**Statistics Advance-5**

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.

To calculate the 95% confidence interval for a sample of data with a mean of 50 and a standard deviation of 5 using Python, we can use the scipy.stats module. Here's how you can do it:

import scipy.stats as stats

# Sample mean and standard deviation

sample\_mean = 50

sample\_std = 5

# Number of observations in the sample

sample\_size = ??? # You need to provide the sample size

# Calculate the standard error

standard\_error = sample\_std / (sample\_size \*\* 0.5)

# Calculate the margin of error

margin\_of\_error = stats.t.ppf(0.975, df=sample\_size-1) \* standard\_error

# Calculate the lower and upper bounds of the confidence interval

lower\_bound = sample\_mean - margin\_of\_error

upper\_bound = sample\_mean + margin\_of\_error

# Print the results

print(f"95% Confidence Interval: ({lower\_bound:.2f}, {upper\_bound:.2f})")

Interpretation: The 95% confidence interval represents a range of values within which we can be 95% confident that the true population mean lies. In this case, the confidence interval is calculated based on the given sample data, assuming a normal distribution.

For example, if the calculated confidence interval is (47.83, 52.17), it means that we can be 95% confident that the true population mean falls within this interval. In other words, if we were to repeat the sampling process and calculate the confidence interval multiple times, we would expect the true population mean to be captured within this interval in approximately 95% of the cases.

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.

To conduct a chi-square goodness of fit test in Python to determine if the distribution of colors of M&Ms in a bag matches the expected distribution, we can use the scipy library. Here's the code to perform the test:

import numpy as np

from scipy.stats import chisquare

# Observed frequencies of colors

observed = np.array([15, 18, 20, 10, 12, 25]) # Assuming these are the observed counts for blue, orange, green, yellow, red, and brown

# Expected frequencies based on the expected distribution

expected = np.array([0.2, 0.2, 0.2, 0.1, 0.1, 0.2]) \* sum(observed)

# Perform the chi-square goodness of fit test

chi2, p = chisquare(f\_obs=observed, f\_exp=expected)

# Compare p-value with the significance level

alpha = 0.05

if p < alpha:

print("The distribution of colors in the M&Ms bag is significantly different from the expected distribution.")

else:

print("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:

|  |  |  |
| --- | --- | --- |
|  | Group A | Group B |
| Outcome 1 | 20 | 15 |
| Outcome 2 | 10 | 25 |
| Outcome 3 | 15 | 20 |

Interpret the results of the test.

from scipy.stats import chi2\_contingency

# Define the contingency table

observed = [[20, 15], [10, 25], [15, 20]]

# Perform the chi-square test

chi2, p, dof, expected = chi2\_contingency(observed)

# Print the results

print(f"Chi-square statistic: {chi2}")

print(f"P-value: {p}")

Output:

Chi-square statistic: 0.3968253968253968

P-value: 0.8202061964422475

Interpretation of the results:

The chi-square statistic value tells us how much the observed frequencies in the contingency table differ from the expected frequencies under the assumption of independence between the groups. A higher chi-square value indicates a larger deviation from independence.

The p-value measures the statistical significance of the chi-square statistic. It indicates the probability of observing such a large or larger chi-square statistic under the assumption of independence. A smaller p-value suggests stronger evidence against the null hypothesis of independence.

If the p-value is less than a chosen significance level (e.g., 0.05), it provides evidence to reject the null hypothesis and suggests that there is a significant association between the groups and the outcomes.

Conversely, if the p-value is greater than the significance level, there is no significant evidence to reject the null hypothesis, and we would conclude that there is no significant association between the groups and the outcomes.

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.

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

Confidence Interval = p̂ ± Z \* √((p̂ \* (1 - p̂)) / n)

Where:

* p̂ is the observed proportion of individuals who smoke in the sample (60/500 = 0.12)
* Z is the Z-score corresponding to the desired confidence level (95% confidence level corresponds to Z = 1.96)
* n is the sample size (500)

import math

# Sample size

n = 500

# Observed proportion of individuals who smoke in the sample

p\_hat = 60 / n

# Z-score for 95% confidence level

z = 1.96

# Calculate the standard error

standard\_error = math.sqrt((p\_hat \* (1 - p\_hat)) / n)

# Calculate the lower and upper bounds of the confidence interval

lower\_bound = p\_hat - z \* standard\_error

upper\_bound = p\_hat + z \* standard\_error

# Display the confidence interval

print(f"Confidence Interval: [{lower\_bound:.4f}, {upper\_bound:.4f}]")

Output:

Confidence Interval: [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.

import scipy.stats as stats

sample\_mean = 75

sample\_std = 12

sample\_size = 50 # Assuming a sample size of 50, but you can change it to your actual sample size

confidence\_level = 0.90

# Calculate the standard error

standard\_error = sample\_std / (sample\_size \*\* 0.5)

# Calculate the margin of error

margin\_of\_error = stats.t.ppf((1 + confidence\_level) / 2, sample\_size - 1) \* standard\_error

# Calculate the lower and upper bounds of the confidence interval

lower\_bound = sample\_mean - margin\_of\_error

upper\_bound = sample\_mean + margin\_of\_error

print(f"Confidence Interval: ({lower\_bound:.2f}, {upper\_bound:.2f})")

Interpreting the results:

The calculated 90% confidence interval is a range of values (lower bound to upper bound) within which we can expect the true population mean to fall with 90% confidence.

In this case, the confidence interval is calculated to be (70.69, 79.31). This means that, based on the sample data, we can be 90% confident that the true population mean lies between 70.69 and 79.31. It's important to note that this is an estimate and there is still a 10% chance that the true population mean is outside of this interval.

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.

import numpy as np

import matplotlib.pyplot as plt

from scipy.stats import chi2

# Set the degrees of freedom

df = 10

# Generate x values

x = np.linspace(0, 30, 500)

# Compute the chi-square probability density function (PDF) for the given degrees of freedom

pdf = chi2.pdf(x, df)

# Create the plot

plt.plot(x, pdf, 'b-', label=f'Chi-square (df={df})')

# Shade the area corresponding to a chi-square statistic of 15

x\_shade = np.linspace(15, 30, 500)

y\_shade = chi2.pdf(x\_shade, df)

plt.fill\_between(x\_shade, y\_shade, color='lightblue', alpha=0.5, label='Area of chi-square statistic 15')

# Label the axes and add a legend

plt.xlabel('x')

plt.ylabel('Probability Density')

plt.legend()

# 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.

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

CI = p̂ ± z \* sqrt((p̂(1 - p̂)) / n)

Where: p̂ is the sample proportion (520/1000 = 0.52 in this case) z is the z-score corresponding to the desired confidence level (for 99% confidence level, z ≈ 2.576) n is the sample size (1000 in this case)

Let's calculate the confidence interval:

CI = 0.52 ± 2.576 \* sqrt((0.52 \* (1 - 0.52)) / 1000)

CI = 0.52 ± 2.576 \* sqrt((0.52 \* 0.48) / 1000)

CI = 0.52 ± 2.576 \* sqrt(0.2496 / 1000)

CI = 0.52 ± 2.576 \* sqrt(0.0002496)

CI = 0.52 ± 2.576 \* 0.015804

CI = 0.52 ± 0.04071

CI ≈ (0.479, 0.561)

Therefore, the 99% confidence interval for the true proportion of people in the population who prefer Coke is approximately (0.479, 0.561).

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.

To conduct a chi-square goodness of fit test for this scenario, we need to compare the observed frequencies with the expected frequencies based on the assumption of a fair coin.

Let's break down the steps to perform the test:

Step 1: State the hypotheses:

Null hypothesis (H0): The coin is fair (not biased towards tails).

Alternative hypothesis (Ha): The coin is biased towards tails.

Step 2: Set the significance level (α): The significance level (α) is given as 0.05 (or 5%).

Step 3: Determine the expected frequencies: Assuming the coin is fair, we would expect an equal chance of getting heads or tails in each flip. Since there are 100 flips, the expected frequency for each outcome would be 50.

Step 4: Calculate the test statistic (chi-square statistic): The chi-square statistic formula for goodness of fit is: χ² = Σ [(Oᵢ - Eᵢ)² / Eᵢ] Where Oᵢ is the observed frequency, Eᵢ is the expected frequency, and Σ denotes the sum across all categories.

In our case, we have two categories (heads and tails). Using the formula, we can calculate the chi-square statistic.

For tails: (Observed frequency) O₁ = 45 (Expected frequency) E₁ = 50

For heads: (Observed frequency) O₂ = 100 - O₁ = 100 - 45 = 55 (Expected frequency) E₂ = 50

Now, we can calculate the chi-square statistic: χ² = [(O₁ - E₁)² / E₁] + [(O₂ - E₂)² / E₂]

Step 5: Determine the degrees of freedom: The degrees of freedom (df) for a goodness of fit test is the number of categories minus 1. In our case, df = 2 - 1 = 1.

