STA 437: Applied Multivariate Statistics

Al Nosedal.
University of Toronto.

Winter 2015

1. Chapter 5. Tests on One or Two Mean Vectors
   - Some important results
   - Comparing Two Mean Vectors
   - Tests on Individual variables conditional on rejection of H_0

"If you can't explain it simply, you don't understand it well enough"

Albert Einstein.

Definition

$\mathbf{A}$ ($k \times k$ symmetric matrix) is positive definite if

$$\mathbf{x}^{'}\mathbf{A}\mathbf{x} > 0$$

for all vectors $\mathbf{x} \neq \mathbf{0}$.

Cauchy-Schwarz Inequality

Let $\mathbf{b}$ and $\mathbf{d}$ be any two $p \times 1$ vectors. Then

$$(\mathbf{b}^{'}\mathbf{d})^2 \leq (\mathbf{b}^{'}\mathbf{b})(\mathbf{d}^{'}\mathbf{d})$$

with equality if and only if $\mathbf{b} = c\mathbf{d}$ (or $\mathbf{d} = c\mathbf{b}$).

Extended Cauchy-Schwarz Inequality

Let $\mathbf{b}$ and $\mathbf{d}$ be any two vectors, and let $\mathbf{B}$ be a positive definite matrix. Then

$$(\mathbf{b}^{'}\mathbf{d})^2 \leq (\mathbf{b}^{'}\mathbf{B}\mathbf{b})(\mathbf{d}^{'}\mathbf{B}^{-1}\mathbf{d})$$

with equality if and only if $\mathbf{b} = c\mathbf{B}^{-1}\mathbf{d}$ (or $\mathbf{d} = c\mathbf{B}\mathbf{b}$) for some constant c.

Maximization Lemma

Let **B** be positive definite and **d** be a given vector. Then, for an arbitrary nonzero vector **x**,

$$\max_{\mathbf{x}\neq 0} \frac{(\mathbf{x}^{'}\mathbf{d})^2}{\mathbf{x}^{'}\mathbf{B}\mathbf{x}} = \mathbf{d}^{'}\mathbf{B}^{-1}\mathbf{d}$$

with the maximum attained when $\mathbf{x} = c\mathbf{B}^{-1}\mathbf{d}$ for any constant $c \neq 0$.

Hotelling's T^2-Test

We assume that a random sample $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$ is available from $N_p(\mu, \mathbf{\Sigma})$, where $\mathbf{y}_i$ contains the p measurements on the ith sampling unit. We estimate μ by $\bar{\mathbf{y}}$ and $\mathbf{\Sigma}$ by $\mathbf{S}$. In order to test $H_0 : \mu = \mu_{\mathbf{0}}$ versus $H_1 : \mu \neq \mu_{\mathbf{0}}$, we use the test statistic

$$T^2 = n(\bar{\mathbf{y}} - \mu_{\mathbf{0}})' \mathbf{S}^{-1} (\bar{\mathbf{y}} - \mu_{\mathbf{0}}).$$

The distribution is indexed by two parameters, the dimension p and degrees of freedom $\nu = n - 1$. We reject H_0 if $T^2 > T^2_{\alpha, p, n-1}$ and "accept" otherwise. Critical values of the T^2-distribution are found in Table A.7.

Development of T^2

The development of this multivariate significance test proceeds as follows:

a) We define a new variable:

$$\mathbf{W}_{n\times 1} = a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + ... + a_p\mathbf{X}_p = \mathbf{X}_{n\times p}\mathbf{a}_{p\times 1}$$

where $\mathbf{X}_j$ is an n-element column vector giving each of the n subjects' score on dependent measure j; $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p]$ is an $n \times p$ data matrix whose ith row gives subject i's scores on each of the outcome variables; $\mathbf{a}$ is a p-element column vector giving the weights by which the dependent measures are to be multiplied before being added together.

Development of T^2

b) Our null hypothesis is that $\mu_1 = \mu_{10}, \mu_2 = \mu_{20}, ..., \mu_p = \mu_{p0}$ are all true. If one or more of these equalities is false, the null hypothesis is false. This hypothesis can be expressed in matrix form as

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} \mu_{10} \\ \mu_{20} \\ \vdots \\ \mu_{p0} \end{pmatrix} = \mu_0$$

and it implies that $\mu_{\mathbf{w}} = \mathbf{a}' \mu_{\mathbf{0}}$.

