Multivariate Analysis
Chapter 6: Inference for Several Sample

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. make inference for mean vectors from paired populations,
2. make inference for mean vectors from independent populations,
3. perform MANOVA.

Paired Data

Suppose that two "treatments" are administered to the same unit. For example,

- two drugs are administrated to the same person (drug 1 versus drug 2).
- a person participates in a vocational training program (before versus after).

Let (X_{j1}, X_{j2}) be the measurements recorded on the jth unit.

- The difference $D_j = X_{j1} - X_{j2}$ reflects the differential effects of the treatments.
- For paired observations, X_{j1} is often correlated with X_{j2}.
- But D_j can be independent of D_i for $j \neq i$.

Univariate Test for Paired Data

Let D_j represent independent observations from an $N\left(\delta, \sigma_d^2\right)$ distribution, where $\delta = \mathbb{E}\left(X_{j1} - X_{j2}\right)$. Then,

$$t = \frac{\bar{D} - \delta}{S_d/\sqrt{n}} \sim t_{n-1},$$

where

$$\bar{D} = \frac{1}{n}\sum_{j=1}^{n} D_j, \quad \text{and} \quad S_d^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(D_j - \bar{D}\right)^2.$$

- A level α test of $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ rejects H_0 if $|t| > t_{n-1}\left(\alpha/2\right)$.
- A $1-\alpha$ confidence interval for δ is

$$\bar{D} - t_{n-1}\left(\frac{\alpha}{2}\right)\frac{S_d}{\sqrt{n}} \leq \delta \leq \bar{D} + t_{n-1}\left(\frac{\alpha}{2}\right)\frac{S_d}{\sqrt{n}}.$$

Multivariate Extension

In the univariate case, X_{jk} corresponds to unit j under treatment k. The difference

$$D_j = X_{j1} - X_{j2} \sim N\left(\delta, \sigma_d^2\right).$$

In the multivariate case, let X_{jki} be the variable i of unit j under treatment k, and

$$D_{ji} = X_{j1i} - X_{j2i}.$$

Let $\boldsymbol{D}_j = \begin{bmatrix} D_{j1} & D_{j2} & \cdots & D_{jp} \end{bmatrix}^T$ and assume, for $j = 1, .., n$, that

$$\boldsymbol{D}_j \sim N_p\left(\boldsymbol{\delta}, \boldsymbol{\Sigma}_d\right).$$

Hotelling's T^2

Result 6.1

Let $\boldsymbol{D}_1$, $\boldsymbol{D}_2$, ..., $\boldsymbol{D}_n$ be a random sample from an $N_p\left(\boldsymbol{\delta}, \boldsymbol{\Sigma}_d\right)$ population. Then,

$$T^2 = n\left(\bar{\boldsymbol{D}} - \boldsymbol{\delta}\right)^T \boldsymbol{S}_d^{-1}\left(\bar{\boldsymbol{D}} - \boldsymbol{\delta}\right)$$

is distributed as an

$$\frac{(n-1)\,p}{n-p} F_{p,n-p}$$

random variable, whatever the true $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}_d$.
If n and $n-p$ are both large, T^2 is approximately distributed as a χ^2_p random variable, regardless of the form of the underlying population of differences.

From Previous Topics: Under Normality

- Consider testing $H_0 : \boldsymbol{\delta} = \mathbf{0}$ versus $H_1 : \boldsymbol{\delta} \neq \mathbf{0}$. A level α test rejects H_0 if

$$n\bar{\boldsymbol{d}}^T \boldsymbol{S}_d^{-1} \bar{\boldsymbol{d}} \quad > \quad \frac{(n-1)\,p}{n-p} F_{p,n-p}(\alpha)\,.$$

- A $1-\alpha$ confidence region for $\boldsymbol{\delta}$ consists of all $\boldsymbol{\delta}$ such that

$$n\left(\bar{\boldsymbol{D}} - \boldsymbol{\delta}\right)^T \boldsymbol{S}_d^{-1}\left(\bar{\boldsymbol{D}} - \boldsymbol{\delta}\right) \quad \leq \quad \frac{(n-1)\,p}{n-p} F_{p,n-p}(\alpha)\,.$$

