Statistics 135

Chapter 6

Comparing Two or More Means

Chris Drake
Department of Statistics
University of California, Davis

Paired Comparisons

1 *Univariate case:* Pairs of data (X_{1i}, X_{2i}) with

$$X_{ji} \sim N(\mu_j, \sigma_j^2) \quad for \quad j = 1, .., n$$

,

2 We wish to test

$$H_0 : \ \ \mu_1 = \mu_2 \qquad vs \qquad H_1 : \ \ \mu_1 \neq \mu_2$$

we form $D_i = X_{1i} - X_{2i}$ and under H_0 the difference

$$D_i \sim N(0, \sigma_d^2)$$

and the test-statistic we use is

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{n-1}$$

1 *Multivariate case:* Our data consists of pairs of vectors and the differences formed from those vectors

$$(\mathbf{X}_{1j}, \mathbf{X}_{2j}) \qquad \textit{and} \qquad \mathbf{D_i} = \mathbf{X}_{1j} - \mathbf{X}_{2j}$$

we also calculate the vector of sample average differences

$$\bar{\mathbf{D}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{D}_j$$

and the sample covariance matrix

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

from this we can calculate the Hotelling's T^2 statistic

$$T^2 = n\bar{\mathbf{D}}'\mathbf{S}^{-1}\bar{\mathbf{D}}$$

to test $H_0: \;\; \mu_1 - \mu_2 = \mu_D = 0.$

2 For a more general hypothesis the data consists of

i X_{ijk} for $i = 1, 2,\ j = 1, .., n$ and $k = 1, ..p$ where X_{ijk} is the k^{th} component of the i^{th} pair member for the j^{th} sampled pair.

ii The interest is in hypotheses about $\delta = \mu_1 - \mu_2$

iii $D_{jk} = X_{1jk} - X_{2jk}$ is the difference for the k^{th} component of the j^{th} pair. If we assume that our data vectors follow multivariate normal distributions then

iv $\mathbf{D}$ has a p-variate normal distribution $N_p(\delta, \mathbf{\Sigma_d})$.

v Hypothesis: $H_0: \quad \delta = 0$ vs $H_1: \quad \delta \neq 0$

vi Test statistic

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

vii H_0 is rejected if $n(\bar{\mathbf{d}} - \delta)'\mathbf{S}_d^{-1}(\bar{\mathbf{d}} - \delta) > \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$

3 Note, that rejection of H_0 means that on average, the mean difference of all components are not the same. That does not imply that all components are different, only that at least one of the components has a mean difference that is not zero.

4 A $(1-\alpha)$ confidence region is given by:

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

if it contains the point zero, the null hypothesis that all p mean differences are zero, is not rejected.

5 If the null hypothesis of all p-components having a mean difference of zero is rejected, we can use Bonferroni or the individual intervals derived from the T^2 statistic to investigate which differences are not zero.

6 If our sample sizes (ie the number of pairs) are large, then T^2 has an approximate χ_p^2 distribution, even if our data does not come from a p-variate normal distribution.

Repeated Measures

1 Repeated measures refers to the design when several measurements on one or more variables are taken over time; we can have repeated measurements on a vector. We can also have a single univariate variable that is observed over time and under different experimental conditions. $\mathbf{X}_j = (X_{1j}, X_{2j}, .., X_{qj})$ are measurements on the same experimental unit and are correlated and we assume

$$\mathbf{X} \sim N_q(\mu, \mathbf{\Sigma}) \qquad where \qquad \sigma_{ij} = Cov(X_i, X_j)$$

2 If we are interested to see if successive measurements differ on average we could test $H_0: \ \ \mu_i = \mu_j$; to see if all successive means are the same, we test $\mu_1 - \mu_2 = ... = \mu_q - \mu_{q-1} = 0$; this leads to a matrix

$$C_{(q-1)q} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}$$

$\mu_i - \mu_j$ is called a contrast and $\mathbf{C}$ is called a contrast matrix.

Recall, if

$$\mathbf{X} \sim N(\mu, \mathbf{\Sigma}) \Longrightarrow \mathbf{CX} \sim N(\mathbf{C}\mu, \mathbf{C\Sigma C}')$$

and the test statistics for $H_0 : \mu_1 - \mu_2 = ... = \mu_p - \mu_{p-1} = 0$ can be tested using

$$T^2 = n\big(\mathbf{C\bar{X}} - \mathbf{C}\mu\big)'\big(\mathbf{CSC}'\big)^{-1}\big(\mathbf{C\bar{X}} - \mathbf{C}\mu\big) \sim \frac{(n-1)(q-1)}{(n-q+1)}F_{(q-1,n-q+1)}$$

where $q - 1$ is the number of contrasts.

