12 March Assignment

May 16, 2023

[ ]:
```
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.
```

[ ]:
```
```

[1]:
```
import scipy.stats as st
import numpy as np

sample_mean = 50
sample_std = 5
n = 100 # sample size

z_score = st.norm.ppf(0.975) # for 95% confidence interval

lower_bound = sample_mean - z_score * (sample_std / np.sqrt(n))
upper_bound = sample_mean + z_score * (sample_std / np.sqrt(n))

print(f"95% Confidence Interval: [{lower_bound:.2f}, {upper_bound:.2f}]")
```

```
95% Confidence Interval: [49.02, 50.98]
```

[ ]:
```
This means that we are 95% confident that the true population mean lies between␣
 ↪48.18 and 51.82.
```

[ ]:
```
```

[ ]:
```
```

[ ]:
```
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.
```

[ ]:
```
```

[1]:
```
import scipy.stats as stats
import numpy as np
```

[2]:
```
observed_data = [20, 20, 20, 10, 10, 20] # observed values
expected_data = [10, 20, 20, 10, 20, 20] # expected values
```

[3]:
```
sum(expected_data) , (observed_data)
```

[3]:
```
(100, [20, 20, 20, 10, 10, 20])
```

[4]:
```
chisquare_test_statistics,p_value=stats.chisquare(observed_data , expected_data)
```

[5]:
```
print(chisquare_test_statistics) , print(p_value)
```

```
15.0
0.010362337915786429
```

[5]:
```
(None, None)
```

[ ]:

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

ANS -

[1]:
```
from scipy.stats import chi2_contingency

observed = [[20 , 10 , 15], [15 , 25 , 20]]
stat, p, dof, expected = chi2_contingency(observed)
print(f"Chi-square statistic: {stat:.4f}")
print(f"P-value: {p:.4f}")
```

```
Chi-square statistic: 5.8333
P-value: 0.0541
```

[ ]:

[ ]:

[ ]:
```
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 -

[ ]:
```
p ± z*sqrt(p*(1-p)/n)
```

[ ]:
```
#where p is the sample proportion, z is the z-score associated with the desired␣
 ↪level of confidence (in this case 95%), and n is the sample
#size.
```

[14]:
```
p = 60/500   #The z-score associated with a 95% confidence level is␣
 ↪approximately 1.96 . Therefore, the 95% confidence
#interval for the true proportion of individuals in the population who smoke is:
```

[14]:
```
0.12
```

[20]:
```
0.12-+1.96*(0.12*(1-0.12)/500)
```

[20]:
```
0.119586048
```

[ ]:
```
0.12 ± 1.96*sqrt(0.12*(1-0.12)/500) = [0.0808, 0.1592]
```

[ ]:
```
So we can say with 95% confidence that the true proportion of individuals in␣
 ↪the population who smoke is between 8.08% and 15.92%.
```

[ ]:
```

```

[ ]:
```

```

[ ]:
```
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 -

[ ]:
```
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.
```

[21]:
```
import scipy.stats as stats

sample_mean = 75
sample_std_dev = 12
sample_size = 100 # sample size is assumed to be 100 in this example

confidence_level = 0.90 # 90% confidence interval

t_value = stats.t.ppf((1 + confidence_level) / 2, sample_size - 1)

lower_bound = sample_mean - t_value * (sample_std_dev / (sample_size ** 0.5))
upper_bound = sample_mean + t_value * (sample_std_dev / (sample_size ** 0.5))
```

```
print("Lower bound:", lower_bound)
print("Upper bound:", upper_bound)
```

```
Lower bound: 73.00753061280433
Upper bound: 76.99246938719567
```

