STATISTICS WORKSHEET

 # 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.

Ans: a) True

Q2. Which of the following theorems 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

Q3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans: b) Modeling bounded count data

Q4. Point out the correct statement.

Ans: d) All of the mentioned

Q5._____Random variables are used to model rates.

Ans: c) Poisson

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

Ans: b) False

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

Ans: b) Hypothesis

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

Ans: a) 0

Q9. Which of the following statements is incorrect with respect to outliers?

Ans: c) Outliers cannot conform to the regression relationship

 # Q10and Q15 are subjective answer type questions. Answer them in your own words briefly.

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

Ans:-Normal Distribution is a probability function used in statistics that tells about how the data values are distributed. It is also known as Gaussian distribution, is a continuous probability distribution that is commonly used to model real-world variables that tend to cluster around a mean value, with the majority of the observations falling close to the mean, and fewer observations further away from the mean.
In a normal distribution, the probability density function is symmetric and bell-shaped, with the mean, median, and mode all equal to each other. The spread of the distribution is controlled by the standard deviation, which determines the width of the curve. Many natural phenomena, such as heights, weights, and IQ scores, follow a normal distribution, and it is widely used in statistics, hypothesis testing, and inferential analysis.
In normally distributed data, there is a constant proportion of data points lying under the curve between the mean and a specific number of standard deviations from the mean. Thus, for a normal distribution, almost all values lie within 3 standard deviations of the mean.

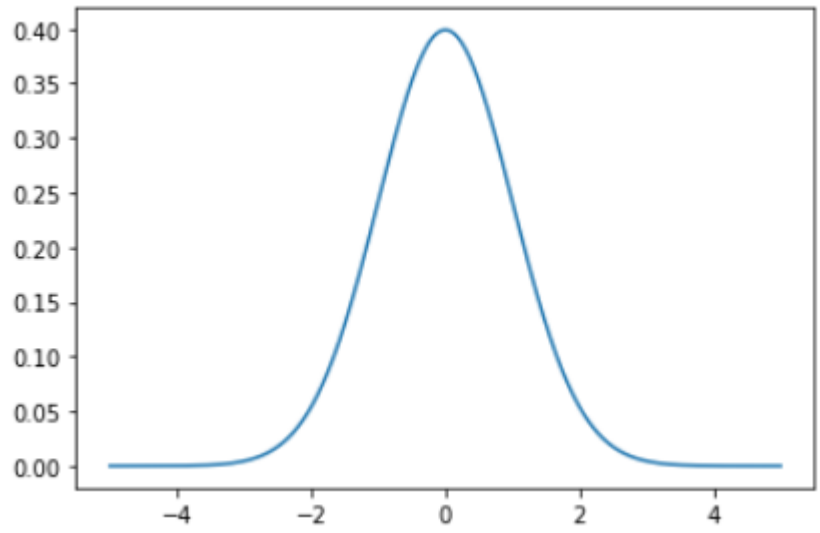
```
import numpy as np
import scipy as sp
from scipy import stats
import matplotlib.pyplot as plt

## generate the data and plot it for an ideal normal curve

## x-axis for the plot
x_data = np.arange(-5, 5, 0.001)

## y-axis as the gaussian, mean= 0 and SD = 1
y_data = stats.norm.pdf(x_axis, 0, 1)

## plot data
plt.plot(x_data, y_data)
plt.show()
```



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

Ans:- Handling missing data is an important task in data analysis as missing data can negatively affect the accuracy and effectiveness of the analysis.In the dataset, the blank shows the missing values.Missing Data can also refer to as NAN(Not a number) values in pandas. In DataFrame sometimes many datasets simply arrive with missing data, either because it exists and was not collected or it never existed. There are various techniques that can be used to handle missing data, some of which are:

1. **List wise Deletion**: In this method, the entire observation with missing data is removed from the dataset. This method is easy to apply but can result in a loss of information.
2. Pair wise Deletion: In this method, only the missing values are removed from the analysis, and the rest of the data is used. This method is less biased than the listwise deletion but can result in an increased variability.
3. Mean/Median/Mode Imputation: In this method, the missing values are replaced by the mean, median or mode value of the non-missing values in the same column. This method is simple to apply but can result in biased estimates.
4. Regression Imputation: In this method, a regression model is used to predict the missing values based on the non-missing values in the same column. This method can provide more accurate estimates than mean/median/mode imputation.
5. Multiple Imputation: In this method, missing values are imputed multiple times using regression imputation, and the results are combined to provide a more accurate estimate.

The imputation technique replaces missing values with substituted values. The missing values can be imputed in many ways depending upon the nature of the data and its problem.

Imputation with constant value:As the title hints — it replaces the missing values with either zero or any constant value. Most Frequent Value- The most frequent value in the column is used to replace the missing values in another popular technique that is effective for both nominal and numerical features.

Imputation with Arbitrary value: If you can replace the missing value with some arbitrary value using fillna(). **Replacing with previous value – Forward fill-** We can impute the values with the previous value by using forward fill. It is mostly used in time series data. Syntax: df.fillna(method='ffill').Replacing with next value – **Backward fill**- the missing value is imputed using the next value. It is mostly used in time series data.

