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

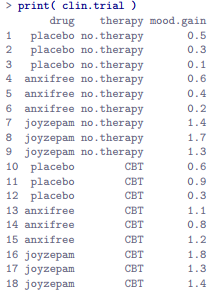
***Part V – Statistical Tools***

**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**

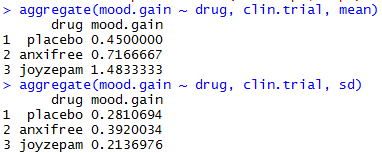
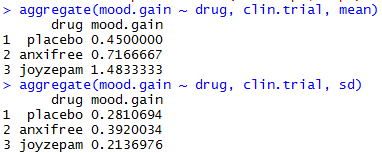
* Suppose you’ve become involved in a clinical trial in testing a new antidepressant drug Joyzepam.
* In order to construct a fair test of drug’s effectiveness, the study involves 3 separate drugs to be administered = yours, a placebo, an existing antidepressant/anti-anxiety drug Anxifree.
* A collection 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



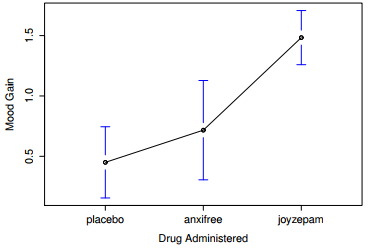
* Interested in the effect of drug on **mood.gain**
* **1st** first thing to do = calculate descriptive statistics + draw some graphs
* See how many people we have in each group:



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



* Plot the average mood gain for all 3 conditions;



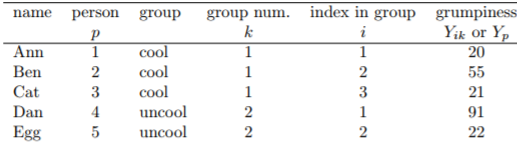
* Error bars show 95% CI’s
* As the plot makes clear 🡪 larger improvement in mood for Joyzepam participants than for either Anxifree or the placebo.
* Anxifree shows a larger mood gain than the control group, but the difference isn’t as large.
* The question that we want to answer is: are these difference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



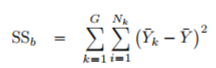
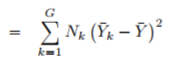
* Using Y(p) is clearly the simpler of the 2, however when doing ANOVA it’s important to *keep track of which participants belong in which groups* 🡪 need to use the Y(i, k) notation to do this.
* **Total sum of squares, SS(tot)** 🡪 instead of *averaging* squared deviations (variance), *add* them up

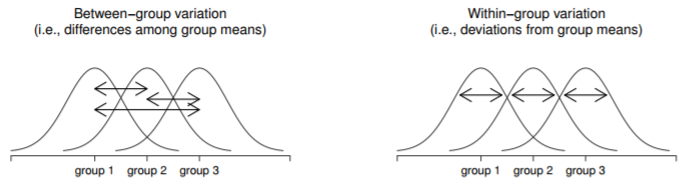


* When talking about analyzing variances in the context of ANOVA, what we’re *really* doing is working w/ the **total sums of squares** rather than the actual variance.
* 1 very nice thing about the SS(tot) is we can break it up into 2 different kinds of variation.
* **Within-group sum of squares** (**SS(w)**) = 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¯



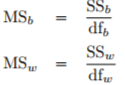
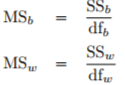
* Not too difficult to show the total variation among people in the experiment SS(tot) is actually the sum of the differences between the groups SS(b) + the variation inside the groups SS(w).



* 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)



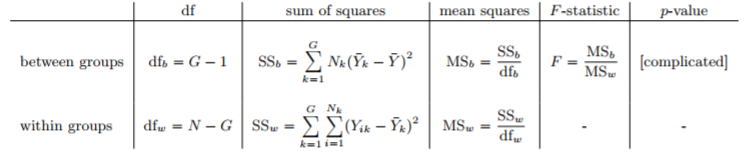
* Then convert summed squares value into a **mean squares value** by dividing by the dF:

* 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:



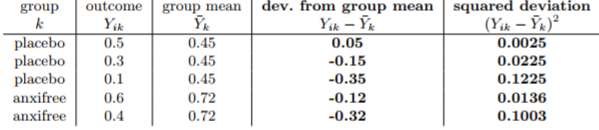
* where, once again, assume the error terms are normally distributed w/ mean 0 + SD σ.
* It’s now straightforward to say what the mean square values are measuring + what this means for the interpretation of F.
* It turns out **within-groups mean square, MS(w),** can be viewed as an estimator (in the 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



