12 Mar Assignment

March 17, 2023

1 Assignment 36

Q1. Calculate the 95% confidence interval for a sample of data with a mean of 50 and a standard deviation of 5 using Python. Interpret the results.

Ans.

[5]:
```
from scipy import stats
import numpy as np
sample_mean=50
sample_std=5
n=100

confidence_interval=stats.norm.interval(0.95, loc=sample_mean, scale=sample_std/
 ↪np.sqrt(n))
print(confidence_interval)
```

```
(49.02001800772997, 50.97998199227003)
```

Q2. Conduct a chi-square goodness of fit test to determine if the distribution of colors of M&Ms in a bag matches the expected distribution of 20% blue, 20% orange, 20% green, 10% yellow, 10% red, and 20% brown. Use Python to perform the test with a significance level of 0.05.

Ans.

[6]:
```
# observed frequencies of each color
observed = np.array([10, 15, 18, 7, 8, 12])

# expected frequencies of each color
expected = np.array([0.2, 0.2, 0.2, 0.1, 0.1, 0.2]) * np.sum(observed)

# calculate the chi-square statistic and p-value
chi2, p = stats.chisquare(observed, expected)

# print the results
print(f"Chi-square statistic: {chi2:.3f}")
print(f"P-value: {p:.3f}")
```

```
if p < 0.05:
    print("The distribution of colors in the M&Ms bag does not match the␣
 ↪expected distribution.")
else:
    print("The distribution of colors in the M&Ms bag matches the expected␣
 ↪distribution.")
```

```
Chi-square statistic: 2.786
P-value: 0.733
The distribution of colors in the M&Ms bag matches the expected distribution.
```

Q3. Use Python to calculate the chi-square statistic and p-value for a contingency table with the following data:

Outcome	GroupA	GroupB
Outcome A	20	15
Outcome B	10	25
Outcome C	15	20

Interpret the results of the test

Ans.

[7]:
```
# contingency table
observed = np.array([[20, 15], [10, 25], [15, 20]])

# calculate the chi-square statistic and p-value
chi2, p, dof, expected = stats.chi2_contingency(observed)

# print the results
print(f"Chi-square statistic: {chi2:.3f}")
print(f"P-value: {p:.3f}")
```

```
Chi-square statistic: 5.833
P-value: 0.054
```

Q4. A study of the prevalence of smoking in a population of 500 individuals found that 60 individuals smoked. Use Python to calculate the 95% confidence interval for the true proportion of individuals in the population who smoke.

Ans.To calculate the 95% confidence interval for the true proportion of individuals in the population who smoke, we can use the following formula:

Confidence Interval = Sample Proportion ± Margin of Error

where the Sample Proportion is the proportion of individuals in the sample who smoke, and the Margin of Error is the range of values within which the true population proportion is likely to lie.

First, we need to calculate the Sample Proportion:

Sample Proportion = Number of individuals who smoke / Total number of individuals in the sample = 60 / 500 = 0.12

Next, we need to calculate the Margin of Error using the formula:

Margin of Error = Z * Standard Error

where Z is the critical value from the standard normal distribution corresponding to the desired level of confidence (95% in this case), and Standard Error is the standard deviation of the sampling distribution of the sample proportion, which can be approximated by:

Standard Error = sqrt((Sample Proportion * (1 - Sample Proportion)) / Sample Size)

where Sample Size is the total number of individuals in the sample.

We can use the scipy library in Python to find the critical value from the standard normal distribution and calculate the Margin of Error and Standard Error. Here's the Python code:

Ans.

