Theoretical Que : Que 1

**Theoretical Questions**

Que 1. Explain the difference between descriptive and inferential statistics. Provide examples of each.

Ans : There are 2 types of Statistics based on their purposes in Data Science :

1. **Descriptive**
2. **Inferential**

Difference between descriptive and inferential statistics :

|  | **Descriptive** | **Inferential** |
| --- | --- | --- |
| **Purpose** | summarize data | make predictions or inferences about a population. |
| **Application** | is used to describe what the data shows | is used to make decisions or predictions about a broader group |
| **Data used** | deals with the entire dataset | typically works with a sample and draws conclusions about the population |

1. **Descriptive Statistics** : It is mainly used for Summarizing, Analysing and Visualising. These statistics provide a clear picture of the data set's basic characteristics, such as central tendency, variation, and distribution.

E.g

1. Measures of central tendency :

a. Mean : its an average of dataset values

*E.g. The average test score of 30 students is 75%*

b. Meridian : its a middle value in sorted dataset

*E.g. If the salaries of 5 employees are 30k, 40k, 50k, 60k, and 100k, the median salary is 50k (the middle value when arranged in order).*

c. Mode : it's a most frequent value in dataset

E.g. *If 100 people were surveyed about their favorite fruit, and 45 chose "apple," the mode of the responses is "apple."*

1. Measures of variability :

a. Range : its a difference between highest and lowest value

*E.g. the range of ages in a group.*

b.Variance : The average of the squared differences from the Mean

*E.g the variance in test scores.*

c.Standard Deviation : its a measure of how spread out the values are in a dataset. The square root of the variance.

*E.g. the standard deviation of heights in a population.*

1. Graphical Representation :

a.Histograms : A bar representation of data distribution

*e.g.A histogram showing the distribution of ages in a population.*

b.Pie charts : A circular graphic representation divided into slices to illustrate numerical proportion

*e.g.A pie chart showing the percentage of different genres in a music library*

**2. Inferential Statistics :** This mainly involves predictions, generalization, inferences about the population based on a sample of data. The goal is to use sample data to estimate parameters, test hypotheses, or make predictions about a larger group.

E.g.

1. Hypothesis testing : Null hypothesis and Alternate Hypothesis

a. t-test : use to compare the means of 2 groups

*E.g. testing if there is a significant difference in test scores between two different teaching methods.*

b. z-test :

E.g.

c. chi square : use to determine if there is a significant association between categorical variables

E.g*. Testing if there is an association between gender and preference for a new product.*

d. anova test(Analysis of variance) : use to compare the means of more 3 and than 3 groups

*E.g. testing if there are significant differences in productivity among different departments in a company.*

1. Confidence Interval : Estimating the range within which a population parameter lies with a certain level of confidence.

*E.g A 95% confidence interval for the average height of a population*

1. Regression Analysis : Examining the relationship between variables.

*E.g. Using linear regression to predict a student's GPA based on their SAT scores.*

Que 2

Que 2 : Define the Central Limit Theorem and discuss its significance in statistical inference.

Ans :

The Central Limit Theorem(CLT) in Statistics states that as the sample size increases and its variance is finite, then the distribution of the sample mean approaches normal distribution irrespective of the shape of the population distribution.

i.e.regardless of the shape of the population distribution, the sampling distribution of the sample mean will tend to be normal (or approximately normal) if the sample size is sufficiently large. This holds true as long as the data are independent and identically distributed (i.i.d.).

Significance of CLT :

1. **Enables Normal Approximation:** One of the major implications of CLT is that it allows statisticians to make inferences about the population, even when the underlying distribution is unknown. By knowing that the sampling distribution of the sample mean will be approximately normal (for large enough sample sizes), we can apply normal probability techniques (e.g., z-scores, confidence intervals) to estimate population parameters.
2. **Hypothesis Testing:** The CLT is fundamental for hypothesis testing. Many statistical tests, such as t-tests and z-tests, rely on the assumption that sample means follow a normal distribution. Without CLT, we would need to know the exact distribution of the population, which is often impractical.
3. **Confidence Intervals:** CLT allows us to construct confidence intervals for population parameters (like the mean) based on the sample mean. This is because the sampling distribution of the sample mean is normal, so we can calculate probabilities and intervals around the sample mean using properties of the normal distribution.
4. **Generalization :** CLT helps in generalizing sample results to the broader population. As long as the sample size is large enough, the sample mean will provide an accurate estimate of the population mean, regardless of the original population's distribution.

