Chapter 1

Chapter 2

Statistical methods related to means

These methods are based on two essential conditions:

- the normality of the populations;
- the random and simple nature of the samples.

For certain tests concerning means, a third condition is required:

- the equality of the population variances.

2.1 Estimation of the Mean

2.1.1 Principle of Estimation

Sampling theory involves determining sample properties from a known population. The principle of estimation reverses this: we have data from samples (surveys, control tests, etc.) and seek to infer properties of the entire population.

Remark 1. From a small sample, exact data about the population cannot be obtained. Therefore, we must provide estimates along with their margin of error or risk.

2.1.2 Types of Estimation

Estimation may be:

- Point estimation
- Interval estimation (confidence interval)

Point Estimation

A point estimate of the mean m is the sample mean of a random sample of size n:

$$\bar{x} \quad \text{is an estimator of } m.$$

2.1.3 Estimation by Confidence Interval

Point estimates, although useful, do not provide any information regarding the precision of the estimates. They do not take into account possible errors due to sampling fluctuations. The theory of **confidence intervals (CI)** consists of constructing, around the point estimate, an interval that has a high probability $(1-\alpha)$ of containing the true value of the parameter.

2.2 Sampling Distribution and Confidence Interval for a Mean

2.2.1 Case of Large Samples $(n \geq 30)$

A. Sampling Distribution of a Mean

Let us consider a population characterized by a mean m and a standard deviation δ_p, corresponding to a quantitative variable. If we randomly draw k samples, each of the same size n, we will notice that the sample means $m_1, m_2, \ldots, m_k$ show variations due to sampling fluctuations.

Let $\bar{X}$ denote the random variable that may take as its value the mean of a randomly drawn sample from the population. This variable $\bar{X}$ is called the **sampling mean**, and its probability law is known as the **sampling distribution of the mean**.

It can be shown that $\bar{X}$ follows a normal distribution with mean μ and variance $\frac{\delta_p^2}{n}$ for $n \geq 30$:

$$\bar{X} \sim \mathcal{N}\left(m, \frac{\delta_p^2}{n}\right)$$

Thus:

$$P\left(m - t_{1-\frac{\alpha}{2}}\frac{\delta_p}{\sqrt{n}} \leq \bar{X} \leq m + t_{1-\frac{\alpha}{2}}\frac{\delta_p}{\sqrt{n}}\right) = 1-\alpha$$

The probability that $\bar{X}$ lies within the interval $\left[m - t_{1-\frac{\alpha}{2}}\frac{\delta_p}{\sqrt{n}},\ m + t_{1-\frac{\alpha}{2}}\frac{\delta_p}{\sqrt{n}}\right]$ is $1-\alpha$, which defines the **confidence interval of the mean**.

- $(1-\alpha)$: Confidence level
- α: Risk of error
- $t_{1-\frac{\alpha}{2}}$: Value from the standard normal table

Generally, $\alpha = 5\%$, and in particular cases, $\alpha = 1\%$.

$$\alpha = 5\% \Rightarrow t_{1-\frac{\alpha}{2}} = 1.96, \quad P\left(m - 1.96\frac{\delta_p}{\sqrt{n}} \leq \bar{X} \leq m + 1.96\frac{\delta_p}{\sqrt{n}}\right) = 0.95$$

$$\alpha = 1\% \Rightarrow t_{1-\frac{\alpha}{2}} = 2.60, \quad P\left(m - 2.60\frac{\delta_p}{\sqrt{n}} \leq \bar{X} \leq m + 2.60\frac{\delta_p}{\sqrt{n}}\right) = 0.99$$

Example. A machine is intended to produce tablets with an average weight of 200 mg and a standard deviation of 10 mg. A random sample of 50 tablets is taken.

Population: $m = 200$ mg, $\delta_p = 10$ mg; Sample size: $n = 50 > 30$.

At $\alpha = 5\%$, $t_{1-\frac{\alpha}{2}} = 1.96$:

$$IC = [197.22,\ 202.77]$$

The average weight of 50 tablets lies between 197.22 mg and 202.77 mg with a 5% risk of error.