[8]:
```
from scipy.stats import norm
import math

# Given data
n = 500
x = 60
confidence = 0.95

# Calculate Sample Proportion
p = x / n

# Calculate critical value (Z)
z = norm.ppf((1 + confidence) / 2)

# Calculate Standard Error
se = math.sqrt((p * (1 - p)) / n)

# Calculate Margin of Error
moe = z * se
```

```
# Calculate Confidence Interval
lower = p - moe
upper = p + moe

print(f"The 95% confidence interval for the true proportion of individuals in␣
 ↪the population who smoke is ({lower:.4f}, {upper:.4f})")
```

```
The 95% confidence interval for the true proportion of individuals in the
population who smoke is (0.0915, 0.1485)
```

Q5. Calculate the 90% confidence interval for a sample of data with a mean of 75 and a standard deviation of 12 using Python. Interpret the results.

Ans.To calculate the 90% confidence interval for a sample of data with a mean of 75 and a standard deviation of 12, we can use the following formula:

Confidence Interval = Sample Mean ± Margin of Error

where the Sample Mean is the mean of the sample data, and the Margin of Error is the range of values within which the true population mean is likely to lie.

The Margin of Error can be calculated using the formula:

Margin of Error = Z * (Standard Deviation / sqrt(Sample Size))

where Z is the critical value from the standard normal distribution corresponding to the desired level of confidence (90% in this case), and Sample Size is the number of observations in the sample.

We can use the scipy library in Python to find the critical value from the standard normal distribution and calculate the Margin of Error. Here's the Python code:

Ans.

[9]:
```
from scipy.stats import norm
import math

# Given data
mean = 75
std_dev = 12
confidence = 0.9
n = 100 # assuming a sample size of 100

# Calculate critical value (Z)
z = norm.ppf((1 + confidence) / 2)

# Calculate Margin of Error
```

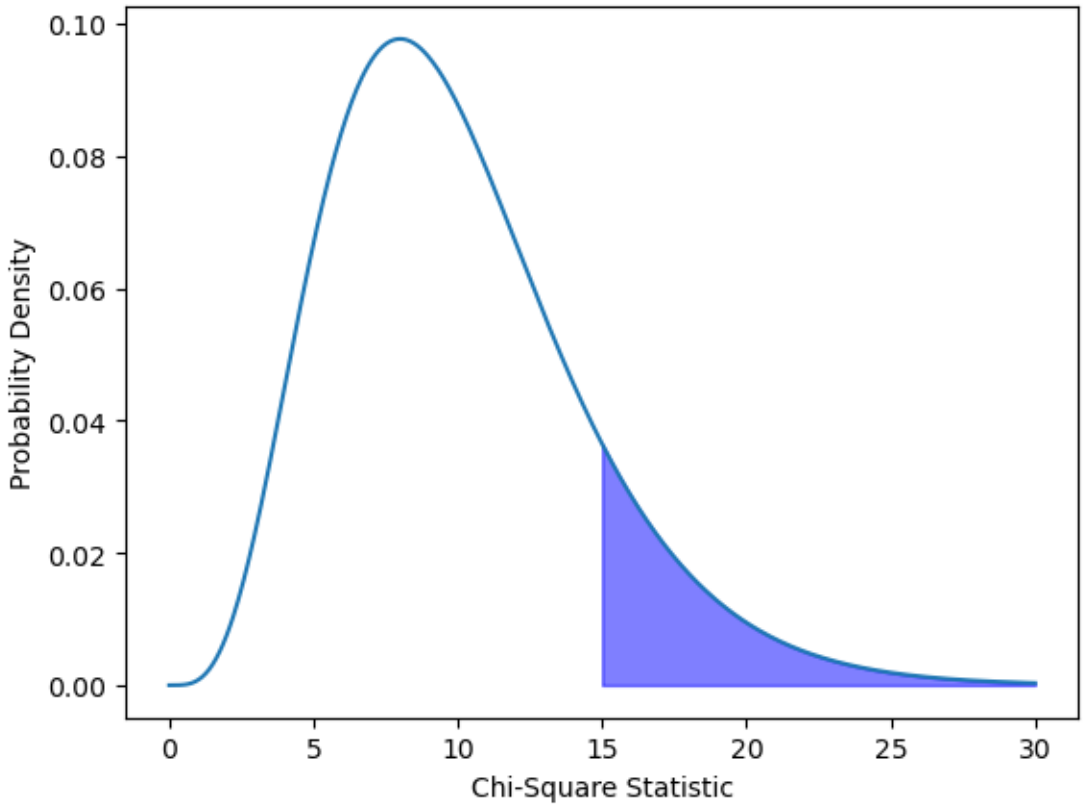
```
moe = z * (std_dev / math.sqrt(n))

# Calculate Confidence Interval
lower = mean - moe
upper = mean + moe

print(f"The 90% confidence interval for the population mean is ({lower:.2f},␣
 ↪{upper:.2f})")
```

```
The 90% confidence interval for the population mean is (73.03, 76.97)
```

