STATISTICS WORKSHEET-1

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

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

Answer=(A)True

2. 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

3. 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

4. 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

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

Answer=(B)False

7. 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

8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.
a) 0
b) 5
c) 1
d) 10

Answer=(A)0

9. 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

WORKSHEET
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-
1) Normal distribution in statistics is 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 graph form, normal distribution will appear as a bell curve.
2) A normal distribution is the proper term for a probability bell curve.
3) 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.In reality, most pricing distributions are not perfectly normal.
4)The normal distribution model is motivated by the Central Limit Theorem. This theory states that averages calculated
from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sampled (provided it has finite variance). Normal distribution is
sometimes confused with symmetrical distribution. Symmetrical distribution is one where a dividing line produces two mirror images, but the actual data could be two humps or a series of hills in addition to the bell curve that indicates
a normal distribution.
USED FOR-normal distribution, also called Gaussian distribution, the most common distribution function for independent,
randomly generated variables. Its familiar bell-shaped curve is ubiquitous in statistical reports, from survey analysis and quality control to resource allocation.

11. How do you handle missing data? What imputation techniques do you recommend?
Answer-
The concept of missing data is implied in the name: it's data that is not captured for a variable for the observationin question. Missing data reduces the statistical power of the analysis, which can distort the validity of the results.
The imputation method develops reasonable guesses for missing data. It's most useful when the percentage of missing data is low.
If the portion of missing data is too high, the results lack natural variation that could result in an effective model.

MEAN,MEDIAN,MODE~

This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a
small number of missing observations, data scientists can calculate the mean or median of the existing observations. However, when there
are many missing variables, mean or median results can result in a loss of variation in the data. This method does not use time-series
characteristics or depend on the relationship between the variables.
TIME-SERIES SPECIFIC METHOD~
Another option is to use time-series specific methods when appropriate to impute data. There are four types
of time-series data:
1)No trend or seasonality.
2)Trend, but no seasonality.
3)Seasonality, but no trend.
4)Both trend and seasonality.
The time series methods of imputation assume the adjacent observations will be like the missing data. These methods work well when that assumption is valid. However, these methods won't always produce reasonable results, particularly in the case of strong seasonality.
LINEAR INTERPOLATION~
Linear interpolation is often used to approximate a value of some function by using two known values of that function at other points.
This formula can also be understood as a weighted average. The weights are inversely related to the distance from the end points to the unknown point. The closer point has more influence than the farther point.
SEASONAL ADJUSTENT WITH LINEAR INTERPOLATION~
When dealing with missing data, you should use this method in a time series that exhibits a trend line, but it's not appropriate for seasonal data.

12. What is A/B testing?
Answer-
A/B testing, also known as split testing, is a way to compare different versions of something to see which performs better.In these experiments,you define a conversion goal to measure, like clicks or completed transactions. Two variations of the same marketing asset (like a web page or email) are then shown to different users at random while measuring the difference in performance.

13. Is mean imputation of missing data acceptable practice?
Answer-
1)The process of replacing null values in a data collection with the data's mean is known as mean imputation.
2)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.
3)Bad practice in general
4)If just estimating means: mean imputation preserves the mean of the observed data
5)Leads to an underestimate of the standard deviation
6)Distorts relationships between variables by "pulling" estimates of the correlation toward zeroSecond, 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.

14. What is linear regression in statistics?
Answer-
Linear regression in statistics 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.
Three major uses for regression analysis are
(1) determining the strength of predictors,
(2) forecasting an effect, and
(3) trend forecasting
~First, 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.
~Second, it 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. A typical question is, "how much additional sales income do I get for each additional $1000 spent on marketing?"
~Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, "what will the price
of gold be in 6 months?"
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)

15. What are the various branches of statistics?
Answer-
BRANCH OF STATISTICS-
There are two main branches of statistics
1.descriptive statistics and 2
.inferential statistics.
Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.
1.DESCRIPTIVE STATISTIC~
It 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.
2.INFERENTIAL STATISTICS~

It 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.