# STATS 206
# Applied Multivariate Analysis
## Lecture 4:
## Comparisons of Several Multivariate Means

Prathapasinghe Dharmawansa

Department of Statistics
Stanford University
Autumn 2013

# Agenda

- Paired comparisons and repeated measures
- Comparing mean vectors from two populations
- Comparing several multivariate population means
  - One-way multivariate analysis of variance (MANOVA)
- Testing for equality of covariance matrices
- Two-way multivariate analysis of variance

# Paired Comparisons

- Treatments applied to the same/identical units
- Aiming to eliminate the influence of extraneous unit-to-unit variations
- The univariate case:
  - $X_{jk}$: measurements on unit $j, j = 1, \ldots, n$, for treatment $k, k = 1, 2$
  - Analysis based on $D_j = X_{j1} - X_{j2}$ using summary statistics:

$$\overline{D} = \frac{1}{n}\sum_{j=1}^{n} D_j, \quad s_d^2 = \frac{1}{n-1}\sum_{j=1}^{n}(D_j - \overline{D})^2$$

  - Given $D_j$: i.i.d. normal, the hypothesis $H_0 : \mathsf{E}(D) = \delta$ is tested using

$$t = \frac{\overline{D} - \delta}{s_d/\sqrt{n}} \sim t\text{-dist. with } (n-1) \text{ deg. of freedom}$$

A $100(1-\alpha)\%$ confidence interval is given by

$$\overline{D} - t_{n-1}(\alpha/2)\frac{s_d}{\sqrt{n}} \le \delta \le \overline{D} + t_{n-1}(\alpha/2)\frac{s_d}{\sqrt{n}}$$

# Paired Comparisons: The Multivariate Case (i)

- $p$ variables, $2$ treatments, $n$ experimental units
  $X_{1jk}(X_{2jk})$: variable $k$ under treatment $1$ $(2)$ within the $j$-th unit
  $k = 1, \ldots, p$; $j = 1, \ldots, n$
- Then

$$D_{jk} = X_{1jk} - X_{2jk}, k = 1, \ldots, p;$$
$$\mathbf{D}_j = [D_{j1},\ D_{j2},\ \ldots,\ D_{jp}]'$$

- Assume that

$$\mathsf{E}(\mathbf{D}_j) = \boldsymbol{\delta} = [\delta_1,\ \delta_2,\ \ldots,\ \delta_p]'$$
$$\mathsf{Cov}(\mathbf{D}_i) = \boldsymbol{\Sigma}_d,\ \ \forall j = 1, \ldots, n$$

## Paired Comparisons: The Multivariate Case (ii)

(cont'd)

- If $\mathbf{D}_1, \ldots, \mathbf{D}_n$: i.i.d. $N_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_d)$, then: derive inferences on $\boldsymbol{\delta}$ using

$$\text{(Hotelling's) } T^2 = \; n(\overline{\mathbf{D}} - \boldsymbol{\delta})'\mathbf{S}_d^{-1}(\overline{\mathbf{D}} - \boldsymbol{\delta})$$

where

$$\overline{\mathbf{D}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{D}_j, \; \mathbf{S}_d = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{D}_j - \overline{\mathbf{D}})(\mathbf{D}_j - \overline{\mathbf{D}})'$$

For all values of $\boldsymbol{\delta}, \boldsymbol{\Sigma}_d$:

$$T^2 \sim \frac{p(n-1)}{n-p} F_{p,n-p}$$

Remark: $n, (n-p)$ large $\Longrightarrow T^2 \approx \chi_p^2$!

# Paired Comparisons: The Multivariate Case (iii)

## Inferences about $\delta$

(cont'd) $\mathbf{D}_1, \ldots, \mathbf{D}_n$: i.i.d. $N_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_d)$; Observing $\mathbf{d}_1, \ldots, \mathbf{d}_n$

- Testing $H_0 : \boldsymbol{\delta} = \mathbf{0}$ vs. $H_1 : \boldsymbol{\delta} \neq \mathbf{0}$

  The $\alpha$-level test: Rejecting $H_0$ in favor of $H_1$ if

$$T^2 = n\overline{\mathbf{d}}'\mathbf{S}_d^{-1}\overline{\mathbf{d}} > \frac{p(n-1)}{n-p}F_{p,n-p}(\alpha)$$

- A $100(1-\alpha)\%$ confidence region for $\boldsymbol{\delta}$ is

$$(\overline{\mathbf{d}} - \boldsymbol{\delta})'\mathbf{S}_d^{-1}(\overline{\mathbf{d}} - \boldsymbol{\delta}) \leq \frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)$$

# Paired Comparisons: The Multivariate Case (iv)

## Inferences about $\delta$

(cont'd) $\mathbf{D}_1, \ldots, \mathbf{D}_n$: i.i.d. $N_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_d)$; Observing $\mathbf{d}_1, \ldots, \mathbf{d}_n$

- A $100(1-\alpha)\%$ simultaneous confidence intervals for individual mean differences ($\delta_i$'s) are

$$\delta_i : \quad \overline{d}_i \pm \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{d_i}^2}{n}}$$

$\overline{d}_i$: the $i$-th element of $\overline{\mathbf{d}}$; $s_{d_i}^2$: the $(i,i)$-th element of $\mathbf{S}_d$

Remark: for large $n$, $(n-p)$: $\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha) \approx \chi_p^2(\alpha)$

$\Longrightarrow$ No normality needed!!!

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

$$\delta_i : \quad \overline{d}_i \pm t_{n-1}\left(\frac{\alpha}{2p}\right) \sqrt{\frac{s_{d_i}^2}{n}}$$

## Example 1: i) Wastewater Treatment Plant Effluent

### Checking for a Mean Difference with Paired Observations

- Wastewater samples for inspection in 2 labs
  - a commercial lab and a state lab
- $n = 11$ samples of wastewater
- $p = 2$ variables

(1) Biochemical oxygen demand (BOD)
(2) Suspended solids (SS)

# Example 1: ii) The Data

## Checking for a Mean Difference with Paired Observations

| Sample $j$ | Commercial Lab | | State Lab of Hygiene | |
|---|---|---|---|---|
| | $x_{1j1}$ (BOD) | $x_{1j2}$ (SS) | $x_{2j1}$ (BOD) | $x_{2j2}$ (SS) |
| 1 | 6 | 27 | 25 | 15 |
| 2 | 6 | 23 | 28 | 13 |
| 3 | 18 | 64 | 36 | 22 |
| 4 | 8 | 44 | 35 | 29 |
| 5 | 11 | 30 | 15 | 31 |
| 6 | 34 | 75 | 44 | 64 |
| 7 | 28 | 26 | 42 | 30 |
| 8 | 71 | 124 | 54 | 64 |
| 9 | 43 | 54 | 34 | 56 |
| 10 | 33 | 30 | 29 | 20 |
| 11 | 20 | 14 | 39 | 21 |

Table 6.1 in the textbook

# Example 1: iii) Analyzing the Data

## Checking for a Mean Difference with Paired Observations

Assuming differences from paired observations are normal !!!

- Differences from paired observations: $d_{jk} = x_{1jk} - x_{2jk}, k = 1, 2$

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $d_{j1}$ | -19 | -22 | -18 | -27 | -4 | -10 | -14 | 17 | 9 | 4 | -19 |
| $d_{j2}$ | 12 | 10 | 42 | 15 | -1 | 11 | -4 | 60 | -2 | 10 | -7 |

- $T^2$ statistic for testing $H_0 : \boldsymbol{\delta} = [\delta_1, \delta_2]' = [0, 0]'$ vs. $H_1 : \boldsymbol{\delta} \neq [0, 0]'$

$$\overline{\mathbf{d}} = \begin{pmatrix} -9.3636 \\ 13.2727 \end{pmatrix}, \ \mathbf{S}_d = \begin{pmatrix} 199.2545 & 88.3091 \\ 88.3091 & 418.6182 \end{pmatrix}$$

$$T^2 = n\overline{\mathbf{d}}'\mathbf{S}_d^{-1}\overline{\mathbf{d}}\Big|_{n=11} = 13.6393$$

$\alpha$-level test: $\alpha = 0.05 \Rightarrow T^2 > \frac{p(n-1)}{(n-p)} F_{p,n-p}(0.05)\Big|_{\substack{n=11 \\ p=2}} = 9.4589$

$\Longrightarrow$ Reject $H_0$

# Example 1: iv) Analyzing the Data (Cont'd)
## Checking for a Mean Difference with Paired Observations

Assuming differences from paired observations are normal !!!

- The $95\%$ simultaneous confidence intervals for $\delta_1$ and $\delta_2$ are:

$$\delta_1 : \ \bar{d}_1 \pm \sqrt{\frac{p(n-1)}{n-p} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{d_1}^2}{n}} = -9.3636 \pm \sqrt{9.4589} \sqrt{\frac{199.2545}{11}}$$

or simply : $(-22.4533, 3.7260)$

$$\delta_2 : \ \bar{d}_2 \pm \sqrt{\frac{p(n-1)}{n-p} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{d_2}^2}{n}} = 13.2727 \pm \sqrt{9.4589} \sqrt{\frac{418.6182}{11}}$$

