STATISTICS WORKSHEET-1

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

Q1 Bernoulli random variables take (only) the values 1 and 0. a) True b) False

Answer- A) True

Q2 Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases? a) Central Limit Theorem b) Central Mean Theorem c) Centroid Limit Theorem d) All of the mentioned

Answer- a)- Central Limit Theorem and c) Centroid Mean Theorem

Q3. Which of the following is incorrect with respect to use of Poisson distribution? a) Modeling event/time data b) Modeling bounded count data c)) Modeling contingency tables d) All of the mentioned

Answer – b) Modeling bounded count data

Q4. Point out the correct statement. a) The exponent of a normally distributed random variables follows what is called the log- normal distribution b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent c) The square of a standard normal random variable follows what is called chi-squared distribution d) All of the mentioned

Answer- d- All of the mentioned

5. _____ random variables are used to model rates. a) Empirical b) Binomial c) Poisson d) All of the mentioned

Answer- c) Poisson

Q6. 10. Usually replacing the standard error by its estimated value does change the CLT. a) True b) False

Answer- b) False

Q7. 1. Which of the following testing is concerned with making decisions using data? a) Probability b) Hypothesis c) Causal d) None of the mentioned

Answer- b) Hypothesis

Q8 Normalized data are centered at ____ and have units equal to standard deviations of the original data.

Answer a) 0

Q9. Which of the following statement is incorrect with respect to outliers? a) Outliers can have varying degrees of influence b) Outliers can be the result of spurious or real processes c) Outliers cannot conform to the regression relationship d) None of the mentioned

Answer- c) Outliers cannot conform to the regression relationship

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

Q10. What do you understand by the term Normal Distribution?

Answer - The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables. The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean taper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely. While the normal distribution is symmetrical, not all symmetrical distributions are normal.

Q 11. How do you handle missing data? What imputation techniques do you recommend?

Q 12. What is A/B testing?

Answer- A/B testing is one of the most popular controlled experiments used to optimize web marketing strategies. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternatives A and B. two alternative designs: A and B. Visitors of a website are randomly served with one of the two. Then, data about their activity is collected by web analytics. Given this data, one can apply statistical tests to determine whether one of the two designs has better efficacy. Now, different kinds of metrics can be used to measure a website efficacy. With

discrete metrics, also called binomial metrics, only the two values 0 and 1 are possible. The following are examples of popular discrete metrics.

• Click-through rate — if a user is shown an advertisement, do they click on it? • Conversion rate — if a user is shown an advertisement, do they convert into customers?

• Bounce rate — if a user is visits a website, is the following visited page on the same website? With continuous metrics, also called non-binomial metric, the metric may take continuous values that are not limited to a set two discrete states. The following are examples of popular continuous metrics.

• Average revenue per user — how much revenue does a user generate in a month?

• Average session duration — for how long does a user stay on a website in a session?

• Average order value — what is the total value of the order of a user?

Q13. Is mean imputation of missing data acceptable practice?

Answer- yes it is acceptable practice. mean imputation (also called mean substitution) really ought to be a last resort. mean imputation is the replacement of a missing observation with the mean of the non-missing observations for that variable. #1: Mean imputation does not preserve the relationships among variables. True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. by imputing the mean, you are able to keep your sample size up to the full sample size. That's good too. This is the original logic involved in mean imputation. Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

Q14. What is linear regression in statistics?

Answer-The regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income. t can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. Types of Linear Regression Simple linear regression 1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous) Multiple linear regression 1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous) Logistic regression 1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous) Ordinal regression 1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous) Multinomial regression 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous) Discriminant analysis 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio) When selecting the model for the analysis, an important consideration is model fitting. Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as R^2). However, overfitting can occur by adding too many variables to the model, which reduces model generalizability. Occam's razor describes the problem extremely well – a

simple model is usually preferable to a more complex model. Statistically, if a model includes a large number of variables, some of the variables will be statistically significant due to chance alone.

Q15. What are the various branches of statistics?

Answer- Two branches, **descriptive statistics and inferential statistics**, comprise the field of statistics.