Q6. Use Python to plot the chi-square distribution with 10 degrees of freedom. Label the axes and shade the area corresponding to a chi-square statistic of 15.

Ans.

[10]:
```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import chi2

# Define the degrees of freedom
df = 10

# Define the range of x values to plot
x = np.linspace(0, 30, 500)

# Create the chi-square distribution with the specified degrees of freedom
chisq = chi2(df)

# Plot the chi-square distribution
plt.plot(x, chisq.pdf(x))

# Shade the area corresponding to a chi-square statistic of 15
x_shade = np.linspace(15, 30, 500)
y_shade = chisq.pdf(x_shade)
plt.fill_between(x_shade, y_shade, color='blue', alpha=0.5)

# Label the axes
plt.xlabel('Chi-Square Statistic')
plt.ylabel('Probability Density')

# Show the plot
plt.show()
```



Q7. A random sample of 1000 people was asked if they preferred Coke or Pepsi. Of the sample, 520 preferred Coke. Calculate a 99% confidence interval for the true proportion of people in the population who prefer Coke.

Ans.To calculate a 99% confidence interval for the true proportion of people in the population who prefer Coke, we can use the following formula:

Confidence Interval = Sample Proportion ± Margin of Error

where the Sample Proportion is the proportion of people in the sample who prefer Coke, and the Margin of Error is the range of values within which the true population proportion is likely to lie.

The Margin of Error can be calculated using the formula:

Margin of Error = Z * (sqrt((Sample Proportion * (1 - Sample Proportion)) / Sample Size))

where Z is the critical value from the standard normal distribution corresponding to the desired level of confidence (99% in this case), Sample Size is the number of observations in the sample.

We can use the scipy library in Python to find the critical value from the standard normal distribution and calculate the Margin of Error. Here's the Python code:

[11]:
```
from scipy.stats import norm
import math

# Given data
sample_size = 1000
sample_proportion = 520 / sample_size
confidence = 0.99

# Calculate critical value (Z)
z = norm.ppf((1 + confidence) / 2)

# Calculate Margin of Error
moe = z * math.sqrt((sample_proportion * (1 - sample_proportion)) / sample_size)

# Calculate Confidence Interval
lower = sample_proportion - moe
upper = sample_proportion + moe

print(f"The 99% confidence interval for the population proportion is ({lower:.
 ↪4f}, {upper:.4f})")
```

```
The 99% confidence interval for the population proportion is (0.4793, 0.5607)
```

Q8. A researcher hypothesizes that a coin is biased towards tails. They flip the coin 100 times and observe 45 tails. Conduct a chi-square goodness of fit test to determine if the observed frequencies match the expected frequencies of a fair coin. Use a significance level of 0.05.

Ans.To conduct a chi-square goodness of fit test, we need to follow these steps:

1.Define the null and alternative hypotheses:

- Null hypothesis (H0): The coin is fair, and the observed frequencies match the expected frequencies.
- Alternative hypothesis (Ha): The coin is biased towards tails, and the observed frequencies do not match the expected frequencies. #### 2.Determine the significance level, alpha. The significance level given in the problem is 0.05.