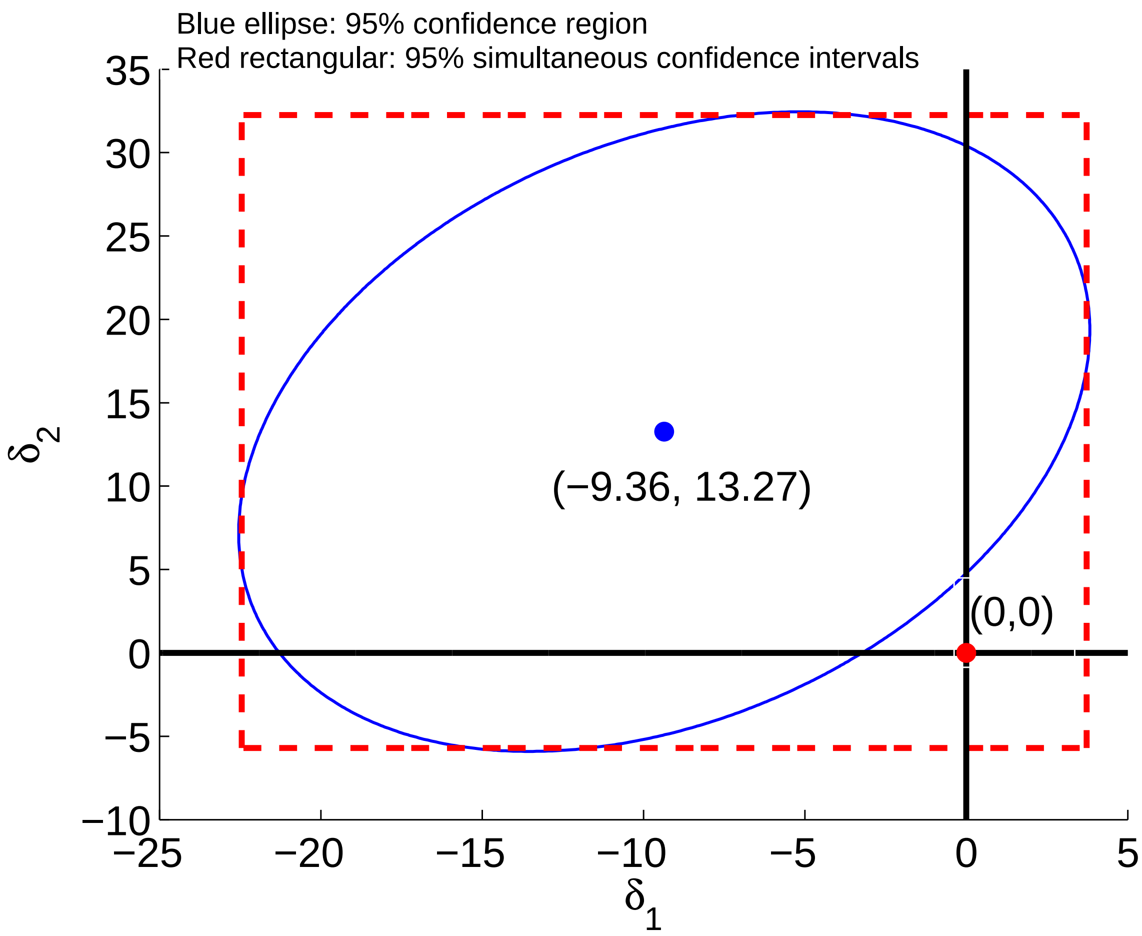
or simply : $(-5.7001, 32.2456)$

$\boldsymbol{\delta} = [\delta_1, \delta_2]' = [0, 0]'$ is contained in the above intervals!

- What to conclude based on $T^2$ and the above?
Check the fundamental definitions again! (Illustration next)

# Example 1: v) Analyzing the Data (Cont'd)

## Checking for a Mean Difference with Paired Observations



# Paired Comparisons: The Multivariate Case (v)

## Contrast Vectors and Contrast Matrices

- A $p$-dim. vector is called a contrast vector if its elements sum to $0$.

  $\Longrightarrow$ Orthogonal to the $p$-dim. all-one vector

  e.g. $(p = 4)$: $\mathbf{c} = [1, 0, -1, 0]'$

  Let $\mathbf{1} = [1, 1, 1, 1]' \Longrightarrow \mathbf{c}'\mathbf{1} = 0$

- A $k \times l$ matrix is called a contrast matrix if

  i) each of its rows is a contrast vector, and

  ii) the rows are linearly independent $(k \leq l)$.

  e.g.,

$$\mathbf{C}_{p\times 2p} = [\mathbf{I}_{p\times p} \quad -\mathbf{I}_{p\times p}] = \left[\begin{array}{cccc|cccc} 1 & 0 & \dots & 0 & -1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & -1 \end{array}\right]$$

  Used in next page

# Paired Comparisons: The Multivariate Case (vi)

## Representation Using a Contrast Matrix

- Concatenate the results from two treatments:

$$\overline{\mathbf{x}}_{2p\times 1} = [\overline{x}_{11}, \ldots, \overline{x}_{1p}, \overline{x}_{21}, \ldots, \overline{x}_{2p}]', \quad \mathbf{S}_{2p\times 2p} = \left[\begin{array}{c:c} \underset{(p\times p)}{\mathbf{S}_{11}} & \underset{(p\times p)}{\mathbf{S}_{12}} \\ \hdashline \underset{(p\times p)}{\mathbf{S}_{21}} & \underset{(p\times p)}{\mathbf{S}_{22}} \end{array}\right]$$

- Use the contrast matrix $\mathbf{C}_{p\times 2p}$ introduced in the previous page:

$$\Longrightarrow \mathbf{d}_j = \mathbf{C}\mathbf{x}_j, j = 1, \ldots, n$$

$$\overline{\mathbf{d}} = \mathbf{C}\overline{\mathbf{x}}$$

$$\mathbf{S}_d = \mathbf{C}\mathbf{S}\mathbf{C}'$$

$$T^2 = n\overline{\mathbf{x}}\mathbf{C}'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}\mathbf{C}\overline{\mathbf{x}}$$

Directly working on $\mathbf{x}_j$'s (no need to calculate $\mathbf{d}_j$'s)

# Repeated Measures (i)

- (Measurements of a single response variable) $q$ treatments, $n$ units

  the $j$-th observation : $\mathbf{X}_j = [X_{j1}, X_{j2}, \ldots, X_{jq}]'$, $j = 1, 2, \ldots, n$

  $X_{ji}$: response to the $i$-th treatment on the $j$-th unit
- Let $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X}_j), \forall j$
- For comparative purposes, consider two contrasts of components of $\boldsymbol{\mu}$:

$$(1).\quad \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_q \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & \ldots & 0 \\ 1 & 0 & -1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} = \mathbf{C}_1 \boldsymbol{\mu}$$

$$(2).\quad \begin{bmatrix} \mu_2 - \mu_1 \\ \mu_3 - \mu_2 \\ \vdots \\ \mu_q - \mu_{q-1} \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} = \mathbf{C}_2 \boldsymbol{\mu}$$

# Repeated Measures (ii)

- Here $\mathbf{C}_1$ and $\mathbf{C}_2$: $(q-1) \times q$ contrast matrices

$$\mathbf{C}_1 = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & -1 \end{bmatrix}, \mathbf{C}_2 = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}$$

- Equal treatment means $\Longleftrightarrow \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$; $\mathbf{C}$: any $(q-1) \times q$ contrast matrix
- For a normal population: to test $\mathbf{C}\boldsymbol{\mu} = \mathbf{0}$
  - Consider contrasts $\mathbf{C}\mathbf{x}_j$ in the observations
  - Use summary statistics $\mathbf{C}\overline{\mathbf{x}}, \mathbf{CSC}'$
  - The $T^2$-statistic: $T^2 = n(\mathbf{C}\overline{\mathbf{x}})'(\mathbf{CSC})^{-1}(\mathbf{C}\overline{\mathbf{x}})$
    Remark:

$T^2$ does not depend on the particular choice of the $(q-1) \times q$ contrast matrix $\mathbf{C}$.

# Repeated Measures (iii)

## Test for Equality of Treatments in a Repeated Measures Design

Consider a normal ($q$-dim., with mean $\boldsymbol{\mu}$) population; $\mathbf{C}$: contrast matrix

- Testing $H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}$ (equal treatment means) vs. $H_1 : \mathbf{C}\boldsymbol{\mu} \neq \mathbf{0}$
  - An $\alpha$-level test rejects $H_0$ if

$$T^2 = n(\mathbf{C}\overline{\mathbf{x}})'(\mathbf{CSC})^{-1}(\mathbf{C}\overline{\mathbf{x}}) > \frac{(q-1)(n-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha)$$

- A confidence region for the contrast $\mathbf{C}\boldsymbol{\mu}$:

$$n(\mathbf{C}\overline{\mathbf{x}} - \mathbf{C}\boldsymbol{\mu})'(\mathbf{CSC})^{-1}(\mathbf{C}\overline{\mathbf{x}} - \mathbf{C}\boldsymbol{\mu}) \leq \frac{(q-1)(n-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha)$$

- Simultaneous $100(1-\alpha)\%$ confidence intervals for single contrasts $\mathbf{c}'\boldsymbol{\mu}$ (for any contrast vector $\mathbf{c}$ of interests):

$$\mathbf{c}'\boldsymbol{\mu} : \quad \mathbf{c}'\overline{\mathbf{x}} \pm \sqrt{\frac{(q-1)(n-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha)} \sqrt{\frac{\mathbf{c}'\mathbf{S}\mathbf{c}}{n}}$$

# Example 2: Anesthetic test with dogs (i)

- $n = 19$ dogs were studied
- $q = 4$ treatments

  Treatment 1 = high $CO_2$ pressure without Halothane (H)
  Treatment 2 = low $CO_2$ pressure without H
  Treatment 3 = high $CO_2$ pressure with H
  Treatment 4 = low $CO_2$ pressure with H

- The same single response: milliseconds between heartbeats

  This is a **repeated measures** design

# Example 2: Anesthetic test with dogs – The Data (ii)

| Dog | Treatment | | | | Dog | Treatment | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | | 1 | 2 | 3 | 4 |
| 1 | 426 | 609 | 556 | 600 | 11 | 349 | 382 | 473 | 497 |
| 2 | 253 | 236 | 392 | 395 | 12 | 429 | 410 | 488 | 547 |
| 3 | 359 | 433 | 349 | 357 | 13 | 348 | 377 | 447 | 514 |
| 4 | 432 | 431 | 522 | 600 | 14 | 412 | 473 | 472 | 446 |
| 5 | 405 | 426 | 513 | 513 | 15 | 347 | 326 | 455 | 468 |
| 6 | 324 | 438 | 507 | 539 | 16 | 434 | 458 | 637 | 524 |
| 7 | 310 | 312 | 410 | 456 | 17 | 364 | 367 | 432 | 469 |
| 8 | 326 | 326 | 350 | 504 | 18 | 420 | 395 | 508 | 531 |
| 9 | 375 | 447 | 547 | 548 | 19 | 397 | 556 | 645 | 625 |
| 10 | 286 | 286 | 403 | 422 | | | | | |

Table 6.2 in the textbook

## Example 2: Anesthetic test with dogs – The Analysis (iii)

- Let $\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3, \mu_4]'$ be the mean responses corresponding to treatments 1, 2, 3 and 4, respectively.
- Assuming normal population and testing different hypotheses:
  - For example, the hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4$ can be written as:

$$\underbrace{\mathbf{C}\boldsymbol{\mu} = \mathbf{0}}_{\text{No treatment differences}}, \quad \text{where } \mathbf{C}_{3\times 4} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix} \quad (q = 4)$$

