Advanced Statistics

Dr. Syed Faisal Bukhari
Associate Professor
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab

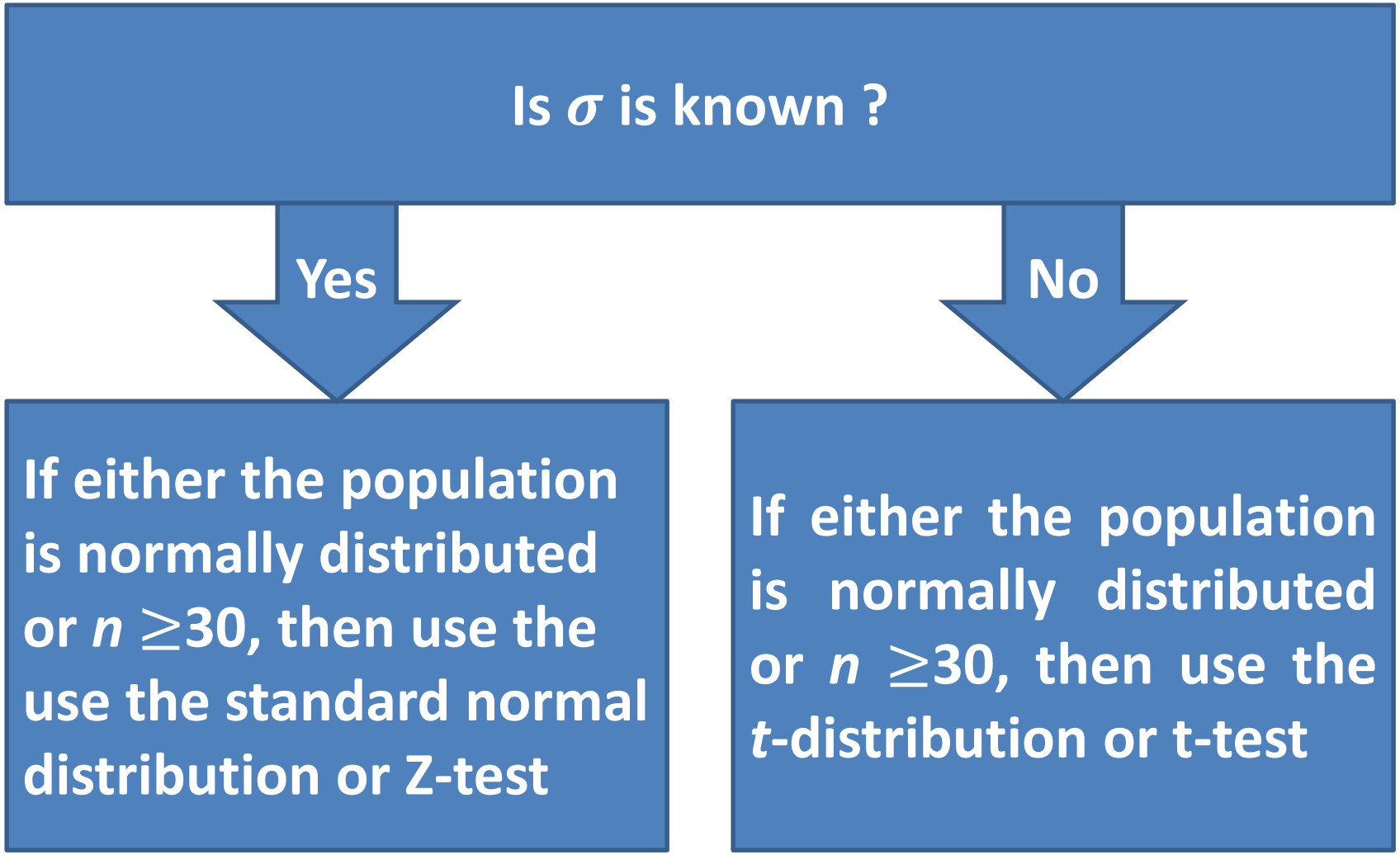
Textbooks

- **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
No
If either the population is normally distributed or $n \geq 30$, then use the use the standard normal distribution or Z-test
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 $t > t_{\alpha/2}$

Example 9.13: TCDD Levels in Plasma and Fat Tissue of Vietnam Veterans

A study published in Chemosphere reported the levels of the dioxin TCDD of 20 Massachusetts Vietnam veterans who were possibly exposed to Agent Orange. The **TCDD levels** in plasma and in **fat tissue** are listed in **Table 9.1**.

Find a **95% confidence interval** for $\mu_1 - \mu_2$, where μ_1 and μ_2 represent the **true mean TCDD levels in plasma** and in **fat tissue**, respectively. Assume the distribution of the differences to be approximately normal.