Imputation using Statistics: Mean or Moving Average or Median Value Median, Mean, or rounded mean are further popular imputation techniques for numerical features. The technique, in this instance, replaces the null values with mean, rounded mean, or median values determined for that feature across the whole dataset. It is advised to utilize the median rather than the mean when your dataset has a significant number of outliers.

Advanced Imputation Technique: Advanced imputation techniques adopt machine learning algorithms to impute the missing values in a dataset. Eg.K Nearest Neighbors - The objective is to find the k nearest examples in the data where the value in the relevant feature is not absent and then substitute the value of the feature that occurs most frequently in the group

The choice of imputation technique depends on the nature and extent of the missing data, and the specific requirements of the analysis. No single imputation technique is universally suitable for all situations. However, multiple imputation is generally recommended as it provides the most accurate estimates and accounts for the uncertainty in the imputed values.

12. What is A/B testing?

Ans: 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/B testing works best when testing incremental changes, such as UX changes, new features, ranking, and page load times. Note, all other variables need to be held constant when performing an A/B test.

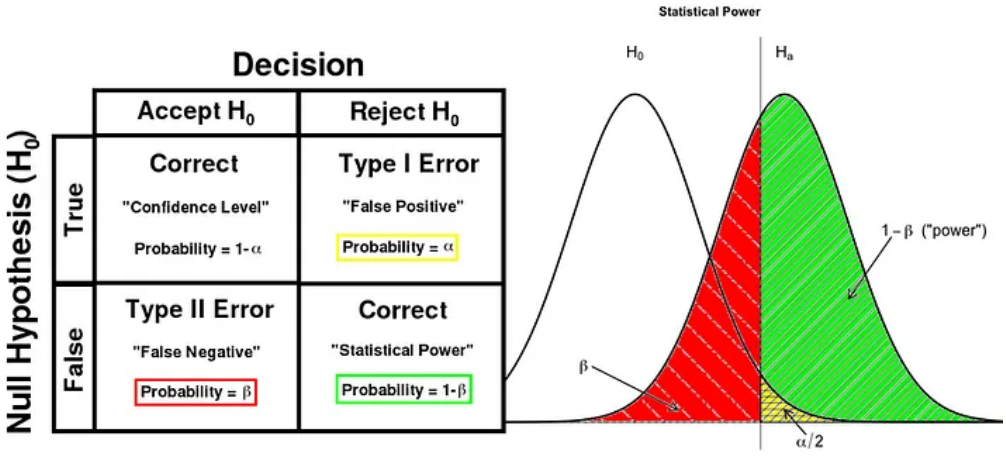
1. Formulate your hypothesis - The null hypothesis is one that states that sample observations result purely from chance. From an A/B test perspective, the null hypothesis states that there is no difference between the control and variant group.

The alternative hypothesis is one that states that sample observations are influenced by some nonrandom cause. From an A/B test perspective, the alternative hypothesis states that there is a difference between the control and variant group.

2.Create your control group and test group : Random sampling - to eliminate bias because you want the results of your A/B test to be representative of the entire population rather than the sample itself. Sample Size- To determine the minimum sample size for your A/B test prior to conducting it so that you can eliminate under coverage bias.

3. Conduct the test, compare the results, and reject or do not reject the null hypothesis :
1. **Significance level (alpha):** The significance level, also denoted as alpha or α, is the probability of rejecting the null hypothesis when it is true. Generally, we use the significance value of 0.05
2**. P-Value:** It is the probability that the difference between the two values is just because of random chance. P-value is evidence against the null hypothesis. The smaller the p-value, the stronger the chances to reject the H0. For the significance level of 0.05, if the p-value is lesser than it hence we can reject the null hypothesis
3. **Confidence interval:** The confidence interval is an observed range in which a given percentage of test outcomes fall. We manually select our desired confidence level at the beginning of our test. Generally, we take a 95% confidence interval

One of the most used hypothesis tests is the two-sample t-test. It is applied to compare the average difference between the two groups.
Now let's say that at the end of the test, we got a p-value of 0.003, which is very less than the general significance level of 0.05. In this case, we do not have enough evidence for the Null hypothesis. And thus, we will reject the Null Hypothesis.
On the other hand, if the p-value were higher than the significance level of 0.05, then we would have failed to reject the null hypothesis.



13. Is mean imputation of missing data acceptable practice?

Ans:- Although imputing missing values by using the mean is a popular imputation technique, there are serious problems with mean imputation. The variance of a mean-imputed variable is always biased downward from the variance of the un-imputed variable. This bias affects standard errors, confidence intervals, and other inferential statistics.

Experts agree that mean imputation should be avoided when possible
1**. Mean substitution leads to bias in multivariate estimates** such as correlation or regression coefficients. Values that are imputed by a variable's mean have, in general, a correlation of zero with other variables. Relationships between variables are therefore biased toward zero.

