TUTORIAL 3
STA437 WINTER 2015

AL NOSEDAL

Contents

1. Toy example

Example.

	Group 1	
Subject	*X*	*Y*
1	1	1
2	2	0
3	2	1
4	2	2
5	3	0
6	3	1
7	3	1.85
8	4	0.5

Group 2		
Subject	*X*	*Y*
1	3	2
2	3	4
3	3	5
4	4	2
5	4	3
6	4	4
7	5	4
8	5	5

1. Enter data for groups 1 and 2.

2. Make a scatterplot for these two groups. Use two different colours, one for each group.

3. Compute, the following three linear combos:
a) $Z_1 = 1X + 0Y$
b) $Z_2 = 0X + 1Y$
c) $Z_3 = 0.51X + 0.86Y$

4. Find the mean of Z_1 for groups 1 and 2. Compute the difference between these two means.

5. Repeat 4 for Z_2 and Z_3.

Solution (R code)

```
## 1. Entering data

g1.x<-c(1,2,2,2,3,3,3,4)

g1.y<-c(1,0,1,2,0,1,1.85,0.5)

g2.x<-c(3,3,3,4,4,4,5,5)

g2.y<-c(2,4,5,2,3,4,4,5)

## Groups

group.1<-matrix(c(g1.x,g1.y),nrow=8,ncol=2)

group.2<-matrix(c(g2.x,g2.y),nrow=8,ncol=2)
```

```
all<-rbind(group.1,group.2)

## 2. Scatterplots

plot(group.1,col="blue",pch=19,xlim=c(0,6),ylim=c(0,6),xlab="X",ylab="Y")

points(group.2,col="red",pch=19)

## 3. Linear combos and
## 4. Computing means

## Z1 = 1 X + 0 Y

a1<-matrix(c(1,0),nrow=2,ncol=1)

z1<-all%*%a1

## group 1

mean(z1[1:8, ])

sd(z1[1:8, ])

## group 2

mean(z1[9:16, ])

sd(z1[9:16, ])

## Z2 = 0 X + 1 Y

a2<-matrix(c(0,1),nrow=2,ncol=1)

z2<-all%*%a2

## group 1

mean(z2[1:8, ])

sd(z2[1:8, ])
```

```
## group 2

mean(z2[9:16, ])

sd(z2[9:16, ])


## Z3 = 0.51 X + 0.86 Y

a3<-matrix(c(0.51,0.86),nrow=2,ncol=1)

z3<-all%*%a3

## group 1

mean(z3[1:8, ])

sd(z3[1:8, ])

## group 2

mean(z3[9:16, ])

sd(z3[9:16, ])

## 5. Difference between means

d1<-mean(z1[1:8, ])-mean(z1[9:16, ])

d1

d2<-mean(z2[1:8, ])-mean(z2[9:16, ])

d2

d3<-mean(z3[1:8, ])-mean(z3[9:16, ])

d3
```

Now, let us try to understand what these linear combos mean, geometrically.

```
plot(group.1,col="blue",pch=19,xlim=c(0,6),ylim=c(0,6),xlab="X",ylab="Y")
```

```
points(group.2,col="red",pch=19)

arrows(0,0,a3[1],a3[2],lty=2)

for(i in 1:8){

points(z3[i]*a3[1],z3[i]*a3[2],col="blue")


}


for(i in 9:16){

points(z3[i]*a3[1],z3[i]*a3[2],col="red")


}
```

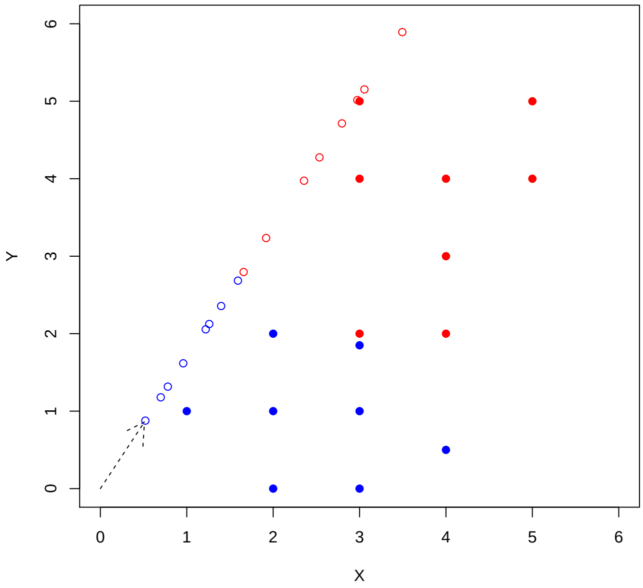

FIGURE 1. Scatterplot and optimal projection.