Note: Agent Orange was a plant-killing chemical (herbicide). The United States military used Agent Orange during the Vietnam conflict from 1962 to 1971 to clear trees, plants and vegetation from U.S. bases and to remove foliage used for cover

A 100(1 − α)% 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

$$s_d = \sqrt{\frac{1}{n(n-1)} \{n \sum d_i^2 - (\sum d_i)^2\}}$$

$d_i = x_{i1} - x_{i2}$

or

$d_i = x_{i2} - x_{i1}$

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

Table 9.1

Veteran	TCDD Levels in Plasma	TCDD Levels in Fat Tissue	$d_i = x_{i1} - x_{i2}$	d^2_i
1	2.5	4.9	-2.4	5.76
2	3.1	5.9	-2.8	7.84
3	5.2	3.4	1.8	3.24
4	3.5	6.9	-3.4	11.56
5	8.1	4.2	3.9	15.21
6	10.2	5.8	4.4	19.36
7	36.0	41.4	-5.4	29.16
8	6.2	10.3	-4.1	16.81
9	4.7	4.4	0.3	0.09
10	4.7	5.4	-0.7	0.49

Table 9.1

Veteran	TCDD Levels in Plasma	TCDD Levels in Fat Tissue	$d_i = x_{i1} - x_{i2}$	d^2_i
11	6.9	7.0	-0.1	0.01
12	6.9	7.1	-0.2	0.04
13	3.3	3.2	0.1	0.01
14	3.5	5.7	-2.2	4.84
15	5.7	5.0	0.7	0.49
16	5.1	5.2	-0.1	0.01
17	3.8	3.5	0.3	0.09
18	6.9	7.6	-0.7	0.49
19	2.5	2.3	0.2	0.04
20	2.4	5.2	-2.8	7.84
			$\sum d_i = -13.20$	$\sum d^2_i = 123.38$

$$\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

$$\overline{d} = \frac{-13.2}{20}$$

$$= -0.63$$

$$s_d = \sqrt{\frac{1}{n(n-1)}\{n\sum d_i^2 - (\sum d_i)^2\}}$$

$$s_d = \sqrt{\frac{1}{20(20-1)}\{(20)(123.38) - (-13.20\}}$$

$$= 2.46$$

$$t_{(\alpha/2,n-1)} = t_{(0.025,\ 19)} = 2.093$$

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

$$-0.63 - 2.093\left(\frac{2.46}{\sqrt{20}}\right) < \mu_d < -0.63 + 2.093\left(\frac{2.46}{\sqrt{20}}\right)$$

$$-1.78 < \mu_d < 0.52$$

There is no significant difference between the mean TCDD level in plasma and the mean TCDD level in fat tissue

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:

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

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

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

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

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

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)

$\overline{d} = ?$

$s_d = ?$

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		$\sum_{i=1}^{n} d_i = 16$	$\sum_{i=1}^{n} d^2{}_i = 42$

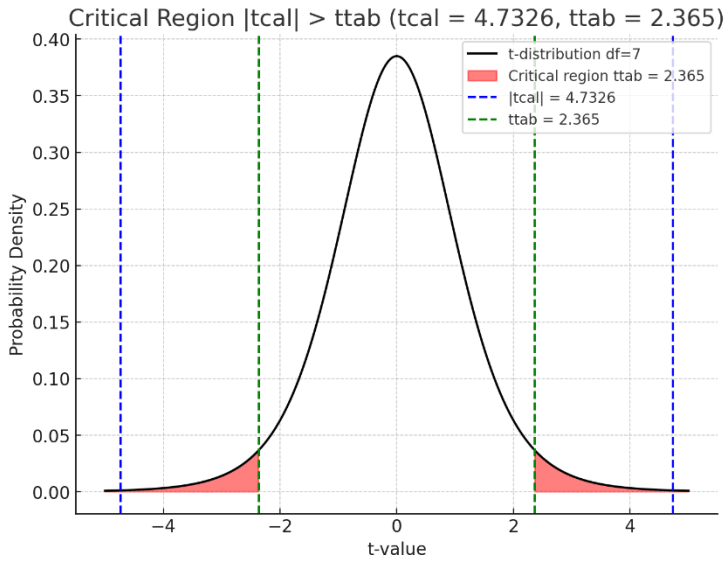
$$\overline{d} = \frac{\sum_{i=1}^{n} \text{di}}{n} = 16/8 = \mathbf{2}$$

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

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

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

$|t_{cal}| = 4.7326$



5. **Critical region:**

$|t_{cal}| = 4.7326$

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

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.

