cs-using-statistics226128147-lab-6

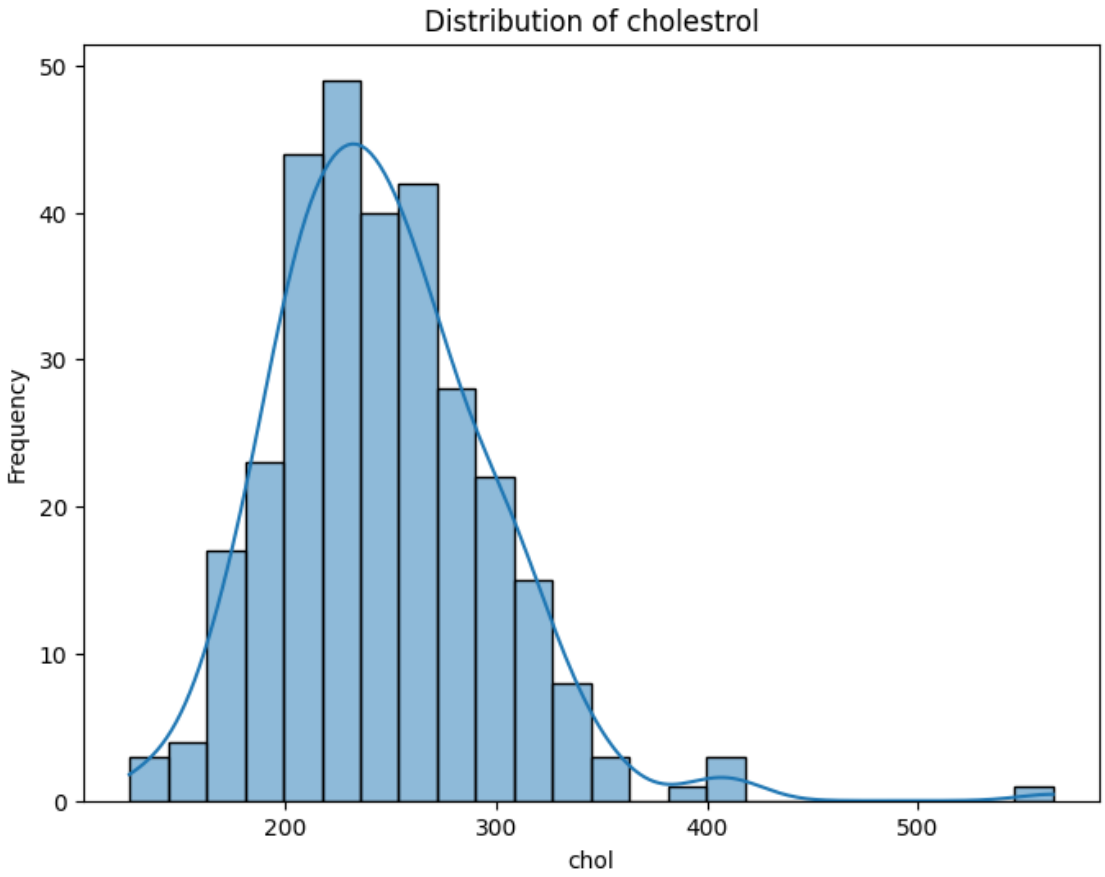
December 13, 2023

Q1.What do you meant by central limit theorem. Simulate using python code with Heart disease dataset.

Ans=he central limit theorem is a statistical concept that describes the behavior of the mean of a random sample drawn from any distribution. It states that as the sample size increases, the distribution of the sample means approaches a normal distribution, regardless of the shape of the original population distribution1.

