STT 843: Multivariate Analysis

4. Inferences about a Mean Vector (Chapter 5.5-5.8)

Guanqun Cao

Department of Statistics and Probability
Michigan State University

Spring 2026

Outline

Large Sample Inference about a Population Mean Vector

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample from a population with mean $\boldsymbol{\mu}$ and positive definite covariance matrix Σ. When $n - p$ is large, the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ is rejected in favor of $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, at a level of significance approximately α, if the observed

$$n\left(\overline{\mathbf{x}} - \boldsymbol{\mu}_0\right)^{\mathsf{T}} \mathbf{S}^{-1} \left(\overline{\mathbf{x}} - \boldsymbol{\mu}_0\right) > \chi_p^2(\alpha)$$

Here $\chi_p^2(\alpha)$ is the upper (100α) th percentile of a chi-square distribution with p.d.f.

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample from a population with mean $\boldsymbol{\mu}$ and positive definite covariance matrix Σ. If $n - p$ is large,

$$\mathbf{a}^\mathsf{T}\overline{\boldsymbol{X}} \pm \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{\mathbf{a}^\mathsf{T}\mathbf{S}\mathbf{a}}{n}}$$

will contain $\mathbf{a}^\mathsf{T}\boldsymbol{\mu}$, for every $\mathbf{a}$, with probability approximately $1 - \alpha$.

Example 4.6 (Constructing large sample simultaneous confidence intervals)

A music educator tested thousands of Finnish students on their native musical ability in order to set national norms in Finland. Summary statistics for part of the data set are given in Table 5.5. These statistics are based on a sample of $n = 96$ Finnish 12 th graders. Construct 90% simultaneous confidence intervals for individual mean components $\mu_i, i = 1, 2, \ldots, 7$.

Paired Comparisons

In the single response (univariate) case, let X_{j1} denote the response to treatment 1 , and let X_{j2} denote the response to treatment 2 for the j th trial. That is, (X_{j1}, X_{j2}) are measurements recorded on the j th unit or j th pair of like units. By design, the n differences

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

Should reflect only the differences D_j represent independent observations from an $N(\delta, \sigma_d^2)$ distribution. the variable

$$t = \frac{\bar{D} - \delta}{s_d / \sqrt{n}}$$

where $\bar{D} = \frac{1}{n}\sum_{j=1}^{n} D_j$ and $s_d^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(D_j - \bar{D}\right)^2$ has a t-distribution with $n - 1$ d.f.

- An α-level test of

$$H_0 : \delta = 0 \quad \text{vs} \quad H_1 : \delta \neq 0$$

may be conducted by comparing $|t|$ with $t_{n-1}(\alpha/2)$-the upper $100(\alpha/2)$ th percentile of a t-distribution with $n-1$ d.f.

- A $100(1-\alpha)\%$ confidence interval for the mean difference $\delta = \mathrm{E}(X_{j1} - X_{j2})$ is provided the statement

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

- Multivariate extension of the paired-comparison procedure to distinguish between p response, two treatments, and n experimental units. The p paired difference random variables become

$$
\begin{aligned}
D_{j1} &= X_{1j1} - X_{2j1} \\
D_{j2} &= X_{1j2} - X_{2j2} \\
&\vdots \\
D_{jp} &= X_{1jp} - X_{2jp}
\end{aligned}
$$

Let $\boldsymbol{D}_j^{\mathsf{T}} = [D_{j1}, D_{j2}, \ldots, D_{jp}]$, and assume for $j = 1, 2, \ldots, n$ that

$$\mathrm{E}(\boldsymbol{D}_j) = \boldsymbol{\delta} = [\delta_1, \delta_2, \cdots, \delta_p]^{\mathsf{T}} \quad \text{and} \quad \mathrm{Cov}(\boldsymbol{D}_j) = \Sigma_d$$

If, in addition, $\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_n$ are independent $N_p(\boldsymbol{\delta}, \Sigma_d)$ random vectors, inference about the vector of mean differences $\boldsymbol{\delta}$ can based upon a T^2 statistics

$$T^2 = n(\overline{\boldsymbol{D}} - \boldsymbol{\delta})^{\mathsf{T}} \mathbf{S}_d^{-1} (\overline{\boldsymbol{D}} - \boldsymbol{\delta})$$

where $\overline{\boldsymbol{D}} = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{D}_j$ and $\mathbf{S}_d = \frac{1}{n-1} \sum_{j=1}^{n} \left(\boldsymbol{D}_j - \overline{\boldsymbol{D}}\right) \left(\boldsymbol{D}_j - \overline{\boldsymbol{D}}\right)^{\mathsf{T}}$.