- $1-\alpha$ simultaneous confidence intervals for δ_i are

$$\bar{D}_i - \sqrt{\frac{(n-1)\,p}{n-p} F_{p,n-p}(\alpha)}\sqrt{\frac{S_{d_i}^2}{n}} \leq \delta_i \leq \bar{D}_i + \sqrt{\frac{(n-1)\,p}{n-p} F_{p,n-p}(\alpha)}\sqrt{\frac{S_{d_i}^2}{n}},$$

  where $S_{d_i}^2$ is the ith diagonal element of $\boldsymbol{S}_d$.
- The Bonferroni $1-\alpha$ simultaneous confidence intervals for δ_i are

$$\bar{D}_i - t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{S_{d_i}^2}{n}} \leq \delta_i \leq \bar{D}_i + t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{S_{d_i}^2}{n}}.$$

From Previous Topics: Large Sample Inference

For n and $n-p$ large,

$$\frac{(n-1)\,p}{n-p}F_{p,n-p}(\alpha) \approx \chi^2_p(\alpha) \quad \text{and} \quad t_{n-1}\left(\frac{\alpha}{2p}\right) \approx z\left(\frac{\alpha}{2p}\right),$$

and normality needs not be assumed.

- A level α test rejects $H_0 : \boldsymbol{\delta} = \mathbf{0}$ if $n\bar{\boldsymbol{d}}^T \boldsymbol{S}_d^{-1}\bar{\boldsymbol{d}} > \chi^2_p(\alpha)$.
- A $1-\alpha$ confidence region for $\boldsymbol{\delta}$ consists of all $\boldsymbol{\delta}$ such that

$$n\left(\bar{\boldsymbol{D}} - \boldsymbol{\delta}\right)^T \boldsymbol{S}_d^{-1}\left(\bar{\boldsymbol{D}} - \boldsymbol{\delta}\right) \quad \leq \quad \chi^2_p(\alpha).$$

- $1-\alpha$ simultaneous confidence intervals for δ_i are

$$\bar{D}_i - \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{S^2_{d_i}}{n}} \leq \delta_i \leq \bar{D}_i + \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{S^2_{d_i}}{n}}.$$

- The Bonferroni $1-\alpha$ simultaneous confidence intervals for δ_i are

$$\bar{D}_i - z\left(\frac{\alpha}{2p}\right)\sqrt{\frac{S^2_{d_i}}{n}} \leq \delta_i \leq \bar{D}_i + z\left(\frac{\alpha}{2p}\right)\sqrt{\frac{S^2_{d_i}}{n}}.$$

Two Populations

When we have two measurements, it is not always the case that they are paired. Instead, they may come from two populations. In the latter case, we often assume

1. the sample $\boldsymbol{X}_{11}$, $\boldsymbol{X}_{12}$, ..., $\boldsymbol{X}_{1n_1}$ is a random sample of size n_1 from a p-variate population with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$,
2. the sample $\boldsymbol{X}_{21}$, $\boldsymbol{X}_{22}$, ..., $\boldsymbol{X}_{2n_2}$ is a random sample of size n_2 from a p-variate population with mean vector $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$,
3. $\boldsymbol{X}_{11}$, $\boldsymbol{X}_{12}$, ..., $\boldsymbol{X}_{1n_1}$ are independent of $\boldsymbol{X}_{21}$, $\boldsymbol{X}_{22}$, ..., $\boldsymbol{X}_{2n_2}$.

If n_1 and n_2 are small, we also need the assumptions

1. Both populations are multivariate normal,
2. $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$.

Pooled Sample Covariance Matrix

Suppose that $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$. Consider

$$
\begin{aligned}
\boldsymbol{S}_{\text{pooled}} &= \frac{\sum_{j=1}^{n_1} \left(\boldsymbol{X}_{1j} - \bar{\boldsymbol{X}}_1\right)\left(\boldsymbol{X}_{1j} - \bar{\boldsymbol{X}}_1\right)^T + \sum_{j=1}^{n_2} \left(\boldsymbol{X}_{2j} - \bar{\boldsymbol{X}}_2\right)\left(\boldsymbol{X}_{2j} - \bar{\boldsymbol{X}}_2\right)^T}{n_1 + n_2 - 2} \\
&= \frac{n_1 - 1}{n_1 + n_2 - 2}\boldsymbol{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\boldsymbol{S}_2.
\end{aligned}
$$

