Multivariate Statistics

Sudipta Das

Assistant Professor,
Department of Data Science,
Ramakrishna Mission Vivekananda University, Kolkata
Slides adapted from Jhonson & Winchern

Outline I

Paired Comparisons I

- To compare the means of two dependent samples.
  - The samples are dependent if one sample is related to the other
- When two samples are dependent, then each data point in one sample can be coupled in some natural, nonrandom fashion with each data point in the second sample.
- This situation occurs when each individual data point within a sample is paired (matched) to an individual data point in the second sample.

Paired Comparisons II

- The pairing may be the result of the individual observations in the two samples:
  1. representing before and after a program (such as weight before and after following a certain diet program),
  2. sharing the same characteristic,
  3. being matched by location,
  4. being matched by time,
  5. control and experimental, and so forth.

Paired Comparisons III

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

$$D_j = X_{1j} - X_{2j},\ j = 1(1)n$$

should reflect only the differential effects of the treatments.

Paired Comparisons IV

Testing for Matched Pairs Experiment *(Univariate Paired t-test)*.

Equality of Mean	1^{st} Mean differs from 2^{nd} Mean
Null Hypothesis	$H_0 : \mu_1 = \mu_2$ or $H_0 : \delta = \mu_1 - \mu_2 = 0$
Alternative Hypothesis	$H_a : \mu_1 \neq \mu_2$ or $H_a : \delta = \mu_1 - \mu_2 \neq 0$
Assumption	The sample differences (D_js) are normally distributed.
Test Statistic	$TS = \frac{\bar{D}-\delta}{S_d/\sqrt{n}}$, where $\bar{D} = \sum_{j=1}^{n} D_j$, $S_d^2 = \frac{1}{n-1}\sum_{j=1}^{n}(D_j - \bar{D})^2$
Distribution of Test Statistic	$TS \sim t_{n-1}$
Computed Test Statistic Under H_0	$ts = \frac{\bar{d}}{s_d/\sqrt{n}}$
Level of significance	α
Critical Value	l_{cv}=$(\alpha/2)^{th}$ quantile of t_{n-1} u_{cv}=$(1-\alpha/2)^{th}$ quantile of t_{n-1}
Rejection Region	$ts < l_{cv}$ or $ts > u_{cv}$
p-value	$p = 2P(TS > \lvert ts \rvert)$
Decision	As Usual

Paired Comparisons V

Multivariate Paired Comparisons

- Additional notation for the paired comparison procedure of n experimental units, with p responses under two treatments.
- We label the p responses within the jth unit as

$$X_{1jk} = \text{ variable } k \text{ under treatment 1, } k = 1, \ldots, p$$

and

$$X_{2jk} = \text{ variable } k \text{ under treatment 2, } l = 1, \ldots, p$$

and

$$D_{jk} = X_{1jk} - X_{2jk}$$

  - the p paired-difference random variables.

Paired Comparisons VI

- We define $\mathbf{D}_j = [D_{j1}, \ldots, D_{jp}]'$ and
- Assume that

$$E[\mathbf{D}_j] = \underline{\delta} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_p \end{bmatrix} \text{ and } Cov(\mathbf{D}_j) = \Sigma_d,$$

for $j = 1, \ldots, n$.

- Notations

$$\bar{\mathbf{D}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{D}_j \text{ and } S_d = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{D}_j - \bar{\mathbf{D}})(\mathbf{D}_j - \bar{\mathbf{D}})'$$

Paired Comparisons VII

- Results
  - Let the differences $D_1, D_2, \ldots, D_n$ be a independent random sample from an $N_p(\delta, \Sigma_d)$ population. Then

$$T^2 = n(\bar{\mathbf{D}} - \delta)' S_d^{-1} (\bar{\mathbf{D}} - \delta) \sim \frac{(n-1)p}{(n-p)} F_{p,n-p}.$$

  - If n and $n-p$ are both large, T^2 is approximately distributed as a χ^2_p random variable, regardless of the form of the underlying population of differences.

