Inference on the means of multiple multivariate samples
Part I - Two samples

This section starts with two sample comparison, which is one of the most useful tests in multivariate analysis.

The notations become cumbersome inevitably for multiple samples of multivariate vectors. The subscripts need to denote vector components, to label populations, as well as to index observations.

1 Paired comparison

First consider the case when two multivariate samples are not independent (conventional way to say dependent). Often, the two groups are term as two "treatments" or two methods.

Review paired comparison for case $p = 1$

Notations (to be consistent with the text by J&W): Denote measurements as $X_{\{observation\ j,\ treatment\ i\}}$.

Let X_{j1} and X_{j2} be the measurement responses of unit j assigned to treatment (*a.k.a.* group) 1 and 2 respectively. For example, the measurements can be math test scores of student j at the beginning (treatment 1) and the end (treatment 2) of a semester.

Assume independence between units $j = 1 \cdots, n$, with variable mean $\mathbb{E}(X_{jk}) = \mu_k, k = 1, 2$.

Assume that there is dependence of the two samples (corresponding to the two treatment methods) on the jth measurements X_{j1} and X_{j2}, for $j = 1, \cdots, n$. For example, the before and after test scores X_{j1} and X_{j2} are of the same student for each $j = 1, \cdots, n$.

To account for the dependence, linking or pairing the jth observations by

$D_j = X_{j1} - X_{j2} =$ difference in measurements of the two treatments 1 and 2 on unit j.

The sample mean and sample variance of $D_j, j = 1, \cdots, n$, are

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

Relevant equal-mean hypotheses for testing the dependent or paired two-samples become

$$\begin{cases} H_o: & \delta = 0 \\ H_a: & \delta \neq 0 \end{cases} \qquad where \quad \delta = \mu_1 - \mu_2.$$

Assuming $i.i.d.\ D_j \sim N(\delta, \sigma^2)$, the paired two-sample t-test statistic is the one-sample t-statistic on the D_j's,

$$t = \frac{\bar{D} - \delta}{s_d/\sqrt{n}} \sim t_{n-1} \qquad under \quad H_o$$

(The notation means the null H_0 is the default $\delta = \mu_1 - \mu_2$.)

With the assumption $D_j \sim N(\delta, \sigma^2)$, we may construct a confidence interval

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

for treatment mean difference

$$\delta = \mu_1 - \mu_2 = \mathbb{E}(X_1 - X_2) = \mathbb{E}(X_{1j} - X_{2j})$$

Note that we may write the squared t-statistic as

$$t_{n-1}^2 = (\bar{D} - \delta)[\widehat{Cov}(\bar{D}) - \delta]^{-1}(\bar{D} - \delta) = n(\bar{D} - \delta)(s_d^2)^{-1}(\bar{D} - \delta)$$

The expression gives a form generalizable to multivariate case.

Generalization to p-dimensional case

Measurements: $X_{\{treatment\ i,\ observation\ j,\ variable\ k\}}$
(note the changed orders of i and j in the subscript)

Let X_{1jk}, X_{2jk} be the measurements of variable k $(k = 1, \cdots, p)$ on unit j with respect to treatment 1 and 2 respectively. For example, the variables can be the pre and post test scores of student j on p subjects at the beginning and the end of a learning period, using teaching method $k = 1, 2$.

Assume independence between units $j = 1 \cdots, n$. Denote the p vectors

$$\boldsymbol{D}_j = \begin{bmatrix} D_{j1} \\ \vdots \\ D_{jp} \end{bmatrix} = \begin{bmatrix} X_{1j1} - X_{2j1} \\ \vdots \\ X_{1jp} - X_{2jp} \end{bmatrix}, \qquad \bar{\boldsymbol{D}} = \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{D}_j = \begin{bmatrix} \frac{1}{n}\sum_{j=1}^{n} D_{j1} \\ \vdots \\ \frac{1}{n}\sum_{j=1}^{n} D_{jp} \end{bmatrix} = \begin{bmatrix} \bar{D}_1 \\ \vdots \\ \bar{D}_p \end{bmatrix}$$

and the matrix

$$\boldsymbol{S}_d = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{D}_j - \bar{\boldsymbol{D}})(\boldsymbol{D}_j - \bar{\boldsymbol{D}})' = \left[\frac{1}{n-1}\sum_{j=1}^{n}(D_{ji} - \bar{D}_i)(D_{jk} - \bar{D}_k)\right]_{p\times p} = [S_{ik}]_{p\times p}$$