Development of T^2

c) The variance of a linear combination of variables can readily be expressed as a linear combination of the variances and covariances of the original variables

$$S^2_{\mathbf{W}} = \mathbf{a}^{'}\mathbf{S}\mathbf{a}$$

where **S** is the covariance matrix of the outcome variables. Thus the univariate t computed on the combined variable **W** is given by

$$t(\mathbf{a}) = \frac{\mathbf{a}^{'}\bar{\mathbf{X}} - \mathbf{a}^{'}\mu_{\mathbf{0}}}{\sqrt{\mathbf{a}^{'}\mathbf{S}\mathbf{a}/n}}$$

Squaring it yields

$$t^2(\mathbf{a}) = n\frac{\mathbf{a}^{'}(\bar{\mathbf{X}} - \mu_{\mathbf{0}})(\bar{\mathbf{X}} - \mu_{\mathbf{0}})^{'}\mathbf{a}}{\mathbf{a}^{'}\mathbf{S}\mathbf{a}}$$

Development of T^2

Note that $t^2(\mathbf{a})$ depends on $\mathbf{a}$, thus we will maximize $t^2(\mathbf{a})$. Using our maximization lemma

$$t^2(\mathbf{a}^*) = T^2 = n(\bar{\mathbf{x}} - \mu_{\mathbf{0}})' \mathbf{S}^{-1}(\bar{\mathbf{x}} - \mu_{\mathbf{0}})$$

where $\mathbf{a}^* = \mathbf{S}^{-1}(\bar{\mathbf{x}} - \mu_{\mathbf{0}})$

Example. Evaluating T^2

Let the data matrix for a random sample of size $n = 3$ from a bivariate Normal population be

$$\mathbf{X} = \begin{pmatrix} 6 & 9 \\ 10 & 6 \\ 8 & 3 \end{pmatrix}$$

Evaluate the observed T^2 for $\mu_0 = [9, 5]'$.

Solution

$$\bar{\mathbf{x}} = \begin{pmatrix} (6+10+8)/3 \\ (9+6+3)/3 \end{pmatrix} = \begin{pmatrix} 8 \\ 6 \end{pmatrix}$$

$$s_{11} = \frac{(6-8)^2+(10-8)^2+(8-8)^2}{2} = 4$$

$$s_{12} = \frac{(6-8)(9-6)+(10-8)(6-6)+(8-8)(3-6)}{2} = -3$$

$$s_{22} = \frac{(9-6)^2+(6-6)^2+(3-6)^2}{2} = 9$$

Solution(cont.)

$$\mathbf{S} = \begin{pmatrix} 4 & -3 \\ -3 & 9 \end{pmatrix}$$

$$\mathbf{S}^{-1} = \frac{1}{27} \begin{pmatrix} 9 & 3 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1/3 & 1/9 \\ 1/9 & 4/27 \end{pmatrix}$$

Solution (cont.)

$$T^2 = n(\bar{\mathbf{y}} - \mu_{\mathbf{0}})' \mathbf{S}^{-1} (\bar{\mathbf{y}} - \mu_{\mathbf{0}}).$$

$$T^2 = \frac{7}{9}$$

Example. Testing a multivariate mean vector

Perspiration from 20 healthy females was analyzed. Three components, X_1= sweat rate, X_2 = sodium content, and X_3 = potassium content, were measured, and the results, which we call the *sweat data*, are given in T5-1.DAT.
Test the hypothesis $H_0 : \mu' = [4, 50, 10]$ against
$H_1 : \mu' \neq [4, 50, 10]$ at the level of significance $\alpha = 0.05$.

Solution

$$\bar{\mathbf{x}} = \begin{pmatrix} 4.64 \\ 45.400 \\ 9.965 \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} 2.879 & 10.010 & -1.810 \\ 10.010 & 199.788 & -5.640 \\ -1.810 & -5.640 & 3.628 \end{pmatrix}$$

$$\mathbf{S}^{-1} = \begin{pmatrix} 0.586 & -0.022 & 0.258 \\ -0.022 & 0.006 & -0.002 \\ 0.258 & -0.002 & 0.402 \end{pmatrix}$$

$$T^2 = n(\bar{\mathbf{y}} - \mu_{\mathbf{0}})'\mathbf{S}^{-1}(\bar{\mathbf{y}} - \mu_{\mathbf{0}}) = 9.74$$

Solution

From Table A.7, we obtain the critical value $T_{0.05,3,19} = 10.719$.
Comparing the observed $T^2 = 9.74$ with the critical value 10.719 we see that $T^2 = 9.74 < 10.719$, and consequently, we **can't** reject H_0 at the 5% level of significance.
Another way of finding the critical value for T^2.

$$\frac{(n-1)p}{(n-p)} F_{p,n-p}(0.05) = \frac{(19)(3)}{17} F_{3,17}(0.05) = \frac{(19)(3)}{17}(3.20) = 10.72941$$

```
### Example 5.2 Sweat data

data<-read.table(file="T5-1.DAT")

x.bar<-apply(data,2,FUN=mean)

x.bar

mu.0<-c(4,50,10)

difference<-x.bar-mu.0

difference<-matrix(difference,ncol=1)

S.inv<-solve(cov(data))
```

```
n<-dim(data)[1]

T.2<-n*t(difference)%*%S.inv%*%difference

T.2

## Critical value

T.alpha<-(19*3/17)*qf(0.95,3,17)

T.alpha
```

Example 5.3.2

In Table 3.4 we have $n = 10$ observations on $p = 3$ variables. Desirable levels for y_1 and y_2 are 15.0 and 6.0, respectively, and the expected level of y_3 is 2.85. We can, therefore, test the hypothesis $H_0 : \mu^{'} = [15, 6.0, 2.85]$ against $H_1 : \mu^{'} \neq [15, 6.0, 2.85]$ at the level of significance $\alpha = 0.05$.

