Chapter 6
Comparing Mean Vectors from Two Populations

Analysts may have data on samples from different populations and may be interested in answering questions as to whether the populations under consideration have different means (or mean vectors in the multivariate case)

Two independent mean differences

(1) When the two populations have equal variances

(A) Univariate case: (small n_1, n_2)

$X_{11}, X_{12}, X_{13}, \dots, X_{1n}$ is a random sample of size n_1 from $N(\mu_1, \sigma_1^2)$

$X_{21}, X_{22}, X_{23}, \dots, X_{2n}$ is a random sample of size n_2 from $N(\mu_2, \sigma_2^2)$

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (Perform Levene's $\overset{F\text{ test}}{\wedge}$ to check this)

Test statistic for the test $H_0: \mu_1 - \mu_2 = \delta_0$ vs $H_a: \mu_1 - \mu_2 \neq \delta_0$ is given as

$$t^* = \frac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\sqrt{s^2_{pooled}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t\left(\alpha/2, n_1 + n_2 - 2\right)$$

where $s^2_{pooled} = \dfrac{S_1^2(n_1-1) + S_2^2(n_2-1)}{n_1+n_2-2}$

with

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad \text{and} \quad S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$$

The $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$ is given by

$$\bar{x}_1-\bar{x}_2 \pm t(\alpha/2, n_1+n_2-2)\sqrt{s^2_{pooled}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

(B) Multivariate case (small n_1, n_2): Assume $\underset{\sim}{x}_{11}, \underset{\sim}{x}_{12}, \ldots, \underset{\sim}{x}_{1n} \sim N_p(\mu_1, \Sigma_1)$
$\underset{\sim}{x}_{21}, \underset{\sim}{x}_{22}, \ldots, \underset{\sim}{x}_{2n} \sim N_p(\mu_2, \Sigma_2)$

To test equality of variance covariance matrices use the Box M test which test $H_0: \Sigma_1 = \Sigma_2$

Will be demonstrated in R.

If the test concludes that $\Sigma_1 = \Sigma_2 = \Sigma$, we have

$\sum_{j=1}^{n_1}(\underset{\sim}{x}_{1j}-\underset{\sim}{\bar{x}}_1)(\underset{\sim}{x}_{1j}-\underset{\sim}{\bar{x}}_1)^T$ as an estimate of $(n_1-1)\Sigma$ and

$\sum_{j=1}^{n_2}(\underset{\sim}{x}_{2j}-\underset{\sim}{\bar{x}}_2)(\underset{\sim}{x}_{2j}-\underset{\sim}{\bar{x}}_2)^T$ as an estimate of $(n_2-1)\Sigma$

so that the pooled covariance matrix is calculated as

$$S_p = S_{pooled} = \frac{\sum_{j=1}^{n_1} (\underset{\sim}{x}_{1j} - \bar{\underset{\sim}{x}}_1)(\underset{\sim}{x}_{1j} - \bar{\underset{\sim}{x}}_1)^T + \sum_{j=1}^{n_2} (\underset{\sim}{x}_{2j} - \bar{\underset{\sim}{x}}_2)(\underset{\sim}{x}_{2j} - \bar{\underset{\sim}{x}}_2)^T}{n_1 + n_2 - 2}$$

$$= \frac{n_1 - 1}{n_1 + n_2 - 2} S_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2$$

where $S_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (\underset{\sim}{x}_{1j} - \bar{\underset{\sim}{x}}_1)(\underset{\sim}{x}_{1j} - \bar{\underset{\sim}{x}}_1)^T$ and

$S_2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (\underset{\sim}{x}_{2j} - \bar{\underset{\sim}{x}}_2)(\underset{\sim}{x}_{2j} - \bar{\underset{\sim}{x}}_2)^T$

To test

$H_0: \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 = \underset{\sim}{\delta}_0$ vs $H_a: \underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 \neq \underset{\sim}{\delta}_0$

