MA26620: Practical 10

One-Way ANOVA, Contrasts, Introduction to Two-Way ANOVA

1 One-Way ANOVA: a recap

Recall from the lectures and Practical 9 that a one-way ANOVA table looks like this:

Source	SS	DF	MS	F-ratio	P-value
Between groups	$m\sum_{i=1}^{k}(Y_{i\bullet}-Y_{\bullet\bullet})^2$	$k-1$	$MS_{GROUPS}=\frac{SS_{GROUPS}}{(k-1)}$	$F_{obs}=\frac{MS_{GROUPS}}{MS_{ERROR}}$	P
Within groups	$\sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij}-Y_{i\bullet})^2$	$k(m-1)$	$MS_{ERROR}=\frac{SS_{ERROR}}{k(m-1)}$		
Total (corr)	$\sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij}-Y_{\bullet\bullet})^2$	$mk-1$			

To test the hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ vs H_1 : not all group means are equal, we reject H_0 if the F-ratio is greater than the appropriate upper percentage point of the F-distribution having $(k-1)$ and $k(m-1)$ degrees of freedom, or equivalently if the p-value is below the significance level.

Load and attach the `ratweightgain.csv` data from last week into RStudio and remind yourself what R's output looks like by running the command `summary(aov(weightgain~diet))`. It doesn't have the total row, but does have everything else. Make sure you know how to interpret the table, especially the p-value. Ask for help if you don't.

2 Contrasts

The one-way ANOVA analyses above are great for concluding whether group means are all equal or not. However, what if we want to say more? Suppose they're not all equal. Then can we delve deeper to get more precise conclusions?

2.1 Definition

This week, we were looking at *contrasts*, which we defined as a quantity

$$\lambda = \sum_{i=1}^{k} c_i \mu_i,$$

where $\sum_{i=1}^{k} c_i = c_1 + c_2 + \ldots + c_k = 0$. That is, a contrast is a linear combination of group means, with the extra property that the coefficients sum to zero.

These allow us to examine *differences* between group means in a more specific way than the usual one-way ANOVA test, which only allows us to judge whether *all* of the group means are equal or not.

2.2 Estimating the value of a contrast

Since $Y_{i\bullet}$ (the observed sample mean of the i-th group) is an unbiased estimate of μ_i (the population mean of the i-th group), we have that $\hat{\lambda} = \sum_{i=1}^{k} c_i Y_{i\bullet}$ is an unbiased estimator of a contrast λ.

2.3 Sum of squares associated with a contrast

The following quantity is called the *sum of squares associated with a contrast,* λ:

$$L(\hat{\lambda}) = \frac{m\hat{\lambda}^2}{\sum_{i=1}^{k} c_i^2}.$$

Recall that m is the number of observations per group. It's possible to derive (though we'll skip the details here) that

$$L(\hat{\lambda})/(SS_{\text{ERROR}}/(n-k)) \sim F_{1,n-k}.$$

2.4 Orthogonal contrasts

The advantage of the contrast method is that we can define the notion of two contrasts to be *orthogonal* if they're uncorrelated (independent). The condition for this is as follows: let

$$\lambda_1 = \sum_{i=1}^{k} c_{1i}\mu_i, \qquad \lambda_2 = \sum_{i=1}^{k} c_{2i}\mu_i.$$

Then λ_1 and λ_2 are *orthogonal* if and only if $\sum_{i=1}^{k} c_{1i}c_{2i} = 0$.

Tests concerning orthogonal contrasts can be carried out independently of each other.

If we have two orthogonal contrasts between k group means, we can talk of $L(\lambda_1)$ and $L(\lambda_2)$ as being two associated sums of squares each having 1 degree of freedom. Other comparisons between the groups will account for the remaining $k-3$ degrees of freedom.

If we can find $k-1$ mutually orthogonal contrasts, then

$$L(\lambda_1) + L(\lambda_2) + \ldots + L(\lambda_{k-1}) = \text{BETWEEN GROUPS SS}.$$

3 Example - Rat Weight Gain Data revisited

Let's begin by making a nice professional-looking ggplot boxplot of the data so we can get a feel for the groups. This is also good revision for the upcoming assessed practical. Tick the ggplot box in the packages tab and run:

```
ggplot(ratweightgain,aes(x=diet,y=weightgain))+geom_boxplot()
```

Remind yourself (from previous practicals/notes) how to add a title and edit axis labels etc. Make your plot look professional!