Paired Comparisons VIII

- Hypothesis & Testing
  Given the observed differences
  $\mathbf{d}_j' = [d_{j1}, d_{j2}, \ldots, d_{jp}],\ j = 1, 2, \ldots, n$, an $\alpha-$level test of

$$H_0 : \delta = 0 \text{ versus } H_a : \delta \neq 0$$

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

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

  where $\bar{\mathbf{d}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{d}_j$ and $s_d = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{d}_j - \bar{\mathbf{d}})(\mathbf{d}_j - \bar{\mathbf{d}})'$
- Example 6.1 (Page 276)

Paired Comparisons IX

- Confidence Region & Intervals
  - A $100(1-\alpha)\%$ confidence region for δ consists of all δs such that

$$(\bar{\mathbf{d}}-\delta)' s_d^{-1} (\bar{\mathbf{d}}-\delta) \leq \frac{(n-1)p}{n(n-p)} F_{p,n-p}(\alpha).$$

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

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

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

Paired Comparisons X

- For $n - p$ large,
$$\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha) = \chi^2_p(\alpha)$$
and normality need not be assumed.
- The Bonferroni $100(1-\alpha)\%$ simultaneous confidence intervals for the individual mean differences are
$$\delta_i : \bar{d}_i \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s^2_{d_i}}{n}}$$
where $t_{n-1}\left(\frac{\alpha}{2p}\right)$ is the upper $100(\frac{\alpha}{2p})$th percentile of a $t-$distribution with $n-1$ d.f.

- Example 6.1 (Page 277)
  - Discussion: The 95% simultaneous confidence intervals include zero, yet the hypothesis $H_0 : \delta = 0$ was rejected at the 5% level. What are we to conclude?

Repeated Measures Design for Comparing Treatments I

- It is another generalization of the univariate paired $t-$statistic arises in situations where q treatments are compared with respect to a single response variable.
- Each subject or experimental unit receives each treatment once over successive periods of time.

Repeated Measures Design for Comparing Treatments II

- The jth observation is

$$\mathbf{X}_j = [X_{j1}, \ldots, X_{jq}]', \; j = 1, \ldots, n$$

  where X_{ji} is the response to the ith treatment on the jth unit.
- The name repeated measures stems from the fact that all treatments are administered to each unit.
- Of course, the experimenter should randomize the order in which the treatments are presented to each subject.

Repeated Measures Design for Comparing Treatments III

- For comparative purposes, we consider contrasts of the components of $\mu = E[\mathbf{X}]$.
- Possible choices are

$$\begin{bmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_q \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} = C_1 \mu$$

or

$$\begin{bmatrix} \mu_2 - \mu_1 \\ \mu_3 - \mu_2 \\ \vdots \\ \mu_q - \mu_{q-1} \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{bmatrix} = C_2 \mu$$

- Both C_1 and C_2 are called contrast matrices,
  - because their $q - 1$ rows are linearly independent and each is a contrast vector.

Repeated Measures Design for Comparing Treatments IV

- In general, the hypothesis that there are no differences in treatments (equal treatment means) becomes

$$C\mu = 0,$$

  irrespective of $C = C_1$ or $C = C_2$.
- Consequently, based on the contrasts $C\mathbf{X}_j$ in the observations,
  - we have means $C\bar{\mathbf{X}}$ and
  - covariance matrix CS_XC', and
  - we test $C\mu = 0$ using the T^2-statistic

$$T^2 = n(C\bar{\mathbf{X}})'(CSC')^{-1}(C\bar{\mathbf{X}}).$$

Repeated Measures Design for Comparing Treatments V

- Hypothesis & Test for Equality of Treatments in a Repeated Measures Design.
  Consider an $N_q(\mu, \Sigma)$ population, and let C be a contrast matrix.
  An $\alpha-$level test of $H_0 : C\mu = 0$ (equal treatment means) versus $H_1 : C\mu \neq 0$ is as follows:
  Reject H_0 if

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

where $\bar{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j$ and $s_x = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'$