[2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

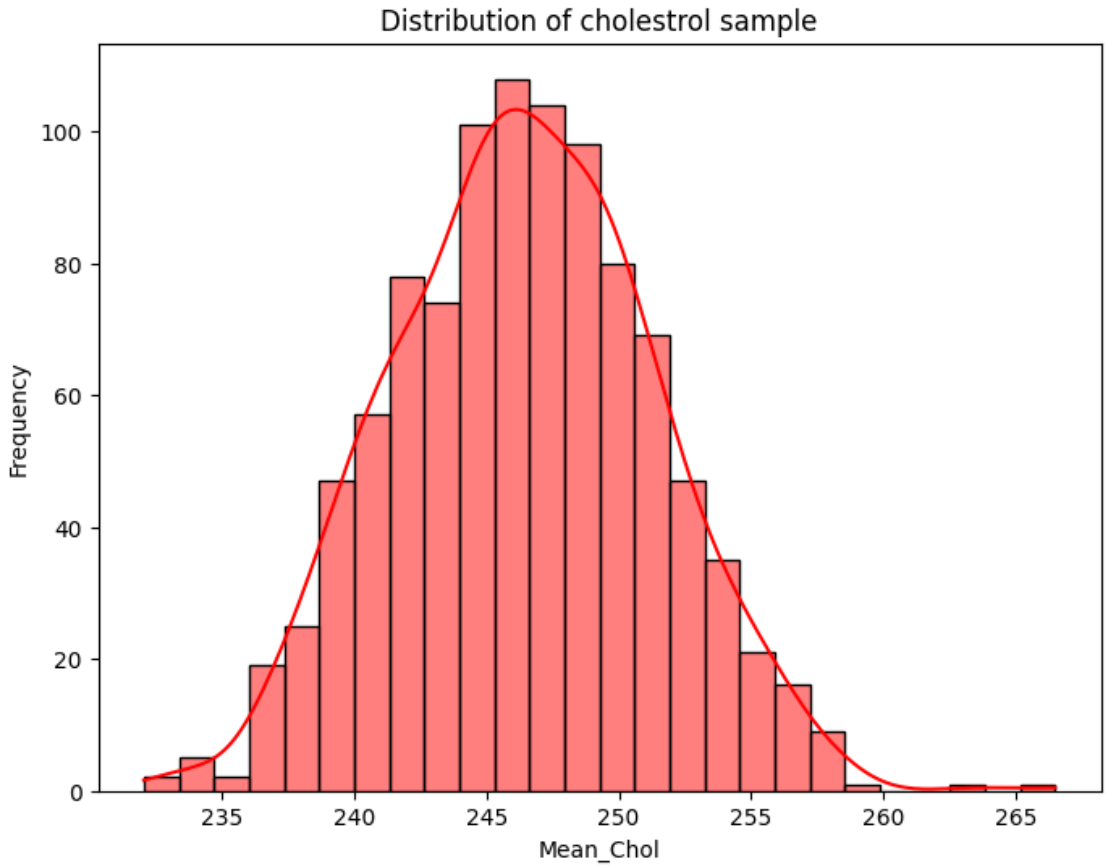
[3]:
```
df= pd.read_csv("heart.csv")
plt.figure(figsize=(8,6))
sns.histplot(df['chol'], kde=True)
plt.title(f"Distribution of cholestrol")
plt.xlabel("chol")
plt.ylabel("Frequency")
plt.show()
```



[4]:
```
sample_set_mean = []

for i in range(1000):
  sample = df["chol"].sample(100, replace=True)
  mean_s = sample.mean()
  sample_set_mean.append(mean_s)

plt.figure(figsize=(8,6))
sns.histplot(sample_set_mean,color ="red", kde=True)
plt.title(f"Distribution of cholestrol sample")
plt.xlabel("Mean_Chol")
plt.ylabel("Frequency")
plt.show()
```



[5]:
```
pop_mean=df["chol"].mean(axis=0)
print("mean of population in chol column",pop_mean)
pop_std=df["chol"].std(axis=0)
print("standard deviation of population in chol column",pop_std)
sample_mean= np.mean(sample_set_mean)

print("*"*70)

print("mean of 1000 sample with 100 rows in chol column",sample_mean)
sample_std= np.std(sample_set_mean)
print("std of 1000 sample with 100 rows in chol column",sample_std)
```

```
mean of population in chol column 246.26402640264027
standard deviation of population in chol column 51.83075098793003
**********************************************************************
mean of 1000 sample with 100 rows in chol column 246.41488999999999
std of 1000 sample with 100 rows in chol column 4.881396438305334
```

