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

***Part V – Statistical Tools***

**Chapter 13 - Comparing Two Means**

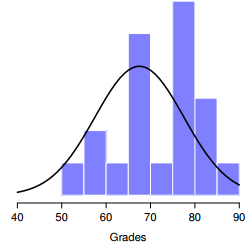
* In the previous chapter we covered the situation when the outcome variable is nominal scale + your predictor variable is *also* nominal scale.
* Lots of real world situations have that character, so you’ll find that chi-square tests in particular are quite widely used.
* However, you’re much more likely to find yourself in a situation where the outcome is interval scale or higher + what you’re interested in is **whether the average value of the outcome variable is higher in 1 group or another.**
* Ex: Psychologist might want to know if anxiety levels are higher among parents than non-parents, or if working memory capacity is reduced by listening to music (relative to not listening to music).
* In a medical context, might want to know if a new drug increases or decreases BP.
* Agricultural scientist might want to know if adding phosphorus to native plants will kill them
* In all these situations, the outcome is a fairly *continuous, interval or ratio scale variable* + our predictor is a binary **grouping** variable.
* In other words, we want to compare means of the 2 groups, + the standard answer to this problem of comparing means is to use a **t-test**, of which there are several varieties depending on exactly the question you want to solve.

**13.1 The one-sample z-test**

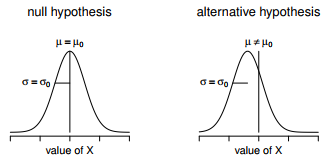
* z-test is almost never used in real life
* Only real purpose = a very convenient stepping stone along the way to the t-test, probably the most (over)used tool in all statistics.

**13.1.1 The Inference Problem That The Test Addresses**

* Ex: Grade introductory stats class on a curve w/ average grade = 67.5 + SD = 9.5.
* Of the many hundreds of students, 20 also take psychology.
* Do the psychology students tend to get the same grades as everyone else (i.e., mean = 67.5) or do they tend to score higher or lower?
* Sample mean of X¯ = 72.3 is a fair bit higher than the hypothesized population mean µ = 67.5
* But on the other hand, sample size n = 20 isn’t all that big. Maybe it’s pure chance.
* To answer the question, it helps to be able to write down what we think we know.
* We know sample mean X¯ = 72.3.
* If willing to assume psychology students have the same SD as the rest of the class, we can say the population SD is σ = 9.5.
* Also assume class is graded to a curve + the psychology student grades are normally distributed.
* Want to be clear about what we want to learn from the data 🡪 our research hypothesis relates to the *population* mean µ for the *psychology* student grades, which is *unknown*.
* Specifically, want to know if µ = 67.5 or not.
* Given what we know, can we devise a hypothesis test to solve our problem?
* The data, along w/ the hypothesized distribution from which they are thought to arise, are here:



* The thin black line is the theoretical distribution from which the psychology student grades (bars) are supposed to have been generated
* Not entirely obvious what the right answer is 🡪 need some statistics
* 1st step in constructing a hypothesis test = *Be clear about what the null + alternative are*.
* H0 = true population mean µ for psychology students = 67.5% (population) 🡺 µ = .675
* H1 = the population mean *isn’t* 67.5% 🡺 µ != .675 or µ <> .675