Que 3

Que 3 : Discuss the concept of sampling and its role in statistical analysis.

Ans :

**Sampling** is the process of selecting a subset of individuals, items, or observations from a larger population in order to make inferences about the entire population. Instead of studying the entire population (which can be time-consuming, expensive, or even impossible), researchers can work with a sample that accurately reflects the population, allowing them to make valid conclusions and predictions about the broader group.

Key Terms used in Sampling :

| Population | The entire group of interest or the complete set of data you're trying to understand (e.g., all voters in a country, all students in a school). |
| --- | --- |
| Sample | A subset of the population that is selected for the actual study |
| Sampling Frame | A list or representation of all the elements in the population from which the sample is drawn. It's important for the sample to be selected from a well-defined sampling frame to ensure that the sample is representative. |
| Sampling Bias | This occurs when the sample selected does not accurately represent the population. Bias can arise from improper sampling methods |
| Sampling Method | The procedure used to select the sample from the population. |

Importance of Sampling in Statistical Analysis :

1. Cost and Time Efficiency : Examining an entire population can be time-consuming and expensive. Sampling allows researchers to obtain meaningful results more quickly and at a lower cost.
2. Feasibility : In many cases, it is impractical or impossible to study an entire population. For example, testing every light bulb produced in a factory for quality control would be impractical. Sampling makes such studies feasible.
3. Accuracy and Precision : Proper sampling methods can yield highly accurate and precise estimates of population parameters. With a well-designed sample, researchers can make reliable inferences about the population.
4. Reduction of Data Volume : Sampling reduces the volume of data that needs to be collected and analyzed, making data management and analysis more manageable.

Types of Sampling Methods :

1. Probability Sampling Methods: i.e. where every member of the population has a known chance of being selected.

These methods are more reliable because they give each member of the population a known chance of being included in the sample.

1. **Simple Random Sampling** : Every individual in the population has an equal chance of being selected.

*e.g.: Drawing names from a hat.*

1. **Stratified Sampling** : The population is divided into subgroups (strata), and random samples are taken from each subgroup. This method ensures that certain characteristics (e.g., age, gender) are proportionally represented.

*e.g. : Dividing a population of students into grade levels and randomly selecting students from each grade.*

1. **Systematic Sampling** : Every nnn-th individual is selected from a list or population. The starting point is randomly chosen.

*e.g. : If you have a list of 1,000 people, you might select every 10th person starting from a randomly chosen position.*

1. **Cluster Sampling :** The population is divided into clusters (e.g., geographic areas, schools), and entire clusters are randomly selected. This is useful when the population is spread out geographically.

*e.g. : Selecting several schools within a district and surveying all students within those schools.*

1. Non-Probability Sampling Methods: i.e. where the chance of selection is not known, which can lead to bias.
2. **Convenience Sampling :** Selecting individuals who are easiest to reach or obtain.

*e.g. : Surveying people at a shopping mall because they are readily accessible.*

1. **Judgmental (Purposive) Sampling** : The researcher selects individuals based on their knowledge or judgment about who will provide the best information for the study.

*e.g. : Interviewing experts in a particular field.*

1. **Quota Sampling** : A non-random sampling method in which the researcher selects people based on a predetermined characteristic (e.g., selecting 100 people with equal numbers of men and women).

*e.g. : Ensuring a set number of male and female participants in a study.*

Role of Sampling in Statistical Analysis :

1. Representativeness : Proper sampling ensures that the sample accurately represents the population, making it possible to generalize findings. A good sample is critical to drawing valid conclusions.
2. Cost-Effectiveness : Collecting data from the entire population can be expensive and time-consuming. Sampling allows researchers to make reliable inferences with fewer resources.
3. Precision and Accuracy : Through statistical techniques like **confidence intervals** and **hypothesis testing**, sampling allows researchers to estimate population parameters (like the mean or proportion) with a certain level of confidence, even though only a portion of the population is studied.
4. Simplifying Complex Data : In many cases, analyzing the entire population may not be feasible or practical. Sampling simplifies the analysis, making it easier to apply statistical methods, compute averages, and evaluate variability without overwhelming complexity.
5. Avoiding Bias : Properly executed sampling techniques (e.g., random sampling) can minimize bias, ensuring that the sample provides an unbiased representation of the population.
6. Statistical Inference : Sampling allows for **inferential statistics**, where conclusions about the population are drawn based on the sample data. This includes creating **confidence intervals**, testing **hypotheses**, and conducting **regression analysis** to predict outcomes for the population.