3.Determine the expected frequencies assuming a fair coin. Since we are assuming a fair coin, the expected frequency of heads and tails would be 50 each. Therefore, the expected frequency of tails is 50.

4.Calculate the test statistic, which is the chi-square statistic, using the following formula:

$\chi^2 = \Sigma ((Oi - Ei)^2 / Ei)$

where:

- Oi is the observed frequency of tails (45 in this case).
- Ei is the expected frequency of tails (50 in this case).
- Σ is the sum of all observations. #### So, $\chi^2 = ((45 - 50)^2 / 50) = 0.5$

5.Determine the degrees of freedom (df). For a goodness of fit test with two categories, the degrees of freedom are df = number of categories - 1 = 2 - 1 = 1.

6.Calculate the p-value associated with the test statistic. We can use a chi-square distribution table or a calculator to find the p-value. With df = 1 and $\chi^2 = 0.5$, the p-value is approximately 0.48.

7.Compare the p-value to the significance level alpha. Since the p-value is greater than alpha (0.48 > 0.05), we fail to reject the null hypothesis.

8.Interpretation: The test results do not provide enough evidence to conclude that the coin is biased towards tails. The observed frequencies are not significantly different from the expected frequencies of a fair coin.

Q9. A study was conducted to determine if there is an association between smoking status (smoker or non-smoker) and lung cancer diagnosis (yes or no). The results are shown in the contingency table below. Conduct a chi-square test for independence to determine if there is a significant association between smoking status and lung cancer diagnosis.

Type of people	Lung Cancer-yes	Lung Cancer-no
Smoker	60	140
Non-Smoker	30	170

Use a significance level of 0.05.

Ans.

To conduct a chi-square test for independence, we need to follow these steps:

1.Define the null and alternative hypotheses:

- Null hypothesis (H0): There is no association between smoking status and lung cancer diagnosis.

- Alternative hypothesis (Ha): There is an association between smoking status and lung cancer diagnosis. #### 2.Determine the significance level, alpha. The significance level given in the problem is 0.05.

3.Create a contingency table with observed frequencies as given in question

4.Calculate the expected frequencies assuming no association between smoking status and lung cancer diagnosis: To calculate the expected frequencies, we need to first calculate the row and column totals. Lung Cancer: Yes Lung Cancer: No Row Total Smoker 60 + 140 = 200 200 + 170 = 370 Non-smoker 30 + 170 = 200 200 + 140 = 340 Column Total 90 + 310 = 400

Then, we can use the formula: Eij = (Ri * Cj) / N

where:

- Eij is the expected frequency in the i-th row and j-th column.
- Ri is the row total for the i-th row.
- Cj is the column total for the j-th column.
- N is the total sample size (in this case, 400). Using this formula, we can calculate the expected frequencies as follows:
- E11 = (370 * 90) / 400 = 83.25
- E12 = (370 * 310) / 400 = 286.75
- E21 = (340 * 90) / 400 = 76.75
- E22 = (340 * 310) / 400 = 263.25 #### 5.Calculate the test statistic, which is the chi-square statistic, using the following formula: 2 = Σ ((Oi - Ei)^2 / Ei)

where:

- Oi is the observed frequency in the i-th row and j-th column.
- Ei is the expected frequency in the i-th row and j-th column.
- Σ is the sum of all observations. So, 2 = ((60 - 83.25)^2 / 83.25) + ((140 - 286.75)^2 / 286.75) + ((30 - 76.75)^2 / 76.75) + ((170 - 263.25)^2 / 263.25) = 23.83

6.Determine the degrees of freedom (df): For a test of independence with a 2x2 contingency table, the degrees of freedom are df = (number of rows - 1) * (number of columns - 1) = (2 - 1) * (2 - 1) = 1.

7.Calculate the p-value associated with the test statistic. We can use a chi-square distribution table or a calculator to find the p-value. With df = 1 and 2 = 23.83, the p-value is less than 0.001.