* In addition to providing us w/ hypotheses, the scenario outlined above provides us w/ a fair amount of background knowledge that might be useful.
* Specifically, there are 2 special pieces of info we can add:
* 1. Psychology grades are *normally distributed*
* 2. The *true SD* of these scores σ is known to be 9.5.
* For the moment, act as if these are absolutely trustworthy facts.
* In real life, this kind of absolutely trustworthy background knowledge doesn’t exist, so to rely on these facts, we have to make the *assumption* these things are true.
* However, since these assumptions may/may not be warranted, might need to check them.
* For now though, keep things simple.
* Next step = figure out what we would be a good choice for a **diagnostic test statistic**;
* i.e. something that would help discriminate between H0 + H1.
* Given that the hypotheses all refer to the psychology population mean µ, you’d feel pretty confident the sample mean X¯ would be a pretty useful place to start
* Could look at the difference between the sample mean X¯ + the value the null predicts for the population mean 🡺 
* If this quantity equals or is very close to 0, things are looking good for the null
* If this quantity is a long way away from 0, it’s less likely the null is worth retaining.
* But how far away from zero should it be for us to reject H0?
* To figure that out, we need to rely on those 2 pieces of background knowledge 🡺 the raw data are normally distributed + we know the value of the population SD σ.
* IF the null is actually *true* + the *true* mean = µ0, these facts together mean we know the complete population distribution of the data: *a normal distribution w/ mean µ0 + SD σ.*
* A statistician might write this as: 
* Okay, if that’s true, what can we say about the distribution of X¯?
* Remember the sampling distribution of the mean X¯ is also normal, + has mean µ.
* But the SD of this sampling distribution SE(X¯), the **standard error of the mean**, is:



* In other words, IF the null is *true*, the sampling distribution of the mean can be written as:



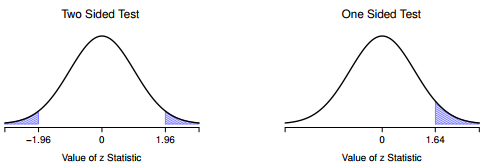
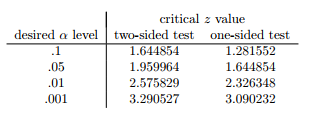
* Now comes the trick to convert the sample mean X¯ into a standard score, z, or z(X¯) = a standardized version *of a sample mean*, NOT a standardized version of a *single observation,* which is what a z-score usually refers to).
* The z-score for our sample mean is:

 OR 

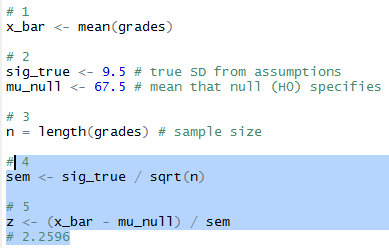
* This z-score is our test statistic + the nice thing about using it as our test statistic is, like all z-scores, *it has a standard normal distribution*:



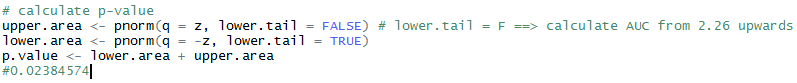
* In other words, regardless of what scale the original data are on, the **z-statistic** itself always has the same interpretation: **the # of SE’s that separate the observed sample mean X¯ from the population mean µ0 predicted by the null.**
* Better yet, regardless of what the population parameters for the raw scores actually are, the 5% critical regions for z-test are always the same



* Z-test is so rarely used in real life that base R doesn’t have a built in function for it.
* However, the test is so incredibly simple that it’s really easy to do one
* 1) Calculate sample mean
* 2) Create variables corresponding to the known population SD + the population mean that the null specifies
* 3) Create a variable for the sample size.
* 4) Calculate the (true) SE of the mean:
* 5) Calculate our z-score:



* At this point, we’d traditionally look up the value 2.26 in our table of critical values.
* Our original hypothesis was 2-sided (didn’t have a theory about whether psych students would be better or worse)
* Can see 2.26 is bigger than the critical value of 1.96 required to be significant at α .05, but smaller than the value of 2.58 required to be significant at a level of α .01.
* Therefore, we can conclude we have a significant effect, which we might write up by saying:
* “With a mean grade of 73.2 in the sample of psychology students, and assuming a true population standard deviation of 9.5, we can conclude that the psychology students have significantly different statistics scores to the class average (z = 2.26, N = 20, p < .05)”
* However, what if want an exact p-value?
* Back in the day, the tables of critical values were huge so you could look up your actual z-value, + find the smallest value of α for which your data would be significant (the very definition of a p-value).
* Notice the α level of a z-test (or any other test, for that matter) defines the total AUC for the critical region
* That is, if we set α = .05 for a 2-sided test, the critical region is set up such that the AUC for the critical region is .05.
* And, for the z-test, the critical value of 1.96 is chosen that way b/c the area in the lower tail (i.e., below -1.96) is exactly .025 + the area under the upper tail (i.e., above 1.96) is exactly .025.
* So, since our observed z-statistic is 2.26, why not calculate the AUC below -2.26 or above 2.26?