Test statistic:

$$T^2 = (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2 - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)) \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pooled} \right]^{-1} (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2 - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2))$$

$$(T^*)^2 = \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - 2)p} T^2 \sim F(\alpha, p, n_1 + n_2 - p - 1)$$

Reject H_0 if $(T^*)^2 > F(\alpha, p, n_1 + n_2 - p - 1)$

The $100(1-\alpha)\%$ confidence region for $\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$ is given by all $\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$ satisfying:

$$(\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2 - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2))^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pooled} \right]^{-1} ((\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2) - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)) \leq$$

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

Note that this region is an ellipsoid centered at the observed $\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2$ and whose axes are determined by the eigen values and eigen vectors of S_{pooled}.

Refer to example 6.3 (Problem 4 in example document)

© Simultaneous confidence intervals for $\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$

We construct these confidence intervals by considering all possible linear combinations of the differences in mean vectors.

The $100(1-\alpha)\%$ simultaneous confidence interval for $\underset{\sim}{a}^T(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$ for all a is given as:

$$\underset{\sim}{a}^T(\bar{\underset{\sim}{X}}_1 - \bar{\underset{\sim}{X}}_2) \pm c\sqrt{\underset{\sim}{a}^T\left(\frac{1}{n_1}+\frac{1}{n_2}\right)S_{pooled}\,\underset{\sim}{a}}$$

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

$1-\alpha$ simultaneous

In particular, we have that the $100(1-\alpha)\%$ confidence interval for $\mu_{1i} - \mu_{2i}$ is given by:

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

The Bonferroni $100(1-\alpha)\%$ simultaneous confidence intervals for the p population mean differences for $\mu_{1i} - \mu_{2i}$ is given as:

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

② Independent samples with $\Sigma_1 \neq \Sigma_2$

Ⓐ Univariate case

X_{1i} $i=1,2,\dots,n$ $\sim N(\mu_1, \sigma_{11})$

X_{2i} $i=1,2,\dots,n$ $\sim N(\mu_2, \sigma_{22})$

$\sigma_{11} \neq \sigma_{22}$ (from Levene's test)

The test of $H_0: \mu_1 - \mu_2 = \delta_0$ vs $H_a: \mu_1 - \mu_2 \neq \delta_0$ has test statistic

$$\frac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(\alpha/2, \nu)$$

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$ (Round this down ALWAYS)

(B) Multivariate case: Independent samples, $\Sigma_1 \neq \Sigma_2$

Result: (When n_1 and n_2 are large)

Let the sample sizes be such that $n_1 - p$ and $n_2 - p$ are large. Then an approximate $100(1-\alpha)\%$ confidence ellipsoid for $\mu_1 - \mu_2$ is given by all $\mu_1 - \mu_2$ satisfying

$$[\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)]^T \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right]^{-1} [\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)] \leq \chi^2(\alpha, p)$$

Test statistic. Reject H_0 if T.S $> \chi^2(\alpha, p)$

- $100(1-\alpha)\%$ simultaneous confidence intervals for all linear combinations $a^T(\mu_1 - \mu_2)$ is given by

$$a^T(\bar{x}_1 - \bar{x}_2) \pm \sqrt{\chi^2(\alpha, p)} \sqrt{a^T\left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right) a}$$

Book example 6.5 (Problem 5 in example document)

When sample sizes are not large

2 populations need to be MVN

n_1 and n_2 each have to be greater than p

$H_0: \mu_1 - \mu_2 = \delta_0$ $\quad$ $H_a: \mu_1 - \mu_2 \neq \delta_0$

Test statistic

$$T^2 = (\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2))^T \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right]^{-1} (\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2))$$

$$(T^*)^2 = \frac{\nu - p + 1}{\nu p} T^2$$

Reject H_0 if $(T^*)^2 > F(\alpha, p, \nu - p + 1)$

ν can be calculated as

$$\frac{p + p^2}{\sum_{i=1}^{2} \frac{1}{n_i} \left\{ tr\left[\left(\frac{1}{n_i} S_i \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} \right)^2 \right] + \left(tr\left[\frac{1}{n_i} S_i \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} \right] \right)^2 \right\}}$$

where $\min(n_1, n_2) \leq \nu \leq n_1 + n_2$

The approximate $100(1-\alpha)\%$ confidence region is given by all $\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$ such that

$$(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2 - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2))^T \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right]^{-1} (\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2 - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)) \leq$$

$$\frac{\nu p}{\nu - p + 1} F(\alpha, p, \nu - p + 1)$$

where ν is as given before.