  - Another example: testing specific effects of $CO_2$ pressure and H

$$(\mu_3 + \mu_4) - (\mu_1 + \mu_2) = (\text{H influence})$$

$$(\mu_1 + \mu_3) - (\mu_2 + \mu_4) = (\text{CO}_2 \text{ pressure influence})$$

$$(\mu_1 + \mu_4) - (\mu_2 + \mu_3) = (\text{H-CO}_2 \text{ pressure ``interaction''})$$

(to be continued on next page)

## Example 2: Anesthetic test with dogs – The Analysis (iv)

### Testing specific effects of $CO_2$ pressure and H

- Contrast matrix $\mathbf{C}_{3\times 4}$

$$\mathbf{C} = \begin{bmatrix} -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1' \\ \mathbf{c}_2' \\ \mathbf{c}_3' \end{bmatrix}, \quad \underbrace{H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{0}}_{\text{No specific effects}}$$

From the data:

$$\mathbf{C}\overline{\mathbf{x}} = [209.32,\ -60.05,\ -12.79]',\ \mathbf{CSC}' = \begin{bmatrix} 9432.2 & 1098.9 & 927.6 \\ 1098.9 & 5195.8 & 914.6 \\ 927.6 & 914.6 & 7557.4 \end{bmatrix}$$

$$T^2 = n(\mathbf{C}\overline{\mathbf{x}})'(\mathbf{CSC})^{-1}(\mathbf{C}\overline{\mathbf{x}}) = 19 \times 6.1061 = 116.0163$$

$$> \frac{(q-1)(n-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha)\Big|_{\alpha=0.05,n=19,q=4} = 10.9312$$

$\Longrightarrow$ Rejecting $H_0$

# Example 2: Anesthetic test with dogs – The Analysis (v)

## Testing specific effects of $CO_2$ pressure and H

- Which ones are responsible for rejecting $H_0$?

  Construct $95\%$ simultaneous confidence intervals for each contrast:

$$\mathbf{c}_1'\boldsymbol{\mu} = (\mu_3 + \mu_4) - (\mu_1 + \mu_2) = \text{H influence}$$

$$\Longrightarrow \mathbf{c}_1'\boldsymbol{\mu} : \quad \underbrace{\mathbf{c}_1'\overline{\mathbf{x}}}_{\overline{x}_3+\overline{x}_4-(\overline{x}_1+\overline{x}_2)} \pm \sqrt{\frac{18 \times 3}{16} F_{3,16}(0.05)} \sqrt{\frac{\mathbf{c}_1'\mathbf{S}\mathbf{c}_1}{19}}$$

$$= 209.32 \pm \sqrt{10.93} \sqrt{\frac{9432.2}{19}} = 209.32 \pm 73.67$$

Zero not included $\Longrightarrow$ there is a halothane (H) influence!

# Example 2: Anesthetic test with dogs – The Analysis (vi)

## Testing specific effects of $CO_2$ pressure and H

Similarly:

$$\mathbf{c}_2'\boldsymbol{\mu} = (\mu_1 + \mu_3) - (\mu_2 + \mu_4) = \text{CO}_2 \text{ pressure influence}$$

$$\Longrightarrow \mathbf{c}_2'\boldsymbol{\mu} : -60.05 \pm \sqrt{10.93}\sqrt{\frac{5195.8}{19}} = -60.05 \pm 54.67$$

$$\mathbf{c}_3'\boldsymbol{\mu} = (\mu_1 + \mu_4) - (\mu_2 + \mu_3) = \text{H-CO}_2 \text{ pressure ``interaction''}$$

$$\Longrightarrow \mathbf{c}_3'\boldsymbol{\mu} : -12.79 \pm \sqrt{10.93}\sqrt{\frac{7557.4}{19}} = -12.79 \pm 65.94$$

$\Longrightarrow$ there is also a $CO_2$ pressure influence

# Comparing Mean Vectors from Two Populations

## Basic Setup and Problem of Interest

- Assumptions:
  1. $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1}$: random sample from $p$-variate Population 1
     Mean: $\boldsymbol{\mu}_1$, covariance matrix: $\boldsymbol{\Sigma}_1$
  2. $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2}$: random sample from $p$-variate Population 2
     Mean: $\boldsymbol{\mu}_2$, covariance matrix: $\boldsymbol{\Sigma}_2$
  3. The above two random samples are independent.
- The problem here:
  - Is $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ (or, is $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$)?
  - If $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \mathbf{0}$, which component means are different?
- Summary statistics to be used later

| sample | summary statistics |
|---|---|
| $\mathbf{x}_{11} \ldots \mathbf{x}_{1n_1}$ | $\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} \mathbf{x}_{1j}$, $\mathbf{S}_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)'$ |
| $\mathbf{x}_{21} \ldots \mathbf{x}_{2n_2}$ | $\bar{\mathbf{x}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \mathbf{x}_{2j}$, $\mathbf{S}_2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)'$ |

# Comparing Mean Vectors from Two Populations

## Small $n_1$ and $n_2$, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ (i)

Further assumptions: 1). normal populations; 2). $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ (Strong!)

- Since $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$

$$\sum_{j=1}^{n_1} (\mathbf{x}_{1j} - \overline{\mathbf{x}}_1)(\mathbf{x}_{1j} - \overline{\mathbf{x}}_1)' : \text{ estimate of } (n_1 - 1)\mathbf{\Sigma}$$

$$\sum_{j=1}^{n_2} (\mathbf{x}_{2j} - \overline{\mathbf{x}}_2)(\mathbf{x}_{2j} - \overline{\mathbf{x}}_2)' : \text{ estimate of } (n_2 - 1)\mathbf{\Sigma}$$

pool info. to get a pooled estimate of $\mathbf{\Sigma}$

$$\begin{aligned}\mathbf{S}_{\text{pooled}} =& \frac{\sum_{j=1}^{n_1} (\mathbf{x}_{1j} - \overline{\mathbf{x}}_1)(\mathbf{x}_{1j} - \overline{\mathbf{x}}_1)' + \sum_{j=1}^{n_2} (\mathbf{x}_{2j} - \overline{\mathbf{x}}_2)(\mathbf{x}_{2j} - \overline{\mathbf{x}}_2)'}{n_1 + n_2 - 2} \\ =& \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{S}_2\end{aligned}$$

# Comparing Mean Vectors from Two Populations

## Small $n_1$ and $n_2$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ (ii)

Further assumptions: 1). normal populations; 2). $\underline{\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2}$

- Problem: testing $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0$ vs. $H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\delta}_0$

  ($\boldsymbol{\delta}_0$: a specified vector)

  Consider the squared distance from $\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2$ to $\boldsymbol{\delta}_0$:

$$\mathsf{E}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$$

$$\mathsf{Cov}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) = \mathsf{Cov}(\overline{\mathbf{X}}_1) + \mathsf{Cov}(\overline{\mathbf{X}}_2) = (1/n_1 + 1/n_2)\,\boldsymbol{\Sigma}$$

$$(1/n_1 + 1/n_2)\,\mathbf{S}_{\text{pooled}} : \text{ estimate of } \mathsf{Cov}(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)$$

$\Longrightarrow$ Reject $H_0$ if

$$T^2 = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 - \boldsymbol{\delta}_0)' \left[ \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{\text{pooled}} \right]^{-1} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 - \boldsymbol{\delta}_0) > c^2$$

# Comparing Mean Vectors from Two Populations

## Small $n_1$ and $n_2$, $\Sigma_1 = \Sigma_2$ (iii)

- If two independent random samples: $\mathbf{X}_{11}, \mathbf{X}_{12}, \ldots, \mathbf{X}_{1n_1} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, and $\mathbf{X}_{21}, \mathbf{X}_{22}, \ldots, \mathbf{X}_{2n_2} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, then

$$\left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \mathbf{S}_{\text{pooled}}\right]^{-1} \left(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)$$

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

$$\Longrightarrow \; P[T^2 \le c^2] = 1 - \alpha, \qquad c^2 \triangleq \frac{(n_1 + n_2 - 2)p}{(n_1 + n_2 - p - 1)} F_{p, n_1 + n_2 - p - 1}(\alpha)$$

Proof:

– $\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 \sim N_p(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, [1/n_1 + 1/n_2]\boldsymbol{\Sigma})$;
– $(n_1 - 1)\mathbf{S}_1 \sim \mathbf{W}_{p, n_1 - 1}(\boldsymbol{\Sigma})$, $(n_2 - 1)\mathbf{S}_2 \sim \mathbf{W}_{p, n_2 - 1}(\boldsymbol{\Sigma})$
– $\mathbf{S}_1, \mathbf{S}_2$ independent $\Longrightarrow (n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 \sim \mathbf{W}_{p, n_1 + n_2 - 2}(\boldsymbol{\Sigma})$

# Comparing Mean Vectors from Two Populations

## Small $n_1$ and $n_2$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, Normal Populations (iv)

- Confidence regions for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$:

$$\{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 : T^2 \leq c^2\}$$

  An ellipsoid centered at $\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2$ with axes determined by eigenvectors and eigenvalues of $\mathbf{S}_{\text{pooled}}$ ($T^2, c^2$: see previous page)

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

$$\mathbf{a}'(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) \pm c\sqrt{\mathbf{a}'\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{\text{pooled}}\mathbf{a}}$$

  will cover $\mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for all $\mathbf{a}$. (Proof: similar to previous analysis)

# Comparing Mean Vectors from Two Populations

## Small $n_1$ and $n_2$, $\Sigma_1 = \Sigma_2$, Normal Populations (v)

(Cont'd from previous page)

- Simultaneous confidence intervals for $\mu_{1i} - \mu_{2i}$:

$$(\overline{X}_{1i} - \overline{X}_{i2}) \pm c\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_{ii,\text{pooled}}}, \quad i = 1, 2, \ldots, p$$

- Bonferroni $100(1-\alpha)\%$ simultaneous confidence intervals for $\mu_{1i} - \mu_{2i}$:

$$(\overline{X}_{1i} - \overline{X}_{i2}) \pm t_{n_1+n_2-2}\left(\frac{\alpha}{2p}\right)\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_{ii,\text{pooled}}}, \quad i = 1, 2, \ldots, p$$

