Statistics **worksheet_1 Answerkey**

1.Bernoulli random variables take (only) the values 1 and 0.

Ans: 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?

Ans: a) Central Limit Theorem

3.Which of the following is incorrect with respect to use of Poisson distribution?

Ans: b) Modeling bounded count data

4. Point out the correct statement.
Ans: d) All of the mentioned

5._____ random variables are used to model rates.

Ans: c) Poisson

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

Ans: b) False

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

b) Hypothesis

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

Ans: d) 10

9.Which of the following statement is incorrect with respect to outliers?

Ans: c) Outliers cannot conform to the regression relationship

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

Ans: A normal distribution is an arrangement of a data set in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme. Height is one simple example of something that follows a normal distribution pattern: Most people are of average height the numbers of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.Here's an example of a normal distribution curve:

Formula: $y=\frac{1}{2\pi\sigma}e^{-\frac{(x-\mu)^2}{2\sigma}}$

Where −

- μ = Mean
- σ = Standard Deviation
- $\pi \approx 3.14159$
- $e \approx 2.71828$

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

Ans: ***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 Listwise Deletion and Pairwise 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 organisations can fill out the gaps manually. Pairwise deletion is the process of eliminating information when a particular data point, vital for testing, is missing. Pairwise 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.

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.

12.What is A/B testing?

Ans: t its most basic, 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.

For example, let's say you wanted to increase the number of clicks on the "Buy now" button on your product pages. You could run an A/B test to find out how button color affects click-through rates, experimenting with a green button and a blue button. You would:

Define your conversion goal. In this example, you want to measure click-through rate.

Design the A/B test. How big of a sample size do you want? Who will participate, new customers or existing customers?

Gather data. Will you run your own test or use A/B testing software? For how long will the test run?

Analyze the results. Which variation had the biggest positive impact on the conversion metric that matters most?

At the end of the A/B test, you'll be able to confidently implement the winning variation without worrying about jeopardizing conversions.

13.Is mean imputation of missing data acceptable practice?

Ans: Mean Imputation: 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.

14. What is linear regression in statistics?

Ans: Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables

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)

15. What are the various branches of statistics?

Ans: There are three real branches of statistics: descriptive statistics and inferential 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

IInferential 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.