Book example 6.6 (Problem 6 in class example)

Dependent Populations

A) Two dependent populations:

- Previously, we conducted hypothesis testing about the difference between two population means when the two samples are drawn independently from two different populations
- Here we deal with inferences for the mean of dependent populations which are also called paired populations.
- Use this test when observations from the two populations of interest are collected in pairs.

Check examples document for an example on when you have paired data.

Let X_{j1} denote the response to treatment 1 (or the response before treatment), and let X_{j2} denote the response to treatment 2 (or response after treatment) for the jth trial.

Note: (X_{j1}, X_{j2}) are measurements recorded on the jth unit or jth pair of like units.

We analyze the n differences:

$$D_j = X_{j1} - X_{j2} \quad , j = 1, 2, \cdots, n$$

The differences D_j represent independent observations from $N(\delta, \sigma_d^2)$ where δ is the population mean difference between the 2 treatments, ie $E(X_{j1} - X_{j2})$

To test:

$H_0: \delta = 0$ (zero mean difference for treatments

$H_1: \delta \neq 0$

Test statistic: $t = \dfrac{\bar{D} - \delta}{S_d/\sqrt{n}} \sim t(\alpha/2, n-1)$

where $\bar{D} = \frac{1}{n}\sum_{j=1}^{n} D_j$ and $S_d^2 = \frac{1}{n-1}\sum_{j=1}^{n}(D_j - \bar{D})^2$

Reject H_0 if $|t| > t(\alpha/2, n-1)$

A $100(1-\alpha)\%$ confidence interval for the mean differences $\delta = E(X_{j1} - X_{j2})$ is given by:

$$\bar{D} \pm t(\alpha/2, n-1)\, S_d/\sqrt{n}$$

Check example (Problem 3) in example document

B) Multivariate Case

- p variables, n experimental units.
- The jth experimental unit has p observations given as

X_{1j1} = variable 1 under treatment 1
X_{1j2} = " 2 " "
:
X_{1jp} = " p under treatment 1

} Population 1

X_{2j1} = Variable 1 under treatment 2
X_{2j2} = " 2 " 2
⋮
X_{2jp} = " p under treatment 2

} Population 2

- The paired difference random variables becomes:

$$D_{j1} = X_{1j1} - X_{2j1}$$
$$D_{j2} = X_{1j2} - X_{2j2}$$
$$\vdots$$
$$D_{jp} = X_{1jp} - X_{2jp}$$

So for the jth experimental unit, we have

$$\underset{\sim}{D}_j^T = (D_{j1}\ D_{j2}\ \cdots\ D_{jp}),$$

$$E(\underset{\sim}{D}_j) = \underset{\sim}{\delta} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_p \end{pmatrix} \quad \text{and} \quad Cov(\underset{\sim}{D}_j) = \Sigma_d \quad j = 1, 2, \ldots, n$$

$\underset{\sim}{D}_1, \underset{\sim}{D}_2, \ldots, \underset{\sim}{D}_n$ are independent $N_p(\underset{\sim}{\delta}, \Sigma_d)$ random vectors.
Given the observed differences $\underset{\sim}{d}_j^T = (d_{j1}\ d_{j2}\ \cdots\ d_{jp})$, $j = 1, 2, \ldots, n$ from $\underset{\sim}{D}_1, \ldots, \underset{\sim}{D}_n$, to test:

$H_0: \underset{\sim}{\delta} = \underset{\sim}{0}$ vs $H_a: \underset{\sim}{\delta} \neq \underset{\sim}{0}$

Test statistic: $T^2 = n(\bar{\underset{\sim}{d}} - \underset{\sim}{\delta})^T S_d^{-1} (\bar{\underset{\sim}{d}} - \underset{\sim}{\delta})$

where

If $T^2 > \frac{(n-1)p}{n-p} f(\alpha, p, n-p)$, reject H_0.