# Example 3 (i)
## Confidence Region for Difference of Two Mean Vectors

- Two kinds of soaps manufactured, each 50 bars ($n_1 = n_2 = 50$)
  Two characteristics measured: $X_1$=lather, $X_2$=mildness ($p = 2$)
  Given summary statistics of two kinds of soaps:
  $$\overline{\mathbf{x}}_1 = \begin{bmatrix} 8.3 \\ 4.1 \end{bmatrix}, \mathbf{S}_1 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}, \overline{\mathbf{x}}_2 = \begin{bmatrix} 10.2 \\ 3.9 \end{bmatrix}, \mathbf{S}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix},$$
  Normal populations: obtain a 95% confidence region for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$

- $\mathbf{S}_1, \mathbf{S}_2$: approx. equal $\Longrightarrow \mathbf{S}_{\text{pooled}} = \frac{49}{98}\mathbf{S}_1 + \frac{49}{98}\mathbf{S}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}$
  $\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 = [-1.9, 0.2]'$
  The confidence ellipse ($p = 2$) is
  $$(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \mathbf{S}_{\text{pooled}}\right]^{-1} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) \leq c^2$$

# Example 3 (ii)
## Confidence Region for Difference of Two Mean Vectors

Alternatively, the ellipse is given by

$$(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))' \, \mathbf{S}_{\text{pooled}}^{-1} \, (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) \leq \left(\frac{1}{n_1} + \frac{1}{n_2}\right) c^2$$

Eigenvalue decomposition of $\mathbf{S}_{\text{pooled}}$

$$\left\{ \left( \lambda_1 = 5.303, \mathbf{e}_1 = \begin{bmatrix} 0.290 \\ 0.957 \end{bmatrix} \right), \left( \lambda_1 = 1.697, \mathbf{e}_1 = \begin{bmatrix} 0.957 \\ -0.290 \end{bmatrix} \right) \right\}$$

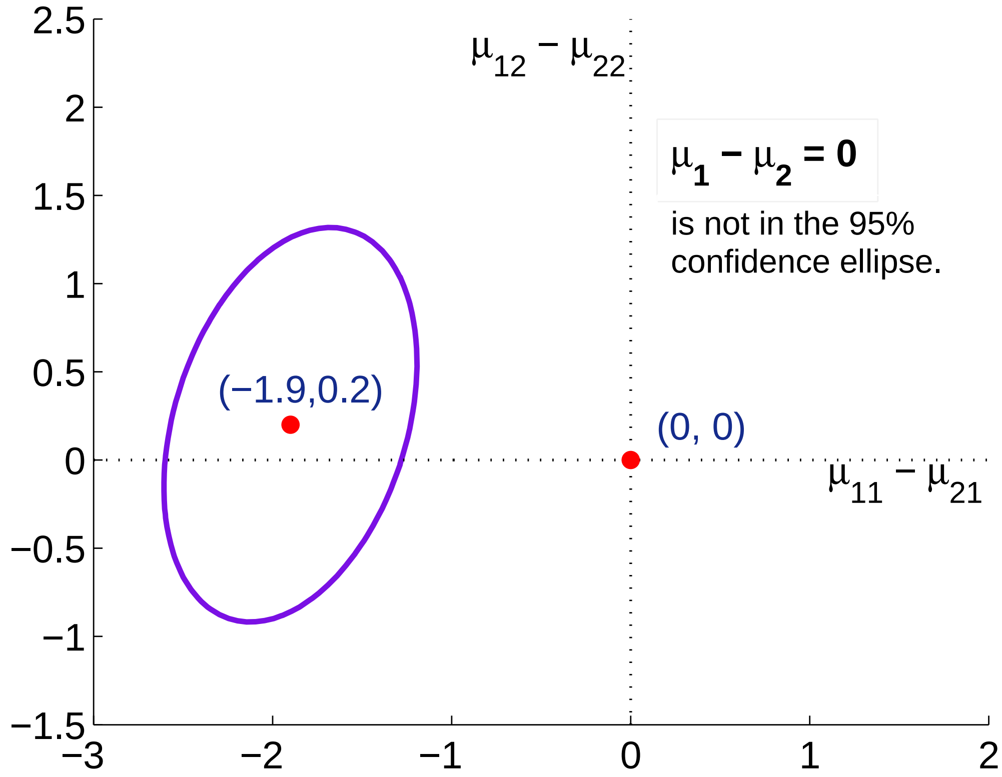
In addition:

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right) c^2 = \left(\frac{1}{50} + \frac{1}{50}\right) \frac{98 \times 2}{97} F_{2,97}(0.05) = 0.25$$

Confidence ellipse extends $\sqrt{\lambda_i} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) c^2} = \sqrt{\lambda_i} \sqrt{0.25}$ along $\mathbf{e}_i$

# Example 3 (iii)

## $95\%$ Confidence Ellipse for $\mu_1 - \mu_2$



# Comparing Mean Vectors from Two Populations

## $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, Large $n_1$ and $n_2$ (i)

- In general ($n_1 \neq n_2$):
  - No pooling for covariance matrix
  - Consider large $n_1$ and $n_2$ ($p$ fixed)

i) Replace $\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \mathbf{S}_{\text{pooled}}$ by $\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2$

$$T^2 = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))' \left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]^{-1} (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$$

ii) Replace $F$-distribution by $\chi^2_p$

# Comparing Mean Vectors from Two Populations

## $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, Large $n_1$ and $n_2$ (ii)

(Cont'd from previous page)

For example: an approximate $100\%(1-\alpha)$ confidence ellipsoid satisfies

$$[\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]' \times \left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right)^{-1} [\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \leq \chi_p^2(\alpha)$$

Another example: the $100(1-\alpha)\%$ simultaneous confidence intervals for all linear combinations $\mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ are provided by

$$\left\{\mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) : \;\; \mathbf{a}'(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\mathbf{a}'\left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right)\mathbf{a}}\right\}$$

# Comparing Mean Vectors from Two Populations

## $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$, Large $n_1$ and $n_2$, $n_1 = n_2$ (iii)

- When $n_1 = n_2 = n$

$$\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2 = \frac{1}{n}(\mathbf{S}_1 + \mathbf{S}_2) = \frac{(n-1)\mathbf{S}_1 + (n-1)\mathbf{S}_2}{n+n-2}\left(\frac{1}{n} + \frac{1}{n}\right)$$
$$= \mathbf{S}_{\text{pooled}}\left(\frac{1}{n} + \frac{1}{n}\right)$$

$n_1 = n_2 = n$ further with large sample size: here the procedure is essentially the same as the one with pooled covariance matrix.

# Example 4 (i): Electrical Usage Data

## Large Sample Procedures for Inferences about Mean Difference

- 2 types of homes, with/without air conditioning: ($n_1 = 45, n_2 = 55$)
- $p = 2$ measurements of electrical usage:
  - $X_1$=on-peak consumption
  - $X_2$=off-peak consumption
- Given the following summary statistics:

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 204.4 \\ 556.6 \end{bmatrix} \quad \mathbf{S}_1 = \begin{bmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{bmatrix}$$

$$\bar{\mathbf{x}}_2 = \begin{bmatrix} 130.0 \\ 355.0 \end{bmatrix} \quad \mathbf{S}_2 = \begin{bmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{bmatrix}$$

obtain 95% simultaneous confidence intervals for components of mean differences using the large sample procedure.

Example 6.5 in the textbook

## Example 4 (ii): Electrical Usage Data Analysis

### Large Sample Procedures for Inferences about Mean Difference

- First: $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 = [\underline{74.4}, \underline{201.6}]'$

$$\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2 = \frac{1}{45}\begin{bmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{bmatrix} + \frac{1}{55}\begin{bmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{bmatrix}$$

$$= \begin{bmatrix} \underline{464.17} & 886.08 \\ 886.08 & \underline{2642.15} \end{bmatrix}$$

- Take $\mathbf{a} = [1\ 0]'$, $\mathbf{a} = [0\ 1]'$ to obtain: $(\chi_2^2(0.05) = 5.99)$

$$\mu_{11} - \mu_{21} : 74.4 \pm \sqrt{5.99}\sqrt{464.17} \text{ or } (21.7, 127.1)$$

$$\mu_{12} - \mu_{22} : 201.6 \pm \sqrt{5.99}\sqrt{2642.15} \text{ or } (75.8, 327.4)$$

- In addition: testing $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0} \Longrightarrow$ Reject $H_0$, since

$$T^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = 15.66 > \chi_2^2(0.05) = 5.99$$

# Comparing Mean Vectors from Two Populations

## $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, Small/Medium $n_1$ and $n_2$ ($p$ Fixed)

Further assumption: normal populations

- Approach: Approximating the dist. of the statistic

$$T^2 = \left[\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]' \left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right)^{-1} \left[\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]$$

$$\overset{\text{approx.}}{\sim} \frac{vp}{v-p+1} F_{p,v-p+1}, \qquad \text{where } v \text{ is estimated as}$$

$$v = \frac{p+p^2}{\sum_{i=1}^{2} \frac{1}{n_i}\left\{\mathrm{tr}\left[\mathbf{A}_i^2\right] + \left(\mathrm{tr}\left[\mathbf{A}_i\right]\right)^2\right\}} \qquad (\min(n_1, n_2) \leq v \leq n_1 + n_2)$$

$$\mathbf{A}_i \triangleq \frac{1}{n_i}\mathbf{S}_i \left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right)^{-1} \qquad i = 1, 2$$

- The procedure:
  - Use the above $T^2$ and the corresponding $\frac{vp}{v-p+1}F_{p,v-p+1}(\alpha)$

## Example 5 (i): Electrical Usage Data (Again)

### Approximating $T^2$ When $\Sigma_1 \neq \Sigma_2$

- 2 types of homes, with/without air conditioning: ($n_1 = 45, n_2 = 55$)
- $p = 2$ measurements of electrical usage:
  - $X_1$=on-peak consumption
  - $X_2$=off-peak consumption
- (Normal Populations) Given the following summary statistics:

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 204.4 \\ 556.6 \end{bmatrix} \quad \mathbf{S}_1 = \begin{bmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{bmatrix}$$

$$\bar{\mathbf{x}}_2 = \begin{bmatrix} 130.0 \\ 355.0 \end{bmatrix} \quad \mathbf{S}_2 = \begin{bmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{bmatrix}$$

test $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ (significance level $\alpha = 0.05$)