Q2. Explain the term z-score and confidence levels with suitable examples. Ans= A z-score, also known as a standard score, is a statistical measure that describes a value's relationship

to the mean of a group of values. It is expressed in terms of standard deviations from the mean. The formula for calculating the z-score of a data point (X) in a dataset with mean ( ) and standard deviation ( ) is given by:

z = − /

The resulting z-score tells you how many standard deviations a particular data point is from the mean. A positive z-score indicates that the data point is above the mean, while a negative z-score indicates that it is below the mean.

Example: Suppose you have a dataset of exam scores with a mean of 75 and a standard deviation of 10. If a student scored 85 on the exam, the z-score would be calculated as:

z =85−75/10 = 1

This means the student's score is 1 standard deviation above the mean.

Confidence Levels: Confidence levels are a way to express the degree of certainty or reliability associated with a statistical inference. In hypothesis testing or constructing confidence intervals, a confidence level represents the probability that the parameter being estimated or tested falls within a certain range.

For example, a 95% confidence level means that if you were to take many samples and construct a confidence interval for each sample, about 95% of those intervals would contain the true population parameter.

Example: Suppose you want to estimate the average height of a population. You take a sample and calculate a 95% confidence interval for the mean height. This interval might be, for instance, 160 cm to 170 cm. The interpretation is that you are 95% confident that the true average height of the population falls within this range.

In summary, a z-score helps quantify how far a data point is from the mean in terms of standard deviations, while confidence levels provide a measure of the certainty or reliability associated with statistical estimates or tests.

Q3. What is null hypothesis? What are the conditions to reject null hypothesis. Ans=The null hypothesis is a statement or assumption that there is no significant difference or effect. It is often denoted as (H0). In statistical hypothesis testing, the null hypothesis represents a default position that there is no change or no effect.

For example, if you are comparing the mean of two groups, the null hypothesis might state that the means are equal. In symbols, for a mean comparison:

H0: 1 = 2

Here, H0 asserts that there is no difference between the means of the two groups.

The conditions to reject the null hypothesis depend on the statistical test being used and the significance level ( ) chosen by the researcher. In general, the rejection of the null hypothesis is based on the p-value obtained from the statistical test.

Common Steps: 1. Formulate the Null Hypothesis H0: Define a hypothesis that there is no effect, no difference, or no relationship. 2. Select the Significance Level ( ): The significance level, often denoted as ( ) is the probability of rejecting the null hypothesis when it is true. Common choices are 0.05, 0.01, or 0.10. 3. Collect and Analyze Data: Collect and analyze the data using an

appropriate statistical test. 4. Calculate the p-value: The p-value is the probability of obtaining results as extreme as, or more extreme than, the observed results under the assumption that the null hypothesis is true. 5.Make a Decision: If the p-value is less than or equal to the significance level ( ), you reject the null hypothesis. If the p-value is greater than the significance level, you fail to reject the null hypothesis.

Conditions to Reject the Null Hypothesis: Small p-value: A small p-value (typically less than the chosen significance level) indicates that the observed data is unlikely under the assumption that the null hypothesis is true.

Compare p-value to Significance Level: If the p-value is less than or equal to the chosen significance level ( ), you reject the null hypothesis.

Strong Evidence Against Null Hypothesis: A smaller p-value provides stronger evidence against the null hypothesis.

It's important to note that failing to reject the null hypothesis does not prove the null hypothesis is true; it simply means there is not enough evidence to reject it based on the data at hand. The conditions to reject the null hypothesis depend on the chosen significance level and the results of the statistical test.

Q4. What do you mean by z-test. Mention the steps with an example.

Ans=A z-test is a statistical test used to determine if there is a significant difference between the means of a sample and a known or hypothesized population mean. It is particularly applicable when the population standard deviation ( ) is known.