2.2.2 B. Confidence Interval for a Mean

Suppose we wish to study a quantitative characteristic within a population. Let m be the population mean and δ_p the population standard deviation (both unknown). We randomly draw a sample of size n and compute the sample mean $\bar{X}$ and standard deviation S.

The goal is to estimate the population mean m based on n, $\bar{X}$, and S, that is, to find an interval in which the true population mean m is expected to lie.

From the previous results:

$$m - t_{1-\frac{\alpha}{2}} \frac{\delta_p}{\sqrt{n}} \leq \bar{X} \leq m + t_{1-\frac{\alpha}{2}} \frac{\delta_p}{\sqrt{n}}$$

Since the sample mean $\bar{X}$ is known, it represents the observed value of the random variable $\bar{X}$ (the sampling mean). We can therefore write:

$$\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\delta_p}{\sqrt{n}} \leq m \leq \bar{X} + t_{1-\frac{\alpha}{2}} \frac{\delta_p}{\sqrt{n}}$$

When the population standard deviation δ_p is unknown, the population variance δ_p^2 must be estimated using the sample variance

$$\frac{n}{n-1} S^2.$$

Thus, the confidence interval for m becomes:

$$\bar{X} - t_{1-\frac{\alpha}{2}} \frac{\delta_e}{\sqrt{n-1}} \leq m \leq \bar{X} + t_{1-\frac{\alpha}{2}} \frac{\delta_e}{\sqrt{n-1}}$$

or equivalently,

$$P\left(\bar{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \leq m \leq \bar{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}}\right) = 1 - \alpha$$

This means that the interval:

$$[\bar{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}},\ \bar{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}}]$$

has a probability of $(1-\alpha)$ of containing the true population mean m. This interval is called the **confidence interval of the mean**.

Common choices are $\alpha = 5\%$ (confidence level $= 95\%$) or $\alpha = 1\%$ (confidence level $= 99\%$). The corresponding critical values are:

$$t_{1-\frac{\alpha}{2}} = 1.96 \quad \text{for } \alpha = 5\%, \qquad t_{1-\frac{\alpha}{2}} = 2.60 \quad \text{for } \alpha = 1\%.$$

In general form, the confidence interval for the population mean is written as:

$$IC(m) = \left[\bar{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}},\ \bar{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}}\right]$$

at the risk level α (or confidence level $1-\alpha$).

Example. In a population of individuals, a random sample of size $n = 40$ has a mean weight of $\bar{X} = 70$ kg and a standard deviation of $\delta_e = 15.4$ kg. Determine, at a 5% risk level, the confidence interval for the mean weight of the population.

Solution:

$$n = 40 > 30, \quad \bar{X} = 70, \quad S = 15.4, \quad t_{1-\frac{\alpha}{2}} = 1.96.$$

Then:

$$IC(m) = \left[\bar{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}},\ \bar{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}}\right]$$

$$IC(m) = [65.16,\ 74.83]$$

At the given risk level $\alpha = 5\%$, we have $t_{1-\frac{\alpha}{2}} = 1.96$.

$$IC(m) = [65.16,\ 74.83] \quad \text{at the 5\% risk level.}$$

This means that there is a 95% probability that the confidence interval $[65.16,\ 74.83]$ contains the true population mean m.

2.2.3 Case of Small Samples ($n < 30$)

A. Sampling Distribution of a Mean

As in the case of large samples, we have:

$$P\left(m - t_{1-\frac{\alpha}{2}} \frac{\delta_p}{\sqrt{n}} \leq \bar{X} \leq m + t_{1-\frac{\alpha}{2}} \frac{\delta_p}{\sqrt{n}}\right) = 1 - \alpha$$

This represents the probability that the sample mean $\bar{X}$ lies within the interval

$$[m - t_{1-\frac{\alpha}{2}} \frac{\delta_p}{\sqrt{n}},\ m + t_{1-\frac{\alpha}{2}} \frac{\delta_p}{\sqrt{n}}].$$

B. Confidence Interval for a Mean

The difference from the case of large samples appears when the population variance δ_p^2 is replaced by its sample estimate

$$\frac{n}{n-1} S^2$$

obtained from the observed sample. This substitution is acceptable only for large samples but not for small ones.