You may have already done this at the start of the practical, but if not, make a one-way ANOVA table by running the command `summary(aov(weightgain~diet))`. The small p-value here would lead us to reject the null hypothesis $H_0 : \mu_A = \mu_B = \mu_C = \mu_D = \mu_E = \mu_F$; we conclude that not all group means are equal.

"But couldn't we have deduced that from the boxplot?", I hear you cry. What was the point of running the ANOVA? After all, you only have to look at groups C and D on the boxplot to see that those groups are different.

In response to those questions: firstly it's not always so clear whether the population group means could feasibly be the same or not. Remember that these are fairly small samples. But also those questions reinforce the importance of contrasts which allow us to ask more specific questions. Does $\mu_B = \mu_C$? From the boxplot we'd probably guess not. Does $\mu_B = \mu_D$? This is trickier to conclude with any confidence.

3.1 Contrasts

Suppose we obtain more information about the other diets. We now learn that the six diets were in fact the combinations of high and low protein forms of three diets based on beef, cereal and pork as follows:

Diet	**A**	**B**	**C**	**D**	**E**	**F**
Protein Level	High	High	High	Low	Low	Low
Diet Base	Beef	Cereal	Pork	Beef	Cereal	Pork

Let's add this information into RStudio:

```
protein<-rep(c("High","Low"),each=30)
dietbase<-rep(c("Beef","Cereal","Pork"),each=10,times=2)
```

We have now created two data variables whose elements are text strings. In our analyses these signify the levels of two factors: one has two levels, High and Low; whilst the other has three levels, Beef, Cereal and Pork. Some routines in R demand that factors are identified as such and we can ensure that we don't encounter any pitfalls by writing:

```
protein<-as.factor(protein)
dietbase<-as.factor(dietbase)
```

Some natural contrasts that may be of interest arise here immediately; let's consider the diet base. For instance, how about beef vs pork (which can be described by the contrast $\lambda_{\text{BvP}} = \mu_{\text{beef}} - \mu_{\text{pork}}$). Also animal vs vegetable i.e. $\frac{1}{2}(\mu_{\text{beef}} + \mu_{\text{pork}}) - \mu_{\text{cereal}}$?

We can estimate these two contrasts (`tapply(weightgain,dietbase,mean)` will help). Do so, and calculate their associated sums of squares, remembering that the group size of each different dietbase is 20, not 10. What do they add up to? Are the contrasts orthogonal?

Run `summary(aov(weightgain~dietbase))` - does the between groups sum of squares look familiar?

3.2 Testing

Let's test whether the beef vs pork contrast is zero. Hopefully you found that $\hat{\lambda}_{\text{BvP}} = \frac{1}{2}$ and $L(\hat{\lambda}_{\text{BvP}}) = 5/2$. According to the theory we saw at the start of the practical:

$$L(\hat{\lambda}_{\text{BvP}})/(SS_{\text{ERROR}}/(n-k)) \sim F_{1,n-k},$$

so in this case we seek the probability of $F_{1,54}$ exceeding

$$\frac{2.5}{15932/(60-3)} = 0.0089442.$$

We can ask R to give us this probability: `pf(df1=1,df2=54,0.0089442,lower.tail=FALSE)`. The p-value that R spits out is huge so it seems that whether the diet base is beef or pork makes no significant difference.

What about the animal vs vegetable contrast? Conduct the F-test and conclude whether or not you'd believe this contrast to be zero.

4 Two-way ANOVA

As we've seen, the rat weight gain data really has two different ways of splitting the observations into groups: by the diet base, or by protein content. That is, there are two different categorical variables (diet base, protein) whose influence upon one continuous dependent variable (weight gain) we may wish to determine. This is the situation that *two-way ANOVA* deals with. We'll see much more detail in the coming lectures, but today we'll see how to implement two-way ANOVA in R.

The structure of our data could be thought of in the following way:

		Diet base		
		Beef	Cereal	Pork
Protein Level	High	10 observations	10 observations	10 observations
	Low	10 observations	10 observations	10 observations

We call diet base the *columns factor*, which has $c = 3$ *levels* (beef, cereal and pork) and we call protein level the *rows factor*, which has $r = 2$ *levels* (high and low). Each group has $m = 10$ observations. Overall there are $r \times c \times m = 60$ observations. We denote the k-th observation in row i, column j by Y_{ijk}.

