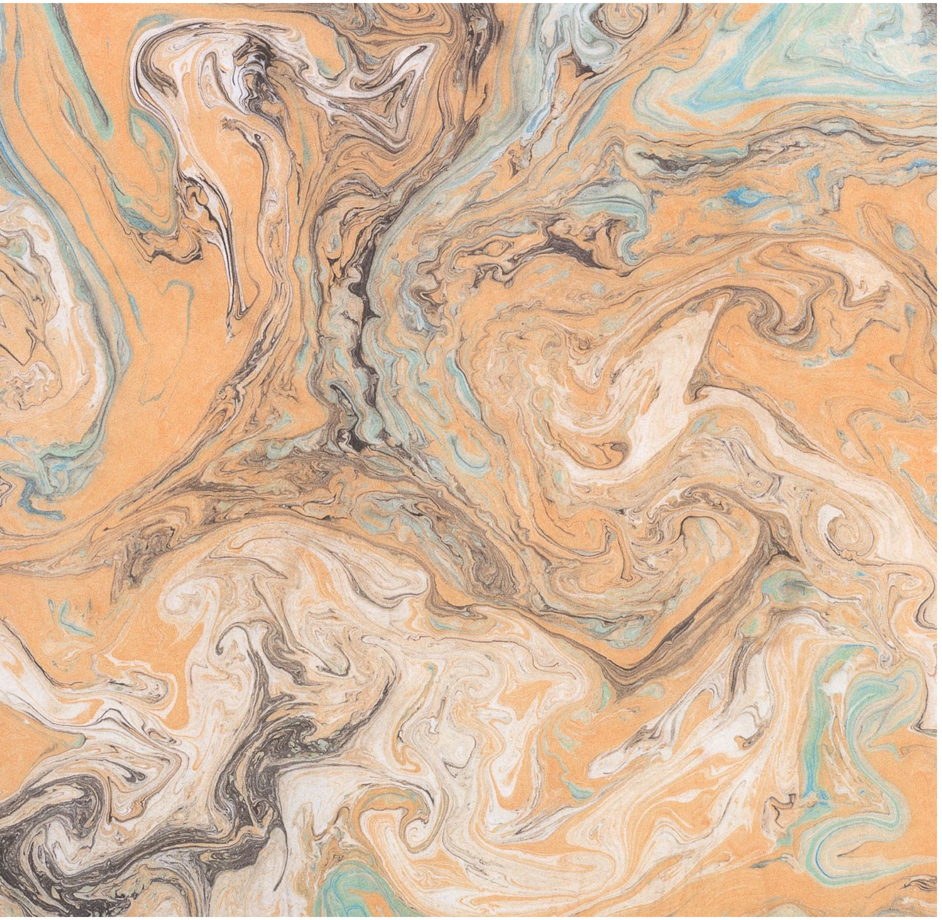

Week 6 – Lecture 2

Comparing Two Samples

Objectives

- Learn methods for comparing samples from distributions that may be different and especially with methods for making inferences about how the distributions differ.
- In many applications, the samples are drawn under different conditions, and inferences must be made about possible effects of these conditions.

Comparing Two Independent Samples On Normal Distribution

- Let $X_1, X_2, \ldots, X_n \sim N(\mu_X, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_m \sim N(\mu_Y, \sigma^2)$ are independent. Then

$$\overline{X} - \overline{Y} \sim N\left[\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right]$$

- If σ^2 is known, a confidence interval for $\mu_X - \mu_Y$ is given by

$$(\overline{X} - \overline{Y}) \pm z(\alpha/2)\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$$

Comparing Two Independent Samples On Normal Distribution (Confidence Interval)

If σ^2 is unknown, the CI for $\mu_X - \mu_Y$ is

$$\mu_X - \mu_Y \in (\bar{X} - \bar{Y}) \pm t_{\alpha/2,n+m-2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}} \qquad (1.2)$$

where $S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}$ and $S_X^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$. These CIs are based on test statistics

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \quad \textit{and} \quad t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

for σ known and σ unknown, respectively.

