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

Ans:-
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

Ans:-
a) Central Limit 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

Ans:-
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

Ans:-
d) All of the mentioned

Q5. _____ random variables are used to model rates.

a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

Ans:-
c) Poisson

Q6. Usually replacing the standard error by its estimated value does change the CLT.

a) True

b) False

Ans:-
b) False

Q7. Which of the following testing is concerned with making decisions using data?

a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

Ans:-
b) Hypothesis

Q8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

a) 0 b) 5 c) 1 d) 10

Ans:-
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

Ans:-

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?

Ans:-

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

In graphical form, the normal distribution appears as a "bell curve".

KEY TAKEAWAYS
The normal distribution is the proper term for a probability bell curve.
In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
Normal distributions are symmetrical, but not all symmetrical distributions are normal.
Many naturally-occurring phenomena tend to approximate the normal distribution.
In finance, most pricing distributions are not, however, perfectly normal.

Q11. How do you handle missing data? What imputation techniques do you recommend?

Ans:-

There are a lot of techniques to treat missing value. I am trying to think what is the best way to organize some of the most commonly used methods, if you use SAS to implement it -

Ignore the records with missing values.
Many tools ignore records with missing values. When the percentage of records with missing values is small, we could ignore those records.

Substitute a value such as mean.
When the percentage is large and also when it makes sense to do something to avoid bias modeling results, substituting a value (e.g. mean, median) is a commonly used way. But this

method could cause bias distribution and variance. That's where the following imputation methods come in.

Predict missing values.
Depending on the type of the imputed variable (i.e. continuous, ordinal, nominal) and missing data pattern (i.e. monotone, non-monotone), below are a few commonly used models. If you plan to do it in SAS, there are SAS codes that you can write to identify the missing data pattern.

Logistic Regression
Discriminant Regression
Markov Chain Monte Carlo (MCMC)
…
Predict missing values - Multiple Imputation. Although there are pros & cons, MI is considered to be superior to single imputation, and it better measures the uncertainty of the missing values. In addition, there are a few required statistical assumptions for multiple imputation:

Whether the data is missing at random (MAR).
Multivariate normal distribution, for some of the modeling methods mentioned above (e.g. regression, MCMC).
…
At last, if you have to think of what to report -

The type of imputation algorithm used.
Some justification for choosing a particular imputation method.
The proportion of missing observations.
The number of imputed datasets (m) created.
The variables used in the imputation model.

Q12. What is A/B testing?

Ans:-

A/B tests, also known as split tests, allow you to compare 2 versions of something to learn which is more effective. Simply put, do your users like version A or version B?

The concept is similar to the scientific method. If you want to find out what happens when you change one thing, you have to create a situation where only that one thing changes.

Think about the experiments you conducted in elementary school. If you put 2 seeds in 2 cups of dirt and put one in the closet and the other by the window, you'll see different results. This kind of experimental setup is A/B testing.

Creating 2 versions of a digital asset to see which one users respond to better. Examples of assets include a landing page, display ad, marketing email , and social post. In an A/B test, half

of your audience automatically receives “version A” and half receives “version B.” The performance of each version is based on conversion rate goals such as the percentage of people who click on a link, complete a form, or make a purchase. A/B testing isn’t a new idea with the advent of digital marketing. At one time, direct mail was the master of “splitting” or “bucketing” offers to see which one worked best. Digital capabilities build on the same idea but enable more specific, reliable, and faster test results.

Q13. Is mean imputation of missing data acceptable practice?

Ans:-

The process of replacing null values in a data collection with the data’s mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

Q14. What Is Linear Regression In Statistics?

Ans:-

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables. There are many names for a regression’s dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

Q15. What are the various branches of statistics?

Ans:-

The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.

While drawing conclusions, one needs to be very careful so as not to draw the wrong or biased conclusions. Even though this appears like a science, there are ways in which one can manipulate studies and results through various means. For example, data dredging is increasingly becoming a problem as computers hold loads of information and it is easy, either intentionally or unintentionally, to use the wrong inferential methods.

Both descriptive and inferential statistics go hand in hand and one cannot exist without the other. Good scientific methodology needs to be followed in both these steps of statistical analysis and both these branches of statistics are equally important for a researcher.