For small samples, the sampling distribution of the mean follows a probability law slightly different from the normal law, called the **Student–Fisher t distribution**, which depends on the sample size n. The probability density curve of the Student distribution is flatter than that of the normal distribution.

Depending on the risk level α and the number of degrees of freedom $df = n - 1$, and since the sample mean $\bar{X}$ is known, the confidence interval for the population mean μ is given by:

$$\bar{X} - t^*_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \leq m \leq \bar{X} + t^*_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}}$$

Confidence Interval of the Population Mean (Small Sample)

We have:

$$P\left(\bar{X} - t^*_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \leq m \leq \bar{X} + t^*_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \right) = 1 - \alpha$$

This represents the probability that the interval

$$\left[\bar{X} - t^*_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}},\ \bar{X} + t^*_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \right]$$

contains the population mean μ.

This interval is called the **confidence interval of the population mean**. In general, we choose $\alpha = 5\%$ or, in some special cases, $\alpha = 1\%$. Thus, in general form:

$$IC(m) = \left[\bar{X} - t^*_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}},\ \bar{X} + t^*_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \right]$$

at the risk level α (or the confidence level $1 - \alpha$).

Example. A sugar concentration test was performed on 8 samples taken from the same population, giving the following results (in g/L):

$$19.5,\ 19.7,\ 19.8,\ 20.2,\ 20.3,\ 20.4,\ 20.4,\ 20.8$$

Tasks:

1. Calculate the sample mean and standard deviation.
2. Determine the 95% confidence interval for the mean.

Solution.
1. *Calculation of the sample mean and standard deviation:*

$$\bar{X} = \frac{1}{8} \sum_{i=1}^{8} x_i = 20.11, \qquad S = \sqrt{\frac{1}{8} \sum_{i=1}^{8} (x_i - \bar{X})^2} = 0.395$$

2. *Confidence interval for the population mean:*

Let m be the population mean of sugar concentration to be estimated. Assuming that sugar concentration in the population follows a normal distribution, the confidence interval for the mean is:

$$IC(m) = \left[\bar{X} - t^*_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}},\ \bar{X} + t^*_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n-1}} \right]$$

At the 5% risk level ($\alpha = 0.05$) and with degrees of freedom $df = n - 1 = 7$, the Student–Fisher table gives:

$$t^*_{1-\frac{\alpha}{2}} = 2.365$$

Hence:

$$IC(m) = [19.75,\ 20.46] \quad \text{at the 5\% risk level.}$$

Interpretation: There is a 95% probability that the true mean sugar concentration m lies between **19.75 g/L** and **20.46 g/L**.

Exercise A — Raw Data (Small Sample, Student's t Distribution)

Problem

In a farmland survey, the nitrate concentration (mg/kg) was measured in 10 soil samples:

$$12.5,\ 13.2,\ 11.8,\ 12.9,\ 13.0,\ 12.4,\ 11.9,\ 12.6,\ 12.7,\ 12.3.$$

Assuming these values come from a normal distribution:

1. Compute the sample mean $\bar{x}$ and standard deviation S.
2. Determine the 95% confidence interval for the population mean.
3. Interpret the result in the environmental context.

Solution

1. Sample mean

$$\sum_{i=1}^{10} x_i = 12.5 + 13.2 + 11.8 + 12.9 + 13.0 + 12.4 + 11.9 + 12.6 + 12.7 + 12.3 = 125.3$$

$$\bar{x} = \frac{125.3}{10} = 12.53 \text{ mg/kg}$$

2. Sample variance and standard deviation (using division by n)

Compute the squared deviations (sum):

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 1.8410$$

Using your requested formula with n:

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{1.8410}{10} = 0.18410$$

Sample standard deviation (population-style):

$$S = \sqrt{S^2} = \sqrt{0.18410} \approx 0.42905 \quad \text{(rounded 0.4290)}$$

