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
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
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
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]
5. ______ random variables are used to model rates.
   a) Empirical
   b) Binomial
   [✓ c) Poisson]
   d) All of the mentioned
6. 10. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   [✓ b) False]
7. 1. Which of the following testing is concerned with making decisions using data?
   a) Probability
   [✓ b) Hypothesis]
   c) Causal
   d) None of the mentioned
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
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

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- The normal distribution, also known as the Gaussian or standard normal distribution, is the probability distribution that plots all of its values in a symmetrical fashion, and most of the results are situated around the probability's mean. Values are equally likely to plot either above or below the mean. The normal distribution is a probability distribution that (roughly) describes many common datasets in the real world. It is the most common type of distribution, and it arises naturally in statistics through random sampling techniques.
Nowadays, it is more common to show up as a model for the "lifespan" of a product, like a light bulb, or the outcome of standardized tests, like IQ. Biological measurements, like height or weight, are often estimated with normal distributions.

Q11 How do you handle missing data? What imputation techniques do you recommend?
Ans-
A)Best techniques to handle missing data-
1)Use deletion methods to eliminate missing data-
The deletion methods only work for certain datasets where participants have missing fields. There are several deleting methods – two common ones include List wise Deletion and Pair wise Deletion. It means deleting any participants or data entries with missing values. This method is particularly advantageous to samples where there is a large volume of data because values can be deleted without significantly distorting readings. Alternatively, data scientists can fill out the missing values by contacting the participants in question. The problem with this method is that it may not be practical for large datasets. Furthermore, some corporations obtain their information from third-party sources, which only makes it unlikely that organizations can fill out the gaps manually. Pair wise deletion is the process of eliminating information when a particular data point, vital for testing, is missing. Pair wise deletion saves more data compared to likewise deletion because the former only deletes entries where variables were necessary for testing, while the latter deletes entire entries if any data is missing, regardless of its importance.

2)Use regression analysis to systematically eliminate data-
Regression is useful for handling missing data because it can be used to predict the null value using other information from the dataset. There are several methods of regression analysis, like Stochastic regression. Regression methods can be successful in finding the missing data, but this largely depends on how well connected the remaining data is. Of course, the one drawback with regression analysis is that it requires significant computing power, which could be a problem if data scientists are dealing with a large dataset.

B)data imputation techniques-
Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. However, a word of caution when using this method – it can artificially reduce the variability of the dataset. Common-point imputation, on the other hand, is when the data scientists utilize the middle point or the most commonly chosen value. For example, on a five-point scale, the substitute value will be 3. Something to keep in mind when utilizing this method is the three types of middle values: mean, median and mode, which is valid for numerical data (it should be noted that for non-numerical data only the median and mean are relevant).

Q12 What is A/B testing?
Ans- A/B testing is basically *statistical hypothesis testing*, or, in other words, statistical inference. It is an analytical method for making decisions that estimates population parameters based on sample statistics.
The population refers to all the visitors coming to your website (or specific group of pages), while the sample refers to the number of visitors that participated in the test.

Let’s say, you make a decision to implement some change on your product pages based on A/B test results that tested a “sample” of the visitors to your website. Ultimately, only a percentage of the visitors saw the challenger, so that of course means not all the visitors. However, with A/B testing, you assume if the challenger (i.e. variation) in the test increased conversions for a group of visitors on product pages, it will thus have the same result for all the 1 visitors of your product pages (we will delve into the accuracy of a variation’s validity later).
To recap, the A/B testing process can be simplified as follows:

1. You start the A/B testing process by making a claim (hypothesis).
2. You launch your test to gather statistical evidence to accept or reject a claim (hypothesis) about your website visitors.
3. The final data shows you whether your hypothesis was correct, incorrect or inconclusive.

Q13 Is mean imputation of missing data acceptable practice?
Ans- 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. That’s a good thing.
Plus, 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.
If all you are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estimate. It *will* still bias your standard error
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?
Ans- Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.
Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable *causes* the other (for example, higher SAT scores do not *cause* higher college grades), but that there is some significant association between the two variables. A scatter plot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatter plot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.
A linear regression line has an equation of the form $Y = a + bX$,
where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when $x = 0$).

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.
1)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.
2)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.