Solution

From Table A.7, we obtain the critical value $T_{0.05,3,9} = 16.766$. Comparing the observed $T^2 = 24.559$ with the critical value 16.766 we see that $T^2 = 24.559 > 16.766$, and consequently, we **reject** H_0 at the 5% level of significance.
Another way of finding the critical value for T^2.

$$\frac{(n-1)p}{(n-p)} F_{p,n-p}(0.05) = \frac{(9)(3)}{7} F_{3,7}(0.05) = \frac{(9)(3)}{7}(4.35) = 16.778$$

```
## Calcium data

data<-read.table(file="T3_4_CALCIUM.DAT")

data<-data[ ,-1]

x.bar<-apply(data,2,FUN=mean)

x.bar

mu.0<-c(15,6,2.85)

difference<-x.bar-mu.0

difference<-matrix(difference,ncol=1)
```

```
S.inv<-solve(cov(data))

n<-dim(data)[1]

T.2<-n*t(difference)%*%S.inv%*%difference

T.2

## Critical value

crit.val<-(9*3)/(7)*qf(0.95,3,7)

crit.val
```

Univariate Two-sample t-Test

In the one-variable case we obtain a random sample $y_{11}, y_{12}, \ldots, y_{1n_1}$ from $N(\mu_1, \sigma_1^2)$ and a second random sample $y_{21}, y_{22}, \ldots, y_{2n_2}$ from $N(\mu_2, \sigma_2^2)$. We assume that the two samples are independent and that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, say, with σ^2 unknown. From the two samples we calculate the pooled variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

where $n_1 + n_2 - 2$ is the sum of the weights $n_1 - 1$ and $n_2 - 1$ in the numerator.

Univariate Two-sample t-Test

To test $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$,
we use

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom when H_0 is true. We therefore reject H_0 if $|t| \geq t_{\alpha/2, n_1+n_2-2}$.

Multivariate Two-Sample T^2-Test

We wish to test $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$.
We obtain a random sample $\mathbf{y}_{11}, \mathbf{y}_{12}, ..., \mathbf{y}_{1n_1}$ from $N_p(\mu_\mathbf{1}, \mathbf{\Sigma_1})$ and a second random sample $\mathbf{y}_{21}, \mathbf{y}_{22}, ..., \mathbf{y}_{2n_2}$ from $N_p(\mu_\mathbf{1}, \mathbf{\Sigma_2})$. We assume that the two samples are independent and that $\mathbf{\Sigma_1} = \mathbf{\Sigma_2} = \mathbf{\Sigma}$, say, with $\mathbf{\Sigma}$ unknown.

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{y}}_\mathbf{1} - \bar{\mathbf{y}}_\mathbf{2})' \mathbf{S_p}^{-1} (\bar{\mathbf{y}}_\mathbf{1} - \bar{\mathbf{y}}_\mathbf{2})$$

where

$$\mathbf{S_p} = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2]$$

We reject H_0 if $T^2 \geq T^2_{\alpha,p,n_1+n_2-2}$. Critical values of T^2 are found in Table A.7.

Example

Four psychological tests were given to 32 men and 32 women. The data are recorded in Table 5.1. The variables are y_1 = pictorial inconsistencies, y_2 = paper from board, y_3 = tool recognition, y_4 = vocabulary.

The mean vectors are

$$\hat{\mathbf{y}}_1 = \begin{pmatrix} 15.97 \\ 15.91 \\ 27.19 \\ 22.75 \end{pmatrix}$$

$$\hat{\mathbf{y}}_2 = \begin{pmatrix} 12.34 \\ 13.91 \\ 16.66 \\ 21.94 \end{pmatrix}$$

Example (cont.)

The covariance matrices of the two samples are

$$\mathbf{S_1} = \begin{pmatrix} 5.192 & 4.545 & 6.522 & 5.250 \\ 4.545 & 13.18 & 6.760 & 6.266 \\ 6.522 & 6.760 & 28.67 & 14.47 \\ 5.250 & 6.266 & 14.47 & 16.65 \end{pmatrix}$$

$$\mathbf{S_2} = \begin{pmatrix} 9.136 & 7.549 & 4.864 & 4.151 \\ 7.549 & 18.60 & 10.22 & 5.446 \\ 4.864 & 10.22 & 30.04 & 13.49 \\ 4.151 & 5.446 & 13.49 & 28.00 \end{pmatrix}$$

Test the hypothesis $H_0 : \boldsymbol{\mu_1} = \boldsymbol{\mu_2}$ versus $H_1 : \boldsymbol{\mu_1} \neq \boldsymbol{\mu_2}$ at the 0.01 significance level.

Univariate t-tests

We give a procedure that could be used to check each variable following rejection of H_0 by a two-sample T^2 test:

$$t_j = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{\sqrt{[(n_1 + n_2)/n_1 n_2] s_{jj}}}, \; j = 1, 2, ..., p,$$

where s_{jj} is the jth diagonal element of $\mathbf{S}_p$. Reject $H_0 : \mu_{1j} = \mu_{2j}$ if $|t_j| > t_{\alpha/2, n_1+n_2-2}$.

Examples

Please, see tutorial 3.