Statistics worksheet-1

1.a
2.a
3.b
4.c
5.c
6.b
7.b
8.a
9.c

10.Normal Distribution in Statistics

The normal distribution, also known as the Gaussian distribution, is the most important probability distribution in statistics for independent, random variables.
its is bell-shaped curve in statistical reports.

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.

11.When dealing with missing data we use two primary methods to solve the error: imputation or the removal of data.
Imputation techniques

- Mean imputation
- Substitution
- Hot deck imputation
- Cold deck imputation
- Regression imputation
- Stochastic regression imputation
- Interpolation and extrapolation

12.A/B testing

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.
A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from who used A and B respectively, you try to decide which is performing better.

1. Make a Hypothesis

In hypothesis testing, we have to make two hypotheses i.e Null hypothesis and the alternative hypothesis. Let's have a look at both.

- Null hypothesis or H0:

The null hypothesis is the one that states that sample observations result purely from chance. From an A/B test perspective, the null hypothesis states that there is no difference between the control and variant groups.

- Alternative Hypothesis or H0:

The alternative hypothesis challenges the null hypothesis and is basically a hypothesis that the researcher believes to be true. The alternative hypothesis is what you might hope that your A/B test will prove to be true.

2. Create Control Group and Test Group

The Control Group is the one that will receive unchanged variable and the Test Group is the one that will receive changed variable

3. Conduct the A/B Test and Collect the Data

One way to perform the test is to calculate daily conversion rates for both the treatment and the control groups.

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

1. Mean imputation is typically considered terrible practice since it ignores feature correlation.
2. 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.Linear Regression

Linear regression models the relationships between at least one explanatory variable and an outcome variable. These variables are known as the independent and dependent variables, respectively. When there is one independent variable (IV), the procedure is known as simple linear regression. When there are more IVs, statisticians refer to it as multiple regression.

Linear regression has two primary purposes—understanding the relationships between variables and forecasting.

- The coefficients represent the estimated magnitude and direction (positive/negative) of the relationship between each independent variable and the dependent variable.

- A linear regression equation allows you to predict the mean value of the dependent variable given values of the independent variables that you specify.

Equation: $y= a0+a1x+ \varepsilon$

Here,

Y= Dependent Variable
X= Independent Variable
a0= intercept of the line
a1 = Linear regression coefficient
ε = random error

15.There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.

- Data collection is all about how the actual data is collected.
- Descriptive statistics is the part of statistics that deals with Summarizing ,describing and presenting the data.
- Inferential statistics is the aspect that deals with making conclusions and decisions using estimation and hypothesis testing about the data.