3. Standard error of the mean (using your formula)

You requested $\text{SE} = \dfrac{S}{\sqrt{n-1}}$. Thus

$$\text{SE} = \frac{0.42905}{\sqrt{9}} = \frac{0.42905}{3} \approx 0.14302 \quad \text{(rounded 0.1430)}$$

4. 95% Confidence Interval

Degrees of freedom (for the t-quantile):

$$df = n - 1 = 9$$

Critical value:

$$t_{0.975,9} \approx 2.262$$

Margin of error:

$$t_{0.975,9} \times \text{SE} = 2.262 \times 0.14302 \approx 0.32351$$

Confidence interval:

$$\boxed{\text{CI}_{95\%} = [\,12.53 - 0.32351,\ 12.53 + 0.32351\,] \approx [12.2065,\ 12.8535]\ \text{mg/kg}}$$

3. Interpretation. We are 95% confident that the true mean nitrate concentration in this soil lies between **12.21** and **12.85 mg/kg**. Since the upper bound is near 13 mg/kg (a possible agronomic threshold), further sampling could improve accuracy.

Exercise B — Grouped (Continuous) Data (Large Sample, z Approximation)

Problem

In the quality control of effervescent tablets, a sample of 150 tablets was tested for sodium bicarbonate content (mg). The data were grouped as follows:

Class interval (mg)	[1610,1615)	[1615,1620)	[1620,1625)	[1625,1630)	[1630,1635)
Frequency f_i	7	8	42	75	18

Use the class midpoints to estimate the sample mean $\bar{x}$ and standard deviation S, and construct a 95% confidence interval for the true mean sodium bicarbonate content.

1. Class centers and sample mean. Class centers:

$$c_i = 1612.5,\ 1617.5,\ 1622.5,\ 1627.5,\ 1632.5$$

Weighted sum:

$$\sum f_i c_i = 7(1612.5) + 8(1617.5) + 42(1622.5) + 75(1627.5) + 18(1632.5) = 243820.0$$

Total frequency:

$$n = 7 + 8 + 42 + 75 + 18 = 150$$

Sample mean:

$$\bar{x} = \frac{\sum f_i c_i}{n} = \frac{243820.0}{150} = 1625.466\ldots \approx 1625.47 \text{ mg}$$

2. Variance and standard deviation (using division by n as requested). Sum of weighted squared deviations:

$$\sum f_i (c_i - \bar{x})^2 = 3254.83$$

Variance (population-style, divide by n):

$$S^2 = \frac{\sum f_i (c_i - \bar{x})^2}{n} = \frac{3254.83}{150} \approx 21.6988667$$

Standard deviation:

$$S = \sqrt{S^2} = \sqrt{21.6988667} \approx 4.65820 \text{ mg}$$

3. 95% confidence interval for the mean (using SE defined with $\sqrt{n-1}$). Standard error (as you specified):

$$\text{SE} = \frac{S}{\sqrt{n-1}} = \frac{4.65820}{\sqrt{149}} \approx \frac{4.65820}{12.2066} \approx 0.381615$$

Using the normal critical value $z_{0.025} = 1.96$ (or $t_{0.975,149}$ nearly the same):

$$\text{margin} = 1.96 \times \text{SE} \approx 1.96 \times 0.381615 \approx 0.74797 \approx 0.748$$

95% CI:

$$\text{CI}_{95\%} = \big[\bar{x} - 0.74797,\ \bar{x} + 0.74797\big] = [1625.4667 - 0.74797,\ 1625.4667 + 0.74797] \approx [1624.7187,\ 1626.2146]$$

Conclusion. Using the formulas you requested, the 95% confidence interval for the mean is approximately $\boxed{[1624.72,\ 1626.21]\ \text{mg}}$.

4. Interpretation. We are 95% confident that the true mean sodium bicarbonate content lies between **1624.72 mg** and **1626.22 mg**. Since the target value is 1625 mg, and it lies within the confidence interval, the production process is well calibrated.

2.3 The One-Sample t-Test (Test of Conformity)

1. Definition

The one-sample t-test is used to determine whether the mean of a population differs significantly from a known or assumed theoretical value m_0.

