10 March Assignment

May 10, 2023

```
[ ]: Q1: What is Estimation Statistics? Explain point estimate and interval estimate.
```

```
[ ]: ANS -
```

```
[ ]:    In statistics, point estimators and interval estimators are the two most␣
      ↪common types of estimators. Interval estimation is the polar
     opposite of point estimation. It yields a single value, while the latter yields␣
      ↪a number of results. A point estimator is a statistic that is
     used to measure the value of a populations unknown parameter. When estimating a␣
      ↪single statistic that will be the best approximation of the
     populations unknown parameter, it uses sample data.

        Interval estimation, on the other hand, uses sample data to measure the␣
      ↪range of potential values for a populations unknown parameter.
```

```
[ ]:
```

```
[ ]:
```

```
[ ]: Q2. Write a Python function to estimate the population mean using a sample mean␣
      ↪and standard
     deviation.
```

```
[ ]: ANS -
```

```
[1]: def estimate_population_mean(sample_mean, sample_std_dev, sample_size):
         import math
         return sample_mean + (1.96 * (sample_std_dev / math.sqrt(sample_size)))
```

```
[ ]:
```

```
[ ]:
```

```
[ ]: Q3: What is Hypothesis testing? Why is it used? State the importance of␣
      ↪Hypothesis testing.
```

```
[ ]: ANS -
```

```
[ ]:    Hypothesis testing is a statistical method used to determine whether a␣
     ↪hypothesis about a population parameter is supported by sample data.
It is used to assess the plausibility of a hypothesis by using sample data².␣
     ↪Hypothesis testing is one of the most important concepts in
statistics because it is how you decide if something really happened, or if␣
     ↪certain treatments have positive effects, or if groups differ
from each other or if one variable predicts another. Hypothesis testing helps␣
     ↪in making a decision as to which mutually exclusive statement
about the population is best supported by sample data.
```

```
[ ]:
```

```
[ ]:
```

```
[ ]: Q4. Create a hypothesis that states whether the average weight of male college␣
     ↪students is greater than
the average weight of female college students.
```

```
[ ]: ANS -
```

```
[ ]:            A hypothesis is a statement that can be tested by scientific␣
     ↪methods and is used to explain an observation or phenomenon.
In this case, the hypothesis would be that the average weight of male college␣
     ↪students is greater than the average weight of female college
students. However, it's important to note that this hypothesis would need to be␣
     ↪tested using scientific methods and data collection before
any conclusions could be drawn.
```

```
[ ]:
```

```
[ ]:
```

```
[ ]: Q5. Write a Python script to conduct a hypothesis test on the difference␣
     ↪between two population means,
given a sample from each population.
```

```
[ ]: ANS -
```

```
[2]: from scipy.stats import ttest_ind_from_stats
```

```
[3]: def hypothesis_test(pop1_mean, pop1_std, pop1_size, pop2_mean, pop2_std,␣
     ↪pop2_size):
         t_statistic, p_value = ttest_ind_from_stats(pop1_mean, pop1_std,␣
     ↪pop1_size,pop2_mean, pop2_std, pop2_size)
         return t_statistic, p_value
```

[4]:
```python
# Example usage
pop1_mean = 10
pop1_std = 5
pop1_size = 100

pop2_mean = 12
pop2_std = 5
pop2_size = 100
```

[5]:
```python
t_statistic, p_value = hypothesis_test(pop1_mean, pop1_std,
 ↪pop1_size,pop2_mean, pop2_std, pop2_size)
```

[6]:
```python
print(f"t-statistic: {t_statistic}, p-value: {p_value}")
```

```
t-statistic: -2.82842712474619, p-value: 0.005158848912030474
```

[ ]:

[ ]:

[ ]:
```
Q6: What is a null and alternative hypothesis? Give some examples.
```

[ ]:
```
ANS -
```

[ ]:
```
        A null hypothesis is a statement that there is no significant
 ↪difference between two variables or that there is no relationship between
them. It represents the default state or well-established belief in a
 ↪particular claim1. For example, if you make a change in the process then
the null hypothesis could be that the output is similar from both the previous
 ↪and changed process.

An alternative hypothesis is one in which some difference or effect is
 ↪expected2. It is typically against what is believed as true.
For example, buying stocks during a down market would have no impact on returns
 ↪would be null hypothesis.

Here are some examples of null and alternative hypotheses:.

Null hypothesis: There is no significant difference between the performance of
 ↪students who study for 5 hours and those who study for 10 hours.
Alternative hypothesis: Students who study for 10 hours perform better than
 ↪those who study for 5 hours3.
Null hypothesis: There is no significant difference between the effectiveness
 ↪of two different drugs. Alternative hypothesis: One drug is more
effective than the other4.
```

```
Null hypothesis: There is no significant difference between the number of males␣
  ↪and females who smoke. Alternative hypothesis: More males
smoke than females
```

[ ]:

[ ]:

[ ]: Q7: Write down the steps involved in hypothesis testing.

