2/25/2010 36-402/608 ADA-II H. Seltman

Breakout #13 Results

Remember the bran study with 20 subjects given three diets (baseline, hi-fiber, and lo-fiber) in random order to see how they affect cholesterol. (We saw that order appeared to have no effect, and will not use it as a variable.) One way to express the scientific hypotheses is that we are simultaneously interested in testing $\mu_B = (\mu_L + \mu_H)/2$ and $\mu_H = \mu_L$. Although we could do two separate paired t-tests, we will do a single overall one-sample test of the two hypotheses.

```
BHL = bran$hifiber-bran$lofiber
BF = (bran$lofiber+bran$hifiber)/2-bran$baseline
means = matrix(c(mean(BF),mean(BHL)), ncol=1)
n = nrow(bran)
T2 = n * as.numeric(t(means) %*% solve(cov(cbind(BF,BHL))) %*% means)
T2 # 19.70
F = (n-2)/2/(n-1)*T2
F # 9.33
1-pf(F, 2, n-2) # 0.00166
```

Question 1: How does the code relate to the formulas in the handout? What conclusion do you reach? What followup testing should be done?

Because no subjects differ in the treatment they recieved from other subjects, the pertinent section is that for K=1 treatment (with p=3 measurements per subject):

We can construct $p-1$ difference variables, then test $\boldsymbol{\mu} = \mathbf{0}$ for the *differences*. Then the test is $T^2 = n\,\mathbf{m}'\,\mathbf{S}^{-1}\mathbf{m}$ and $\frac{n-2}{2(n-1)}T^2$ follows the $F_{2,n-2}$ distribution when $\boldsymbol{\mu} = \mathbf{0}$.

The first two lines construct difference variables. The "means" variable is $\boldsymbol{\mu}$. The code `cov(cbind(BF,BHL))` generates $\mathbf{S}$ (the estimated variance-covariance matrix), and `solve()` computes the inverse. The `%*%` operator does matrix multiplication, and `t()` does matrix transpose. The code `pf()` looks up the p-value in the F table.

We reject the null hypothesis that both $\mu_B = (\mu_L + \mu_H)/2$ and $\mu_H = \mu_L$. This is equivalent to rejecting the hypothesis that all three population means (of the cholesterol value for the three diets) are equal. We could do paired t-tests to test each individual hypothesis as a directed follow-up.

Recall the flea beetle study in which two different measurements are taken on two similar species of beetles. The question of interest is whether the collection of p=2 measurements are a distinguishing feature between the species (though not necessarily useful for distinguishing individuals).

```
> anova(aov(beet[,1:2]~species, data=beet), test="Hotelling")
```

```
# Error in model.frame.default(formula = beet[, 1:2] ~ species, data = beet,  :
#   invalid type (list) for variable 'beet[, 1:2]'

anova(aov(as.matrix(beet[,1:2])~species, data=beet), test="Hotelling")
# Analysis of Variance Table
#             Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
# (Intercept)  1           344.15   5678.4      2     33 < 2.2e-16 ***
# species      1             4.81     79.4      2     33 2.455e-13 ***
# Residuals   34
```

Question 2: Why did the first attempt fail? How does the code relate to the formulas in the handout? How does `anova()` "know" to do MANOVA? What conclusion do you reach? How can you get a small p-value and then perhaps find that these measurements are not very useful for categorizing individual beetles?

The error message is about an "invalid type" and was fixed by using `as.matrix()`, so apparently aov cannot handle a data.frame as input; it needs a matrix.

The fact that we use two column as our response variable in the model formula says we have p=2 (this is a manova, not an anova), and the fact that species has two levels tells us that K=2, so we use the formulas of section 2b of the handout. The multicolumn response tells R to run MANOVA.

The p-value for species is small, so we reject the null hypothesis that both species have the same population mean vector.

If the two ellipsoids that describe the two bivariate normal distributions of the two measurements have a lot of overlap but do have different population means, then with sufficient subjects we will get a small p-value and correctly reject the null hypothesis that the two species have the same two-valued mean. But the overlap indicates that individuals will be hard to classify just on the basis of these two measurements.

Recall the monkeys being tested for short and long term memory with and without brain surgery on the hippocampus.

```
mem$SL = with(mem,cbind(short=(week2+week4)/2,
                        long=(week8+week12+week16)/3))
anova(aov(SL~treatment, data=mem), test="Hotelling")
# Analysis of Variance Table
#             Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
# (Intercept)  1          250.795  1880.96      2     15 < 2.2e-16 ***
# treatment    1            1.643    12.32      2     15 0.0006831 ***
# Residuals   16
```

Question 3: What is the first statement doing? What are the values of p and K? What conclusion do you reach? What could you do to check assumptions?

This is an example of a “column” in a data.frame that is actually a matrix. (Technically this is possible because data.frames are lists of their columns, so the type of data in each column are is restricted. The multivariate normal dimension is reduced from 5 to 2 (short vs. long) based on the scientists advice. We have K=2 treatment groups (control vs. brain surgery). We reject the null hypothesis that the population mean vectors (correct rate for short and long) are identical for control vs. treated monkeys. Identically we can express the null hypothesis that the single length-two mean vector $(\mu_{TS} - \mu_{CS},\ \mu_{TL} - \mu_{CL}) = \mathbf{0}$, where $\mathbf{0} = (0, 0)$.

We can make a quantile normal plot (e.g., with qqn()) of the `$residuals` of the `aov` object to check for normality. There is no assumption of independent errors (except across subjects – they are not allowed to collaborate or cheat off each other). There is not assumption of equal variances across the p measurements; the variance-covariance matrix accommodates unequal values on the diagonal. There is an assumption of equal (population) variance-covariance matrices across treatments, and we can roughly check that by looking at separate estimated variance covariance matrices, e.g., using:

```
round(cov(mem$SL[mem$treatment=="CONTROL",]),3)
round(cov(mem$SL[mem$treatment!="CONTROL",]),3)
```

From previous EDA we know that these are likely unrepresentative due to a far outlier in each group. I’d rerun the analysis without the outliers (and report both analyses unless there is a good reason to drop the out strange values).