Example 6.6 in the textbook

# Example 5 (ii): Electrical Usage Data Analysis

## Approximating $T^2$ When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

- First:

$$\mathbf{B} \triangleq \left[\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right]^{-1} = 10^{-4}\begin{bmatrix} 59.874 & -20.080 \\ -20.080 & 10.519 \end{bmatrix}$$

$$\mathbf{A}_1 = \frac{1}{n_1}\mathbf{S}_1\mathbf{B} = \begin{bmatrix} 0.776 & -0.060 \\ -0.092 & 0.646 \end{bmatrix}, \mathbf{A}_1^2 = \begin{bmatrix} 0.608 & -0.085 \\ -0.131 & 0.423 \end{bmatrix}$$

$$\mathbf{A}_2 = \frac{1}{n_2}\mathbf{S}_2\mathbf{B} = \begin{bmatrix} 0.224 & -0.060 \\ 0.092 & 0.354 \end{bmatrix}, \mathbf{A}_2^2 = \begin{bmatrix} 0.055 & 0.035 \\ 0.053 & 0.131 \end{bmatrix}$$

$$\frac{1}{n_i}\left\{\text{tr}\left[\mathbf{A}_i^2\right] + \left(\text{tr}\left[\mathbf{A}_i\right]\right)^2\right\} = \begin{cases} 0.0678 & i = 1 \\ 0.0095 & i = 2 \end{cases}$$

## Example 5 (iii): Electrical Usage Data Analysis

### Approximating $T^2$ When $\Sigma_1 \neq \Sigma_2$

- Estimated degrees of freedom

$$v = \frac{2 + 2^2}{0.0678 + 0.0095} = 77.6$$

Testing $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$ (significance level $\alpha = 0.05$)

$$\underbrace{T^2 = 15.66}_{\text{calculated before}} > \frac{vp}{v - p + 1} F_{p, v-p+1}(0.05) \Big|_{v=77.6, p=2} = 6.32$$

$\Longrightarrow$ Reject $H_0$

[same conclusion as in Example 4 (large sample procedure)]

# Comparing Several Multivariate Population Means

## One-way MANOVA Setup: Structure and Assumptions

- Setup: $g$ populations; $n_l$ observations for population $l, l = 1, \ldots, g$
- Assumptions:
  1. $\mathbf{X}_{l1}, \mathbf{X}_{l2}, \ldots, \mathbf{X}_{ln_l}$:
     - random sample of size $n_l$
     - from a population with mean $\boldsymbol{\mu}_l$ $(l = 1, 2, \ldots, g)$

     Random samples from different populations: independent
  2. All populations: common covariance matrix $\boldsymbol{\Sigma}$ (positive definite)
  3. Each population: multivariate normal ($p$-dimensional)
- Problem of interest:

  One-way multivariate analysis of variance (MANOVA)
  - Testing

    $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_g$ vs. $H_1 : \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j, i \neq j, 1 \leq i, j \leq g$
  - If rejecting $H_0$, find out which mean components differ significantly

# Univariate ANOVA $(p = 1)$ (i)

## Normal Samples with Common Variance $\sigma^2$

- Assumptions in univariate analysis of variance (ANOVA)
  - $X_{l1}, \ldots, X_{ln_l}$: random sample $\sim N(\mu_l, \sigma^2)$, $l = 1, \ldots, g$
  - Random samples from $g$ different populations: independent

- Null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$

  Alternatively, let

$$\underbrace{\mu_l}_{l\text{-th population mean}} = \underbrace{\mu}_{\text{overall mean}} + \underbrace{\tau_l}_{\substack{l\text{-th population} \\ \text{(treatment) effect}}}, \quad l = 1, 2, \ldots, g$$

  then

$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_g = 0$$

# Univariate ANOVA (ii)

## Normal Samples with Common Variance $\sigma^2$

- The response $X_{lj} \sim N(\mu + \tau_l, \sigma^2)$ $(j = 1, \ldots, n_l; l = 1, \ldots, g)$

$$X_{lj} = \underbrace{\mu}_{\text{overall mean}} + \underbrace{\tau_l}_{\substack{l\text{-th population}\\ \text{(treatment) effect}}} + \underbrace{e_{lj}}_{\substack{\text{random error}\\ \sim N(0,\sigma^2)}}$$

(Assuming $\sum_{l=1}^{g} n_l \tau_l = 0$ for unique identification of parameters)

- The observation $x_{lj}$

$$\underbrace{x_{lj}}_{\text{observation}} = \underbrace{\overline{x}}_{\substack{\text{overall}\\ \text{sample mean}}} + \underbrace{(\overline{x}_l - \overline{x})}_{\substack{l\text{-th estimated}\\ \text{(treatment) effect}}} + \underbrace{(x_{lj} - \overline{x}_l)}_{\text{residual}}$$

  - $\overline{x}$: estimate of $\mu$ $\quad \overline{x} = \frac{1}{n} \sum_{l=1}^{g} \sum_{j=1}^{n_l} x_{lj}, \quad n = \sum_{l=1}^{g} n_l$
  - $\widehat{\tau}_l = (\overline{x}_l - \overline{x})$: estimate of $\tau_l$ $\quad \overline{x}_l = \frac{1}{n_l} \sum_{j=1}^{n_l} x_{lj}, \quad \sum_{l=1}^{g} n_l \widehat{\tau}_l = 0$
  - $(x_{lj} - \overline{x}_l)$: estimate of $e_{lj}$

# Univariate ANOVA (iii)

- It can be shown that

$$\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\overline{x}_l - \overline{x})(x_{lj} - \overline{x}_l) = 0$$

$$\sum_{l=1}^{g}\sum_{j=1}^{n_l}\overline{x}(x_{lj} - \overline{x}_l) = 0, \quad \sum_{l=1}^{g}\sum_{j=1}^{n_l}\overline{x}(\overline{x}_l - \overline{x}) = 0$$

$\Longrightarrow$ mean $\overline{x}$, treatment effect $\overline{x}_l - \overline{x}$, and residual $x_{lj} - \overline{x}_l$: orthogonal

- In addition

$$\underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj} - \overline{x})^2}_{\substack{\text{Sum of Squares of Total} \\ \text{Variations (Corrected for the Mean)}}} = \underbrace{\sum_{l=1}^{g} n_l(\overline{x}_l - \overline{x})^2}_{\substack{\text{Sum of Squares} \\ \text{of Treatments}}} + \underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj} - \overline{x}_l)^2}_{\substack{\text{Sum of Squares} \\ \text{of Residual}}}$$

# Univariate ANOVA (iv)

- Let $\mathbf{y}$ collect all the observations: $\Longrightarrow \mathbf{y}$: in the $n = \sum_{l=1}^{g} n_l$ dimensions

$$\mathbf{y} = [x_{11}, \ldots, x_{1n_1}, x_{21}, \ldots, x_{2n_2}, \ldots, x_{g1}, \ldots, x_{gn_g}]'$$

- Mean vector: $\overline{x}\mathbf{1}_{n\times 1}$: a vector
- Treatment effect vector: $(\overline{x}_1 - \overline{x})\mathbf{u}_1 + (\overline{x}_2 - \overline{x})\mathbf{u}_2 + \ldots + (\overline{x}_g - \overline{x})\mathbf{u}_g$ where the $n \times 1$ vectors $\mathbf{u}_1$, . . . , $\mathbf{u}_g$ are given by:

$$\mathbf{u}_1 = [\underbrace{1, \ldots, 1}_{n_1}, 0, \ldots, 0, 0, \ldots, 0]', \quad \mathbf{u}_2 = [0, \ldots, 0, \underbrace{1, \ldots, 1}_{n_2}, 0, \ldots, 0]', \ldots$$

$$\mathbf{u}_g = [0, \ldots, 0, 0, \ldots, 0, \underbrace{1, \ldots, 1}_{n_g}]', \quad \mathbf{u}_1 + \mathbf{u}_2 + \ldots + \mathbf{u}_g = \mathbf{1}$$

  Treatment effect vector: perpendicular to $\overline{x}\mathbf{1}$, in the $(g-1)$-D subspace

- Residual vector: $\widehat{\mathbf{e}} = \mathbf{y} - \overline{x}\mathbf{1} - \sum_{l=1}^{g}(\overline{x}_l - \overline{x})\mathbf{u}_l$

  Perpendicular to mean and treatment vectors

  Lying in the $[n - (g-1) - 1] = (n - g)$-dim. subspace

# Univariate ANOVA (v)

## Normal Samples with Common Variance $\sigma^2$

- Summary

| Source of variations | Sum of squares (SS) | Degrees of freedom (d.f.) |
|---|---|---|
| Treatments | $\text{SS}_{\text{tr}} = \sum_{l=1}^{g} n_l(\overline{x}_l - \overline{x})^2$ | $g - 1$ |
| Residual | $\text{SS}_{\text{res}} = \sum_{l=1}^{g} \sum_{j=1}^{n_l} (x_{lj} - \overline{x}_l)^2$ | $\sum_{l=1}^{g} n_l - g$ |
| Total (corrected for the mean) | $\text{SS}_{\text{total}} = \sum_{l=1}^{g} \sum_{j=1}^{n_l} (x_{lj} - \overline{x})^2$ | $\sum_{l=1}^{g} n_l - 1$ |

- Equality of means? Check if treatments are large relative to residuals

  The usual $F$-test rejects $H_0 : \tau_1 = \tau_2 = \ldots = \tau_g = 0$ at the $\alpha$ level if

$$F = \frac{\text{SS}_{\text{tr}}/(g-1)}{\text{SS}_{\text{res}}/(\sum_{l=1}^{g} n_l - g)} > F_{g-1, \sum n_l - g}(\alpha)$$

# One-way MANOVA ($p > 1$) (i)

## Normal Samples with Common Covariance $\boldsymbol{\Sigma}$

- MANOVA model for comparing $g$ population mean vectors

$$\mathbf{X}_{lj} = \underbrace{\boldsymbol{\mu}}_{\substack{\text{overall} \\ \text{mean vector}}} + \underbrace{\boldsymbol{\tau}_l}_{\substack{l\text{-th population} \\ \text{(treatment) effect}}} + \underbrace{\mathbf{e}_{lj}}_{\sim N_p(\mathbf{0},\boldsymbol{\Sigma})}$$

$$l = 1, 2, \ldots, g; \quad j = 1, 2, \ldots, n_l; \qquad \sum_{l=1}^{g} n_l \boldsymbol{\tau}_l = \mathbf{0}$$

- The vector of observations

$$\underbrace{\mathbf{x}_{lj}}_{\text{observation}} = \underbrace{\overline{\mathbf{x}}}_{\substack{\text{overall} \\ \text{sample mean} \\ \widehat{\boldsymbol{\mu}}}} + \underbrace{(\overline{\mathbf{x}}_l - \overline{\mathbf{x}})}_{\substack{\text{estimated} \\ \text{treatment effect} \\ \widehat{\boldsymbol{\tau}}_l}} + \underbrace{(\mathbf{x}_{lj} - \overline{\mathbf{x}}_l)}_{\substack{\text{residual} \\ \widehat{\mathbf{e}}_{lj}}}$$