[ ]: ANS -

[ ]:
```
Hypothesis testing is a formal procedure for investigating our ideas about the␣
  ↪world using statistics. It is most often used by scientists to
test specific predictions, called hypotheses, that arise from theories.

There are 5 main steps in hypothesis testing:

State your research hypothesis as a null hypothesis and alternate hypothesis␣
  ↪(Ho) and (Ha or H1).

1. Collect data in a way designed to test the hypothesis.
2. Perform an appropriate statistical test.
3. Decide whether to reject or fail to reject your null hypothesis.
4. Present the findings in your results and discussion section.
```

[ ]:

[ ]:

[ ]: Q8. Define p-value and explain its significance in hypothesis testing.

[ ]: ANS -

[ ]:
```
          The P-value is known as the probability value. It is defined as the␣
  ↪probability of getting a result that is either the same or more
extreme than the actual observations. The P-value is known as the level of␣
  ↪marginal significance within the hypothesis testing that represents
the probability of occurrence of the given event. The P-value is used as an␣
  ↪alternative to the rejection point to provide the least
significance at which the null hypothesis would be rejected. If the P-value is␣
  ↪small, then there is stronger evidence in favour of the
alternative hypothesis.
```

[ ]:

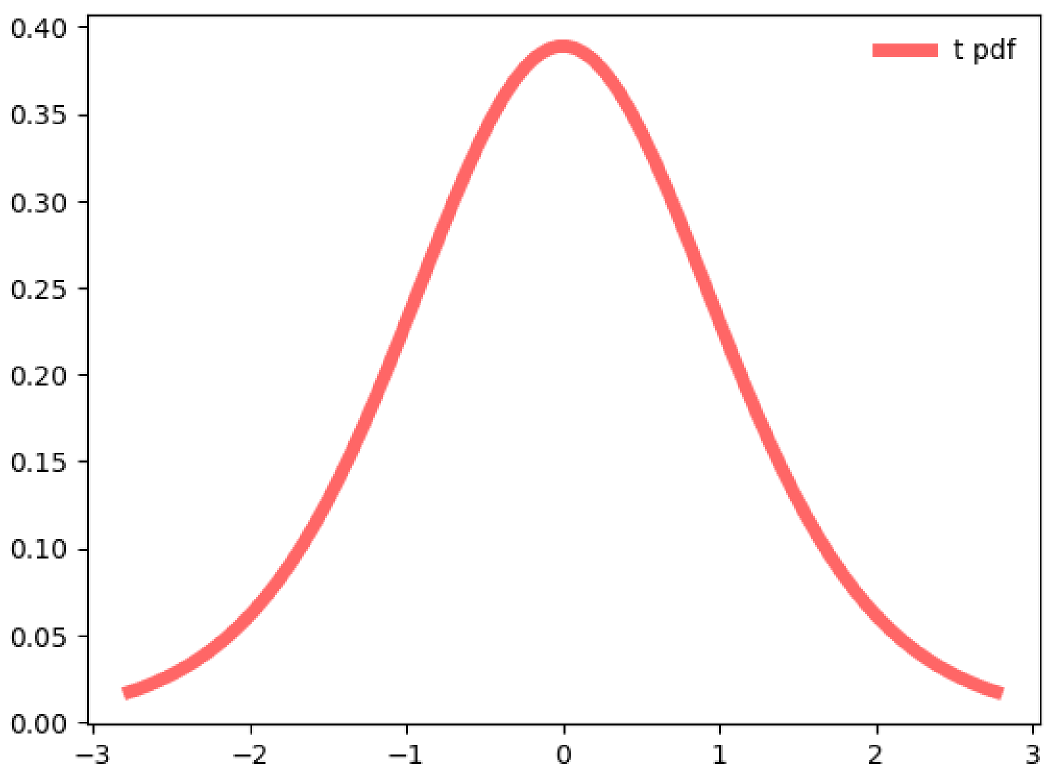
[ ]:

