Multivariate Inference - II

Two multivariate sample tests

STAT 32950-24620

Spring 2023 (4/6)

Groups of random vectors

$$X = \begin{bmatrix} X_1 \\ \cdots\cdots \\ X_2 \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ X_g \end{bmatrix},$$

Each random vector

$$X_k = [X_{k1} \ \cdots \ X_{ki} \ \cdots \ X_{kp}]' = (X_{k1}, \ \cdots, X_{ki}, \ \cdots, X_{kp})$$

may have several observations

$$(X_{k,11}, \ \cdots, X_{k,1i}, \ \cdots, X_{k,1p})$$
$$\vdots$$
$$(X_{k,n_k 1}, \ \cdots, X_{k,n_k i}, \ \cdots, X_{k,n_k p})$$

Observed data

$$X = \begin{bmatrix} X_1 \\ \cdots\cdots \\ X_2 \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ X_g \end{bmatrix} = \left[\begin{array}{cccc} x_{1,11} & x_{1,12} & \cdots & x_{1,1p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1,n_1 1} & x_{1,n_1 2} & \cdots & x_{1,n_1 p} \\ \hline x_{2,11} & x_{2,12} & \cdots & x_{2,1p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{2,n_2 1} & x_{2,n_1 2} & \cdots & x_{2,n_2 p} \\ \hline \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \hline x_{g,11} & x_{g,12} & \cdots & x_{g,1p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{g,n_g 1} & x_{g,n_g 2} & \cdots & x_{g,n_g p} \end{array}\right]$$

(notice different usages of notations)

Example

Data: Turtle shell measurements (n = 48 obs., each of 4 variables)

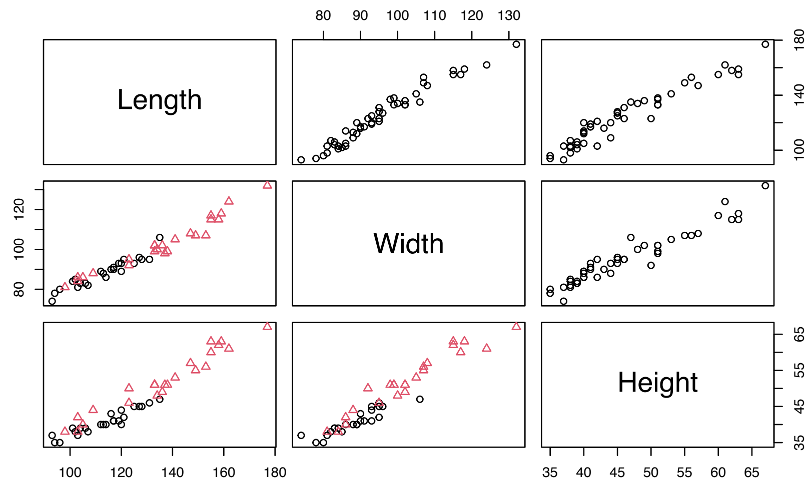
```
load(file="turtles.rda")
attach(turtles)  # from package Flury
#str(turtles)    #original data n=48, var=4: Sex,L,W,H
#summary(turtles)
```

Plot

```
pairs(turtles[,-1],
   lower.panel = function(x, y){
      points(x, y, pch = unclass(turtles[,1]),
             col = as.numeric(turtles[,1]))},
   main =
    'Data "turtles": male (circle), female (triangle)')
```

Data pairwise plots

Data "turtles": male (circle), female (triangle)



Subset observations into two samples

Subset observations by Gender ($g = 2$ "treatment" groups)

```
male=subset(turtles[,2:4],Gender=="Male")
female=subset(turtles[,2:4],Gender=="Female")
```

Random vectors

$$X = \left[\begin{array}{c} X_{male} \\ \hline X_{female} \end{array} \right]$$

Observation data by groups

$$X = \left[\begin{array}{c} X_{male} \\ \hline X_{female} \end{array} \right] = \left[\begin{array}{ccc} x_{m,11} & x_{m,12} & x_{m,13} \\ \vdots & \vdots & \vdots \\ x_{m,n_1 1} & x_{m,n_1 2} & x_{m,n_1 3} \\ \hline x_{f,11} & x_{f,12} & x_{f,13} \\ \vdots & \vdots & \vdots \\ x_{f,n_2 1} & x_{f,n_2 2} & x_{f,n_2 3} \end{array} \right]$$

Two independent samples with common covariance

Assuming two samples from X_1 and X_2 are **independent**.

$$X_1 \sim N_p(\mu_1, \Sigma_1), \quad X_2 \sim N_p(\mu_2, \Sigma_2)$$

Consider the equal-covariance case

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

To estimate Σ, use the **pooled sample covariance matrix**

$$S_{pool} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

S_{pool} is an unbiased estimator of the population covariance matrix.

$$E(S_{pool}) = \Sigma$$

Two independent samples: mean vectors $\mu_1 = \mu_2$?