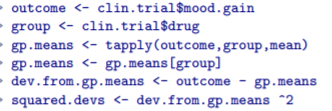
* It turns out **Q** refers to a **weighted mean of the squared treatment effects**



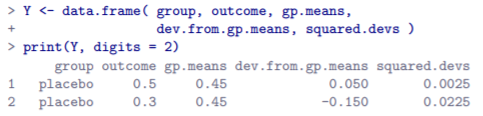
* The true value Q = 0 if the null is true, + Q > 0 if the alternative is true
* Therefore, at a bare minimum the F value must be > 1 to have any chance of rejecting the null.
* This doesn’t mean it’s impossible to get an F-value less than 1.
* What it means is if the null is true, the sampling distribution 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

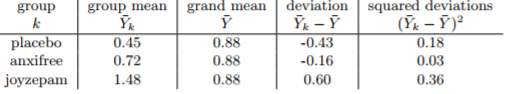




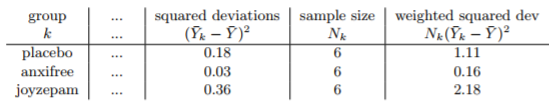




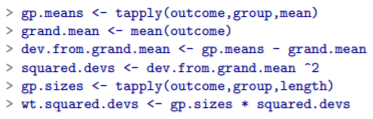
* Now for **between-group sum of squares, SS(b)** but instead of calculating differences between an observation Y(i, k) + a group mean Y¯(k) for all of the observations, calculate differences between group means Y¯(k) + the grand mean Y¯ (in this case 0.88) for all groups

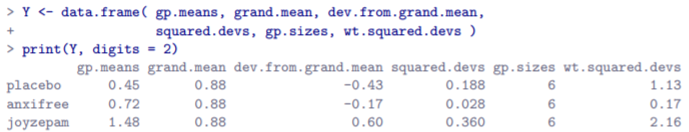


* However, for between group calculations we need to multiply each squared deviation by N(k), the number of observations in the group.
* Do this b/c every observation in a group (N(k) of them) is associated w/ a between group difference
* So if there are 6 people in the placebo group + placebo group 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



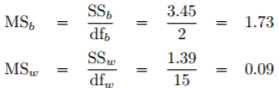




* Now that we’ve calculated sums of squares values, SS(b) + SS(w), the rest of the ANOVA is painless
* Next, calculate the dF 🡪 G = 3 groups + N = 18 observations in total



* Now obtain the mean square values



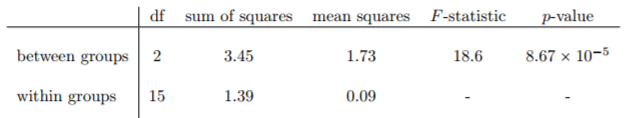
* The mean square values can be used to calculate the F-value, the test statistic we’re interested in.



* 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:

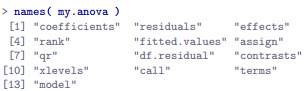


**14.3 Running an ANOVA in R**

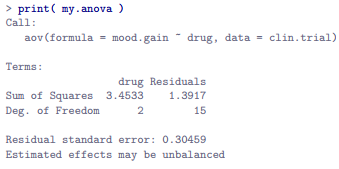
* Several arguments to **aov()** but only interested in 2 = formula (outcome + grouping vars) + data

* 
* my.anova actually has 2 classes
* 1st class tells us it’s an analysis of variance object + the 2nd tells us it’s also a linear model object
* There is a pretty deep statistical relationship between ANOVA + regression 🡪 any function in R for dealing w/ regressions can also be applied to aov objects, which is neat
* aov objects = rather complicated beasts.
* Print out the names of all the stored quantities.



* aov() does a lot of calculations for you, not just basic ones outlined in previous sections.
* This means it’s generally a good idea to create a variable like my.anova to store output of aov() b/c later on, you can use my.anova as an input to lots of other functions that pull out bits + pieces from an aov object + calculate various other things you might need.
* Simplest thing you can do w/ an aov object = print it + see a few of the key quantities of interest



* Notice R doesn’t use the names “between-group” and “within-group” + instead tries to assign more meaningful names
* Between-groups variance = the effect a **drug** has on the outcome variable = SS(b) = 3.45
* Within-groups variance = the “leftover” variability = **residuals** = SS(w) = 1.39
* To get F-ratio + p-value, use a different function 🡪 ask for a summary (or **anova(my.anova)**)





* Get the SS’s, dF’s, mean squares, F-statistic, + p-value organized into ANOVA table.

