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) 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
ANS: a) Discrete

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) pdf

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: c) mean

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) empirical mean

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) variance

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) 0 and 1

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: b) bootstrap

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) summarized

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: A histogram is a type of bar chart that graphically displays the frequencies of a data set. Similar to a bar chart, a histogram plots the frequency, or raw count, on the Y-axis (vertical) and the variable being measured on the X-axis (horizontal).

The only difference between a histogram and a bar chart is that a histogram displays frequencies for a group of data, rather than an individual data point; therefore, no spaces are present between the bars. Typically, a histogram groups data into small chunks (four to eight values per bar on the horizontal axis), unless the range of data is so great that it easier to identify general distribution trends with larger groupings.

A box plot, also called a box-and-whisker plot, is a chart that graphically represents the five most important descriptive values for a data set. These values include the minimum value, the first quartile, the median, the third quartile, and the maximum value. When graphing this five-number summary, only the horizontal axis displays values. Within the quadrant, a vertical line is placed above each of the summary numbers. A box is drawn around the middle three lines (first quartile, median, and third quartile) and two lines are drawn from the box's edges to the two endpoints (minimum and maximum).

11. How to select metrics?
ANS:

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: 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.

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

ANS: When there are a number of outliers that positively or negatively skew the data.

15. What is the Likelihood?

ANS: The likelihood returns the probability density of a random variable realization as a function of the associated distribution statistical parameter. For instance, 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 this observed data more probable as a realization.

THE END