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

***Part V – Statistical Tools***

**Chapter 13 - Comparing Two Means**

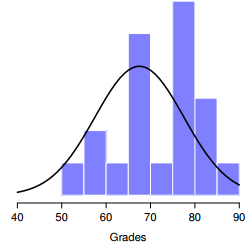
* In the previous chapter we covered the situation when the outcome variable is nominal scale + your predictor variable is *also* nominal scale.
* Lots of real world situations have that character, so you’ll find that chi-square tests in particular are quite widely used.
* However, you’re much more likely 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**

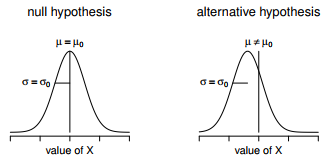
* z-test is almost never used in real life
* Only real purpose = a very convenient stepping stone along the way to the t-test, probably the most (over)used tool in all statistics.

**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:



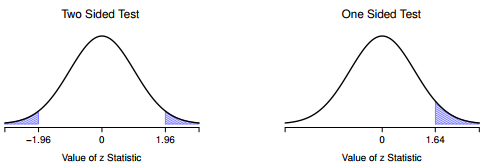
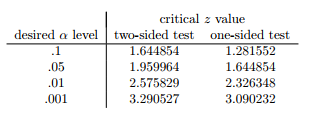
* Now comes the trick to convert the sample mean X¯ into a standard score, z, or z(X¯) = a standardized version *of a sample mean*, NOT a standardized version of a *single observation,* which is what a z-score usually refers to).
* The z-score for our sample mean is:

 OR 

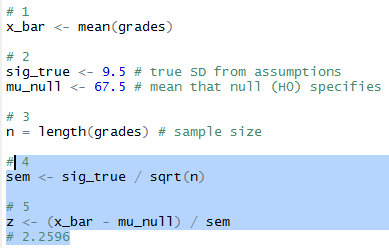
* This z-score is our test statistic + the nice thing about using it as our test statistic is, like all z-scores, *it has a standard normal distribution*:



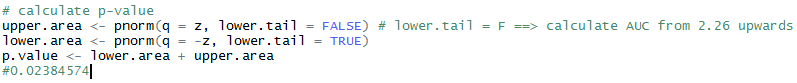
* In other words, regardless of what scale the original data are on, the **z-statistic** itself always has the same interpretation: **the # of SE’s that separate the observed sample mean X¯ from the population mean µ0 predicted by the null.**
* Better yet, regardless of what the population parameters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

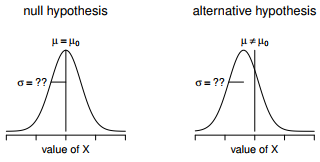
* After some thought, decide it might not be safe to assume psychology student grades necessarily have the same SD as the other students in the class.
* After all, if I’m entertaining the hypothesis that they *don’t* have the same mean, why believe they *absolutely* have the same DF?
* In view of this 🡪 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:



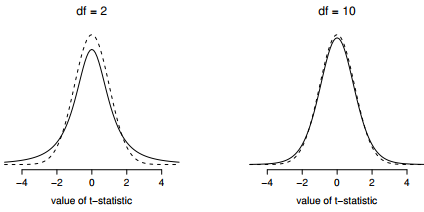
* Can’t say we *know* σ = 9.5 but *can* say σ^ = 9.52
* Running a z-test but w/ the estimated SD of 9.52 instead of relying on the assumption the true SD = 9.5 still gives a significant result, + this approach is close, but not quite correct.
* Now are relying on an *estimate* of the population SD, we 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:



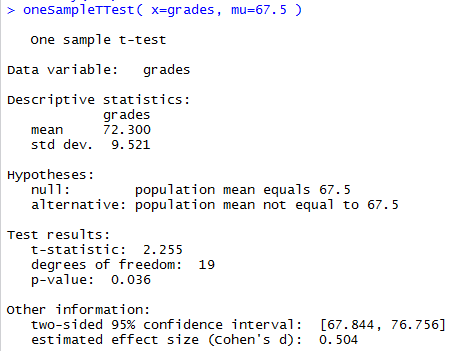
* Only thing that changed in the equation = instead of using known true value σ, use the estimate ˆσ.



* 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*.



