as the result of rolling a fair die or the temperature at which a chemical reaction occurs. This type of data follows a uniform distribution. |
| | |
| 14 | Give an example where the median is a better measure than the mean. |
| | The median is a better measure than the mean when dealing with skewed data or when extreme values (outliers) are present. In such cases, the mean can be greatly influenced by these extreme values and may not be a good representation of the typical value of the data. The median, on the other hand, is less affected by extreme values and can provide a more accurate representation of the typical value.<br><br>One example where the median is a better measure than the mean is in household income data. Income data is often skewed to the right, meaning that there are a few households with very high income that greatly influence the mean. In this case, the mean income may not be a good representation of the typical income of households, as most households have income below the mean. Instead, the median income, which is the income value that falls in the middle of the income range when all households are ranked by income, may provide a more accurate representation of the typical income of households. |
| | |
| 15 | What is the Likelihood? |
| | In statistics, likelihood refers to the probability of obtaining a particular set of observations or data, given a specific hypothesis or model.<br><br>The likelihood function is a function of the parameters of the model and measures how well the model fits the observed data. It is defined as the conditional probability of the observed data, given the values of the model parameters. The likelihood function plays a crucial role in statistical inference, as it helps us to estimate the unknown parameters of a model by maximizing the likelihood function.<br><br>The likelihood is often confused with probability, but they are not the same. Probability refers to the chance of an event occurring, whereas likelihood refers to the fit of a model to the data.<br><br>To summarize, likelihood is a measure of how well a model fits the observed data, and it is used in statistical inference to estimate the parameters of the model. |