The two-way ANOVA model is as follows:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \qquad i = 1, \ldots, r; \quad j = 1, \ldots, c; \quad k = 1, \ldots, m.$$

Let's break that down:

- μ is an overall grand mean, the mean of all rcm observations.
- α_i denotes the mean effect of being in the i-th row.
- β_j denotes the mean effect of being in the j-th column.
- γ_{ij} denotes the *interaction* effect. More on this below.
- ε_{ijk} are random independent errors with expected value 0 and variance σ^2.

So what are these "interaction" terms? Well, interaction effects represent the combined effects of factors on the dependent variable. When an interaction effect is present, the impact of one factor depends on the level of the other factor. Part of the power of ANOVA is the ability to estimate and test interaction effects. In our example, it could be the case that the effect of having high protein has a different effect on the rats' weight gains depending on which diet base is present – a two-way ANOVA will allow us to analyse this.

It turns out that (we'll see why in the lectures), much like in the one-way case, the TOTAL SS can be decomposed into the sum of sums of squares which correspond to row effects, column effects and interaction effects (plus a residual sum of squares). R can fit these models easily. Simply run `summary(aov(weightgain~protein*dietbase))` to fit the two-way ANOVA model.

In general, a two-way ANOVA table looks like this:

Source	SS	DF	MS	F-ratio
Row factor	Row SS	$r-1$	Row SS$/(r-1)$	MS_{ROW}/MS_{ERROR}
Column factor	Col SS	$c-1$	Col SS$/(c-1)$	MS_{COL}/MS_{ERROR}
Interaction Row:Column	R:C SS	$(r-1)(c-1)$	R:C SS$/((r-1)(c-1))$	$MS_{R:C}/MS_{ERROR}$
Error	Error SS	$rc(m-1)$	Error SS$/(rc(n-1))$	
Total (corr)	Total SS	$mcr-1$		

We'll cover how to calculate the various sums of squares in the coming lectures. However, much like in the one-way case, MS_{ERROR} is an unbiased estimator of σ^2 regardless of whether or not the α_i, β_i and γ_{ij} are zero. The other mean-squares (see MS column) have expected value σ^2 only if the corresponding factors are 0 and grow if not. Consequently the three given F ratios can respectively be used to test whether all row effects are zero (i.e. $\alpha_i = 0 \forall i$), all column effects are zero (i.e. $\beta_j = 0 \forall j$ or all interaction effects are zero ($\gamma_{ij} = 0 \forall i, j$). R gives us a p-value although note that it doesn't give us the 'Total' line of the table above. For our rat weight gain data, we see that dietbase has a large p-value; we would not reject the hypothesis that all the diet bases are equal. It seems therefore that the level of diet base (i.e. pork, beef, or cereal) does not have a significant effect on the rats' weight gains. What would you conclude about protein? How about the interactions? Are they significant?

One final tool for us to meet today is the *interaction plot*. Again, we'll discuss these more in the lectures. With this data, we seem to have borderline significant interactions (since the p-value is not too far from 5%). We can make an interaction plot by running `interaction.plot(protein,dietbase,weightgain)`. This simply plots the cell means for each cell in table on page 3.

We would interpret this plot as follows: we see that the lines for beef and pork are very nearly parallel; this suggests that moving from high to low protein affects the weight gain in a similar way for both diet bases (reducing from around 100g to around 80g in both cases). The cereal line is not parallel to the other two however. This suggests that the reduction in weight gained for the two levels of protein (High/Low) depends upon the level of diet base (beef/pork/cereal), with cereal behaving differently from the meat-based diets.

5 Exercises

1. (See data in `pollution.csv`) Nitrogen dioxide is a known pollutant, but its effects are not well known. A study was carried out of protein leakage in the lungs of mice exposed to nitrogen dioxide for A=10, B=12, and C=14 days. Half the mice were exposed to the nitrogen dioxide; the other half were not, and served as a control group. The response variable is the percent of serum fluorescence, with high values indicating more protein leakage. One third of the animals in the exposed and control groups had their serum fluorescence measured at 10, 12 and 14 days.

   Analyse the data.

2. (See data in `seniors.csv`) As part of a long term study of senior citizens, sociologists an physicians studied the relationship between geographical location and health status with regard to depression. The worksheet contains depression test scores on 120 individuals, of which

   - 60 were in good general health whilst 60 suffered from an existing medical condition;
   - 40 resided in Scotland, 40 in England and 40 in Wales.

   Analyse the data by using a two-way ANOVA model and report on your conclusions.