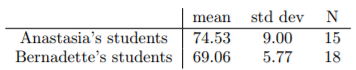
* Output is trying to lead you through the entire data analysis process.
* 1st 2 lines = what kind of test was run + what data were used.
* Then see some basic info about the sample 🡪 sample mean + SD
* Then moves towards inferential statistics 🡪 telling you what the null + alternative were + reports 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**,

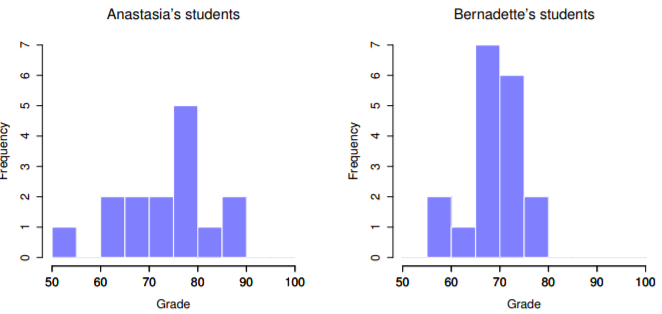


* Assumptions of the one sample t-test (same assumptions as the z-test, minus the one about the known SD)
* **Normality** 🡪 still assuming the population distribution is normal + there are standard tools you can use to check to see if this assumption is met + other tests you can do in its place if violated
* 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.

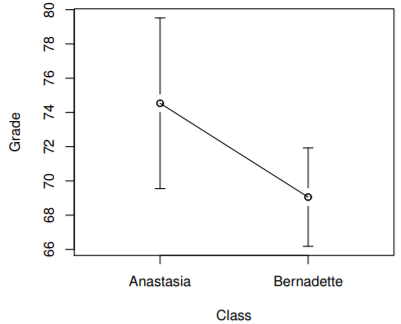
**13.3 The independent samples t-test (Student test)**

* 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



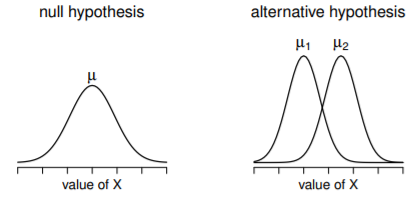
* Error bars depict 95% CI’s around the mean.
* On the basis of visual inspection, it looks like there’s a real difference between the groups, though it’s hard to say for sure.
* The independent samples t-test comes in two different forms, Student’s and Welch’s.
* The original **Student t-test** = simpler 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.



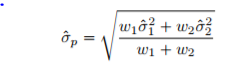
* The null assumes both groups have the same mean µ, whereas the alternative assumes they have different means µ1 + µ2.
* Notice it is assumed the population distributions are normal, + that, although the alternative allows the group to have different means, it assumes they have the same SD
* How should 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



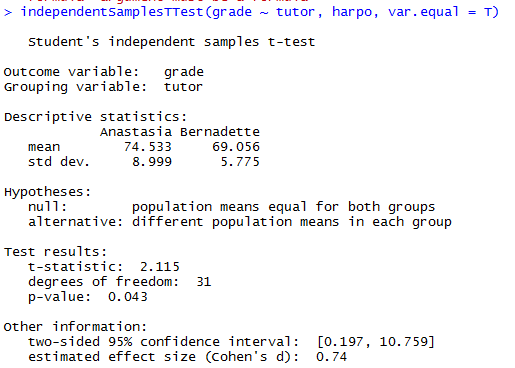
* If you take the square root of this then you get ˆσ(p), the **pooled SD estimate**
* Pooled SD calculation is it’s not terribly different to the regular SD calculation
* From now on, drop the p subscript, + just refer to this estimate as ˆσ.
* Whole reason for calculating this pooled estimate was we 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

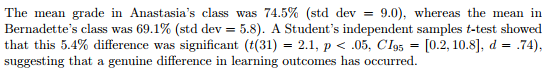


* Our **t-statistic** is therefore



* 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()**



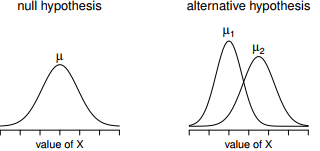
* Output has a very familiar form.
* 1st tells you what test was run, names of the variables used, reports the sample means + SDs for both groups, states the null + alternative in a fairly explicit form, then reports the test results: a t-statistic, the dF, + the p-value.
* Final section reports 2 things: 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**

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

* The biggest problem w/ using the Student test in practice is the 3rd assumption, **homoscedasticity,** which assumes both groups have the same SD.
* Rarely true in real life: if 2 samples don’t have the same means, why expect the same SD?
* Really no reason to expect this assumption to be true.
* **Welch t-test =** a different form of the t-test that does NOT rely on this assumption.