Que 4

Que 4 : Explain the process of hypothesis testing and the key components involved.

Ans :

It is a statistical procedure used to make inferences or draw conclusions about a population based on sample data. The goal is to test whether there is enough evidence in the sample data to support or reject a claim or assumption about the population.

The process involves formulating a **null hypothesis** and an **alternative hypothesis**, conducting an appropriate statistical test, and then making a decision based on the data and a predefined significance level.

Key Components of Hypothesis Testing :

1. Null Hypothesis (H₀):The **null hypothesis** represents the default assumption or status quo. It suggests that there is no effect, no difference, or no relationship between variables. It’s the hypothesis that we attempt to test against.

*e.g. : H₀: The average height of adults in a city is 5'7".*

1. Alternative Hypothesis (H₁ or Ha):The **alternative hypothesis** is what you want to prove or support. It represents a claim that there is an effect, a difference, or a relationship between the variables. The alternative hypothesis is typically the opposite of the null hypothesis.

*e.g. : H₁: The average height of adults in a city is not 5'7".*

1. Significance Level (α):  
    The **significance level (α)** is the threshold for rejecting the null hypothesis. It represents the probability of committing a **Type I error** (rejecting a true null hypothesis). Common significance levels are 0.05 (5%), 0.01 (1%), and 0.10 (10%).

*e.g. : If α = 0.05, it means you are willing to accept a 5% risk of incorrectly rejecting the null hypothesis when it is actually true.*

1. Test Statistic:  
    A **test statistic** is a numerical value calculated from the sample data, which is used to determine whether to reject the null hypothesis. The specific test statistic depends on the type of hypothesis test being conducted. Some common test statistics are:
   1. **Z-statistic** (for large sample sizes, when population standard deviation is known),
   2. **t-statistic** (for smaller sample sizes, when population standard deviation is unknown),
   3. **Chi-square statistic** (for categorical data analysis),
   4. **Anova / F-statistic** (for comparing variances).

The test statistic is calculated using sample data and then compared to a critical value from the relevant statistical distribution (e.g., normal distribution, t-distribution).

1. P-Value:The **p-value** is the probability of obtaining the observed results (or more extreme results) if the null hypothesis were true. It is a measure of evidence against the null hypothesis. The lower the p-value, the stronger the evidence against H₀.

* If **p ≤ α**, you reject the null hypothesis.
* If **p > α**, you fail to reject the null hypothesis.

*e.g. : If a p-value of 0.03 is obtained and α = 0.05, the null hypothesis is rejected.*

1. Critical Value(s):  
    The **critical value(s)** are the threshold values that the test statistic must exceed in order to reject the null hypothesis. Critical values are determined based on the significance level (α) and the type of test being performed (one-tailed or two-tailed test).

*e.g. : In a* ***two-tailed test*** *at α = 0.05, the critical values might be -1.96 and +1.96 (for a Z-test), meaning that if the test statistic is outside this range (either below -1.96 or above +1.96), the null hypothesis is rejected.*

1. Decision:  
    After comparing the test statistic to the critical value or examining the p-value, you make a decision:
2. **Reject H₀:** If the evidence from the sample is strong enough (i.e., the test statistic exceeds the critical value, or the p-value is less than α), you reject the null hypothesis and conclude that the alternative hypothesis is likely.
3. **Fail to Reject H₀:** If the evidence is not strong enough (i.e., the test statistic is within the critical value range, or the p-value is greater than α), you fail to reject the null hypothesis. This does **not** prove that the null hypothesis is true; it just means there is insufficient evidence to support the alternative hypothesis.

8.Type I and Type II Errors:

1. **Type I Error (False Positive):** Rejecting the null hypothesis when it is actually true. The probability of making a Type I error is the significance level (α).
2. **Type II Error (False Negative):** Failing to reject the null hypothesis when it is actually false. The probability of making a Type II error is denoted by β.