The basic idea for simplifying the problem of analyzing differences in location simultaneously in several variables consists of finding suitable linear combinations of all variables. Using an arbitrary linear combination

$$Z = a_1X_1 + a_2X_2 + ... + a_pX_p$$

we obtain in both groups a new variable Z, whose mean and standard deviation is denoted by $\bar{Z}_1$ and s_1 (for group 1), and by $\bar{Z}_2$ and s_2 (for group 2), respectively. For the linear combination Z we can compute the associated standard distance

$$D(Z) = D(a_1, a_2, ..., a_p) = \frac{|\bar{Z}_1 - \bar{Z}_2|}{s}$$

with

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

We can now define the *multivariate standard distance* as the maximum standard distance that can be obtained from any linear combination of X_1 to X_p. Formally,

$$D_p = maxD(a_1, a_2, ..., a_p)$$

over all possible choices of the coefficients a_1 to a_p. The index p in D_p indicates that the measure of distance is based on p variables. The linear combination for which the maximum is achieved will be called **discriminant function** or **discriminant variable**. Multivariate standard distance is therefore nothing but univariate standard distance for a particular linear combination called the discriminant function.

Finally, let us compute the "discriminant function"

R code

```
## S1

S1<-cov(group.1)

y.bar.1<-apply(group.1,2,FUN=mean)

## S2

S2<-cov(group.2)

y.bar.2<-apply(group.2,2,FUN=mean)

## Sp
```

```
Sp<-(1/14)*(7*S1+7*S2)

diff<-y.bar.1-y.bar.2

diff<-matrix(diff,nrow=2,ncol=1)

## Finding Discriminant Function

a.star<-solve(Sp)%*%diff

a.star<-(-1)*a.star

norm<-t(a.star)%*%a.star

## Normalized Discriminant Function

disc.fun<-a.star/sqrt(norm[1])

disc.fun
```

2. Comparing two mean vectors

2.1. **Review of Univariate Two-Sample t-Test.** In the one-variable case we obtain a random sample $y_{11},\ y_{12}, ...,\ y_{1n_1}$ from $N(\mu_1, \sigma_1^2)$ and a second random sample $y_{21},\ y_{22}, ...,\ y_{2n_2}$ 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. 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 $|t| \geq t_{\alpha/2, n_1+n_2-2}$.

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 $N_p(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$ and a second random sample $\mathbf{y}_{21}, \mathbf{y}_{22}, ..., \mathbf{y}_{2n_2}$ from $N_p(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_2})$. We assume that the two samples are independent and that $\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \boldsymbol{\Sigma}$, say, with $\boldsymbol{\Sigma}$ 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

$$\bar{\mathbf{y}}_\mathbf{1} = \begin{pmatrix} 15.97 \\ 15.91 \\ 27.19 \\ 22.75 \end{pmatrix}$$

$$\bar{\mathbf{y}}_\mathbf{2} = \begin{pmatrix} 12.34 \\ 13.91 \\ 16.66 \\ 21.94 \end{pmatrix}$$

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.

Solution

The pooled covariance matrix is

$$\mathbf{S_p} = \frac{1}{32+32-2}[(32-1)\mathbf{S}_1 + (32-1)\mathbf{S}_2] = \begin{pmatrix} 7.164 & 6.047 & 5.693 & 4.701 \\ 6.047 & 15.89 & 8.492 & 5.856 \\ 5.693 & 8.492 & 29.36 & 13.98 \\ 4.701 & 5.856 & 13.98 & 22.32 \end{pmatrix}$$

$$T^2 = \frac{n_1 n_2}{n_1+n_2}(\mathbf{\bar{y}_1} - \mathbf{\bar{y}_2})' \mathbf{S_p}^{-1}(\mathbf{\bar{y}_1} - \mathbf{\bar{y}_2}) = 97.6015$$