Comparing Two Independent Samples On Normal Distribution (Hypothesis Testing)

$$\begin{cases} H_0 : \mu_X - \mu_Y = 0 \\ H_1 : \mu_X - \mu_Y > 0 \end{cases} \qquad \begin{cases} H_0 : \mu_X - \mu_Y = 0 \\ H_1 : \mu_X - \mu_Y < 0 \end{cases} \qquad \begin{cases} H_0 : \mu_X - \mu_Y = 0 \\ H_1 : \mu_X - \mu_Y \neq 0 \end{cases}$$

$$Z_0 > Z_\alpha \qquad Z_0 < -Z_\alpha \qquad |Z_0| > Z_{\alpha/2}$$

$$t_0 > t_{n+m-2}(\alpha) \qquad t_0 < -t_{n+m-2}(\alpha) \qquad |t_0| > t_{n+m-2}(\alpha/2)$$

$$Z_0 = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \qquad t_0 = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

- Use z-test statistic if σ is known.
- Use t-test statistic if σ is unknown.

Comparing Two Independent Samples On Normal Distribution

Example 1:

$n = 13, m = 8, \bar{X} = 80.02, \bar{Y} = 79.98, S_X = 0.024, S_Y = 0.031$ and σ is unknown.

a) What is CI for $\mu_X - \mu_Y$?

$b)\ \begin{cases} H_0: \mu_X - \mu_Y = 0 \\ H_a: \mu_X - \mu_Y \neq 0 \end{cases} \qquad \alpha = 0.01$

Perform the above test.

Some notes

- In the case of the testing and CI methods for $\mu_X - \mu_Y$, the test rejects if and only if the CI does not include zero.
- The test of H_0 vs H_1 defined here is equivalent to a likelihood ratio test.
- If the distributions are **NOT normal**:

  1) $n + m - 2 > 30$ → By CLT we can use Z-test statistic.

  2) Transform data so we have normal distributions.

A Non-Parametric Method - The Mann-Whitney Test

- Non-parametric methods do not assume that the data follow any particular distributional form.
- **Mann-Whitney Test (Wilcoxon rank sum test):** This test is a non-parametric test that allows two groups to be compared without making the assumption that values are following any specific distribution.

Let $X_1, X_2, \dots, X_n$ be a sample from some probability distribution F and $Y_1, Y_2, \dots, Y_m$ be a sample from some probability distribution G.

Let n_1 be the smaller sample size.
R = sum of the ranks from sample with size n_1

$$R' = n_1(m + n + 1) - R$$

- Step 1: $H_0: F = G$ vs $H_1: F \neq G$
- Step 2: $R^* = \min(R, R')$
- Step 3: Reject H_0 if $R^* \leq R_{table}$ at α level (Table 8 of Appendix B).

A Non-Parametric Method -The Mann-Whitney Test – Example 2

Table: Method A

Values	rank
79.98	7.5
80.04	19.0
80.02	11.5
80.04	19.0
80.03	15.5
80.03	15.5
80.04	19.0
⟶ 79.97	4.5
80.05	21.0
80.03	15.5
80.02	11.5
80.00	9.0
80.02	11.5

Table: Method B

Values	rank
80.02	11.5
79.94	1.0
79.98	~~7.5~~
⟶ 79.97	4.5
⟶ 79.97	4.5
80.03	15.5
79.95	2.0
⟶ 79.97	4.5

Comparing Paired Samples on Normal Distribution

Let $X_1, \cdots, X_n \sim N(\mu_X, \sigma_X^2)$ and $Y_1, \cdots, Y_n \sim N(\mu_Y, \sigma_Y^2)$ are dependent or paired. Define $D_i = X_i - Y_i,\ i = 1, \cdots, n$. Then

$$E(D_i) = \mu_X - \mu_Y = \mu_D \quad Var(D_i) = \sigma_D^2$$

If σ_D^2 is unknown, then $t = \frac{\bar{D} - \mu_D}{S_{\bar{D}}} \sim t_{(n-1)}$

If σ_D^2 is known $(or\ n \uparrow)$ $\quad Z = \frac{\bar{D} - \mu_D}{\sigma_{\bar{D}}} \sim N(0,1)$ and CI for μ_D is

$$\bar{D} \pm t_{(n-1)}(\alpha/2) S_{\bar{D}} \quad \text{and} \quad \bar{D} \pm Z_{\alpha/2} \sigma_{\bar{D}}$$

- $\overline{D}$ is the sample mean of the set $\{D_1, D_2, \ldots, D_n\}$.
- s_D is the sample standard deviation of the set $\{D_1, D_2, \ldots, D_n\}$.
- $s_{\overline{D}} = \frac{s_D}{\sqrt{n}}$