# One-way MANOVA (ii)

## Normal Samples with Common Covariance $\Sigma$

- Similar to the univariate case, we have: (cross-product terms sum to $\mathbf{0}$)

$$\underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{lj}-\overline{\mathbf{x}})(\mathbf{x}_{lj}-\overline{\mathbf{x}})'}_{\substack{\text{total sum of squares and cross products} \\ \text{matrix } (\mathbf{B}+\mathbf{W}) \text{ (Corrected for the Mean)}}} = \underbrace{\sum_{l=1}^{g} n_l(\overline{\mathbf{x}}_l-\overline{\mathbf{x}})(\overline{\mathbf{x}}_l-\overline{\mathbf{x}})'}_{\substack{\text{between-population sum of squares} \\ \text{and cross products matrix } \mathbf{B} \text{ (Treatment)}}}$$

$$+ \underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{lj}-\overline{\mathbf{x}}_l)(\mathbf{x}_{lj}-\overline{\mathbf{x}}_l)'}_{\substack{\text{within-population sum of squares} \\ \text{and cross products matrix } \mathbf{W} \text{ (Residual)}}}$$

Note that

$$\mathbf{W} = \sum_{l=1}^{g}\underbrace{\sum_{j=1}^{n_l}(\mathbf{x}_{lj}-\overline{\mathbf{x}}_l)(\mathbf{x}_{lj}-\overline{\mathbf{x}}_l)'}_{(n_l-1)\mathbf{S}_l} : \text{ Generalizing } \underbrace{(n_1+n_2-2)\mathbf{S}_{\text{pooled}}}_{\text{in the two-sample case}}$$

$\mathbf{S}_l$: Sample covariance matrix for the $l$-th sample

## One-way MANOVA (iii)

- MANOVA Table

| Source of variations | Matrix of sum of squares and cross products (SSP) | Degrees of freedom (d.f.) |
|---|---|---|
| Treatments | $\mathbf{B} = \sum_{l=1}^{g} n_l(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})'$ | $g - 1$ |
| Residual (Error) | $\mathbf{W} = \sum_{l=1}^{g} \sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)'$ | $\sum_{l=1}^{g} n_l - g$ |
| Total (corrected for the mean) | $\mathbf{B} + \mathbf{W} = \sum_{l=1}^{g} \sum_{j=1}^{n_l} (\mathbf{x}_{lj} - \bar{\mathbf{x}})(\mathbf{x}_{lj} - \bar{\mathbf{x}})'$ | $\sum_{l=1}^{g} n_l - 1$ |

# One-way MANOVA (iv)

## Normal Samples with Common Covariance $\Sigma$

- No treatment effects? Null hypothesis: $H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \ldots = \boldsymbol{\tau}_g = \mathbf{0}$
- Test based on generalized variances: determinant of covariance matrix

  Reject $H_0$ if $\Lambda^*$ is too small, where the test statistic $\Lambda^*$ is given by

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} = \frac{\left|\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{lj}-\overline{\mathbf{x}}_l)(\mathbf{x}_{lj}-\overline{\mathbf{x}}_l)'\right|}{\left|\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{lj}-\overline{\mathbf{x}})(\mathbf{x}_{lj}-\overline{\mathbf{x}})'\right|} = \frac{1}{|\mathbf{W}^{-1}\mathbf{B}+\mathbf{I}|}$$

  $\Lambda^*$: Wilk's lambda: related to the likelihood ratio criterion
- Other statistics for checking equality of several means:
  - Lawley-Hotelling trace $=$ tr$[\mathbf{B}\mathbf{W}^{-1}]$
  - Pillai's trace $=$ tr$[\mathbf{B}(\mathbf{B}+\mathbf{W})^{-1}]$
  - Roy's largest root $=$ maximum eigenvalue of $\mathbf{W}(\mathbf{B}+\mathbf{W})^{-1}$

Other statistics than $\Lambda^*$: see P.336 in the textbook

# One-way MANOVA (v)

## Normal Samples with Common Covariance $\Sigma$

- For some special cases: exact distributions

| # of variables | # of groups | Distribution |
|---|---|---|
| $p = 1$ | $g \geq 2$ | $\left(\frac{\sum n_l - g}{g-1}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{g-1, \sum n_l - g}$ |
| $p = 2$ | $g \geq 2$ | $\left(\frac{\sum n_l - g - 1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1), 2(\sum n_l - g - 1)}$ |
| $p \geq 1$ | $g = 2$ | $\left(\frac{\sum n_l - p - 1}{p}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{p, \sum n_l - p - 1}$ |
| $p \geq 1$ | $g = 3$ | $\left(\frac{\sum n_l - p - 2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p, 2(\sum n_l - p - 2)}$ |

Table 6.3 in the textbook

# One-way MANOVA (vi)

- For other cases and large sample sizes:
  - a modification of $\Lambda^*$ due to Bartlett

  If $H_0$ is true and $\sum n_l = n$ is large, then

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

$$\overset{\text{approx. dist. as}}{\approx} \quad \chi^2_{p(g-1)}$$

- Therefore, for large $\sum n_l = n$, reject $H_0$ at significance level $\alpha$ if

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

# Example 6: MANOVA (i)

## Multivariate Analysis of Wisconsin Nursing Home Data

- Investigating the effects of ownership/certification/both on $p = 4$ costs
  - $X_1$: nursing labor
  - $X_2$: dietary labor
  - $X_3$: plant operation and maintenance labor
  - $X_4$: housekeeping and laundry labor
- A total of $n = 516$ observations for $g = 3$ groups
- Summary statistics: (next page)

---

Example 6.10 in the textbook

ANOVA and MANOVA packages available in many computing softwares

Ownership including private party, nonprofit organization, government; Certification including Skilled nursing facilities, intermediate care facilities, etc

# Example 6: MANOVA (ii)

## Multivariate Analysis of Wisconsin Nursing Home Data

- Summary statistics: $g = 3, p = 4, n = \sum_{l=1}^{3} n_l = 516$
  (Group 1: Private; Group 2: Nonprofit; Group 3: Government)

| Group | # of observ. | Sample mean vectors |
|---|---|---|
| $l = 1$ | $n_1 = 271$ | $\overline{\mathbf{x}}_1 = \begin{bmatrix} 2.066 & 0.480 & 0.082 & 0.360 \end{bmatrix}'$ |
| $l = 2$ | $n_2 = 138$ | $\overline{\mathbf{x}}_2 = \begin{bmatrix} 2.167 & 0.596 & 0.124 & 0.418 \end{bmatrix}'$ |
| $l = 3$ | $n_3 = 107$ | $\overline{\mathbf{x}}_3 = \begin{bmatrix} 2.273 & 0.521 & 0.125 & 0.383 \end{bmatrix}'$ |

Sample covariance matrices: $\mathbf{S}_1 = \begin{bmatrix} 0.291 & & & \\ -0.001 & 0.011 & & \\ 0.002 & 0.000 & 0.001 & \\ 0.010 & 0.003 & 0.000 & 0.010 \end{bmatrix}$

$\mathbf{S}_2 = \begin{bmatrix} 0.561 & & & \\ 0.011 & 0.025 & & \\ 0.001 & 0.004 & 0.005 & \\ 0.037 & 0.007 & 0.002 & 0.019 \end{bmatrix}$ $\mathbf{S}_3 = \begin{bmatrix} 0.261 & & & \\ 0.030 & 0.017 & & \\ 0.003 & -0.000 & 0.004 & \\ 0.018 & 0.006 & 0.001 & 0.013 \end{bmatrix}$

# Example 6: MANOVA (iii)

## Multivariate Analysis of Wisconsin Nursing Home Data

- We can obtain: (assuming all $\mathbf{S}_l$'s compatible for now)

$$\mathbf{W} = \sum_{l=1}^{3}(n_l - 1)\mathbf{S}_l = \begin{bmatrix} 182.962 & & & \\ 4.408 & 8.200 & & \\ 1.695 & 0.633 & 1.484 & \\ 9.581 & 2.428 & 0.394 & 6.538 \end{bmatrix}$$

$$\overline{\mathbf{x}} = \frac{n_1\overline{\mathbf{x}}_1 + n_2\overline{\mathbf{x}}_2 + n_3\overline{\mathbf{x}}_3}{n_1 + n_2 + n_3} = [2.136 \;\; 0.519 \;\; 0.102 \;\; 0.380]'$$

$$\mathbf{B} = \sum_{l=1}^{3} n_l(\overline{\mathbf{x}}_l - \overline{\mathbf{x}})(\overline{\mathbf{x}}_l - \overline{\mathbf{x}})' = \begin{bmatrix} 3.475 & & & \\ 1.111 & 1.225 & & \\ 0.821 & 0.453 & 0.235 & \\ 0.584 & 0.610 & 0.230 & 0.304 \end{bmatrix}$$

$$\Lambda^* = |\mathbf{W}|/|\mathbf{B} + \mathbf{W}| = 0.7714$$

# Example 6: MANOVA (iv)

## Multivariate Analysis of Wisconsin Nursing Home Data

- Testing $H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \boldsymbol{\tau}_3$ (no effect of ownership on costs) $(\alpha = 0.01)$

1. Since $p = 4, g = 3$, we can use the exact distribution table (last row):

$$\frac{(\sum n_l - p - 2)}{p} \cdot \frac{(1 - \sqrt{\Lambda^*})}{\sqrt{\Lambda^*}} = \frac{(516 - 4 - 2)}{4} \cdot \frac{(1 - \sqrt{0.7714})}{\sqrt{0.7714}} = 17.67$$

$$> F_{2p, 2(\sum n_l - p - 2)}(0.01) = F_{8,1020}(0.01) = 2.5287$$

$\Longrightarrow$ Reject $H_0$ at $\alpha = 0.01$ (Conclusion: average costs differ)

2. Since $n = 516$ is large here, we may also use Bartlett's method:

$$-(n - 1 - (p + g)/2) \ln \Lambda^* = -(511.5) \ln 0.7714 = 132.76$$

$$> \chi^2_{p(g-1)}(0.01) = \chi^2_8(0.01) = 20.0902$$

$\Longrightarrow$ Reject $H_0$ at $\alpha = 0.01$

(Two results are consistent here.)