We reject H_0 if $(0, .., 0)$ is not in the confidence region given by

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

As before, confidence intervals can be derived from the test statistic above.

Comparing Two Means

In this section we will consider inference when we have two independent samples from two populations.

Sample		Sample Statistics	
pop 1	$\mathbf{x}_{11}, \mathbf{x}_{12}...\mathbf{x}_{1n_1}$	$\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} \mathbf{x}_{1j}$	$\mathbf{S}_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \bar{\mathbf{x}}_{\mathbf{1}})(\mathbf{x}_{1j} - \bar{\mathbf{x}}_{\mathbf{1}})'$
pop 2	$\mathbf{x}_{21}, \mathbf{x}_{22}...\mathbf{x}_{2n_2}$	$\bar{\mathbf{x}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \mathbf{x}_{2j}$	$\mathbf{S}_2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\mathbf{x}_{2j} - \bar{\mathbf{x}}_{\mathbf{2}})(\mathbf{x}_{2j} - \bar{\mathbf{x}}_{\mathbf{2}})'$

Assumptions

1 $\mathbf{X}_{i1}, ..., \mathbf{X}_{in_i}$ are samples from a p-variate (not necessarily) normal distribution with mean vector μ_i and covariance matrix $\mathbf{\Sigma}_i$ for $i = 1, 2$

2 The samples are taken independently ie $\mathbf{X}_{1j}$ is independent of $\mathbf{X}_{2l}$ for $j = 1, ..., n_1$ and $l = 1, ..., n_2$

3 For small n_1 and n_2 we assume multivariate normality of both populations from which the samples are taken.

4 We also assume $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$ and calculate a pooled estimate of the covariance matrix.

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

Since

$$Cov(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2) = Cov(\bar{\mathbf{X}}_{\mathbf{1}}) + Cov(\bar{\mathbf{X}}_{\mathbf{2}}) = \frac{1}{n_1}\mathbf{\Sigma}_1 + \frac{1}{n_2}\mathbf{\Sigma}_2$$

a test statistic for testing $H_0:\ \mu_1-\mu_2-\delta_0=0$ that expands the T^2 statistic to two samples is

$$T^2 = (\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2-\delta_0)'\left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathbf{S}_{pooled}\right]^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2-\delta_0)$$

1 If the samples are from normally distributed populations, then

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

where $p = dim(\mathbf{X})$.

2 A $(1 - \alpha) \times 100\%$ confidence region is given by

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\mu_1 - \mu_2))' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_p \right]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - (\mu_1 - \mu_2)) < c^2$$

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

3 The confidence region is an ellipsoid centered at $\mathbf{x}_1 - \mathbf{x}_2$ and all $\mu_1 - \mu_2$ within c^2 of $\mathbf{x}_1 - \mathbf{x}_2$ constitute the confidence region.

Simultaneous confidence intervals

1 Bonferroni method for p confidence intervals $\mu_{1i} - \mu_{2i}$ for $i = 1, .., p$ is given by

$$(\bar{x}_{1i} - \bar{x}_{2i} \pm t_{n_1+n_2-2;(\alpha/2p)} \sqrt{\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big) s_{ii,pooled}}$$

2 With $\mathbf{a}$ proportional to $\mathbf{S}^{-1}_{pooled}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ we have

$$P(T^2 \leq c^2) = P(t^2_{\mathbf{a}} \leq c^2 \text{ for all } \mathbf{a}) = 1 - \alpha$$

where

$$t^2_{\mathbf{a}} = \frac{\Big[\mathbf{a}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \mathbf{a}'(\mu_1 - \mu_2)\Big]^2}{\mathbf{a}'\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)\mathbf{S}_{pooled}\mathbf{a}}$$

and

$$\bar{x}_{1i} - \bar{x}_{2i} \pm \sqrt{\frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p,n_1+n_2-p-1}} \sqrt{(\frac{1}{n_1} + \frac{1}{n_2}) s_{ii,pooled}}$$