Statistics 135

Chapter 6

Manova

Chris Drake
Department of Statistics
University of California, Davis

Analysis of Variance

Analysis of variance deals with the comparison of two or more (k) means.

1 The model is

$$X_{lj} = \mu + \tau_l + \epsilon_{lj} \qquad \epsilon_{lj} \sim N(0, \sigma^2)$$

for $l = 1, .., g$ and $j = 1, .., n_l$

2 The null hypothesis is

$$H_0: \ \tau_1 = \tau_2 = ... = \tau_g = 0 \qquad vs \qquad H_1: \ \tau_l \neq 0 \text{at least one } l$$

3 The test statistic is $F = MS_{treatments}/MS_{residual} \sim F_{g-1,n-g}$, where $n = \sum_{l=1}^{g} n_l$.

4 The mean squares treatments is given by

$$MS_{treatments} = \frac{\sum_{l=1}^{g} n_l(\bar{X}_l - \bar{X})^2}{g-1}$$

5 The mean squares residual is given by

$$MS_{error} = \frac{\sum_{l=1}^{g} \sum_{j=1}^{n_l} (X_{lj} - \bar{X}_l)^2}{n - g}$$

The results are summarized in an ANOVA table as follows:

Source	df	SS	MS	F
treatments	g-1	$\sum_{l=1}^{g} n_l(\bar{X}_l - \bar{X})^2$	$\frac{SS_{treat}}{g-1}$	$\frac{MS_{treat}}{MS_{residual}}$
residual	$n - g$	$\sum_{l=1}^{g}(n_g - 1)s_l^2$	$\frac{SS_{residual}}{n-g}$	
Total	$n - 1$	$\sum_{l,j=1}(X_{lj} - \bar{X})^2$		

MANOVA

Sample		Sample Statistics	
pop 1	$\mathbf{x}_{11}, \mathbf{x}_{12} ... \mathbf{x}_{1n_1}$	$\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} \mathbf{x}_{1j}$	$\mathbf{S}_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \bar{\mathbf{x}}_\mathbf{1})(\mathbf{x}_{1j} - \bar{\mathbf{x}}_\mathbf{1})'$
pop 2	$\mathbf{x}_{21}, \mathbf{x}_{22} ... \mathbf{x}_{2n_2}$	$\bar{\mathbf{x}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \mathbf{x}_{2j}$	$\mathbf{S}_2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\mathbf{x}_{2j} - \bar{\mathbf{x}}_\mathbf{2})(\mathbf{x}_{2j} - \bar{\mathbf{x}}_\mathbf{2})'$
.	.	.	
.	.	.	
pop g	$\mathbf{x}_{g1}, \mathbf{x}_{g2} ... \mathbf{x}_{gn_g}$	$\bar{\mathbf{x}}_g = \frac{1}{n_g}\sum_{j=1}^{n_g} \mathbf{x}_{gj}$	$\mathbf{S}_g = \frac{1}{n_g-1}\sum_{j=1}^{n_g}(\mathbf{x}_{gj} - \bar{\mathbf{x}}_\mathbf{g})(\mathbf{x}_{gj} - \bar{\mathbf{x}}_\mathbf{g})'$

Assumptions

1 $\mathbf{x}_{l1}, \mathbf{x}_{l2} ... \mathbf{x}_{ln_l} \sim N_p(\mu_l, \mathbf{\Sigma})$ for $l = 1, .., g$; populations being sampled are normal with different means and common covariance matrix.

2 $\mathbf{x}_{lj}$ and $\mathbf{x}_{kj'}$ are independent for any $j \neq j'$ and $l \neq k$. Observations from different samples are independent, observation vectors from the l^{th} sample are independent but variables observed on the same experimental unit are correlated with covariance matrix $\mathbf{\Sigma}$.

3 The model we assume is

$$\mathbf{X}_{lj} = \mu_l + \mathbf{e}_{lj} = \mu + (\mu_l - \mu) + \mathbf{e}_{lj} = \mu + \tau_l + \mathbf{e}_{lj} \quad for \quad l = 1, .., g$$

where τ_l is the l^{th} population effect, μ is the overall mean and $\mathbf{e}_{lj}$ is the random error vector.

4 The hypothesis to be tested is

$$H_0 : \ \tau_1 = \tau_2 = ... = \tau_g = 0 \quad vs \quad H_1 : \ \tau_l \neq 0 \text{ for at least some } l$$

5 To define the parameters uniquely, we need to impose restrictions, typically $\sum_{l=1}^{g} n_l \tau_l = 0$.

6 Decompose an observation vector $\mathbf{x}_{lj}$ as follows

$$\begin{array}{rcccccc} \mathbf{x}_{lj} & = & \bar{\mathbf{x}} & + & (\bar{\mathbf{x}}_l - \bar{\mathbf{x}}) & + & (\bar{\mathbf{x}}_{lj} - \bar{\mathbf{x}}_l) \\ observation & = & \text{overall} & + & \text{treatment} & + & \text{residual} \\ & & \text{mean} & & \text{effect} & & \end{array}$$

7 Subtract $\bar{\mathbf{x}}$ from both sides and multiply the column vectors on both sides by $(\mathbf{x}_{lj} - \bar{\mathbf{x}})'$ (note: this is equivalent to squaring in the univariate case) and sum over all observations, to get

$$\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{lj} - \bar{\mathbf{x}})(\mathbf{x}_{lj} - \bar{\mathbf{x}})' = \sum_{l=1}^{g} n_l(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})' + \sum_{l=1}^{g}\sum_{j=1}^{n_l}(\bar{\mathbf{x}}_{lj} - \bar{\mathbf{x}}_l)(\bar{\mathbf{x}}_{lj} - \bar{\mathbf{x}}_l)'$$

The first sum is the total corrected sum of squares and cross products, the second sum is between treatments sum of squares and cross products and the third sum is the residual sum of squares and cross products; it can be shown that it is a weighted average of the within each sample estimated covariance matrix.

MANOVA table:

Source of variation	Matrix of sum of squares and Cross products	Degrees of freedom (df)
treatment	$\mathbf{B}=\sum_{l=1}^{g} n_l(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})(\bar{\mathbf{x}}_l - \bar{\mathbf{x}})'$	$g-1$
residual	$\mathbf{W}=\sum_{l=1}^{g}\sum_{j=1}^{n_l}(\bar{\mathbf{x}}_{lj} - \bar{\mathbf{x}}_l)(\bar{\mathbf{x}}_{lj} - \bar{\mathbf{x}}_l)'$	$\sum_{l=1}^{g} n_l - g$
Total	$\mathbf{T} = \sum_{l=1}^{g}\sum_{j=1}^{n_l}(\mathbf{x}_{lj} - \bar{\mathbf{x}})(\mathbf{x}_{lj} - \bar{\mathbf{x}})'$	$\sum_{l=1}^{g} n_l - 1$

8 A test statistic for H_0 is given by

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

This statistic is Wilks' Lambda. A transformation of Λ^* is a multiple of an F-distribution. The exact transformation depends on the number of groups to be compared and the dimension of $\mathbf{X}$.

Two-Way Analysis of Variance

Analysis of the effects of two factors on a response variable.

1 The model is

$$X_{lkr} = \mu + \tau_l + \beta_k + \gamma_{lk} + \epsilon_{lkr} \qquad \epsilon_{lkr} \sim N(0, \sigma^2)$$

for $l = 1, .., g$, $k = 1, .., b$ and $r = 1, .., n$;

2 The design assumes equal number of replicates for each of the $g \times b$ factor combinations

3 The null hypotheses are

$$\begin{array}{rlcl} H_{01}: & \tau_1 = \tau_2 = ... = \tau_g = 0 & vs & H_{11}: \ \tau_l \neq 0 \ \text{at least one } l \\ H_{02}: & \beta_1 = \beta_2 = ... = \beta_b = 0 & vs & H_{12}: \ \beta_k \neq 0 \ \text{at least one } k \\ H_{03}: & \gamma_{11} = \gamma_{12} = ... = \gamma_{gb} = 0 & vs & H_{13}: \ \gamma_{lk} \neq 0 \ \text{some } (l, k) \end{array}$$

4 The hypotheses should be tested in the order H_{03} followed by either H_{01} or H_{02} if the hypothesis of no interaction is not rejected.

The results are summarized in an ANOVA table as follows:

Source	df	SS
Factor 1	g-1	$\sum_{l=1}^{g} bn(\bar{X}_{l\cdot} - \bar{X})^2$
Factor 2	b-1	$\sum_{k=1}^{b} gn(\bar{X}_{\cdot k} - \bar{X})^2$
Interaction	$(g-1) \times (b-1)$	$\sum_{l,k}(\bar{X}_{lk} - \bar{X}_{l\cdot} - \bar{X}_{\cdot k} + \bar{X})^2$
residual	$gb(n-1)$	$\sum_{l=1}^{g}\sum_{k=1}^{b}\sum_{r=1}^{n}(X_{lkr} - \bar{X}_{lk})^2$
Total	$gbn - 1$	$\sum_{l,j=1}(X_{lj} - \bar{X})^2$

The ratios of mean squares for factors 1, 2 or interaction divided by the mean square residual are used to test for H_{01}, H_{02} and H_{03}.

Two-Way MANOVA

Assumptions

1 $\mathbf{x}_{lkr} \sim N_p(\mu_{lkr}, \mathbf{\Sigma})$ for $l = 1, .., g$, $k = 1, .., b$ and $r = 1, .., n$; populations being sampled are normal with different means and common covariance matrix.

2 $\mathbf{x}_{lkr}$ and $\mathbf{x}_{l',k',r'}$ are independent for any $l \neq l'$ and $k \neq k'$. Observations at different levels of the same factor are independent, observation vectors at the same level of a factor are independent but variables observed on the same experimental unit are correlated with covariance matrix $\mathbf{\Sigma}$.

3 The model we assume is

$$\mathbf{X}_{lkr} = \mu + \tau_l + \beta_k + \gamma_{lk} + \mathbf{e}_{lkr}$$

for $l = 1, .., g$, $k = 1, .., b$ and $r = 1, .., n$ where τ_l is the l^{th} factor 1 effect, β_k is the kth factor 2 effect and γ_{lk} is the interaction between factors 1 and 2; μ is the overall mean and $\mathbf{e}_{lkr}$ is the random error vector.

4 The hypotheses to be tested are the same as in the univariate case.

MANOVA table:

Source of variation	Matrix of sum of squares and Cross products	Degrees of freedom (df)
Factor 1	$\mathbf{SSP}_{fac1} = \sum_{l=1}^{g} bn(\bar{\mathbf{x}}_{l\cdot} - \bar{x})(\bar{\mathbf{x}}_{l\cdot} - \bar{x})'$	$g-1$
Factor 2	$\mathbf{SSP}_{fac2} = \sum_{l=1}^{b} gn(\bar{\mathbf{x}}_{\cdot k} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{\cdot k} - \bar{\mathbf{x}})'$	$b-1$
Interaction	$\mathbf{SSP}_{int} = \sum_{l,k} n(\bar{\mathbf{x}}_{lk} - \bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}}_{\cdot k} + \bar{\mathbf{x}})(\bar{\mathbf{x}}_{lk} - \bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}}_{\cdot k} + \bar{\mathbf{x}})'$	$(g-1)(b-1)$
residual	$\mathbf{SSP}_{res} = \sum_{l,k,r}(\mathbf{x}_{lkr} - \bar{\mathbf{x}}_{lk})(\mathbf{x}_{lkr} - \bar{\mathbf{x}}_{lk})'$	$gb(n-1)$
Total	$\mathbf{T} = \sum_{lkr}(\mathbf{x}_{lkr} - \bar{\mathbf{x}})(\mathbf{x}_{lkr} - \bar{\mathbf{x}})'$	$gbn-1$

5 The hypothesis of no interaction $H_{03}: \ \gamma_{11} = \gamma_{12} = ... = \gamma_{gb} = 0$ is rejected for small Wilks' lambda

$$\Lambda^* = \frac{\mid \mathbf{SSP}_{res} \mid}{\mid \mathbf{SSP}_{int} + \mathbf{SSP}_{res} \mid}$$

for large samples a χ^2 approximation can be used as follows:

$$-\left[gb(n-1) - \frac{p+1-(g-1)(b-1)}{2}\right] ln\Lambda^* > \chi^2_{(g-1)(b-1)p}(\alpha)$$

6 The hypothesis for main effects, if H_{03} is not rejected, can proceed in either order. Usually, $H_{01}: \ \tau_1 = \tau_2 = .. = \tau_g = 0$ is tested first with

$$\Lambda^* = \frac{\mid \mathbf{SSP}_{res} \mid}{\mid \mathbf{SSP}_{fac1} + \mathbf{SSP}_{res} \mid}$$

for large samples a χ^2 approximation can be used as follows:

$$-\left[gb(n-1) - \frac{p+1-(g-1)}{2}\right] ln\Lambda^* > \chi^2_{(g-1)p}(\alpha)$$