Let the differences $\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_n$ be a random sample from $N_p(\boldsymbol{\delta}, \Sigma_d)$ population. Then

$$T^2 = n(\overline{\boldsymbol{D}} - \boldsymbol{\delta})^{\mathsf{T}} \mathbf{S}_d^{-1} (\overline{\boldsymbol{D}} - \boldsymbol{\delta})$$

is distributed as an $[(n-p)p/(n-p)F_{p,n-p}]$ random variable, whatever the true $\boldsymbol{\delta}$ and Σ_d.

Given the observed differences
$\boldsymbol{D}_j^{\mathsf{T}} = [D_{j1}, D_{j2}, \ldots, D_{jp}], j = 1, 2, \ldots, n,$

- an α-level test of

$$H_0 : \boldsymbol{\delta} = 0 \quad \text{versus} \quad H_1 : \boldsymbol{\delta} = 0$$

for an $N_p(\boldsymbol{\delta}, \Sigma_d)$ population rejects H_0 if the observed

$$T^2 = n\overline{\boldsymbol{D}}^{\mathsf{T}}\mathbf{S}_d^{-1}\overline{\boldsymbol{D}} > \frac{(n-p)p}{n-p}F_{p,n-p}(\alpha)$$

where $F_{p,n-p}(\alpha)$ is the upper $(100\alpha)\%$ th percentile of an F-distribution with p and $n-p$ d.f.

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

$$(\overline{\boldsymbol{D}}-\boldsymbol{\delta})^{\mathsf{T}}\mathbf{S}_d^{-1}(\overline{\boldsymbol{D}}-\boldsymbol{\delta}) > \frac{(n-p)p}{n(n-p)}F_{p,n-p}$$

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

$$\delta_i : \overline{\boldsymbol{D}}_i \pm \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{D_i}^2}{n}}$$

where $\overline{\boldsymbol{D}}_i$ is the i th element of $\overline{\boldsymbol{D}}$ and $s_{D_i}^2$ is the i th diagonal element of $\mathbf{S}_d^{22}$.

Example 4.7 (Checking for a mean difference with paired observations) Muicipal wasterwater treatment plants are required by law to monitor their discharges into rivers and streams on a regular basis. Concern about the reliability of data from one of these self-monitoring programs led to a study in which samples of effluent were divided and sent to two laboratories for testing. One-half was sent to a private commercial laboratory routinely used in the monitoring program. Measurements of biochemical oxygen demand (BOD) and suspended solid (SS) were obtained, for $n = 11$ sample splits, from the two laboratories.

Do the two laboratories's chemical analyses agree? If differences exist, what is their nature?

Alternatively, think of each observation

$$\underset{2p\times 1}{\mathbf{x}_i} = \begin{bmatrix} \mathbf{x}_{1i} \\ \mathbf{x}_{2i} \end{bmatrix} \begin{matrix} \leftarrow \text{ pre-tests} \\ \leftarrow \text{ post-tests} \end{matrix}$$

$$\underset{2p\times 1}{\bar{\mathbf{x}}} = \begin{bmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \end{bmatrix}$$

$$\underset{2p\times 2p}{\mathbf{S}} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}$$

Interest is in $\mathbf{C}\mathbf{x}_i$, where

$$\underset{p\times 2p}{\mathbf{C}} = \begin{bmatrix} 1 & & & \mathbf{0} & -1 & & & \mathbf{0} \\ & 1 & & & & -1 & & \\ & & \ddots & & & & \ddots & \\ \mathbf{0} & & & 1 & \mathbf{0} & & & -1 \end{bmatrix}$$

Note

$$\mathbf{d}_i = \mathbf{C}\mathbf{x}_i$$
$$\bar{\mathbf{d}} = \mathbf{C}\bar{\mathbf{x}}$$
$$\mathbf{S}_d = \mathbf{C}\mathbf{S}\mathbf{C}'$$

$$T^2 = n\bar{\mathbf{x}}'\mathbf{C}'\left(\mathbf{C}\mathbf{S}\mathbf{C}'\right)^{-1}\mathbf{C}\bar{\mathbf{x}} \sim T^2_{p,n-1}$$

and

$$= \frac{(n-1)p}{(n-1-p+1)}F_{p,n-1-p+1}$$
$$\underline{=}^{q} \frac{(n-1)p}{n-p}F_{p,n-p}$$