## Simultaneous Confidence Intervals for Treatment Effects

- Investigating mean differences when the equal mean hypothesis is rejected
- Bonferroni approach for pairwise comparisons ($\boldsymbol{\tau}_k - \boldsymbol{\tau}_l$ or $\boldsymbol{\mu}_k - \boldsymbol{\mu}_l$)
  - Let $\tau_{ki}$ be the $i$-th element of $\boldsymbol{\tau}_k$ (estimated by $\widehat{\boldsymbol{\tau}}_k = \overline{\mathbf{x}}_k - \overline{\mathbf{x}}$)
  - Independent random samples between populations

$$\Longrightarrow \mathsf{Var}(\widehat{\tau}_{ki} - \widehat{\tau}_{li}) = \mathsf{Var}(\overline{X}_{ki} - \overline{X}_{li}) = (1/n_k + 1/n_l)\sigma_{ii}$$

  $\sigma_{ii}$: the $i$-th diagonal element of $\boldsymbol{\Sigma}$

  - Furthermore, let $w_{ii}$ be the $i$-th diagonal element of $\mathbf{W}$

$$\widehat{\sigma}_{ii} = \frac{w_{ii}}{n-g}, n = \sum_{l=1}^{g} n_l \quad \Longrightarrow \quad \widehat{\mathsf{Var}}(\overline{X}_{ki} - \overline{X}_{li}) = \left(\frac{1}{n_k} + \frac{1}{n_l}\right)\frac{w_{ii}}{n-g}$$

  - $g$ groups with $p$-dim. data $\Longrightarrow$ $p$ variables and $g(g-1)/2$ pairwise differences; so each two-sample $t$-interval in the Bonferroni approach will use the critical value $t_{n-g}(\alpha/2m)$ where $m = pg(g-1)/2$.

# Simultaneous Confidence Intervals for Treatment Effects

(Cont'd from previous page)

- Bonferroni simultaneous confidence intervals

  Let $n = \sum_{l=1}^{g} n_l$. For the one-way MANOVA model, with probability at least $(1-\alpha)$, $\tau_{ki} - \tau_{li}$ belongs to

$$\overline{x}_{ki} - \overline{x}_{li} \pm t_{n-g}\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\frac{w_{ii}}{n-g}\left(\frac{1}{n_k}+\frac{1}{n_l}\right)}$$

  for all components $i = 1, 2, \ldots, p$ and all differences $l \leq k = 1, 2, \ldots, g$. Here $w_{ii}$ is the $i$-th diagonal element of $\mathbf{W}$.

- To see an example, check p. 309 (Example 6.11 there) in the textbook (based on nursing homes data).

# Testing for Equality of Covariance Matrices (i)

- Previous assumption: equal covariance matrix; this needs to be tested.
- Setup: $g$ populations; $p$ variables; $\mathbf{\Sigma}_l$: positive definite, $l = 1, \ldots, g$
  Testing the null hypothesis: $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_g = \mathbf{\Sigma}$
  against the alternative: $H_1 : \mathbf{\Sigma}_k \neq \mathbf{\Sigma}_j$ for some $1 \leq k \neq j \leq g$
- (Normal populations) Likelihood ratio statistic for testing $H_0$ vs. $H_1$:

$$\Lambda = \prod_{l=1}^{g} \left( \frac{|\mathbf{S}_l|}{|\mathbf{S}_{\text{pooled}}|} \right)^{\frac{n_l - 1}{2}}$$

$$\mathbf{S}_{\text{pooled}} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \ldots + (n_g - 1)\mathbf{S}_g}{\sum_{l=1}^{g}(n_l - 1)}$$

  - $n_l$: sample size for the $l$-th population
  - $\mathbf{S}_l$: sample cov. matrix for the $l$-th population
  - $\mathbf{S}_{\text{pooled}}$: pooled sample cov. matrix

## Testing for Equality of Covariance Matrices (ii)

### Box's $M$-Test

- Box's $M$ statistic:

$$M \triangleq -2\ln\Lambda = \left[\sum_{l=1}^{g}(n_l-1)\right]\ln|\mathbf{S}_{\text{pooled}}| - \sum_{l=1}^{g}\left[(n_l-1)\ln|\mathbf{S}_l|\right]$$

Under $H_0$: $\mathbf{S}_l$ close to $\mathbf{S}_{\text{pooled}}$, $\Lambda$ close to 1 $\Longrightarrow$ $M$-statistic: small

- Box's $M$-test: based on a $\chi^2$ approx. to sampling dist. of $M = -2\ln\Lambda$

Set ($p$: number of variables; $g$: number of populations)

$$u \triangleq \left[\sum_l \frac{1}{(n_l-1)} - \frac{1}{\sum_l(n_l-1)}\right]\left[\frac{2p^2+3p-1}{6(p+1)(g-1)}\right]$$

Then

$$C = (1-u)M \overset{\text{approx. dist. as}}{\approx} \chi^2_v, \quad \text{where} \quad v = \frac{1}{2}p(p+1)(g-1)$$

$\Longrightarrow$ Reject $H_0$ if $C > \chi^2_{p(p+1)(g-1)/2}(\alpha)$ at significance level $\alpha$

(The above $\chi^2$ approx. works well if $p, g \le 5$ and $n_l \ge 20, \forall l$)

## Example 7: Box's $M$-Test (i)
### Nursing Home Data in Example 6

- Use the nursing home data and test $H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3 = \mathbf{\Sigma}$
- Recall:

  $(p = 4, g = 3, n_1 = 271, n_2 = 138, n_3 = 107, n = \sum_{l=1}^{3} n_l = 516)$

  Sample covariance matrices:

$$\mathbf{S}_1 = \begin{bmatrix} 0.291 & & & \\ -0.001 & 0.011 & & \\ 0.002 & 0.000 & 0.001 & \\ 0.010 & 0.003 & 0.000 & 0.010 \end{bmatrix}$$

$$\mathbf{S}_2 = \begin{bmatrix} 0.561 & & & \\ 0.011 & 0.025 & & \\ 0.001 & 0.004 & 0.005 & \\ 0.037 & 0.007 & 0.002 & 0.019 \end{bmatrix} \mathbf{S}_3 = \begin{bmatrix} 0.261 & & & \\ 0.030 & 0.017 & & \\ 0.003 & -0.000 & 0.004 & \\ 0.018 & 0.006 & 0.001 & 0.013 \end{bmatrix}$$

# Example 7: Box's $M$-Test (ii)

- We can calculate the following:

$$\ln|\mathbf{S}_1| = -17.397, \qquad \ln|\mathbf{S}_2| = -13.926$$

$$\ln|\mathbf{S}_3| = -15.741, \qquad \ln|\mathbf{S}_{\text{pooled}}| = -15.564$$

$$u = \left(\frac{1}{270} + \frac{1}{137} + \frac{1}{106} + \frac{1}{516-3}\right)\frac{2(4^2)+3(4)-1}{6(4+1)(3-1)} = 0.0132$$

$$M = 513(-15.564) - [270(-17.397) + 137(-13.926) + 106(-15.741)]$$

$$= 289.266$$

$$C = (1-u)M = (1-0.0132)289.266 = 285.4$$

$$v = p(p+1)(g-1)/2 = 4(5)(2)/2 = 20$$

Set $\alpha$ at any reasonable level $\Longrightarrow C > \chi^2_{20}(\alpha) \Longrightarrow$ Reject $H_0$
Conclusion: the three covariance matrices in nursing home data are not equal.

# Two-way Univariate/Multivariate Analysis of Variance

- Starting with the univariate case (Two-way ANOVA)
- Then proceeding with the multivariate case (Two-way MANOVA)

# Two-way ANOVA (i)

- The model

$$X_{lkr} = \mu + \tau_l + \beta_k + \gamma_{lk} + \underbrace{e_{lkr}}_{\text{i.i.d. } N(0,\sigma^2)}$$

$$(l = 1, \ldots, g; \quad k = 1, \ldots, b; \quad r = 1, \ldots, n)$$

$$\left( \sum_{l=1}^{g} \tau_l = \sum_{k=1}^{b} \beta_k = \sum_{l=1}^{g} \gamma_{lk} = \sum_{k=1}^{b} \gamma_{lk} = 0 \right)$$

  - $2$ factors: factor 1 and factor 2
  - $g$ levels of factor 1 and $b$ levels of factor 2
  - $n$ independent observations of $gb$ combinations of levels
  - $X_{lkr}$: the $r$-th observ. at level $l$ of factor 1 and level $k$ of factor 2
  - $\mu$: overall mean (general level of response)
  - $\tau_l$: fixed effect of factor 1; $\beta_k$: fixed effect of factor 2
  - $\gamma_{lk}$: interaction between factor 1 and factor 2

# Two-way ANOVA (ii)

- The expected response at level $l$ of factor 1 and level $k$ of factor 2:

$$\mathsf{E}(X_{lkr}) = \mu + \tau_l + \beta_k + \gamma_{lk} \quad (l = 1, \ldots, g; \quad k = 1, \ldots, b)$$

- The data:

$$x_{lkr} = \overline{x} + (\overline{x}_{l\bullet} - \overline{x}) + (\overline{x}_{\bullet k} - \overline{x}) + (\overline{x}_{lk} - \overline{x}_{l\bullet} - \overline{x}_{\bullet k} + \overline{x}) + (x_{lkr} - \overline{x}_{lk})$$

  - $\overline{x}$: overall average (overall sample mean)
  - $\overline{x}_{l\bullet}$: average for level $l$ of factor 1;
  - $\overline{x}_{\bullet k}$: average for level $k$ of factor 2;
  - $\overline{x}_{lk}$: average for level $l$ of factor 1 and level $k$ of factor 2

$$\overline{x}_{l\bullet} = \frac{1}{bn}\sum_{k=1}^{b}\sum_{r=1}^{n} x_{lkr}, \; \overline{x}_{\bullet k} = \frac{1}{gn}\sum_{l=1}^{g}\sum_{r=1}^{n} x_{lkr}, \; \overline{x}_{lk} = \frac{1}{n}\sum_{r=1}^{n} x_{lkr}$$