* Graph of the null + alternative *assumed by the Welch t-test*.
* Like the Student’s t-test: assume both samples are drawn from a normal population*; but the alternative no longer requires the 2 populations to have equal variance.*
* Welch test = default t-test in R + is very similar to Student’s t-test.
* **t-statistic** used in Welch is calculated in much the same way as for Student 🡪 take the difference between sample means + divide by some estimate of the SE of that difference:



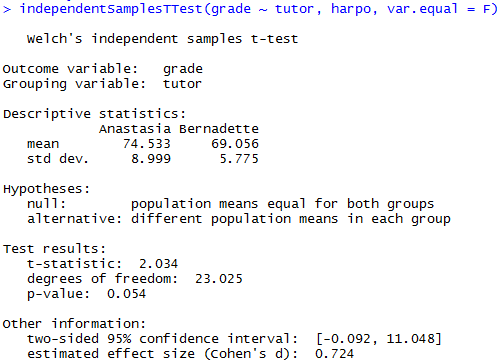
* The main difference = the SE calculations are different.
* If 2 two populations have *different SDs*, it’s complete nonsense to try to calculate a pooled SD estimate, b/c you’re averaging apples + oranges
* Can still estimate the SE of the difference between sample means; it just ends up looking different:

**STUDENT** =  **WELCH** = 

* So, the t-statistic from Welch is actually somewhat different than from Student.
* 2nd difference between Welch + Student = dF are calculated in a very different way.
* In Welch, the dF *DOESN’T have to be a whole number any more* + *DOESN’T correspond all that closely to the number of DPs - minus the number of constraints heuristic used up to this point.*
* The dF are, in fact:



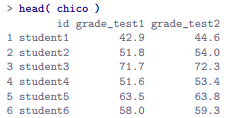
* What matters = see the df value for a Welch test tends to be a bit smaller than in the Student test, + it doesn’t have to be a whole number.
* To run a Welch test in R is pretty easy 🡪 don’t bother telling R to assume equal variances = take the command used to run a Student’s t-test + drop **var.equal = TRUE**



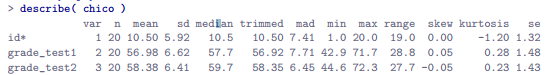
* Output is exact same format as last time but very 1st line is different, b/c it’s a Welch test
* *Except our result isn’t significant anymore.*
* The fact that 1 t-test is significant + the other isn’t doesn’t itself mean very much
* As a general rule, it’s not a good idea to go out of your way to try to interpret/explain the difference between a p-value of .049 + a p-value of .051.
* **If this sort of thing happens in real life, the difference in these p-values is almost certainly due to chance.**
* What DOES matter is you take a little bit of care in thinking about WHICH test you use.
* Student + Welch t-tests have different strengths + weaknesses.
* If 2 populations really DO have equal variances, Student t-test = slightly more powerful (lower Type II error rate)
* If they DON’T have the same variances, the assumptions of the Student test are violated + you may not be able to trust it 🡪 might end up w/ a higher Type I error rate.
* It’s a trade-off.
* IRL, tend to prefer the Welch test b/c almost no-one actually believes population variances are identical
* Assumptions of the Welch test are very similar to those made by the Student t-test except the Welch test does NOT assume **homoscedasticity** (homogeneity of variance).
* This leaves only the assumption of **normality**, + the assumption of **independence**.
* The specifics of these assumptions are the same for the Welch test as for the Student test.

