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
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
3. Which of the following function is associated with a continuous random variable?
   a) pdf
   b) pmv
   c) pmf
   d) all of the mentioned
4. The expected value or_____of a random variable is the center of its distribution.
   a) mode
   b) median
   c) mean
   d) bayesian inference
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
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
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
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) one of the mentioned

9. Data that summarize all observations in a category are called_____data.
   a) frequency
   b) summarized
   c) raw
   d) none of the mentioned

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

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. They are less detailed than histograms and take up less space.

Although histograms are better in displaying the distribution of data, you can use a box plot to tell if the distribution is symmetric or skewed. In a symmetric distribution, the mean and median are nearly the same, and the two whiskers has almost the same length.

Histograms and box plots are very similar in that they both help to visualize and describe numeric data. Although histograms are better in determining the underlying distribution of the data, box plots allow you to compare multiple data sets better than histograms as they are less detailed and take up less space.

11. How to select metrics?

The metrics are chosen on terms of nature of the problem. Classification , Regression and unsupervised learning all have different metrics. Also based on the problem given to decide if we want specificity or sensitivity also where and how the results would would be applied in real word. To know distribution of the target variable.

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

Statistical significance can be accessed using hypothesis testing: – Stating a null hypothesis which is usually the opposite of what we wish to test. we choose a suitable statistical test and statistics used to reject the null hypothesis and choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)
We then calculate the observed test statistics from the data and check whether it lies in the critical region. There are multiple test we performed based on the nature of the problem and features of our data set.

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

- Allocation of wealth among individuals
- Values of oil reserves among oil fields (many small ones, a small number of large ones)
- Life table is example of exponential distribution
- Wind speed is Weibull distribution
- Surgery patient's stay in hospital is gamma distribution
- Another example is the location of the centers of raindrop ripples on a pond; they are not uniformly spaced in (say) the east-west direction, but they are uniformly distributed.
- A histogram of the stars' right ascensions or declinations (the azimuthal and elevation angles used in astronomy) looks a lot like a step function, i.e., flat within the image boundaries. The positions are not uniformly spaced, but they are distributed in the same way as a uniformly distributed random variable for any size postage stamp, including the entire image.
- The simplest example is the distribution of numbers that show up on the top of a fair die after a large number of throws. Each number from 1 to 6 will occur with approximately equal frequency. Increasing the number of throws will not tend to produce a bell-shaped histogram, in fact the fractional occurrence will approach a constant 1/6 over the possible numbers.

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

When there are way to many outliers, in those cases if we use mean. We will be way off as mean is drastically affected by outliers. Thus, in such cases it is preferable to use mean as the metric for central tendency than mean. . Another time when we usually prefer the median over the mean (or mode) is when our data is skewed

15. What is the Likelihood?

A likelihood function takes the data set as a given, and represents the likeliness of different parameters for your distribution. The Likelihood function gives us an idea of how well the data summarizes these parameters.