From interpolation in Table A.7, we obtain $T^2_{0.01,4,62} = 15.373$, and we therefore reject $H_0 : \boldsymbol{\mu_1} = \boldsymbol{\mu_2}$.

```
## "Reading" data

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

## breaking down data

males<-data[1:32 ,-1]

females<-data[33:64 ,-1]

## sample sizes

n1<-dim(males)[1]

n2<-dim(females)[1]

## mean vectors

y.bar.1<-apply(males,2,FUN=mean)

y.bar.2<-apply(females,2,FUN=mean)
```

```
## covariance matrices

S.1<-cov(males)

S.2<-cov(females)

Sp<-(n1+n2-2)^(-1)*((n1-1)*S.1 + (n2-1)*S.2)

## Hotelling's T^2

T.2<-(n1*n2/(n1+n2))*t(y.bar.1-y.bar.2)%*%solve(Sp)%*%(y.bar.1-y.bar.2)

T.2

## Critical value

p<-dim(males)[2]

crit.val<-((n1+n2-2)*p/(n1+n2-p-1))*qf(0.99,p,n1+n2-p-1)

crit.val
```

3. Test on Individual variables conditional on Rejection of H_0 by the T^2-test

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 $|t_j| > t_{\alpha/2, n_1+n_2-2}$.

Example. For the psychological data in Table 5.1, calculate t-tests for the individual variables at the 0.01% significance level.

Solution (R code)

```
## Test on Individual variables conditional on Rejection of H0

## Univariate t-tests

# c = constant
```

```
c<-(n1*n2/(n1+n2))

## first variable

t1.2<-c*t(y.bar.1[1]-y.bar.2[1])%*%solve(Sp[1,1])%*%(y.bar.1[1]-y.bar.2[1])

t1<-sqrt(t1.2)

t1

# you should get 5.41

## Critical value

crit.val.1<-qt(1-(0.01/2),n1+n2-2)

crit.val.1

# you should get 2.65

## second variable

t2.2<-c*t(y.bar.1[2]-y.bar.2[2])%*%solve(Sp[2,2])%*%(y.bar.1[2]-y.bar.2[2])

t2<-sqrt(t2.2)

t2

# you should get 2.00

## third variable

t3.2<-c*t(y.bar.1[3]-y.bar.2[3])%*%solve(Sp[3,3])%*%(y.bar.1[3]-y.bar.2[3])

t3<-sqrt(t3.2)

t3

# you should get 7.77

## fourth variable
```

```
t4.2<-c*t(y.bar.1[4]-y.bar.2[4])%*%solve(Sp[4,4])%*%(y.bar.1[4]-y.bar.2[4])

t4<-sqrt(t4.2)

t4

# you should get 0.68
```

4. Computation of T^2

4.1. **Obtaining T^2 from Multiple Regression.** We illustrate the regression approach to computation of T^2 using the psychological data in Table 5.1. We set $w = \frac{n_2}{n_1+n_2} = \frac{32}{64} = \frac{1}{2}$ for each observation in the first group (males) and equal to $-\frac{n_1}{n_1+n_2} = \frac{-1}{2}$ in the second group (females). When w is regressed on the 64 y's, we obtain

$$\begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} -0.751 \\ 0.051 \\ -0.020 \\ 0.047 \\ -0.031 \end{pmatrix}$$

$R^2 = 0.6115$. And $T^2 = (n_1 + n_2 - 2)\frac{R^2}{1-R^2} = \frac{62(0.6115)}{1-0.6115} = 97.601$, as we obtained before.

R code

```
## categorical variable for males

w1<-n2/(n1+n2)

## categorical variable for females

w2<-(-1)*n1/(n1+n2)

## vector with categorical variables

w<-c(rep(w1,n1),rep(w2,n2))

w

## Regressing w on variables
```

```
new.data<-matrix(unlist(data[ ,-1]),nrow=64,ncol=4)

## saving results from regression

model<-lm(w~new.data)

## let us see what we have in model

sum.mod<-summary(model)

sum.mod

names(sum.mod)

## Finding Hotelling's T^2

## We need R^2 (Multiple R-squared)

R.2<-sum.mod$r.squared

T.2<-(n1+n2-2)*(R.2/(1-R.2))

T.2
```

5. Paired Observations Test

5.1. **Univariate case.** Suppose that two samples are not independent because there exists a natural pairing between the ith observation y_i in the first sample and the ith observation x_i in the second sample for all i, as, for example, when a treatment is applied twice to the same individual. With such pairing, the samples are often referred to as *paired observations or matched pairs.* The two samples thus obtained are correlated, and a two-sample test statistic is not appropriate. We reduce the two samples to one by working with the difference between the paired observations, as in the following layout for two treatments applied to the same subject:

Pair Number	*Treatment 1*	*Treatment 2*	$d_i = y_i - x_i$
1	y_1	x_1	d_1
2	y_2	x_2	d_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n	d_n

To obtain a t-test it is not sufficient to assume individual Normality for each of y and x. To allow for the covariance between y and x, we need the additional assumption that y and x have a bivariate Normal distribution with

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix}$$

It then follows that $d_i = y_i - x_i$ is $N(\mu_y - \mu_x, \sigma_d^2)$, where $\sigma_d^2 = \sigma_y^2 + \sigma_x^2 - 2\sigma_{yx}$. From $d_1, d_2, ..., d_n$ we calculate

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \qquad \text{and} \qquad s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2.$$

To test $H_0 : \mu_y = \mu_x$, that is, $H_0 : \mu_d = 0$, we use the one-sample statistic

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

which is distributed as t_{n-1} if H_0 is true. We reject H_0 in favor of $H_1 : \mu_d \neq 0$ if $|t| > t_{\alpha/2,n-1}$.

5.2. **Multivariate case.** Here we assume the same natural pairing of sampling units as in the univariate case, but we measure p variables on each sampling unit. In terms of two treatments applied to each sampling unit, this situation is as follows:

Pair Number	*Treatment 1*	*Treatment 2*	$d_i = y_i - x_i$
1	$\mathbf{y}_1$	$\mathbf{x}_1$	$\mathbf{d}_1$
2	$\mathbf{y}_2$	$\mathbf{x}_2$	$\mathbf{d}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$\mathbf{y}_n$	$\mathbf{x}_n$	$\mathbf{d}_n$

Given the observed differences $\mathbf{d}_j^{'} = [d_{j1}, d_{j2}, ..., d_{jp}]$, $j = 1, 2, ..., n$, an α-level test of $H_0 : \delta = \mathbf{0}$ versus $H_0 : \delta \neq \mathbf{0}$ for an $N_p(\delta, \mathbf{\Sigma}_d)$ population rejects H_0 if the observed

$$T^2 = n\bar{\mathbf{d}}'\mathbf{S^{-1}}_d\bar{\mathbf{d}} > \frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)$$

where $F_{p,n-p}(\alpha)$ is the upper (100α)th percentile of an F-distribution with p and $n-p$ d.f.

Example. Municipal wastewater treatment plants are required by law to monitor their discharges into rivers and streams on a regular basis. Concern about the reliability of data from one of these self-monitoring programs led to a study in which samples of effluent were divided and sent to two laboratories for testing. One-half of each sample was sent to the Wisconsin State Laboratory of Hygiene, and one-half was sent to a private commercial laboratory routinely used in the monitoring program. Measurements of biochemical oxygen demand (BOD) and suspended solids (SS) were obtained for $n = 11$ sample splits, from the two laboratories. The data are available in water.DAT. First and third columns of water.DAT correspond to BOD for Commercial lab and State lab, respectively. Second and fourth columns of water.DAT correspond to SS for Commercial lab and State lab, respectively. Do the two laboratories' chemical analyses agree? (Use $\alpha = 0.05$).

Solution (R code)

```
water<-read.table(file="water.DAT")

water<-matrix(unlist(water),nrow=11,ncol=4)

# vector with difference for BOD

a1<-matrix(c(1,0,-1,0),nrow=4,ncol=1)

d1<-water%*%a1

# vector with difference for SS

a2<-matrix(c(0,1,0,-1),nrow=4,ncol=1)

d2<-water%*%a2

# matrix of differences

diffs<-cbind(d1,d2)

# vector of differences

diff.bar<-apply(diffs,2,FUN=mean)

# covariance matrix

S.d<-cov(diffs)

# n= number of rows of diffs
```

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

# Test statistic

T.2<-n*t(diff.bar)%*%solve(S.d)%*%diff.bar

T.2

# p = number of columns of diffs

p<-dim(diffs)[2]

alpha<-0.05

# critical value

crit.val<-(p*(n-1)/(n-p))*qf(1-alpha,p,n-p)

crit.val
```

Taking $\alpha = 0.05$, we find that $[p(n-1)/(n-p)]F_{p,n-p}(0.05) = [2(10)/9]F_{2,9}(0.05) = 9.47$. Since $T^2 = 13.6 > 9.47$, we reject H_0 and conclude that there is a nonzero mean difference between the measurements of the two laboratories.