**14.4 Effect size**

* A few different ways to measure effect size in an ANOVA, but most common = **η2 (eta squared)** + **partial η2**
* For a one way ANOVA, they’re identical, so for the moment just look at η2

* Straightforward interpretation of η2 = the proportion of the variability in the outcome variable (mood.gain) that can be explained in terms of the predictor (drug).
* η2 = 0 means there is no relationship at all between the 2, whereas η2 = 1 means the relationship is perfect.
* Better yet, η2 is *very closely related to a squared correlation (i.e., r2).*
* So, if trying to figure out whether a particular value of η2 is big or small, it’s sometimes useful to remember that:



* Can be interpreted as if it referred to the *magnitude of a Pearson correlation*.
* η2 = .71 corresponds to η = .84.
* If we think about this as being *equivalent to a* ***correlation*** *of about .84*, we’d conclude the relationship between drug + mood.gain is strong.
* Core R packages don’t include any functions for calculating η2 but it’s straightforward to calculate directly from the numbers in an ANOVA table.



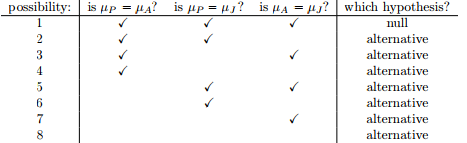
* Since it’s tedious to do this the long way (especially when running more complicated ANOVAs) lsr has etaSquared() + only care about **x** argument = aov object corresponding to an ANOVA



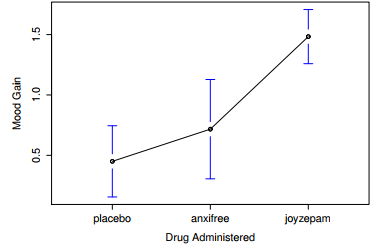
* Shows 2 different numbers
* 1st corresponds to η2 statistic
* 2nd refers to **partial η2**, a somewhat different measure of effect size.
* For our simple drug ANOVA, they’re the same number 🡪 won’t always be true w/ complicated ANOVAs

**14.5 Multiple comparisons and post hoc tests**

* Any time you run an ANOVA w/ > 2 groups + end up w/ a significant effect, 1st thing to ask = *which groups are actually different from one another?*
* Ex: Our null = all 3 drugs (placebo, Anxifree and Joyzepam) have exact same effect on mood.
* But if you think about it, the null is actually claiming 3 different things all at once here.
* Specifically, it claims that:
* Competitor’s drug (Anxifree) is no better than a placebo (i.e., µA = µP )
* Your drug (Joyzepam) is no better than a placebo (i.e., µJ = µP )
* Anxifree + Joyzepam are equally effective (i.e., µJ = µA)
* If *ANY* 1 of those 3 claims is false, the null is also false.
* We’ve rejected our null 🡪 at least 1 of those things isn’t true. But WHICH?
* All 3 of these propositions are of interest 🡪 certainly want to know if your new drug Joyzepam is better than a placebo + would be nice to know how well it stacks up against an existing commercial alternative (Anxifree).
* Would even be useful to check performance of Anxifree against the placebo (even if Anxifree has already been extensively tested against placebos by other researchers, it can still be very useful to check that your study is producing similar results to earlier work)
* When we characterize the null in terms of these 3 distinct propositions, it becomes clear there are 8 possible **states of the world** we need to distinguish between:



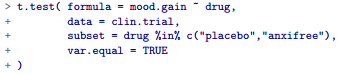
* By rejecting the null, we’ve decided we don’t believe #1 is the true state of the world.
* Next question = *WHICH* of the other 7 possibilities do we think is right?
* When faced with this situation, usually helps to look at the data.
* Ex: Look at plots



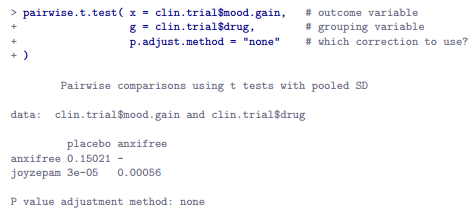
* Tempting to conclude Joyzepam is better than both a placebo + Anxifree, but there’s no real difference between Anxifree + the placebo.
* However, to get a clearer answer about this, it might help to run some tests
* Running pairwise t-tests
* Given we’ve got 3 separate pairs of means (placebo vs. Anxifree, placebo vs. Joyzepam, + Anxifree vs. Joyzepam) to compare, could do is run 3 separate t-tests + see what happens.
* A couple of ways to do this.
* 1) Construct new variables corresponding groups you want to compare (e.g., anxifree, placebo + joyzepam), + then run a t-test on these new variables