The z-test involves calculating a z-score, which represents how many standard deviations a data point or sample mean is from the population mean. The z-score is then compared to critical values or used to calculate a p-value to determine whether to reject the null hypothesis.

Steps for a One-Sample Z-Test: 1. Formulate the Hypotheses: - Null Hypothesis (H0): There is no significant difference between the sample mean ( ) and the population mean ( ). H0: ( ) = - Alternative Hypothesis (H1) or (Ha): There is a significant difference between the sample mean and the population mean. (H1) : ( ) (two-tailed test) (H1) : ( ) > (right-tailed test) (H1) : ( ) < (left-tailed test) 2. Choose the Significance Level ( ): - Common choices include 0.05, 0.01, or 0.10. 3. Collect and Analyze Data: - Collect a sample of data and calculate the sample mean ( ). 4. Calculate the Z-Score: - Use the formula: = ( ) − √ - ( ) is the sample mean. - is the population mean under the null hypothesis. - is the known population standard deviation. - n is the sample size. 5. Determine Critical Values or P-Value: - For a two-tailed test, find the critical values or calculate the p-value. - For a one-tailed test, determine whether the z-score falls in the critical region (reject H0): 6. Make a Decision: - If the z-score falls into the critical region or if the p-value is less than the significance level ( ), reject the null hypothesis. - If the z-score is not in the critical region and the p-value is greater than ( ), fail to reject the null hypothesis. Example: Suppose you want to test whether the average height of a sample of students ( ) is significantly different from the known population mean height ( ) of 65 inches. The population standard deviation ( ) is 5 inches. You collect a sample of 25 students, and the sample mean is 67 inches. - Null Hypothesis: H0: ( ) = - Alternative Hypothesis: (H1) : ( ) (two-tailed test) - Significance Level: ( )= 0.05.

You then calculate the z-score using the formula mentioned above and compare it to critical values or calculate a p-value to make a decision about the null hypothesis.

Q5. Generate a random array of 50 numbers having mean 110 and sd 15. Do the z-test 3 with 5% significance level.(Use python modulestatsmodels.stats.weightstats.ztest and function z-test)

[6]:
```
import numpy as np
from statsmodels.stats.weightstats import ztest

# Generate random array
random_array = np.random.normal(110,15,50)

# Perform z-test
null_hypothesis_mean = 110 # Assuming the null hypothesis that the mean is 110
significance_level = 0.05

# Perform one-sample z-test
z_stat, p_value = ztest(random_array, value=null_hypothesis_mean)
print("z score value:",z_stat)
print("p alue:",p_value)

# Check if the null hypothesis is rejected
if p_value < significance_level:
  print(f"Reject the null hypothesis.")
else:
  print(f"Fail to reject the null hypothesis.")
```

```
z score value: -0.02279541839103585
p alue: 0.9818134626707882
Fail to reject the null hypothesis.
```

6. Generate an array with values: [88, 92, 94, 94, 96, 97, 97, 97, 99, 99,105, 109, 109,109, 110, 112, 112, 113, 114, 115], with mean=100 and sd=15. Do the z-test with 5% significance level.

[7]:
```
# Given data
data = np.array([88, 92, 94, 94, 96, 97, 97, 97, 99, 99,
105, 109, 109, 109, 110, 112, 112, 113, 114, 115])

# Given mean and standard deviation
given_sd = 15
null_hypothesis_mean = 100
significance_level = 0.05

# Perform one-sample z-test
z_stat2, p_value2 = ztest(data, value=null_hypothesis_mean)
print("z score value:",z_stat2)
print("p alue:",p_value2)

# Check if the null hypothesis is rejected
```

```
if p_value2 < significance_level:
  print(f"Reject the null hypothesis.")
else:
  print(f"Fail to reject the null hypothesis.")
```

```
z score value: 1.5976240527147705
p alue: 0.1101266701438426
Fail to reject the null hypothesis.
```

Q7. A new toll road is being build on the expectation that 8500 cars will use it per day.In the 300 days of its operation a daily average of 8120 cars where found to have used the toll road. Using 1% level of significance test whether the expectation was incorrect?(Assume that distribution of daily road users is normally distributed with sd 950, Use python code for calculation.)