A $100(1-\alpha)\%$ confidence region for $\underset{\sim}{\delta}$ consists of all $\underset{\sim}{\delta}$ such that

$$(\bar{\underset{\sim}{D}} - \underset{\sim}{\delta})^T S_d^{-1} (\bar{\underset{\sim}{D}} - \underset{\sim}{\delta}) \leq \frac{(n-1)p}{n(n-p)} F(\alpha, p, n-p)$$

Also, $100(1-\alpha)\%$ simultaneous confidence intervals for the individual mean differences δ_i are given by

$$\bar{d}_i \pm \sqrt{\frac{(n-1)p}{(n-p)} F(\alpha, p, n-p)} \sqrt{\frac{s_{d_i}^2}{n}} \quad ; i = 1, 2, \ldots, p$$

where $\bar{d}_i$ is the ith element of $\bar{d}$ and $s_{d_i}^2$ is the ith diagonal element of S_d

- The Bonferroni $100(1-\alpha)\%$ simultaneous confidence intervals for q individual mean differences are

$$\bar{d}_i \pm t(\alpha/2q, n-1) \sqrt{\frac{s_{d_i}^2}{n}}$$

Note: (1) If n and $n-p$ are both large, T^2 is approximately $\chi^2(\alpha, p)$ regardless of the form of the underlying population of the differences.

(2) for $n-p$ large, $\frac{(n-1)p}{n-p} F(\alpha, p, n-p) \cong \chi^2(\alpha, p)$ and normality need not be assumed.

Refer to example document (Problem 7) - Book example 6.1

Question: What if we have more than 2 populations?

Comparing several population means

Univariate ANOVA

Suppose we are interested in comparing 3 or more population means. As usual, we can draw independent samples from each population, calculate their sample means and use them in the analyses.

Eg. ① We want to compare average weight of chickens from 4 different poultry farms [One variable of interest: weight, 4 populations: Location 1-4]

② Checking differences in average maths scores for high schools in the Madison county.

—

Now, suppose we wish to compare means of g different populations, we can test

$H_0: \mu_1 = \mu_2 = \cdots = \mu_g$ against

H_a: At least one group mean is different from the others.

Assumption: $X_{i1}, X_{i2}, \ldots, X_{n_i}$ is a random sample from $N(\mu_i, \sigma^2)$ (Constant variance) $i = 1, 2, \ldots, g$ and the random samples from each population are independent.

- Each population mean, μ_i can be decomposed into the overall mean (grand mean) and a component which is due to the specific population, that is:

$$\mu_i = \mu + \tau_i$$

(μ_i: ith population mean; μ: Grand mean; τ_i: ith population treatment effect)

H_0 and H_a can be written in terms of the individual treatment effects as

H_0: $\tau_1 = \tau_2 = \dots = \tau_g = 0$ (There is no treatment effect)

H_a: At least one of the $\tau_i \neq 0$ (There is a possible treatment effect)

- Each individual observation X_{ij} can be decomposed into

$$X_{ij} = \mu_i + e_{ij} = \underbrace{\mu}_{\text{overall mean}} + \underbrace{\tau_i}_{\text{treatment effect}} + \underbrace{e_{ij}}_{\text{random error}} \qquad (\mu_i - \mu)$$

Subject to $\sum_{i=1}^{g} n_i \tau_i = 0$ · $e_{ij} \sim N(0, \sigma^2)$ (independent)

- We estimate the model from the sample as follows:

$$\underbrace{X_{ij}}_{\text{Observation}} = \underbrace{\bar{X}}_{\substack{\text{overall sample mean} \\ \text{- estimate of } \mu}} + \underbrace{(\bar{X}_i - \bar{X})}_{\substack{\text{estimated treatment effect} \\ \text{- estimate of } \tau_i}} + \underbrace{(X_{ij} - \bar{X})}_{\substack{\text{residual} \\ \text{estimate of } e_{ij}}}$$

- Question: How much of the total variation in the data comes from variation within the group or variation between the group?