* Some tests make reasonable assumptions, other tests do not.
* The 1 sample z-test makes 3 basic assumptions
* **Normality**.
* z-test assumes the true population distribution is normal
* This is often pretty reasonable + it’s an assumption that we can check if we feel worried about it
* **Independence**
* The observations in your data set are not correlated w/ each other, or related to each other in some funny way.
* Isn’t as easy to check statistically + relies a bit on good experimental design.
* Have to ask yourself if it’s really plausible to imagine each observation is a completely random sample from the population you’re interested in.
* In practice, this assumption is never met; but we try our best to design studies that minimize the problems of correlated data.
* **Known standard deviation**
* True SD of the population is known to the researcher.
* This is just stupid. In no real world data analysis problem do you know the SD σ of some population, but are completely ignorant about the mean µ.
* In other words, this assumption is always wrong.

**13.2 The one-sample t-test**

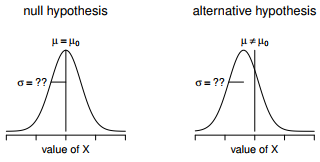
* After some thought, decide it might not be safe to assume psychology student grades necessarily have the same SD as the other students in the class.
* After all, if I’m entertaining the hypothesis that they *don’t* have the same mean, why believe they *absolutely* have the same DF?
* In view of this 🡪 stop assuming we the true value of σ = violates assumptions of a z-test
* Still got raw data which give an estimate of the population SD:



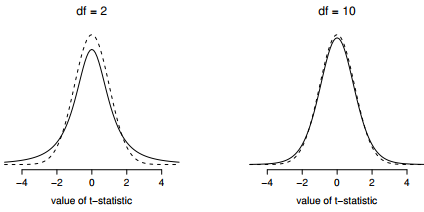
* Can’t say we *know* σ = 9.5 but *can* say σ^ = 9.52
* Running a z-test but w/ the estimated SD of 9.52 instead of relying on the assumption the true SD = 9.5 still gives a significant result, + this approach is close, but not quite correct.
* Now are relying on an *estimate* of the population SD, we need to *make some adjustment for the fact that we have some uncertainty about what the true population SD actually is*.
* Maybe our data are just a fluke + maybe the true population SD = 11
* If that were actually true + we ran the z-test assuming σ = 11, the result would end up being non-significant
* This annoying ambiguity was resolved in 1908 by William Sealy Gosset, a chemist for Guinness brewery at the time who published the work under the pseudonym “A Student” 🡪 **Student’s t-test**
* Partially helped us accommodate the fact that we aren’t completely sure what the true SD is
* Fisher found general solution to the problem
* Not knowing the true SD subtly changes the sampling distribution.
* **t-statistic** is calculated in *exactly the same way* mentioned above.
* If our null is “the true mean = µ”, but our sample has mean X¯ + our estimate of the population SD is σˆ, then our t statistic is:



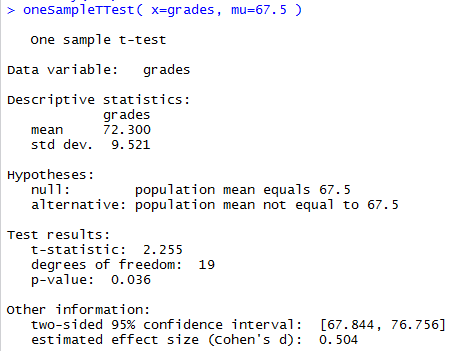
* Only thing that changed in the equation = instead of using known true value σ, use the estimate ˆσ.



* Graphical illustration of the null + alternative assumed by the (2-sided) 1 sample t-test.
* Note the similarity to the z-test
* The null = the population mean µ is equal to some specified value µ0,
* The alternative = It is not.
* Like the z-test, assume the data are normally distributed; but do NOT assume the population SD σ is known in advance
* If this estimate has been constructed from n observations, the sampling distribution turns into a **t-distribution** w/ n - 1 dF.
* The t-distribution is very similar to the normal distribution, but has heavier tails



* Notice the t-distribution’s heavier tails (higher **kurtosis**) than a normal distribution which is quite exaggerated when the dF are very small, but negligible for larger values.
* In other words, for large dF, the t-distribution is essentially identical to a normal distribution.
* This is as it should be: if you have a sample size of n = 70M, your estimate of the SD would be pretty much perfect
* **So, expect that for large n, the t-test behaves exactly the same way as a z-test**
* Mechanics of the t-test are almost identical to the mechanics of the z-test = not much point in going through the tedious exercise of showing you how to do the calculations using low level commands
* Pretty much identical to calculations did earlier except w/ the *estimated* SD
* Then test our hypothesis using the t-distribution rather than the normal distribution (**pt()** rather than **pnorm())**
* **t.test() =** very flexible (can run lots of different kinds of t-tests) + is somewhat terse (output = quite compressed).
* Simpler functions in **lsr**
* 1-sample t-test 🡪 **oneSampleTTest()** 🡪 specify x (variable containing the data) + mu, the true population mean *according to the null*.



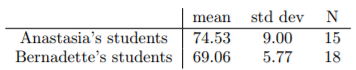
* Output is trying to lead you through the entire data analysis process.
* 1st 2 lines = what kind of test was run + what data were used.
* Then see some basic info about the sample 🡪 sample mean + SD
* Then moves towards inferential statistics 🡪 telling you what the null + alternative were + reports the results of the test:
* The t-statistic, the dF, + the p-value.
* Finally reports 2 other things you might care about: the **confidence interval** for the mean + a measure of **effect size**
* Result is statistically significant (i.e. p < .05) + could report the result by saying:
* “With a mean grade of 72.3, psychology students scored slightly higher than the average grade of 67.5 (t(19) = 2.25, p< .05); the 95% confidence interval is [67.8, 76.8].
* t(19) = shorthand for a t-statistic w/ 19 dF
* It’s often the case that people don’t report the CI, or do so using a much more compressed form
* Not uncommon to see the CI included as part of the **stat block**,

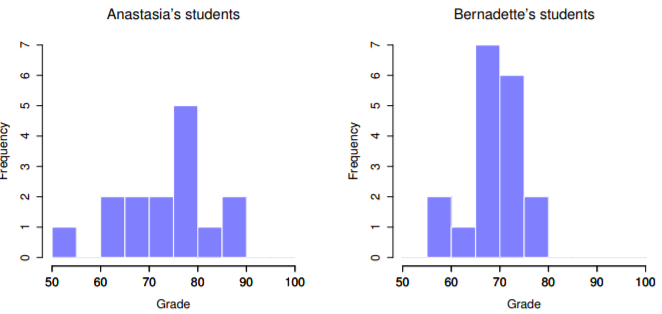


* Assumptions of the one sample t-test (same assumptions as the z-test, minus the one about the known SD)
* **Normality** 🡪 still assuming the population distribution is normal + there are standard tools you can use to check to see if this assumption is met + other tests you can do in its place if violated
* In the same way we can weaken the assumptions of a z-test so that we’re only talking about the *sampling* distribution, we can weaken the t-test assumptions so that we don’t have to assume normality of the *population*.
* However, for the t-test, it’s trickier to do this.
* As before, replace the assumption of population normality w/ an assumption of sampling distribution of X¯ normality.
* However, remember we’re also relying on a *sample estimate* of the DF + we also require the sampling distribution of σˆ to be chi-square.
* That makes things nastier this version is rarely used in practice
* Fortunately, if the population is normal, then both of these 2 assumptions are met.
* **Independence** 🡪 must assume observations in our sample are generated independently of one another. See the earlier discussion about the z-test for specifics
* Overall, these 2 assumptions aren’t terribly unreasonable, + as a consequence a 1-sample t-test is widely used in practice as a way of comparing a sample mean vs. a hypothesized population mean.