[8]:
```
import numpy as np
from scipy.stats import norm

# Given data
expected_mean = 8500
sample_mean = 8120
sample_size = 300
sample_sd = 950
significance_level = 0.01

# Calculate the standard error of the mean
deno = sample_sd / np.sqrt(sample_size)

# Calculate the z-score
z_score = (sample_mean - expected_mean) / deno

# Calculate the p-value
p_value = 2 * norm.cdf(-np.abs(z_score)) # Multiply by 2 for a two-tailed test
print("z score:",z_score)
print("p value:",p_value)

# Check if the null hypothesis is rejected
if p_value < significance_level:
  print("Reject the null hypothesis.")
else:
  print("Fail to reject the null hypothesis.")
```

```
z score: -6.92820323027551
p value: 4.262191597843629e-12
Reject the null hypothesis.
```

Q8. Suppose a botanist wants to know if the mean height of a certain species of plant is equal to 15 inches. She collects a random sample of 12 plants and records each of their heights in inches. Use the one sample t-test to determine if the mean height for

this species of plant is actually equal to 15 inches.(data= [14, 14, 16, 13, 12, 17, 15, 14, 15, 13, 15, 14], library: ttest_1samp() function from the scipy.stats library).

[9]:
```
from scipy.stats import ttest_1samp

# Given data
data = [14, 14, 16, 13, 12, 17, 15, 14, 15, 13, 15, 14]

# Given null hypothesis mean
null_hypothesis_mean = 15

# Perform one-sample t-test
t_stat, p_value3 = ttest_1samp(data, null_hypothesis_mean)
print("t test value:",t_stat)
print("p value:",p_value3)

# Check if the null hypothesis is rejected
significance_level = 0.05
if p_value3 < significance_level:
  print("Reject the null hypothesis.")
else:
  print("Fail to reject the null hypothesis.")
```

```
t test value: -1.6848470783484626
p value: 0.12014460742498101
Fail to reject the null hypothesis.
```

Q9. An online fashion store called showdonkey advertises that its average delivery time is less than six hours for local deliveries. A random sample of the amount of time taken to deliver a package to an address is Stanmore produced the delivery times as shown:7,3,4,6,10,5,6,4,3,8. Is there sufficient evidence to support showdonkeys advertisement with 5% level of significance. (Use python code).

[10]:
```
# Given data
delivery_times = [7, 3, 4, 6, 10, 5, 6, 4, 3, 8]

# Given null hypothesis mean
null_hypothesis_mean = 6

# Perform one-sample t-test
t_stat1, p_value4 = ttest_1samp(delivery_times, null_hypothesis_mean)
print("t test value:",t_stat1)
print("p value:",p_value4)

# Check if the null hypothesis is rejected
significance_level = 0.05
if p_value < significance_level:
  print("Reject the null hypothesis.")
```

```
else:
  print("Fail to reject the null hypothesis.")
```

```
t test value: -0.5570860145311561
p value: 0.5910512317836043
Reject the null hypothesis.
```

Q10. Take the heart csv file do univariate sampling on the chol column, do bivariate analysis on the columns on your choice and find highly correlated columns.

[11]:
```
#univariate sampling on the chol column

pop_mean_uni=df["chol"].mean(axis=0)
print("mean of population in chol column",pop_mean_uni)
pop_std_uni=df["chol"].std(axis=0)
print("standard deviation of population in chol column",pop_std_uni)

print("*"*75)

sample_uni = df["chol"].sample(100, replace=True)
mean_sample_uni = sample_uni.mean()
print("mean of sample with 100 rows in chol column",mean_sample_uni)
sample_std_uni= np.std(sample_set_mean)
mean_sample_std = sample_uni.std()
print("std of sample with 100 rows in chol column",sample_std_uni)
```

```
mean of population in chol column 246.26402640264027
standard deviation of population in chol column 51.83075098793003
***************************************************************************
mean of sample with 100 rows in chol column 244.71
std of sample with 100 rows in chol column 4.881396438305334
```