Example (Check example ___ in examples document)

	Observations				
Population 1:	9	6	9	$n_1 = 3$	$\bar{X}_1 = 8$
Population 2:	0	2		$n_2 = 2$	$\bar{X}_2 = 1$
Population 3:	3	1	2	$n_3 = 3$	$\bar{X}_3 = 2$

Overall mean: $\frac{9+6+9+0+2+3+1+2}{8} = 4 = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3}{n_1+n_2+n_3}$

We wish to find the 3 Sum of squares

a) SST = Sum of squares total
= Sum of squared deviation of each observation from the overall mean
$= (9-4)^2 + (6-4)^2 + (9-4)^2 + (0-4)^2 + (2-4)^2 + (3-4)^2 + (1-4)^2 + (2-4)^2$

$= 88$

$= \sum_{i=1}^{g} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$

= SS_{cor} (Total corrected SS)

$df(SS_{cor}) = \left(\sum_{i=1}^{g} n_i\right) - 1$ — Why?

(b) SSW = "Sum of squares within the individual groups"
= Sum of squared deviation of each observation from its corresponding population mean.
$= (9-8)^2 + (6-8)^2 + (9-8)^2 + (0-2)^2 + (2-2)^2 + (3-2)^2 + (1-2)^2 + (2-2)^2$

$= 10$

$= \sum_{i=1}^{g} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$

= SS_{res}

$df(SS_{res}) = \left(\sum_{i=1}^{g} n_i\right) - g$ — Why?

c) SSB = "Sum of squares between the groups"

$= (8-4)^2 + (8-4)^2 + (8-4)^2 + (2-4)^2 + (2-4)^2 + (3-4)^2 + (1-4)^2 + (2-4)^2$

$= 78$

$= \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})^2$

$= SS_{tr}$

$df(SS_{tr}) = g-1$

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i} x_{ij}^2}_{(SS_{obs})} = \underbrace{SS_{tr}}_{(SS_{mean})} + \underbrace{\sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})^2}_{(SS_{tr})} + \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}_{(SS_{res})}$$

ANOVA Table

Source of Variation	Sum of squares (SS)	Degrees of freedom (df)
Treatments	$SS_{tr} = \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})^2$	$g-1$
Residual (error)	$SS_{res} = \sum_{i=1}^{g}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$	$\sum_{i=1}^{g} n_i - g$
Total	$SS_{cor} = \sum_{i=1}^{g}\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$	$\sum_{i=1}^{g} n_i - 1$

$$F^* = \frac{SS_{tr}/(g-1)}{SS_{res}/(\sum_{i=1}^{g} n_i - g)} \sim F(\alpha, g-1, \sum_{i=1}^{g} n_i - g)$$

Reject H_0 if $F^* > F(\alpha, g-1, \sum_{i=1}^{g} n_i - g)$

Back to our example (Testing at $\alpha = 0.01$)

$$F^* = \frac{78/2}{10/5} = 19.5$$

$F^* = 19.5 > F(0.01, 2, 5) = 13.27$ so reject H_0.

Multivariate ANOVA

- Used when comparing more than 2 populations.
- Random samples from each of g populations are arranged as

Population 1 : $\underset{\sim}{X}_{11}, \underset{\sim}{X}_{12}, \cdots, \underset{\sim}{X}_{1n_1}$

Population 2 : $\underset{\sim}{X}_{21}, \underset{\sim}{X}_{22}, \cdots, \underset{\sim}{X}_{2n_2}$

$\vdots$ $\vdots$

Population g : $\underset{\sim}{X}_{g1}, \underset{\sim}{X}_{g2}, \cdots, \underset{\sim}{X}_{gn_g}$

- We use MANOVA to investigate whether the population mean vectors are the same. If not, there are methods to investigate further which mean components differ significantly.

