

Lecture 6

Comparisons of Several Mean Vectors

DSA 8070 Multivariate Analysis

September 26 - September 30, 2022

Whitney Huang
Clemson University

Agenda

Comparisons of
Several Mean Vectors

CLEMSON
UNIVERSITY

Comparisons of Two
Mean Vectors

Multivariate Analysis of
Variance

1 Comparisons of Two Mean Vectors

2 Multivariate Analysis of Variance

Motivating Example: Swiss Bank Notes (Source: PSU stat 505)

Suppose there are two distinct populations for 1000 franc Swiss Bank Notes:

- The first population is the population of Genuine Bank Notes
- The second population is the population of Counterfeit Bank Notes

For both populations the following measurements were taken:

- 1 Length of the note
- 2 Width of the Left-Hand side of the note
- 3 Width of the Right-Hand side of the note
- 4 Width of the Bottom Margin
- 5 Width of the Top Margin
- 6 Diagonal Length of Printed Area

We want to determine if counterfeit notes can be distinguished from the genuine Swiss bank notes

Review: Two Sample t-Test

Suppose we have data from a single variable from population 1: $X_{11}, X_{12}, \dots, X_{1n_1}$ and population 2: $X_{21}, X_{22}, \dots, X_{2n_2}$. Here we would like to draw inference about their population means μ_1 and μ_2 .

Assumptions:

- **Homoskedasticity**: The data from both populations have common variance σ^2
- **Independence**: The subjects from both populations are independently sampled. Also $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{2j}\}_{j=1}^{n_2}$ are independent to each other
- **Normality**: The data from both populations are normally distributed (not that crucial for “large” sample)

Here we are going to consider testing $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$

Review: Two Sample t-Test

We define the sample means for each population using the following expression:

$$\bar{x}_1 = \frac{\sum_{j=1}^{n_1} x_{1j}}{n_1}, \quad \bar{x}_2 = \frac{\sum_{j=1}^{n_2} x_{2j}}{n_2}.$$

We denote the sample variance

$$s_1^2 = \frac{\sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)^2}{n_1 - 1}, \quad s_2^2 = \frac{\sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2}{n_2 - 1}.$$

Under the **homoskedasticity** assumption, we can “pool” two samples to get the pooled sample variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \stackrel{H_0}{\sim} t_{n_1 + n_2 - 2}$$

We can use this result to construct confidence intervals and to perform hypothesis tests

The Two Sample Problem: The Multivariate Case

Now we would like to use two independent samples $\{X_{11}, \dots, X_{12}, \dots, X_{1n_1}\}$ and $\{X_{21}, \dots, X_{22}, \dots, X_{2n_2}\}$, where

$$X_{ij} = \begin{bmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{bmatrix}$$

to infer the relationship between μ_1 and μ_2 , where

$$\mu_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{ip} \end{bmatrix}$$

Assumptions

- Both populations have common covariance matrix, i.e.,
 $\Sigma_1 = \Sigma_2$
- Independence**: The subjects from both populations are independently sampled
- Normality**: Both populations are normally distributed

The Multivariate Two-Sample Problem

Here we are testing

$$H_0 : \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1p} \end{bmatrix} = \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ \vdots \\ \mu_{2p} \end{bmatrix}, \quad H_a : \mu_{1k} \neq \mu_{2k} \text{ for at least one } k \in \{1, 2, \dots, p\}$$

Under the **common covariance** assumption we have

$$\mathbf{S}_p = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2},$$

where

$$\mathbf{S}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T, \quad i = 1, 2$$

The Two-Sample Hotelling's T-Square Test Statistic

The two-sample t test is equivalent to

$$t^2 = (\bar{x}_1 - \bar{x}_2)^T \left[s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]^{-1} (\bar{x}_1 - \bar{x}_2).$$

Under H_0 , $t^2 \sim F_{1, n_1 + n_2 - 2}$. We can use this result to perform a hypothesis test

We can extend this to the multivariate situation:

$$T^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \left[\mathbf{S}_p \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

