MA26620: Practical 9

Practical 9: One-way ANOVA

Good morning. We saw in the lecture this week about how different estimates of variances behaved differently if group means were unequal and said that ANOVA (ANalysis Of VAriance) exploits this. Today, we'll take this further and see how to use R to practically conduct one-way ANOVA procedures.

1 Recap of the theory

Most of Section 1 of this practical is a recap of what we've covered in the lectures. I'd recommend giving it a careful read and do ask questions on any bits that are hard to follow. Sections 1.4 and 1.5 are new; we haven't yet met those results in the lectures.

1.1 Expectations and variances of group means

As we covered in the lecture, the one-way ANOVA model deals with situations where observations occur in groups. We will examine situations where there are k groups, each consisting of m observations, making $n = mk$ observations in all.

The model used in this situation asserts that each observation in the i-th group has expectation μ_i. We further assume that the Y_{ij} are uncorrelated with constant variance σ^2, that is:

$$\mathbb{E}[Y_{ij}] = \mu_i, \qquad \mathrm{Var}(Y_{ij}) = \sigma^2, \qquad Y_{ij} \text{ uncorrelated, where } i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, m.$$

Recall that dots denote averaging over the respective subscript, so for instance $Y_{3\bullet}$ denotes the sample mean of the third group. This gives an unbiased estimate of the *population* mean for the third group, μ_3. Moreover, since each group is a random sample (we studied these extensively in semester one), the variance of the sample mean for each group is σ^2/m. Group means are uncorrelated, due to our assumption that observations Y_{ij} are uncorrelated.

Combining the above observations:

$$\mathbb{E}[Y_{i\bullet}] = \mu_i, \qquad \mathrm{Var}(Y_{i\bullet}) = \frac{\sigma^2}{m}, \qquad \mathrm{Cov}(Y_{i\bullet}, Y_{h\bullet}) = 0\ \forall i \neq h.$$

These results allow us to find the variance (and thus ESE) of any linear combination of the group means, for example:

$$\begin{aligned}\mathrm{Var}(Y_{1\bullet} - Y_{2\bullet}) &= \mathrm{Var}(Y_{1\bullet}) + \mathrm{Var}(Y_{2\bullet}) - 2\mathrm{Cov}(\mathrm{Var}(Y_{1\bullet}), \mathrm{Var}(Y_{2\bullet})) \\ &= \frac{\sigma^2}{m} + \frac{\sigma^2}{m} - 2 \times 0 = \frac{2\sigma^2}{m},\end{aligned}$$

or

$$\mathrm{Var}(Y_{1\bullet} - 2Y_{2\bullet} + 3Y_{3\bullet}) = \mathrm{Var}(Y_{1\bullet}) + 4\mathrm{Var}(Y_{2\bullet}) + 9\mathrm{Var}(Y_{3\bullet}) = \frac{14\sigma^2}{m}.$$

These expressions are in terms of σ^2, so **we need a way of estimating this** in an unbiased fashion from the data. To do this, we'll consider sums of squares.

1.2 Sums of squares

Various quantities can be thought of as corrected sums of squares:

- The (corrected) sum of squares of the **first group** is $\sum_{j=1}^{m}(Y_{1j} - Y_{1\bullet})^2$.
- This can be defined for **every other group** too.

- We can define a sum of squares **between** the $Y_{i\bullet}$. Since $Y_{\bullet\bullet}$ is the average of the $Y_{i\bullet}$s, we get $\sum_{i=1}^{k}(Y_{i\bullet} - Y_{\bullet\bullet})^2$.
- The **total** (corrected) sum of squares that considers all mk observations is:

$$\sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - Y_{\bullet\bullet})^2.$$

- It turns out that these are related by an incredibly useful algebraic identity (which we proved in the lectures):

$$\sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - Y_{\bullet\bullet})^2 = m\sum_{i=1}^{k}(Y_{i\bullet} - Y_{\bullet\bullet})^2 + \sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - Y_{i\bullet})^2,$$

  i.e. TOTAL SS = BETWEEN GROUPS SS + WITHIN GROUPS SS.

1.3 Expectations of sums of squares

So, "*what are the expectations of these sums of squares*", I hear you cry? *"Presumably something in terms of σ^2?"*

Well, you're quite right. We saw the gory details in this week's lecture on where these results come from, but the actual results were as follows:

$$\mathbb{E}\left[\text{WITHIN GROUPS SS}\right] = k(m-1)\sigma^2.$$

regardless of the values of the μ_i.

If we make the additional assumption that all μ_i are equal to a single value μ:

$$\mathbb{E}\left[\text{BETWEEN GROUPS SS}\right] = (k-1)\sigma^2,$$

and so

$$\mathbb{E}\left[\frac{\text{BETWEEN GROUPS SS}}{k-1}\right] = \sigma^2, \qquad \mathbb{E}\left[\frac{\text{WITHIN GROUPS SS}}{k(m-1)}\right] = \sigma^2.$$

It follows from the useful algebraic identity above that

$$\mathbb{E}[\text{TOTAL SS}] = (km-1)\sigma^2.$$

Thus **if all the group means are equal**, we have three ways of estimating σ^2 (since BETWEEN, WITHIN, and TOTAL SS are all computed from the observations).