$\boldsymbol{S}_d$ is the sample covariance matrix on the measurements of the difference vector. The diagonal entry S_{kk} is the sample variance of the kth (difference) variable, the (i,k)th entry S_{ik} is the sample covariance of the ith and kth variables. Denote

$$E(\boldsymbol{D}) = \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_p \end{bmatrix}, \qquad Cov(\boldsymbol{D}) = \Sigma_d$$

The equal-mean hypotheses for the paired two-samples is $\begin{cases} H_o: & \boldsymbol{\delta} = 0 \\ H_a: & \boldsymbol{\delta} \neq 0 \end{cases}$. If furthermore, $\mathbf{D} \sim N_p(\boldsymbol{\delta}, \Sigma_d)$, it is natural to use the Hotelling's T^2 statistic

$$T^2 = n(\bar{\mathbf{D}} - \boldsymbol{\delta})'\mathbf{S}^{-1}(\bar{\mathbf{D}} - \boldsymbol{\delta}) \qquad \sim \quad \frac{(n-1)p}{n-p}F_{p,n-p} \qquad under \quad H_o.$$

Using small letters for observations, we may use observed differences $\boldsymbol{d}_j$ to construct confidence region for $\boldsymbol{\delta}$

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

$(1-\alpha)100\%$ confidence intervals for δ_k, the difference of the two population mean parameters of the kth variable, can be constructed by observed $\bar{d}_k$ using T^2's F-distribution.

$$\bar{d}_k \pm \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{d_k}^2}{n}}, \qquad k = 1, \cdots, p,$$

or using the Bonferroni method using t distribution

$$\bar{d}_k \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{d_k}^2}{n}}, \qquad k = 1, \cdots, p.$$

These are simultaneous confidence intervals for all $k = 1, \cdots, p$. For large $n - p$, T^2 is approximately χ^2_p, analogous to the one sample case. The asymptotic simultaneous confidence intervals are

$$\bar{d}_k \pm \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{s^2_{d_k}}{n}} \qquad \text{(approximately)}, \qquad where \quad P\left(\chi^2_p \geq \chi^2_p(\alpha)\right) = \alpha$$

2 Contrast matrix

2.1 Contrast matrix for comparisons of component means

Consider the original data of two samples have the jth observation $\boldsymbol{X}_j = (X_{1j1}, \cdots, X_{1jp},\ X_{2j1}, \cdots, X_{2jp})'$, for $j = 1, \cdots, n$, with the population mean vector $(\mu_{11}, \cdots, \mu_{1p},\ \mu_{21}, \cdots, \mu_{2p})'$. Often we are interested in comparing $\mu_{1k} - \mu_{2k}$, $k = 1, \cdots, p$.

Other than taking the difference of the data as in the previous section, a more general way to compare component means is by constructing contrast matrix.

For example, consider the matrix $\boldsymbol{C}$ below applied to the original jth observation.

$$\boldsymbol{C}\boldsymbol{X}_j = \begin{bmatrix} 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & -1 & \cdots & 0 \\ \vdots & & & \vdots & \vdots & & & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & -1 \end{bmatrix}_{p \times 2p} \begin{bmatrix} X_{1j1} \\ \vdots \\ X_{1jp} \\ X_{2j1} \\ \vdots \\ X_{2jp} \end{bmatrix} = \begin{bmatrix} X_{1j1} - X_{2j1} \\ X_{1j2} - X_{2j2} \\ \vdots \\ X_{1jp} - X_{2jp} \end{bmatrix}$$

The resulting data is the jth observed difference vector.

$$\boldsymbol{C}\boldsymbol{X}_j = \boldsymbol{D}_j, \qquad \boldsymbol{C}\bar{\boldsymbol{X}} = \bar{\boldsymbol{D}}$$

So instead of calculating $\bar{\boldsymbol{d}}$ from the data and then applying T^2, we may use $\boldsymbol{C}\bar{x} = \bar{d}$ directly. The test of equal means can be written as

$$T^2 = n(\boldsymbol{C}\bar{\boldsymbol{x}})'(\boldsymbol{C}\boldsymbol{S}_x\boldsymbol{C}')^{-1}(\boldsymbol{C}\bar{\boldsymbol{x}}) \qquad \sim \quad \frac{(n-1)p}{n-p}F_{p,n-p} \quad under \quad H_o.$$

The T^2 value is the same as using the difference data. The confidence region for $\boldsymbol{C}\boldsymbol{\mu} = \boldsymbol{\delta}$ can be constructed similarly.

$$n(\boldsymbol{C}\bar{\boldsymbol{x}} - \boldsymbol{C}\boldsymbol{\mu})'(\boldsymbol{C}\boldsymbol{S}_x\boldsymbol{C}')^{-1}(\boldsymbol{C}\bar{\boldsymbol{x}} - \boldsymbol{C}\boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$$

The matrix $\boldsymbol{C}$ is a contrast matrix.