Repeated Measures Design for Comparing Treatments VI

- A confidence region for contrasts $C\mu$, with μ the mean of a normal population, is determined by the set of all $C\mu$ such that

$$n(C\bar{\mathbf{x}} - C\mu)'(Cs_xC')^{-1}(C\bar{\mathbf{x}} - C\mu) \leq \frac{(n-1)(q-1)}{(n-q+1)}F_{q-1,n-q+1}(\alpha).$$

- Consequently, simultaneous $100(1-\alpha)\%$ confidence intervals for single contrasts $c'\mu$ for any contrast vectors of interest are given by

$$c'\mu : c'\bar{\mathbf{x}} \pm \sqrt{\frac{(n-1)(q-1)}{(n-q+1)}F_{q-1,n-q+1}(\alpha)}\sqrt{\frac{c's_xc}{n}}.$$

- Example 6.2 (Page 281)

Comparing Mean Vectors from Two Populations I

- Consider a random sample of size n_1 from Population 1 and a sample of size n_2 from Population 2.
- The observations on p variables can be arranged as follows:

Sample	Summary Statistics	
(Population 1)		
$(x_{11}, \ldots, x_{1n_1})$	$\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{x}_{1j}$	$S_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j}-\bar{\mathbf{x}}_1)(\mathbf{x}_{1j}-\bar{\mathbf{x}}_1)'$
(Population 2)		
$(x_{21}, \ldots, x_{2n_2})$	$\bar{\mathbf{x}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{x}_{2j}$	$S_2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\mathbf{x}_{2j}-\bar{\mathbf{x}}_2)(\mathbf{x}_{2j}-\bar{\mathbf{x}}_2)'$

- We want to make inferences about

$$\text{(mean vector of population 1)} - \text{(mean vector of population 2)} = \mu_1 - \mu_2.$$

Comparing Mean Vectors from Two Populations II

- Assumptions Concerning the Structure of the Data
  1. The sample $X_{11}, X_{12}, \ldots, X_{1n_1}$, is a random sample of size n_1 from a $p-$variate population with mean vector μ_1 and covariance matrix Σ_1.
  2. The sample $X_{21}, X_{22}, \ldots, X_{2n_2}$, is a random sample of size n_2 from a $p-$variate population with mean vector μ_2 and covariance matrix Σ_2.
  3. Also, $X_{11}, X_{12}, \ldots, X_{1n_1}$, are independent of $X_{21}, X_{22}, \ldots, X_{2n_2}$.

Comparing Mean Vectors from Two Populations with Equal Variance I

- Further Assumptions When n_1 and n_2 are small.
  1. Both populations are multivariate normal.
  2. Also, $\Sigma_1 = \Sigma_2$ (same covariance matrix).
- Pooled sample covariance matrix

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

Comparing Mean Vectors from Two Populations with Equal Variance II

- Results
  1. $E(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = \mu_1 - \mu_2.$
  2. $Cov(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = Cov(\bar{\mathbf{X}}_1) + Cov(\bar{\mathbf{X}}_2) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma.$
  3. $(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \sim N_p\left(\mu_1 - \mu_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma\right)$

Comparing Mean Vectors from Two Populations with Equal Variance III

4. $(n_1 - 1)S_1$ is distributed as $W_{n_1-1}(\Sigma)$ and $(n_2 - 1)S_2$ is distributed as $W_{n_2-1}(\Sigma)$
5. $(n_1 - 1)S_1$ and and $(n_2 - 1)S_2$ are independent, thus $[(n_1 - 1)S_1 + (n_2 - 1)S_2]$ is distributed as $W_{n_1+n_2-2}(\Sigma)$
6. Statistic

$$
\begin{aligned}
T^2 &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\mu_1 - \mu_2))' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pooled} \right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\mu_1 - \mu_2)) \\
&= \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-\frac{1}{2}} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\mu_1 - \mu_2))' \left[\frac{(n_1 + n_2 - 2)S_{pooled}}{n_1 + n_2 - 2} \right]^{-1} \\
&\quad \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-\frac{1}{2}} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\mu_1 - \mu_2))
\end{aligned}
$$

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

- Example 6.3 (Page 287)