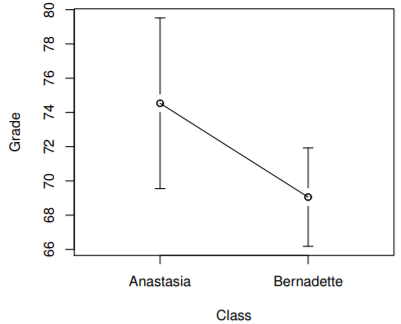
**13.3 The independent samples t-test (Student test)**

* 1-sample t-test has its uses but is not the most typical example of a t-test
* Much more common situation = 2 different groups of observations.
* Ex: psychology 🡪 2 different groups of participants corresponding to different conditions in a study
* For each person in study: measure an outcome variable of interest + research question = whether or not the 2 groups have the same population mean = **independent samples t-test**
* Suppose 33 students in a statistics lectures not graded to a curve + don’t know anything about the average grade for the class as a whole.
* There are 2 tutors for the class, A + B w/ N1 = 15 students in A’s tutorials + N2 = 18 in B’s tutorials.
* Research question = whether A or B is a better tutor, or if it doesn’t make much of a difference.





* A’s students may be getting slightly better grades on average, but also seem a bit more variable



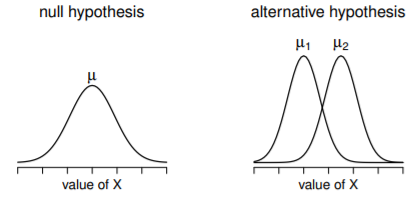
* Error bars depict 95% CI’s around the mean.
* On the basis of visual inspection, it looks like there’s a real difference between the groups, though it’s hard to say for sure.
* The independent samples t-test comes in two different forms, Student’s and Welch’s.
* The original **Student t-test** = simpler of the 2, but relies on *much more restrictive assumptions*
* Assuming you want to run a 2-sided test, the goal = to determine whether 2 independent samples of data are drawn from populations w/ the same mean (the null) or different means (the alternative).
* When we say “independent” samples = really mean there’s no special relationship between observations in the 2 samples.
* If we have an experimental design where participants are randomly allocated to 1 of 2 groups, + we want to compare the 2 groups’ mean performance on some outcome measure, then an **independent samples t-test** (rather than a *paired samples* t-test) is what we’re after.
* µ1 = true population mean for group 1 (A’s students), + µ2 = true population mean for group 2 (B’s)
* What is the population being referred to in this case?
* Is it the set of students actually taking the class? The set of people who *might* take the class (an unknown number) of them? Or something else? Does it matter which of these we pick?
* Yes, it does matter: if you change your definition of what the “real world” population actually is, the sampling distribution of your observed mean X¯ changes too.
* The t-test relies on an assumption that observations are sampled at random from an infinitely large population
* To the extent that real life isn’t like that, the t-test can be wrong.
* In practice, however, this isn’t usually a big deal
* Even though the assumption is almost always wrong, it doesn’t lead to a lot of pathological behavior from the test, so we tend to just ignore it
* Let X¯ 1 + X¯ 2 = the observed sample means for both groups.
* Our null = The 2 population means are identical (µ1 = µ2) + the alternative is they’re not (µ1 != µ2).



* To construct a hypothesis test that handles this scenario, start by noting that if the null is true, then the difference between the population means is *exactly* 0
* As a consequence, a **diagnostic test statistic** will *be based on the difference between the 2 sample means.*
* This is b/c if the null is true, we’d expect X¯1 - X¯2 to be pretty close to 0.
* However, just w/ our 1-sample tests (z-test or t-tests), must be precise about exactly *how close to zero* this difference should be
* Calculate a SE estimate just like last time + divide the difference between means by this estimate 🡪 **t-statistic:**



* Need to figure out what this SE estimate actually is, a bit trickier than for the 2 prior tests
* A **pooled estimate** of the SD
* In the original Student t-test: make the assumption that the 2 groups have the *same* population SD
* That is, regardless of whether the population means are the same, assume the population SD’s are identical, σ1 = σ2.
* Since we’re assuming the 2 SDs are the same, we drop the subscripts + refer to both of them as σ.



