Statistical and Mathematical Methods for Data Analysis

Dr. Faisal Bukhari

Punjab University College of Information Technology (PUCIT)

Textbooks

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

References

- **Probability & Statistics for Engineers & Scientists**, Ninth edition, Ronald E. Walpole, Raymond H. Myer
- *https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_1samp.html*
- *https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html*
- *https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html*

These notes contain material from the above resources.

Is σ is known ?

Yes

If either the population is normally distributed or $n \geq 30$, then use the use the standard normal distribution or Z-test

No

If either the population is normally distributed or $n \geq 30$, then use the *t*-distribution or t-test

Inferences on a Population Mean

- Inference methods on a population mean based upon the *t*-procedure are appropriate for large **sample sizes $n \geq 30$** and also for **small sample sizes** as long as the data can reasonably be taken to be **approximately normally distributed**.

- **Nonparametric techniques** can be employed for **small sample sizes with data** that are clearly **not normally distributed.**

- In some circumstances an experimenter may wish to use a **"known"** value of the **population standard deviation σ** in place of the **sample standard deviation s**. In this case, the **standard normal distribution Z** is used.

Independent and Dependent Samples.

- Two samples are **independent** if the sample values selected from **one population** are **not related to or somehow paired or matched** with the sample values selected from the other population.

- Two samples are **dependent** (or consist of **matched pairs**) if the members of one sample can be used to determine the members of the other sample. [Samples consisting of **matched pairs** (such as husband wife data) are **dependent**.

- In addition to **matched pairs of sample data**, **dependence** could also occur with samples related **through associations** such as **family members**.]

Confidence Interval for $\mu_D = \mu_1 - \mu_2$ for **Paired Observations**

If $\bar{d}$ and s_d are the **mean** and **standard deviation**, respectively, of the normally distributed differences of n **random pairs of measurements**, a $100(1 - \alpha)\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is

$$\bar{d} - t_{(\alpha/2, n-1)} \frac{s_d}{\sqrt{n}} < \mu_d < \bar{d} + t_{(\alpha/2, n-1)} \frac{s_d}{\sqrt{n}}$$

Where,

$$s_d = \sqrt{\frac{\sum(d - \bar{d})^2}{n-1}} \text{ OR } s_d = \sqrt{\frac{1}{n(n-1)}\left\{n \sum_{i=1}^{n} d^2{}_i - \left(\sum_{i=1}^{n} d^2{}_i\right)^2\right\}}$$

$$s_d^2 = \frac{\sum(d - \bar{d})^2}{n-1} \text{ OR } s_d^2 = \frac{1}{n(n-1)}\left\{n \sum_{i=1}^{n} d^2{}_i - \left(\sum_{i=1}^{n} d^2{}_i\right)^2\right\}$$

$$d_i = x_{1i} - x_{2j} \text{ OR } d_i = x_{2i} - x_{1j}, \ \bar{d} = \frac{\sum_{i=1}^{n} di}{n}$$

H_0	Value of Test Statistic	H_1	Critical Region
$\mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$; σ known	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$; $v = n - 1$, σ unknown	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$; σ_1 and σ_2 known	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$; $v = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ but unknown, $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t' = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; $v = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$, $\sigma_1 \neq \sigma_2$ and unknown	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t' < -t_\alpha$ $t' > t_\alpha$ $t' < -t_{\alpha/2}$ or $t' > t_{\alpha/2}$
$\mu_D = d_0$ paired observations	$t = \dfrac{\overline{d} - d_0}{s_d/\sqrt{n}}$; $v = n - 1$	$\mu_D < d_0$ $\mu_D > d_0$ $\mu_D \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

Testing Hypothesis about Paired Observation

a) H_o: $\mu_d = 0$

H_1:$\mu_d < 0$ (One tailed test)

b) H_o: $\mu_d = 0$

H_1:$\mu_d > 0$ (One tailed test)

c) H_o: $\mu_d = 0$

H_1:$\mu_d \neq 0$ (Two tailed test)

Test statistic:

$t_{cal} = \dfrac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$,

Where $d_i = x_{1i} - x_{2j}$ OR $d_i = x_{2i} - x_{1j}$

$$\bar{d} = \frac{\sum_{i=1}^{n} \mathrm{di}}{n}$$

$$s_d = \sqrt{\frac{\sum(\mathrm{d} - \bar{d})^2}{n-1}} \text{ OR}$$

$$s_d = \sqrt{\frac{1}{n(n-1)}\{n \sum_{i=1}^{n} \mathrm{d}^2{}_i - (\sum_{i=1}^{n} \mathrm{d}_i)^2\}}$$

The *t* Test for Dependent Samples: An Example

Eight individuals indicated their attitudes toward socialized medicine before and after listening to a pro–socialized medicine lecture. Attitudes were assessed on a scale from 1 to 7, with higher scores indicating more positive attitudes. The attitudes before and after listening to the lecture were as indicated in the second and third columns of the table. Test for a relationship between the time of assessment and attitudes toward socialized medicine using a correlated groups *t* test.

Individual	Before speech	After speech
1	3	6
2	4	6
3	3	3
4	5	7
5	2	4
6	5	6
7	3	7
8	4	6

Solution

$\mu_D = 0$ (Population mean)

$n = 8$ (Sample size)

$\alpha = 0.05$ (Level of significance)

$\bar{d} = ?$

$s_d = ?$

1. We state our hypothesis as:

$H_o: \mu_d = 0$

$H_1: \mu_d \neq 0$ (Two tailed test)

2. The level of significance is set $\alpha = 0.05$

3. Test statistic to be used is

$$t_{cal} = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

4. Calculations:

Before speech	After speech	$d_i = x_{2i} - x_{1j}$	$d^2{}_i$
3	6	3	9
4	6	2	4
3	3	0	0
5	7	2	4
2	4	2	4
5	6	1	1
3	7	4	16
4	6	2	4
Sum		$\sum_{i=1}^{n}$ **di= 16**	$\sum_{i=1}^{n} \mathbf{d^2{}_i = 42}$

$$\bar{d} = \frac{\sum_{i=1}^{n} \mathrm{di}}{n} = 16/2 = \mathbf{2}$$

$$s_d = \sqrt{\frac{1}{n(n-1)}\left\{n\sum_{i=1}^{n} d^2{}_i - \left(\sum_{i=1}^{n} d_i\right)^2\right\}}$$

$$s_d = \sqrt{\frac{1}{8(8-1)}\left\{8(42) - (16)^2\right\}} = \sqrt{\frac{80}{8(8-1)}} = 1.1952$$

$$\mathbf{t_{cal}} = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = t_{cal} = \frac{2 - 0}{\frac{1.1952}{\sqrt{8}}} = \frac{2}{0.4226}$$

$|\mathbf{t_{cal}}| = 4.7326$

5. **Critical region:**

$|\mathbf{t_{cal}}| > \mathbf{t_{tab}}$**, where** $\mathbf{t_{tab} = t_{(\alpha/2,\, n-1)}}$

Where $t_{tab} = t_{(\alpha/2,\, n-1)} = t_{(0.0250,\, 7)} = 2.365$

6. **Conclusion:** Since calculated value of t_{cal} is greater than t_{tab}, so we reject H_O

Interpret your results.

After the **pro-socialized medicine lecture**, individuals' attitudes toward **socialized medicine** were significantly more positive than before the lecture.

DataFrame in Python

What is a DataFrame?

A Pandas DataFrame is a **2 dimensional data structure**, like a 2 dimensional array, or a table with rows and columns.

