Multivariate Inference II

Multiple sample tests, MANOVA

STAT 32950-24620

Spring 2023 (4/11-13)

MANOVA (Multivariate Inference II)

Groups of random vectors

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

Each X_k $(k = 1, \cdots, g)$ is a vector with p component variables.

Notation for a multivariate random sample

Each group k 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, g multivariate samples

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

(notice various usages of indeces; check context of notations)

Review ANOVA Example

Example: Univariate **Analysis of Variance** (ANOVA)

```
trt = as.factor(c(1,1,1,2,2,3,3,3)) # not as numeric
y1 = c(9,6,9, 0,2, 3,1,2)
cbind(trt,y1)
```

```
##      trt y1
## [1,]   1  9
## [2,]   1  6
## [3,]   1  9
## [4,]   2  0
## [5,]   2  2
## [6,]   3  3
## [7,]   3  1
## [8,]   3  2
```


ANOVA table

Source of variation	SS (sum of squares)	d.f.	Variance ratio (F-value)
Treatments	$SS_{trt} = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(\bar{x}_\ell - \bar{x})^2$	$g-1$	$\dfrac{SS_{trt}/(g-1)}{SS_{res}/(n-g)}$
Residuals	$SS_{res} = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j} - \bar{x}_\ell)^2$	$n-g$	
Total	$SS_{tot} = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(x_{\ell j} - \bar{x})^2$	$n-1$	

where

$$n = n_1 + \cdots + n_g$$

Example ANOVA on data

```
summary(aov(y1 ~ trt)); aov(y1 ~ trt) # transposed anova
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## trt          2     78      39    19.5 0.0044 **
## Residuals    5     10       2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

## Call:
##    aov(formula = y1 ~ trt)
##
## Terms:
##                 trt Residuals
## Sum of Squares   78        10
## Deg. of Freedom   2         5
##
## Residual standard error: 1.414
## Estimated effects may be unbalanced
```


Multivariate Analysis of Variance (MANOVA)

Example:

```
y2=c(3,2,7, 4,0, 8,9,7); #trt=as.factor(c(1,1,1,2,2,3,3,3),
y = cbind(y1,y2); cbind(trt,y)
```

```
##      trt y1 y2
## [1,]   1  9  3
## [2,]   1  6  2
## [3,]   1  9  7
## [4,]   2  0  4
## [5,]   2  2  0
## [6,]   3  3  8
## [7,]   3  1  9
## [8,]   3  2  7
```

$g = 3$ (trt) samples, $p = 2$ dimensions of measurements.
Samples sizes $n_1 = 3, n_2 = 2, n_3 = 3$.

Multivariate data forms

Example: $g = 3, p = 2, n_1 = 3, n_2 = 2, n_3 = 3$

$$
\begin{bmatrix} 9 & 3 \\ 6 & 2 \\ 9 & 7 \\ \cdots & \cdots \\ 0 & 4 \\ 2 & 0 \\ \cdots & \cdots \\ 3 & 8 \\ 1 & 9 \\ 2 & 7 \end{bmatrix}
\quad Or \quad
\left(\begin{array}{ccc}
\begin{bmatrix} 9 \\ 3 \end{bmatrix} & \begin{bmatrix} 6 \\ 2 \end{bmatrix} & \begin{bmatrix} 9 \\ 7 \end{bmatrix} \\
\begin{bmatrix} 0 \\ 4 \end{bmatrix} & \begin{bmatrix} 2 \\ 0 \end{bmatrix} & \\
\begin{bmatrix} 3 \\ 8 \end{bmatrix} & \begin{bmatrix} 1 \\ 9 \end{bmatrix} & \begin{bmatrix} 2 \\ 7 \end{bmatrix}
\end{array} \right)
\quad Or \quad
\begin{array}{c}
\begin{bmatrix} 9 & 6 & 9 \\ 0 & 2 & \\ 3 & 1 & 2 \end{bmatrix} \\
\begin{bmatrix} 3 & 2 & 7 \\ 4 & 0 & \\ 8 & 9 & 7 \end{bmatrix}
\end{array}
$$

MANOVA table

Source of variation	Matrix of sums of squares and cross products (SSCP)	d.f.
Treatments	$B = \sum_{t=1}^{g} \sum_{j=1}^{n_t} (\bar{x}_t - \bar{x})(\bar{x}_t - \bar{x})'$	$g-1$
Residuals	$W = \sum_{t=1}^{g} \sum_{j=1}^{n_t} (x_{tj} - \bar{x}_t)(x_{tj} - \bar{x}_t)'$	$n-g$
Total	$B + W = \sum_{t=1}^{g} \sum_{j=1}^{n_t} (x_{tj} - \bar{x})(x_{tj} - \bar{x})'$	$n-1$

where B, W are $p \times p$ matrices, $n = n_1 + \cdots + n_g$.

R MANOVA example

```
manova(y ~ trt) # show ANOVA of component variables
```

```
## Call:
##    manova(y ~ trt)
##
## Terms:
##                   trt Residuals
## y1                 78        10
## y2                 48        24
## Deg. of Freedom     2         5
##
## Residual standard errors: 1.414 2.191
## Estimated effects may be unbalanced
```

Remarks: Compare with the univariate ANOVA.

R MANOVA test

```
summary(manova(y ~ trt))
```

```
##           Df Pillai approx F num Df den Df Pr(>F)
## trt        2   1.54     8.39      4     10 0.0031 **
## Residuals  5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Between group SS matrix B

The between treatment (group) matrix of sum of squares and cross products is

$$B = \sum_{t=1}^{3} n_t(\bar{\mathbf{y}}_t - \bar{\mathbf{y}})(\bar{\mathbf{y}}_t - \bar{\mathbf{y}})'$$

```
# Btw trt SS matrix B=(n-1)*cov(fitted)
B=7*cov(manova(y~trt)$fitted)
B
```

```
##      y1  y2
## y1  78 -12
## y2 -12  48
```

```
det(B)
```

```
## [1] 3600
```


Within group SS matrix W

The within treatment (group) matrix of sum of squares and cross products is

$$W = \sum_{t=1}^{3} \sum_{j=1}^{n_t} (\mathbf{y}_{tj} - \bar{\mathbf{y}}_t)(\bar{\mathbf{y}}_{tj} - \bar{\mathbf{y}}_t)'$$

```
#  within trt or residual SS matrix W=(n-1)*cov(residual)
W=7*cov(manova(y~trt)$residual)
W
```

```
##    y1 y2
## y1 10  1
## y2  1 24
```

```
det(W)
```

```
## [1] 239
```


Total SS matrix

The matrix of total sum of squares and cross products is

B + W = total SS matrix

```
7*cov(y)
```

```
##      y1  y2
## y1  88 -11
## y2 -11  72
```

```
det(B+W)
```

```
## [1] 6215
```


Wilks' lambda

$$\Lambda^* = \frac{det(W)}{det(B+W)}$$

```
det(W)/det(B+W)    # Same as Wilks test  0.038
```

```
## [1] 0.03846
```

For $g = 3$ $(p = 3 \geq 1)$, $n = \sum_{t=1}^{g} n_t = 8$,

$$\left(\frac{n-p-2}{p}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2p,2(n-p-2)}$$

```
((8-2-2)/2)*(sqrt(det(B+W)/det(W))-1)    # F = 8.19886
```

```
## [1] 8.199
```

Verify Wilks' lambda test

```
### Test using distribution Wilks' lambda  ###
summary(manova(y~trt), test="Wilks")
```

```
##           Df  Wilks approx F num Df den Df Pr(>F)
## trt        2 0.0385      8.2      4      8 0.0062 **
## Residuals  5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.:
```