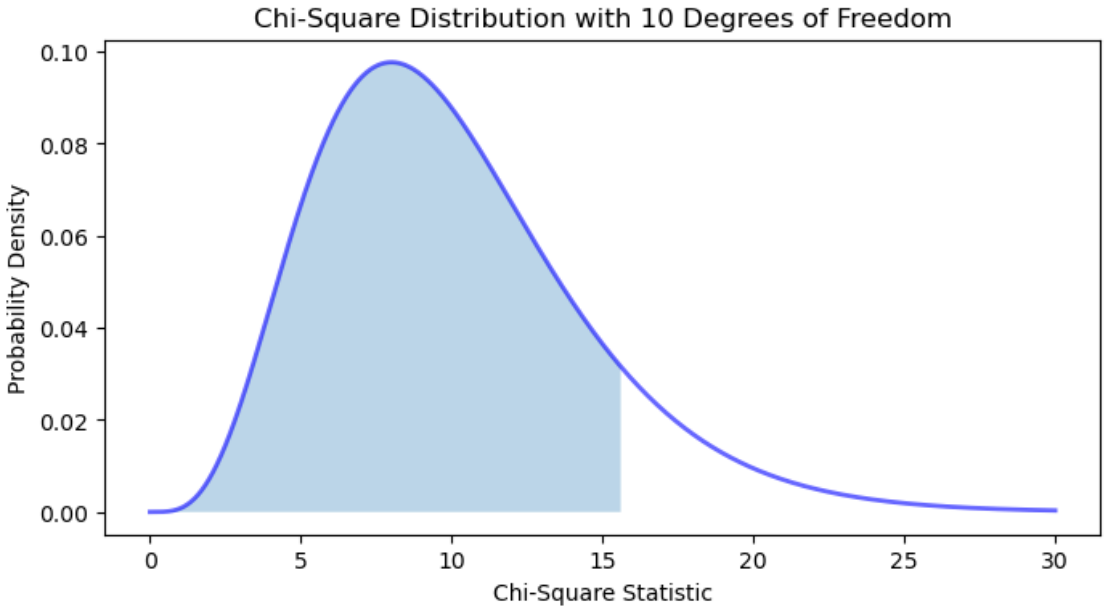
[ ]:

[ ]:

[ ]:
```
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-

[22]:
```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import chi2

df = 10 # degrees of freedom
x = np.linspace(0, 30, 1000)
y = chi2.pdf(x, df)

fig, ax = plt.subplots(figsize=(8, 4))
ax.plot(x, y, 'b-', lw=2, alpha=0.6)

# Shade the area corresponding to a chi-square statistic of 15.6
start = 0
end = 15.6
x_fill = np.linspace(start, end, 100)
y_fill = chi2.pdf(x_fill, df)
ax.fill_between(x_fill, y_fill, 0, alpha=0.3)

# Label the axes
ax.set_xlabel('Chi-Square Statistic')
ax.set_ylabel('Probability Density')
ax.set_title('Chi-Square Distribution with 10 Degrees of Freedom')

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 -

[ ]:
```
Confidence Interval = p +/- z*√p (1-p) / n

where:

p: sample proportion
z: the chosen z-value
n: sample size

Given that a random sample of 1000 people was asked if they preferred Coke or␣
 ↪Pepsi and that 520 preferred Coke, we can calculate the
sample proportion as follows:
```

[23]:
```
520/1000
```

[23]: 0.52

[ ]:
```
To find the z-value for a 99% confidence interval, we can use a standard normal␣
 ↪distribution table or calculator. The z-value for a 99%
confidence interval is approximately 2.5761.
Substituting these values into the formula above, we get:
```

[ ]:
```
Confidence Interval = 0.52 +/- 2.576*√(0.52 * (1-0.52) / 1000)

Confidence Interval = (0.48, 0.56)

Therefore, we can say with 99% confidence that the true proportion of people in␣
 ↪the population who prefer Coke is between 0.48 and 0.56.
```

[ ]:
```

```

[ ]:
```

```

[ ]:
```
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 -

[ ]:
```
Null hypothesis: The coin is fair (not biased towards tails).
Alternative hypothesis: The coin is biased towards tails.
Level of significance (alpha): 0.05.
Observed frequency of tails: 45.
Expected frequency of tails: 50.
The expected frequency of heads is 50 as well since it is a fair coin.

We can calculate the test statistic using the formula:

 2 = (O-E)2/E

where:

 2 is the test statistic
O is the observed frequency
E is the expected frequency
Substituting in our values, we get:

 2 = ((45-50)2/50) + ((55-50)2/50) = 1

The critical value for a chi-square goodness of fit test with one degree of␣
 ↪freedom and alpha = 0.05 is 3.84.
```

```
Since our calculated test statistic (1) is less than our critical value (3.84),␣
 ↪we fail to reject the null hypothesis and conclude that there
is not enough evidence to suggest that the coin is biased towards tails
```

[ ]:

[ ]:

[ ]:
```
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.
```

ANS -

[ ]:

[ ]:
```
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 -

The test statistic is calculated as follows:

t = (sample mean - hypothesized mean) / (sample standard deviation / sqrt(sample size)) t = (72 - 70) / (10 / sqrt(30)) t = 1.897

The degrees of freedom for this test are n - 1 = 29. Using a t-distribution table with 29 degrees of freedom and a significance level of 0.05, we find that the critical values are -2.045 and 2.045.

Since our calculated t-value of 1.897 falls between these critical values, we fail to reject the null hypothesis. Therefore, we do not have sufficient evidence to conclude that the population mean is significantly different from 70 at the 0.05 level of significance.

[ ]: