***Learning Statistics with R - University of Adelaide***

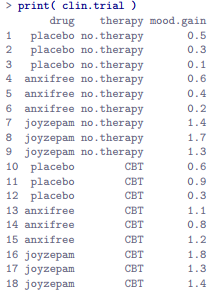
***Part V – Statistical Tools***

**14. Comparing Several Means (one-way ANOVA)**

* 1 of the most widely used tools in statistics = **the analysis of variance** = **ANOVA**.
* Basic technique = developed by Sir Ronald Fisher in early 20th century
* Term “ANOVA” is a little misleading, in 2 respects
* Although name refers to variances, ANOVA is concerned w/ investigating differences in means
* There are several different things that are all referred to as ANOVAs, some of which have only a very tenuous connection to one another.
* Range of different ANOVA methods that apply in quite different situations,
* Simplest form of ANOVA = several different groups of observations + are interested in finding out whether those groups differ in terms of some outcome variable of interest = **one-way ANOVA.**

**14.1 An illustrative data set**

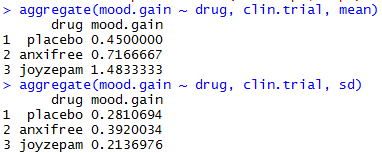
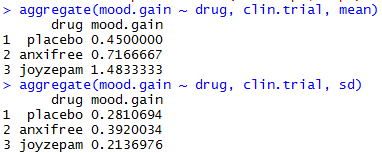
* Suppose you’ve become involved in a clinical trial in testing a new antidepressant drug Joyzepam.
* In order to construct a fair test of drug’s effectiveness, the study involves 3 separate drugs to be administered = yours, a placebo, an existing antidepressant/anti-anxiety drug Anxifree.
* A collection of 18 participants w/ moderate to severe depression are recruited for initial testing.
* B/c the drugs are sometimes administered in conjunction w/ psychological therapy, study includes 9 people undergoing cognitive behavioral therapy (CBT) + 9 who are not.
* Participants are randomly assigned (doubly blinded) a treatment, such that there are 3 CBT people + 3 no-therapy people assigned to each of the 3 drugs.
* A psychologist assesses mood of each person after a 3 month run w/ each drug + overall improvement in each person’s mood is assessed on a scale ranging from -5 to 5



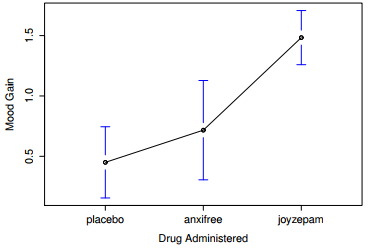
* Interested in the effect of drug on **mood.gain**
* **1st** first thing to do = calculate descriptive statistics + draw some graphs
* See how many people we have in each group:



* Calculate means + SDs for mood.gain variable broken down by drug



* Plot the average mood gain for all 3 conditions;



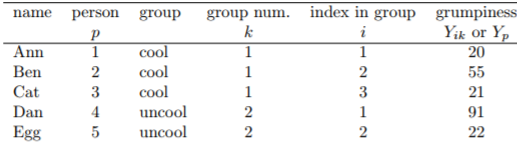
* Error bars show 95% CI’s
* As the plot makes clear 🡪 larger improvement in mood for Joyzepam participants than for either Anxifree or the placebo.
* Anxifree shows a larger mood gain than the control group, but the difference isn’t as large.
* The question that we want to answer is: are these difference “real”, or are they just due to chance?

