

Comparisons of Several Multivariate Means

Paired Comparisons:

There are two scenarios in which paired comparison should be applied. 1). Same experimental units are measured before and after treatment. 2). Two different treatments are randomly assigned to each pair of identical or similar experimental units.

Original Data:

| | treatment 1 | treatment 2 |
|----------|------------------------------------|------------------------------------|
| units | x_1, x_2, \dots, x_p | x_1, x_2, \dots, x_p |
| 1 | $x_{111}, x_{112}, \dots, x_{11p}$ | $x_{211}, x_{212}, \dots, x_{21p}$ |
| 2 | $x_{121}, x_{122}, \dots, x_{12p}$ | $x_{221}, x_{222}, \dots, x_{22p}$ |
| \vdots | \vdots | \vdots |
| n | $x_{1n1}, x_{1n2}, \dots, x_{1np}$ | $x_{2n1}, x_{2n2}, \dots, x_{2np}$ |

Working Data:

| units | 1 | 2 | \dots | p |
|----------|------------------------------|------------------------------|----------|------------------------------|
| 1 | $D_{11} = x_{111} - x_{211}$ | $D_{12} = x_{112} - x_{212}$ | \dots | $D_{1p} = x_{11p} - x_{21p}$ |
| 2 | $D_{21} = x_{121} - x_{221}$ | $D_{22} = x_{122} - x_{222}$ | \dots | $D_{2p} = x_{12p} - x_{22p}$ |
| \vdots | \vdots | \vdots | \vdots | \vdots |
| n | $D_{n1} = x_{1n1} - x_{2n1}$ | $D_{n2} = x_{1n2} - x_{2n2}$ | \dots | $D_{np} = x_{1np} - x_{2np}$ |

Let

$$\vec{D}_i = (D_{i1}, D_{i2}, \dots, D_{ip})$$

Assuptions:

1.

$$E(\vec{D}_i) = \delta = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_p \end{pmatrix}, \text{Cov}(\vec{D}_i) = \Sigma_D$$

2. $\vec{D}_1, \vec{D}_2, \dots, \vec{D}_n$ is a random sample from a multivariate normal distribution.

Inferences about δ :

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n \vec{D}_i$$

$$S_D = \frac{1}{n-1} \sum_{i=1}^n (\vec{D}_i - \bar{D})(\vec{D}_i - \bar{D})'$$

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

Hypothesis testing:

$$H_0 : \delta = \delta_0 \text{ v.s } H_1 : \delta \neq \delta_0$$

$$\text{Reject } H_0, \text{ if } T^2 = n(\bar{d} - \delta_0)'(S_d)^{-1}(\bar{d} - \delta_0) > \frac{(n-1)p}{n-p} F_{p, n-p}(\alpha)$$

100(1 - α)% confidence region:

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

Simultaneous confidence intervals for $\delta_1, \delta_2, \dots, \delta_p$

$$\delta_i: \bar{d}_i \pm \sqrt{\frac{(n-1)p}{n-p} F_{p, n-p}(\alpha)} \sqrt{\frac{(S_d)_{ii}}{n}}$$

We can also construct simultaneous confidence intervals for $a'\delta$. And the Bonferroni 100(1 - α)% simultaneous confidence intervals for $\delta_1, \delta_2, \dots, \delta_p$ are:

$$\delta_i: \bar{d}_i \pm t_{n-1}(\frac{\alpha}{2p}) \sqrt{\frac{(S_d)_{ii}}{n}}$$

Repeated measures

Design: Same experimental units with more than two treatments (measurements).

Data:

| units | treatment 1 | treatment 2 | ... | treatment q |
|----------|-------------|-------------|-----|---------------|
| 1 | X_{11} | X_{12} | ... | X_{1q} |
| 2 | X_{21} | X_{22} | ... | X_{2q} |
| \vdots | \vdots | \vdots | | \vdots |
| n | X_{n1} | X_{n2} | ... | X_{nq} |

Let

$$\vec{X}' = (X_1, X_2, \dots, X_q)$$

$$E(\vec{X}') = (\mu_1, \mu_2, \dots, \mu_q)$$

Various types of Differences:

1.

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

2.

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

C_1 and C_2 are called contrast matrices. Let C be any contrast matrix.