2. Objective

Tests of conformity aim to verify whether a sample can be considered as drawn from a given population or representative of that population, with respect to a parameter such as the mean.

3. Hypotheses

- Null hypothesis: $H_0 : m = m_0$
- Alternative hypothesis:
  - Two-tailed: $H_1 : m \neq m_0$
  - Right-tailed: $H_1 : m > m_0$
  - Left-tailed: $H_1 : m < m_0$

4. Test Statistic

The test statistic is given by:

$$t = \frac{\bar{X} - m_0}{S/\sqrt{n-1}}$$

where:

- $\bar{X}$: sample mean
- S : sample standard deviation
- n : sample size
- m_0 : theoretical mean

The statistic t follows a **Student's t-distribution** with $(n - 1)$ degrees of freedom.

5. Testing Procedure

1. Formulate the hypotheses H_0 and H_1
2. Choose the significance level α (commonly 5%)
3. Compute the test statistic t
4. Determine the critical value $t_{1-\frac{\alpha}{2},n-1}$
5. Make a decision:
   - If $|t| > t_{1-\frac{\alpha}{2}}$: Reject H_0
   - Otherwise: Do not reject H_0

6. Numerical Example

An agroecologist wants to test whether the mean soil organic matter content in a field differs from the national reference value

$$m_0 = 4.5\%.$$

A random sample of $n = 12$ soil samples yields the following percentages of organic matter:

$$4.2,\ 4.8,\ 5.0,\ 4.1,\ 4.4,\ 4.7,\ 4.3,\ 4.9,\ 4.5,\ 4.6,\ 4.0,\ 4.8$$

We perform a two-tailed one-sample t-test at significance level $\alpha = 0.05$.

Data (table)

4.2	4.8	5.0	4.1	4.4	4.7	4.3	4.9	4.5	4.6	4.0	4.8

Step-by-step numerical calculation

1. Sample size and sum

$$n = 12 \qquad \sum_{i=1}^{n} x_i = 54.3$$

2. Sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{54.3}{12} = 4.525$$

3 variance and squared deviations

We compute $d_i = x_i - \bar{X}$ and d_i^2. For clarity we list them:

x_i	$d_i = x_i - \bar{X}$	d_i^2
4.2	-0.325	0.105625
4.8	$+0.275$	0.075625
5.0	$+0.475$	0.225625
4.1	-0.425	0.180625
4.4	-0.125	0.015625
4.7	$+0.175$	0.030625
4.3	-0.225	0.050625
4.9	$+0.375$	0.140625
4.5	-0.025	0.000625
4.6	$+0.075$	0.005625
4.0	-0.525	0.275625
4.8	$+0.275$	0.075625
Sum of squared deviations	$\sum_{i=1}^{n} d_i^2 = 1.1825$	

4. Sample variance and standard deviation

Sample variance (using n):

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n} = \frac{1.1825}{12} = 0.09854$$

Sample standard deviation:

$$S = \sqrt{S^2} = \sqrt{0.09854} \approx 0.3140 \quad \text{(rounded 0.314)}$$

5. Standard error of the mean (SE)

$$\text{SE} = \frac{S}{\sqrt{n-1}} = \frac{0.3140}{\sqrt{11}} = \frac{0.3140}{3.3166} \approx 0.0946$$

6. Test statistic

Difference between sample mean and reference:

$$\bar{X} - m_0 = 4.525 - 4.5 = 0.025$$

t statistic:

$$t = \frac{\bar{X} - m_0}{\frac{S}{\sqrt{n-1}}} = \frac{0.025}{0.0946484724} \approx 0.2641352719$$

Degrees of freedom:

$$df = n - 1 = 11$$

7. p-value and critical value

For a two-tailed test with $df = 11$:

$$t_{0.975,11} \approx 2.2009851601 \quad \text{(often rounded 2.201)}$$

8. Decision

Since $|t| = 0.2641 < 2.201$, we **do not reject** H_0.