DataFrame in Python

```
import pandas as pd
data = {
  "calories": [420, 380, 390],
  "duration": [50, 40, 45]
}

#load data into a DataFrame object:
df = pd.DataFrame(data)
print(df)
```

calories	duration
420	50
380	40
390	45

```
import pandas as pd
from scipy import stats
# Create a DataFrame with the provided data
data = pd.DataFrame({
    'Before speech': [3, 4, 3, 5, 2, 5, 3, 4],
    'After speech': [6, 6, 3, 7, 4, 6, 7, 6]
})
# Calculate the differences i.e., d_i = x_2i - x_1i
differences = data['After speech'] - data['Before speech']
# Compute mean and standard deviation of differences d̄ and s_d
mean_diff = differences.mean()
std_diff = differences.std(ddof=1)
# Use ddof=1 for sample standard deviation
```

```
# Calculate t-statistic
```
$t_{cal} = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$

```
n = len(differences)
t_statistic = mean_diff / (std_diff / (n**0.5))

# Degrees of freedom
df = n - 1
# Calculate the p-value
p_value = 2 * (1 - stats.t.cdf(abs(t_statistic), df))
# Set the significance level (alpha)
alpha = 0.05
# Make a decision i.e., P-value ≤ α
if p_value <= alpha:
    conclusion = "Reject the null hypothesis"
else:
    conclusion = "Fail to reject the null hypothesis"
```

```python
# Print the results
print("Mean Difference:", mean_diff)
print("Standard Deviation of Differences:",
std_diff)
print("t-statistic:", t_statistic)
print("Degrees of Freedom:", df)
print("p-value:", p_value)
print("Conclusion:", conclusion)
```

Mean Difference: 1.25

Standard Deviation of Differences: 1.479019945774904

t-statistic: 3.3721277801018267

Degrees of Freedom: 7

p-value: 0.009308586649079443

Conclusion: Reject the null hypothesis

Conclusion: Reject the null hypothesis **($p < \alpha$)**

Since the p-value (0.0093) is less than the chosen significance level ($\alpha = 0.05$), we reject the null hypothesis.

There is sufficient evidence to conclude that there is a significant difference between the "Before speech" and "After speech" scores

Scipy

The scipy.stats is the SciPy sub-package. It is mainly used for probabilistic distributions and statistical operations.

1. The ttest_1samp function from scipy.stats calculates the **t-statistic** and corresponding **p-value**.

2. The ttest_ind function from scipy.stats calculates the **t-statistic** and corresponding **p-value** for this two-sample **t-test**.

3. The ttest_rel function from scipy.stats calculates the **t-statistic** and corresponding **p-value** for this **paired t-test**.

If the **p-value** is **less than** the chosen **significance level (alpha),** we reject the null hypothesis.

Independent One-Sample T-Test

scipy.stats.ttest_1samp(*a*, *popmean*, *axis=0*, *nan_policy='propagate'*, *alternative='two-sided'*, *, *keepdims=False*)**

Or confined form

`ttest_1samp(a, popmean, axis=0, alternative='two-sided')`

One of the important parameter:

alternative*{‘two-sided’, ‘less’, ‘greater’}, optional*

*For detail signature of the method **ttest_1samp** refer to*

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_1samp.html

Independent One-Sample T-Test

```
import scipy.stats as stats
# Sample data
data = [28, 32, 35, 30, 25, 29, 27, 32, 34, 31]
# Define the null hypothesis value
null_mean = 30
# Perform a one-sample t-test
t_statistic, p_value = stats.ttest_1samp(data, null_mean)
# Set the significance level (alpha)
alpha = 0.05
# Print the results
print("Sample Mean:", sum(data) / len(data))
print("t-statistic:", t_statistic)
print("p-value:", p_value)
# Make a decision
if p_value < alpha:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

Independent Two Sample T-Test

```
scipy.stats.ttest_ind(a, b, axis=0, equal_var=True,
nan_policy='propagate', permutations=None,
random_state=None, alternative='two-sided', trim=0, *,
keepdims=False)
```

Or confined form

```
stats.ttest_ind(group1, group2)
```

One of the important parameter:

alternative*{'two-sided', 'less', 'greater'} , optional*

For detail signature of the method `ttest_ind` *refer to*

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

Independent Two-Sample T-Test