Step 6: Determine the critical value: Using the significance level (α) and degrees of freedom (df), we can find the critical value from the chi-square distribution table. For α = 0.05 and df = 1, the critical value is approximately 3.841.

Step 7: Compare the test statistic with the critical value: If the test statistic is greater than the critical value, we reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

Let's calculate the chi-square statistic and compare it with the critical value:

χ² = [(O₁ - E₁)² / E₁] + [(O₂ - E₂)² / E₂]

= [(45 - 50)² / 50] + [(55 - 50)² / 50]

= (5² / 50) + (5² / 50)

= (25 / 50) + (25 / 50)

= 0.5 + 0.5

= 1.0

Since the test statistic (χ² = 1.0) is less than the critical value (3.841), we fail to reject the null hypothesis.

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.

|  |  |  |
| --- | --- | --- |
|  | Lung Cancer: Yes | Lung Cancer: No |
| Smoker | 60 | 140 |
| Non-smoker | 30 | 170 |

Use a significance level of 0.05.

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

Step 1: Set up the hypotheses.

* Null Hypothesis (H₀): There is no association between smoking status and lung cancer diagnosis.
* Alternative Hypothesis (H₁): There is an association between smoking status and lung cancer diagnosis.

Step 2: Define the significance level. The significance level (α) is given as 0.05.

Step 3: Calculate the expected frequencies. To conduct a chi-square test, we need to calculate the expected frequencies for each cell in the contingency table. The expected frequency for each cell is calculated using the formula: Expected Frequency = (Row Total \* Column Total) / Grand Total

The expected frequencies for the given contingency table are as follows:

Lung Cancer: Yes Lung Cancer: No Total

Smoker 60 \* (200/400) 140 \* (200/400) 200 Non-smoker 30 \* (200/400) 170 \* (200/400) 200 Total 90 310 400

Step 4: Calculate the chi-square test statistic. The chi-square test statistic is calculated using the formula: χ² = Σ [(O-E)² / E]

Where:

* = Observed frequency
* E = Expected frequency

Using the observed and expected frequencies from the contingency table, we can calculate the chi-square test statistic.

Lung Cancer: Yes Lung Cancer: No

Smoker 60 140 Non-smoker 30 170

Expected Frequencies: Lung Cancer: Yes Lung Cancer: No Smoker 50 150 Non-smoker 40 160

Now, we can calculate the chi-square test statistic (χ²):

χ² = [(60-50)² / 50] + [(140-150)² / 150] + [(30-40)² / 40] + [(170-160)² / 160]

= (10² / 50) + (-10² / 150) + (-10² / 40) + (10² / 160)

= 2 + 0.6667 + 5 + 0.625

= 8.2917

Step 5: Determine the degrees of freedom. The degrees of freedom (df) for a chi-square test of independence are calculated using the formula: df = (Number of rows - 1) \* (Number of columns - 1)

In this case, the contingency table has 2 rows and 2 columns, so the degrees of freedom are: df = (2-1) \* (2-1) = 1

Step 6: Determine the critical value. To determine the critical value for a chi-square test at a given significance level and degrees of freedom, we refer to a chi-square distribution table or use statistical software. Since the significance level (α) is 0.05 and the degrees of freedom (df) are 1, the critical value is approximately 3.841.

Step 7: Compare the test statistic with the critical value. Compare the calculated chi-square test statistic (8.2917) with the critical value (3.841).

Since the calculated test statistic (8.2917) is greater than the critical value (3.841), we reject the null hypothesis (H₀) and conclude that there is a significant 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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Milk Chocolate | Dark  Chocolate | White Chocolate |
| U.S. (n=500) | 200 | 150 | 150 |
| U.K. (n=500) | 225 | 175 | 100 |

Use a significance level of 0.05.

: Step 1: Set up the contingency table with observed frequencies

U.S. (n=500) 200 150 150 U.K. (n=500) 225 175 100

Step 2: Calculate the expected frequencies under the assumption of independence.

To calculate the expected frequencies, we need to find the row and column totals and use them to determine the expected frequencies.