7 The hypothesis for the second main effect, $H_{02}: \ \beta_1 = \beta_2 = .. = \beta_b = 0$ is tested with

$$\Lambda^* = \frac{|\ \mathbf{SSP}_{res}\ |}{|\ \mathbf{SSP}_{fac2} + \mathbf{SSP}_{res}\ |}$$

for large samples a χ^2 approximation can be used as follows:

$$-\left[gb(n-1) - \frac{p+1-(b-1)}{2}\right] ln\Lambda^* > \chi^2_{(b-1)p}(\alpha)$$

8 For $\mathbf{SSP}_{res}$ to be positive definite it is necessary that $p \leq gb(n-1)$.

9 $(1-\alpha)$ simultaneous confidence intervals for $\tau_{li} - \tau_{mi}$ are given by

$$(\bar{x}_{l\cdot i} - x_{m\cdot i}) \pm t_\nu \left(\frac{\alpha}{pg(n-1)}\right) \sqrt{\frac{E_{ii}}{\nu}\frac{2}{bn}}$$

where $\nu = gb(n-1)$, E_{ii} is the i^{th} diagonal element of $\mathbf{E} = \mathbf{SSP}_{res}$ and $\bar{x}_{l\cdot i} - \bar{x}_{m\cdot i}$ is the i^{th} component of $\bar{\mathbf{x}}_{l\cdot} - \bar{\mathbf{x}}_{m\cdot}$.

Other Test Statistics

1 Bartlett-Lawley-Hotelling trace, on SAS output it is Lawley-Hotellin Trace

$$Lawley - Hotelling \;\; = \;\; tr(BW^{-1})$$

with critical value

$$c^2(\alpha) \approx \nu_e \left[\frac{s\nu_1}{\nu_2} \right] F_{\nu-1,\nu_2}(\alpha)$$

a ν_e are the residual df

b $\nu_1 = s(2M + s + 1)$

c $\nu_2 = 2(sN + 1)$

d $s = min(\nu_h, p)$ where ν_h are the degrees of freedom for the particular hypothesis (factor 1, factor 2 or interaction for a two-way Manova.

e $M = (\mid \nu_h - p \mid -1)/2$

f $N = (\nu_e - p - 1)/2$

2 Bartlett, Nanda, Pillai trace criterion, on SAS output it is Pillai's Trace

$$Pillai's\ Trace\ = tr[\mathbf{B}(\mathbf{B}+\mathbf{W})^{-1}]$$

with critical value

$$c^2 \approx \nu_e \frac{V(\alpha)}{1-V(\alpha)} \qquad V(\alpha) = \frac{\frac{s\nu_1}{\nu_2} F_{\nu_1,\nu_2}(\alpha)}{(1+\frac{\nu_1}{\nu_2})F_{\nu_1,\nu_2}(\alpha)}$$

and $\nu_1 = s(2M+s+1)$, $\nu_2 = s(2N+s+1)$, M,N as before.

3 Roy's maximum root criterion

$$\lambda_{max} = maximum\ \ eigenvalue\ \ of\ \ \mathbf{W}(\mathbf{B}+\mathbf{W})^{-1}$$

with critical value

$$c^2 \approx \nu_e \left[\frac{\nu_1}{\nu_e - \nu_1 + \nu_h}\right] F_{\nu_1,\nu_e-\nu_1+\nu_h}(\alpha)$$

where $\nu_1 = max(\nu_h, p)$ and ν_h are the df for the respective hypothesis.

Remarks:

1 Here **W** is the cross-product of residuals and **B** is the matrix corresponding to the hypothesis to be tested. For H_{01} this is the cross-product matrix for the interactions from the Manova table and $\nu_h = (g-1)(b-1)$ where g = number of levels of factor A, b = number of levels of factor B.

2 The critical values given here are approximations but presumably somewhat better than the χ^2 approximations given in the textbook for large n, when sample sizes are small. They agree with the SAS output from PROC GLM.

3 When $s = 1$ the test statistics based on the three criteria give the same value and the degrees of freedom for numerator and denominator are the same.

