**STATISTICS WORKSHEET- 6**

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

Q1. 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

ANS1. ‘D’

Q2. 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

ANS2. ‘A’-Continuous random variable can take any value on some subset of the real line.

Q3. Which of the following function is associated with a continuous random variable?

a) pdf

b) pmv

c) pmf

d) all of the mentioned

ANS3. ‘A’- Probability density function.

Q4. The expected value or \_\_\_\_\_\_\_ of a random variable is the center of its distribution.

a) mode

b) median

c) mean

d) bayesian inference

ANS4. ‘C’

Q5. 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

ANS5.’A’

Q6. The \_\_\_\_\_\_\_\_\_ of the Chi-squared distribution is twice the degrees of freedom.

a) variance

b) standard deviation

c) mode

d) none of the mentioned

ANS6. ‘A’

Q7. 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

ANS7.’C’

Q8. 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

ANS8.’B’

Q9. Data that summarize all observations in a category are called \_\_\_\_\_\_\_\_\_\_ data.

a) frequency

b) summarized

c) raw

d) none of the mentioned

ANS9. ‘B’

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

Q10. What is the difference between a boxplot and histogram?

Ans10. Histograms are favored for determining a data's underlying probability distribution. Box plots, on the other hand, are more helpful in comparing data sets. They are less descriptive and take up less room than histograms.

Q11. How to select metrics?

ANS11. It majorly depends on the type of algorithm you are using

For ex for classification we use

1. AUC-ROC
2. Precision recall
3. Accuracy score
4. Log loss

For Regression we might use

1. MSAE
2. Adjusted R square
3. R square

Q12. How do you assess the statistical significance of an insight?

Ans12. Hypothesis testing can be used to determine statistical significance:

Specifying a null hypothesis, which is normally the inverse of what we want to evaluate (classifiers A and B perform equally well, Treatment A is identical to Treatment B) – Then, we choose a suitable statistical test and statistics to refute the null hypothesis – we also select a specific area for the statistics to lie in that is extreme enough to reject the null hypothesis (p-value)

Q13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal.

ANS13. Poisson distribution

Binomial distribution

Exponential distribution

Q14. Give an example where the median is a better measure than the mean.

Ans14. Usually when we have ordinal data , then median is a better measure of central tendancy than mean

Assume you conduct a customer satisfaction survey with a group of 9 people and rate their overall satisfaction on a scale of 1 to 10. You get a 5.22 score. You're aware that, on average, you attract customers with a score of 3 or higher, so you're pleased because this means that you're still above where you want to be. But then you lose six of those nine customers all of a sudden. You return to your data and discover the following scores:1–3, 3–3, 3–5, 9–10–10–10–10–10–10–10–

Q15. What is the Likelihood

Ans15. The goodness of fit of a statistical model to a sample of data for given values of unknown parameters is measured by likelihood. It is derived from the sample's joint probability distribution, but is interpreted and used as a function of the parameters only, thereby treating the random variables as fixed at the observed values.