An extension to a comparison of p treatments given to each subject over time

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \begin{matrix} \\ \leftarrow \text{ evaluation after day 2 dosage} \\ \leftarrow \text{ evaluation after day j dosage} \\ \\ \end{matrix}$$

Interest may lie in comparisons of treatment means

$$\underset{(p-1)\times p}{\mathbf{C}}\boldsymbol{\mu} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} \mu_2 - \mu_1 \\ \mu_3 - \mu_2 \\ \vdots \\ \mu_p - \mu_{p-1} \end{bmatrix}$$

$$T^2 = n(\mathbf{C}\bar{\mathbf{x}})' \left(\mathbf{CSC}'\right)^{-1} \mathbf{C}\bar{\mathbf{x}} \sim T^2_{p-1,n-1}$$
$$\overset{q}{=} \frac{(n-1)(p-1)}{(n-1-(p-1)+1)} F_{(p-1),n-1-(p-1)+1}$$
$$\overset{q}{=} \frac{(n-1)(p-1)}{n-p+1} F_{p-1,n-p+1}$$

e.g., if comparing 3 days, we might use

$$\underset{2\times 3}{\mathbf{C}} = \begin{bmatrix} -1 & 0 & 1 \\ 1 & -2 & 1 \end{bmatrix} \begin{matrix} \leftarrow \text{ linear} \\ \leftarrow \text{ quadradic} \end{matrix}$$

e.g., if comparing 4 days, we might use

$$\underset{3\times 4}{\mathbf{C}} = \begin{bmatrix} -3 & -1 & 1 & 3 \\ 1 & -1 & -1 & 1 \\ 1 & -3 & 3 & -1 \end{bmatrix} \begin{matrix} \leftarrow \text{linear} \\ \leftarrow \text{quadradic} \\ \leftarrow \text{cubic} \end{matrix}$$

Sample	Summary statistics	
(Population 1)		
$\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}$	$\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{x}_{1j}$	$\mathbf{S}_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j}-\bar{\mathbf{x}}_1)(\mathbf{x}_{1j}-\bar{\mathbf{x}}_1)^{\mathsf{T}}$
(Population 2)		
$\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_2}$	$\bar{\mathbf{x}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{x}_{2j}$	$\mathbf{S}_2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\mathbf{x}_{2j}-\bar{\mathbf{x}}_2)(\mathbf{x}_{2j}-\bar{\mathbf{x}}_2)^{\mathsf{T}}$

Two-sample comparison

1. The sample $\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1}$ is a random sample of size n_1 from a p-variate population with mean $\boldsymbol{\mu}_1$ and covariance matrix Σ_1.
2. The sample $\boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_2}$ is a random sample of size n_1 from a p-variate population with mean μ_2 and covariance matrix Σ_2.
3. Also, $\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1}$ are independent of $\boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_2}$.

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0 \quad \text{vs} \quad H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq 0$$

Further Assumption When n_1 and n_2 Are Small

1. Both populations are multivariate normal.
2. Also, $\Sigma_1 = \Sigma_2$ (same covariance matrix)

Set the estimate of Σ as

$$\mathbf{S}_{\text{pooled}} = \frac{\sum_{j=1}^{n_1} (\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)^{\mathsf{T}} + \sum_{j=1}^{n_2} (\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)^{\mathsf{T}}}{n_1 + n_2 - 2}$$

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

$$\text{Cov}\left(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2\right) = \text{Cov}\left(\boldsymbol{X}_1\right) + \text{Cov}\left(\boldsymbol{X}_2\right) = \frac{1}{n_1}\Sigma + \frac{1}{n_2}\Sigma = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma$$

Hence

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma$$

is an estimator of $\text{Cov}\left(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2\right)$.