Assumptions about the structure of the data for one-way MANOVA

1. $\underset{\sim}{X}_{l1}, \underset{\sim}{X}_{l2}, \ldots, \underset{\sim}{X}_{ln_l}$ is a random sample of size n_l from a population with mean $\underset{\sim}{\mu}_l$, $l=1,2,\ldots,g$. The random samples from the different populations are independent.
2. All populations have a common covariance matrix, Σ.
3. Each population is multivariate normal. (Can be relaxed when sample sizes n_l are large.

Refer to problem 9 (Book example 6.9)

Model

$$\underset{\sim}{X}_{lj} = \underset{\sim}{\mu} + \underset{\sim}{\tau}_l + \underset{\sim}{e}_{lj}, \quad j=1,2,\ldots,n_l \text{ and } l=1,2,\ldots,g$$

subject to $\sum_{l=1}^{g} n_l \underset{\sim}{\tau}_l = 0$

- $\underset{\sim}{e}_{lj} \sim N_p(0, \Sigma)$
- $\underset{\sim}{\mu}$ is the overall mean,
- $\underset{\sim}{\tau}_l$ is the lth treatment effect

The model is estimated from the sample as:

$$\underset{\sim}{X}_{lj} = \underset{\sim}{\bar{X}} + (\underset{\sim}{\bar{X}}_l - \underset{\sim}{\bar{X}}) + (\underset{\sim}{X}_{lj} - \underset{\sim}{\bar{X}}_l)$$

(Observation) (overall mean) + (estimated treatment effect, $\hat{\tau}_l$) + (residual $\hat{e}_{lj}$)

As in the univariate case, the total sum of squares is decomposed into the sum of squares between groups and the sum of squares within groups.

MANOVA Table for Comparing Population Mean Vectors

Source of variation	Matrix of sum of squares and cross products	Degrees of freedom
Treatment (Between)	$B = \sum_{i=1}^{g} n_i (\bar{\underset{\sim}{x}}_i - \bar{\underset{\sim}{x}})(\bar{\underset{\sim}{x}}_i - \bar{\underset{\sim}{x}})^T$ $= (n_1-1)S_1 + (n_2-1)S_2 + \cdots + (n_g-1)S_g$	$g-1$
Residual (Error)	$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}}_i)(\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}}_i)^T$	$\sum_{i=1}^{g} n_i - g$
Total	$B + W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_{ij} - \bar{\underset{\sim}{x}})^T$	$\sum_{i=1}^{g} n_i - 1$

Testing $H_0: \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \cdots = \underset{\sim}{\mu}_g$

$H_a: \underset{\sim}{\mu}_i \neq \underset{\sim}{\mu}_j$ for at least one $i \neq j$

We calculate

$$\Lambda^* = \frac{|W|}{|B+W|} = \frac{\left| \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}}_{i})(\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}}_{i})^T \right|}{\left| \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}})(\underset{\sim}{X}_{ij} - \overline{\underset{\sim}{X}})^T \right|}$$

Λ^* is known as Wilk's lambda

- The exact distribution of Λ^* can be derived for the special cases as listed in the table below:

for any other case and large sample sizes, (i.e $\left(\sum_{i=1}^{g} n_i = n\right)$) a modification of Λ^* due to Bartlett can be used in the hypothesis test.

$$-\left(n-1-\frac{p+g}{2}\right)\ln\Lambda^* = -\left(n-1-\frac{p+g}{2}\right)\ln\left(\frac{|W|}{|B+W|}\right)$$

$$\sim \chi^2(\alpha, p(g-1))$$

Reject H_0 if

$$-\left(n-1-\frac{p+g}{2}\right)\ln\left(\frac{|W|}{|B+W|}\right) > \chi^2(\alpha, p(g-1))$$