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

Ans : D

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

Ans : A

3. Which of the following function is associated with a continuous random variable?

a) pdf

b) pmv

c) pmf

d) all of the mentioned

Ans : A

4. The expected value or _____ of a random variable is the center of its distribution.

a) mode

b) median

c) mean

d) bayesian inference

Ans : B

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

Ans : C

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

Ans : A

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

Ans : C

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

Ans : A

9. Data that summarize all observations in a category are called _____ data.

a) frequency

b) summarized

c) raw

d) none of the mentioned

Ans : B

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

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

Ans: Both histograms and box plots are used to explore and present the data in an easy and understandable manner. 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.

11. How to select metrics?

- 1. Use standards. prefer metrics that have been tested by others;
- 2. Measure yourself the way your customer measures you
- 3. Only measure metrics that have an owner

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

Ans : Statistical significance can be accessed using hypothesis testing:
– 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)
– Then, we choose a suitable statistical test and statistics used to reject the null hypothesis
– 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)
– 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 (i.e: is at least one of the predictor useful in predicting the response?)

13. Give examples of data that does not 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.

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

Ans : When a distribution is skewed, the median does a better job of describing the center of the distribution than the mean. For example, consider the following distribution of salaries for residents in a certain city: The median does a better job of capturing the "typical" salary of a resident than the mean.

15. What is the Likelihood

Ans: Likelihood function. Due to the introduction of a probability structure on the parameter space or on the collection of models, it is a possible occurrence that a parameter value or a statistical model have a large likelihood value for a given specified observed data, and yet have a low probability, or vice versa.