**14.2 How ANOVA works**

* The experimental design described strongly suggests we’re interested in comparing the *average mood change* for 3 different drugs = an analysis similar to the t-test but involving > 2 groups.
* Let µP = population mean for the mood change induced by the placebo + let µA + µJ = corresponding means for our 2 drugs, Anxifree + Joyzepam
* Testing H0 = all 3 population means are identical (neither of the 2 drugs is any more effective than a placebo) 🡺 **H0: it is true that µP = µA = µJ**
* Our alternative = *at* *least one* of the 3 different treatments is different from the others.
* There are quite a few different ways in which the null can be false.
* For now just write the alternative as **H1: it is not true that µP = µA = µJ**
* This null is a trickier to test than any of the ones we’ve seen previously.
* Start out by playing around w/ variances + this gives us a useful tool for investigating means
* Use G = total # of groups = 3 + N = total sample size = 18 w/ N(k) = # of people in k-th group = 6 for all 3 groups.
* When all groups have the same # of observations, the experimental design is said to be **balanced**
* Not a big deal for one-way ANOVA but becomes more important for more complicated ANOVAs.
* Finally, use Y = outcome variable = mood change 🡪 Specifically, use Y(i, k) = mood change experienced by the i-th member of the k-th group.
* Similarly, use Y¯ = average mood change taken across all 18 people in experiment, + Y¯(k) = average mood change experienced by the 6 people in group k. Excellent
* Recall the formula for the sample variance but applied for Y



* Pretty much identical to the regular formula for the variance but only difference is this time = 2 summations 🡪 over groups (i.e., values for k) + over the people *within* groups (i.e., values for i).
* The only reason we have a double summation 🡪 b/c we classified people into groups, + then assigned numbers to people within groups.
* Consider a table w/ total N = 5 people sorted into G = 2 groups.



* See a person variable p = it would be perfectly sensible to refer to Y(p) as grumpiness of p-th person in the sample.
* Dan 🡪 p = 4 🡪 could say Y(p) = 91
* Alternative 🡪 note Dan belongs to the group k = 2 + is the 1st person listed in the group (i = 1)
* So it’s equally valid to refer to Dan’s by saying Y(i, k) = 91
* Each person p corresponds to a unique (i, k) combo,



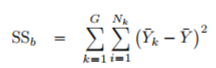
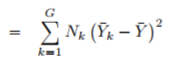
* Using Y(p) is clearly the simpler of the 2 , however when doing ANOVA it’s important to keep track of which participants belong in which groups 🡪 need to use the Y(i, k) notation to do this.
* **Total sum of squares, SS(tot)** 🡪 instead of *averaging* squared deviations (variance), *add* them up

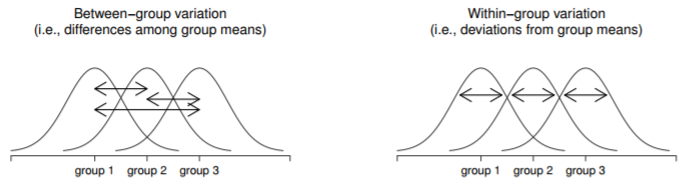


* When talking about analysing variances in the context of ANOVA, what we’re *really* doing is working w/ the total sums of squares rather than the actual variance.
* 1 very nice thing about the SS(tot) is we can break it up into 2 different kinds of variation.
* **Within-group sum of squares** (**SS(w)**) = how different each individual person is from their own group mean where Y¯(k) = a group mean (average mood change for the k-th drug)



* Instead of comparing individuals to the average of ALL people in the experiment, only comparing them to those people in the *same group*.
* As a consequence, you’d expect SS(w) to be smaller than SS(tot) b/c it’s completely ignoring any group differences (the fact that drugs (if they work) will have different effects on people’s moods)
* **Between-group sum of squares** (**SS(b)**) = looking at differences between group means Y¯(k) + **grand mean** Y¯



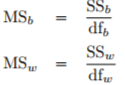
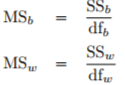
* Not too difficult to show the total variation among people in the experiment SS(tot) is actually the sum of the differences between the groups SS(b) + the variation inside the groups SS(w).



* Total variability associated w/ the outcome variable (SS(tot)) can be mathematically carved up into the sum of the variation due to differences in sample means for different groups (SS(b)) + all the rest of the variation (SS(w)).
* If the null is true, you’d expect all sample means to be pretty similar to each other 🡪 implies SS(b) is really small (or at least a lot smaller than the variation associated w/ everything else, SS(w))
* Qualitative idea behind ANOVA = compare the 2 sums of squares values SS(b) + SS(w) to each other
* If SS(b) is large relative to SS(w), we have reason to suspect the population means for the different groups *aren’t* identical to each other.
* To convert this into a workable hypothesis test, there’s a little bit of fiddling around needed.
* 1st calculate our test statistic, **an F ratio** to get a feel for why we do it this way.
* To convert SS values into an F-ratio, 1st calculate is dF associated w/ the SS(b) + SS(w) values.
* As usual, dF = # of unique DP’s that contribute to a particular calculation minus the # of constraints they need to satisfy.
* For SS(w), calculate variation of individual observations (N DP’s) around the group means (G constraints)
* In contrast, for SS(b), interested in variation of group means (G DP’s) around the grand mean (1 constraint)



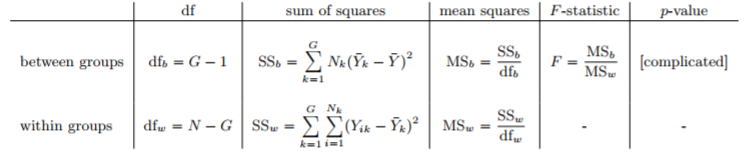
* Then convert summed squares value into a **mean squares value** by dividing by the dF:

* Calculate the F-ratio by *dividing* the *between*-groups MS by the *within*-groups MS



* At a general level, intuition behind the F statistic is straightforward 🡺 bigger values of F = between-groups variation is large relative to within-groups variation.
* Larger value of F = more evidence we have against the null.
* But how large does F have to be in order to actually reject H0?
* To understand this, need a slightly deeper understanding of what ANOVA is + what mean squares actually are.
* To complete hypothesis test, need to know the sampling distribution for F if the null is true = an **F distribution** 🡪 has 2 parameters, corresponding to the 2 dF involved (between + w/in groupds)
* Summary of all the key quantities involved in a one-way ANOVA



* At a fundamental level, ANOVA = competition between 2 different statistical models, H0 + H1.
* Recall, our null = all group means are identical to 1 another.
* If so, a natural way to think about the outcome variable Y(i, k) is to *describe individual scores in terms of a single population mean µ*, plus the *deviation from that population mean*. This deviation is usually denoted ε(i, k) = **error/residual** associated w/ that observation.
* Just like w/ the word “significant”, the word “error” has a technical meaning in statistics that isn’t quite the same as its everyday English definition.
* In statistics, it means *leftover variability* = stuff the model can’t explain.
* In any case, here’s the null when we write it as a statistical model:



* where we make the assumption that the residual values ε(i, k) are normally distributed (mean 0 + a SD σ that is the same for all groups)



* The only difference between the null + the alternative is we *allow each group to have a different population mean.*
* Let µk = population mean for the k-th group, then the statistical model corresponding to H1 is:



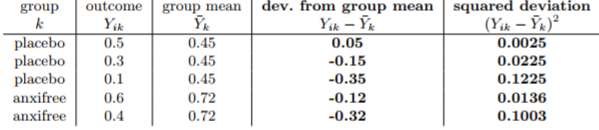
* where, once again, assume the error terms are normally distributed w/ mean 0 + SD σ.
* It’s now straightforward to say what the mean square values are measuring + what this means for the interpretation of F.
* It turns out **within-groups mean square, MS(w),** can be viewed as an estimator (in the technical sense) of the **error variance** σ^2 .
* The **between-groups mean square MS(b)** is also an estimator; but of the error variance *plus a quantity that depends on the true differences among the group means*.
* Call this quantity **Q**, then we can see the F-statistic is basically



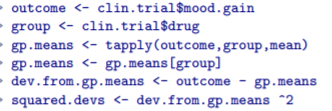
* It turns out **Q** refers to a **weighted mean of the squared treatment effects**



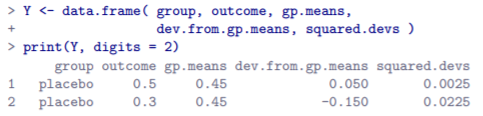
* The true value Q = 0 if the null is true, + Q > 0 if the alternative is true
* Therefore, at a bare minimum the F value must be > 1 to have any chance of rejecting the null.
* This doesn’t mean it’s impossible to get an F-value less than 1.
* What it means is if the null is true, the sampling distribution of the F ratio has mean = 1, + so we need to see F-values > 1 in order to safely reject the null.
* Or, if we want to be sticklers for accuracy, 
* To be a bit more precise about the sampling distribution, notice that if the null is true, both MS(b) + MS(w) are estimators of the variance of the residuals ε(i, k)
* If those residuals are normally distributed, you might suspect the estimate of the variance of ε(i, k) is chi-square distributed, b/c (Section 9.6) that’s what a chi-square distribution is = what you get when you square a bunch of normally-distributed things + add them up
* Since the F distribution is, by definition, what you get when you take the ratio between 2 things that are χ2 distributed, we have our sampling distribution.
* Worked Example 🡪 back to clinical trial data introduced at the start of the chapter.
* Descriptive statistics calculated tell us our group means: average mood gain = 0.45 for placebo, 0.72 for Anxifree, + 1.48 for Joyzepam.
* For 1st 5 observations, start by calculating SS(w)
* Next write down, for each person in the study, the corresponding group mean, Y¯(k).
* Then calculate – again for every person – the deviation from the corresponding group mean
* Then square everything.



* Now add up squared deviations across all observations: **SS(w) = 0.0025 + 0.0225 + 0.1225 + 0.0136 + 0.1003 = 0.2614**



* Look closely at these commands 1 at a time. Every single 1 is something you’ve seen before

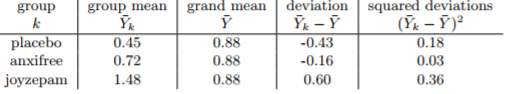




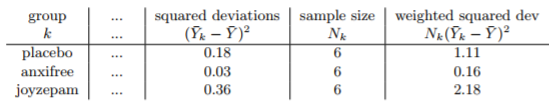




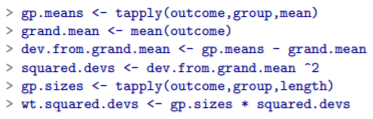
* Now for between-group sum of squares, SS(b) but instead of calculating differences between an observation Y(i, k) + a group mean Y¯(k) for all of the observations, calculate differences between group means Y¯(k) + the grand mean Y¯ (in this case 0.88) for all groups

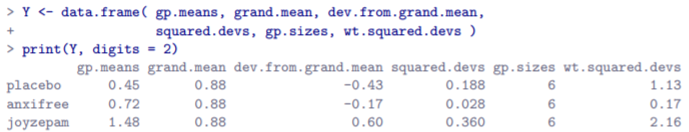


* However, for between group calculations we need to multiply each squared deviation by N(k), the number of observations in the group.
* Do this b/c every observation in a group (N(k) of them) is associated w/ a between group difference
* So if there are 6 people in the placebo group + placebo group mean differs from the grand mean by 0.19, the total between group variation associated w/ these 6 people is 6ˆ0.16 = 1.14.