Case Study 10.1: Blood Sample Data

Study by J. A. Wesson, Forestry and Wildlife Department, Virginia Tech

Study Overview

J. A. Wesson examined the influence of succinylcholine on androgen levels in deer blood.

- Conducted in the Forestry and Wildlife Department at Virginia Tech.

- Blood samples were taken from wild, free-ranging deer.

- Samples were collected after administration of succinylcholine via dart and capture gun.

Methodology

1. Intramuscular injection of succinylcholine was administered to deer.

2. First blood sample was taken immediately after injection.

3. A second blood sample was taken 30 minutes after injection.

Assuming that the populations of androgen levels at time of injection and 30 minutes later are normally distributed, test at the 0.05 level of significance whether the androgen concentrations are altered after 30 minutes.

Table for Case Study 10.1 (Part 1)

Deer	At Time of Injection (ng/mL)	30 Minutes after Injection (ng/mL)	$d_i = x_{1i} - x_{2j}$
1	2.76	7.02	-4.26
2	5.18	3.10	-2.08
3	2.68	5.44	2.76
4	3.05	3.99	0.94
5	4.10	5.21	1.11
6	7.05	10.26	3.21
7	6.60	13.91	7.31

Table for Case Study 10.1 (Part 2)

Deer	At Time of Injection (ng/mL)	30 Minutes after Injection (ng/mL)	$d_i = x_{1i} - x_{2j}$
8	4.79	18.53	13.74
9	7.39	7.91	0.52
10	7.30	4.85	-2.45
11	11.78	11.10	-0.68
12	3.90	3.22	-0.68
13	20.60	94.03	68.03
14	67.48	94.03	26.55
15	17.04	41.70	24.66

$\mu_D = 0$ (Population mean)

$n = 15$ (Sample size)

$\alpha = 0.05$ (Level of significance)

$\overline{d} = ?$

$s_d = ?$

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:

Deer	x_{1i}	x_{2i}	d_i	$d^2{}_i$
1	2.76	7.02	-4.26	18.15
2	5.18	3.10	-2.08	4.33
3	2.68	5.44	2.76	7.62
4	3.05	3.99	0.94	0.88
5	4.10	5.21	1.11	1.23
6	7.05	10.26	3.21	10.3
7	6.60	13.91	7.31	53.44

Deer	x_{1i}	x_{2i}	d_i	$d^2{}_i$
8	4.79	18.53	13.74	188.79
9	7.39	7.91	0.52	0.27
10	7.30	4.85	-2.45	6.0
11	11.78	11.10	-0.68	0.46
12	3.90	3.22	-0.68	0.46
13	20.60	94.03	68.03	4628.08
14	67.48	94.03	26.55	704.9
15	17.04	41.70	24.66	608.12
		Total	138.68	6233.03

$$\bar{d} = \frac{\sum_{i=1}^{n} \text{di}}{n} = 138.2/15 = \mathbf{9.21}$$

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

$$s_d = \sqrt{\frac{1}{15(15-1)}\{15(6233.03) - (138.2)^2\}} = \mathbf{18.82}$$

$$t_{cal} = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = t_{cal} = \frac{9.21 - 0}{\frac{18.82}{\sqrt{15}}} = 1.90$$

$|t_{cal}| = 1.90$

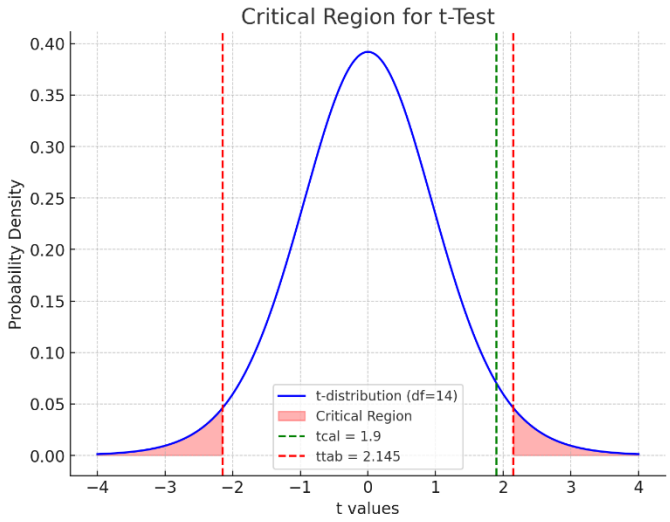
5. **Critical region:**

$|t_{cal}| = 1.90$

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

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

$= t_{(0.0250,\, 14)} = 2.145$



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

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
```

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

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.

```
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.
```