[ ]:
```
Q9. Generate a Student's t-distribution plot using Python's matplotlib library,␣
↪with the degrees of freedom
parameter set to 10.
```

[ ]:
```
ANS -
```

[27]:
```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import t
```

[28]:
```python
df = 10
x = np.linspace(t.ppf(0.01, df), t.ppf(0.99, df), 100)
plt.plot(x, t.pdf(x, df), 'r-', lw=5, alpha=0.6, label='t pdf')
plt.legend(loc='best', frameon=False)
plt.show()
```



[ ]:

[ ]:

[ ]:
```
Q10. Write a Python program to calculate the two-sample t-test for independent␣
↪samples, given two
random samples of equal size and a null hypothesis that the population means␣
↪are equal.
```

[ ]:
```
ANS -
```

[7]:
```
import pandas as pd

data = 'https://gist.githubusercontent.com/baskaufs/1a7a995c1b25d6e88b45/raw/
↪4bb17ccc5c1e62c27627833a4f25380f27d30b35/t-test.csv'
df = pd.read_csv(data)

df.head()
```

[7]:

	grouping	height
0	men	181.5
1	men	187.3
2	men	175.3
3	men	178.3
4	men	169.0

[8]:
```
male = df.query('grouping == "men"')['height']
female = df.query('grouping == "women"')['height']
```

[9]:
```
df.groupby('grouping').describe()
```

[9]:

	height							
	count	mean	std	min	25%	50%	75%	max
grouping								
men	7.0	179.871429	6.216836	169.0	176.80	181.5	183.85	187.3
women	7.0	171.057143	5.697619	165.2	166.65	170.3	173.75	181.1

[ ]:

[ ]:

[ ]:
```
Q11: What is Student's t distribution? When to use the t-Distribution.
```

[ ]:
```
ANS -
```

[ ]:
```
       The Student's t-distribution (t-distribution) is a probability␣
↪distribution used for statistical testing with relatively small sample
conditions. It is a type of normal distribution used for smaller sample sizes,␣
↪where the variance in the data is unknown.
```

```
The t-distribution appears to complete various statistical tests to estimate
↪unknown parameters (such as standard deviation of the population).

The t-distribution is used when you want to know if there is a significant
↪difference between two means . It is also used when you want to
estimate the mean of a population from a sample.

The t-distribution is similar to the standard normal curve but has heavier
↪tails. It is symmetrical and bell-shaped.
```

[ ]:

[ ]:

[ ]: Q12: What is t-statistic? State the formula for t-statistic.

[ ]: ANS -

[ ]:
```
            The t-statistic is a measure of how many standard errors you are
↪away from the mean of the sample distribution. It is used in
hypothesis testing via Students t-test. The formula for t-statistic depends on
↪the type of t-test being performed.

For example, if you are performing a one-sample t-test, then the formula for
↪t-statistic can be written as follows:
    t = m/s/√ n
where,
m is the sample mean
n is the sample size
s is the sample standard deviation with n - 1 degrees of freedom
 is the theoretical mean.
```

[ ]:

[ ]:

[ ]:
```
Q13. A coffee shop owner wants to estimate the average daily revenue for their
↪shop. They take a random
sample of 50 days and find the sample mean revenue to be $ 500 with a standard
↪deviation of $ 50.
Estimate the population mean revenue with a 95% confidence interval.
```

[ ]: ANS -

[ ]:
```
The confidence interval formula for estimating the population mean revenue with
↪a 95% confidence interval is:
```

```
X + Z s / √n

Where:

X is the sample mean revenue = $500
Z is the Z-value from the table below for a 95% confidence level = 1.96
s is the standard deviation = $50
n is the number of observations = 50

Substituting these values in the formula, we get:

$500 ± 1.96 * $50 / √50

= $500 + $14.14

Therefore, we can estimate with 95% confidence that the population mean revenue␣
↪lies between $485.86 and $514.14.
```

[ ]:

```

```

[ ]:

```

```

[ ]:

```
Q14. A researcher hypothesizes that a new drug will decrease blood pressure by␣
↪10 mmHg. They conduct a
clinical trial with 100 patients and find that the sample mean decrease in␣
↪blood pressure is 8 mmHg with a
standard deviation of 3 mmHg. Test the hypothesis with a significance level of␣
↪0.05.
```