The row totals are: U.S. (n=500): 200 + 150 + 150 = 500 U.K. (n=500): 225 + 175 + 100 = 500

The column totals are: milk chocolate: 200 + 225 = 425 dark chocolate: 150 + 175 = 325 white chocolate: 150 + 100 = 250

To calculate the expected frequencies, we multiply the row totals by the column totals and divide by the overall sample size:

Expected frequency for U.S. and milk chocolate: (500/1000) \* 425 = 212.5

Expected frequency for U.K. and milk chocolate: (500/1000) \* 425 = 212.5

Expected frequency for U.S. and dark chocolate: (500/1000) \* 325 = 162.5

Expected frequency for U.K. and dark chocolate: (500/1000) \* 325 = 162.5

Expected frequency for U.S. and white chocolate: (500/1000) \* 250 = 125

Expected frequency for U.K. and white chocolate: (500/1000) \* 250 = 125

The expected frequencies are:

U.S. (n=500) 212.5 162.5 125 U.K. (n=500) 212.5 162.5 125

Step 3: Calculate the chi-square statistic.

The chi-square statistic is calculated using the formula:

χ² = Σ [(Oᵢ - Eᵢ)² / Eᵢ]

Where Oᵢ is the observed frequency and Eᵢ is the expected frequency for each cell.

Calculating the chi-square statistic for each cell:

For milk chocolate: χ²₁ = [(200 - 212.5)² / 212.5] + [(225 - 212.5)² / 212.5]

For dark chocolate: χ²₂ = [(150 - 162.5)² / 162.5] + [(175 - 162.5)² / 162.5]

For white chocolate: χ²₃ = [(150 - 125)² / 125] + [(100 - 125)² / 125]

Summing up the values: χ² = χ²₁ + χ²₂ + χ²₃

Calculate the chi-square statistic.

Step 4: Determine the degrees of freedom.

Degrees of freedom (df) can be calculated using the formula:

df = (number of rows - 1) \* (number of columns - 1)

In our case, df = (2 - 1) \* (3 - 1) = 2.

Step 5: Determine the critical value.

Using a significance level of 0.01 and the degrees of freedom (df = 2), we can consult a chi-square distribution table or use a statistical calculator to find the critical chi-square value. In this case, the critical value is approximately 9.210.

Step 6: Compare the calculated chi-square value with the critical value.

If the calculated chi-square value is greater than the critical value, we reject the null hypothesis. Otherwise, we fail to reject the null hypothesis.

Step 7: Interpret the results.

If we reject the null hypothesis, we conclude that there is a significant association between chocolate preference and country of origin. If we fail to reject the null hypothesis, we do not have enough evidence to suggest a significant association.

Performing the calculations, we find that the calculated chi-square value exceeds the critical value. Therefore, we reject the null hypothesis. This indicates 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.

To conduct a hypothesis test to determine if the population mean is significantly different from 70, we can use a one-sample t-test. Here are the steps to perform the test:

Step 1: State the null hypothesis (H0) and alternative hypothesis (H1):

* Null hypothesis (H0): The population mean is equal to 70 (μ = 70)
* Alternative hypothesis (H1): The population mean is significantly different from 70 (μ ≠ 70)

Step 2: Set the significance level (α):

* The significance level is given as 0.05, which corresponds to a 5% level of significance.

Step 3: Calculate the test statistic:

* We can calculate the test statistic using the formula: t = (sample mean - hypothesized mean) / (sample standard deviation / √sample size)
* In this case:

Sample mean (x̄) = 72

Hypothesized mean (μ) = 70

Sample standard deviation (s) = 10

Sample size (n) = 30

* Plugging in these values, we get:

t = (72 - 70) / (10 / √30)

t ≈ 2.68

Step 4: Determine the critical value:

* Since this is a two-tailed test (μ ≠ 70), we need to find the critical value(s) based on the significance level and degrees of freedom.
* With a sample size of 30, the degrees of freedom are (n - 1) = 29.
* Looking up the critical t-values in the t-distribution table (or using software), we find the critical values for a 5% significance level with 29 degrees of freedom are approximately ±2.045.

Step 5: Make a decision:

* If the test statistic falls within the critical region, we reject the null hypothesis.
* In this case, the absolute value of the test statistic (2.68) is greater than the critical value (2.045).
* Thus, we reject the null hypothesis.

Step 6: Draw a conclusion:

* Since we rejected the null hypothesis, we can conclude that there is sufficient evidence to suggest that the population mean is significantly different from 70 at the 5% level of significance.