Under H_0 , we have

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

We can use this result to perform inferences for multivariate cases

Two-Sample Test for Swiss Bank Notes

```
> (xbar1 <- colMeans(dat[real, -1]))
      V2      V3      V4      V5      V6      V7
214.969 129.943 129.720   8.305  10.168 141.517
> (xbar2 <- colMeans(dat[fake, -1]))
      V2      V3      V4      V5      V6      V7
214.823 130.300 130.193  10.530  11.133 139.450
> Sigma1 <- cov(dat[real, -1])
> Sigma2 <- cov(dat[fake, -1])
> n1 <- length(real); n2 <- length(fake); p <- dim(dat[, -1])[2]
> Sp <- ((n1 - 1) * Sigma1 + (n2 - 1) * Sigma2) / (n1 + n2 - 2)
> # Test statistic
> T.squared <- as.numeric(t(xbar1 - xbar2) %*% solve(Sp * (1 / n1 + 1
/ n2)) %*% (xbar1 - xbar2))
> Fobs <- T.squared * ((n1 + n2 - p - 1) / ((n1 + n2 - 2) * p))
> # p-value
> pf(Fobs, p, n1 + n2 - p - 1, lower.tail = F)
[1] 3.378887e-105
```

Conclusion

The counterfeit notes can be distinguished from the genuine notes on at least one of the measurements \Rightarrow which ones?

$$\bar{x}_{1k} - \bar{x}_{2k} \pm \sqrt{\frac{p(n_1 + n_2 - 2)}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1, \alpha}} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_{k,p}^2},$$

where $s_{k,p}^2$ is the pooled variance for the variable k

Variable	95% CI
Length of the note	(-0.04, 0.34)
Width of the Left-Hand note	(-0.52, -0.20)
Width of the Right-Hand note	(-0.64, -0.30)
Width of the Bottom Margin	(-2.70, -1.75)
Width of the Top Margin	(-1.30, -0.63)
Diagonal Length of Printed Area	(1.81, 2.33)

Assumptions:

- **Homoskedasticity:** The data from both populations have common covariance matrix Σ

Will return to this in next slide

- **Independence:**

This assumption may be violated if we have clustered, time-series, or spatial data

- **Normality:**

Multivariate QQplot, univariate histograms, bivariate scatter plots

Testing for Equality of Mean Vectors when $\Sigma_1 \neq \Sigma_2$

- Bartlett's test can be used to test if $\Sigma_1 = \Sigma_2$ but this test is sensitive to departures from normality
- As a crude rule of thumb: if $s_{1,k}^2 > 4s_{2,k}^2$ or $s_{2,k}^2 > 4s_{1,k}^2$ for some $k \in \{1, 2, \dots, p\}$, then it is likely that $\Sigma_1 \neq \Sigma_2$
- Life gets difficult if we cannot assume that $\Sigma_1 = \Sigma_2$. However, if both n_1 and n_2 are "large", we can use the following approximation to conduct inferences:

$$T^2 = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T \left[\frac{1}{n_1} \mathbf{S}_1 + \frac{1}{n_2} \mathbf{S}_2 \right]^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \stackrel{H_0}{\sim} \chi_p^2$$

Comparing More Than Two Populations: Romano-British Pottery Example (source: PSU stat 505)

- Pottery shards are collected from four sites in the British Isles:
 - Llanedyrn (L)
 - Caldicot (C)
 - Isle Thorns (I)
 - Ashley Rails (A)
- The concentrations of five different chemicals were be used
 - Aluminum (Al)
 - Iron (Fe)
 - Magnesium (Mg)
 - Calcium (Ca)
 - Sodium (Na)
- **Objective:** to determine whether the chemical content of the pottery depends on the site where the pottery was obtained

Review: Analysis of Variance (ANOVA)

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$
 $H_a : \text{At least one mean is different}$

Source	df	SS	MS	F statistic
Treatment	$g - 1$	$SSTr$	$MSTr = \frac{SSTr}{g-1}$	$F = \frac{MSTr}{MSE}$
Error	$N - g$	SSE	$MSE = \frac{SSE}{N-g}$	
Total	$N - 1$	$SSTo$		

- Test Statistic: $F^* = \frac{MSTr}{MSE}$. Under H_0 , $F^* \sim F_{df_1=g-1, df_2=N-g}$

- **Assumptions:**

- The distribution of each group is normal with equal variance (i.e. $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_g^2$)
- Responses for a given group are independent to each other

One-way Multivariate Analysis of Variance (One-way MANOVA)