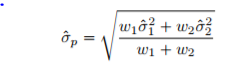
* The null assumes both groups have the same mean µ, whereas the alternative assumes they have different means µ1 + µ2.
* Notice it is assumed the population distributions are normal, + that, although the alternative allows the group to have different means, it assumes they have the same SD
* How should we estimate this? How should we construct a single estimate of a SD when we have 2 samples?
* The answer: Take a *weighted average of the variance estimates* =**pooled estimate of the variance**.
* The weight assigned to each sample = # of observations in that sample - 1.



* Now that we’ve assigned **weights** to each sample, calculate the **pooled estimate of the variance** by taking the weighted average of the 2 variance estimates



* Finally, convert a pooled variance estimate to a **pooled standard deviation estimate** by taking the square root 🡪 ˆσ(p):



* Our data set actually corresponds to a set of N observations, sorted into 2 groups.
* Use X(i, k) = refers to the grade received by the i-th student in the k-th tutorial group
* X(1,1) = grade received by the 1st student in A’s class, X(2, 1) = her 2nd student, + so on.
* We have 2 separate group means X¯1 + X¯2, which we could generically refer to using the notation X¯k (mean grade for the k-th tutorial group)
* Since every single student falls into 1 of the 2 tutorials, + we can describe their deviation from the group mean as the difference X(i, k) - X¯(k)
* Why not just use *these* deviations (extent to which each student’s grade differs from the mean grade in *their* tutorial)?
* Remember, a **variance** = just the average of a bunch of squared deviations



* But, calculating the variance by dividing by N produces a *biased estimate of the population variance.*
* Need to divide by N - 1 to fix this
* The reason this bias exists is b/c the *variance estimate* relies on the *sample mean +* to the extent the sample mean *isn’t* equal to the population mean, *it can systematically bias our estimate of variance*
* 2 sample means = more bias 🡪 now need to divide by N - 2 instead of N – 1 to calculate a pooled variance estimate



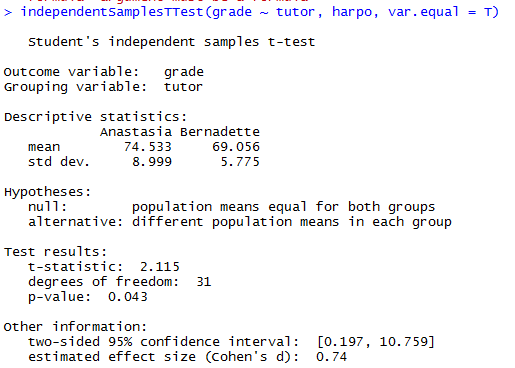
* If you take the square root of this then you get ˆσ(p), the **pooled SD estimate**
* Pooled SD calculation is it’s not terribly different to the regular SD calculation
* From now on, drop the p subscript, + just refer to this estimate as ˆσ.
* Whole reason for calculating this pooled estimate was we knew it would be helpful when calculating our SE estimate, but, SE of what?
* In the 1-sample t-test 🡪 SE of the *sample mean*, SE(X¯) = σ / Sqrt(N)
* This time we have 2 sample means + are interested in the difference between the two X¯1 - X¯2’s
* As a consequence, the SE we need to divide by is in fact the **SE of the *difference between means***.
* As long as the 2 variables really do have the same SD, our estimate for the SE is