Que 5

Que 5 : Describe the T-distribution and how it differs from the normal distribution.

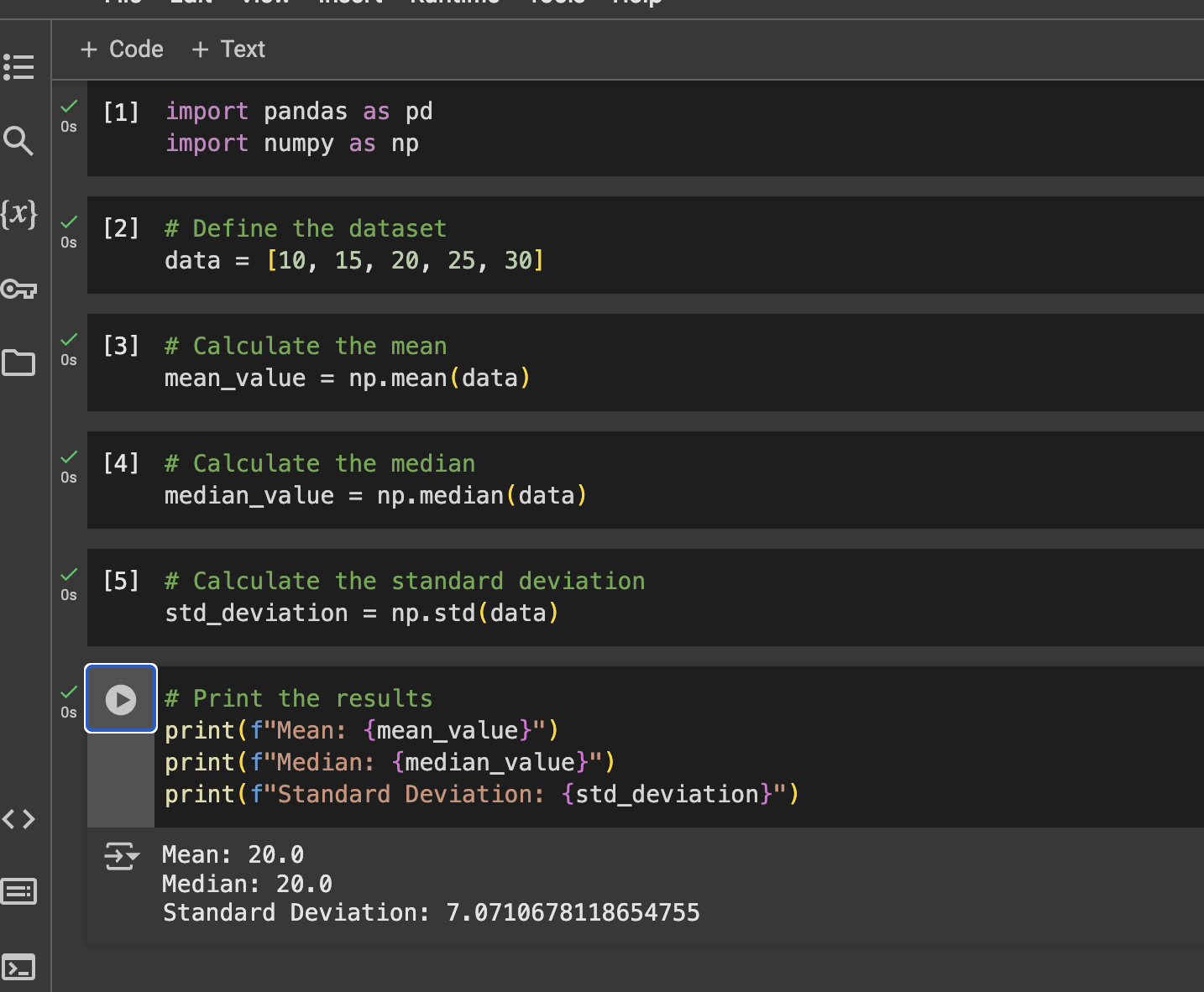
Applied Que : Que 6

**Applied Questions:**

Que 6 : Calculate the mean, median, and standard deviation for the following dataset: [10, 15, 20, 25, 30].