Inference about $C\mu$:

$$\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j, S = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})(X_j - \bar{X})'$$

Hypothesis testing:

$$H_0 : C\mu = 0 \text{ v.s. } H_1 : C\mu \neq 0$$

Reject H_0 , if $T^2 = n(C\bar{X} - 0)(CSC')^{-1}(C\bar{X} - 0) > \frac{(n-1)(q-1)}{n-q+1} F_{q-1, n-q+1}(\alpha)$

Confidence region:

$$\{C\mu : n(C\bar{X} - C\mu)'(CSC')^{-1}(C\bar{X} - C\mu) \leq \frac{(n-1)(q-1)}{n-q+1} F_{q-1, n-q+1}(\alpha)\}$$

Simultaneous confidence intervals:

$$\vec{c}'\mu: \vec{c}'\bar{X} \pm \sqrt{\frac{(n-1)(q-1)}{n-q+1} F_{q-1, n-q+1}(\alpha)} \sqrt{\frac{\vec{c}'S\vec{c}}{n}}$$

Comparing Two Independent Samples

Assumptions:

1. $\vec{X}_{11}, \dots, \vec{X}_{1n_1}$: random sample from p variate population with mean μ_1 and variance matrix Σ_1 .
2. $\vec{X}_{21}, \dots, \vec{X}_{2n_2}$: random sample from p variate population with mean μ_2 and variance matrix Σ_2 .
3. These two samples are independent of each other.

Further assumptions:

1. Both populations are multivariate normal.
2. $\Sigma_1 = \Sigma_2$

Summary statistics: $\bar{X}_1, S_1; \bar{X}_2, S_2$ Pooled variance matrix:

$$S_{\text{pool}} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

Estimation:

estimate for $\mu_1 - \mu_2$: $\bar{X}_1 - \bar{X}_2$

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

estimate for $\text{Cov}(\bar{X}_1 - \bar{X}_2)$: $\left(\frac{1}{n_1} + \frac{1}{n_2}\right)S_{\text{pool}}$

Hypothesis testing:

$$H_0 : \mu_1 - \mu_2 = \delta_0 \text{ v.s. } H_1 : \mu_1 - \mu_2 \neq \delta_0$$

Reject H_0 if

$$T^2 = (\bar{x}_1 - \bar{x}_2 - \delta_0)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{\text{pool}} \right]^{-1} (\bar{x}_1 - \bar{x}_2 - \delta_0) > c^2$$

where

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

Simultaneous confidence intervals for $a'(\mu_1 - \mu_2)$:

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

Bonferroni confidence intervals for $a'_1(\mu_1 - \mu_2), \dots, a'_m(\mu_1 - \mu_2)$

$$a'_i(\bar{X}_1 - \bar{X}_2) \pm t_{n_1+n_2-2}\left(\frac{\alpha}{2m}\right) \sqrt{a'_i\left(\frac{1}{n_1} + \frac{1}{n_2}\right)S_{pool}a_i}$$

when $\Sigma_1 \neq \Sigma_2$:

Let $n_1 - p$ and $n_2 - p$ be large. Then an approximate $100(1 - \alpha)\%$ confidence ellipsoid for $\mu_1 - \mu_2$ is given by all $\mu_1 - \mu_2$ satisfying

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

Simultaneous confidence interval for $a'(\mu_1 - \mu_2)$

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

One-Way MANOVA

Data and Assumptions:

$$\begin{array}{ll} \text{Population 1:} & \vec{X}_{11}, \vec{X}_{12}, \dots, \vec{X}_{1n_1} \\ \text{Population 2:} & \vec{X}_{21}, \vec{X}_{22}, \dots, \vec{X}_{2n_2} \\ & \vdots \\ \text{Population } g: & \vec{X}_{g1}, \vec{X}_{g2}, \dots, \vec{X}_{gn_g} \end{array}$$

1. $\vec{X}_{11}, \vec{X}_{12}, \dots, \vec{X}_{1n_1}$: random sample \sim population with mean μ_l and covariance matrix Σ
2. Samples are independent.
3. All populations have a common covariance matrix Σ .
4. Each population is multivariate normal

Univariate case:

Model:

$$X_{lj} = \mu + \tau_l + e_{lj}$$

with the constraint: $\sum_{l=1}^g n_l \tau_l = 0$

Hypotheses:

$H_0 : \tau_1 = \tau_2 = \dots = \tau_g = 0$ v.s. $H_1 : \text{not all of them equal to } 0$

Estimation equation:

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

Variation decomposition:

$$\sum_{l=1}^g \sum_{j=1}^{n_l} (x_{lj} - \bar{x})^2 \underset{(\text{SS}_{\text{tot}})}{=} \sum_{l=1}^g n_l (\bar{x}_l - \bar{x})^2 \underset{(\text{SS}_{\text{tr}})}{=} + \sum_{l=1}^g \sum_{j=1}^{n_l} (x_{lj} - \bar{x}_l) \underset{(\text{SS}_{\text{res}})}{=}$$

ANOVA Table:

| Source of Variation | Sum of Squares | Degree of Freedom |
|---------------------|--------------------------|------------------------|
| Treatment | SS_{tr} | $g - 1$ |
| Residual | SS_{res} | $\sum_{l=1}^g n_l - g$ |
| Total | SS_{tot} | $\sum_{l=1}^g n_l - 1$ |

F -test for H_0 :

Reject H_0 if

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

Multivariate Case (MANOVA):

Model:

$$\vec{X}_{lj} = \mu + \tau_l + e_{lj}$$

with the constraint: $\sum_{l=1}^g n_l \tau_l = 0$.

Estimation equation:

$$\vec{x}_{lj} = \bar{x} + (\bar{x}_l - \bar{x}) + (x_{lj} - \bar{x}_l)$$

Variance matrix decomposition:

$$\begin{aligned} \sum_{l=1}^g \sum_{j=1}^{n_l} (x_{lj} - \bar{x})(x_{lj} - \bar{x})' &= \sum_{l=1}^g n_l (\bar{x}_l - \bar{x})(\bar{x}_l - \bar{x})' + \sum_{l=1}^g \sum_{j=1}^{n_l} (x_{lj} - \bar{x}_l)(x_{lj} - \bar{x}_l)' \\ &= \text{B}_{\text{SSP}} + \text{W}_{\text{SSP}} \end{aligned}$$

Hypothesis:

$H_0: \tau_1 = \tau_2 = \dots = \tau_g = 0$ v.s. $H_1 : \text{Not all of them are equal to } 0$'s

MANOVA Table:

| Source of Variation | Sum of Squares | Degree of Freedom |
|---------------------|----------------|------------------------|
| Treatment | BSSP | $g - 1$ |
| Residual | WSSP | $\sum_{l=1}^g n_l - g$ |
| Total | B + W | $\sum_{l=1}^g n_l - 1$ |

Let $\lambda_1, \dots, \lambda_s$ be the eigenvalues of $W^{-1}B$, and we have the following test statistics for one-way MANOVA:

1. Wilks' Lambda:

$$\Lambda^* = \prod_{l=1}^s \frac{1}{1 + \lambda_l}$$

2. Pillai-Bartlett trace:

$$V = \sum_{l=1}^s \frac{\lambda_l}{1 + \lambda_l}$$

3. Roy's greatest root:

$$GCR = \frac{\lambda_1}{1 + \lambda_1}$$

4. Hotelling-Lawley trace:

$$T = \sum_{l=1}^s \lambda_l$$

where $s = \min(g - 1, p)$. Note that: 1). When $s = 1$, all these tests are equivalent, i.e., they give the same p-value. 2). Wilk's Lambda test is based on the likelihood ratio test. 3). The null distributions of these test statistics are complicated but programmed into most good statistical packages such as splus and sas. 4). No one test dominates the other. Which is the best depends on the nature of the differences expected.

Pairwise comparison (Bonferroni approach):

We need to construct simultaneous confidence intervals for $\tau_{ki} - \tau_{li}$ where $1 \leq k, l \leq g$, $1 \leq i \leq p$ and $l \neq k$. The total number of pairs is $m = pg(g - 1)/2$. Based on the estimation equation, the estimats for τ_{ki} and τ_{li} are

$$\hat{\tau}_{ki} = (\bar{x}_k)_i - (\bar{x})_i \text{ and } \hat{\tau}_{li} = (\bar{x}_l)_i - (\bar{x})_i$$