		Treatment			
		1	2	...	g
Subject	1	$\mathbf{Y}_{11} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ \vdots \\ Y_{11p} \end{pmatrix}$	$\mathbf{Y}_{21} = \begin{pmatrix} Y_{211} \\ Y_{212} \\ \vdots \\ Y_{21p} \end{pmatrix}$...	$\mathbf{Y}_{g1} = \begin{pmatrix} Y_{g11} \\ Y_{g12} \\ \vdots \\ Y_{g1p} \end{pmatrix}$
	2	$\mathbf{Y}_{21} = \begin{pmatrix} Y_{121} \\ Y_{122} \\ \vdots \\ Y_{12p} \end{pmatrix}$	$\mathbf{Y}_{22} = \begin{pmatrix} Y_{221} \\ Y_{222} \\ \vdots \\ Y_{22p} \end{pmatrix}$...	$\mathbf{Y}_{g2} = \begin{pmatrix} Y_{g21} \\ Y_{g22} \\ \vdots \\ Y_{g2p} \end{pmatrix}$
	\vdots	\vdots	\vdots		\vdots
	n_i	$\mathbf{Y}_{1n_1} = \begin{pmatrix} Y_{1n_11} \\ Y_{1n_12} \\ \vdots \\ Y_{1n_1p} \end{pmatrix}$	$\mathbf{Y}_{2n_2} = \begin{pmatrix} Y_{2n_21} \\ Y_{2n_22} \\ \vdots \\ Y_{2n_2p} \end{pmatrix}$...	$\mathbf{Y}_{gn_g} = \begin{pmatrix} Y_{gn_g1} \\ Y_{gn_g2} \\ \vdots \\ Y_{gn_gp} \end{pmatrix}$

Source: PSU stat 505

- **Notation:** \mathbf{Y}_{ij} is the vector of variables for subject j in group i ; n_i is the sample size in group i ;
 $N = n_1 + n_2 + \dots + n_g$ the total sample size
- **Assumptions:** 1) common covariance matrix Σ ; 2) Independence; 3) Normality

Test Statistics for MANOVA

- We are interested in testing the null hypothesis that the group mean vectors are all equal

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

The alternative hypothesis:

$H_a : \mu_{ik} \neq \mu_{jk}$ for at least one $i \neq j$ and at least one variable k

- **Mean vectors:**

- **Sample Mean Vector:** $\bar{\mathbf{y}}_{i.} = \frac{1}{n_i} \mathbf{Y}_{ij}, \quad i = 1, \dots, g$
- **Grand Mean Vector:** $\bar{\mathbf{y}}_{..} = \frac{1}{N} \sum_{i=1}^g \sum_{j=1}^{n_i} \mathbf{Y}_{ij}$

- **Total Sum of Squares:**

$$\mathbf{T} = \sum_{i=1}^g \sum_{j=1}^{n_i} (\mathbf{Y}_{ij} - \bar{\mathbf{y}}_{..})(\mathbf{Y}_{ij} - \bar{\mathbf{y}}_{..})^T$$

MANOVA Decomposition and MANOVA Table

$$\begin{aligned} T &= \sum_{i=1}^g \sum_{j=1}^{n_i} (\mathbf{Y}_{ij} - \mathbf{y}_{..})(\mathbf{Y}_{ij} - \bar{\mathbf{y}})^T \\ &= \sum_{i=1}^g \sum_{j=1}^{n_i} [(\mathbf{Y}_{ij} - \bar{\mathbf{y}}_{i.}) + (\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})][(\mathbf{Y}_{ij} - \bar{\mathbf{y}}_{i.}) + (\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})]^T \\ &= \underbrace{\sum_{i=1}^g \sum_{j=1}^{n_i} (\mathbf{Y}_{ij} - \bar{\mathbf{y}}_{i.})(\mathbf{Y}_{ij} - \bar{\mathbf{y}}_{i.})^T}_E + \underbrace{\sum_{i=1}^g n_i (\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})^T}_H \end{aligned}$$

MANOVA Table

Source	df	SS
Treatment	$g - 1$	H
Error	$N - g$	E
Total	$N - 1$	T

Reject $H_0 : \mu_1 = \mu_2 = \cdots = \mu_g$ if the matrix H is “large” relative to the matrix E

Test Statistics for MANOVA

There are several different test statistics for conducting the hypothesis test:

- Wilks Lambda

$$\Lambda^* = \frac{|E|}{|H + E|}$$

Reject H_0 if Λ^* is “small”

- Hotelling-Lawley Trace

$$T_0^2 = \text{trace}(HE^{-1})$$

Reject H_0 if T_0^2 is “large”

- Pillai Trace

$$V = \text{trace}(H(H + E)^{-1})$$

Reject H_0 if V is “large”

Romano-British Pottery Example

```
> dat <- read.table("pottery.txt", header = F)
> out <- manova(cbind(V2, V3, V4, V5, V6) ~ V1, data = dat)
> summary(out, test = "Wilks")
              Df      Wilks approx F num Df den Df    Pr(>F)
V1              3 0.012301   13.088     15 50.091 1.84e-12 ***
Residuals 22
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(out)
              Df Pillai approx F num Df den Df    Pr(>F)
V1              3 1.5539   4.2984     15     60 2.413e-05 ***
Residuals 22
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```