To test

$$H_o : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

Apply Hotelling's T^2 to $\bar{X}_1 - \bar{X}_2$.

Use

$$\widehat{Cov}(\bar{X}_1 - \bar{X}_2) = \left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pool}$$

(compared with S/n for one sample)

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

Hotelling's T^2 for two samples

Under $H_o : \mu_1 - \mu_2 = 0$,

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

where

$$T^2 = [(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pool} \right]^{-1} [(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)]$$

Apply Hotelling's T^2 on mean difference

Example (by steps)

Comparing two p-variate means ($p = 3, n_1 = 24, n_2 = 24$)

$$H_o : \mu_{male} = \mu_{female}$$

$$H_a : \mu_{male} \neq \mu_{female}$$

```
male=subset(turtles[,2:4],Gender=="Male")
female=subset(turtles[,2:4],Gender=="Female")
mbar=colMeans(male); fbar=colMeans(female);
diffmean = mbar - fbar
n1=24; n2=24; p=3
Spool=
(n1-1)/(n1+n2-2)*cov(male)+(n2-1)/(n1+n2-2)*cov(female)
T2=t(diffmean)%*%solve((1/n1+1/n2)*Spool)%*%diffmean
```


Assumption $\Sigma_1 = \Sigma_2$

$g = 2, p = 3, n_1 = 24, n_2 = 24$.

```
cov(male)
```

```
##        Length Width Height
## Length 138.77 79.15  37.38
## Width   79.15 50.04  21.65
## Height  37.38 21.65  11.26
```

```
cov(female)
```

```
##        Length Width Height
## Length  451.4 271.2 168.70
## Width   271.2 171.7 103.29
## Height  168.7 103.3  66.65
```

Assuming $\Sigma_1 = \Sigma_2$, for now.

Check descriptive sample statistics

```
diffmean; Spool
```

```
## Length  Width Height
## -22.62 -14.29 -11.25
```

```
##        Length  Width Height
## Length  295.1 175.16 103.04
## Width   175.2 110.89  62.47
## Height  103.0  62.47  38.95
```

```
T2   # 66.8
```

```
##       [,1]
## [1,] 66.76
```

```
# F test p-value
1-pf((n1+n2-1-p)*T2/(p*(n1+n2-2)),df1=p,df2=n1+n2-1-p)
```

```
##           [,1]
## [1,] 1.141e-08
```

Hotelling's T^2 using "manova" in R (Turtle data, $g = 2, p = 3$)

```
data=cbind(Length, Width, Height)
summary(manova(data~Gender),test="Hotelling")
```

```
##           Df Hotelling-Lawley approx F num Df den Df  Pr
## Gender     1             1.45     21.3      3     44 1.:
## Residuals 46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.:
```

Remark: Given F critical, T^2 value can be recovered by

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

```
# F*(p*(n1+n2-2))/(n1+n2-1-p), F(3,44)=21.284
21.3*(3*(24+24-2))/(24+24-1-3)
```

```
## [1] 66.8
```


Univariate marginal test of mean diff. on variable Length

Test each pair of component means by univariate t-test. On Length:

$$H_o : \mu_{male,Length} = \mu_{female,Length}$$
$$H_a : \mu_{male,Length} \neq \mu_{female,Length}$$

```
t.test(Length~Gender) # Marginal t=test
```

```
##
##  Welch Two Sample t-test
##
## data:  Length by Gender
## t = -4.6, df = 36, p-value = 6e-05
## alternative hypothesis: true difference in means betweer
## 95 percent confidence interval:
##  -32.68 -12.57
## sample estimates:
##   mean in group Male mean in group Female
##                113.4                136.0
```


Marginal test on variable Width

$$H_o : \mu_{male,Width} = \mu_{female,Width}$$
$$H_a : \mu_{male,Width} \neq \mu_{female,Width}$$

```
t.test(Width~Gender)
```

```
##
##  Welch Two Sample t-test
##
## data:  Width by Gender
## t = -4.7, df = 35, p-value = 4e-05
## alternative hypothesis: true difference in means betweer
## 95 percent confidence interval:
##  -20.461  -8.123
## sample estimates:
##   mean in group Male mean in group Female
##                88.29               102.58
```


Marginal test on variable Height

$$H_o : \mu_{male,Height} = \mu_{female,Height}$$
$$H_a : \mu_{male,Height} \neq \mu_{female,Height}$$

```
t.test(Height~Gender)
```

```
##
##  Welch Two Sample t-test
##
## data:  Height by Gender
## t = -6.2, df = 31, p-value = 7e-07
## alternative hypothesis: true difference in means betweer
## 95 percent confidence interval:
##  -14.927  -7.573
## sample estimates:
##   mean in group Male mean in group Female
##                40.71                51.96
```