```
## verify
1-pf(8.1989,df1=4,df2=8)  # = 0.006234
```

```
## [1] 0.006234
```


Bartlett's approximation

For $n = \sum n_t$ large, under H_o of equal mean vectors,

$$-\left(n-1-\frac{p+g}{2}\right)\ln\Lambda^* \sim \chi^2_{p(g-1)}$$

```
### Using Bartlett's approximation ###
n=8
p=2
g=3
-(n-1-(p+g)/2)*log(0.038455)   # = 14.66
```

```
## [1] 14.66
```

```
1-pchisq(14.6622, df=p*(g-1))   # = 0.0054556
```

```
## [1] 0.005456
```


Verify Bartlett's approximation

```
summary(manova(y~trt))  # another approximation of F
```

```
##           Df Pillai approx F num Df den Df Pr(>F)
## trt        2   1.54     8.39      4     10 0.0031 **
## Residuals  5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.:
```

```
1-pf(8.3882,df1=4,df2=10) # = 0.003096
```

```
## [1] 0.003096
```


Check equal-covariance assumption

Check covariance structure using Box's M

```
source("BoxM.R")    # Test Sigma's are equal
Box_M(y,c(3,2,3))   # Not equal, due to Sigma 2 sigular
```

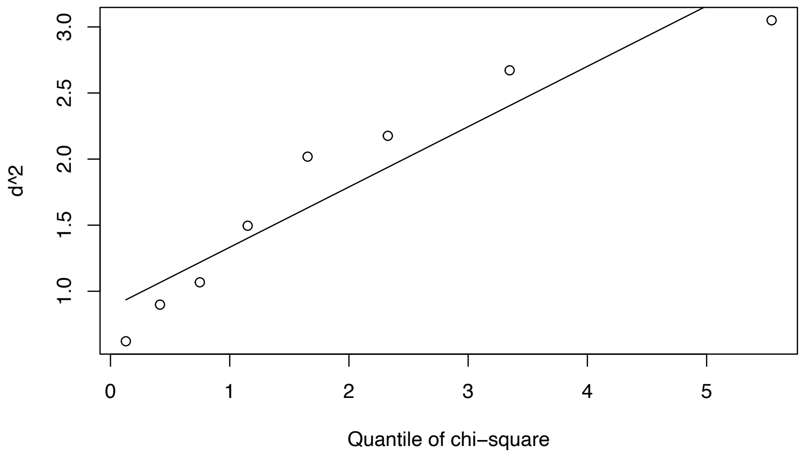
```
## [1] "Test result:"
##           [,1]
## Box.M-C   Inf
## p.value     0
```

```
Box_M(y[c(1:3,6:8),],c(3,3))     # Sigma1 = Sigma3 ok
```

```
## [1] "Test result:"
##             [,1]
## Box.M-C 1.3998
## p.value 0.7056
```

Check normality assumption