Then,

$$
\begin{aligned}
\mathbb{E}\left(\boldsymbol{S}_{\text{pooled}}\right) &= \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{\Sigma} + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{\Sigma} \\
&= \mathbf{\Sigma}.
\end{aligned}
$$

This means that $\boldsymbol{S}_{\text{pooled}}$ is an unbiased estimator of $\mathbf{\Sigma}$.

Two-Sample T^2 Statistic

For testing

$$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\delta}_0,$$

we can consider $\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - \boldsymbol{\delta}_0$. Under the independence assumption,

$$\operatorname{cov}\left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2\right) = \operatorname{cov}\left(\bar{\boldsymbol{X}}_1\right) + \operatorname{cov}\left(\bar{\boldsymbol{X}}_2\right) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \boldsymbol{\Sigma}.$$

We can estimate $\operatorname{cov}\left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2\right)$ by

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \boldsymbol{S}_{\text{pooled}}.$$

Hence, we can consider

$$T^2 = \left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - \boldsymbol{\delta}_0\right)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \boldsymbol{S}_{\text{pooled}}\right]^{-1} \left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - \boldsymbol{\delta}_0\right).$$

Distribution of Two-Sample T^2 Statistic

Result 6.2: Distribution and Confidence Region

Let $\boldsymbol{X}_{11}$, $\boldsymbol{X}_{12}$, ..., $\boldsymbol{X}_{1n_1}$ be a random sample of size n_1 from $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, and $\boldsymbol{X}_{21}$, $\boldsymbol{X}_{22}$, ..., $\boldsymbol{X}_{2n_2}$ be a random sample of size n_2 from $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Then

$$T^2 = \left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - \boldsymbol{\delta}\right)^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \boldsymbol{S}_{\text{pooled}}\right]^{-1} \left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - \boldsymbol{\delta}\right)$$

is distributed as

$$\frac{(n_1 + n_2 - 2)\, p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1},$$

where $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.
Consequently,

$$T^2 \leq \frac{(n_1 + n_2 - 2)\, p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1}(\alpha)$$

occurs with probability $1 - \alpha$.

Simultaneous Confidence Intervals

Result 6.3

Let

$$c^2 = \frac{(n_1+n_2-2)\,p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1}(\alpha).$$

With probability $1-\alpha$,

$$\boldsymbol{a}^T\left(\bar{\boldsymbol{X}}_1-\bar{\boldsymbol{X}}_2\right) \pm c\sqrt{\boldsymbol{a}^T\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\boldsymbol{S}_{\text{pooled}}\boldsymbol{a}}$$

will cover $\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ for all $\boldsymbol{a}$. In particular, a confidence interval for $\mu_{1i}-\mu_{2i}$ is

$$\bar{X}_{1i}-\bar{X}_{2i} \pm c\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_{ii,\text{pooled}}}.$$

Individual Interval

Under the normality assumption, we have

$$\begin{aligned}\bar{X}_{1i} - \bar{X}_{2i} &\sim N\left(\delta_i, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma_{ii}\right), \\ \frac{(n_1 + n_2 - 2)\, S_{ii,\text{pooled}}}{\sigma_{ii}} &\sim \chi^2_{n_1+n_2-2}.\end{aligned}$$

Hence,

$$\frac{\frac{\bar{X}_{1i} - \bar{X}_{2i} - \delta_i}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sigma_{ii}}}}{\sqrt{\frac{(n_1+n_2-2)S_{ii,\text{pooled}}}{\sigma_{ii}} / (n_1 + n_2 - 2)}} = \frac{\bar{X}_{1i} - \bar{X}_{2i} - \delta_i}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{ii,\text{pooled}}}} \sim t_{n_1+n_2-2}$$

A $1 - \alpha$ confidence interval for δ_i is

$$\bar{X}_{1i} - \bar{X}_{2i} \pm t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right)\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_{ii,\text{pooled}}}.$$

