STA 437: Applied Multivariate Statistics

Al Nosedal.
University of Toronto.

Winter 2015

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

Albert Einstein.

Decomposition for univariate ANOVA

Consider the following independent samples.
Population 1: 9, 6, 9.
Population 2: 0, 2.
Population 3: 3, 1, 2

Means

$$\bar{x}_1 = \frac{9+6+9}{3} = 8$$
$$\bar{x}_2 = \frac{0+2}{2} = 1$$
$$\bar{x}_3 = \frac{3+1+2}{3} = 2$$
$$\bar{x} = \frac{(9+6+9+0+2+3+1+2}{8} = 4$$

"Decomposition" of observations

$$9 = x_{11} = (\bar{x}) + (\bar{x}_1 - \bar{x}) + (x_{11} - \bar{x}_1)$$
$$9 = 4 + (8 - 4) + (9 - 8)$$
$$9 = 4 + (4) + (1)$$
$$3 = x_{31} = (\bar{x}) + (\bar{x}_3 - \bar{x}) + (x_{31} - \bar{x}_3)$$
$$3 = 4 + (2 - 4) + (3 - 2)$$
$$3 = 4 + (-2) + (1)$$

Decomposing all our obs.

Repeating this operation for each observation, we obtain the arrays:

Observations

$$\begin{pmatrix} 9 & 6 & 9 \\ 0 & 2 & \\ 3 & 1 & 2 \end{pmatrix}$$

Mean

$$\begin{pmatrix} 4 & 4 & 4 \\ 4 & 4 & \\ 4 & 4 & 4 \end{pmatrix}$$

Decomposing all our obs.

Treatment Effect

$$\begin{pmatrix} 4 & 4 & 4 \\ -3 & -3 & \\ -2 & -2 & -2 \end{pmatrix}$$

Residual

$$\begin{pmatrix} 1 & -2 & 1 \\ -1 & 1 & \\ 1 & -1 & 0 \end{pmatrix}$$

Sums of squares

Under H_0, each $\hat{\tau}_l$ is an estimate of zero. If the treatment contribution is large, H_0 should be rejected. The "size" of an array is quantified by stringing the rows and of the array out into a vector and calculating its squared length. This quantity is called the sum of squares (SS). For the observations,

$SS_{obs} = 9^2 + 6^2 + 9^2 + 0^2 + 2^2 + 3^2 + 1^2 + 2^2 = 216$

Similarly,

$SS_{mean} = 4^2 + 4^2 + 4^2 + 4^2 + 4^2 + 4^2 + 4^2 + 4^2 = 8(4^2) = 128$

$SS_{tr} = 4^2 + 4^2 + 4^2 + (-3)^2 + (-3)^2 + (-2)^2 + (-2)^2 + (-2)^2 = 78$

$SS_{res} = 1^2 = (-2)^2 + 1^2 + (-1)^2 + 1^2 + 1^2 + (-1)^2 + 062 = 10$

Sums of squares

The sums of squares satisfy the same decomposition, as the observations

$SS_{obs} = SS_{mean} + SS_{tr} + SS_{res}$

$216 = 128 + 78 + 10$

ANOVA Table

Source of of variation	Sum of Squares (SS)	Degrees of freedom (d.f.)
Treatments	$SS_{tr} = \sum_{l=1}^{g} n_l(\bar{x}_l - \bar{x})^2$	$g - 1$
Residual (Error)	$SS_{res} = \sum_{l=1}^{g} n_l \sum_{j=1}^{n_l}(x_{lj} - \bar{x}_l)^2$	$\sum_{l=1}^{g} n_l - g$
Total (corrected)	$SS_{cor} = \sum_{l=1}^{g} n_l \sum_{j=1}^{n_l}(\bar{x}_{lj} - \bar{x})^2$	$\sum_{l=1}^{g} n_l - 1$