# Two-way ANOVA (iii)

- It can be shown that

$$\underbrace{\sum_{l=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(x_{lkr}-\overline{x})^2}_{\text{SS}_{\text{total}}} = \underbrace{\sum_{l=1}^{g} bn(\overline{x}_{l\bullet}-\overline{x})^2}_{\text{SS}_{\text{fac1}}} + \underbrace{\sum_{k=1}^{b} gn(\overline{x}_{\bullet k}-\overline{x})^2}_{\text{SS}_{\text{fac2}}}$$

$$+ \underbrace{\sum_{l=1}^{g}\sum_{k=1}^{b} n(\overline{x}_{lk}-\overline{x}_{l\bullet}-\overline{x}_{\bullet k}+\overline{x})^2}_{\text{SS}_{\text{int}}} + \underbrace{\sum_{l=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(x_{lkr}-\overline{x}_{lk})^2}_{\text{SS}_{\text{res}}}$$

- Corresponding degrees of freedom:

$$\underbrace{\text{SS}_{\text{total}}}_{gbn-1} = \underbrace{\text{SS}_{\text{fac1}}}_{g-1} + \underbrace{\text{SS}_{\text{fac2}}}_{b-1} + \underbrace{\text{SS}_{\text{int}}}_{(g-1)(b-1)} + \underbrace{\text{SS}_{\text{res}}}_{gb(n-1)}$$

# Two-way ANOVA (iv)

- Two-way ANOVA table (SS: sum of squares, d.f.: degree of freedom, MS: mean squares, SoVAR: source of variation)

| SoVAR | SS | d.f. | MS | $F$-ratio |
|---|---|---|---|---|
| Factor 1 | $\text{SS}_{\text{fac1}}$ | $g-1$ | $\text{MS}_{\text{fac1}} = \frac{\text{SS}_{\text{fac1}}}{g-1}$ | $\frac{\text{MS}_{\text{fac1}}}{\text{MS}_{\text{res}}}$ |
| Factor 2 | $\text{SS}_{\text{fac2}}$ | $b-1$ | $\text{MS}_{\text{fac2}} = \frac{\text{SS}_{\text{fac2}}}{b-1}$ | $\frac{\text{MS}_{\text{fac2}}}{\text{MS}_{\text{res}}}$ |
| Interaction | $\text{SS}_{\text{int}}$ | $(g-1)(b-1)$ | $\text{MS}_{\text{int}} = \frac{\text{SS}_{\text{int}}}{(g-1)(b-1)}$ | $\frac{\text{MS}_{\text{int}}}{\text{MS}_{\text{res}}}$ |
| Residual | $\text{SS}_{\text{res}}$ | $gb(n-1)$ | $\text{MS}_{\text{res}} = \frac{\text{SS}_{\text{res}}}{gb(n-1)}$ | |
| Total | $\text{SS}_{\text{total}}$ | $gbn-1$ | | |

- To test the hypothesis of no interaction $H_0 : \gamma_{11} = \gamma_{12} = \ldots = \gamma_{gb} = 0$ vs. $H_1$: at least one $\gamma_{lk} \neq 0$ (for some $l, k$), we can use the $F$-ratio $\frac{\text{MS}_{\text{int}}}{\text{MS}_{\text{res}}}$. (Similar tests for the factor effects)

# Two-way MANOVA (i)

- Parallel to the univariate case, here the model is: (vectors are $p \times 1$)

$$\mathbf{X}_{lkr} = \boldsymbol{\mu} + \boldsymbol{\tau}_l + \boldsymbol{\beta}_k + \boldsymbol{\gamma}_{lk} + \underbrace{\mathbf{e}_{lkr}}_{\text{i.i.d. } N_p(\mathbf{0},\boldsymbol{\Sigma})}$$

$$(l = 1, \ldots, g; \quad k = 1, \ldots, b; \quad r = 1, \ldots, n)$$

$$\left( \sum_{l=1}^{g} \boldsymbol{\tau}_l = \sum_{k=1}^{b} \boldsymbol{\beta}_k = \sum_{l=1}^{g} \boldsymbol{\gamma}_{lk} = \sum_{k=1}^{b} \boldsymbol{\gamma}_{lk} = \mathbf{0} \right)$$

# Two-way MANOVA (ii)

- The data can be written as:

$$\mathbf{x}_{lkr} = \overline{\mathbf{x}} + (\overline{\mathbf{x}}_{l\bullet} - \overline{\mathbf{x}}) + (\overline{\mathbf{x}}_{\bullet k} - \overline{\mathbf{x}}) + (\overline{\mathbf{x}}_{lk} - \overline{\mathbf{x}}_{l\bullet} - \overline{\mathbf{x}}_{\bullet k} + \overline{\mathbf{x}}) + (\mathbf{x}_{lkr} - \overline{\mathbf{x}}_{lk})$$

  – $\overline{\mathbf{x}}$: overall average of observ. vectors (overall sample mean)
  – $\overline{\mathbf{x}}_{l\bullet}$: average of observ. vectors for level $l$ of factor 1;
  – $\overline{\mathbf{x}}_{\bullet k}$: average of observ. vectors for level $k$ of factor 2;
  – $\overline{\mathbf{x}}_{lk}$: average of observ. vectors at the $l$-th level of factor 1 and the $k$-th level of factor 2

# Two-way MANOVA (iii)

- It can be shown that

$$\sum_{l=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(\mathbf{x}_{lkr}-\overline{\mathbf{x}})(\mathbf{x}_{lkr}-\overline{\mathbf{x}})'$$

$$= \sum_{l=1}^{g} bn(\overline{\mathbf{x}}_{l\bullet}-\overline{\mathbf{x}})(\overline{\mathbf{x}}_{l\bullet}-\overline{\mathbf{x}})' + \sum_{k=1}^{b} gn(\overline{\mathbf{x}}_{\bullet k}-\overline{\mathbf{x}})(\overline{\mathbf{x}}_{\bullet k}-\overline{\mathbf{x}})'$$

$$+ \sum_{l=1}^{g}\sum_{k=1}^{b} n(\overline{\mathbf{x}}_{lk}-\overline{\mathbf{x}}_{l\bullet}-\overline{\mathbf{x}}_{\bullet k}+\overline{\mathbf{x}})(\overline{\mathbf{x}}_{lk}-\overline{\mathbf{x}}_{l\bullet}-\overline{\mathbf{x}}_{\bullet k}+\overline{\mathbf{x}})'$$

$$+ \sum_{l=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(\mathbf{x}_{lkr}-\overline{\mathbf{x}}_{lk})(\mathbf{x}_{lkr}-\overline{\mathbf{x}}_{lk})'$$

or

$$\text{SSP}_{\text{total}} = \text{SSP}_{\text{fac1}} + \text{SSP}_{\text{fac2}} + \text{SSP}_{\text{int}} + \text{SSP}_{\text{res}}$$

# Two-way MANOVA (iv)

- Corresponding degrees of freedom:

$$\underbrace{\mathsf{SSP}_{\mathsf{total}}}_{gbn-1} = \underbrace{\mathsf{SSP}_{\mathsf{fac1}}}_{g-1} + \underbrace{\mathsf{SSP}_{\mathsf{fac2}}}_{b-1} + \underbrace{\mathsf{SSP}_{\mathsf{int}}}_{(g-1)(b-1)} + \underbrace{\mathsf{SSP}_{\mathsf{res}}}_{gb(n-1)}$$

or in a table form: (SSP: matrix of sum of squares and cross products)

| Source of variation | SSP | d.f. |
|---|---|---|
| Factor 1 | $\mathsf{SSP}_{\mathsf{fac1}}$ | $g-1$ |
| Factor 2 | $\mathsf{SSP}_{\mathsf{fac2}}$ | $b-1$ |
| Interaction | $\mathsf{SSP}_{\mathsf{int}}$ | $(g-1)(b-1)$ |
| Residual (Error) | $\mathsf{SSP}_{\mathsf{res}}$ | $gb(n-1)$ |
| Total | $\mathsf{SSP}_{\mathsf{total}}$ | $gbn-1$ |

- Tests: based on generalized variances (see next page)

# Two-way MANOVA (v)

- Effects of interaction:

  Testing $H_0 : \boldsymbol{\gamma}_{11} = \boldsymbol{\gamma}_{12} = \ldots = \boldsymbol{\gamma}_{gb} = \mathbf{0}$ vs. $H_1$: at least one $\boldsymbol{\gamma}_{lk} \neq \mathbf{0}$

  (Likelihood ratio test; $H_0$: No interaction effects)

  Reject $H_0$ if the following likelihood ratio statistic is too small:

$$\Lambda^*_{\mathsf{int}} \triangleq \frac{|\mathsf{SSP}_{\mathsf{res}}|}{|\mathsf{SSP}_{\mathsf{int}} + \mathsf{SSP}_{\mathsf{res}}|}$$

  For large samples, use Bartlett's correction:

  $\Longrightarrow$ Reject $H_0$ at level $\alpha$ if

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

# Two-way MANOVA (vi)

- Effects of factor 1:

  Testing $H_0 : \boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \ldots = \boldsymbol{\tau}_g = \mathbf{0}$ vs. $H_1$: at least one $\boldsymbol{\tau}_l \neq \mathbf{0}$

  ($H_0$: No factor 1 effects)

  Let

$$\Lambda^*_{\mathsf{fac1}} \triangleq \frac{|\mathsf{SSP}_{\mathsf{res}}|}{|\mathsf{SSP}_{\mathsf{fac1}} + \mathsf{SSP}_{\mathsf{res}}|}$$

  For large samples, use Bartlett's correction again:

  $\Longrightarrow$ Reject $H_0$ at level $\alpha$ if

$$-\left[gb(n-1) - \frac{p+1-(g-1)}{2}\right] \ln \Lambda^*_{\mathsf{fac1}} > \chi^2_{(g-1)p}(\alpha)$$

## Two-way MANOVA (vii)

- Effects of factor 2:

  Testing $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \ldots = \boldsymbol{\beta}_b = \mathbf{0}$ vs. $H_1$: at least one $\boldsymbol{\beta}_k \neq \mathbf{0}$

  ($H_0$: No factor 2 effects)

  Similarly, let

$$\Lambda^*_{\text{fac2}} \triangleq \frac{|\text{SSP}_{\text{res}}|}{|\text{SSP}_{\text{fac2}} + \text{SSP}_{\text{res}}|}$$

  For large samples, use Bartlett's correction:

  $\Longrightarrow$ Reject $H_0$ at level $\alpha$ if

$$-\left[gb(n-1) - \frac{p+1-(b-1)}{2}\right] \ln \Lambda^*_{\text{fac2}} > \chi^2_{(b-1)p}(\alpha)$$

- When a null hypothesis is rejected, we may use Bonferroni method to obtain simultaneous confidence intervals for further analysis.