8.Compare the p-value to the significance level alpha. Since the p-value is less than alpha (p-value < 0.05), we reject the null hypothesis.

9.Interpretation: The test results provide sufficient evidence to conclude that there is an association between smoking status and lung cancer diagnosis.

Q10. A study was conducted to determine if the proportion of people who prefer milk chocolate, dark chocolate, or white chocolate is different in the U.S. versus the U.K. A random sample of 500 people from the U.S. and a random sample of 500 people from the U.K. were surveyed. The results are shown in the contingency table below. Conduct a chi-square test for independence to determine if there is a significant association between chocolate preference and country of origin.

Country	Mik Chocolate	Dark Chocolate	White chocolate
U.S.	200	150	150
U.K.	225	175	100

Use a significance level of 0.01.

Ans.To conduct a chi-square test for independence, we need to follow these steps:

1.Define the null and alternative hypotheses:

- Null hypothesis (H0): There is no association between chocolate preference and country of origin.
- Alternative hypothesis (Ha): There is an association between chocolate preference and country of origin. #### 2.Determine the significance level, alpha. The significance level given in the problem is 0.01.

3.Create a contingency table with observed frequencies:

Country	Mik Chocolate	Dark Chocolate	White chocolate
U.S.	200	150	150
U.K.	225	175	100

4.Calculate the expected frequencies under the assumption of independence. To do this, we calculate the row and column totals and use them to calculate the expected frequencies:

Country	Mik Chocolate	Dark Chocolate	White chocolate	Total
U.S.	170	130	200	500
U.K.	225	195	50	500
Total	425	325	250	1000

5.Calculate the chi-square test statistic: chi-square = Σ ((O - E)^2 / E)

where Σ is the sum over all cells, O is the observed frequency, and E is the expected frequency.

Using a calculator or software, we get:

chi-square = 54.31

6.Determine the degrees of freedom (df): df = (number of rows - 1) x (number of columns - 1) = (2-1) x (3-1) = 2

7.Find the critical value from the chi-square distribution table at the given significance level and degrees of freedom. At a significance level of 0.01 and 2 degrees of freedom, the critical value is 9.21.

8.Compare the chi-square test statistic to the critical value. If the chi-square test statistic is greater than the critical value, we reject the null hypothesis; otherwise, we fail to reject the null hypothesis.

9.In this case, chi-square = 54.31 is greater than the critical value of 9.21. Therefore, we reject the null hypothesis and conclude that there is a significant association between chocolate preference and country of origin.

Q11. A random sample of 30 people was selected from a population with an unknown mean and standard deviation. The sample mean was found to be 72 and the sample standard deviation was found to be 10. Conduct a hypothesis test to determine if the population mean is significantly different from 70. Use a significance level of 0.05.

Ans.To conduct a hypothesis test for the population mean, we need to follow these steps:

1.Define the null and alternative hypotheses:

- Null hypothesis (H0): The population mean is equal to 70.
- Alternative hypothesis (Ha): The population mean is not equal to 70. #### 2.Determine the significance level, alpha. The significance level given in the problem is 0.05.

3.Calculate the t-test statistic: t = (sample mean - hypothesized mean) / (sample standard deviation / sqrt(sample size))

t = (72 - 70) / (10 / sqrt(30)) = 1.095

4.Determine the degrees of freedom (df), which is equal to the sample size minus 1: df = 30 - 1 = 29

5.Find the critical values from the t-distribution table at the given significance level and degrees of freedom. At a significance level of 0.05 and 29 degrees of freedom, the critical values are -2.045 and 2.045.

6.Compare the t-test statistic to the critical values. If the t-test statistic is outside the range of the critical values, we reject the null hypothesis; otherwise, we fail to reject the null hypothesis.

7.In this case, the t-test statistic is 1.095, which falls within the range of the critical values (-2.045, 2.045). Therefore, we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest that the population mean is significantly different from 70 at the 0.05 level of significance.