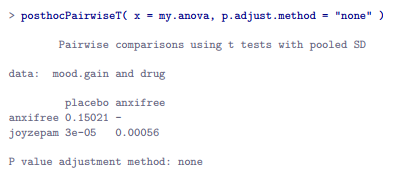
* 2) Use subset argument in **t.test()** to select only observations corresponding to 1 of the 2 groups we’re interested in:



* Regardless of which version we do, R will print out the results of the t-test
* If we go on to do this for all possible pairs of variables, can look to see which (if any) pairs of groups are significantly different to each other.
* This “lots of t-tests idea” isn’t a bad strategy, but there are some problems w/ it.
* For the moment, bigger problem = it’s a pain to type in such a long command over + over again
* If an experiment has 10 groups, you have to run 45 t-tests
* To keep typing to a minimum, R provides **pairwise.t.test()** toautomatically runs all t-tests for you.
* 3 arguments to specify = outcome variable x, group variable g, + **p.adjust.method** argument, which adjusts the p-value in 1 way or another (for now just set p.adjust.method = none)



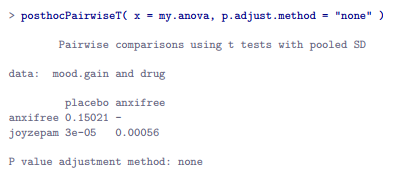
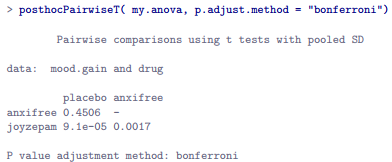
* Can’t just give it an aov object, + have it produce this output even though R has actually stored enough info inside it should just be able to get it to run all the pairwise tests using one as an input.
* There’re other functions in R for running multiple comparisons + at least 1 works this way
* **TukeyHSD()** takes an aov object as its input + outputs **Tukey’s honestly significant difference tests**
* **posthocPairwiseT()** in lsr package lets you do this.
* Can just input an **aov object** itself + get the pairwise tests as an output.
* Actually just a simple way of calling pairwise.t.test() function, but be aware of changes later on



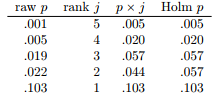
* Suppose you’ve run an ANOVA + stored the results in my.anova, + you’re happy using the **Holm correction** (default method in pairwise.t.test()).
* In that case, all you have to do is type  + R will output the test results
* Corrections for multiple testing
* The concern w/ just running lots + lots of t-tests is when running these analyses, what we’re doing is going on a fishing expedition 🡺 running lots of tests w/out much theoretical guidance in the hope that some of them come up significant.
* This kind of “theory-free search” for group differences = **post hoc analysis** (“after this”)
* If you DO have some theoretical basis for wanting to investigate SOME comparisons but not others, it’s a different story.
* In those circumstances you’re not really running “post hoc” analyses at all but making **planned comparisons** (Section 16.9).
* It’s okay to run post-hoc analyses, but a lot of care is required.
* Analysis ran in the previous section is actually pretty dangerous 🡺 each individual t-test is designed to have a 5% Type I error rate (α = .05), + we ran 3 of these tests.
* Imagine what would have happened if an ANOVA involved 10 different groups + we decided to run 45 post hoc t-tests to try to find out which ones were significantly different from each other
* You’d expect 2 or 3 to come up significant by chance alone.
* As in Chapter 11, the central organizing principle behind **null hypothesis testing** is *we seek to control our Type I error rate*
* Butnow, running lots of t-tests at once, in order to determine the source of ANOVA results, the *actual* Type I error rate *across this whole family of tests has gotten completely out of control.*
* Usual solution to this problem = introduce an **adjustment** to the p-value = aims to control **total error rate** across the family of tests
* An adjustment of this form (usually (but not always) applied b/c one is doing post hoc analysis\_ is often referred to as a **correction for multiple comparisons**, or sometimes as **simultaneous inference**
* In any case, there are quite a few different ways of doing this adjustment, but be aware there are many other methods out there
* The simplest of these adjustments = the **Bonferroni correction**
* Suppose a post-hoc analysis consists of m separate tests + we want to ensure the total probability of making any Type I errors at all is *at most* α = .9
* Worth noting in passing that NOT all adjustment methods try to do this.
* Thus is an approach for controlling **family-wise Type I error rate**
* However, there are other post hoc tests seek to control the **false discovery rate**, a somewhat different thing
* If so, the **Bonferroni correction** just says multiply all raw p-values by m.
* If we let p = the original p-value, + let p`(j) = the *corrected* value, the Bonferroni correction tells that:



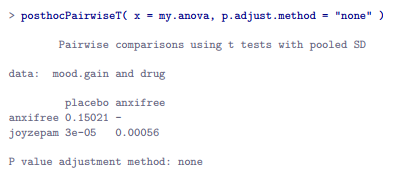
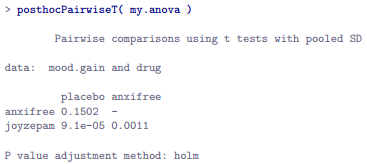
* Therefore, if using the Bonferroni correction, you’d reject the null if p` < α.
* The logic behind this correction is very straightforward.
* We’re doing m different tests, so if we arrange it so that each test has a Type I error rate of *at most* α/m, then the total Type I error rate across these tests *cannot be larger than α*.
* That’s pretty simple, so much so that in the original paper, the author writes*: The method given here is so simple + so general that I am sure it must have been used before this.*
* To use a Bonferroni correction in R, use **pairwise.t.test()**, making sure you set **p.adjust.method = bonferroni.**
* There’s also **p.adjust() 🡪** input vector of raw p-values + it outputs a vector of adjusted p-values
* This can be handy sometimes
* Also note more advanced users may wish to consider using tools provided by the **multcomp** package.
* Alternatively, since the whole reason we’re doing these pairwise tests in the 1st place is b/c we have an ANOVA we’re trying to understand, it’s probably more convenient to use **posthocPairwiseT()** in the lsr, since we can use an aov object as the input:

* Compare these p-values to those when we made no adjustment 🡪 clear that the only thing R has done is *multiply them by 3*.
* Although the Bonferroni correction is the simplest adjustment out there, it’s not usually the best
* 1 method often used instead = **Holm correction**
* Idea behind Holm correction = to pretend you’re doing tests sequentially, starting w/ the smallest (raw) p-value + moving onto the largest one.
* For the j-th largest of the p-values, the adjustment is EITHER
*  (i.e., the biggest p-value remains unchanged, the 2nd biggest p-value is doubled, the 3rd biggest p-value is tripled, + so on),
* OR 
* Which adjustment is performed = *whichever one is smaller.*
* What the Holm correction *does*:
* 1st: Sort all of p-values in order, from smallest to largest.
* For the smallest p-value, all you do is multiply it by m, + you’re done.
* However, for all others, it’s a 2-stage process.
* Ex: when you move to the 2nd smallest p value, 1st multiply it by m -1
* If this produces a number > the adjusted p-value you got last time, then keep it.
* But if it’s smaller than the last one, copy the *last p-value*.
* To illustrate how this works, consider the table below, which shows calculations of a Holm correction for a collection of 5 p-values:



* Although it’s a little harder to calculate, the Holm correction has some very nice properties:
* More powerful than Bonferroni (i.e., lower Type II error rate), but, counterintuitive as it might seem, has the same Type I error rate.
* As a consequence, in practice = never any reason to use a simpler Bonferroni correction since it is always outperformed by the slightly more elaborate Holm correction
* B/c of this, the Holm correction is the default one used by pairwise.t.test() + posthocPairwiseT().

* As you can see, the biggest p-value (comparison between Anxifree + the placebo) is unaltered at a value of .15 = exactly the same as when we applied no correction at all.
* In contrast, the smallest p-value (Joyzepam vs. placebo) has been multiplied by 3
* Finally, having run the post-hoc analysis to determine which groups are significantly different to one another, you might write up the result like this: 
* Or, if you don’t like the idea of reporting exact p-values, you’d change those numbers to p < .01, p < .001 + p > .05 respectively.
* Either way, the key thing = indicate that you used Holm’s correction to adjust the p-values
* And of course, that elsewhere in the write up you’ve included relevant descriptive statistics (i.e., group means + SD’s), since these p-values on their own aren’t terribly informative

**14.6 Assumptions of one-way ANOVA**

* Like any statistical test, ANOVA relies on some assumptions about the data.
* 3 key assumptions you need to be aware of**: normality, homogeneity of variance, independence**.
* Remember the statistical models underpinning ANOVA:



* where µ = a single population mean (same for all groups) + µk = population mean for group k
* Up to this point, we’ve been mostly interested in whether our data are best described in terms of a *single grand mean* (the null) or in terms of *different group-specific means* (the alternative).
* *This makes sense, of course, as that’s actually the important research question*
* However, all our testing procedures have (implicitly) relied on a specific assumption about the residuals, ε(i, k), namely that  (that they’re defined by a normal distribution)
* *None of the math works properly without this bit*.
* Or, to be precise, you can still DO all the calculations + end up w/ an F-statistic, but you have *no guarantee THIS F-statistic actually measures what you think it’s measuring*
* So, any conclusions you might draw on the basis of that F test might be wrong.
* How do we check whether this assumption about the residuals is accurate?
* 3 distinct claims buried in this 1 statement
* **Normality =** residuals are assumed to be normally distributed.
* can assess this w/ **QQ plots** or running a **Shapiro-Wilk test**.
* **Homogeneity of variance/homoscedasticity** = only have 1 value for population SD (i.e., σ), rather than allowing each group to have its own value (σk).
* ANOVA assumes the population SD = the same for all groups.
* **Independence** = this assumption is a little trickier.
* It basically means that knowing 1 residual tells you NOTHING about *any other residual*.
* All ε(i, k) values are assumed to have been generated w/out any “regard for”/“relationship to” any of the other ones.
* No obvious/simple way to test for this, but some situations that are clear violations of this:
* Ex: In a repeated-measures design (each participant in a study appears in more than 1 condition), independence doesn’t hold
* There’s a special relationship between some observations 🡪 those that correspond to the same person
* When that happens, you need to use something like **repeated measures ANOVA**.
* 1 question people often want to know the answer to = the extent to which you can trust the results of an ANOVA if the assumptions are violated.
* Or, to use technical language, *how robust is ANOVA to violations of the assumptions* (later version)*.*

**14.7 Checking the homogeneity of variance assumption**

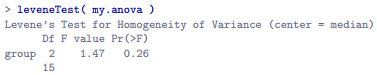
* There’s more 1 way to test the homogeneity of variance assumption
* Most commonly used tests = **Levene test** + the closely-related **Brown-Forsythe test**
* Could also use the **Bartlett test**, implemented in R via **bartlett.test()** if you’re interested.
* **Levene’s test =** shockingly simple.
* Suppose we have our outcome variable Y(i, k)
* All we do is define a new variable, Z(i, k), corresponding to the **absolute deviation from the group mean**



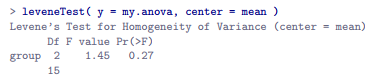
* Take a moment to think about what Z(i, k) actually is + what we’re trying to test.
* The value of **Z(i, k)** *= a measure of how the i-th observation in the k-th group deviates from its group mean.*
* Our null = all groups have the same variance = same overall deviations from the group means
* So, the null in a Levene’s test = *the population means of Z are identical for all groups.*
* So, we need a statistical test of the null of “all group means are identical” = ANOVA
* So, all Levene’s test does = run an ANOVA on the *new variable* Z(i, k).
* Brown-Forsythe test doesn’t do anything particularly different?
* Only change = it constructs the transformed variable Z in a slightly different way, using deviations from the group MEDIANS rather than from the group means.



* Regardless of whether doing the Levene or Brown-Forsythe test, the test statistic (sometimes denoted **F**, sometimes as **W**) is calculated in exactly the same way the F-statistic for a regular ANOVA is calculated 🡪 just using a Z(i, k) rather than Y(i, k).
* Obviously, since a Levene test is just an ANOVA, it would be easy enough to manually create the transformed variable Z(i, k) + use aov() to run an ANOVA on that.
* However, that’s tedious way 🡪 better way = use **leveneTest()** from the **car package**.



* Main argument = y
* can do this in lots of different ways, + probably the simplest way = actually input the original aov object
* See the test is non-significant, F(2,15) = 1.47, p = .26, so it looks like the homogeneity of variance assumption is fine (groups deviations aren’t significantly different)
* Remember, although R reports the test statistic as an F-value, it could equally be called W, in which case just write W(2,15) = 1.47.
* Also, note the output says *center = median* 🡪 tells you that, by default, *leveneTest()actually does the Brown-Forsythe test*.
* To use mean instead, explicitly set the **center** argument

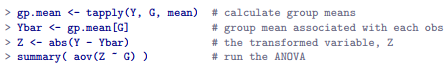


* That being said, in most cases it’s probably best to stick to the default value, since Brown-Forsythe is a bit more robust than the original Levene test.
* 2 more quick comments
* 1) As mentioned above, there are other ways of calling leveneTest().
* Although vast majority of situations that call for a Levene test involve checking the assumptions of an ANOVA (in which case you probably have a variable like my.anova), sometimes you might find yourself wanting to specify variables directly.
* 2 different ways that you can do this:



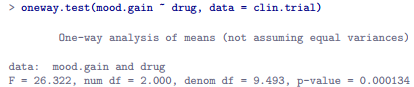
* > 2) It’s possible to run a Levene test just using aov()
* Here’s the code that creates the new variables + runs an ANOVA.
* If interested, run this to verify it produces same answers as a Levene test (w/ center = mean)