Comparing Mean Vectors from Two Populations with Equal Variance IV

- Simultaneous Confidence Intervals
  - With probability $1-\alpha$,

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

    will cover $\mathbf{a}'(\mu_1-\mu_2)$ for all $\mathbf{a}$, where $c^2 = \frac{(n_1+n_2-2)p}{(n_1+n_2-1-p)}F_{p,n_1+n_2-1-p}$.
  - In particular $\mu_1-\mu_2$ will be covered by

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

- Example 6.4 (Page 289)

Comparing Mean Vectors from Two Populations with Unequal Variance I

- Assumptions
  1. Sample sizes n_1 and n_2 are large.
  2. Also $n_1 - p$ and $n_2 - p$ are large.
  3. $\Sigma_1 \neq \Sigma_2$.

Comparing Mean Vectors from Two Populations with Unequal Variance II

- Results
  1. $E(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = \mu_1 - \mu_2$.
  2. $Cov(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) = Cov(\bar{\mathbf{X}}_1) + Cov(\bar{\mathbf{X}}_2) = \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2$.
  3. By the central limit theorem, $(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \sim N_p\left(\mu_1 - \mu_2, \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2\right)$
  4. Thus,
     $$(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\mu_1 - \mu_2))' \left[\frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2\right]^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\mu_1 - \mu_2)) \sim \chi^2_p$$
  5. Approximately,
     $$T^2 = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\mu_1 - \mu_2))' \left[\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right]^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\mu_1 - \mu_2)) \sim \chi^2_p$$

Comparing Mean Vectors from Two Populations with Unequal Variance III

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

$$(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\mu_1 - \mu_2))' \left[\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right]^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2 - (\mu_1 - \mu_2)) \leq \chi^2_p(\alpha)$$

7. Also, $100(1-\alpha)\%$ simultaneous confidence intervals for all linear combinations $\mathbf{a}'(\mu_1 - \mu_2)$ are provided by

$$\mathbf{a}'(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \pm \sqrt{\chi^2_p(\alpha)}\sqrt{\mathbf{a}'\left[\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2\right]\mathbf{a}}$$

- Example 6.5 (Page 293)

Comparing Several Multivariate Population Means (One-Way MANOVA)

- Assumptions about the Structure of the Data for One-Way MANOVA
  1. $X_{l1}, X_{l2}, \ldots, X_{ln_l}$, is a random sample of size n_l from the lth population with mean μ_l, $l = 1, 2, \ldots, g$.
  2. The random samples from different populations are independent.
  3. All populations have a common covariance matrix.
  4. Each population is multivariate normal.
     - This condition can be relaxed by appealing to the central limit theorem when the sample sizes n_l are large.

Univariate One-Way ANOVA I

- Null hypothesis

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

- Alternate hypothesis

$$H_A : \text{ at least one of the } \mu_l s \text{ is different.}$$

Univariate One-Way ANOVA II

- Reparameterization of the null and alternate hypothesis
  - Let

$$\underbrace{\mu_l}_{l\text{th population mean}} = \underbrace{\mu}_{\text{overall mean}} + \underbrace{\tau_l}_{l\text{th population (treatment effect)}},$$

    where $\tau_l = \mu_l - \mu$.
  - Null hypothesis

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_g = 0.$$

  - Alternate hypothesis

$$H_A : \text{ at least one of the } \tau_l s \text{ is not equal to zero.}$$

Univariate One-Way ANOVA III

- ANOVA Model For Comparing g Population Mean Vectors
  - Under the assumption of normality, the jth response from lth group X_{lj}, can be represented as

$$X_{lj} = \underbrace{\mu}_{\text{overall mean}} + \underbrace{\tau_l}_{\text{treatment effect}} + \underbrace{e_{lj}}_{\text{random error}},$$

    where $e_{lj} \sim N(0, \sigma^2)$.
  - To define uniquely the model parameters and their least square estimates, it is customary to impose the constraint $\sum_{l=1}^{g} n_l \tau_l = 0$.