F-test

The usual F-test rejects $H_0 : \tau_1 = \tau_2 = ... = \tau_g = 0$ at level α if

$$F = \frac{SS_{tr}/(g-1)}{SS_{res}/(\sum_{l=1}^{g} n_l - g)} > F_{g-1, \sum n_l - g}(\alpha)$$

where $F_{g-1, \sum n_l - g}(\alpha)$ is the upper (100α)th percentile of the F-distribution with $g-1$ and $\sum_{l=1}^{g} n_l - g$ degrees of freedom. This is equivalent to rejecting H_0 for large values of SS_{tr}/SS_{res} or for large values of $1 + SS_{tr}/SS_{res}$. The statistic appropriate for a multivariate generalization rejects H_0 for **small** values of the reciprocal

$$\frac{1}{1 + SS_{tr}/SS_{res}} = \frac{SS_{res}}{SS_{res} + SS_{tr}}$$

Example

Using the information in our previous example, we have the following ANOVA table (use $\alpha = 0.01$ to test H_0).

Source of of variation	Sum of Squares (SS)	Degrees of freedom (d.f.)
Treatments	$SS_{tr} = 78$	$g - 1 = 3 - 1 = 2$
Residuals (Error)	$SS_{res} = 10$	$\sum_{l=1}^{g} n_l - g = 8 - 3 = 5$
Total (corrected)	$SS_{cor} = 88$	$\sum_{l=1}^{g} n_l - 1 = 8 - 1 = 7$

F statistic

Consequently

$$F = \frac{78/2}{10/5} = \frac{39}{2} = 19.5$$

Since $F = 19.5 > F_{2,5}(0.01) = 13.27$, we reject $H_0 : \tau_1 = \tau_2 = \tau_3 = 0$ (no treatment effect) at 1% level of significance.

MANOVA Table

Source of of variation	Matrix of sum of squares (SS)	d.f.
Treat	$\mathbf{B} = \sum_{l=1}^{g} n_l (\bar{\mathbf{x}}_{\mathbf{l}} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{\mathbf{l}} - \bar{\mathbf{x}})'$	$g - 1$
Res (Error)	$\mathbf{W} = \sum_{l=1}^{g} n_l \sum_{j=1}^{n_l} (\mathbf{x}_{\mathbf{lj}} - \bar{\mathbf{x}}_{\mathbf{l}})(\mathbf{x}_{\mathbf{lj}} - \bar{\mathbf{x}}_{\mathbf{l}})'$	$\sum_{l=1}^{g} n_l - g$
Total (corrected)	$\mathbf{B} + \mathbf{W} = \sum_{l=1}^{g} n_l \sum_{j=1}^{n_l} (\bar{\mathbf{x}}_{\mathbf{lj}} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{\mathbf{lj}} - \bar{\mathbf{x}})'$	$\sum_{l=1}^{g} n_l - 1$

Example. Manova table and Wilk's lambda

Suppose an additional variable is observed along with the variable introduced in our previous example. The sample sizes are $n_1 = 3$, $n_2 = 2$, and $n_3 = 3$. Arranging the observation pairs $\mathbf{x}_{lj}$ in rows, we obtain

Population 1:

$$\begin{pmatrix} 9 \\ 3 \end{pmatrix}$$

$$\begin{pmatrix} 6 \\ 2 \end{pmatrix}$$

$$\begin{pmatrix} 9 \\ 7 \end{pmatrix}$$

Example

Population 2:

$$\begin{pmatrix} 0 \\ 4 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

Example

Population 3:

$$\begin{pmatrix} 3 \\ 8 \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ 9 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 7 \end{pmatrix}$$

Example

We have already expressed the observations on the first variable as the sum of an overall mean, treatment effect, and residual in our discussion of univariate ANOVA. We found that

$SS_{obs} = SS_{mean} + SS_{tr} + SS_{res}$

$216 = 128 + 78 + 10$

Total SS (corrected) $= SS_{obs} - SS_{mean} = 216 - 128 = 88$

Example

Repeating this operation for the observations on the second variable, we have

$$\begin{pmatrix} 3 & 2 & 7 \\ 4 & 0 & \\ 8 & 9 & 7 \end{pmatrix}$$

(**observation**)

Example

$$\begin{pmatrix} 5 & 5 & 5 \\ 5 & 5 & \\ 5 & 5 & 5 \end{pmatrix}$$

(mean)