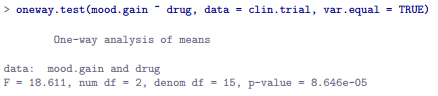


**14.8 Removing the homogeneity of variance assumption**

* In our example, the **homogeneity of variance**/**homoscedasticity** assumption turned out to be a pretty safe one 🡪 Levene test came back non-significant, so we probably don’t need to worry.
* However, in real life we aren’t always that lucky.
* How do we save an ANOVA when **homoscedasticity** is violated?
* We’ve seen this problem before 🡺 Student’s t-test assumes equal variances, so if broken, we used the Welch’s t-test, which does not assume this.
* In fact, Welch also showed how we can solve this problem for ANOVA too (**Welch one-way test**), implemented in R using **oneway.test()** w/ args = model formula (w/ outcome variable on left + grouping variable on the right), data frame containing the variables, + var.equal.
* If var.equal is FALSE (default) a Welch one-way test is run.
* If TRUE, it just runs a regular ANOVA.
* The function also has a **subset** argument to analyze only some observations + a **na.action** argument that tells it how to handle missing data (aren’t necessary for our purposes)
* So, to run the **Welch one-way ANOVA**:



* To understand what’s happening here, compare these numbers to what we got earlier w/ our original ANOVA (set var.equal = TRUE)

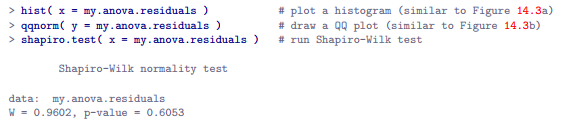


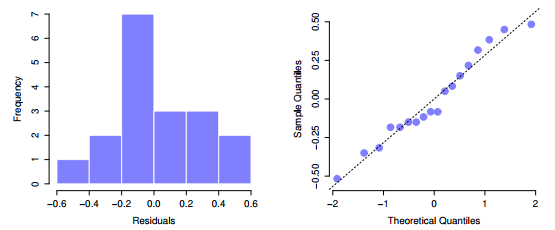
* Originally ANOVA gave us F(2, 15) = 18.6, whereas the Welch one-way test gave us F(2, 9) = 26.32
* In other words, the Welch one-way test reduced the within-groups dF from 15 to 9.49, + the F-value increased from 18.6 to 26.32

**14.9 Checking the normality assumption**

* Testing the normality assumption is relatively straightforward. (Section 13.9).
* Only thing we really need to know how to do is pull out the **residuals** (ε(i, k) values) so we can draw QQ plots + run Shapiro-Wilk tests.
* First, extract the residuals w/ **residuals()** on the aov object, draw plots, + run a hypothesis test







* The histogram + QQ plot both look pretty normal, which is supported by the results of Shapiro-Wilk test (W = .96, p < .61) which finds no indication that normality is violated

**14.10 Removing the normality assumption**

* Now we’ve seen how to check for normality, but what we can do to address violations of normality?
* In context of a 1-way ANOVA, easiest solution = probably to switch to **non-parametric tests** (don’t rely on any particular assumption about the kind of distribution involved).
* We’ve seen non-parametric tests before (Chapter 13)
* When you only have 2 groups, **Wilcoxon test** provides a non-parametric alternative you need.
* W/ 3+ groups, can use the **Kruskal-Wallis rank sum test**
* The **Kruskal-Wallis test** = surprisingly similar to ANOVA, in some ways.
* In ANOVA, started w/ Y(i, k), the value of the outcome variable, for the i-th person in the k-th group.
* For Kruskal-Wallis, *rank order* all these Y(i, k) values + *conduct analysis on the ranked data.*
* Let R(i, k) = the ranking given to the i-th member of the k-th group + calculate R¯k, the *average* rank given to observations in the k-th group:



* Also calculate R¯, the grand mean rank



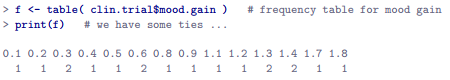
* Now calculate the squared deviations from the grand mean rank R¯.
* When we do this for *individual* scores (i.e. calculate (R(i, k) - ´R¯)^2 ), then we have a nonparametric measure of *how far the (i, k)-th observation deviates from the grand mean rank.*
* When calculating squared deviation of *group means from the grand means* (calculate (R¯k - R¯)^2, we have a nonparametric measure of *how much group k deviates from the grand mean rank.*
* W/ this in mind, follow same logic as in ANOVA
* Define ranked sums of squares measures in much the same way.
* 1) **Total ranked sums of squares** = 
* Define the **between-groups ranked sums of squares:** 
* So, if the null is true (there are no true group differences at all), you’d expect the between-group rank sums RSS(b) to be very small, much smaller than the total rank sums RSS(tot).
* Qualitatively, the Kruskal-Wallis test-statistic, K, is very much the same as when constructing the ANOVA F-statistic, but for technical reasons, it’s constructed in a slightly different way:



* If the null is true, the sampling distribution of K is approximately chi-square w/ G - 1 dF (where G = number of groups).
* The larger the value of K = less consistent the data are w/ the null = this is a one-sided test: reject H0 when K is sufficiently large.
* At a conceptual level, this is the right way to think about how the Kruskal-Wallis test works.
* However, from a purely mathematical perspective, it becomes needlessly complicated
* *Could* show that the equation for K can be rewritten as



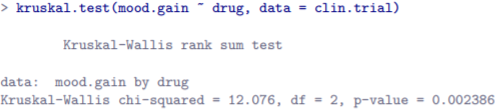
* Sometimes see this given for K = way easier to calculate than described in the previous section, but is totally meaningless to actual humans.
* Probably best to think of K as described it earlier 🡺 *as an analogue of ANOVA based on ranks.*
* But keep in mind the *test statistic that gets calculated ends up w/ a rather different look to it than the one we used for our original ANOVA.*
* ALSO, **this story is only actually true when there are no ties in the raw data** 🡪 no 2 observations have the exact same value.
* If there ARE ties, must introduce a **correction factor** to these calculations.
* Suppose we construct a frequency table for raw data, + let f(j) = # of observations w/ the j-th unique value.



* Looking at this table, notice the 3rd entry in the frequency table has a value of 2
* This corresponds to a mood.gain = 0.3 🡪 tells us 2 people’s mood increased by 0.3.
* More to the point, we can say that f[3] = 2, or f(3) = 2.
* Now that we know this, the **tie correction factor (TCF)** is



* The *tie-corrected value* of the Kruskal-Wallis statistic = obtained by dividing the value of K by this quantity, TCF 🡪 it is this tie-corrected version that R calculates.
* Running the test is pretty painless, since R has **kruskal.test(),** which is pretty flexible, + allows you to input data in a few different ways.
* Most of the time you’ll have data like our clinical trial data set = have an outcome variable (mood.gain), + a grouping variable (drug).



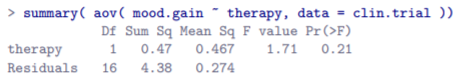
* High KW-1-way-ANOVA statistic, significant difference (p < .01)
* Sometimes it can be useful to specify x as a list
* Suppose you actually had data as 3 separate variables 🡪 placebo, anxifree + joyzepam.
* If your data are in that format, it’s convenient to know you can bundle all 3 together as a list



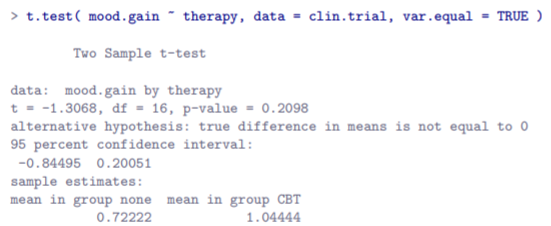
* This would give the exact same results as above

**14.11 On the relationship between ANOVA and the Student t test**

* 1 last thing to point out before finishing (something people find surprising but worth knowing about)
* An ANOVA w/ 2 groups = identical to the Student t-test.
* It’s not just that they’re *similar*, but are actually equivalent in every meaningful way.
* Single concrete demonstration 🡺 Suppose instead of running an ANOVA on the mood.gain ~ drug model, instead do it using **therapy** as the predictor.
* If we run this ANOVA, here’s what we get:



* Overall, looks like there’s no significant effect at all, BUT this is actually a misleading answer (Ch. 16)
* In any case, it’s irrelevant to our current goals: our interest here = the F-statistic + the p-value
* The F-statistic = F(1,16) = 1.71 + p-value = .21.
* Since we only have 2 groups 🡪 *didn’t actually need to resort to an ANOVA* 🡪 could’ve just decided to run a Student t-test.



* Curiously, the p-values are identical = .21.
* Having run a *t-test* instead of an ANOVA, we get a somewhat different answer, t(16) = -1.3068.
* However, there is a fairly straightforward relationship here.
* If we *square* the t-statistic 🡪 1.3068^2 = 1.7077,**we get the F-statistic from before.**
* Haven’t yet discussed any analog of the paired samples t-test for more than two groups = **repeated measures ANOVA** (will appear in a later version of this book) or how to run an ANOVA when interested in more than 1 grouping variable = Chapter 16.