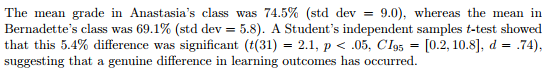


* Our **t-statistic** is therefore



* Just as we saw with our 1-sample test, the sampling distribution of this t-statistic is a t-distribution as long as the null is true + all of the assumptions of the test are met.
* The dF, however, is slightly different: As usual, think of dF = # of DP’s - # of constraints.
* Here = N observations (N1 in sample 1, + N2 in sample 2) + 2 constraints (the sample means)
* So the total dF for this test are N - 2.
* Can run an independent samples t-test using **t.test()**
* A somewhat simpler function in lsr package = **independentSamplesTTest()**



* Output has a very familiar form.
* 1st tells you what test was run, names of the variables used, reports the sample means + SDs for both groups, states the null + alternative in a fairly explicit form, then reports the test results: a t-statistic, the dF, + the p-value.
* Final section reports 2 things: a CI + an effect size.
* It’s pretty important to be clear on what *this* CInterval actually refers to: it is a *CI for the difference between the group means.*
* In our example, A’s students had an average = 74.5, + B’s students had an average = 69.1, so the difference between the 2 sample means is 5.4.
* But of course the difference between *population* means might be bigger or smaller than this.
* The CI reported says there’s a 95% chance the *true* difference between means lies between 0.2 + 10.8
* In any case, the difference between the 2 groups is significant (just barely), so we might write up the result using text like this
* 
* Notice that I’ve included the CI + effect size in the stat block. People don’t always do this.
* At a bare minimum, you’d expect to see the t-statistic, dF + p value.
* If statisticians had their way, everyone would also report the CI + probably the effect size measure too, b/c they are useful things to know
* Make a judgment based on whether you think it will help your readers (+ if you’re writing a scientific paper, the editorial standard for the journal in question)
* 1 very common worry students have when they start running t-tests is they often end up w/ negative values for the t-statistic + don’t know how to interpret it
* Assuming a 2-sided test, the p-values will be identical for positive + negative t-values w/ the same absolute value
* On closer inspection, notice the CI’s also have opposite signs.
* Whenever this happens, the 2 versions arise from slightly different ways of running the t-test.
* A t-statistic R is calculating here is always of the form



* If mean 1 > mean 2, the t-statistic will be positive, if mean 2 is larger the t statistic will be negative.
* Similarly, the CI R reports is the CI for the difference mean1 – mean2, which will be the reverse of what you’d get if the CI for the difference mean2 – mean1
* Which we should call “mean 1” + “mean 2” is arbitrary but you *need to designate them*
* Whenever I get a significant t-test result + want to figure out which mean is larger, just look at the actual group means, not the t-statistic
* It really doesn’t matter what R printed out, try to report the t-statistic in such a way that the numbers match up w/ the text.
* Suppose we want to write in a report is “A’s class had higher grades than B’s class”.
* The phrasing here implies A’s group comes first, so it makes sense to report the t-statistic as if A’s class corresponded to group 1.
* If so, write “A’s class had higher grades than B’s class (t(31) = 2.1, p = .04)”
* On the other hand, suppose I wanted to use B’s class listed first 🡪 makes more sense to treat *her* class as group 1, + if so, write: “B’s class had lower grades than A’s class (t(31) = -2.1, p < .04).
* **Please note that you can’t do this for other types of test statistics.**
* It works for t-tests, but wouldn’t be meaningful for chi-square tests, F-tests, or indeed for most tests
* *Don’t overgeneralize this advice*!
* As always, our hypothesis test relies on some assumptions. For the Student t-test there are 3
* **Normality 🡪** Like a 1-sample t-test, it is assumed the data are normally distributed.
* Specifically, assume both groups are normally distributed
* **Independence 🡪** assumed the observations are independently sampled.
* In the context of the Student test this has 2 aspects to it:
* 1) assume the observations w/in each sample are independent of 1 another (same as for the 1-sample test).
* 2) Also assume there are NO **cross-sample dependencies.**
* If, for instance, it turns out you included some participants in both *experimental conditions* of a study (e.g., accidentally allowing the same person to sign up to different conditions), there are some cross sample dependencies you’d need to take into account.
* **Homogeneity of variance** (**homoscedasticity**) = the population SD is the same in both groups.
* Can test this assumption using the **Levene test**