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) none 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

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Answer:--

1) Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data.

2) Boxplots may also depict values that are far outside of the normal range of responses (referred to as outliers). A histogram is a graphical representation of the spread of data points.
3) In the univariate case, box-plots do provide some information that the histogram does not (at least, not explicitly).

11. How to select metrics?

Answer:--

1. Good metrics are important to your company growth and objectives. Your key metrics should always be closely tied to your primary objective. ...
2. Good metrics can be improved. Good metrics measure progress, which means there needs to be room for improvement. ...
3. Good metrics inspire action.

KEY STEPS TO SELECTING EVALUATION METRICS

1. Classification. This algorithm will predict data type from defined data arrays. For example, it may respond with yes/no/not sure.
2. Regression. The algorithm will predict some values. For example, weather forecast for tomorrow.
3. Ranking. The model will predict an order of items.

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

Answer:---

Statistical significance can be accessed using hypothesis testing:
1)Stating a null hypothesis which is usually the opposite of what we

wish to test (classifiers A and B perform equivalently, Treatment A is equal of treatment B)
2)Then, we choose a suitable statistical test and statistics used to reject the null hypothesis
3) Also, we choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)
4)We calculate the observed test statistics from the data and check whether it lies in the critical region
Common tests:
– One sample Z test
– Two-sample Z test
– One sample t-test
– paired t-test
– Two sample pooled equal variances t-test
– Two sample unpooled unequal variances t-test and unequal sample sizes (Welch's t-test)
– Chi-squared test for variances
– Chi-squared test for goodness of fit
– Anova (for instance: are the two regression models equals? F-test)
– Regression F-test

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

Answer:-- Any type of categorical data won't have a gaussian distribution or lognormal distribution. Exponential distributions - eg. the amount of time that a car battery lasts or the amount of time until an earthquake occurs.

Many random variables have distributions that
are asymptotically Gaussian but may be significantly non-Gaussian for small numbers. For example the Poisson Distribution, which describes (among other things) the number of unlikely events occurring after providing a sufficient opportunity for a few events to

occur. It is pretty non-Gaussian unless the mean number of events is very large. The mathematical form of the distribution is still Poisson, but a histogram of the number of events after many trials with a large average number of events eventually looks fairly Gaussian.

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

Answer:-- 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.

Example-- Suppose, every month you spend INR 700 on personal care, INR 100 to pay the water bill, INR 800 on snacks, INR 500 to pay the electricity bill, and INR 6000 to pay the house rent. If you calculate the average expenditure, it comes out to be INR 1,620 by the notion of mean, and INR 700 by employing the median concept.

15. What is the Likelihood?

Answer:---

The likelihood function (often simply called the likelihood) represents the probability of random variable realizations conditional on particular values of the statistical parameters. Thus, when evaluated on a given sample, the likelihood function indicates which parameter values are more likely than others, in the sense that they would have made the observed data more probable. Consequently, the likelihood is often written as instead of to emphasize that it is to be understood as a function of the parameters instead of the random variable .