A contrast matrix $\boldsymbol{C}$ consists of sum-to-zero row vectors, called contrast vectors.

Let $\mathbf{1} = [1\ \cdots\ 1]'$ be the vector of all 1's with length equal to the number of columns of a contrast matrix $\boldsymbol{C}$, then by definition, matrix $\boldsymbol{C}$ is orthogonal to the vector $\mathbf{1}$ in the sense that $\boldsymbol{C}\mathbf{1} = \mathbf{0}$.

2.2 Contrast matrix for repeated measurements

Contrast matrix is particularly useful in the analysis of repeated measurements, in which different treatments are applied on the same subjects, ofter over successive periods of time.

3 Compare mean vectors of two independent populations with common Σ

The following two independent sample case and related tests are among the most used in practice. Consider:

Random sample 1 of size n_1 is from a random vector of mean $\boldsymbol{\mu}_1$ covariance matrix Σ_1 (in population 1)
Random sample 2 of size n_2 is from a random vector of mean $\boldsymbol{\mu}_2$ covariance matrix Σ_2 (in population 2)
The two samples are independent.

If the samples sizes n_1, n_2 are not large, assume further that

Both population random variables are of multivariate normal with common covariance matrix

$$\Sigma_1 = \Sigma_2 = \Sigma$$

We wish to test $\begin{cases} H_o: & \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \\ H_a: & \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 \end{cases}$ without necessarily specifying the exact value of the common mean vector.

The H_o of $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ implies that all p components of the two population mean vectors are equal.

Suppose that the sample means and sample covariance matrices are $\bar{\boldsymbol{X}}_i$, and $\boldsymbol{S}_i$, for sample $i = 1.2$. A natural estimator of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is the difference of the sample mean vectors $\boldsymbol{X}_1 - \bar{\boldsymbol{X}}_2$, which is an unbiased estimator.

$$\mathbb{E}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$$

Under the equal-covariance assumption $\Sigma_1 = \Sigma_2 = \Sigma$, we can obtain an estimate of the common covariance matrix Σ by the pooled sample covariance matrix, an unbiased estimator.

$$\boldsymbol{S}_{pool} = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2}{n_1 + n_2 - 2}, \qquad \mathbb{E}(\boldsymbol{S}_{pool}) = \Sigma$$

By the independence between the two samples,

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

which can be estimated via $\boldsymbol{S}_{pool}$,

$$\mathbb{E}\left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{S}_{pool}\right] = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma = Cov(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)$$

With the sample variance estimate in hand, the rest inference steps are straightforward generalizations of the one sample case and the univariate two-sample case.

If $\boldsymbol{X}_{11} \cdots, \boldsymbol{X}_{1n_1}$ are i.i.d. p-variate random observations from $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $\boldsymbol{X}_{21}, \cdots, \boldsymbol{X}_{2n_2}$ i.i.d. p-variate random observations from $N_p(\boldsymbol{\mu}_2, \Sigma)$, and if the two samples are independent, then the test statistic

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

has the property

$$T^2 \sim \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1}F_{p,n_1+n_2-p-1} \qquad under \quad H_o: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2.$$

The result is from the normal distribution of $\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2$, the degrees of freedom of F is from the independence of the two samples and the distribution property of $\boldsymbol{S}_{pool}$ (Wishart distribution; proof omitted).

Note that T^2 is often written as $T^2 = \frac{n_1 n_2}{n_1 + n_2}[(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]'\ \boldsymbol{S}^{-1}_{pool}\ [(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]$.

As in the case of paired samples, the Maximization Lemma (extension of Cauchy-Schwarz Inequality) implies that

$$\boldsymbol{b}'(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) \pm c\sqrt{\boldsymbol{b}'\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\boldsymbol{S}_{pool}\boldsymbol{b}} \qquad with \quad c^2 = \frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}(\alpha)$$

is a $(1-\alpha)\%$ confidence region of $\boldsymbol{b}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for any $\boldsymbol{b} \in \mathbb{R}^p$. A particularly useful application is the simultaneous confidence intervals for $\mu_{1i} - \mu_{2i}$, the difference of the component means:

$$\bar{X}_{1i} - \bar{X}_{2i} \pm c\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)s_{ii(pool)}}, \quad i = 1, \cdots, p,$$

where $s_{ii(pool)}$ is the ith diagonal element in $\boldsymbol{S}_{pool}$. The p simultaneous confidence intervals form a hyper-rectangle in $\mathbb{R}^p$.