2. **Standard errors and variance of imputed variables are biased.**- Mean imputation can lead to biased estimates of the mean and standard deviation of the variable. This is because mean imputation reduces the variance of the variable by artificially inflating the sample size. As a result, the standard errors of the estimates are underestimated, and the significance levels of the statistical tests are overestimated.
For instance, let's assume that we would like to calculate the standard error of a mean estimation of an imputed variable. Since all imputed values are exactly the mean of our variable, we would be too sure about the correctness of our mean estimate. In other words, the confidence interval around the point estimation of our mean would be too narrow.

3. One major limitation of mean imputation is that it assumes that the missing values are missing completely at random (MCAR), which means that the probability of a value being missing is unrelated to its actual value. In many cases, this assumption may not hold, and missing values may be related to other variables in the dataset.

Mean imputation is typically considered terrible practice since it ignores feature correlation. It preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased

In summary, mean imputation can be a quick and easy solution for handling missing data, but it should be used with caution and only when the assumption of MCAR is justifiable. In cases where the assumption is not met , more sophisticated imputation techniques, such as multi[le imputation or maximum likelihood estimation, should be used.

14. What is linear regression in statistics?

Ans:- Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. The goal of linear regression is to estimate the parameters of the linear equation, such as the slope and intercept, which can then be used to predict the value of the dependent variable based on the values of the independent variables.

Linear regression models have many real-world applications in an array of industries such as economics (e.g. predicting growth), business (e.g. predicting product sales, employee performance), social science (e.g. predicting political leanings from gender or race), healthcare (e.g. predicting blood pressure levels

from weight, disease onset from biological factors), and more.

A linear regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.

It does this by essentially fitting a best-fit line and seeing how the data is dispersed around this line.

In order for regression results to be properly interpreted, several assumptions about the data and the model itself must hold.

Linear Regression Equation

The measure of the relationship between two variables is shown by the correlation coefficient. The range of the coefficient lies between -1 to +1. This coefficient shows the strength of the association of the observed data between two variables. Linear Regression Equation is given below: **Y=a+bX**

where X is the independent variable and it is plotted along the x-axis

Y is the dependent variable and it is plotted along the y-axis

Here, the slope of the line is b, and a is the intercept (the value of y when x = 0).

Key Ideas of Linear Regression

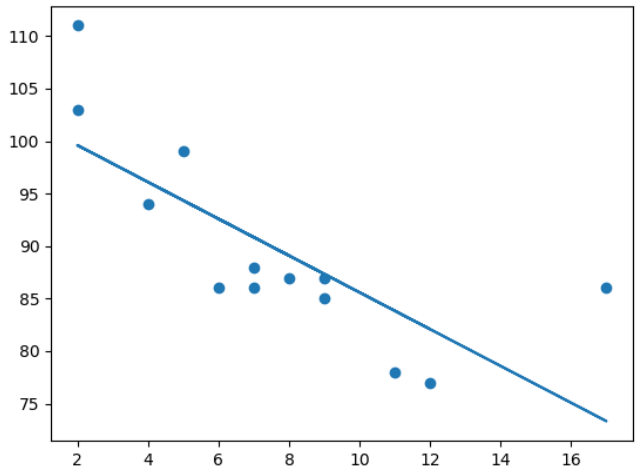
Correlation explains the interrelation between variables within the data.

Variance is the degree of the spread of the data.

Standard deviation is the dispersion of mean from a data set by studying the variance's square root.

Residual (error term) is the actual value found within the dataset minus the expected value that is predicted in linear regression.

```
import matplotlib.pyplot as plt
from scipy import stats
x = [5,7,8,7,2,17,2,9,4,11,12,9,6]
y = [99,86,87,88,111,86,103,87,94,78,77,85,86]
slope, intercept, r, p, std_err = stats.linregress(x, y)
def myfunc(x): return slope * x + intercept
mymodel = list(map(myfunc, x))
plt.scatter(x, y)
plt.plot(x, mymodel)
plt.show()
```



15.What are the various branches of statistics?

Ans: Statistics have majorly categorised into two types: 1. Descriptive statistics 2. Inferential statistics

Descriptive Statistics: Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.Descriptive statistics are also categorised into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

The frequency measurement displays the number of times a particular data occurs. Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data. Central tendencies are the mean, median and mode of the data. And the measure of position describes the percentile and quartile ranks.

Inferential Statistics This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population. It grants us permission to give statements that go beyond the available data or information.

Apart from these two main branches, there are various other sub-branches of statistics, such as:

- Biostatistics: This branch applies statistical methods to analyze biological and medical data.
- Econometrics: This branch applies statistical methods to analyze economic data.
- Social statistics: This branch deals with the analysis of social phenomena and data.

- Business statistics: This branch deals with the application of statistical methods in the field of business and commerce.
- Environmental statistics: This branch deals with the analysis of environmental data.
- Psychometrics: This branch applies statistical methods to analyze psychological data.
- Educational statistics: This branch deals with the application of statistical methods in the field of education.
- Statistical genetics: This branch deals with the analysis of genetic data.
- Quality control: This branch deals with the application of statistical methods to improve the quality of products and services.
- Time series analysis: This branch deals with the analysis of time-based data, such as
- stock prices, weather data, etc.