[12]:
```
#bivariate analysis on age and chol column.
import random
x=df['chol']
y=df['age']
pop_corr= np.corrcoef(x, y)[0,1]
print("Correlation coefficient wrt population:",pop_corr)
bi_variate_sample = list(zip(df["chol"].sample(100), df["age"].sample(100)))
sample_correlation = np.corrcoef(bi_variate_sample, rowvar=False)[0,1]
print("Correlation coefficient wrt sample",sample_correlation)
```

```
Correlation coefficient wrt population: 0.21367795655956182
Correlation coefficient wrt sample -0.050820462517509234
```

[13]:
```
df.corr()
```

[13]:

	age	sex	cp	trestbps	chol	fbs
age	1.000000	-0.098447	-0.068653	0.279351	0.213678	0.121308
sex	-0.098447	1.000000	-0.049353	-0.056769	-0.197912	0.045032
cp	-0.068653	-0.049353	1.000000	0.047608	-0.076904	0.094444
trestbps	0.279351	-0.056769	0.047608	1.000000	0.123174	0.177531
chol	0.213678	-0.197912	-0.076904	0.123174	1.000000	0.013294
fbs	0.121308	0.045032	0.094444	0.177531	0.013294	1.000000
restecg	-0.116211	-0.058196	0.044421	-0.114103	-0.151040	-0.084189
thalach	-0.398522	-0.044020	0.295762	-0.046698	-0.009940	-0.008567
exang	0.096801	0.141664	-0.394280	0.067616	0.067023	0.025665
oldpeak	0.210013	0.096093	-0.149230	0.193216	0.053952	0.005747
slope	-0.168814	-0.030711	0.119717	-0.121475	-0.004038	-0.059894
ca	0.276326	0.118261	-0.181053	0.101389	0.070511	0.137979
thal	0.068001	0.210041	-0.161736	0.062210	0.098803	-0.032019
target	-0.225439	-0.280937	0.433798	-0.144931	-0.085239	-0.028046

	restecg	thalach	exang	oldpeak	slope	ca
age	-0.116211	-0.398522	0.096801	0.210013	-0.168814	0.276326
sex	-0.058196	-0.044020	0.141664	0.096093	-0.030711	0.118261
cp	0.044421	0.295762	-0.394280	-0.149230	0.119717	-0.181053
trestbps	-0.114103	-0.046698	0.067616	0.193216	-0.121475	0.101389
chol	-0.151040	-0.009940	0.067023	0.053952	-0.004038	0.070511
fbs	-0.084189	-0.008567	0.025665	0.005747	-0.059894	0.137979
restecg	1.000000	0.044123	-0.070733	-0.058770	0.093045	-0.072042
thalach	0.044123	1.000000	-0.378812	-0.344187	0.386784	-0.213177
exang	-0.070733	-0.378812	1.000000	0.288223	-0.257748	0.115739
oldpeak	-0.058770	-0.344187	0.288223	1.000000	-0.577537	0.222682
slope	0.093045	0.386784	-0.257748	-0.577537	1.000000	-0.080155
ca	-0.072042	-0.213177	0.115739	0.222682	-0.080155	1.000000
thal	-0.011981	-0.096439	0.206754	0.210244	-0.104764	0.151832
target	0.137230	0.421741	-0.436757	-0.430696	0.345877	-0.391724

	thal	target
age	0.068001	-0.225439
sex	0.210041	-0.280937
cp	-0.161736	0.433798
trestbps	0.062210	-0.144931
chol	0.098803	-0.085239
fbs	-0.032019	-0.028046
restecg	-0.011981	0.137230
thalach	-0.096439	0.421741
exang	0.206754	-0.436757
oldpeak	0.210244	-0.430696
slope	-0.104764	0.345877
ca	0.151832	-0.391724
thal	1.000000	-0.344029
target	-0.344029	1.000000