If the group means are **NOT** equal, the within groups sum of squares can still be used, as WITHIN GROUPS SS divided by $k(m-1)$ is still an unbiased estimator of σ^2. It turns out (we proved this in yesterday's lecture) that without the equal group means assumption:

$$\mathbb{E}[\text{BETWEEN GROUPS SS}] = m\sum_{i=1}^{k}(\mu_i - \mu_{\bullet})^2 + (k-1)\sigma^2.$$

This gives us some motivation for testing whether the group means are in fact equal by comparing the quantities

$$\frac{\text{BETWEEN GROUPS SS}}{k-1} \text{ and } \frac{\text{WITHIN GROUPS SS}}{k(m-1)}.$$

In the case of group means all being equal, both should be about the same, but if the groups have different means, then the former tends to be bigger.

1.4 Normality

If we make the additional assumption of Normality, the following results hold:

WITHIN GROUPS SS $\sim \sigma^2\chi^2_{k(m-1)}$;

If the μ_i are all equal, then BETWEEN GROUPS SS $\sim \sigma^2\chi^2_{(k-1)}$.

If the μ_i are all equal, then TOTAL SS $\sim \sigma^2\chi^2_{(mk-1)}$.

Moreover, if the μ_i are all equal, then

$$\frac{\text{BETWEEN GROUPS SS}}{k-1} \bigg/ \frac{\text{WITHIN GROUPS SS}}{k(m-1)}$$

has an $F_{k-1,k(m-1)}$ distribution (if you haven't met the F distribution before, you might like to do a quick online search and read a bit about it. It has two parameters called the numerator degrees of freedom and the denominator degrees of freedom, and appears in your stats tables). The ratio will tend to grow if the means are not equal.

1.5 The Analysis of Variance Table

Phew, what a lot of stuff to remember! Fortunately, the whole procedure can be summarised in a so-called *one-way ANOVA table*:

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)	$\sum_{i=1}^{k}\sum_{j=1}^{m}(Y_{ij} - Y_{\bullet\bullet})^2$	$mk-1$			

Here, MS stands for mean square and the p-value labelled P is $P(F_{k-1,k(m-1)} > F\text{-ratio})$. Note also that the "Within groups" may also be referred to as "Error" or "Residual".

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.

2 Creating ANOVA tables in R

It probably goes without saying that computing an ANOVA table by hand would be a lengthy, tedious, and probably error-prone task (and yet one that a statistics student 40 years ago would have had to undertake). Fortunately R can come to the rescue and calculate all these sums of squares, degrees of freedom, mean squares, the F-ratio and P-value, pretty much instantly.

Let's see how.

2.1 Preliminaries

Download the `ratweightgain.csv` data from the module webpages and attach it. This data concerns 60 male rats of similar physical size who were fed one of six different diets (A,B,C,D,E,F). Ten rats were assigned to each diet and their weight gain (in grams) was measured.

Before we jump to an ANOVA table, let's investigate the data a little. Begin by finding the mean weight gain for each type of diet: `ybar <- tapply(weightgain,diet,mean)`. This command applies the R-function `mean` to the `weightgain` data, grouped according to the value of `diet`. Look at `ybar` to see the results. Adapt the `tapply` command to calculate the sample variances (`var`), standard deviations (`sd`) and group sizes (`length`) for each group.

It's a good idea to plot the data to get a feel for what's going on. Something like `boxplot(weightgain~diet)` will give you a quick boxplot (remember ggplot can make much nicer ones and if we were using the plot for anything more than a quick glance at the data, it'd need a title and well labelled axes).

2.2 The ANOVA model

To fit the one-way ANOVA model $\mathbb{E}[\text{weightgain}] = \mu_i$, with a different μ_i for each diet type, simply enter the command:

```
model1<-aov(weightgain~diet)
```

and then `summary(model1)` and try to interpret this.

We see from this table that our estimate of the error variance $\hat{\sigma}^2$ is 214.6 and this has 54 degrees of freedom (why 54?). We therefore see that our estimate of the variance of each group mean is 214.6/10 and the corresponding standard error is $\sqrt{214.6/10} = 4.6325$.

Would you believe that all the mean weight gains for all diets were the same? To do this we'd have to conduct the usual one-way ANOVA hypothesis test of testing H_0 : all means are equal against the alternative that they are not all equal. R has computed the relevant p-value for you; make sure you're confident with how to interpret it.