**13.5 The Paired-Samples T-Test**

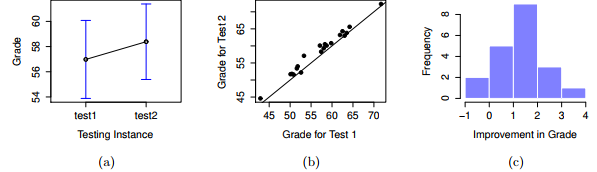
* Regardless of the Student or Welch t-test, an **independent samples t-test** is intended to be *used in a situation where you have 2 samples that are independent of one another.*
* Arises naturally when participants are assigned randomly to 1 of 2 experimental conditions
* But provides very poor approximation to other sorts of research designs.
* In particular, a **repeated measures design** (each participant is measured, w/ respect to the same outcome variable, in BOTH experimental conditions) is NOT suited for analysis using independent samples t-tests.
* Might be interested in whether listening to music reduces people’s working memory capacity
* Could measure each person’s working memory capacity in 2 conditions: w/ + w/out music
* In this experimental design, each participant appears in both groups = requires us to approach the problem in a different way 🡪 **the paired samples t-test**
* This design = very similar to the one that motivated the McNemar test.
* Shouldn’t be a surprise 🡪 both = standard repeated measures designs w/ 2 measurements
* Only difference = this time our outcome variable is interval scale (working memory capacity) rather than a binary, nominal scale variable (yes-or-no question).
* Dataset = class w/ students taking 2 major tests, 1 early in the semester + 1 later in the semester.
* Very hard class 🡪 asserts students are encouraged to work harder
* Theory = 1st test is a bit of a wake-up call for students + when they realise how hard class really is, they’ll work harder for the 2nd test + get a better mark.



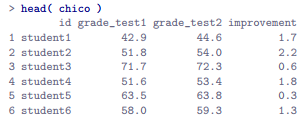
* At a glance 🡪 does seem like class is a hard (most grades between 50-60%), but it does look like there’s an improvement from the 1st test to the 2nd one.

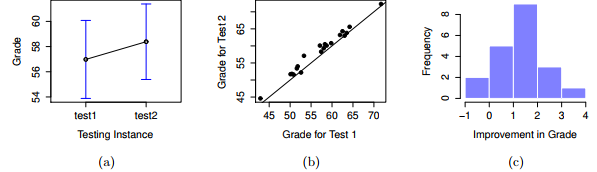


* See that this impression seems to be supported.
* Across all 20 students, 1st test mean grade = 57%, but rises to 58% for the 2nd
* Although, given that the SDs = 6.6% + 6.4% respectively, it’s starting to feel like maybe the improvement is just illusory; maybe just random variation.
* This impression is reinforced when you see the means + confidence intervals plotted



* A) = Mean grade for test 1 + test 2, with associated 95% CI
* B) = Scatterplot showing individual grades for test 1 + test 2
* If we were to rely on plot A) alone, we’d come to the same conclusions we got from looking at the descriptive statistics from describe()
* Looking at how wide those CI’s are 🡪 tempted to think the apparent improvement in student performance is pure chance.
* Nevertheless, this impression is WRONG.
* Look at the scatterplot of the grades for test 1 vs. grades for test 2
* If their grade for test 1 (x) = grade for test 2 (y), the dot falls on the line + points above it = students that performed better on the 2nd test.
* Critically, almost all DP’s fall above the line = almost all students seem to have an improved grade, if only by a small amount.
* This suggests we should be *looking at the improvement* made by each student from 1 test to the next + *treating that as our raw data*





* C) Histogram showing improvement made by each student
* notice almost entire distribution is above 0 = vast majority of students did improve performance from the 1st test to the 2nd
* Very clear there’s a real improvement here + vast majority of students scored higher on test 2,
* This is reflected in the fact that almost the entire histogram is above 0.



* See it’s 95% certain the true (population-wide) average improvement would lie between 0.95-1.86%
* So we can see, *qualitatively*, what’s going on 🡪 there is a *real* **within-group improvement** (everyone improves by about 1%), but it’s very small when set against the quite large **between-group differences** (student grades vary by about 20% or so)
* Think about how to construct an appropriate t-test.
* 1 possibility = Try to run an independent samples t-test using grade\_test1 + grade\_test2 as variables of interest = clearly the wrong thing to do
* Independent samples t-test assumes there is no particular relationship between the 2 samples yet clearly that’s not true in this case b/c of the repeated measures structure to the data.
* W/ an independent samples t-test, we’d be conflating the **within-subject variability** (what we’re interested in testing) w/ the **between-subject variability** (which we are not).
* Instead of running an independent samples t-test on grade\_test1 + grade\_test2, run a **one-sample t-test** on the **within-subject** **variability** variable, **improvement**.
* To formalize this slightly: if X(i)1 = score the i-th participant obtained on the 1st variable + X(i)2 = score the *same* person obtained on the 2nd one, the **difference** score is:



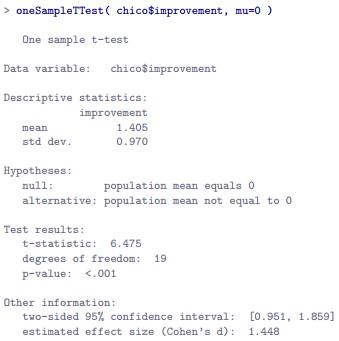
* Notice the difference score = variable 1 - variable 2 + NOT the other way around
* If we want **improvement** to correspond to a *positive*-valued difference, we actually want test 2 to be variable 1.
* Equally, we’d say  = the population mean for this difference variable.
* So, to convert this to a hypothesis test, our null = this mean difference = 0 + the alternative = it is not**,** assuming we’re talking about a 2-sided test here



* This is more or less identical to the way we described the hypotheses for the 1-sample t-test 🡪 only difference = the specific value the null predicts = 0
* t-statistic is defined in more or less the same way too. If D¯ = mean of the difference scores, then:

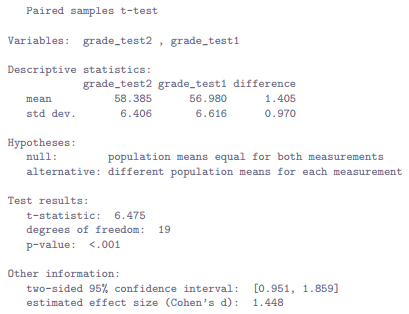
 = 

* Since this is just an ordinary, 1-sample t-test, w/ nothing special about it, the dF are still N – 1
* **Paired samples t-test** really isn’t a new test 🡪 its a 1-sample t-test, *but applied to the difference between 2 variables*.
* Very simple but need to be able to recognize when a paired samples test is appropriate + to understand why it’s better than an independent samples t test.
* 1 way to do this in R = follow the process outlined above = create a difference variable + run a 1-sample t-test it.

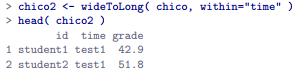


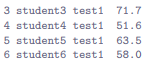
* The output here is (obviously) formatted exactly the same was as w/ the oneSampleTTest() + it confirms our intuition.
* Average improvement = 1.4% from test 1 to test 2 which is significantly different from 0 (t(19) = 6.48, p < .001).
* Suppose you’re lazy + don’t want to go to all the effort of creating a new variable or just want to keep the difference between 1-sample + paired-samples tests clear in your head.
* If so, use **pairedSamplesTTest()** in **lsr**



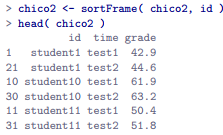


* Output is identical but a bit more detailed
* This time around the descriptive **statistics block** shows the means + SD’s for the *original* variables, as well as for the difference variable (notice it always defines the difference as the 1st listed variable - the 2nd listed one).
* The null + alternative are now framed in terms of the original variables rather than the difference score, but keep in mind that in a paired samples test, it’s still the **difference** score being tested.
* The statistical information at the bottom about the test result is of course the same as before.
* Paired samples t-test is different from other t-tests b/c it is used in **repeated measures designs**
* Ex: Every student is “measured” twice, once for the 1st test, + again for the 2nd.
* Repeated measures data can be expressed in 2 standard ways, **wide form** + **long form**.
* Our class data frame is in **wide form** =every row corresponds to a unique person (form you’re most used to seeing + also the format you’re most likely to receive data in)
* However, majority of tools in R for dealing w/ repeated measures data expect to receive data in **long form**.
* Paired samples t-test is a bit of an exception that way
* How to apply pairedSamplesTTest() to long form data.
* 1st use **wideToLong()** to create a long form version of the data frame.