4 Note also, for all models to be identifiable, we usually impose the restrictions $\sum_l \tau_l = \sum_k \beta_k = \sum_{l,k} \gamma_{lk} = 0$.

Profile Analysis

P treatments given to two or more groups; μ_{ik} is the mean response in group k to the i^{th} treatment for $i = 1, .., p$; x_{ijk} is the response for the j^{th} subject in the k^{th} group to the i^{th} treatment. We'll consider the case of 2 groups.

Def: the profile for a group is the means connected by straight lines

There are three hypothesis to be tested.

1 Are the profiles parallel, ie are the line segments connected the means parallel for the two groups?

2 If the profiles are parallel, are they coincident, ie are the levels of the means at each treatment the same?

3 If the profiles are parallel and coincident, are they flat; ie are the treatments the same? If there is a baseline measurement or control measurement, then flat profiles mean there is no treatment effect.

1. Parallel profiles: $\mu_{i,1} - \mu_{i-1,1} = \mu_{i,2} - \mu_{i-1,2}$ for $i = 1, .., p$
The contrast matrix for this hypothesis is

$$C_{(p-1)p} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}$$

and the hypothesis is

$$H_{01}: \ \mathbf{C}\mu_{\mathbf{1}} = \mathbf{C}\mu_{\mathbf{2}}$$

We reject the null hypothesis if

$$\left(\mathbf{C}\bar{\mathbf{x}}_{\mathbf{1}} - \mathbf{C}\bar{\mathbf{x}}_{\mathbf{2}}\right)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \mathbf{C}\mathbf{S}_{\mathbf{pooled}}\mathbf{C}' \right]^{-1} \left(\mathbf{C}\bar{\mathbf{x}}_{\mathbf{1}} - \mathbf{C}\bar{\mathbf{x}}_{\mathbf{2}}\right) > c^2$$

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

2. Coincident lines: $\sum_{i=1}^{p} \mu_{1i} = \sum_{i=1}^{p} \mu_{2i}$

for parallel lines to be coincident, all points for the two groups are at the same height and their sums must be equal. Now $\sum_{i=1}^{p} \mu_{1i} = \mathbf{1}'\mu_1$ and the hypothesis is

$$H_{02}: \ \mathbf{1}'\mu_1 = \mathbf{1}'\mu_2$$

The test is rejected if

$$\begin{aligned} T^2 =& \mathbf{1}'\big(\bar{\mathbf{x}}_\mathbf{1} - \bar{\mathbf{x}}_\mathbf{2}\big)' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \mathbf{1}\mathbf{S}_\mathbf{pooled}\mathbf{1}' \right]^{-1} \mathbf{1}'\big(\bar{\mathbf{x}}_\mathbf{1} - \bar{\mathbf{x}}_\mathbf{2}\big) \\ =& \left(\frac{\mathbf{1}'\big(\bar{\mathbf{x}}_\mathbf{1} - \bar{\mathbf{x}}_\mathbf{2}\big)}{\sqrt{\left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{1}\mathbf{S}_\mathbf{pooled}\mathbf{1}'\right)}} \right)^2 > t^2_{n_1+n_2-2}(\alpha) \end{aligned}$$

3. Level (flat) profiles, $\mu_{k1} = \mu_{k2} = .. = \mu_{kp}$ for $k = 1, 2$

$$H_{03}: \ \mathbf{C}\mu = 0$$

We use the same contrast matrix here as for H_{01} but the hypothesis asks if $\mu_i - \mu_{i-1} = 0$ for $i = 1, .., p$; if H_{01} and H_{02} are not rejected then the two groups have the same mean and the samples can be combined for testing H_{03}. The common mean is estimated as

$$\bar{\mathbf{x}} = \frac{n_1}{n_1 + n_2}\bar{\mathbf{x}}_\mathbf{1} + \frac{n_2}{n_1 + n_2}\bar{\mathbf{x}}_\mathbf{2}$$

and the test is given by

$$(n_1 + n_2)\bar{\mathbf{x}}'\mathbf{C}'\left[\mathbf{CSC}'\right]^{-\mathbf{1}}\mathbf{C}\bar{\mathbf{x}} > \frac{(n_1 + n_2 - 1)(p-1)}{(n_1 + n_2 - p + 1)}F_{p-1,n_1+n_2-p+1}(\alpha)$$