Comparing Paired Samples on Normal Distribution

Hypothesis Testing:

$$\begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D > 0 \end{cases} \qquad \begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D < 0 \end{cases} \qquad \begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D \neq 0 \end{cases}$$

Reject H_0 if: $\quad t > t_{(n-1)}(\alpha) \qquad t < -t_{(n-1)}(\alpha) \qquad |t| > t_{(n-1)}(\alpha/2)$

$$Z > Z_\alpha \qquad Z < -Z_\alpha \qquad |Z| > Z_{\alpha/2}$$

Comparing Paired Samples on Normal Distribution – Example 3

Before	*After*	*Difference*
25	27	2
25	29	4
27	37	10
44	56	12
30	46	16
67	82	15
53	57	4
53	80	27
52	61	9
60	59	1
28	43	15

$$\bar{D} = \frac{2+4+\cdots+15}{11} = 10.27$$

$$S_{\bar{D}} = \sqrt{\frac{1}{11-1}\left[(2-10.27)^2+\cdots+(15-10.27)^2\right]} = 2.40$$

90% CI: $\sigma_{\bar{D}}$ is unknown

$$t_{(n-1)}(\alpha/2) = t_{10}(0.05) = 1.812$$

$$\mu_{\bar{D}} \in \bar{D} \pm t_{(n-1)}(\alpha/2)S_{\bar{D}} = 10.27 \pm (1.812)(2.40)$$

$$\mu_{\bar{D}} \in (5.9, 14.6)$$

$$\alpha = 0.01 \quad \begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_{\bar{D}} \neq 0 \end{cases} \quad |t| = \left|\frac{\bar{D}-0}{2.40}\right| = 4.28 > t_{10}(0.005) = 3.169$$

Decision: reject H_0

A Non-Parametric Method - The Wilcoxon Signed Rank Test

This is a non parametric test based on the ranks of observations for two dependent random variables.

Let (X_i, Y_i) $i = 1, \cdots, n$ are paired observations with $D_i = X_i - Y_i$.

$$\begin{cases} H_o : \text{the distribution of } D_i \text{ is symmetric about zero} \\ H_1 : \text{Not } H_0 \end{cases}$$

Step 1: Obtain D_i and $|D_i|$

Step 2: rank $|D_i|$

Step 3: Assign the sign of D_i to the step 2

Step 4: W_+ =sum of ranks with + signs $\quad$ W_-=sum of ranks with - signs

Step 5: test statistic $W = \min(W_-, W_+)$

Step 6: reject H_0 if $W \leq W_{Table}$ at α level (Table 9 of Appendix B)

The Wilcoxon Signed Rank Test - Example 4

Before	*After*	D_i	$\lvert D_i \rvert$	$rank\lvert D_i \rvert$	*signed rank*
25	27	2	2	2	2
25	29	4	4	3.5	3.5
27	37	10	10	6	6
44	56	12	12	7	7
30	46	16	16	10	10
67	82	15	15	8.5	8.5
53	57	4	4	3.5	3.5
53	80	27	27	11	11
52	61	9	9	5	5
60	59	-1	1	1	-1
28	43	15	15	8.5	8.5

$$n = 11$$
$$W_+ = 60$$
$$W_- = 1$$

$$w = \min(1, 60) = 1 \quad \begin{cases} H_0 : \text{dist of } D_i \text{ is symm about zero} \\ H_1 : \text{Not } H_0 \text{(two-sided)} \end{cases}$$

$$W = 1 \leq 5 = W_{Table} \qquad \alpha = 0.01$$

Decision: reject H_0 at $\alpha = 0.01$ level

Summary

Comparing Two Independent Samples

Comparing Two Paired Samples

$$
\begin{cases}
\text{Population Normal} \begin{cases} \sigma \text{ known} \longrightarrow Z \\ \sigma \text{ unknown and n>30} \longrightarrow Z \\ \sigma \text{ unknown and n<30} \longrightarrow t \end{cases} \\
\text{Population Non-Normal} \begin{cases} n > 30 \ (n+m-2 > 30) \xrightarrow{CLT} Z \\ n < 30 \ (n+m-2 < 30) \longrightarrow * \end{cases}
\end{cases}
$$