Worksheet-Set 6: Statistics Solutions

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.**

   b) median

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.**

   d) none of the mentioned

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

Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets

11. How to select metrics?

The metrics selection depends upon the the target variable, if it is categorical or continuous, then in categorical what is more important to us TPR, FPR and in regression if data distribution.

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

Statistical significance is often calculated with statistical hypothesis testing, which tests the validity of a hypothesis by figuring out the probability that your results have happened by chance.

Here, a "hypothesis" is an assumption or belief about the relationship between your datasets. The result of a hypothesis test allows us to see whether this assumption holds under scrutiny or not.

A standard hypothesis test relies on two hypotheses.

Null hypothesis: The default assumption of a statistical test that you're attempting to disprove (e.g., an increase in cost won't affect the number of purchases).

Alternative hypothesis: An alternate theory that contradicts your null hypothesis (e.g., an increase in cost will reduce the number of purchases). This is the hypothesis you hope to prove.

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

Bimodal distribution: *A bimodal distribution has two distinct peaks, which means it is not Gaussian or log-normal. An example of bimodal data could be the heights of people in a population with two distinct subpopulations, such as adults and children.*

Exponential distribution: *An exponential distribution describes the time between events that occur at a constant rate. It is not Gaussian or log-normal because it is skewed to the right, with a long tail of large values. An example of exponential data could be the time between successive radioactive decays.*

Poisson distribution: *A Poisson distribution describes the number of events that occur in a fixed interval of time, given that events occur at a constant rate. It is not Gaussian or log-normal because it is a discrete distribution and has a different shape. An example of Poisson data could be the number of accidents at a particular intersection in a fixed time interval.*

***Uniform distribution**: A uniform distribution has a constant probability density function over a specific range of values. It is not Gaussian or log-normal because it has a rectangular shape, with equal probabilities across the range. An example of uniform data could be the result of rolling a fair six-sided die.*

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

The median is often a better measure than the mean when dealing with skewed distributions or data sets that contain outliers. In such cases, the mean may be heavily influenced by extreme values and not accurately reflect the central tendency of the data.

Here's an example to illustrate this point:

Suppose we want to measure the salaries of employees at a small company. Most employees earn a salary between $30,000 and $70,000 per year, but there are also a few high-level executives who earn much more. One of the executives earns $1,000,000 per year. We can represent this data set as follows:

$30,000, $40,000, $50,000, $60,000, $70,000, $1,000,000

The mean salary for this data set is:

(30,000 + 40,000 + 50,000 + 60,000 + 70,000 + 1,000,000) / 6 = $198,333.33

As we can see, the mean is heavily influenced by the executive's salary, which is an outlier in the data set. This value does not represent the typical salary at the company.

In this case, the median is a better measure of central tendency than the mean. The median is the middle value when the data set is ordered from smallest to largest. For this data set, the median is $55,000. This value better represents the typical salary at the company, as it is not affected by the extreme value of the executive's salary.

Therefore, in this example, the median is a better measure than the mean.

15. What is the Likelihood?

In statistics, likelihood is a concept that describes how well the parameters of a statistical model explain the observed data. Specifically, it is a function that quantifies the probability of obtaining the observed data given a particular set of values for the parameters of the model.

The likelihood function is usually denoted by $L(\vartheta \mid X)$, where ϑ is the vector of parameters for the model and X is the observed data. The likelihood function is defined as the joint probability density function of the observed data, given the parameters of the model. This means that the likelihood function gives us the probability of observing the data that we did, given the values of the parameters.

In simple terms, the likelihood tells us how probable the data is, given a particular set of parameter values. The goal of maximum likelihood estimation is to find the set of parameter values that maximize the likelihood of the observed data.

It is important to note that likelihood is not the same as probability. Probability describes the chance of an event occurring given a specific set of conditions, whereas likelihood describes the support of the observed data for a particular set of parameter values.