Univariate One-Way ANOVA IV

- Decomposition of observations

$$x_{lj} = \underbrace{\bar{x}}_{\text{overall sample mean}} + \underbrace{(\bar{x}_l - \bar{x})}_{\text{estimated treatment effect}} + \underbrace{(x_{lj} - \bar{x}_l)}_{\text{residual}},$$

where

- $\bar{x}$ is an estimate of μ,
- $\hat{\tau}_l = (\bar{x}_l - \bar{x})$ is an estimate of τ_l, and
- $(x_{lj} - \bar{x}_l)$ is an estimate of the error e_{lj}.

Univariate One-Way ANOVA V

- Some algebra
  1. Subtracting $\bar{x}$ from both sides and squaring gives

$$(x_{lj} - \bar{x})^2 = (\bar{x}_l - \bar{x})^2 + (x_{lj} - \bar{x}_l)^2 + 2(\bar{x}_l - \bar{x})(x_{lj} - \bar{x}_l).$$

  2. Summing both sides over j gives

$$\sum_{j=1}^{n_l} (x_{lj} - \bar{x})^2 = n_l (\bar{x}_l - \bar{x})^2 + \sum_{j=1}^{n_l} (x_{lj} - \bar{x}_l)^2.$$

Univariate One-Way ANOVA VI

3. Summing both sides over l we get

$$\underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj}-\bar{x})^2}_{SS_{cor}} = \underbrace{\sum_{l=1}^{g} n_l(\bar{x}_l-\bar{x})^2}_{SS_{tr}} + \underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj}-\bar{x}_l)^2}_{SS_{res}}.$$

- SS_{cor} : Total (corrected) sum of squares
- SS_{tr} : Between (samples) sum of squares
- SS_{res} : Within (samples) sum of squares

4. Equivalently,

$$\underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}x_{lj}^2}_{SS_{obs}} = \underbrace{(n_1+\cdots+n_g)\,\bar{x}^2}_{SS_{mean}} + \underbrace{\sum_{l=1}^{g} n_l(\bar{x}_l-\bar{x})^2}_{SS_{tr}} + \underbrace{\sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj}-\bar{x}_l)^2}_{SS_{res}}.$$

- SS_{obs} : Total sum of squares

Univariate One-Way ANOVA VII

- Some observations
  - Data vector

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

    lies anywhere in $n = n_1 + \cdots + n_g$ dimensional plane.
  - Mean vector

$$\bar{x}\mathbf{1} = [\bar{x}, \ldots, \bar{x}]'$$

    must lie in the equianular line of $\mathbf{1}_{n\times 1}$

Univariate One-Way ANOVA VIII

- Treatment effect vector

$$\left[\underbrace{\bar{x}_1 - \bar{x}, \ldots, \bar{x}_1 - \bar{x}}_{n_1}, \underbrace{\bar{x}_2 - \bar{x}, \ldots, \bar{x}_2 - \bar{x}}_{n_2}, \ldots, \underbrace{\bar{x}_g - \bar{x}, \ldots, \bar{x}_g - \bar{x}}_{n_g}\right]'$$

$$= (\bar{x}_1 - \bar{x})\mathbf{u_1} + (\bar{x}_2 - \bar{x})\mathbf{u_2} + \cdots + (\bar{x}_g - \bar{x})\mathbf{u_g}$$

lies in the hyperplane of linear combinations of the g vectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_g}$, where

- $\mathbf{u_1} = [\underbrace{1, \ldots, 1}_{n_1}, 0, \ldots, 0]'$
- $\mathbf{u_2} = [0, \ldots, 0, \underbrace{1, \ldots, 1}_{n_2}, 0, \ldots, 0]'$
- $\vdots$
- $\mathbf{u_g} = [0, \ldots, 0, \underbrace{1, \ldots, 1}_{n_g}]'$,

Univariate One-Way ANOVA IX

- The mean vector $\bar{x}\mathbf{1}$, also lies in this hyperplane, since $\mathbf{1} = \mathbf{u_1} + \mathbf{u_2} + \cdots + \mathbf{u_g}$.
- In addition, the mean vector $\bar{x}\mathbf{1}$, is always perpendicular to the treatment vector $(\bar{x}_1 - \bar{x})\mathbf{u_1} + (\bar{x}_2 - \bar{x})\mathbf{u_2} + \cdots + (\bar{x}_g - \bar{x})\mathbf{u_g}$.
- Thus,
  - the mean vector has the freedom to lie anywhere along the one-dimensional equiangular line, and
  - the treatment vector has the freedom to lie anywhere in the other $g - 1$ dimension.

