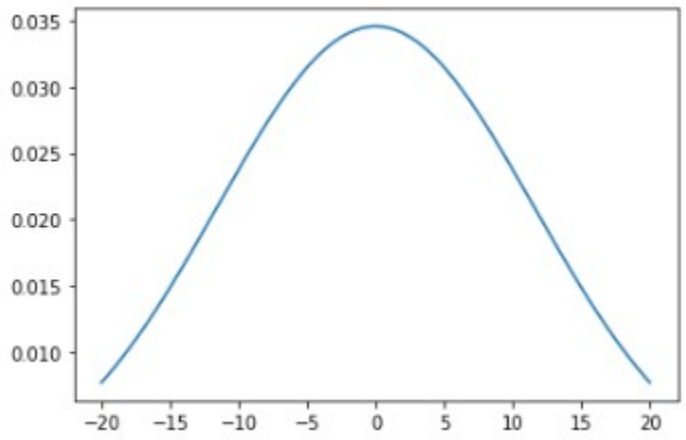
ASSIGNMENT-2

STATISTICS

1. a).True
2. A). Central Limit Theorem
3. B). Modeling bounded count data
4. D). All of the mentioned
5. C). Poisson Distribution
6. B). False
7. B). Hypothesis Testing
8. A). ZERO
9. C). Outliers cannot conform to the regression relationship

10). Normal Distribution is a probability function used in statistics that tells about how the data values are distributed. It is generally observed that data distribution is normal when there is a random collection of data from independent sources. The graph produced is bell-shaped and signifies that the peak point is the mean of the data set and half of the values of data set lie on the left side of the mean and other half lies on the right part of the mean telling about the distribution of the values. The graph is symmetric distribution. For example, the height of the population, shoe size, IQ level, rolling a die, and many more

```
import numpy as np

import matplotlib.pyplot as plt

from scipy.stats import norm

import statistics

# Plot between -10 and 10 with .001 steps.

x_axis = np.arange(-20, 20, 0.01)

# Calculating mean and standard deviation

mean = statistics.mean(x_axis)

sd = statistics.stdev(x_axis)

plt.plot(x_axis, norm.pdf(x_axis, mean, sd))

plt.show()
```



11). Missing data appear when no value is available in one or more variables of an individual**.** Due to Missing data, the statistical power of the analysis can reduce, which can impact the validity of the results.The major reason behind the missing data is item non response.

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 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.

12). 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 non-random 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 stronger the chances to reject the H_0. 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

Next, we can calculate our t statistics using the below formula:

$$T - statistic = \frac{Observed\ value - hypothesized\ value}{Standard\ Error}$$

$$Stamdard\ Error = \sqrt{\frac{2 * Variance(sample)}{N}}$$

13). 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.

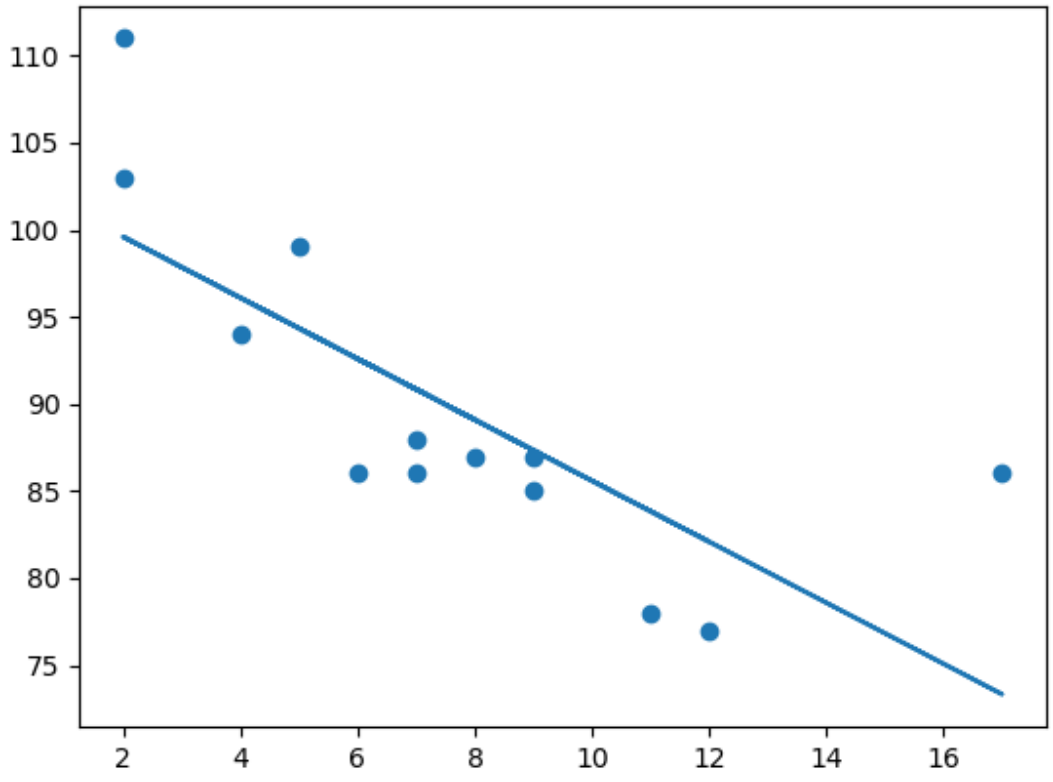
2. **Standard errors and variance** of imputed variables are biased. For instance, let's assume that we would like to calculate the standard error of a mean estimation of an imputed variable. Since all imputed values are exactly the mean of our variable, we would be too sure about the correctness of our mean estimate. In other words, the confidence interval around the point estimation of our mean would be too narrow.

3. If the response mechanism is MAR or MNAR, even the **sample mean of your variable is biased** (compare that with point 3 above). Assume that you want to estimate the mean of a population's income and people with high income are less likely to respond; Your estimate of the mean income would be biased downwards.

14). 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. 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.

```
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). 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 goes beyond the available data or information. For example, deriving estimates from hypothetical research.

Statistics Example

In a class, the collection of marks obtained by 50 students is the description of data. Now when we take out the mean of the data, the result is the average of marks of 50 students. If the average mark obtained by 50 students is 88 out of 100, then we can reach to a conclusion or give a judgment on the basis of the result.