* **SS(b) = 1.11 + 0.16 + 2.18 = 3.45**
* SS(b) calculations are a lot shorter



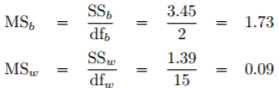




* Now that we’ve calculated sums of squares values, SS(b) + SS(w), the rest of the ANOVA is painless
* Next, calculate the dF 🡪 G = 3 groups + N = 18 observations in total



* Now obtain the mean square values



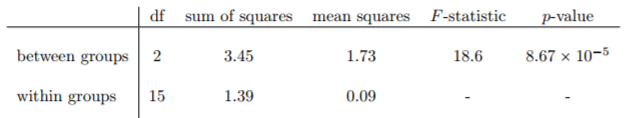
* The mean square values can be used to calculate the F-value, the test statistic we’re interested in.



* Now to find out whether the test gives us a significant result.
* What we really ought to do is choose an **α level** (i.e., acceptable Type I error rate) ahead of time, construct a rejection region, etc.
* In practice it’s just easier to directly calculate the p-value.
* F-test is always 1 sided + we only reject the null for very large F-values = only interested in the upper tail of the F-distribution.



* Therefore, p-value = 0.0000867, or 8.67ˆ10-5 in scientific notation.
* Unless we’re being extremely conservative about Type I error rate, we’re pretty much guaranteed to reject the null.
* At this point, basically done + having completed calculations, it’s traditional to organize all these numbers into an ANOVA table



* Get used to reading them + although software will output a *full* ANOVA table, there’s almost never a good reason to include the *whole* table in a write up.
* A pretty standard way of reporting this result would be to write something like this:



**14.3 Running an ANOVA in R**

* I’m pretty sure I know what you’re thinking after reading the last section, especially if you followed my advice and tried typing all the commands in yourself.... doing the ANOVA calculations yourself sucks. There’s quite a lot of calculations that we needed to do along the way, and it would be tedious to have to do this over and over again every time you wanted to do an ANOVA. One possible solution to the problem would be to take all these calculations and turn them into some R functions yourself. You’d still have to do a lot of typing, but at least you’d only have to do it the one time: once you’ve created the functions, you can reuse them over and over again. However, writing your own functions is a lot of work, so this is kind of a last resort. Besides, it’s much better if someone else does all the work for you... - 437 - 14.3.1 Using the aov() function to specify your ANOVA To make life easier for you, R provides a function called aov(), which – obviously – is an acronym of “Analysis Of Variance”.5 If you type ?aov and have a look at the help documentation, you’ll see that there are several arguments to the aov() function, but the only two that we’re interested in are formula and data. As we’ve seen in a few places previously, the formula argument is what you use to specify the outcome variable and the grouping variable, and the data argument is what you use to specify the data frame that stores these variables. In other words, to do the same ANOVA that I laboriously calculated in the previous section, I’d use a command like this: > aov( formula = mood.gain ~ drug, data = clin.trial ) Actually, that’s not quite the whole story, as you’ll see as soon as you look at the output from this command, which I’ve hidden for the moment in order to avoid confusing you. Before we go into specifics, I should point out that either of these commands will do the same thing: > aov( clin.trial$mood.gain ~ clin.trial$drug ) > aov( mood.gain ~ drug, clin.trial ) In the first command, I didn’t specify a data set, and instead relied on the $ operator to tell R how to find the variables. In the second command, I dropped the argument names, which is okay in this case because formula is the first argument to the aov() function, and data is the second one. Regardless of how I specify the ANOVA, I can assign the output of the aov() function to a variable, like this for example: > my.anova <- aov( mood.gain ~ drug, clin.trial ) This is almost always a good thing to do, because there’s lots of useful things that we can do with the my.anova variable. So let’s assume that it’s this last command that I used to specify the ANOVA that I’m trying to run, and as a consequence I have this my.anova variable sitting in my workspace, waiting for me to do something with it... 14.3.2 Understanding what the aov() function produces Now that we’ve seen how to use the aov() function to create my.anova we’d better have a look at what this variable actually is. The first thing to do is to check to see what class of variable we’ve created, since it’s kind of interesting in this case. When we do that... > class( my.anova ) [1] "aov" "lm" ... we discover that my.anova actually has two classes! The first class tells us that it’s an aov (analysis of variance) object, but the second tells us that it’s also an lm (linear model) object. Later on, we’ll see that this reflects a pretty deep statistical relationship between ANOVA and regression (Chapter 15) and it means that any function that exists in R for dealing with regressions can also be applied to aov objects, which is neat; but I’m getting ahead of myself. For now, I want to note that what we’ve created is an aov object, and to also make the point that aov objects are actually rather complicated beasts. I won’t be trying to explain everything about them, since it’s way beyond the scope of an introductory statistics subject, but to give you a tiny hint of some of the stuff that R stores inside an aov object, let’s ask it to print out the names() of all the stored quantities... 5Actually, it also provides a function called anova(), but that works a bit differently, so let’s just ignore it for now. - 438 - > names( my.anova ) [1] "coefficients" "residuals" "effects" [4] "rank" "fitted.values" "assign" [7] "qr" "df.residual" "contrasts" [10] "xlevels" "call" "terms" [13] "model" As we go through the rest of the book, I hope that a few of these will become a little more obvious to you, but right now that’s going to look pretty damned opaque. That’s okay. You don’t need to know any of the details about it right now, and most of it you don’t need at all... what you do need to understand is that the aov() function does a lot of calculations for you, not just the basic ones that I outlined in the previous sections. What this means is that it’s generally a good idea to create a variable like my.anova that stores the output of the aov() function... because later on, you can use my.anova as an input to lots of other functions: those other functions can pull out bits and pieces from the aov object, and calculate various other things that you might need. Right then. The simplest thing you can do with an aov object is to print() it out. When we do that, it shows us a few of the key quantities of interest: > print( my.anova ) Call: aov(formula = mood.gain ~ drug, data = clin.trial) Terms: drug Residuals Sum of Squares 3.4533 1.3917 Deg. of Freedom 2 15 Residual standard error: 0.30459 Estimated effects may be unbalanced Specificially, it prints out a reminder of the command that you used when you called aov() in the first place, shows you the sums of squares values, the degrees of freedom, and a couple of other quantities that we’re not really interested in right now. Notice, however, that R doesn’t use the names “between-group” and “within-group”. Instead, it tries to assign more meaningful names: in our particular example, the between groups variance corresponds to the effect that the drug has on the outcome variable; and the within groups variance is corresponds to the “leftover” variability, so it calls that the residuals. If we compare these numbers to the numbers that I calculated by hand in Section 14.2.5, you can see that they’re identical... the between groups sums of squares is SSb “ 3.45, the within groups sums of squares is SSw “ 1.39, and the degrees of freedom are 2 and 15 repectively. 14.3.3 Running the hypothesis tests for the ANOVA Okay, so we’ve verified that my.anova seems to be storing a bunch of the numbers that we’re looking for, but the print() function didn’t quite give us the output that we really wanted. Where’s the F-value? The p-value? These are the most important numbers in our hypothesis test, but the print() function doesn’t provide them. To get those numbers, we need to use a different function. Instead of asking R to print() out the aov object, we should have asked for a summary() of it.6 When we do that... > summary( my.anova ) Df Sum Sq Mean Sq F value Pr(>F) 6 It’s worth noting that you can get the same result by using the command anova( my.anova ). - 439 - drug 2 3.45 1.727 18.6 8.6e-05 \*\*\* Residuals 15 1.39 0.093 --- Signif. codes: 0 \*\*\* 0.001 \*\* 0.01 \* 0.05 . 0.1 1 ... we get all of the key numbers that we calculated earlier. We get the sums of squares, the degrees of freedom, the mean squares, the F-statistic, and the p-value itself. These are all identical to the numbers that we calculated ourselves when doing it the long and tedious way, and it’s even organised into the same kind of ANOVA table that I showed in Table 14.1, and then filled out by hand in Section 14.2.5. The only things that are even slightly different is that some of the row and column names are a bit different.