THE UNIVERSITY OF MINNESOTA

Statistics 5401 September 26, 2005

Hotelling's T^2 Examples

This handout contains examples of a two-sample Hotelling's T^2 and a one-sample paired Hotelling's T^2. We work with the Fisher iris data in data set `T11_05`. The variables are x_1 = sepal length, x_2 = sepal width, x_3 = petal length, x_4 = petal width.

Two-sample T^2

We first test the hypothesis that the population mean flower measurements are the same for *I. setosa* and *I. versicolor*, that is, the average petal and sepal dimensions are the same for the two varieties. Thus we are comparing two populations – a two sample situation. We assume that both varieties have the *same variance matrix* $\boldsymbol{\Sigma}$.

```
Cmd> fisher <- read("","t11_05") # read from JWData5.txt
) Data from Table 11.5 p. 657-658 in
) Applied Mulivariate Statistical Analysis, 5th Edition
) by Richard A. Johnson and Dean W. Wichern, Prentice Hall, 2002
) These data were edited from file T11-5.DAT on disk from book
) The variety number was moved to column 1
) Measurements on petals of 4 varieties of Iris.  Originally published
in
) R. A. Fisher, The use of mltiple measurements in taxonomic problems,
) Annals of Eugenics, 7 (1936) 179-198
) Col. 1: variety number (1 = I. setosa, 2 = I. versicolor,
)                         3 = I. virginica)
) Col. 2: x1 = sepal length
) Col. 3: x2 = sepal width
) Col. 4: x3 = petal length
) Col. 5: x4 = petal width
) Rows 1-50:    group 1 = Iris setosa
) Rows 51-100:  group 2 = Iris versicolor
) Rows 101-150: group 3 = Iris virginica
Read from file "TP1:Stat5401:Data:JWData5.txt"

Cmd> varieties <- fisher[,1]

Cmd> setosa <- fisher[varieties == 1,-1]

Cmd> versicolor <- fisher[varieties == 2,-1]

Cmd> virginica <- fisher[varieties == 3,-1 ]

Cmd> stuff1 <- tabs(setosa, mean:T, covar:T)

Cmd> compnames(stuff1) #components of stats computed by tabs()
(1) "mean"
(2) "covar"

Cmd> stuff2 <- tabs(versicolor,mean:T,covar:T)

Cmd> xbar1 <- stuff1$mean # col vector

Cmd> xbar2 <- stuff2$mean # col vector

Cmd> s1 <- stuff1$covar  # setosa variance matrix (4 x 4)

Cmd> s2 <- stuff2$covar  # versicolor variance matrix
```

```
Cmd> print(xbar1,s1)  # setosa statistics
xbar1:
(1)       5.006       3.428       1.462       0.246
s1:
(1,1)     0.12425    0.099216    0.016355    0.010331
(2,1)    0.099216     0.14369    0.011698    0.009298
(3,1)    0.016355    0.011698    0.030159   0.0060694
(4,1)    0.010331    0.009298   0.0060694    0.011106

Cmd> print(xbar2,s2)  # versicolor statistics
xbar2:
(1)       5.936        2.77        4.26       1.326
s2:
(1,1)     0.26643    0.085184      0.1829     0.05578
(2,1)    0.085184    0.098469    0.082653    0.041204
(3,1)      0.1829    0.082653     0.22082    0.073102
(4,1)     0.05578    0.041204    0.073102    0.039106

Cmd> n1 <- nrows(setosa); n2 <- nrows(versicolor) # sample sizes

Cmd> df1 <- n1 - 1; df2 <- n2 - 1 # degrees of freedom in s1 & s2

Cmd> vector(df1,df2) # error degrees of freedom in each sample
(1)          49          49

Cmd> dfpooled <- df1 + df2; spooled <- (df1*s1 + df2*s2)/dfpooled

Cmd> print(dfpooled,spooled)
dfpooled:               pooled  degrees  of  freedom
(1)          98
spooled:                pooled  variance  matrix
(1,1)     0.19534      0.0922    0.099627    0.033055
(2,1)      0.0922     0.12108    0.047176    0.025251
(3,1)    0.099627    0.047176     0.12549    0.039586
(4,1)    0.033055    0.025251    0.039586    0.025106

Cmd> vhat <- (1/n1 + 1/n2) * spooled; print(vhat)
vhat:                     estimated var matrix of xbar1-xbar2
(1,1)   0.0078136    0.003688   0.0039851   0.0013222
(2,1)    0.003688   0.0048432    0.001887     0.00101
(3,1)   0.0039851    0.001887   0.0050195   0.0015834
(4,1)   0.0013222     0.00101   0.0015834   0.0010042

Cmd> diff <- xbar1 - xbar2 # difference of mean vectors

Cmd> se <- sqrt(diag(vhat)) # univariate standard errors

Cmd> # Note that se is a plain (column) vector

Cmd> print(diff,se)  # differences of means and standard errors
diff:
(1)       -0.93       0.658      -2.798       -1.08
se:
(1)    0.088395    0.069593    0.070849     0.03169

Cmd> tstats <- diff/se;print(tstats) # univariate t-statistics
tt:
(1)     -10.521       9.455     -39.493      -34.08
```

```
Cmd> t_sq_12 <- diff' %*% (solve(vhat) %*% diff); write (t_sq_12)
t_sq_12:
(1,1)       2580.83855

Cmd> diff' %*% solve(vhat, diff) # alternate use of solve()
(1,1)      2580.8

Cmd> diff' %*% (vhat %\% diff) # equivalent to preceding
(1,1)      2580.8

Cmd> p <- ncols(setosa); print(p) # number of variables
p:
(1)           4

Cmd> fe <- dfpooled; f <- (fe - p + 1)*t_sq_12/(fe*p); print(f)
f:                                F-statistic form of T^2
(1,1)      625.46

Cmd> # f on 4 and 98 - 4 + 1 = 95 df, the Null distribution

Cmd> cumF(f,p,fe-p+1,upper:T)  # P-value = P(F(4,95) > 625.5)
(1,1)  2.6649e-67
```