Bonferroni Interval

For each i, a $1-\alpha$ confidence interval for δ_i is

$$\bar{X}_{1i} - \bar{X}_{2i} \pm t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right)\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_{ii,\text{pooled}}}.$$

We can also apply the Bonferroni correction and the $1-\alpha$ simultaneous confidence intervals become

$$\bar{X}_{1i} - \bar{X}_{2i} \pm t_{n_1+n_2-2}\left(\frac{\alpha}{2p}\right)\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_{ii,\text{pooled}}}.$$

Two Samples When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$,

$$T^2 \quad \sim \quad \frac{(n_1 + n_2 - 2)\, p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1},$$

whose distribution does not depend on $\boldsymbol{\Sigma}$.
When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ we are unable to find a statistic whose distribution does not depend on the unknowns $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$.

- This is known as the Behrens-Fisher problem: Test equal mean of two normal populations when the population covariance matrices are unequal.

However, for n_1 and n_2 large, we can avoid the complexities due to unequal convariance matrices.

Large Sample Test When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

Result 6.4

Let the sample sizes be such that $n_1 - p$ and $n_2 - p$ are large. Then, an approximate $1 - \alpha$ confidence ellipsoid for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is given by all $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ satisfying

$$\left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - \boldsymbol{\delta}\right)^T \left[\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2\right]^{-1} \left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - \boldsymbol{\delta}\right) \leq \chi_p^2\left(\alpha\right).$$

The $1 - \alpha$ simultaneous confidence intervals for all linear combinations $\boldsymbol{a}^T\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)$ are provided by $\boldsymbol{a}^T\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)$ belongs to

$$\boldsymbol{a}^T\left(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2\right) \pm \sqrt{\chi_p^2\left(\alpha\right)}\sqrt{\boldsymbol{a}^T\left(\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2\right)\boldsymbol{a}}.$$

Compare Several Multivariate Population Means

Sometimes we have more than two populations as

$$\begin{aligned} \text{Population } 1: &\quad \boldsymbol{X}_{11}, \boldsymbol{X}_{12}, ..., \boldsymbol{X}_{1n_1} \\ \text{Population } 2: &\quad \boldsymbol{X}_{21}, \boldsymbol{X}_{22}, ..., \boldsymbol{X}_{2n_2} \\ &\vdots \\ \text{Population } p: &\quad \boldsymbol{X}_{p1}, \boldsymbol{X}_{p2}, ..., \boldsymbol{X}_{pn_p} \end{aligned}$$

Multivariate analysis of variance (MANOVA) is used to investigate whether the population mean vectors are the same. It becomes ANOVA if we only consider the univariate case.

Assumptions

We need the following assumptions:

1. $\boldsymbol{X}_{\ell 1}$, $\boldsymbol{X}_{\ell 2}$, ..., $\boldsymbol{X}_{\ell n_\ell}$ is a random sample of size n_ℓ , from a population with mean $\boldsymbol{\mu}_\ell$, $\ell = 1, 2, ..., g$. Then random sample from different populations are independent.
2. Each population is multivariate normal.
3. All populations have a common covariance matrix $\boldsymbol{\Sigma}$

MANOVA Model

The one-way MANOVA model for comparing g population mean vectors is

$$\boldsymbol{X}_{\ell j} = \underbrace{\boldsymbol{\mu}}_{\text{overall mean}} + \underbrace{\boldsymbol{\tau}_{\ell}}_{\text{treatment effect}} + \underbrace{\boldsymbol{e}_{\ell j}}_{\text{random eror}},$$

for $j = 1, 2, ..., n_{\ell}$ and $\ell = 1, 2, ..., g$, where $\boldsymbol{e}_{\ell j} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$.