Example

$$\begin{pmatrix} -1 & -1 & -1 \\ -3 & -3 & \\ 3 & 3 & 3 \end{pmatrix}$$

(treatment effect)

Example

$$\begin{pmatrix} -1 & -2 & 3 \\ 2 & -2 & \\ 0 & 1 & -1 \end{pmatrix}$$

(residual)

Example

$SS_{obs} = SS_{mean} + SS_{tr} + SS_{res}$
$272 = 200 + 48 + 24$
Total SS (corrected) $= SS_{obs} - SS_{mean} = 272 - 200 = 72$

Example

These two single-component analyses must be augmented with the sum of entry-by-entry **cross products** in order to complete the entries in the MANOVA table. Proceeding row by row in the arrays for the two variables, we obtain the cross product contributions:

Mean: $4(5) + 4(5) + ... + 4(5) = 8(4)(5) = 160$

Treatment: $3(4)(-1) + 2(-3)(-3) + 3(-2)(3) = -12$

Residual: $1(-1) + (-2)(-2) + 1(3) + (-1)(2) + ... + 0(-1) = 1$

Total: $9(3) + 6(2) + 9(7) + 0(4) + ... + 2(7) = 149$

Total (corrected) cross product = total cross product - mean cross product $= 149 - 160 = -11$.

MANOVA table

Source of of variation	Matrix of sum of squares and cross products	d.f.
Treatment	$\mathbf{B} = \begin{pmatrix} 78 & -12 \\ -12 & 48 \end{pmatrix}$	3-1 = 2
Residual	$\mathbf{W} = \begin{pmatrix} 10 & 1 \\ 1 & 24 \end{pmatrix}$	8 - 3 = 5
Total (corrected)	$\begin{pmatrix} 88 & -11 \\ -11 & 72 \end{pmatrix}$	7

Wilk's lambda

One test of $H_0 : \tau_1 = \tau_2 = ... = \tau_p = \mathbf{0}$ involves generalized variances. We reject H_0 if the ratio of generalized variances

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B} + \mathbf{W}|}$$

is too small. The quantity Λ^*, proposed originally by Wilks, corresponds to the equivalent form of the F-test of H_0: no treatment effects in the univariate case. The exact distribution of Λ^* can be derived for special cases listed in Table 6.1 (page 176).

Example (Wilk's lambda)

Using Λ^*, we get

$$\Lambda^* = \frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|} = \frac{10(24)-(1)^2}{88(72)-(-11)^2} = \frac{239}{6215} = 0.0385$$

Since $p = 2$ and $g = 3$, Table 6.1 indicates that an exact test is available. To carry out the test, we compare the test statistic

$$\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right)\frac{(\sum n_l - g - 1)}{(g-1)} = \left(\frac{1-\sqrt{0.0385}}{\sqrt{0.0385}}\right)\left(\frac{8-3-1}{3-1}\right) = 8.19$$

with a percentage point of an F-distribution having $\nu_1 = 2(g-1) = 4$ and $\nu_2 = 2(\sum n_l - g - 1) = 8$ d.f. Since $8.19 > F_{4,8}(0.01) = 7.01$, we reject H_0 at the $\alpha = 0.01$ level and conclude that treatment differences exist.

R code

```
## entering measurements

y1<-c(9,6,9,0,2,3,1,2)
y2<-c(3,2,7,4,0,8,9,7)

## combining measurements

Y<-cbind(y1,y2)

Y

## creating covariate

group<-factor(c(rep(1,3),rep(2,2),rep(3,3)) )
```

```
group

## fitting MANOVA

fit<-manova(Y~group)

## showing MANOVA table

summary(fit,test="Wilks")

## matrix W and matrix B

sum.wilks<-summary(fit,test="Wilks")

sum.wilks

## B
```

```
sum.wilks$SS[1]
```

```
## W
```

```
sum.wilks$SS[2]
```

Roy's Test

Please, see Tutorial 4.

Pillai and Lawley-Hotelling Tests

Please, see Tutorial 4.