If $\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1}$ is a random sample of size n_1 from $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $\boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_2}$ is an independent random sample size n_2 from $N_p(\boldsymbol{\mu}_2, \Sigma)$, then

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

is distributed as

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

Consequently,

$$P\left(T^2 \leq c^2\right) = 1 - \alpha$$

where

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

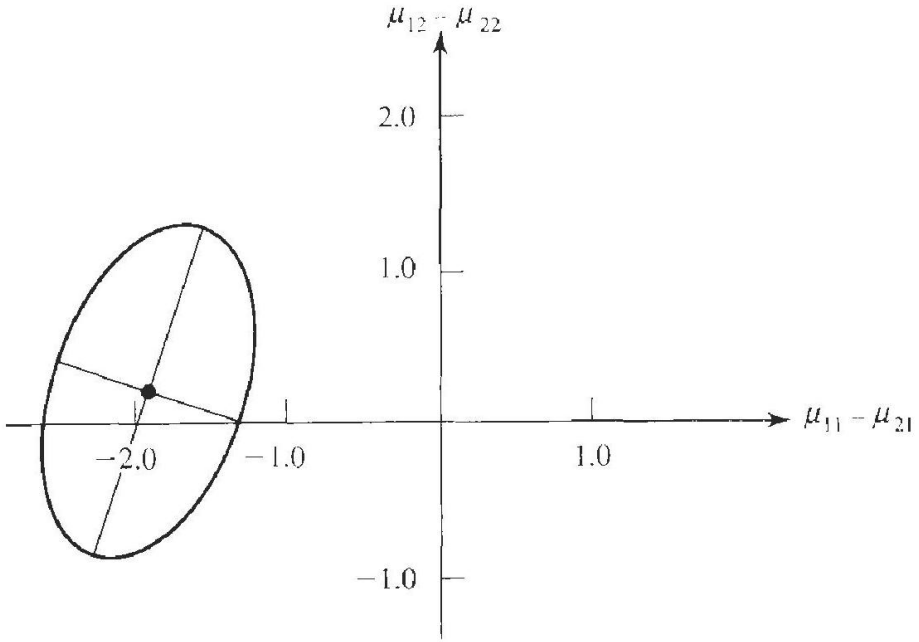
Example 4.8 (Constructing a confidence region for the difference of two mean vectors)

Fifty bars of soap are manufactured in each of two ways. Two characteristics, $X_1 =$ lather and $X_2 =$ mildness, are measured. The summary statistics for bars produced by method 1 and 2 are

$$\mathbf{x}_1 = \begin{bmatrix} 8.3 & 4.1 \end{bmatrix}^T, \mathbf{x}_2 = \begin{bmatrix} 10.2 & 3.9 \end{bmatrix}^T$$

$$\mathbf{S}_1 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

Obtain a 95% confidence region for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.



Simultaneous Confidence Intervals

Let $c^2 = [(n_1 + n_2 - 2)\, p / (n_1 + n_2 - p - 1)]\, F_{p, n_1+n_2-p-1}(\alpha)$. With probability $1 - \alpha$.

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

will cover $\mathbf{a}^{\mathsf{T}} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for all a. In particular $\mu_{1i} - \mu_{2i}$ will be covered by

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

The Two-Sample Situation When $\Sigma_1 \neq \Sigma_2$

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 $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is given by all $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ satisfying

$$T^2 = \left[\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]^{\mathsf{T}} \left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right)^{-1} \left[\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right] \leq \chi^2_p(\alpha)$$

where $\chi^2_p(\alpha)$ is the upper (100α) th percentile of a chi-square distribution with p d.f. Also $100(1-\alpha)\%$ simultaneous confidence interval for all linear combinations $\mathbf{a}^{\mathsf{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ are provided by

$$\mathbf{a}^{\mathsf{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

belongs to $\mathbf{a}^{\mathsf{T}}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) \pm \sqrt{\chi^2_p(\alpha)}\sqrt{\mathbf{a}^{\mathsf{T}}\left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right)\mathbf{a}}$

Testing for Equality of Covariance Matrics

With g populations, the null hypothesis is

$$H_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_g = \Sigma$$

where Σ_l is the covariance matrix for the l th population, $l = 1, 2, \ldots, g$, and Σ is the presumed common covariance matrix. The alternative hypothesis is that at least two of the covariance matrices are not equal.