Interpretation: There is no significant difference between the observed mean soil organic matter content (4.55%) and the national standard of 4.5%. Therefore, the soils in this area can be considered to conform to the standard organic matter level.

2.4 Significance Test and Confidence Interval for the Difference of Two Means: Independent Samples

The objective is to compare two populations based on two independent samples in order to determine if there is a significant difference between their means.

Examples

- Compare the average grades between two classes;
- Compare the average weight between men and women;
- Compare the effectiveness of two treatments.

2.5 Hypotheses

The usual hypotheses are:

$$H_0 : m_1 = m_2 \qquad \text{vs} \qquad H_1 : m_1 \neq m_2$$

(We can also consider one-sided hypotheses: $H_1 : m_1 > m_2$ or $H_1 : m_1 < m_2$.)

Student's t-test for two independent samples

Case 1: $n_1 \neq n_2$

The test statistic is:

$$t_{\text{obs}} = \frac{|\overline{X}_1 - \overline{X}_2|}{\sqrt{\dfrac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

or in terms of sum of squared deviations:

$$t_{\text{obs}} = \frac{|\overline{X}_1 - \overline{X}_2|}{\sqrt{\frac{SCE_1 + SCE_2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\text{If } t_{\text{obs}} \geq t_{1-\alpha/2} \Rightarrow \text{Reject } H_0 \Rightarrow m_1 \neq m_2 \quad \text{for} \quad \begin{cases} \alpha = 0.05 \\ (n_1 + n_2 - 2) \text{ degrees of freedom.} \end{cases}$$

Case 2: $n_1 = n_2 = n$

$$t_{\text{obs}} = \frac{|\overline{X}_1 - \overline{X}_2|}{\sqrt{\frac{SCE_1 + SCE_2}{n(n-1)}}}$$

$$\text{If } t_{\text{obs}} \geq t_{1-\alpha/2} \Rightarrow \text{Reject } H_0 \Rightarrow m_1 \neq m_2 \quad \text{for} \quad \begin{cases} \alpha = 0.05 \\ 2(n-1) \text{ degrees of freedom.} \end{cases}$$

Note: If the variances of the two samples are unequal, the **Welch test** should be used, which is more appropriate in this case.

Example

We want to compare the average exam scores of two classes to see if there is a significant difference.

- Class A ($n_1 = 8$): 78, 82, 85, 90, 76, 88, 84, 79
- Class B ($n_2 = 7$): 72, 75, 80, 77, 74, 79, 70

Significance level: $\alpha = 0.05$ (two-tailed)

Step 1: Compute Sample Means

$$\overline{X}_1 = \frac{78 + 82 + 85 + 90 + 76 + 88 + 84 + 79}{8} = 82.75$$

$$\overline{X}_2 = \frac{72 + 75 + 80 + 77 + 74 + 79 + 70}{7} = 75.29$$

Step 2: Compute Sample Variances

$$S_1^2 = \frac{\sum(x_i - \overline{x}_1)^2}{n_1 - 1} = \frac{167.48}{7} \approx 23.926$$

$$S_2^2 = \frac{\sum(x_i - \overline{x}_2)^2}{n_2 - 1} = \frac{80.61}{6} \approx 13.435$$

Step 3: Compute Pooled Variance Component

$$S_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} = \frac{8 \cdot 23.926 + 7 \cdot 13.435}{13} \approx 21.957$$

$$\frac{1}{n_1} + \frac{1}{n_2} = \frac{1}{8} + \frac{1}{7} \approx 0.267857$$

$$\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = \sqrt{21.957 \cdot 0.267857} \approx 2.425$$

Step 4: Compute t-Statistic

$$t_{\text{obs}} = \frac{|\overline{X}_1 - \overline{X}_2|}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{82.75 - 75.29}{2.425} \approx 3.07$$

Step 5: Critical Value and Degrees of Freedom

$$df = n_1 + n_2 - 2 = 13$$

Two-tailed test at $\alpha = 0.05$: $t_{0.975,13} \approx 2.160$

Step 6: Decision

$$t_{\text{obs}} = 3.07 > t_{0.975,13} = 2.160 \quad \Rightarrow \quad \text{Reject } H_0$$

Conclusion: There is a significant difference between the mean scores of Class A and Class B.