- The MANOVA model is unidentified. We often impose the identification restriction $\sum_{\ell=1}^{g} n_{\ell}\boldsymbol{\tau}_{\ell} = \mathbf{0}$, or equivalent.
- Equal treatment effect means that $\boldsymbol{\tau}_{\ell} = \mathbf{0}$ for all ℓ because of the identification restriction.
- The sample decomposition is

$$\boldsymbol{x}_{\ell j} = \underbrace{\bar{\boldsymbol{x}}}_{\text{overall sample mean}} + \underbrace{\bar{\boldsymbol{x}}_{\ell} - \bar{\boldsymbol{x}}}_{\text{estimated treatment effect}} \underbrace{+\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}}_{\ell}.}_{\text{residual}}$$

Sum of Squares Decomposition

The total sum of squares and cross products satisfy

$$\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}\left(\boldsymbol{x}_{\ell j}-\bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_{\ell j}-\bar{\boldsymbol{x}}\right)^T = \boldsymbol{B}+\boldsymbol{W},$$

where

$$\begin{aligned}\boldsymbol{W} &= \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}\left(\boldsymbol{x}_{\ell j}-\bar{\boldsymbol{x}}_\ell\right)\left(\boldsymbol{x}_{\ell j}-\bar{\boldsymbol{x}}_\ell\right)^T \\ &= \sum_{\ell=1}^{g}\left(n_\ell-1\right)\boldsymbol{S}_\ell\end{aligned}$$

is the within sum of squares and cross products, and

$$\boldsymbol{B} = \sum_{\ell=1}^{g} n_\ell\left(\bar{\boldsymbol{x}}_\ell-\bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}_\ell-\bar{\boldsymbol{x}}\right)^T$$

is the between sum of squares and cross products.

One-Way MANOVA Table

Source of variation	Sum of Squares	Degrees of Freedom
Treatment	$\boldsymbol{B}$	$g-1$
Residual	$\boldsymbol{W}$	$\sum_{\ell=1}^{g}(n_\ell - 1)$
Total	$\boldsymbol{B}+\boldsymbol{W}$	$\sum_{\ell=1}^{g} n_\ell - 1$

Source of variation	Sum of Squares	Degrees of Freedom
Mean	$\sum_{\ell=1}^{g} n_\ell \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T$	1
Treatment	$\boldsymbol{B}$	$g-1$
Residual	$\boldsymbol{W}$	$\sum_{\ell=1}^{g}(n_\ell - 1)$
Total	$\boldsymbol{B}+\boldsymbol{W}$	$\sum_{\ell=1}^{g} n_\ell - 1$

Test Using Wilks' Lambda

If the treatment effects are equal,

$$\boldsymbol{B} = \sum_{\ell=1}^{g} n_\ell \left(\bar{\boldsymbol{x}}_\ell - \bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}_\ell - \bar{\boldsymbol{x}}\right)^T$$

should be dominated by

$$\boldsymbol{W} = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} \left(\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}}_\ell\right)\left(\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}}_\ell\right)^T.$$

We reject H_0 of equal treatment effects if the Wilks' lambda

$$\Lambda^* = \frac{\det\left(\boldsymbol{W}\right)}{\det\left(\boldsymbol{B} + \boldsymbol{W}\right)} = \frac{\det\left(\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}\left(\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}}_\ell\right)\left(\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}}_\ell\right)^T\right)}{\det\left(\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}\left(\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_{\ell j} - \bar{\boldsymbol{x}}\right)^T\right)}$$

is too small.

Distribution of Wilks' Lambda

Let $n = \sum_\ell n_\ell$. The sampling distributions are

p	g	Sampling distribution
$p = 1$	$g \geq 2$	$\frac{n-g}{g-1} \frac{1-\Lambda^*}{\Lambda^*} \sim F_{g-1,n-g}$ (ANOVA)
$p = 2$	$g \geq 2$	$\frac{n-g-1}{g-1} \frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}} \sim F_{2(g-1),2(n-g-1)}$
$p \geq 1$	$g = 2$	$\frac{n-p-1}{p} \frac{1-\Lambda^*}{\Lambda^*} \sim F_{p,n-p-1}$
$p \geq 1$	$g = 3$	$\frac{n-p-2}{p} \frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}} \sim F_{2p,2(n-p-2)}$

For other cases, with the Bartlett correction,

$$-\left(n - 1 - \frac{p+g}{2}\right) \log \Lambda^*$$

is approximately $\chi^2_{p(g-1)}$ if n is sufficiently large.