So the estimate for $\tau_{ki} - \tau_{li}$ is

$$\tau_{ki} - \tau_{li} = (\bar{x}_k)_i - (\bar{x}_l)_i$$

And the variance of the above estimate is

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

Recall Σ is the population covariance matrix. The estimate for Σ is

$$\hat{\Sigma} = \frac{W}{n - g}$$

Hence $\hat{\sigma}_{ii} = \frac{w_{ii}}{n-g}$ where w_{ii} is the $i \times i$ diagonal entry of W . We have:

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

where $n = \sum_{l=1}^g n_l$. The Bonferroni confidence interval for $\tau_{ki} - \tau_{li}$ is

$$\bar{x}_{ki} - \bar{x}_{li} \pm t_{n-g}\left(\frac{\alpha}{2m}\right)\sqrt{\frac{w_{ii}}{n - g}\left(\frac{1}{n_k} + \frac{1}{n_l}\right)}$$

Two-Way MANOVA

Two factors:

Factor 1: g levels; Factor 2: b levels

Model:

$$\vec{X}_{lkr} = \mu + \tau_l + \beta_k + \gamma_{lk} + e_{lkr}$$

$$l = 1, 2, \dots, g; k = 1, 2, \dots, b; r = 1, 2, \dots, n$$

with constraints:

$$\sum \tau_l = \sum \beta_k = \sum_l \gamma_{lk} = \sum_k \gamma_{lk} = 0$$

$$e_{lkr} \sim N_p(0, \Sigma)$$

Estimation equation:

$$x_{lkr} = \bar{x} + (\bar{x}_{l.} - \bar{x}) + (\bar{x}_{.K} + (\bar{x}_{lk} - \bar{x}_{l.} - \bar{x}_{.k} + \bar{x})) + (x_{lkr} - \bar{x}_{lk})$$

Sample covariance matrix decomposition (SSP decomposition)

$$\begin{aligned} ((\text{SSP}_{\text{tot}})) &= \sum_l \sum_k \sum_r (x_{lkr} - \bar{x})(x_{lkr} - \bar{x})' && gbn - 1 \\ (\text{SSP}_{\text{fac.1}}) &= \sum_l bn(\bar{x}_{l.} - \bar{x})(\bar{x}_{l.} - \bar{x})' && g - 1 \\ (\text{SSP}_{\text{fac.2}}) &= \sum_l gn(\bar{x}_{.k} - \bar{x})(\bar{x}_{.k} - \bar{x})' && b - 1 \\ (\text{SSP}_{\text{int}}) &= \sum_l \sum_l n(\bar{x}_{lk} - \bar{x}_{l.} - \bar{x}_{.k} + \bar{x})(\bar{x}_{lk} - \bar{x}_{l.} - \bar{x}_{.k} + \bar{x})' && (g - 1)(b - 1) \\ (\text{SSP}_{\text{resi}}) &= \sum_l \sum_k \sum_r (\bar{x}_{lkr} - \bar{x}_{lk.})(\bar{x}_{lkr} - \bar{x}_{lk.})' && gb(n - 1) \end{aligned}$$

where

$$gbn - 1 = (g - 1) + (b - 1) + (g - 1)(b - 1) + gb(n - 1)$$

Two-way MANOVA Table:

| Source of Variation | Sum of Squares | Degree of Freedom |
|---------------------|-----------------------------|-------------------|
| Factor 1 | $\text{SSP}_{\text{fac.1}}$ | $g - 1$ |
| Factor 2 | $\text{SSP}_{\text{fac.2}}$ | $b - 1$ |
| Interaction | SSP_{int} | $(g - 1)(b - 1)$ |
| Residual | SSP_{res} | $gb(n - 1)$ |
| Total | SSP_{tot} | $gbn - 1$ |

Wilks' test statistics:

For interaction:

$$\Lambda^* = \frac{|\text{SSP}_{\text{res}}|}{|\text{SSP}_{\text{int}} + \text{SSP}_{\text{res}}|}$$

For Factor 1:

$$\Lambda^* = \frac{|\text{SSP}_{\text{res}}|}{|\text{SSP}_{\text{fac.1}} + \text{SSP}_{\text{res}}|}$$

For Factor 2:

$$\Lambda^* = \frac{|\text{SSP}_{\text{res}}|}{|\text{SSP}_{\text{fac.2}} + \text{SSP}_{\text{res}}|}$$

Other test statistics include Pillai-Bartlett trace, Roy's greatest root, and Lawley-Hotelling trace, etc.