- Assuming multivariate normal populations, a likelihood ratio statistic for testing above is given by

$$\Gamma = \prod_l \left(\frac{|\mathbf{S}_l|}{|\mathbf{S}_{\text{pooled}}|} \right)^{(n_l - 1)/2}$$

Here n_l is the sample size for the l th group, $\mathbf{S}_l$ is the l th group sample covariance matrix and $\mathbf{S}_{pool}$ is the pooled sample covariance matrix given by

$$\mathbf{S}_{\text{pool}} = \frac{1}{\sum_l (n_l - 1)} \{(n_1 - 1)\,\mathbf{S}_1 + (n_2 - 1)\,\mathbf{S}_2 + \cdots + (n_g - 1)\,\mathbf{S}_g\}$$

- Box's test is based on this χ^2 approximation to the sampling distribution of $-2 \ln \Gamma$. Setting $-2 \ln \Gamma = M$ (Box's M statistics) gives

$$M = \left[\sum_l (n_l - 1)\right] \ln |\mathbf{S}_{\text{pooled}}| - \sum_l [(n_l - 1) \ln |\mathbf{S}_l|]$$

Set

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

where p is the number of variables and g is the number of groups. Then $C = (1-u)M = (1-u)\left\{\left[\sum_l (n_l - 1)\right] \ln |\mathbf{S}_{\text{pooled}}| - \sum_l [(n_l - 1) \ln |\mathbf{S}_l|]\right\}$ has an approximate χ^2 distribution with

$$\nu = g\frac{1}{2}p(p+1) - \frac{1}{2}p(p+1) = \frac{1}{2}p(p+1)(g-1)$$

degrees of freedom. At significance level α, reject H_0 if $C > \chi^2_{p(p+1)(g-1)/2}(\alpha)$.

Example 4.9 (Testing equality of covariance matrices-nursing homes)

The Wisconsin Department of Health and Social Services reimburse nursing homes in the state for the services provided. The department develops a set of formulas for the rates for each facility, based on factors such as level of care, mean wage rate, and average wage rate in the state.

Nursing homes can be classified on the basis of ownership (private party, nonprofit organization, and government) and certification (skilled nursing facility, intermediate care facility, or combination of the two).

One purpose of a recent study was to investigate the effects of ownership or certification (or both) on cost.s. Four costs, computed on a per-patient-day basis and measured in hours per patient day, were selected for analysis $X_1 =$ cost of nursing labor, $X_2 =$ cost dietary labor, $X_3 =$ cost of plant operation and maintenance labor, and $X_4 =$ cost of housekeeping and laundry labor. A total of $n = 516$ observations on each of the $p = 4$ cost variables were initially separated according to ownership.

Summary statistics for each of the $g = 3$ groups are given in the following table.

Sample covariance matrices

$$\mathbf{S}_1 = \begin{bmatrix} .291 & & & \\ -.001 & .011 & & \\ .002 & .000 & .001 & \\ .010 & .003 & .000 & .010 \end{bmatrix}$$

$$\mathbf{S}_2 = \begin{bmatrix} .561 & & & \\ -.011 & .025 & & \\ .001 & .004 & .005 & \\ .037 & .007 & .002 & .019 \end{bmatrix}; \mathbf{S}_3 = \begin{bmatrix} .261 & & & \\ -.030 & .017 & & \\ .003 & -.000 & .004 & \\ .018 & .006 & .001 & .013 \end{bmatrix}$$

Assuming multivariate normal data, test hypothesis

$H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma.$

	Raw score	
Variable	Mean $(\bar{x}_i)$	Standard deviation $\left(\sqrt{s_{ii}}\right)$
$X_1 =$ melody	28.1	5.76
$X_2 =$ harmony	26.6	5.85
$X_3 =$ tempo	35.4	3.82
$X_4 =$ meter	34.2	5.12
$X_5 =$ phrasing	23.6	3.76
$X_6 =$ balance	22.0	3.93
$X_7 =$ style	22.7	4.03

Table 6.1 Effluent Data				
	Commercial lab		State lab of hygiene	
Sample j	x_{1j1}(BOD)	x_{1j2}(SS)	x_{2j1}(BOD)	x_{2j2} (SS)
	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

Control regions for future individual observations

Let $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i.i.d., $i = 1, \ldots, n$. Let $\mathbf{X}$ be a future observation from the same distribution. Then

$$T^2 = \frac{n}{n+1}(\mathbf{X} - \bar{\mathbf{X}})^{\mathsf{T}} \mathbf{S}^{-1} (\mathbf{X} - \bar{\mathbf{X}}) \sim \frac{(n-1)p}{n-p} F_{p,n-p}$$

and a $100(1-\alpha)\%$ p-dim prediction ellipsoid is given by all $\mathbf{x}$ satisfying

$$\frac{n}{n+1}(\mathbf{x} - \bar{\mathbf{X}})^{\mathsf{T}} \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{X}}) \leq \frac{(n^2-1)}{n(n-p)} F_{p,n-p}(\alpha)$$