Estimate Σ

Let $X_{\ell i}$ be the ith variable of population ℓ, and $X_{\ell ij}$ be the jth unit of $X_{\ell i}$. Under the independence and normality assumption,

$$\begin{aligned} \frac{\sum_{j=1}^{n_\ell} \left(X_{\ell ij} - \bar{X}_{\ell i}\right)^2}{\sigma_{ii}} &\sim \chi^2_{n_\ell - 1}, \\ \frac{\sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} \left(X_{\ell ij} - \bar{X}_{\ell}\right)^2}{\sigma_{ii}} &\sim \chi^2_{n-g}. \end{aligned}$$

Hence, an estimator of σ_{ii} is

$$\frac{1}{n-g} W_{ii} = \frac{1}{n-g} \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} \left(X_{\ell ij} - \bar{X}_{\ell}\right)^2.$$

Individual Interval

$$\begin{aligned} \bar{X}_{ki} - \bar{X}_{\ell i} &\sim N\left(\tau_{ki} - \tau_{\ell i}, \left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\sigma_{ii}\right), \\ \frac{W_{ii}}{\sigma_{ii}} &\sim \chi^2_{n-g}. \end{aligned}$$

Hence,

$$\frac{\frac{\bar{X}_{ki} - \bar{X}_{\ell i} - (\tau_{ki} - \tau_{\ell i})}{\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\sigma_{ii}}}}{\sqrt{\frac{W_{ii}}{\sigma_{ii}} / (n-g)}} = \frac{\bar{X}_{ki} - \bar{X}_{\ell i} - (\tau_{ki} - \tau_{\ell i})}{\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right) W_{ii} / (n-g)}} \sim t_{n-g}.$$

A $1-\alpha$ confidence interval for $\tau_{ki} - \tau_{\ell i}$ is

$$\bar{X}_{ki} - \bar{X}_{\ell i} \pm t_{n-g}\left(\frac{\alpha}{2}\right)\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\frac{W_{ii}}{n-g}}.$$

Simultaneous Confidence Interval

Result 6.5: Bonferroni Interval

Let $n = \sum_{\ell=1}^{g} n_\ell$. For the MANOVA model

$$\boldsymbol{X}_{\ell j} = \boldsymbol{\mu} + \boldsymbol{\tau}_\ell + \boldsymbol{e}_{\ell j},$$

with confidence at least $1-\alpha$, $\tau_{ki} - \tau_{\ell i}$ belongs to

$$\bar{X}_{ki} - \bar{X}_{\ell i} \pm t_{n-g}\left(\frac{\alpha}{pg\left(g-1\right)}\right)\sqrt{\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\frac{W_{ii}}{n-g}}$$

for all components $i = 1, ..., p$ and all differences $\ell < k = 1, ..., g$. Here W_{ii} is the ith diagonal element of $\boldsymbol{W}$.

Multivariate Two-Way Fixed Effects Model with Interaction

A two-way MANOVA model is

$$\boldsymbol{X}_{\ell kr} = \boldsymbol{\mu} + \boldsymbol{\tau}_\ell + \boldsymbol{\beta}_k + \boldsymbol{\gamma}_{\ell k} + \boldsymbol{e}_{\ell kr},$$

for rth observation at level ℓ of factor 1 and level k of factor 2 by $X_{\ell kr}$, $\ell = 1, 2, ..., g$, $k = 1, 2, ..., b$, $r = 1, 2, ..., n$, and $\boldsymbol{e}_{\ell kr}$ are independent $N_p(0, \boldsymbol{\Sigma})$.

- Similar to one-way MANOVA, the two-way MANOVA model is unidentified. We often impose the identification restriction

$$\sum_{\ell=1}^{g} n_\ell \boldsymbol{\tau}_\ell = \sum_{k=1}^{b} \boldsymbol{\beta}_k = \sum_{\ell=1}^{g} \boldsymbol{\gamma}_{\ell k} = \sum_{k=1}^{b} \boldsymbol{\gamma}_{\ell k} = \boldsymbol{0}.$$

- $\boldsymbol{\gamma}_{\ell k}$ is the interaction term between factor 1 and factor 2.
- Here, we only consider the balanced case.

