STATISTICS WORKSHEET- 6

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

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

3. Which of the following function is associated with a continuous random variable?
   Ans. d) all of the mentioned

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

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

6. The _________ of the Chi-squared distribution is twice the degrees of freedom.
   Ans. b) standard deviation

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

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

9. Data that summarize all observations in a category are called __________ data.
   Ans. b) summarized

10. What is the difference between a boxplot and histogram?
    Ans. A histogram, a visual representation, uses bars to display the data value to display the frequency of data items in a series of equal-sized numerical intervals. The X and Y axes display the interval sizes and frequencies, respectively. Each bar's height indicates the frequency of each interval size. It presents data in a way that makes it simpler to identify a process's dispersion and central tendency. The histogram allows us to look at the distribution and shape of the data. When the sample size is greater than 50, it is most effective.

A Box plot is a way to visualize the distribution of the data by using a box and some vertical lines. It is known as the whisker plot. The data can be distributed between five key ranges, which are as follows:

1. **Minimum: Q1-1.5*IQR**
2. **1st quartile (Q1): 25th percentile**
3. **Median:50th percentile**
4. **3rd quartile(Q3):75th percentile**
5. **Maximum: Q3+1.5*IQR**

Here IQR represents the InterQuartile Range which starts from the first quartile (Q1) and ends at the third quartile (Q3).

11. How to select metrics?

Ans. First of all, metrics which we optimise tweaking a model and performance evaluation metrics in machine learning are not typically the same. Below, we discuss metrics used to optimise Machine Learning models. For performance evaluation, initial business metrics can be used.

Understanding the task
Based on prerequisites, we need to understand what kind of problems we are trying to solve. Here is a list of some common problems in machine learning:

- **Classification. This algorithm will predict data type from defined data arrays. For example, it may respond with yes/no/not sure.**
- **Regression. The algorithm will predict some values. For example, weather forecast for tomorrow.**
- **Ranking. The model will predict an order of items. For example, we have a student group and need to rank all the students depending on their height from the tallest to the shortest.**

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

Ans. 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.

Researchers use a measurement known as the p-value to determine statistical significance: if the p-value falls below the significance level, then the result is statistically significant. The p-value is a function of the means and standard deviations of the data samples.

Statistical significance is most practically used in hypothesis testing. For example, you want to know whether changing the color of a button on your website from red to green will result in more people clicking on it.

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

Ans. Exponential distributions do not have a log-normal distribution or a Gaussian distribution. In fact, any type of data that is categorical will not have these distributions as well. Example: Duration of a phone car, time until the next earthquake, etc.

Any distribution of money or value will be non--Gaussian. For example: distributions of income; distributions of house prices; distributions of bets placed on a sporting event. These distributions cannot have negative values and will usually have extended right hand tails.

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

Ans. The mean is used for normal distributions. The median is generally used for skewed distributions. The mean is not a robust tool since it is largely influenced by outliers. The median is better suited for skewed distributions to derive at central tendency since it is much more robust and sensible.

Income is the classic example of when to use the median instead of the mean because its distribution tends to be skewed. The median indicates that half of all incomes fall below 27581, and half are above it. For these data, the mean overestimates where most household incomes fall.

15. What is the Likelihood?

    Ans. Likelihood, being the outcome of a likelihood function thus defined, describes the plausibility, under a certain statistical model (the null hypothesis in hypothesis testing), of a certain parameter value after observing a particular outcome.

There didn't seem much likelihood of it happening. There is every likelihood that sanctions will work. If something is a likelihood, it is likely to happen. The likelihood is that your child will not develop diabetes.