2.3 Post-ANOVA analysis

2.3.1 Pairwise t-tests

You might like to run the command `pairwise.t.test(weightgain,diet,p.adjust="none")`, which runs a t-test on each pair of diets (so low p-values suggest unequal means). However, tests applied will all have correlated results and the overall significance level (called the familywise error rate (FWER)) can be greatly different from that of individual tests. For example, with 6 groups, there are 15 ($={}^6C_2$) pairwise comparisons; the overall FWER could be as high as $1 - (1 - 0.05)^{15} = 0.54$.

Methods exist to adjust the error rate of each test to make the FWER more reasonable. We won't go into detail of how they work, but using `p.adjust="Bonferroni"` is a much more conservative test, while `p.adjust="Holm"` is similar to the Bonferroni test but tends to find more significant differences.

2.3.2 Tukey's Honest Significant Difference Test

This rather wonderfully named test also aims to find which means are significantly different to each other.

Run the command `TukeyHSD(model1)`.

This test sets the FWER at 95% (by default) and adjusts the individual significance values accordingly, and gives a 95% confidence interval for the differences in the means (the lower and upper ends of which are labelled `lwr` and `upr`). If 0 lies in this interval for a particular pair, then the difference between these two groups is insignificant.

The interpretation of this output can be quite challenging –it's best to use a structured approach:

(i) rank order the group averages from highest to lowest.

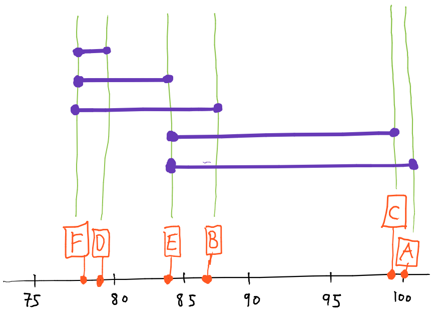
(ii) compare the largest with the smallest; if significant, compare the largest with the second smallest etc, continuing until the largest has been compared with the second largest, or a non-signifcant difference has been found.

(iii) continue comparing the second largest with the smallest, etc, as above, then the third largest etc.

(iv) at each stage, once a non-significant difference is located, conclude that there is no significant difference between any means enclosed by a non-signifcant pair,

(v) represent the overall conclusions about similarities and differences among the means using a schematic diagram involving the ordered sample averages, with lines drawn to connect means that are not significantly different.

R allows us to make a plot that will assist us in these endeavours. Run `plot(TukeyHSD(model1),las=1)` (the `las=1` here enables all the axis labels to fit on). So following the above procedure the groups in decreasing order of mean are A, C, B, E, D, F. Then $\mu_A - \mu_F$ and $\mu_A - \mu_D$ are significantly different from zero, but $\mu_A - \mu_E$ isn't (since zero lies within the 95% CI) so we'd draw a line joining A,C,B and E. Then we'd focus our attention on C. $\mu_C - \mu_F$ and $\mu_C - \mu_D$ are significantly different from zero, but $\mu_C - \mu_E$ isn't, so we'd draw a line joining C, B, E. And so on... check that you can arrive at a diagram like the one below.



Make sure you can reproduce this from R's output. We can then interpret this easily. For instance, we'd believe that $\mu_A = \mu_B$, because a line joins them. We wouldn't believe that $\mu_A = \mu_D$, nor that $\mu_C = \mu_D$ since a line doesn't join them.

3 Exercises

1. An engineer conducts an experiment with the aim of evaluating the maximum force that steel rods manufactured in different countries can withstand before breaking. She tests five rods from each of six countries. The following ANOVA table was computed:

Source	**SS**	**DF**	**MS**	**F-ratio**	**P**
Between countries					0.115
			10		
Total (corr)	340				

   (a) Copy and complete the table.

   (b) State a model underlying the analysis and carry out the usual hypothesis test, stating clearly and carefully your conclusions.

2. Researchers for the American Trading Standards journal Consumer Reports, June 1986, pp. 366-367 analysed results of a laboratory analysis of calories and sodium content of major hot dog brands. The researchers analyzed three types of hot dog: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat). The data are contained in the file `hotdogs.csv` on the module webpages under the variable names:

   - hotdog: Type of hotdog (beef, meat, or poultry)
   - calories: Calorific content per hot dog
   - sodium: Sodium content per hot dog

   Investigate for any differences in the mean calorific content for each type of hotdog.

3. Due to increased energy shortages and costs, utility companies are stressing ways in which bills can be cut. One company reached an agreement with the owner of a new apartment block to conduct a test of energy saving plans. Identical apartments were chosen for the study, similar in size, amount of shade, direction faced, etc. Four plans were to be tested, and the thermostat was set at a fixed level in each apartment. Monthly utility bills (£) were recorded.

Treatment A (no insulation)	**Treatment B** (wall and ceiling insulation)	**Treatment C** (no insulation, awnings for windows)	**Treatment D** (insulation and awnings)
74.44	68.75	71.34	65.47
89.96	73.47	83.62	72.33
82.00	71.23	79.98	70.87

   Analyse the data fully.