Univariate One-Way ANOVA X

- The residual vector,

$$\hat{\mathbf{e}} = \mathbf{y} - \bar{x}\mathbf{1} - (\bar{x}_1 - \bar{x})\mathbf{u_1} - (\bar{x}_2 - \bar{x})\mathbf{u_2} - \cdots - (\bar{x}_g - \bar{x})\mathbf{u_g}$$

  is perpendicular to both the mean vector and the treatment effect vector.
- Hence, the residual vector has the freedom to lie anywhere in the subspace of dimension

$$n - (g - 1) - 1 = n - g$$

  that is perpendicular to the hyperplane of the g vectors $\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_g}$.

Univariate One-Way ANOVA XI

- To summarize we compute the following ANOVA table showing the calculations of the sums of squares and the associated degrees of freedom.

Source of variation	Sum of squares (SS)	Degrees of freedom (d.f.)
Treatments	$SS_{tr} = \sum_{l=1}^{g} n_l(\bar{x}_l - \bar{x})^2$	$g-1$
Residuals (error)	$SS_{res} = \sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj} - \bar{x}_l)^2$	$n-g$
Total	$SS_{cor} = \sum_{l=1}^{g}\sum_{j=1}^{n_l}(x_{lj} - \bar{x})^2$	$n-1$

Univariate One-Way ANOVA XII

- Test statistics under $H_0 : \tau_1 = \cdots = \tau_I$,

$$TS = \frac{SS_{tr}/(g-1)}{SS_{res}/(n-g)} \sim F_{g-1,n-g}.$$

- At α level of significance the rejection region is

$$TS \geq F_{g-1,n-g}(\alpha).$$

- Equivalently a small value of

$$\frac{SS_{res}}{SS_{res} + SS_{tr}}$$

will reject the null hypothesis.

- Example 6.8 (Page 301)

Multivariate One-Way ANOVA I

- MANOVA Model For Comparing g Population Mean Vectors

$$\mathbf{X}_{lj} = \mu + \tau_l + \mathbf{e}_{lj},\ l = 1, \ldots, g \text{ and } j = 1, \ldots, n_l$$

where

- $\mathbf{e}_{lj}$ are independent $N_p(0, \Sigma)$,
- parameter vector μ is an overall mean (level) and
- τ_l represents the lth treatment effect with $\sum_{l=1}^{g} n_l \tau_l = \mathbf{0}$.

Multivariate One-Way ANOVA II

- Decomposition of vector observations

$$\mathbf{x}_{lj} = \underbrace{\bar{\mathbf{x}}}_{\text{overall sample mean}} + \underbrace{(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})}_{\text{estimated treatment effect}} + \underbrace{(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)}_{\text{residual}},$$

where

- $\bar{\mathbf{x}}$ is an estimate of μ,
- $\hat{\tau}_l = (\bar{\mathbf{x}}_l - \bar{\mathbf{x}})$ is an estimate of τ_l, and
- $(\mathbf{x}_{lj} - \bar{\mathbf{x}}_l)$ is an estimate of the error $\mathbf{e}_{lj}$.

Multivariate One-Way ANOVA III

- Some algebra
  - Subtracting overall sample mean and multiply by its transpose

$$\begin{aligned}(\mathbf{x}_{lj}-\bar{\mathbf{x}})(\mathbf{x}_{lj}-\bar{\mathbf{x}})' &= [(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)+(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})][(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)+(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})]' \\ &= (\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)'+(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})' \\ &\quad +(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})'+(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)'\end{aligned}$$

  - Summing both sides over j gives

$$\sum_{j=1}^{n_l}(\mathbf{x}_{lj}-\bar{\mathbf{x}})(\mathbf{x}_{lj}-\bar{\mathbf{x}})' = \sum_{j=1}^{n_l}(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)' + n_l(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})'$$