Step 7: 95% Confidence Interval

$$IC_{95\%} = (\overline{x}_1 - \overline{x}_2) \pm t_{0.975,13} \cdot \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$IC_{95\%} = 7.46 \pm 2.160 \cdot 2.425 \approx 7.46 \pm 5.23$$

$$IC_{95\%} \approx [2.23, 12.69]$$

Interpretation: The true difference in mean scores is likely between 2.23 and 12.69 points.

2.6 Significance Test and Confidence Interval for the Difference of Two Means: Paired Samples

1. Principle

The paired-sample t-test is used to compare two related measurements, for example:

- Before / After a treatment on the same individuals

- Math and French scores for the same students
- Weight before and after a diet

We test whether the mean of the differences is significantly different from zero.

1. Hypothesis

The test of equality of means for two paired samples is written as:

$$H_0 : m_1 = m_2 \quad \text{or} \quad H_0 : \delta = 0$$

We then calculate the differences for each pair:

$$d_i = x_i - y_i \quad \text{for } i = 1, \ldots, n$$

—

2. Test Statistic

$$t_{\text{obs}} = \frac{\overline{d}}{S_d/\sqrt{n}}$$

where:

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad S_d^2 = \frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}$$

$$\text{df} = n - 1$$

—

3. Decision

$$\text{If } |t_{\text{obs}}| \geq t_{1-\alpha/2,n-1} \quad \Rightarrow \quad \text{Reject } H_0$$

This test is called the **paired-sample Student's t-test**.

—

4. Confidence Interval for the Difference

If H_0 is rejected, we can estimate the mean difference δ and calculate a confidence interval:

$$\hat{\delta} = \overline{d}, \quad CI_{1-\alpha} = \overline{d} \pm t_{1-\alpha/2,n-1}\frac{S_d}{\sqrt{n}}$$

—

Example: Effect of Fertilizer on Plant Biomass (Paired Samples)

We measure the biomass (in g) of 6 plants before and after applying a fertilizer.

Plant	Before x_1	After x_2	$d_i = x_1 - x_2$	$d_i - \overline{d}$	$(d_i - \overline{d})^2$
1	12.5	14.0	-1.5	-0.533	0.284
2	10.2	11.5	-1.3	-0.333	0.111
3	15.0	15.8	-0.8	0.167	0.028
4	11.7	12.4	-0.7	0.267	0.071
5	13.3	14.2	-0.9	0.067	0.004
6	12.0	12.6	-0.6	0.367	0.135

Step 1: Mean of Differences

$$\overline{d} = \frac{-1.5 - 1.3 - 0.8 - 0.7 - 0.9 - 0.6}{6} = \frac{-5.8}{6} \approx -0.967$$

Step 2: Variance and Standard Deviation

$$S_d^2 = \frac{\sum (d_i - \overline{d})^2}{n-1} = \frac{0.633}{5} \approx 0.127$$

$$S_d = \sqrt{0.127} \approx 0.356$$

Step 3: t-Statistic

$$t_{\text{obs}} = \frac{\overline{d}}{S_d/\sqrt{n}} = \frac{-0.967}{0.356/\sqrt{6}} \approx -6.66$$

Step 4: Degrees of Freedom and Critical Value

$$\text{df} = n - 1 = 6 - 1 = 5, \quad t_{0.975,5} \approx 2.571$$

Step 5: Decision

$$|t_{\text{obs}}| = 6.66 > 2.571 \quad \Rightarrow \quad \text{Reject } H_0$$

Conclusion: The fertilizer has a significant effect on increasing plant biomass.

Step 6: 95% Confidence Interval

$$CI_{95\%} = \overline{d} \pm t_{0.975,5} \cdot \frac{S_d}{\sqrt{n}} = -0.967 \pm 2.571 \cdot 0.145 \approx [-1.340, -0.594]$$

Interpretation: The average increase in biomass is likely between 0.594 and 1.340 g.