There is a black box way to find T^2 and its associated P-value using macro `hotell2val()`:

```
Cmd> usage(hotell2val)
hotell2val(x1, x2 [, pval:T]), REAL matrices x1 and x2 with no MISSING
  elements and ncols(x1) = ncols(x2)

Cmd> hotell2val(setosa,versicolor,pval:T)
WARNING: searching for unrecognized macro hotell2val near
hotell2val(
component: hotelling
(1,1)      2580.8
component: pvalue
(1,1)          0
```

Paired T^2 Next we consider the question as to whether the *I. setosa* sepal and petal *lengths* differ from the sepal and petal *widths*. This asks a question about the shape of the flowers. The hypothesis to be tested can be stated as

$$H_0: \mu_1 = \mu_2 \text{ and } \mu_3 = \mu_4,$$

where μ_i is the population mean of x_i for *I. setosa*.

This is a comparison of the means of two variables measured on the same flower and is sometimes called a "within-subject" comparison. It should be viewed as a bivariate *paired* comparison of length and width, that is of $\mathbf{y}_1 = [x_1, x_3]'$ and $\mathbf{y}_2 = [x_2, x_4]'$. The comparison reduces to testing the hypothesis that $E[\mathbf{d}] = 0$, where $\mathbf{d} = \mathbf{y}_1 - \mathbf{y}_2 = \mathbf{Cx}$, where

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

Rather than computing explicitly the vector of differences, $\mathbf{d} = (x_1 - x_2, x_3 - x_4)'$, we compute the sample mean vector and variance matrix of the differences by transforming the *I. Setosa* mean vector and variance matrix using the matrix $\mathbf{C}$ with $\bar{\boldsymbol{d}} = \mathbf{C}\bar{\boldsymbol{x}}$ and $\mathbf{S}_d = \mathbf{CSC}'$.

```
Cmd> c <- matrix(vector(1,-1,0,0, 0,0,1,-1),4)';print(c)
c:
(1,1)           1          -1           0           0
(2,1)           0           0           1          -1

Cmd> sd <- c %*% s1 %*% c' ; print(sd)
sd:
(1,1)    0.069506   0.0036245
(2,1)   0.0036245    0.029127

Cmd> dbar <- c %*% xbar1; print(dbar)
dbar:
(1,1)       1.578
(2,1)       1.216

Cmd> vhatdbar <- (1/n1) * sd ; print(vhatdbar)
vhatdbar:
(1,1)   0.0013901   7.249e-05
(2,1)   7.249e-05  0.00058253

Cmd> t2d <- dbar' %*% (vhatdbar %\% dbar); print(t2d)
t2d:
(1,1)      4012.1

Cmd> p <- 2; fe <- n1-1;f <- (fe - p+1)*t2d/(fe*p)

Cmd> f  # 1% critical value = 5.077
(1,1)      1965.1

Cmd> cumF(f,p,fe-p+1,upper:T)
(1,1)  9.0628e-47         < .01, therefore reject H_0

Cmd> hotellval(setosa %*% c',pval:T) # black box using hotellval()
component: hotelling
(1,1)      4012.1
component: pvalue
(1,1)           0
```