  - Summing both sides over l gives

$$\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{lj}-\bar{\mathbf{x}})(\mathbf{x}_{lj}-\bar{\mathbf{x}})' = \sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)' + \sum_{l=1}^{g} n_l(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})'$$

Multivariate One-Way ANOVA IV

- Therefore,

$$\begin{bmatrix}\text{total(corrected) sum}\\ \text{of squares \& cross product}\end{bmatrix} = \sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{lj}-\bar{\mathbf{x}})(\mathbf{x}_{lj}-\bar{\mathbf{x}})'$$
$$= \sum_{l=1}^{g} n_l(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})(\bar{\mathbf{x}}_l-\bar{\mathbf{x}})' + \sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)(\mathbf{x}_{lj}-\bar{\mathbf{x}}_l)'$$
$$= \begin{bmatrix}\text{treatment(Between) sum of}\\ \text{squares \& cross product}\end{bmatrix} + \begin{bmatrix}\text{residual(Within) sum of}\\ \text{squares \& cross product}\end{bmatrix}$$
$$= \mathbf{B}+\mathbf{W}$$

  - Note that: $\mathbf{W} = (n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2 + \cdots + (n_g-1)\mathbf{S}_g$

Multivariate One-Way ANOVA V

- Thus the MANOVA Table for Comparing Population Mean Vectors

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

Multivariate One-Way ANOVA VI

- Under the null hypothesis $H_0 : \tau_1 = \tau_2 = \cdots = \tau_g = 0$, we construct the test statistics as the ratio of generalized variances

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

- We reject H_0 if this ratio is too small.

Multivariate One-Way ANOVA VII

- The exact distribution of Λ^* can be derived for the special cases listed in the following table.

Number of variables	Number of groups	Sampling distribution for multivariate normal data
$p = 1$	$g \geq 2$	$\left(\frac{n-g}{g-1}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{g-1,n-g}$
$p = 2$	$g \geq 2$	$\left(\frac{n-g-1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1),2(n-g-1)}$
$p \geq 1$	$g = 2$	$\left(\frac{n-p-1}{p}\right)\left(\frac{1-\Lambda^*}{\Lambda^*}\right) \sim F_{p,n-p-1}$
$p \geq 1$	$g = 3$	$\left(\frac{n-p-2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p,2(n-p-2)}$

- Example 6.9 (Page 304)

Multivariate One-Way ANOVA VIII

- Bartlett has shown that if H_0 is true and $\sum_{l=1}^{g} n_l = n$ is large,

$$-\left(n-1-\frac{p+g}{2}\right)\ln\Lambda^* \sim \chi^2_{p(g-1)}$$

- Consequently, for large n, we reject H_0 at significance level α 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 6.10 (Page 306)

Simultaneous Confidence Intervals for Treatment Effects I

- When the hypothesis of equal treatment effects is rejected, those effects that led to the rejection of the hypothesis are of interest.
- For pairwise comparisons, the Bonferroni approach can be used to construct simultaneous confidence intervals for the components of the differences $\tau_k - \tau_l$ (or $\mu_k - \mu_l$).
  - These intervals are shorter than those obtained for all contrasts, and they require critical values only for the univariate t-statistic.

Simultaneous Confidence Intervals for Treatment Effects II

- Note that:
  - $\tau_k = [\tau_{k1}, \ldots, \tau_{ki}, \ldots, \tau_{kp}]'$,
  - τ_k is estimated by $\hat{\tau}_k = \bar{\mathbf{x}}_k - \bar{\mathbf{x}}$, and
  - $\hat{\tau}_{ki} - \hat{\tau}_{li} = \bar{x}_{ki} - \bar{x}_{li}$.
- Therefore,

$$Var(\hat{\tau}_{ki} - \hat{\tau}_{li}) = Var(\bar{x}_{ki} - \bar{x}_{li}) = \left(\frac{1}{n_k} + \frac{1}{n_l}\right)\sigma_{ii},$$

where σ_{ii} is the ith diagonal element of Σ.