```
import numpy as np
from scipy import stats

# Sample data for two groups
group1 = np.array([85, 90, 88, 92, 78])
group2 = np.array([79, 82, 85, 88, 90])

# Perform independent two-sample t-test
t_statistic, p_value = stats.ttest_ind(group1, group2)
# Define significance level (alpha)
alpha = 0.05
# Compare p-value to alpha
if p_value < alpha:
    print(f"p-value ({p_value}) is less than alpha
({alpha}). Reject the null hypothesis.")
else:
    print(f"p-value ({p_value}) is greater than or equal to
alpha ({alpha}). Fail to reject the null hypothesis.")
```

Independent Two-Sample Paired T-Test

scipy.stats.ttest_rel(a, b, axis=0, nan_policy='propagate', alternative='two-sided', *, keepdims=False)[source]

Calculate the t-test on TWO RELATED samples of scores, a and b.

Or confined form

t_statistic, p_value = stats.ttest_rel(before, after)

One of the important parameter:

alternative{*'two-sided', 'less', 'greater'} , optional*

For detail signature of the method ttest_rel *refer to*

[***https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html***](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html)

Independent Two-Sample Paired T-Test

```
import numpy as np
from scipy import stats

# Sample data for two groups
    before = np.array([85, 90, 88, 92, 78])
    after = np.array([80, 88, 86, 94, 77])

# Perform paired t-test
    t_statistic, p_value = stats.ttest_rel(before, after)
# Define significance level (alpha)
alpha = 0.05
# Compare p-value to alpha
if p_value < alpha:
    print(f"p-value ({p_value}) is less than alpha 
({alpha}). Reject the null hypothesis.")
else:
    print(f"p-value ({p_value}) is greater than or equal to 
alpha ({alpha}). Fail to reject the null hypothesis.")
```

In class quiz

The federal government awarded grants to the agricultural departments of 9 universities to test the yield capabilities of two new varieties of wheat. Each variety was planted on a plot of equal area at each university, and the yields, in kilograms per plot, were recorded as follows:

	University								
Variety	1	2	3	4	5	6	7	8	9
1	38	23	35	41	44	29	37	31	38
2	45	25	31	38	50	33	36	40	43

Find a 95% confidence interval for the mean difference between the yields of the two varieties, assuming the differences of yields to be approximately normally distributed. Also apply paired t-test. Explain why pairing is necessary in this problem.

```python
import numpy as np
from scipy import stats

# Data
variety_1 = np.array([38, 23, 35, 41, 44, 29, 37, 31, 38])
variety_2 = np.array([45, 25, 31, 38, 50, 33, 36, 40, 43])

# Calculate the differences
differences = variety_1 - variety_2

# Calculate the sample mean and standard error
mean_diff = np.mean(differences)
std_err = stats.sem(differences)

# Confidence level and degrees of freedom
alpha = 0.05
df = len(differences) - 1

# Calculate the margin of error
margin_of_error = stats.t.ppf(1 - alpha / 2, df) * std_err

# Calculate the confidence interval
lower_bound = mean_diff - margin_of_error
upper_bound = mean_diff + margin_of_error
```

```python
# Print results
print(f"Sample Mean Difference: {mean_diff}")
print(f"Standard Error of the Mean Difference: {std_err}")
print(f"Degrees of Freedom: {df}")
print(f"95% Confidence Interval: ({lower_bound}, {upper_bound})")
```

Sample Mean Difference: -2.7777777777777777

Standard Error of the Mean Difference: 1.5255033575273311

Degrees of Freedom: 8

95% Confidence Interval: (-6.295594828243093, 0.7400392726875378)

```python
import numpy as np
from scipy import stats
# Data
variety_1 = np.array([38, 23, 35, 41, 44, 29, 37, 31, 38])
variety_2 = np.array([45, 25, 31, 38, 50, 33, 36, 40, 43])
# Perform paired t-test
t_statistic, p_value = stats.ttest_rel(variety_1, variety_2)
# Significance level (alpha)
alpha = 0.05
# Interpret the results
if p_value < alpha:
    print("p-value is less than alpha. Reject the null hypothesis.")
else:
    print("p-value is greater than or equal to alpha. Fail to reject the null hypothesis.")

# p-value is greater than or equal to alpha. Fail to reject the null hypothesis.
```