```r
source("qqchi2.R"); qqchi2(y)
```




```
## [1] "correlation coefficient:"
## [1] 0.9456
```


Post MANOVA

If the null hypothesis of MANOVA is rejected, which treatments have significant effects?

Write the kth treatment mean as

$$\mu_k = \mu + \tau_k = \text{ Overall mean } + \text{ Effect of treatment } k$$

To compare the effect of treatment k and treatment ℓ, the quantity of interests is the difference of the vectors

$$\tau_k - \tau_\ell$$

which is equivalent to

$$\mu_k - \mu_\ell$$

Comparison of sample means, one variable at a time

For the ith component variable,

the sample estimate of the mean difference

between sample k and sample ℓ is

$$\hat{\tau}_{ki} - \hat{\tau}_{\ell i} = \hat{\mu}_{ki} - \hat{\mu}_{\ell i} = \bar{X}_{ki} - \bar{X}_{\ell i}$$

A confidence interval for $\tau_{ki} - \tau_{\ell i}$ will have the form

$$\hat{\tau}_{ki} - \hat{\tau}_{\ell i} \pm c \times \sqrt{\widehat{var}(\hat{\tau}_{ki} - \hat{\tau}_{\ell i})}$$

Evaluate $var(\hat{\tau}_{ki} - \hat{\tau}_{\ell i})$

Assuming independence between the samples,

$$var(\hat{\tau}_{ki} - \hat{\tau}_{\ell i}) = var\left(\bar{X}_{ki} - \bar{X}_{\ell i}\right) = var\left(\bar{X}_{ki}\right) + var\left(\bar{X}_{\ell i}\right)$$

Under the assumption that the g populations are of equal covariance structure,

$$\Sigma_1 = \cdots = \Sigma_g = \Sigma = [\sigma_{ij}]_{i,j=1,\cdots,p}$$

For variable i, sample k and sample ℓ,

$$var\left(\bar{X}_{ki}\right) = \frac{\sigma_{ii}}{n_k}, \qquad var\left(\bar{X}_{\ell i}\right) = \frac{\sigma_{ii}}{n_\ell}$$

Estimation of $var(\hat{\tau}_{ki} - \hat{\tau}_{\ell i})$

The sample estimate of σ_{ii} is the ith diagonal element of the pooled sample covariance matrix

$$S_{pool} = \frac{1}{\sum_{\ell=1}^{g}(n_\ell - 1)}\left[(n_1 - 1)S_1 + \cdots + (n_g - 1)S_g\right] = \frac{1}{n-g}W$$

which gives

$$var(\hat{\tau}_{ki} - \hat{\tau}_{\ell i}) = var\left(\bar{X}_{ki}\right) + var\left(\bar{X}_{\ell i}\right) = \left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)\frac{w_{ii}}{n-g}$$

with w_{ii} the ith diagonal element of W used in MANOVA.

Bonferroni CI test level determination

By the Bonferroni method, the $(1-\alpha)100\%$ confidence interval for the ith component of the difference vector has the form

$$\hat{\tau}_{ki} - \hat{\tau}_{\ell i} \pm t_d(\alpha/2m)\sqrt{\widehat{var}(\hat{\tau}_{ki} - \hat{\tau}_{\ell i})}$$

where $m = p\binom{g}{2} = pg(g-1)/2$ is the number of simultaneous confidence intervals, which gives the confidence level at the component level

$$\frac{\alpha}{2m} = \frac{\alpha}{pg(g-1)}$$

Bonferroni CI test statistic and its d.f.

Under the null hypothesis $\tau_{ki} - \tau_{\ell i} = 0$,

$$\frac{(\hat{\tau}_{ki} - \hat{\tau}_{\ell i}) - 0}{\sqrt{\widehat{var}(\hat{\tau}_{ki} - \hat{\tau}_{\ell i})}} \sim t_d$$

where the degrees of freedom d is determined by the degrees of freedom of

$$\widehat{var}(\hat{\tau}_{ki} - \hat{\tau}_{\ell i})$$

Note that $n - g = \sum_{\ell=1}^{g} n_\ell - 1$ is the degrees of freedom of W and S_{pool}.

Bonferroni simultaneous confidence intervals

Therefore we have obtain Bonferroni simultaneous component-wise confidence intervals for the treatment group differences $\tau_k - \tau_\ell$,

$$\bar{x}_{ki} - \bar{x}_{\ell i} \pm t_{n-g}(\alpha/2m)\sqrt{s_{ii}\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)}$$

or

$$\bar{x}_{ki} - \bar{x}_{\ell i} \pm t_{n-g}(\alpha/pg(g-1))\sqrt{\frac{w_{ii}}{n-g}\left(\frac{1}{n_k} + \frac{1}{n_\ell}\right)}$$

for all $i = 1, \cdots, p$ and all $k, \ell = 1, \cdots, g$.