Name of the Assignment: Statistics (Worksheet – 6)	
Submitted by	: Shalini Joshi
Designation	: Data Science Intern
Date of Submission	: 26th Feb,2023

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: Histograms and **box plots** are graphical representations for the frequency of numeric data values. 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.

11. How to select metrics?

Ans: In order to select **metrics**, we should firstly lay out how all parts of a given business process interact. This is called as **Process Mapping.** This will help us to know the metrics we need to improve the process.

- Regression metrics look for cause-and-effect relationships. They can be used to estimate the effect of one or more continuous variables on another variable.
- Comparison tests look for differences among group means.
- Correlation tests check whether variables are related without hypothesizing a cause-and-effect relationship.
- Non-parametric tests don't make as many assumptions about the data, and are useful when one or more of the common statistical assumptions are violated. However, the inferences they make aren't as strong as with parametric tests.

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

Ans: Statistical significance can be assessed using hypothesis testing:

- Stating a null hypothesis which is usually the opposite of what we wish to test
- Then, we chose a suitable statistical test and statistics used to reject the null hypothesis
- Also, we use 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*
- ✓ *Ch-squared test for variances*
- ✓ *Ch-squared test for goodness of fit*

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

Ans: Examples of data that does not have a Gaussian distribution, nor log-normal are as follows:

- 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: Likelihood refers to how well a sample provides support for particular values of a parameter in a model.
when we calculate likelihood we're trying to determine if we can trust the parameters in a model based on the sample data that we've observed.

********** **The End**********