Sum of Squares Decomposition

The sum of squares can be decomposed as

$$
\begin{aligned}
& \sum_{\ell=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}\left(\boldsymbol{x}_{\ell kr}-\bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_{\ell kr}-\bar{\boldsymbol{x}}\right)^{T} & \\
= & \sum_{\ell=1}^{g} bn\left(\bar{\boldsymbol{x}}_{\ell\cdot}-\bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}_{\ell\cdot}-\bar{\boldsymbol{x}}\right)^{T} & \mathrm{SSP}_{\text{fac1}} \\
& +\sum_{k=1}^{b} gn\left(\bar{\boldsymbol{x}}_{\cdot k}-\bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}_{\cdot k}-\bar{\boldsymbol{x}}\right)^{T} & \mathrm{SSP}_{\text{fac2}} \\
& +\sum_{\ell=1}^{g}\sum_{k=1}^{b} n\left(\bar{\boldsymbol{x}}_{\ell k}-\bar{\boldsymbol{x}}_{\ell\cdot}-\bar{\boldsymbol{x}}_{\cdot k}+\bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}_{\ell k}-\bar{\boldsymbol{x}}_{\ell\cdot}-\bar{\boldsymbol{x}}_{\cdot k}+\bar{\boldsymbol{x}}\right)^{T} & \mathrm{SSP}_{\text{int}} \\
& +\sum_{\ell=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}\left(\boldsymbol{x}_{\ell kr}-\bar{\boldsymbol{x}}_{\ell k}\right)\left(\boldsymbol{x}_{\ell kr}-\bar{\boldsymbol{x}}_{\ell k}\right)^{T}, & \mathrm{SSP}_{\text{res}}
\end{aligned}
$$

where $\bar{\boldsymbol{x}}_{\ell\cdot}$ is the average for the ℓth level of factor 1, $\bar{\boldsymbol{x}}_{\cdot k}$ is the average for the kth level of factor 2.

Two-Way MANOVA Table

Source of variation	Sum of Squares	Degrees of Freedom
Factor 1	$\mathrm{SSP}_{\mathrm{fac1}}$	$g-1$
Factor 2	$\mathrm{SSP}_{\mathrm{fac2}}$	$b-1$
Interaction	$\mathrm{SSP}_{\mathrm{int}}$	$(g-1)(b-1)$
Residual	$\mathrm{SSP}_{\mathrm{res}}$	$gb(n-1)$
Total	$\sum_{\ell=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(\boldsymbol{x}_{\ell kr}-\bar{\boldsymbol{x}})(\boldsymbol{x}_{\ell kr}-\bar{\boldsymbol{x}})^T$	$ngb-1$

Test of No Interaction

The likelihood ratio test of

$$H_0 : \boldsymbol{\gamma}_{11} = \boldsymbol{\gamma}_{12} = \cdots = \boldsymbol{\gamma}_{gb} = \mathbf{0} \text{ (no interaction)}$$
$$H_1 : \text{at least one } \boldsymbol{\gamma}_{\ell k} \neq \mathbf{0}$$

rejects H_0 if the Wilks' lambda

$$\Lambda^* = \frac{|\mathrm{SSP}_{\mathrm{res}}|}{|\mathrm{SSP}_{\mathrm{int}} + \mathrm{SSP}_{\mathrm{res}}|}$$

is too small. For large enough sample size, we reject H_0 at the α level if

$$-\left[gb\left(n-1\right) - \frac{p+1-\left(g-1\right)\left(b-1\right)}{2}\right]\log\Lambda^* > \chi^2_{(g-1)(b-1)p}\left(\alpha\right).$$

Test of Main Effects

The likelihood ratio test of

$$H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \cdots = \boldsymbol{\tau}_g = \mathbf{0} \text{ (no interaction)}$$
$$H_1 : \text{at least one } \boldsymbol{\tau}_\ell \neq \mathbf{0}$$

rejects H_0 if the Wilks' lambda

$$\Lambda^* = \frac{|\mathrm{SSP}_{\mathrm{res}}|}{|\mathrm{SSP}_{\mathrm{fac1}} + \mathrm{SSP}_{\mathrm{res}}|}$$

is too small. For large enough sample size, we reject H_0 at the α level if

$$-\left[gb\left(n-1\right) - \frac{p+1-(g-1)}{2}\right] \log \Lambda^* > \chi^2_{(g-1)p}\left(\alpha\right).$$

