Comparisons of Two Mean Vectors

Multivariate Analysis of Variance

Lecture 6

Comparisons of Several Mean Vectors

Readings: Johnson & Wichern 2007, Chapter 6.3-6.5

DSA 8070 Multivariate Analysis

Whitney Huang
Clemson University

Agenda

1. Comparisons of Two Mean Vectors
2. Multivariate Analysis of Variance

Comparisons of Two Mean Vectors
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: $X_{11}, X_{12}, \cdots, X_{1n_1}$ and population 2: $X_{21}, X_{22}, \cdots, X_{2n_2}$. 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 $\Rightarrow \{X_{1i}\}_{i=1}^{n_1}$ 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

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

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 $\{\boldsymbol{X}_{11}, \cdots \boldsymbol{X}_{12}, \cdots \boldsymbol{X}_{1n_1}\}$ and $\{\boldsymbol{X}_{21}, \cdots \boldsymbol{X}_{22}, \cdots \boldsymbol{X}_{2n_2}\}$, where

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

to infer the relationship between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, where

$$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{ip} \end{bmatrix}$$

Assumptions

- Both populations have common covariance matrix, i.e., $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$
- Independence: The subjects from both populations are independently sampled
- Normality: Both populations are normally distributed

The Multivariate Two-Sample Problem

Comparisons of Two Mean Vectors
Multivariate Analysis of Variance

Here we are testing

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

Under the common covariance assumption we have

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

where

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

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

The two-sample t test is equivalent to

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

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:

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

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?

Simultaneous Confidence Intervals

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

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	$(1.81, 2.33)$

Checking Model Assumptions

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 $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ but this test is sensitive to departures from normality
- As as crude rule of thumb: if $s^2_{1,k} > 4s^2_{2,k}$ or $s^2_{2,k} > 4s^2_{1,k}$ for some $k \in \{1, 2, \cdots, p\}$, then it is likely that $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$
- Life gets difficult if we cannot assume that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ However, if both n_1 and n_2 are "large", we can use the following approximation to conduct inferences:

$$T^2 = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^T \left[\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2\right]^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) \overset{H_0}{\sim} \chi^2_p$$

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: (Univariate) Analysis of Variance (ANOVA)

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

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

- Test Statistic: $F^* = \frac{\text{MSTr}}{\text{MSE}}$. 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)

Subject \ Group	1	2	⋯	g
1	$\boldsymbol{Y}_{11} = \begin{bmatrix} Y_{111} \\ Y_{112} \\ \vdots \\ Y_{11p} \end{bmatrix}$	$\boldsymbol{Y}_{21} = \begin{bmatrix} Y_{211} \\ Y_{212} \\ \vdots \\ Y_{21p} \end{bmatrix}$	⋯	$\boldsymbol{Y}_{g1} = \begin{bmatrix} Y_{g11} \\ Y_{g12} \\ \vdots \\ Y_{g1p} \end{bmatrix}$
2	$\boldsymbol{Y}_{21} = \begin{bmatrix} Y_{121} \\ Y_{122} \\ \vdots \\ Y_{12p} \end{bmatrix}$	$\boldsymbol{Y}_{22} = \begin{bmatrix} Y_{221} \\ Y_{222} \\ \vdots \\ Y_{22p} \end{bmatrix}$	⋯	$\boldsymbol{Y}_{g2} = \begin{bmatrix} Y_{g21} \\ Y_{g22} \\ \vdots \\ Y_{g2p} \end{bmatrix}$
⋮	⋮	⋮	⋯	⋮
n_i	$\boldsymbol{Y}_{1n_i} = \begin{bmatrix} Y_{1n_i1} \\ Y_{1n_i2} \\ \vdots \\ Y_{1n_ip} \end{bmatrix}$	$\boldsymbol{Y}_{2n_i} = \begin{bmatrix} Y_{2n_i1} \\ Y_{2n_i2} \\ \vdots \\ Y_{2n_ip} \end{bmatrix}$	⋯	$\boldsymbol{Y}_{gn_i} = \begin{bmatrix} Y_{gn_i1} \\ Y_{gn_i2} \\ \vdots \\ Y_{gn_ip} \end{bmatrix}$

- **Notation**: $\boldsymbol{Y}_{ij}$ is the vector of variables for subject j in group i; n_i is the sample size in group i; $N = n_1 + n_2 + \cdots + n_g$ the total sample size

- **Assumptions**: 1) common covariance matrix $\boldsymbol{\Sigma}$; 2) Independence; 3) Normality

Test Statistics for MANOVA

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

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g.$$

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: $\bar{\boldsymbol{y}}_{i.} = \frac{1}{n_i}\boldsymbol{Y}_{ij}, \quad i = 1, \cdots, g$
  - Grand Mean Vector: $\bar{\boldsymbol{y}}_{..} = \frac{1}{N}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\boldsymbol{Y}_{ij}$
- **Total Sum of Squares**:

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

MANOVA Decomposition and MANOVA Table

$$
\begin{aligned}
\boldsymbol{T} &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{Y}_{ij} - \boldsymbol{y}_{..})(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{y}})^T \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left[(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{y}}_{i.}) + (\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})\right]\left[(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{y}}_{i.}) + (\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})\right]^T \\
&= \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{y}}_{i.})(\boldsymbol{Y}_{ij} - \bar{\boldsymbol{y}}_{i.})^T}_{\boldsymbol{E}} + \underbrace{\sum_{i=1}^{g} n_i(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})(\bar{\boldsymbol{y}}_{i.} - \bar{\boldsymbol{y}}_{..})^T}_{\boldsymbol{H}}
\end{aligned}
$$

MANOVA Table

Source	df	SS
Treatment	$g-1$	$\boldsymbol{H}$
Error	$N-g$	$\boldsymbol{E}$
Total	$N-1$	$\boldsymbol{T}$

Reject $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g$ if the matrix $\boldsymbol{H}$ is “large” relative to the matrix $\boldsymbol{E}$

Test Statistics for MANOVA

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

- Wilks Lambda

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

  Reject H_0 if Λ^* is "small"

- Hotelling-Lawley Trace

$$T_0^2 = \mathrm{trace}(\boldsymbol{H}\boldsymbol{E}^{-1})$$

  Reject H_0 if T_0^2 is "large"

- Pillai Trace

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

  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
```

$\Rightarrow$ at least one of the chemicals differs among the sites

Summary

In this lecture, we learned about:

- Hypothesis Testing for Two Mean Vectors
- MANOVA

In the next lecture, we will learn about Multivariate Linear Regression