The $(1-\alpha)\%$ Bonferroni simultaneous confidence interval for the difference of the component means $\mu_{1i} - \mu_{2i}$ is

$$\bar{X}_{1i} - \bar{X}_{2i} \pm t_{n_1+n_2-2,\alpha/2p}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)s_{ii(pool)}}, \quad i = 1, \cdots, p.$$

As often the case, Bonferroni confidence intervals are rather conservative. Since T^2 induced simultaneous confidence interval are specially cases of a more general result, they can be too conservative as well. These intervals sometimes are too wide to be practically useful.

4 Comparing mean vectors of two independent populations with $\Sigma_1 \neq \Sigma_2$

When $\Sigma_1 \neq \Sigma_2$, that is, the covariance matrices of the two populations can not be assumed equal, we no longer have a statistics distance measure like the pivot T^2 in the equal covariance case that does not depend on the unknown Σ_i's. Approximations or asymptotic methods are needed to compare the means of two multivariate samples.

In the following we present two approximation methods.

Asymptotic method for large sample sizes when $\Sigma_1 \neq \Sigma_2$

To use asymptotic methods, typically we need to require the sample sizes are large relative to the number of variables, that is, when $n_i - p$ $(i = 1, 2)$ are sufficiently large. By the independence of the two samples,

$$Cov(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) = \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2$$

It is natural to estimate the covariance matrix by

$$\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2$$

The Central Limit Theorem implies that, asymptotically,

$$\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 \;\sim\; N_p\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2\right) \qquad \text{(approximately)}$$

which leads to the next approximation that, for large n_i's, under the null H_o (that $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are the true means),

$$T^2 = [(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]'\left(\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2\right)^{-1}[(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \quad \sim \chi^2_p \quad \text{(approximately)}$$

Confidence regions and simultaneous confidence intervals can be constructed accordingly.

Remarks

- When the sample sizes are moderate to large, Hotelling's T^2 is known to be remarkably unaffected by slight departures from normality and outliers.
- In the special case $n_1 = n_2 = n$, one can derive that

  $$\frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2 = \left(\frac{1}{n}+\frac{1}{n}\right)\boldsymbol{S}_{pool}$$

  which means that under equal sample sizes, the procedure is basically the same as the one using the pooled sample covariance matrix.
- Therefore, the effect of $\Sigma_1 \neq \Sigma_2$ is the least when $n_1 = n_2$, and the effect is more severe when the sample size difference $|n_1 - n_2|$ is large.

Approximation method for moderate to small sample sizes when $\Sigma_1 \neq \Sigma_2$

When sample sizes are not large, we assume that the underlying populations are of multivariate normal distribution. In addition, each sample size $n_i > p$, $i = 1, 2$. Then the testing statistic and its approximate null distribution are known as the multivariate **Behrens-Fisher problem**.

The approach (proof omitted) uses F distribution to approximation the null distribution.

Under the null $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$,

$$T^2 \sim \frac{vp}{v-p+1}F_{p,v-p+1} \qquad \text{(approximately)}$$

where the test statistics is as before,

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

The value of v in the multiplier and in the second degree of freedom of the F needs to be estimated by

$$v = \frac{p(p+1)}{\sum_{i=1}^{2}\frac{1}{n_i}\left\{tr\left[\left(\frac{1}{n_i}\boldsymbol{S}_i\boldsymbol{S}_0^{-1}\right)^2\right] + \left(tr\left[\frac{1}{n_i}\boldsymbol{S}_i\boldsymbol{S}_0^{-1}\right]\right)^2\right\}}$$

where

$$\boldsymbol{S}_0 = \frac{1}{n_1}\boldsymbol{S}_1 + \frac{1}{n_2}\boldsymbol{S}_2.$$

Remarks

- When $n_1 = n_2, S_1 = S_2$, then $\nu = n_1 + n_2$.
- Testing the equality of two covariance matrices is often sensitive to the assumption of multivariate normality of the underlying distributions.
- The complexity of unequal covariance matrices can largely avoided when the sample sizes are large by using asymptotic methods.

Note: Related sections in Johnson and Wichern: 6.1-6.3.