Chapter 6: MANOVA

Multivariate analysis of variance (MANOVA) generalizes ANOVA to allow multivariate responses.

We'll start by reviewing ANOVA (the balanced case), particularly to develop the notation consistent with the MANOVA presentation.

ANOVA review

In balanced one-way ANOVA, there are k samples one from each of k different populations, each with n observations. The populations being sampled might be individuals subjected to different treatments in a medical experiment, crops being given different fertlizer/watering regimes. If it is not an experiment, the different populations might represent different groups, such as different varieties of a crop, or different ethnicities/nationalities for people.

Values from observations within a particular group are denoted by y_{ij}, where $i = 1, \ldots, k$ denotes the sample and $j = 1, \ldots n$ denotes the observation within the sample. Note that this is different notation from the previous chapter where the first index represented the row (observation) and the second index represented the column (variable).

ANOVA review

	Sample 1 from $N(\mu_1, \sigma^2)$	Sample 2 from $N(\mu_2, \sigma^2)$	...	Sample k from $N(\mu_k, \sigma^2)$
	y_{11}	y_{21}	$\cdots$	y_{k1}
	y_{12}	y_{22}	$\cdots$	y_{k2}
	$\vdots$	$\vdots$		$\vdots$
	y_{1n}	y_{2n}	$\cdots$	y_{kn}
Total	$y_{1.}$	$y_{2.}$	$\cdots$	$y_{k.}$
Mean	$\overline{y}_{1.}$	$\overline{y}_{2.}$	$\cdots$	$\overline{y}_{k.}$
Variance	s_1^2	s_2^2	$\cdots$	s_k^2

ANOVA review

The ith group has mean

$$\overline{y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} y_{yij}$$

and total

$$y_{i.} = \sum_{j=1}^{n} y_{yij}$$

The ANOVA model is that each observation is due to an overall mean, a treatment (or population) mean, and an unobserved error term

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij}$$

where $i = 1, \ldots, k$ and $j = 1, \ldots, n$.

ANOVA review

The null hypothesis is $H_0 : \mu_1 = \cdots = \mu_k$, and the alternative is $H_1 : \mu_i \neq \mu_j$ for some $i \neq j$. (alternatively we could express this interms of α_is instead of μ_is). The model assumes that all populations have the same variance σ^2. Assuming this, we wish to test whether the means differ for the different populations.

The basic idea of ANOVA is that if the null hypothesis is true, then the common variance σ^2 can be estimated either by averaging the variances of the separate samples, or by using the sample standard deviation of the sample means.

ANOVA review

The pooled standard deviation is

$$s_e^2 = \frac{1}{k}\sum_{i=1}^{k} s_i^2 = \frac{1}{k(n-1)}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{y}_i)^2$$

The sample standard deviation of the sample means is

$$s_{\overline{y}}^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\overline{y}_{i.} - \overline{y}_{..})^2$$

where $\overline{y}_{..}$ is the mean of the observations over both groups (samples) and observations. You can also think of it as the mean of the sample means

$$\overline{y}_{..} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \frac{1}{k}\sum_{i=1}^{k}\overline{y}_{i.}$$

ANOVA review

Under the null hypothesis and model assumptions, both sample variances are related to σ^2:

$$E(s_e^2) = \sigma^2; \quad E(ns_{\bar{y}}^2) = \sigma^2$$

If the null hypothesis is false (but all populations have the same variance), then it is still the case that $E(s_e^2) = \sigma^2$. However, the variability of $ns_{\bar{y}}^2$ is higher because there is variability due to both the sample means and the variability of the population means themselves. In this case

$$E(ns_{\bar{y}}^2) = \sigma^2 + \frac{n}{k-1}\sum_{i=1}^{k}\alpha_i$$

ANOVA review

The ratio of $ns^2_{\overline{y}}$ and s^2_e has an F distribution under the hypothesis. It is partly feasible to work out the distribution because $ns^2_{\overline{y}}$ and s^2_e are independent random variables (under H_0). This is a consequence of $\overline{y}$ and s^2 being independent for samples from a normal distribution and from the assumption that each of the k samples is independent. The numerator and denominator are therefore each related to χ^2 random variables, and the ratio of χ^2 random variables is related to the F distribution.

ANOVA review

$$F = \frac{ns_{\bar{y}}^2}{s_e^2} = \frac{SSH/(k-1)}{SSE/(k(n-1))} = \frac{MSH}{MSE}$$

has an $F_{k-1,k(n-1)}$ distribution. Note that the expected value of the numerator divided by the expected value of the denominator is equal to 1; however the expected value of a ratio is typically not the ratio of the expected values, and we have

$$E(F_{k-1,k(n-1)}) = \frac{k(n-1)}{k(n-1)-2} = \frac{n-k}{n-k-2}$$

This is close to 1 for large nk when n is much larger than k. Note that the expected value of an F random variable only depends on the denominator degrees of freedom.

ANOVA review

Hypothesis testing is done as a one-sided test, only rejecting H_0 for sufficiently large F. The F distribution is skewed to the right, and the p-value is the area under the curve to the right of the observed F value.

MANOVA

MANOVA generalizes both the Hotelling T^2, which allows two populations with multiple variables on each, and ANOVA, which allows one variable but with two or more populations.

For the MANOVA set up, we have observation vectors $\mathbf{y}_{ij}$ from sample $i = 1, \ldots k$, with $j = 1, \ldots, n$ indexing the observation. Each observation vector $\mathbf{y}_{ij}$ is a p-dimensional multivariate normal vector with mean vector $\boldsymbol{\mu}_i$ and common covariance matrix $\boldsymbol{\Sigma}$. The set up can be written in a way analgous to balanced one-way ANOVA with individual observations replaced with observation vectors.

MANOVA

	Sample 1 from $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$	Sample 2 from $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$	$\cdots$	Sample k from $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$
	$\mathbf{y}_{11}$	$\mathbf{y}_{21}$	$\cdots$	$\mathbf{y}_{k1}$
	$\mathbf{y}_{12}$	$\mathbf{y}_{22}$	$\cdots$	$\mathbf{y}_{k2}$
	$\vdots$	$\vdots$		$\vdots$
	$\mathbf{y}_{1n}$	$\mathbf{y}_{2n}$	$\cdots$	$\mathbf{y}_{kn}$
Total	$\mathbf{y}_{1.}$	$\mathbf{y}_{2.}$	$\cdots$	$\mathbf{y}_{k.}$
Mean	$\overline{\mathbf{y}}_{1.}$	$\overline{\mathbf{y}}_{2.}$	$\cdots$	$\overline{\mathbf{y}}_{k.}$

Totals and means are defined as follows:

Total of the ith sample: $\mathbf{y}_{i.} = \sum_{j=1}^{n} \mathbf{y}_{ij}$.

Overall total: $\mathbf{y}_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n} \mathbf{y}_{ij}$.

Mean of the ith sample: $\overline{\mathbf{y}}_{i.} = \mathbf{y}_{i.}/n$.

Overall mean: $\overline{\mathbf{y}}_{..} = \mathbf{y}_{..}/kn$.

MANOVA

The model can be written as

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij}$$

where we assume

$$\mathbf{y}_{ij} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$$

MANOVA

For $r = 1, \ldots, p$, we can also write the model as

$$\begin{pmatrix} y_{ij1} \\ y_{ij2} \\ \vdots \\ y_{ijr} \end{pmatrix} = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{ir} \end{pmatrix} + \begin{pmatrix} \varepsilon_{ij1} \\ \varepsilon_{ij2} \\ \vdots \\ \varepsilon_{ijr} \end{pmatrix}$$

so that for each variable $r = 1, \ldots, p$, the model is

$$y_{ijr} = \mu_{ir} + \varepsilon_{ijr}$$

The null and alternative hypotheses are

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_k, \quad H_1 : \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \text{ for at least one pair } i \neq j$$

i.e., that each population has the same mean vector and that that at least two populations have different mean vectors, or that at least two populations have at least one variable with different means.

MANOVA

The null hypothesis can be written also as p sets of $k-1$ equalities:

$$\begin{aligned}
\mu_{11} &= \mu_{21} = \cdots = \mu_{k1} \\
\mu_{12} &= \mu_{22} = \cdots = \mu_{k2} \\
\vdots &= \vdots = \quad = \vdots \\
\mu_{1p} &= \mu_{1p} = \cdots = \mu_{kp}
\end{aligned}$$

This is a total of $p(k-1)$ equalities, and any one of these failing is sufficient to make H_0 false.

MANOVA

Analogous to the SSH (Sums of squares hypothesis) and SSE (sums of squares for error), we have

$$\mathbf{H} = n\sum_{i=1}^{k}(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})' = \sum_{i=1}^{k}\frac{1}{n}\mathbf{y}_{i.}\mathbf{y}_{i.}' - \frac{1}{kn}\mathbf{y}_{..}\mathbf{y}_{..}'$$

$$\mathbf{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})' = \sum_{ij}\mathbf{y}_{ij}\mathbf{y}_{ij}' - \frac{1}{n}\sum_{i}\mathbf{y}_{i.}\mathbf{y}_{i.}'$$

MANOVA

The **E** and **H** matrices are both $p \times p$, but not necessarily full rank. The rank of **H** is $\min(p, v_H)$, where v_H is the degrees of freedom associated with the hypothesis, i.e. $k - 1$.
We can think the pooled covariance matrix as

$$S_{\text{pl}} = \frac{\mathbf{E}}{(n-1)k}$$

with

$$E(S_{\text{pl}}) = \mathbf{\Sigma}$$

However, if the sample mean vectors were equal for each population, then we would have $\mathbf{H} = \mathbf{0}$.

Thus $\mathbf{H}$ has the form

$$\mathbf{H} = \begin{pmatrix} \mathrm{SSH}_{11} & \mathrm{SPH}_{12} & \cdots & \mathrm{SPH}_{1p} \\ \mathrm{SPH}_{12} & \mathrm{SSH}_{22} & \cdots & \mathrm{SPH}_{2p} \\ \vdots & \vdots & & \vdots \\ \mathrm{SPH}_{1p} & \mathrm{SPH}_{2p} & \cdots & \mathrm{SSH}_{pp} \end{pmatrix},$$

where, for example,

$$\mathrm{SSH}_{22} = n \sum_{i=1}^{k} (\overline{y}_{i.2} - \overline{y}_{..2})^2 = \sum_{i} \frac{y_{i.2}^2}{n} - \frac{y_{..2}^2}{kn},$$

$$\mathrm{SPH}_{12} = n \sum_{i=1}^{k} (\overline{y}_{i.1} - \overline{y}_{..1})(\overline{y}_{i.2} - \overline{y}_{..2}) = \sum_{i} \frac{y_{i.1} y_{i.2}}{n} - \frac{y_{..1} y_{..2}}{kn}.$$

MANOVA

$$\mathbf{E} = \begin{pmatrix} \text{SSE}_{11} & \text{SPE}_{12} & \cdots & \text{SPE}_{1p} \\ \text{SPE}_{12} & \text{SSE}_{22} & \cdots & \text{SPE}_{2p} \\ \vdots & \vdots & & \vdots \\ \text{SPE}_{1p} & \text{SPE}_{2p} & \cdots & \text{SSE}_{pp} \end{pmatrix},$$

where, for example,

$$\text{SSE}_{22} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij2} - \overline{y}_{i.2})^2 = \sum_{ij} y_{ij2}^2 - \sum_{i}\frac{y_{i.2}^2}{n},$$

$$\text{SPE}_{12} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij1} - \overline{y}_{i.1})(y_{ij2} - \overline{y}_{i.2}) = \sum_{ij} y_{ij1}y_{ij2} - \sum_{i}\frac{y_{i.1}y_{i.2}}{n}$$

MANOVA

The **E** and **H** matrices can be used in different ways to test the null hypothesis. Wilks' Test Statistic is

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|}$$

The null is rejected if $\Lambda < \Lambda_{\alpha,p,v_H,v_E}$ where v_H is the degrees of freedom for the hypothesis, $k-1$, and v_E is degrees of freedom for error, $k(n-1)$. Critical values are in Table A9. The test statistic can instead be converted to an F, but there are different cases.

Table 6.1. Transformations of Wilks' Λ to Exact Upper Tail F-Tests

Parameters p, ν_H	Statistic Having F-Distribution	Degrees of Freedom
Any $p, \nu_H = 1$	$\frac{1-\Lambda}{\Lambda}\frac{\nu_E - p + 1}{p}$	$p, \nu_E - p + 1$
Any $p, \nu_H = 2$	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{\nu_E - p + 1}{p}$	$2p, 2(\nu_E - p + 1)$
$p = 1$, any ν_H	$\frac{1-\Lambda}{\Lambda}\frac{\nu_E}{\nu_H}$	ν_H, ν_E
$p = 2$, any ν_H	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{\nu_E - 1}{\nu_H}$	$2\nu_H, 2(\nu_E - 1)$

Properties of Wilks' Λ

- We need $v_E = (n-1)k \geq p$ for the determinants to be positive
- The degrees of freedom for error and hypothesis are the same as for univariate ANOVA
- The distribution of Λ_{p,v_H,v_E} is the same as Λ_{p,v_E,v_H}. This saves some space for the table of critical values.
- Wilks' Λ can be written as

$$\Lambda = \prod_{i=1}^{\min(p,v_H)} \frac{1}{1+\lambda_i}$$

where λ_i is the ith eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$. Here $s = \min(p, v_H)$ is the rank of s, which is also the number of nonzero eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$.

- Λ is in the interval [0,1]. If the sample mean vectors were all equal (for example, if they were all equal to their expected values under the null), then $\mathbf{H} = \mathbf{0}$.

Properties of Wilks' Λ

- Increasing the number of variables p decreases the critical value for Λ needed to reject the null hypothesis. This means that it is more difficult to reject H_0 (since we reject for small Λ) unless the null hypothesis is false for the new variables. I.e., adding new variables for which the populations are equal makes it harder to reject the null hypothesis.
- When $v_H = 1, 2$ or $p = 1, 2$, Wilks' Λ is equivalent to an F statistic. Otherwise, an approximate transformation to an F can be used:

Properties of Wilks' Λ

For values of p and ν_H other than those in Table 6.1, an approximate F-statistic is given by

$$F = \frac{1 - \Lambda^{1/t}}{\Lambda^{1/t}} \frac{\mathrm{df}_2}{\mathrm{df}_1}, \tag{6.15}$$

with df_1 and df_2 degrees of freedom, where

$$\mathrm{df}_1 = p\nu_H, \qquad \mathrm{df}_2 = wt - \frac{1}{2}(p\nu_H - 2),$$

$$w = \nu_E + \nu_H - \frac{1}{2}(p + \nu_H + 1), \qquad t = \sqrt{\frac{p^2\nu_H^2 - 4}{p^2 + \nu_H^2 - 5}}.$$

When $p\nu_H = 2$, t is set equal to 1. The approximate F in (6.15) reduces to the exact F-values given in Table 6.1, when either ν_H or p is 1 or 2.

Properties of Wilks' Λ

If the null hypothesis is rejected, then follow up tests could be made. Fixing $r \in \{1, \ldots, p\}$, one could test

$$H_{0r} : \mu_{1r} = \mu_{2r} = \cdots = \mu_{kr}$$

which would be a univariate ANOVA test to see if the k populations differ on variable r.

As usual, testing all variables simultaneously and then testing individual variables has better type I error than just testing all variables separately to begin with. It is also possible that the simultaneous test rejects H_0 but that each H_{0r} for $r = 1, \ldots, p$ fails to be rejected.

Example where Wilks' Λ rejects but individual ANOVAs don't reject

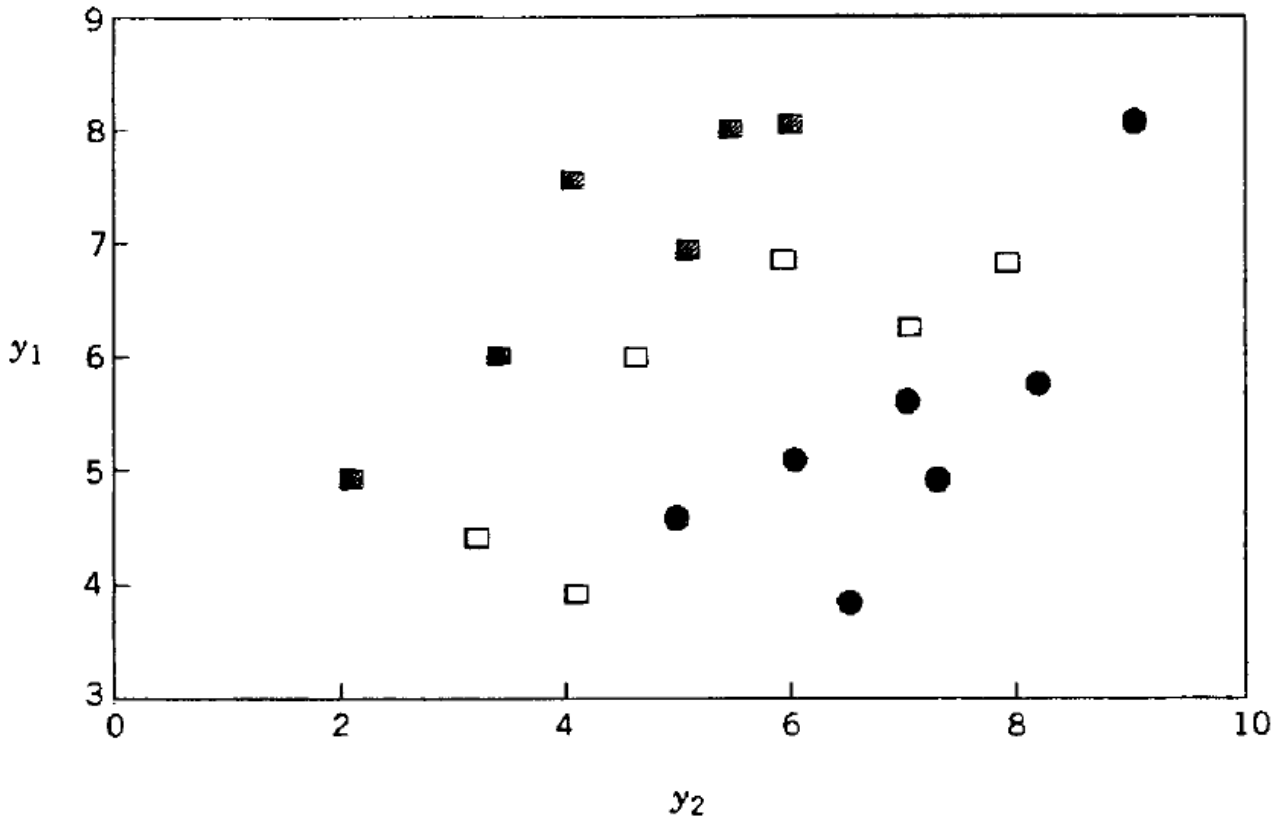


Figure 6.1. Three samples with significant Wilks' Λ but nonsignificant F's.

Other statistics

There are alternatives to Wilks' Λ, but my impression is that Wilks' Λ is the most widely used. Common alternatives are

- Hotelling's Trace statistic, $\text{tr}(\mathbf{E}^{-1}\mathbf{H}) = \sum_{i=1}^{s} \lambda_i$ considered more liberal than Wilks' Λ
- Pillai's Trace statistic:

$$\text{tr}[(\mathbf{E}+\mathbf{H})^{-1})\mathbf{H}] = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i}$$

  considered more conservative than Wilks' Λ
- Roy's largest root:

$$\frac{\lambda_1}{1+\lambda_1}$$

  uses the variance from the variable that separates the group most based on the largest eigenvalue.

Other statistics

For our purposes, we can just use Wilks' Λ, but it is good to be aware of other statistics if they are output from software. These generally can be related to an F distribution, except Roy's largest root test, which is just bounded by an F statistic. In other words, the F statistic bounding Roy's largest root test essentially gives a lower bound on the p-value, so that if this bound is above α, then you can safely not reject H_0, but if the bound is below α, then it is not clear whether you should reject (based on the F alone).

Chile example

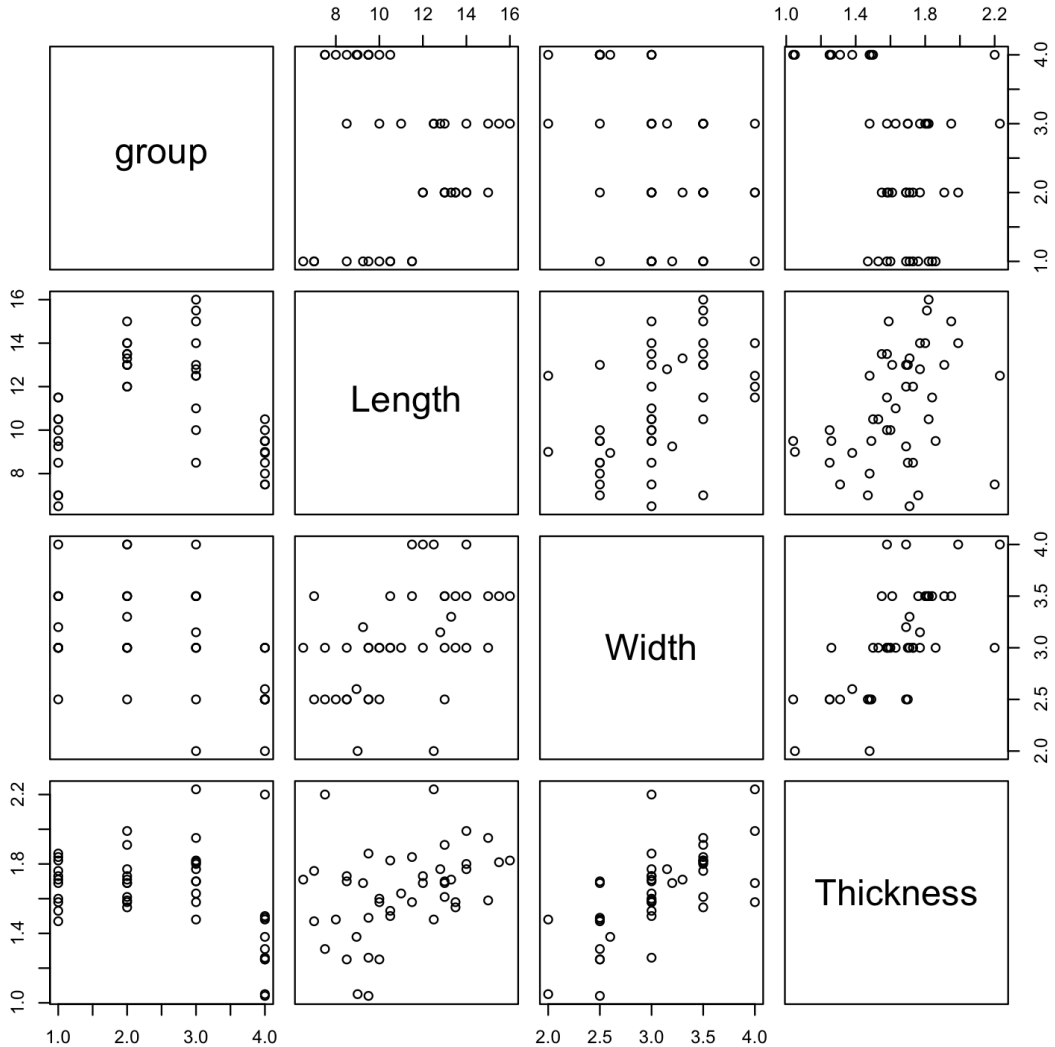
As an example, we'll use part of a data set on chile varieties grown in New Mexico. The variables included here are length, width, and thickness for individual chile pods randomly selected from three varieties: Alcalde, Casados, Chimayo, and Cochiti. The question is whether the chile pods differ in any of the variables at the four locations.

As a first step, we might try to plot the data. The R code is (assuming that the file is in your working directory for R):

```
> y <- read.table("chile.txt",header=T)
> plot(y)
```

```
> x
      group Length Width Thickness
1   Alcalde  10.50  3.00      1.53
2   Alcalde   7.00  3.50      1.76
3   Alcalde  10.50  3.50      1.82
4   Alcalde  11.50  4.00      1.58
5   Alcalde  11.50  3.50      1.84
6   Alcalde   9.50  3.00      1.86
7   Alcalde   6.50  3.00      1.71
8   Alcalde   8.50  3.00      1.73
9   Alcalde  10.00  3.00      1.60
10  Alcalde   7.00  2.50      1.47
11  Alcalde   9.25  3.20      1.69
12 Casados  12.00  3.00      1.73
13 Casados  12.00  4.00      1.69
14 Casados  13.50  3.50      1.55
15 Casados  14.00  3.00      1.77
16 Casados  15.00  3.00      1.59
17 Casados  13.00  3.50      1.61
18 Casados  13.50  3.00      1.58
```

Chile example



Chile example

To do MANOVA for this example without relying on built-in procedures, we need to construct the **H** and **E** matrices. Recall that

$$\mathbf{H} = n\sum_{i=1}^{k}(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})' = \sum_{i=1}^{k}\frac{1}{n}\mathbf{y}_{i.}\mathbf{y}_{i.}' - \frac{1}{kn}\mathbf{y}_{..}\mathbf{y}_{..}'$$

$$\mathbf{E} = \sum_{i=1}^{k}\sum_{j=1}^{n}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i.})' = \sum_{ij}\mathbf{y}_{ij}\mathbf{y}_{ij}' - \frac{1}{n}\sum_{i}\mathbf{y}_{i.}\mathbf{y}_{i.}'$$

Chile example

First, It will be convenient to relabel the groups as 1, 2, 3, and 4.

```
> y$group = 1*(y$group=="Alcalde") + 2*(y$group=="Casados")
+ 3*(y$group=="Chimayo") + 4*(y$group=="Cochiti")
> y
   group Length Width Thickness
1      1  10.50  3.00      1.53
2      1   7.00  3.50      1.76
3      1  10.50  3.50      1.82
4      1  11.50  4.00      1.58
5      1  11.50  3.50      1.84
6      1   9.50  3.00      1.86
7      1   6.50  3.00      1.71
8      1   8.50  3.00      1.73
9      1  10.00  3.00      1.60
10     1   7.00  2.50      1.47
11     1   9.25  3.20      1.69
```

Chile example

For this data, $k = 4$ and $p = 3$.

We need to define $\overline{\mathbf{y}}_{i.}$ for $i = 1, 2, 3, 4$. Note that $\overline{\mathbf{y}}_{i.}$ is a vector of length 3 because of the three variables. As an example, $\overline{\mathbf{y}}_{4.}$ represents the average length, width, and thickness for the Cochiti Peublo green chiles.

Chile example

```
> y1. <- colMeans(y[y$group==1,2:4])
> y2. <- colMeans(y[y$group==2,2:4])
> y3. <- colMeans(y[y$group==3,2:4])
> y4. <- colMeans(y[y$group==4,2:4])
> y1.
   Length     Width Thickness
     9.25      3.20      1.69
> y2.
   Length     Width Thickness
13.300000  3.300000  1.710909
> y.. <- colMeans(y[,2:4])
> y..
   Length     Width Thickness
11.075000  3.062500  1.638409
```

Chile example

```
> H <- 10*((y1.-y..) %*% t(y1.-y..) + (y2.-y..) %*% t(y2.-y..)
(y3.-y..) %*% t(y3.-y..) + (y4.-y..) %*% t(y4.-y..))
> H
         Length     Width Thickness
[1,] 173.49750 15.523750 9.2122500
[2,]  15.52375  3.265625 1.6948750
[3,]   9.21225  1.694875 0.9966795
```

Chile example

```
> E <- matrix(rep(0,9),ncol=3)
> for(i in 1:4) { # using second equation for E
+ for(j in 1:11) {
+ E <- E + Y[(i-1)*11+j,2:4] %*% t(Y[(i-1)*11+j,2:4])
+ }}
> E # still need to subtract some terms
        Length     Width Thickness
[1,] 5673.3950 1518.3300  808.2775
[2,] 1518.3300  423.5625  224.3895
[3,]  808.2775  224.3895  120.8541
> E <- E- 11*(y1. %*% t(y1.) + y2. %*% t(y2.) +
 y3. %*% t(y3.) + y4. %*% t(y4.))
> E
       Length  Width Thickness
[1,] 103.0500 10.450  0.668500
[2,]  10.4500  7.625  1.919000
[3,]   0.6685  1.919  1.744509
```

Chile example

To construct different test statistics, we need the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$. These are

```
> lambda <- eigen(solve(E) %*% H)
> lambda
$values
[1] 2.06508208 0.34193415 0.05798507

$vectors
           [,1]       [,2]        [,3]
[1,] -0.2289322 -0.1022592  0.00809076
[2,]  0.2804863  0.5704725  0.45630477
[3,] -0.9321574  0.8149259 -0.88978677
```

Chile example

Note that in this case, $p = 3$ and $v_H = k - 1 = 3$, so we should get full rank **H** and **E** with three positive eigenvectors. The following are the test statistics:

```
> b <- lambda$values
> b1 <- 1/(1+b)
> prod(b1)
[1] 0.2297985 #Wilks' Lambda
> b2 <- b/(1+b)
> sum(b2)
[1] 0.9833585  #Pillai's trace
> sum(b)
[1] 2.465001 #Hotelling's trace
> b[1]/(1+b[1])
[1] 0.6737445 #Roy's largest root
```

MANOVA in R

```
> a <- manova(Y[,2:4] ~ Y[,1])
> summary(a,test="W")
          Df   Wilks approx F num Df den Df  Pr(>F)
Y[, 1]     1 0.76113   4.1845      3     40 0.01146 *

> summary(a,test="H")
          Df Hotelling-Lawley approx F num Df den Df  Pr(>F)
Y[, 1]     1          0.31384   4.1845      3     40 0.01146 *

> summary(a,test="P")
          Df  Pillai approx F num Df den Df  Pr(>F)
Y[, 1]     1 0.23887   4.1845      3     40 0.01146 *

> summary(a,test="R")
          Df     Roy approx F num Df den Df  Pr(>F)
Y[, 1]     1 0.31384   4.1845      3     40 0.01146 *
```

MANOVA in R

The results disagree with my calculations. However, I don't trust that I set things up correctly in R. In particular, if I look for the eigenvalues, I get only one non-zero eigenvalue:

```
> summary(a)$Eigenvalues
             [,1]          [,2]          [,3]
Y[, 1] 0.3138357 1.076934e-17 1.076934e-17
```

and this doesn't square with the theory, so something isn't quite right, but I'm not sure what! Unfortunately, I don't see a way of getting the **H** and **E** matrices out of the R output.

MANOVA in R

Trust me, I was quite annoyed that I couldn't get this working last night. Any ideas for how I set up the model incorrectly?

MANOVA in R: grouping variables shouldn't be numeric!

```
> a <- manova(Y[,2:4]~factor(Y[,1]))
> summary(a,test="W")
               Df  Wilks approx F num Df den Df    Pr(>F)
factor(Y[, 1])  3 0.2298   8.5413      9 92.633 3.307e-09 ***

> summary(a,test="H")
               Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
factor(Y[, 1])  3            2.465   10.043      9    110 3.836e-11

> summary(a,test="P")
               Df  Pillai approx F num Df den Df    Pr(>F)
factor(Y[, 1])  3 0.98336   6.5016      9    120 1.663e-07 ***

> summary(a,test="R")
               Df    Roy approx F num Df den Df    Pr(>F)
factor(Y[, 1])  3 2.0651   27.534      3     40 7.967e-10 ***
```

Interpretation

Consistent with the idea that the Pillai test is conservative and Hotelling is liberal, these two tests have the highest and lowest p-values, respectively, although all tests agree that the chile peppers are different on the three variables.

You could do additional tests to determine which populations are different or which variables contribute most to differences between the chile varieties. Some subsets of the data will not show evidence of a difference between two groups.

If you don't specify a particular test, then the default output is the Pillai's trace test only. Unfortunately, you seem to need to call the summary function once for each test.

Assumptions

We went ahead and proceeded with the data analysis without testing assumptions. In particular, the data might not be multivariate normal for the different groups. Visual tests don't show anything alarming. In the scatterplot matrix, bivariate plots look roughly like clouds, although Length versus Thickness might have some multivariate outliers. However, individual tests of normality do fail using `shapiro.test()`, so there is evidence against normality and therefore multivariate normality as well.

In spite of the fact that the data are not normal, averages will be closer to multivariate normal than individual data points, and so it might not be a problem to use methods assuming multivariate normality.

Unbalanced MANOVA

If the sample sizes are unequal, then the MANOVA is called unbalanced. Here the ith sample has sample size n_i. The computation of the test statistics is very similar, with

$$N = \sum_{i=1}^{k} n_i, \quad \bar{\mathbf{y}}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{y}_{ij}, \quad \bar{\mathbf{y}}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \mathbf{y}_{ij}$$

$$\mathbf{H} = \sum_{i=1}^{k} n_i (\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..})' = \sum_{i=1}^{k} \frac{1}{n_i} \mathbf{y}_{i.} \mathbf{y}_{i.}' - \frac{1}{N} \mathbf{y}_{..} \mathbf{y}_{..}'$$

$$\mathbf{E} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{\mathbf{y}}_{ij} - \bar{\mathbf{y}}_{i.})(\bar{\mathbf{y}}_{ij} - \bar{\mathbf{y}}_{i.})' = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \bar{\mathbf{y}}_{i.}) \bar{\mathbf{y}}_{i.})' - \sum_{i=1}^{k} \frac{1}{n_i} \mathbf{y}_{i.} \mathbf{y}_{i.}'$$

η^2

The quantity

$$\eta^2 = \frac{\text{between sum of squres}}{\text{total sum of squares}}$$

is called Fisher's correlation ratio for ANOVA, and is similar to R^2. It is also a measure of model fit, since if it is large, then this means that the sum of squares error is small.

For MANOVA, we can get similar expressions for η^2 based on Wilks' Λ and Roy's root test with

$$\eta^2_\Lambda = 1 - \Lambda$$

and

$$\eta^2_\theta = \frac{\lambda_1}{1 + \lambda_1}$$

Canonical correlation

The value $\sqrt{\lambda_1/(1+\lambda_1)}$ is the maximum correlation between a linear combination of the p variables and a linear combination of dummy variables representing the groups.

If you only had two variables, then you would have $x = 0, 1$ depending on whether an observation belonged to one of two groups. In this case, the value is the maximum correlation between a linear combination of the p response variables and x. If there were three groups, you could have $x_1 = 1$ if an observations was from group 1; otherwise $x_1 = 0$. Similarly, let $x_2 = 1$ if an observation is from group 2; otherwise $x_2 = 0$. Gerenally, let $x_i = 1$ if an observation is from group i. Then we only need $k - 1$ dummy variables since an observation belongs to group k if and only if $x_1 = \cdots = x_{k-1} = 0$.

These correlations between groups and variables are called canonical correlations.

Canonical correlation

Generally define $r_i^2 = \lambda_i/(1+\lambda_i)$. Then r_i^2s is called the ith squared canonical correlation, which will play a role later in canonical correlation analysis.

The test statistics for MANOVA such as Wilks' Λ and Pillai's trace can be expressed in terms of the r_i values as

$$\Lambda = \prod_{i=1}^{s} 1 - r_i^2$$

$$\text{Pillais trace} = \prod_{i=1}^{2} r_i^2$$

Two-way ANOVA/MANOVA

Analogous to two-way ANOVA, we can do two-way MANOVA as well. This is when we have separate samples for combinations of two factors. For the ANOVA, the model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

where γ_{ij} is an interaction term.

The book points out that other books recommend testing for an interaction first, and if it is significant, then include both main effects (i.e. both α and β), and only test for significance of main effects if the interaction is not significant. The author takes the interesting position that it is reasonable to test for main effects even in the presence of an interaction. This seems to be a minority view, and not one that I was taught, but I actually have no opinion on this and don't feel I understand it well enough...

Two-way ANOVA/MANOVA

The value α_i represents the average effect of the ith level of the first factor, averaging over the levels of the second factor. We can also interpret this as the average when the first factor is at the ith level, minus the overal average

$$\alpha_i = \overline{\mu}_{i.} - \mu_{..}$$

SImilarly

$$\beta_j = \overline{\mu}_{.j} - \mu_{..}$$

Two-way balanced ANOVA/MANOVA

Table 6.4. Univariate Two-Way Analysis of Variance

Source	Sum of Squares	df
A	$\text{SSA} = nb\sum_i(\overline{y}_{i..} - \overline{y}_{...})^2$	$a-1$
B	$\text{SSB} = na\sum_j(\overline{y}_{.j.} - \overline{y}_{...})^2$	$b-1$
AB	$\text{SSAB} = n\sum_{ij}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2$	$(a-1)(b-1)$
Error	$\text{SSE} = \sum_{ijk}(y_{ijk} - \overline{y}_{ij.})^2$	$ab(n-1)$
Total	$\text{SST} = \sum_{ijk}(y_{ijk} - \overline{y}_{...})^2$	$abn-1$

Two-way balanced ANOVA/MANOVA

To test main effects or the interaction term, the appropriate sum of squares is divided by degrees of freedom to obtain the mean squares, the mean squaure for the effect is divided by mean squared error for an F test. For example, to test whether factor A is significant, i.e.,

$$H_{0A} : \boldsymbol{\alpha} = \mathbf{0}$$

use

$$\frac{SSA/(a-1)}{SSE/(ab(n-1))}$$

where a is the number of levels of factor A and b is the number of levels of b.

Two-way balanced ANOVA/MANOVA

MANOVA is analogous to ANOVA, with the model being

$$\mathbf{y}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij} + \boldsymbol{\varepsilon}_{ijk} = \boldsymbol{\mu}_{ij} + \boldsymbol{\varepsilon}_{ijk}$$

where, for example $\boldsymbol{\alpha}_i$ is a p-dimension vector which is the effect of ith treatment on each of the p variables. All vectors in the model are p-dimensional.

Again we have $\boldsymbol{\alpha}_i = \overline{\boldsymbol{\mu}}_{i\cdot} - \overline{\boldsymbol{\mu}}_{\cdot\cdot}$

The total sum of squares can be partitioned as

$$\mathbf{T} = \mathbf{H}_A + \mathbf{H}_B + \mathbf{H}_{AB} + \mathbf{E}$$

Two-way balanced MANOVA

Table 6.5. Multivariate Two-Way Analysis of Variance

Source	Sum of Squares and Products Matrix	df
A	$\mathbf{H}_A = nb\sum_i(\bar{\mathbf{y}}_{i\cdot\cdot} - \bar{\mathbf{y}}_{\cdot\cdot\cdot})(\bar{\mathbf{y}}_{i\cdot\cdot} - \bar{\mathbf{y}}_{\cdot\cdot\cdot})'$	$a-1$
B	$\mathbf{H}_B = na\sum_j(\bar{\mathbf{y}}_{\cdot j\cdot} - \bar{\mathbf{y}}_{\cdot\cdot\cdot})(\bar{\mathbf{y}}_{\cdot j\cdot} - \bar{\mathbf{y}}_{\cdot\cdot\cdot})'$	$b-1$
AB	$\mathbf{H}_{AB} = n\sum_{ij}(\bar{\mathbf{y}}_{ij\cdot} - \bar{\mathbf{y}}_{i\cdot\cdot} - \bar{\mathbf{y}}_{\cdot j\cdot} + \bar{\mathbf{y}}_{\cdot\cdot\cdot}) \times (\bar{\mathbf{y}}_{ij\cdot} - \bar{\mathbf{y}}_{i\cdot\cdot} - \bar{\mathbf{y}}_{\cdot j\cdot} + \bar{\mathbf{y}}_{\cdot\cdot\cdot})'$	$(a-1)(b-1)$
Error	$\mathbf{E} = \sum_{ijk}(\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{ij\cdot})(\bar{\mathbf{y}}_{ijk} - \bar{\mathbf{y}}_{ij\cdot})'$	$ab(n-1)$
Total	$\mathbf{T} = \sum_{ijk}(\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{\cdot\cdot\cdot})(\mathbf{y}_{ijk} - \bar{\mathbf{y}}_{\cdot\cdot\cdot})'$	$abn-1$

Two-way balanced MANOVA

Tests can be based on Wilks' Λ using

$$\Lambda_A = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}_A|} \sim \Lambda_{p,a-1,ab(n-1)}$$

$$\Lambda_B = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}_B|} \sim \Lambda_{p,b-1,ab(n-1)}$$

$$\Lambda_{AB} = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}_{AB}|} \sim \Lambda_{p,(a-1)(b-1),ab(n-1)}$$

Or you can use eigenvalues of $\mathbf{E}^{-1}\mathbf{H}_A$, $\mathbf{E}^{-1}\mathbf{H}_B$, $\mathbf{E}^{-1}\mathbf{H}_{AB}$

Two-way balanced MANOVA example

An example is a test on bars of steel measuring torque and strain when bars of steel are rotated either fast or slow (factor A) and using four different lubricants (factor B). This is a bivariate example with $p = 2$, $a = 2$, and $b = 4$. There are $2 \times 4 = 8$ samples, but they are not indepedent because we expect the slow rotating examples might be more similar to each other than fast rotating examples, and similarly test results for the same lubricant might be related.

Two-way balanced MANOVA

Table 6.6. Two-Way Classification of Measurements on Bar Steel

	A_1		A_2	
Lubricant	y_1	y_2	y_1	y_2
B_1	7.80	90.4	7.12	85.1
	7.10	88.9	7.06	89.0
	7.89	85.9	7.45	75.9
	7.82	88.8	7.45	77.9
B_2	9.00	82.5	8.19	66.0
	8.43	92.4	8.25	74.5
	7.65	82.4	7.45	83.1
	7.70	87.4	7.45	86.4
B_3	7.28	79.6	7.15	81.2
	8.96	95.1	7.15	72.0
	7.75	90.2	7.70	79.9
	7.80	88.0	7.45	71.9
B_4	7.60	94.1	7.06	81.2
	7.00	86.6	7.04	79.9
	7.82	85.9	7.52	86.4
	7.80	88.8	7.70	76.4

Two-way balanced MANOVA

$$\mathbf{H}_A = \begin{pmatrix} 1.205 & 27.208 \\ 27.208 & 614.251 \end{pmatrix} \qquad \mathbf{H}_B = \begin{pmatrix} 1.694 & -9.862, \\ -9.862 & 74.874 \end{pmatrix}$$

$$\mathbf{H}_{AB} = \begin{pmatrix} .132 & 1.585 \\ 1.585 & 32.244 \end{pmatrix} \qquad \mathbf{E} = \begin{pmatrix} 4.897 & -1.890 \\ -1.890 & 736.390 \end{pmatrix}$$

$$\Lambda_A = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_A|} = \frac{3602.2}{7600.2} = .474 < \Lambda_{.05,2,1,24} = .771,$$

$$\Lambda_B = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}_B|} = \frac{3602.2}{5208.6} = .6916 > \Lambda_{.05,2,3,24} = .591.$$

Two-way balanced MANOVA

From these results, there is no interaction between lubricant and speed (so the lubricants do not perform differently at different speeds for those speeds in the experiment). Also, speed had an effect on torque and strain, but lubricant did not. From a manufacturing point of view, this might lead to a decision about using a cheaper lubricant.

Two-way balanced MANOVA example

```
> x <- read.table("steel.txt",header=T)
> x
   speed lube torque strain
1      1    1   7.80   90.4
2      1    1   7.10   88.9
3      1    1   7.89   85.9
4      1    1   7.82   88.8
5      1    2   9.00   82.5
6      1    2   8.43   92.4
7      1    2   7.65   82.4
8      1    2   7.70   87.4
9      1    3   7.28   79.6
10     1    3   8.96   95.1
11     1    3   7.75   90.2
12     1    3   7.80   88.0
13     1    4   7.60   94.1
14     1    4   7.00   86.6
15     1    4   7.82   85.9
16     1    4   7.80   88.8
17     2    1   7.12   85.1
18     2    1   7.06   89.0
19     2    1   7.45   75.9
20     2    1   7.45   77.9
```

Two-way balanced MANOVA in R

```
> a <- manova(cbind(x$torque,x$strain) ~ x$speed + x$lube + x$
> summary(a,test="W")
                Df   Wilks approx F num Df den Df     Pr(>F)
x$speed          1 0.49222  13.9266      2     27 6.985e-05 ***
x$lube           1 0.98899   0.1503      2     27     0.8612
x$speed:x$lube   1 0.99321   0.0923      2     27     0.9121
Residuals       28
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Two-way balanced MANOVA in R

Consistent with the previous results, this suggests that the interaction isn't important, and lubricant also doesn't seem important. You could use backward selection two settle upon a model, and this won't change the results in this case:

```
> a <- manova(cbind(x$torque,x$strain) ~ x$speed + x$lube)
> summary(a,test="R")
           Df     Roy approx F num Df den Df    Pr(>F)
x$speed     1 1.02826  14.3956      2     28 5.015e-05 ***
x$lube      1 0.01113   0.1558      2     28    0.8564
Residuals 29
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> a <- manova(cbind(x$torque,x$strain) ~ x$speed)
> summary(a,test="R")
           Df    Roy approx F num Df den Df    Pr(>F)
x$speed     1 1.0237   14.843      2     29 3.639e-05 ***
Residuals 30
---
```

Higher-order models

The two-way approach can be extended to have more main effects and more interaction terms, and this is easy to implement in R using the modeling notation.

Other designs from ANOVA, such as split plot designs, random effects models, mixed models, and so on can also be generalized to the multivariate setting.

Checking assumptions

We discussed how you could visually look for violations of the assumption of multivariate normality. In particular, you can check for each combination of factors whether the responses appear to be multivariate normal, and whether individual variables tend to be normally distributed for each combination of factors.

A limitation of this approach is if each of the samples is small. In the steel bar example, there were four replicates per combination of treatments.

A more thorough approach is to examine the residuals of the model. The residuals should be multivariate normal from the same distribution, you could look at scatterplot matrices and test univariate normality of each vector of the residuals.

Checking assumptions

The residuals from the model are

$$\widehat{\boldsymbol{\varepsilon}}_{ijk} = \mathbf{y}_{ijk} - \overline{\mathbf{y}}_{ij.}$$

The residuals should be distributed as $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ These are easily available from R. A formal test of univariate normality shows that the residual vectors fail in one of the dimensions, so multivariate normality is not formally passed for this data.

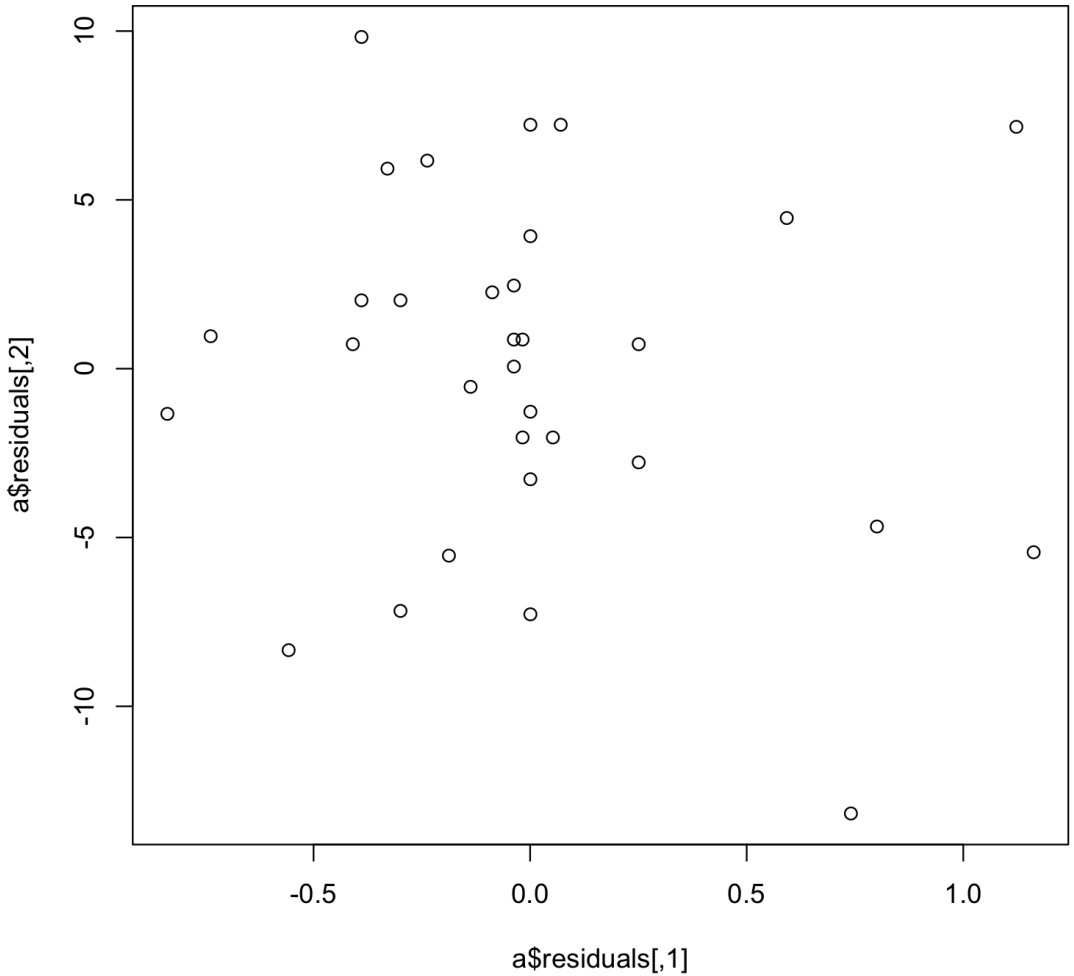
Checking assumptions

```
> names(a)
 [1] "coefficients"  "residuals"      "effects"        "rank"
 [5] "fitted.values" "assign"         "qr"             "df.resid
 [9] "xlevels"       "call"           "terms"          "model"
> a$residuals
         [,1]     [,2]
1  -0.037500   2.4625
2  -0.737500   0.9625
3   0.052500  -2.0375
4  -0.017500   0.8625
5   1.162500  -5.4375
6   0.592500   4.4625
...
32  0.250625  -2.7750
```

Checking assumptions

```
> shapiro.test(a$residuals[,1])

Shapiro-Wilk normality test

data:  a$residuals[, 1]
W = 0.9075, p-value = 0.009693

> shapiro.test(a$residuals[,2])

Shapiro-Wilk normality test

data:  a$residuals[, 2]
W = 0.9813, p-value = 0.8377
```

Two-way balanced MANOVA



Two-way balanced MANOVA

The residuals here are NOT a typical residual plot where you plot the residuals against the fitted values. Here we've plotted just the residual vectors, one component against the other. This is more like plotting a histogram of the residuals in a univariate ANOVA. The plot here seems more spread out than it should be for a bivariate normal, but there aren't any huge outliers.

Chapter 6: overview of remainder

The rest of Chapter 6 deals with profile analysis, repeated measures versions of MANOVA, and growth curves in a multivariate setting. For repeated measures.

Profile analysis and repeated measures MANOVA can be extended to $k \geq 2$ groups instead of having a single group, and can have additional factors in the model (e.g., sex can be a factor within two separate groups).

For growth curves, you can test for linear, quadratic, and other polynomial trends in quantitative factors, such as years of age, in addition to having multiple groups. An example of a type of problem here might be testing whether growth curves for kids are the same for kids who were either nursed or formula-fed, where sex of the child is treated as a covariate.

Chapter 7: tests of covariance matrices

We'll have just a one-day review of chapter 7, then we'll get to other procedures over the break such as principle components, multidimensional scaling, and discriminant analysis that have a different flavor from the first half of the course.

First, in testing properties of covariance matrices, there are many types of questions we might be interesting in testing about covariance matrices such as

- does a covariance matrix equal a hypothesized matrix: $\mathbf{\Sigma} = \mathbf{\Sigma}_0$?
- are the variables independent? (Covariances equal to 0 for multivariate normal)
- does a covariance matrix have a special structure, such as all covariances being equal?
- do two covariance matrices come from the same population, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$?

Tests of covariance matrices

To test $H_0 : \mathbf{\Sigma} = \mathbf{\Sigma}_0$ versus $H_1 : \mathbf{\Sigma} \neq \mathbf{\Sigma}_0$, it isn't necessary to specify $\boldsymbol{\mu}$. For this hypothesis all variances and covariances are specified under the null.

To test the hypothesis, we see if an observed sample covariance $\mathbf{S}$ is significantly different from $\mathbf{\Sigma}_0$. The test statistic can either be expressed in terms of determinants or eigenvalues:

$$\begin{aligned} u &= v\left[\ln|\mathbf{\Sigma}_0| - \ln|\mathbf{S}| + \text{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) - p\right] \\ &= v\left[\sum_{i=1}^{p}(\lambda_i - \ln\lambda_i) - p\right] \end{aligned}$$

where $v = n - 1$ for a one-sample problem and $v = \sum_{i=1}^{k} n_k - k = N - k$ for a pooled covariance matrix obtained from k samples.

Tests of covariance matrices

Under H_0, when v is large, the test has an approximate χ^2 distribution with $\binom{p}{2} = p(p+1)/2$ degrees of freedom, which is also the number of off-diagonal elements in the upper or lower triangle of the matrix. (I.e., there are $\binom{p}{2}$ terms σ_{ij} with $i < j$ in a covariance matrix.) For smaller v, there is a correction to make the test perform a little better.

A special case is the test of the hypothesis $H_0 : \mathbf{\Sigma} = \mathbf{I}$, testing whether the a set of variables has unit variance and are uncorrelated. If you are not interested in whether the variances are equal to 1, but are interested in the covariances (correlations), you could standardize the variables first (get their z-scores) and then do the tests, so that there variances will be equal to 1. In this case, you are really just testing whether all the correlations are equal to 0. In a two variable case, this amounts to testing whether the correlation between two variables is 0.

Tests of covariance matrices

We can take the chile data as an example, just using one variety, those grown on Casados farms. Testing whether the variances are equal to 1 or not is not very interesting, so we'll standardize the data first.

```
> y <- read.table("chile.txt",header=T)
> y2 <- y[y$group=="Casados",2:4] # Casados only
> cor(y2)
              Length      Width   Thickness
Length     1.00000000 -0.1627035 -0.03993501
Width     -0.16270351  1.0000000  0.37510947
Thickness -0.03993501  0.3751095  1.00000000
> z <- scale(y2,center=T,scale=T) #R does the z-scores for you
> cov(z)  #note that the covariance of the z-scores is the correlati
# on the original scale...
              Length      Width   Thickness
Length     1.00000000 -0.1627035 -0.03993501
Width     -0.16270351  1.0000000  0.37510947
Thickness -0.03993501  0.3751095  1.00000000
```

Tests of covariance matrices

You might first consider doing pairwise tests of correlations. The sample size is small here, only 11 observations, so there isn't much power to detect correlations that exist. So it might be better to test all correlations simultaneously first. We'll see what happens if we test individual correlations afterward. First we need to define $\mathbf{\Sigma}_0$.

```
> I <- diag(3)  # the identity matrix is our null covariance matrix
> p <- 3
> S <- cov(z)
# sum(diag(S)) is the trace of S
> u = v*(log(det(I)) - log(det(S)) + sum(diag(S)) - 3)
> u
[1] 1.790068
# modification for small samples suggested by book
> uprime <- (1 - (1/(6*v-1))*(2*p+1-2/(p+1)))*u
> uprime
[1] 1.592857
```

Tests of covariance matrices

We compare this to a χ^2 with $\binom{3}{2} = 3$ degrees of freedom, which has a mean of 3 (since the expected value of a χ^2 is its degrees of freedom. This means that if the data came from a $N_3(\boldsymbol{\mu}, \mathbf{I})$, then you'd expect to get a larger test statistic than this on average, so the data is quite consistent with length, width, and thickness being uncorrelated.

To quantify how consistent the data is with the null hypothesis, you can use a pvalue:

```
> 1-pchisq(1.593,3)
[1] 0.6609781
```

Keep in mind that this was a small sample, so there wasn't much power to detect correlations.

Tests of covariance matrices: sphericity

A slightly stronger condition to test is that the variables are independent AND have the same variance, although the variance might not be 1.

Here the null hypothesis is $H_0 : \mathbf{\Sigma} = \sigma^2 \mathbf{I}$ and the alternative is $H_1 : \mathbf{\Sigma} \neq \sigma^2 \mathbf{I}$.

If the variables are multivariate normal, then

$$(\mathbf{y} - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) = c^2$$

describes an ellipsoid, while if H_0 is true, then plugging in $\Sigma = \sigma^2 \mathbf{I}$ leads to

$$(\mathbf{y} - \boldsymbol{\mu})' (\mathbf{y} - \boldsymbol{\mu}) = \sigma^2 c^2,$$

which is the equation for a p-dimensional "sphere".

Tests of covariance matrices: sphericity

One could instead test the hypothesis $H_0 : \mathbf{C\Sigma C}' = \sigma^2\mathbf{I}$, where $\mathbf{C}$ is a contrast matrix, which is useful for repeated measures.

A test statistic based on the likelihood ratio is

$$-2\ln(LR) = -n\left[\frac{|\mathbf{S}|}{(\text{tr }\mathbf{S}/p)^p}\right]^n$$

which can be improved by

$$u' = -\left(v - \frac{2p^2 + p + 2}{6p}\right)(-2\ln(LR))$$

Then u' is approximately χ^2 with $binomp + 12 - 1$ degrees of freedom. The degrees of freedom comes from the number of parameters under the alternative minus the number of parameters under the null. Under the alternative, there are p variances and $\binom{p}{2}$ covariances with $p + \binom{p}{2} = \binom{p+1}{2}$. Under the null, covariances are 0 and there is a common variance, so there is 1 parameter.

Tests of covariance matrices: sphericity

To apply this to the chile data from Casados, we would expect the test to reject the null because thickness is much less variable than length or width, with the variances being 0.760, 0.210, and 0.019 for length, width, and thickness, respectively. But, just to illustrate, we get

```
> n <- 11
> LR <- (det(S)/(sum(diag(S)))^p)^(n/2)
> LR
[1] 6.347235e-15
> uprime <- -(v - (2*p^2+p+2)/(6*p))*log(LR^(2/n))
> uprime
[1] 51.84292
```

Tests of covariance matrices: sphericity

We compare this number to a χ^2 with $\binom{4}{2} - 1 = 5$ degrees of freedom, so the mean is 5. The variance is $2k$ for k degrees of freedom, so in this case the variance is 10, and the standard deviation is a little more than 3. So the test statistic is more than 15 standard deviations above the mean. Again, to quantify this as a p-value,

```
> 1 - pchisq(51.84,5)
[1] 5.818157e-10
```

So, although we don't have sufficient evidence to conclude that the variables are uncorrelated, we have sufficient evidence to conclude that that it's not the case that they are independent with a common variance. This null hypothesis could be false either due to different variances or due to correlation.

Tests of covariance matrices: common covariance and variance

The test for a common covariance and variance throughout the covariance matrix is an important covariance structure that is often used in repeated measures. The idea is that the covariance matrix looks like this

$$\sigma^2 \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & \vdots & \vdots & & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix}$$

The idea is that all variables are correlated to the same degree, and all variables have the same covariance. When this assumption is met, you can analyze repeated measures data using ANOVA. This covariance structure is called **compound symmetry**, **uniform**, or **intraclass correlation model**. Often in software, such as in mixed models in SAS, there are a limited number of covariances structures that you can assume for the data and analyze the data assuming that particular structure.

Tests of covariance matrices: common covariance and variance

The null hypothesis can be written as

$$H_0 : \mathbf{\Sigma} = \sigma^2[(1-\rho)\mathbf{I} + \rho\mathbf{J}]$$

We can't really state what the covariance matrix is under the null exactly, but we can estimate it under the null by using the average of the p variances and the average of the $\binom{p}{2}$ covariances. Thus, let

$$s^2 = \frac{1}{p}\sum_{i=1}^{p} s_{ii}, \quad r = \frac{1}{\binom{p}{2}}\sum_{i>j} s_{ij}$$

Tests of covariance matrices: common covariances and variance

The estimated covariance matrix under the null is

$$\mathbf{S}_0 = \begin{pmatrix} s^2 & s^2 r & \cdots & s^2 r \\ s^2 r & s^2 & \cdots & s^2 r \\ \vdots & \vdots & & \vdots \\ s^2 r & s^2 r & \cdots & s^2 \end{pmatrix} = s^2[(1-r)\mathbf{I} + r\mathbf{J}]$$

Here $r = s^2 r / s^2$ estimates the correlation, and $s^2 r$ estimates the covariance.

Let $\mathbf{S}$ be the usual sample estimate of the covariance matrix without the constraint that variances are equal to each other and covariances are equal to each other. Then let

$$u = \frac{|\mathbf{S}|}{|\mathbf{S}_0|}$$

Tests of covariance matrices: common covariances and variance

and the test statistic is

$$u' = -\left[v - \frac{p(p+1)^2(2p-3)}{6(p-1)(p^2+p-4)}\right]$$

and this is approximately χ^2 with $\binom{p+1}{2} - 2$ degrees of freedom. We have -2 instead of -1 becuase under the null, there are two parameters being estimated instead of 1.

Tests of covariance matrices: common covariances and variance

To test the chile data for compound symmetry,

```
> S <- cov(y2)
> (S[1,2] + S[1,3] + S[2,3])/3
[1] -0.01536667
> rs2 <- (S[1,2] + S[1,3] + S[2,3])/3
> s2 <- (var(y2$Length) + var(y2$Width) + var(y2$Thickness))/3
> r <- rs2/s2
> J <- matrix(rep(1,9),ncol=3)
> S0 <- s2*((1-r)*I + r*J)
> S0
             [,1]        [,2]        [,3]
[1,]  0.32966970 -0.01536667 -0.01536667
[2,] -0.01536667  0.32966970 -0.01536667
[3,] -0.01536667 -0.01536667  0.32966970
```

Tests of covariance matrices: common covariances and variance

```
> u <- det(S)/det(S0)
> u
[1] 0.07127613
> uprime <- -(v-p*(p+1)^2*(2*p-3)/(6*(p-1)*(p^2+p-4)))*log(u)
> uprime
[1] 22.45015
```

This gives us a large χ^2 value with 4 degrees of freedom, so there is strong evidence against compound symmetric structure in the green chile data.

Comparing covariance matrices

MANOVA assumes that covariances are equal between different populations that are being sampled. Although MANOVA is considered to be fairly robust against this assumption, we can test this assumption by testing

$$H_0 : \mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \cdots = \mathbf{\Sigma}_k$$

A univariate analogue is

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

Comparing covariance matrices

The univariate, multi-sample case can be testing using

$$c = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{v_i} - \frac{1}{\sum_{i=1}^{k} v_i}\right]$$

$$s^2 = \frac{\sum_{i=1}^{k} v_i s_i^2}{\sum_{i=1}^{k} v_i}$$

$$m = \left(\sum_{i=1}^{k} v_i\right) \ln s^2 - \sum_{i=1}^{k} v_i \ln s_i^2$$

Then m/c is roughly χ^2_{k-1}. The test assumes that the k samples are indepedent, so it is not a test of whether the diagonals of a covariance matrix are equal in a multivariate setting where variables are correlated.

Comparing covariance matrices

The multivariate analogue for Bartlett's test uses

$$\mathbf{S}_{\text{pl}} = \frac{\sum_{i=1}^{k} v_i \mathbf{S}_i}{\sum_{i=1}^{k} v_i} = \frac{\mathbf{E}}{v_E}$$

$$M = \frac{|\mathbf{S}_1|^{v_1/2}|\mathbf{S}_2|^{v_2/2}\cdots|\mathbf{S}_k|^{v_k/2}}{|\mathbf{S}_{\text{pl}}|^{\sum_i v_i/2}}$$

$$c = \left[\sum_{i=1}^{k}\frac{1}{v_i} - \frac{1}{\sum_{i=1}^{k} v_i}\right]\left[\frac{2p^2+3p-1}{6(p+1)(k-1)}\right]$$

$$u = -2(1-c)\ln M$$

Then u is approximately χ^2 with degrees of freedom $(k-1)\binom{p+1}{2}$.

Comparing covariance matrices

A warning about Bartlett's test in the univariate case is that it is very sensitive to departures from normality. In particular, if two indepednent populations have the same variance but are not normally distributed, then Bartlett's test might reject the null hypothesis of equal variance much more often than α for an α-level test.

In particular, if you are sampling from non-normal distributions with equal variances and apply Bartlett's test, then the type I error rate can increase with increasing sample sizes. We tried this last semester in the SAS class with two independent exponential samples. This is a very bad situation, since we usually expect our inferences to improve with more data. So this caveat about Bartlett's test should apply to the multivariate version also. Departures from multivariate normality could lead to too easily rejecting the null hypothesis.

Comparing covariance matrices

For this reason Bartlett's test is often not used and tests of equality of variance are often not done very formally. Instead informal measures are often used such as looking at side-by-side box plots to look for gross violations of equal variance.

Comparing covariance matrices: chile example

TWe'll check the equality of the green chile covariance matrices:

```
> y1 <- y[y$group=="Alcalde",2:4]; y2 <- y[y$group=="Casados",2:4]
> y3 <- y[y$group=="Chimayo",2:4]; y4 <- y[y$group=="Cochiti",2:4]
> S1 <- cov(y1); S2 <- cov(y2); S3 <- cov(y3); S4 <- cov(y4)
> Spl <- S1 + S2 + S3 + S4
> M <- det(S1)^5 * det(S2)^5 + det(S3)^5 + det(S4)^5
> M <- M/det(Spl)^20
> c <- (4/10 - 1/40)*(2*3^2 + 3*3-1)/(6*(3+1)*(4-1))
> u <- -2*(1-c)*log(M)
> u
[1] 31.36207
> 1-pchisq(u,18) #18 degrees of freedom
[1] 0.02613031
```

There is evidence, but not strong evidence against equal variances. Particular since the data might not be multivariate normal, this is not very strong evidence. I would not be uncomfortable using MANOVA based on this result.