**Solution :**

Given dataset : [10, 15, 20, 25, 30]



Que 7

7. A researcher wants to estimate the average height of students in a university. She samples 50 students and finds the mean height to be 65 inches with a standard deviation of 3 inches. Construct a 95% confidence interval for the population mean height.

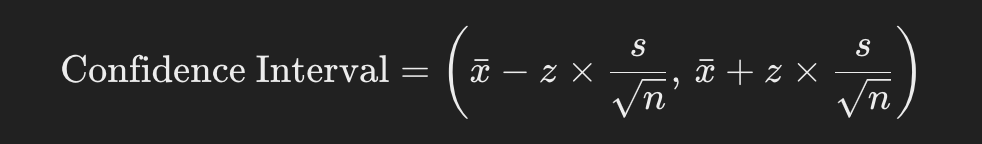
**Solution :**

Given data :

Sample mean () = 65

Sample size (n) = 50

Sample Standard deviation (s) = 3



Since sample size > 30, we will use z-score test here

Confidence Interval is 95%, so this leaves 2.5% in each tail.

Hence the corresponding upper tail is 97.5% i.e. 0.975 which is 1.960 from z-table

So the Critical z-value = 1.960

**Standard Error** = s /

= 3 /

= 0.4243

**Margin of Error** = z \* Standard Error

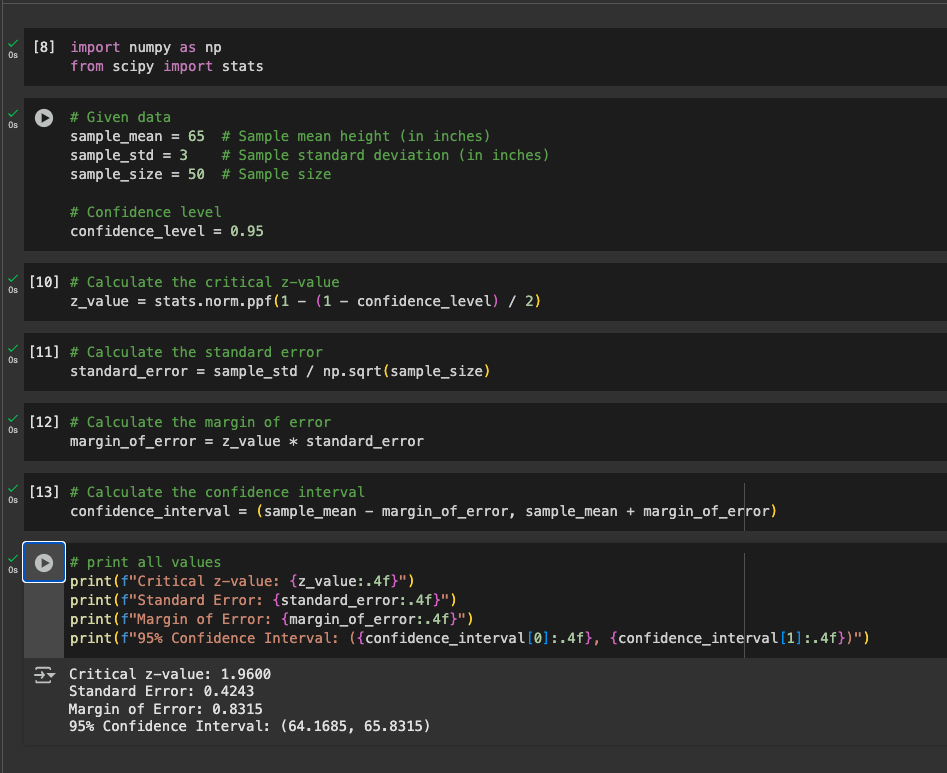
= 1.960 \* 0.4243

= 0.8316

**Confidence Interval** =

= 65 + 0.8316 or 65 - 0.8316

= 65.831 or 64.1684



Que 8

8. A manufacturer claims that the average lifespan of its light bulbs is 1000 hours. A random sample of 50 light bulbs has a mean lifespan of 980 hours with a standard deviation of 50 hours.

Test the manufacturer's claim at a significance level of 0.05 using a right-tailed hypothesis test.