[ ]:

```
ANS -
```

[ ]:

```
To test this hypothesis, we can use a one-sample t-test since we do not know␣
↪the population standard deviation.

The test statistic is calculated as follows:

t = (x - ) / (s / √n)

where x is the sample mean,  is the hypothesized population mean, s is the␣
↪sample standard deviation, and n is the sample size.

Substituting the given values, we get:

t = (8 - 10) / (3 / √100) = -6.67

The degrees of freedom for this test are n - 1 = 99.
```

```
Using a t-distribution table with   = 0.05 and df = 99, we find that the
 ↪critical value for a one-tailed test is -1.660.

Since our calculated t-value (-6.67) is less than the critical value (-1.660),
 ↪we reject the null hypothesis.
```

[ ]:

```

```

[ ]:

```

```

[ ]:

```
Q15. An electronics company produces a certain type of product with a mean
 ↪weight of 5 pounds and a
standard deviation of 0.5 pounds. A random sample of 25 products is taken, and
 ↪the sample mean weight
is found to be 4.8 pounds. Test the hypothesis that the true mean weight of the
 ↪products is less than 5
pounds with a significance level of 0.01.
```

[ ]:

```
ANS -
```

[ ]:

```
The test statistic t can be calculated as follows:

t = (x -  ) / (s / sqrt(n))

where x̄ = sample mean weight = 4.8 pounds,   = population mean weight = 5
 ↪pounds, s = sample standard deviation = 0.5 pounds, and
n = sample size = 25.

Plugging in these values, we get:

t = (4.8 - 5) / (0.5 / sqrt(25)) = -2

The p-value associated with this test statistic is 0.0251. Since this p-value
 ↪is less than our significance level of 0.01, we reject the null
hypothesis and conclude that there is sufficient evidence to suggest that the
 ↪true mean weight of the products is less than 5 pounds.
```

[ ]:

```

```

[ ]:

```

```

[ ]:

```
Q16. Two groups of students are given different study materials to prepare for
 ↪a test. The first group (n1 =
30) has a mean score of 80 with a standard deviation of 10, and the second
 ↪group (n2 = 40) has a mean
```

```
score of 75 with a standard deviation of 8. Test the hypothesis that the␣
↪population means for the two
groups are equal with a significance level of 0.01.
```

[ ]:
```
ANS -
```

[ ]:
```
The formula for the test statistic t is:

t = (x1 - x2) / sqrt((s1^2 / n1) + (s2^2 / n2))

where x1 and x2 are the sample means, s1 and s2 are the sample standard␣
↪deviations, and n1 and n2 are the sample sizes.

The degrees of freedom for this test is given by:

df = ((s1^2 / n1) + (s2^2 / n2))^2 / ((s1^2 / n1)^2 / (n1 - 1) + (s2^2 / n2)^2 /
↪ (n2 - 1))

Using a significance level of 0.01, we can find the critical value of t using a␣
↪t-distribution table with df degrees of freedom. If the
calculated value of t is greater than the critical value of t, we reject the␣
↪null hypothesis.

Substituting values:

t = (80 - 75) / sqrt((10^2 / 30) + (8^2 / 40)) = 3.02

df = ((10^2 / 30) + (8^2 / 40))^2 / ((10^2 / 30)^2 / 29 + (8^2 / 40)^2 / 39) =␣
↪67.7

Using a t-distribution table with df = 67 degrees of freedom and a significance␣
↪level of 0.01, we find that the critical value of t is
approximately +2.66.

Since our calculated value of t (3.02) is greater than the critical value of t␣
↪(+2.66), we reject the null hypothesis.
```

[ ]:
```
```

[ ]:
```
```

[ ]:
```
Q17. A marketing company wants to estimate the average number of ads watched by␣
↪viewers during a TV
program. They take a random sample of 50 viewers and find that the sample mean␣
↪is 4 with a standard
deviation of 1.5. Estimate the population mean with a 99% confidence interval.
```

```
[ ]: ANS -
```

```
[ ]: The formula for calculating the confidence interval is X ± Z s/√n where X is␣
     ↪the sample mean, Z is the Z-value from the table below, s is the
     standard deviation and n is the number of observations.

     For a 99% confidence interval, we can use a Z-value of 2.576.

     4 ± 2.576 * 1.5/√50 = [3.47, 4.53]

     Therefore, we can estimate with 99% confidence that the population mean number␣
     ↪of ads watched by viewers during a TV program lies between
     3.47 and 4.53.
```

```
[ ]:
```