Estimate $\boldsymbol{\Sigma}$

Let $X_{\ell kri}$ be the ith variable of rth unit at level ℓ of factor 1 and level k of factor 2. Under the independence and normality assumption,

$$
\begin{aligned}
\frac{\sum_{r=1}^{n}\left(X_{\ell kri}-\bar{X}_{\ell ki}\right)^{2}}{\sigma_{ii}} &\sim \chi^{2}_{n-1}, \\
\frac{\sum_{\ell=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}\left(X_{\ell kri}-\bar{X}_{\ell ki}\right)^{2}}{\sigma_{ii}} &\sim \chi^{2}_{gb(n-1)}.
\end{aligned}
$$

Hence, an estimator of σ_{ii} is

$$
\frac{1}{gb\left(n-1\right)}\mathrm{SSP}_{\mathrm{res},ii} = \frac{1}{n-g}\sum_{\ell=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}\left(X_{\ell kri}-\bar{X}_{\ell ki}\right)^{2}.
$$

Confidence Interval For $\tau_{\ell i} - \tau_{mi}$

$$
\begin{aligned}
\bar{X}_{\ell \cdot i} - \bar{X}_{m \cdot i} &\sim N\left(\tau_{\ell i} - \tau_{mi}, \frac{2\sigma_{ii}}{bn}\right), \\
\frac{\mathrm{SSP}_{\mathrm{res},ii}}{\sigma_{ii}} &\sim \chi^2_{gb(n-1)},
\end{aligned}
$$

where $\bar{X}_{\ell \cdot i}$ is the ith element of $\bar{\boldsymbol{X}}_{\ell \cdot}$. Hence,

$$
\frac{\frac{\bar{X}_{\ell \cdot i} - \bar{X}_{m \cdot i} - (\tau_{\ell i} - \tau_{mi})}{\sqrt{2\sigma_{ii}/(bn)}}}{\sqrt{\frac{\mathrm{SSP}_{\mathrm{res},ii}}{\sigma_{ii}} / \left[gb\left(n-1\right)\right]}} = \frac{\bar{X}_{\ell \cdot i} - \bar{X}_{m \cdot i} - (\tau_{\ell i} - \tau_{mi})}{\sqrt{\frac{2}{bn}\mathrm{SSP}_{\mathrm{res},ii} / \left[gb\left(n-1\right)\right]}} \sim t_{gb(n-1)}.
$$

A $1-\alpha$ confidence interval for $\tau_{\ell i} - \tau_{mi}$ is

$$
\bar{X}_{\ell \cdot i} - \bar{X}_{m \cdot i} \pm t_{gb(n-1)}\left(\frac{\alpha}{2}\right)\sqrt{\frac{2}{bn}\mathrm{SSP}_{\mathrm{res},ii} / \left[gb\left(n-1\right)\right]}.
$$

Simultaneous Confidence Interval

A $1-\alpha$ Bonferroni simultaneous confidence interval for $\tau_{\ell i}-\tau_{mi}$ is

$$\bar{X}_{\ell \cdot i}-\bar{X}_{m \cdot i} \pm t_{gb(n-1)}\left(\frac{\alpha}{pg\left(g-1\right)}\right)\sqrt{\frac{2}{bn}\mathrm{SSP}_{\mathrm{res},ii}/\left[gb\left(n-1\right)\right]}.$$

A $1-\alpha$ Bonferroni simultaneous confidence interval for $\beta_{ki}-\beta_{mi}$ is

$$\bar{X}_{\cdot ki}-\bar{X}_{\cdot mi} \pm t_{gb(n-1)}\left(\frac{\alpha}{pb\left(b-1\right)}\right)\sqrt{\frac{2}{gn}\mathrm{SSP}_{\mathrm{res},ii}/\left[gb\left(n-1\right)\right]}.$$