Simultaneous Confidence Intervals for Treatment Effects III

- The estimated variance is

$$\hat{Var}(\bar{x}_{ki} - \bar{x}_{li}) = \left(\frac{1}{n_k} + \frac{1}{n_l}\right)\frac{w_{ii}}{n-g},$$

  where w_{ii} is the ith diagonal element of $\mathbf{W}$, the pooled sample variance.
- Apportioning the overall error rate (α), over the numerous Bonferroni confidence statements.
  - There are p variables and $g(g-1)/2$ pairwise differences, so each two-sample $t-$interval will employ the critical value

$$t_{n-g}\left(\frac{\alpha}{2m}\right),$$

    where $m = pg(g-1)/2$.

Simultaneous Confidence Intervals for Treatment Effects IV

- Thus for the MANOVA model with confidence at least $(1-\alpha)$,

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

  for all component $i = 1, \ldots, p$ and all differences $l < k = 1, \ldots, g$.

  - Here, $n = \sum_{k=1}^{g} n_k$ and w_{ii} is the ith diagonal element of $\mathbf{W}$.
- Example 6.11 (Page 309)

Testing for Equality of Covariance Matrices I

- One of the assumptions made when comparing two or more multivariate mean vectors is that the covariance matrices of the potentially different populations are the same.
- Before pooling the variation across samples to form a pooled covariance matrix when comparing mean vectors, it can be worthwhile to test the equality of the population covariance matrices.
- One commonly employed test for equal covariance matrices is Box's $M-$test, discussed subsequently.

Testing for Equality of Covariance Matrices II

- Null hypothesis

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

where Σ_l is the covariance matrix for the lth population, $l = 1, 2, \ldots, g$, and Σ is the presumed common covariance matrix.

- Alternate hypothesis

H_A : at least two of the covariance matrices are not equal.

Testing for Equality of Covariance Matrices III

- Assumption: Populations are multivariate normal
- Test statistics under H_0 is

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

where

- n_l is the sample size for the lth group,
- $\mathbf{S}_l$ is the lth group sample covariance matrix and
- $\mathbf{S}_{pooled}$ is the pooled sample covariance matrix given by

$$\mathbf{S}_{pooled} = \frac{1}{\sum_{l=1}^{g}(n_l - 1)} \left[(n_1 - 1)\mathbf{S}_1 + \cdots + (n_g - 1)\mathbf{S}_g \right].$$

Testing for Equality of Covariance Matrices IV

- Distribution of the test statistics under H_0 :

$$\begin{aligned} C &= (1-u)M = (1-u)(-2\ln\Lambda) \\ &= (1-u)\left[\left(\sum_{l=1}^{g}(n_l-1)\right)\ln|\mathbf{S}_{pooled}| - \left(\sum_{l=1}^{g}(n_l-1)\ln|\mathbf{S}_l|\right)\right] \\ &\sim \chi^2_\nu, \end{aligned}$$

where

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

Testing for Equality of Covariance Matrices V

- At level of significance α, the rejection region is

$$C > \chi^2_{p(p+1)(g-1)/2}(\alpha).$$

- Box's χ^2 approximation works well if each n_l exceeds 20 and if p and g do not exceed 5.
- In situations where these conditions do not hold, Box has provided a more precise F approximation to the sampling distribution of M.
- Example 6.12 (Page 311)

Testing for Equality of Covariance Matrices VI

- Box's $M-$test is routinely calculated in many statistical computer packages that do MANOVA and other procedures requiring equal covariance matrices.
- It is known that the $M-$test is sensitive to some forms of non-normality.
  - More broadly, in the presence of non-normality, normal theory tests on covariances are influenced by the kurtosis of the parent populations.
- However, with reasonably large samples, the MANOVA tests of means or treatment effects are rather robust to nonnormality.

Testing for Equality of Covariance Matrices VII

- Thus the $M-$test may reject H_0 in some non-normal cases where it is not damaging to the MANOVA tests.
- Moreover, with equal sample sizes, some differences in covariance matrices have little effect on the MANOVA tests.
- To summarize, we may decide to continue with the usual MANOVA tests even though the $M-$test leads to rejection of H_0.