**Solution :**

Given data :

Sample mean () = 980

Sample size (n) = 50

Sample Standard deviation (s) = 50

Population mean (

Since the Significance level is 0.05, hence Confidence level is 0.95

To perform Hypothesis test we need to follow below steps :

1. Define hypothesis :

Null Hypothesis (H0) : (

Alternate Hypothesis (H1) : (

1. Identify statistic test we can perform :

Since sample size is more than 30, we perform z-test

= (980 - 1000)/(50/

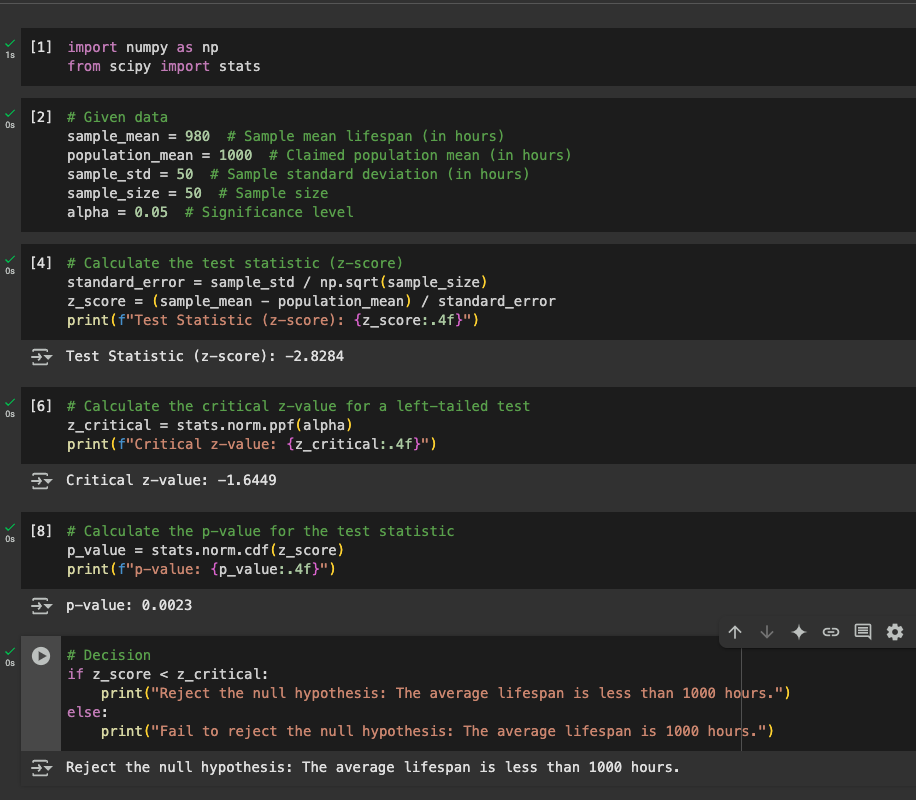
=-2.828

1. Define the value for

Critical value = -1.645

1. Compare the z statistic value and Critical value

-2.828 < -1.645 Null hypothesis falls in Rejection region



Que 9

9. A pharmaceutical company is testing a new drug for lowering blood pressure. They want to determine if the drug is effective in reducing blood pressure levels. State the null and alternative hypotheses for this study.

**Solution :**

Null Hypothesis (H0) : The new drug is effective for lowering the blood pressure

Alternate Hypothesis (H1) : The new drug is not effective for lowering the blood pressure

Que 10

10. A quality control manager at a factory wants to ensure that the average weight of products coming off the production line is 500 grams. She takes a random sample of 30 products and finds the mean weight to be 495 grams with a standard deviation of 10 grams. Test the manager's claim at a significance level of 0.01 using a left-tailed hypothesis test.

**Solution :**

Given data :

Sample mean () = 495

Sample size (n) = 30

Sample Standard deviation (s) = 10

Population mean (

Since the Significance level is 0.01, hence Confidence level is 0.99

To perform Hypothesis test we need to follow below steps :

1. Define hypothesis :

Null Hypothesis (H0) : (

Alternate Hypothesis (H1) : (

1. Identify statistic test we can perform :

Since sample size is 30, we perform t-test

= 495 - 500 / (10/

= -2.739

1. Define the value for

Critical value = -2.462

1. Compare the t statistic value and Critical value

-2.739 < 2.462 Null hypothesis falls in Rejection region