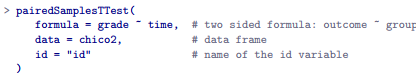




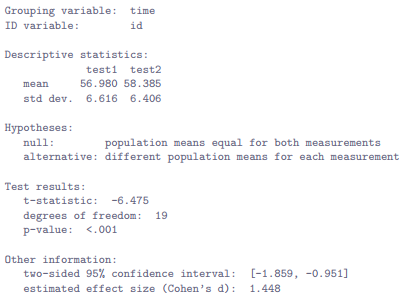
* See a new data frame containing 3 variables: an **id** indicating a person who provided the data, a **time** indicating which test the data refers to, + a **grade** that records scores a person got on a test
* Notice this data frame is in **long form** 🡪 every row corresponds to a unique measurement.
* B/c every person provides TWO observations (test 1 + 2), there are TWO rows for every person.



* How would we run our paired samples t-test now?
* 1 possibility = use **longToWide()** to force the data *back into wide form +* do the same thing as before.
* There’s an easier way
* Think about what how the chico2 data frame is structured 🡪 3 variables + they all matter.
* Outcome measure is stored as the grade + we effectively have 2 “groups” of measurements (test 1 + test 2) defined by the time points at which a test is given.
* Finally, b/c we want to keep track of *which measurements should be paired together*, we need to know *which student obtained each grade* (what id gives us)
* So, when your data are presented to you in long form, we want specify a 2-sided formula + a data frame, in the same way that we do for an independent samples t-test
* Formula specifies the outcome variable + the groups 🡪 grade ~ time
* data frame = chico2.
* However, also need to tell it the id variable







* As you can see, it’s a bit more detailed than output from **oneSampleTTest().**
* It gives the descriptive statistics for the original variables, states the null in a fashion a bit more appropriate for a repeated measures design, + then reports all nuts + bolts from the hypothesis test itself.
* Not surprisingly the numbers the same as the ones that we saw last time.
* 1 final comment about pairedSamplesTTest():
* Get comfortable thinking about repeated measures data in both wide + long formats, + also become familiar w/ the different ways R functions tend to specify models + tests for repeated measures data.
* A slightly different way of thinking about what the paired samples t-test is doing.
* There’s a sense in which what you’re *really* trying to do is look at how the outcome variable (grade) *is related to the grouping variable* (time), AFTER taking account of the fact there are individual differences between people (id).
* So there’s a sense in which id is actually a *second predictor* 🡪 trying to predict grade on the basis of the time AND the id.
* With that in mind, pairedSamplesTTest() lets you specify a formula like **grade ~ time + (id)** which tells R everything it needs to know:
* Variable on the left (grade) = outcome variable, bracketed term on the right (id) = id variable, + the other term on the right = the grouping variable (time).
* If you specify your formula that way, you only need to specify the formula + data frame + so can get away w/ using a command as simple as:



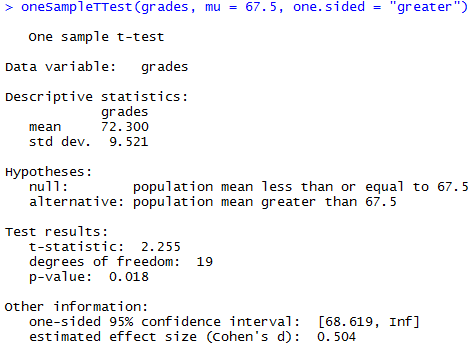
* or you can drop the argument names and just do:



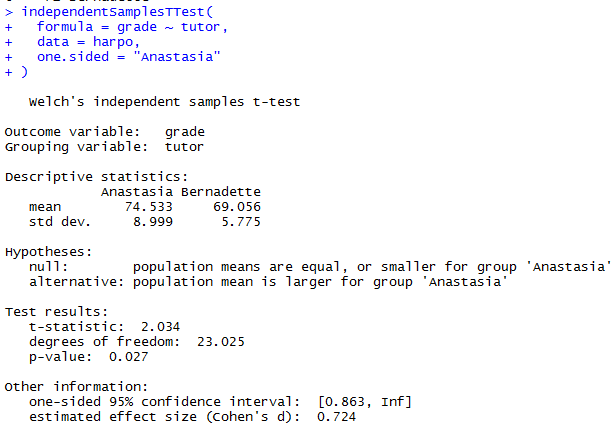
* These commands produce the same output as the last one, but is more elegant.
* The main reason for allowing you to write formulas that way = they’re quite similar to the way **mixed models** (fancy repeated measures analyses) are specified in the **lme4** package

**13.6 One sided tests**

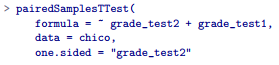
* W/ null hypothesis tests, there are some situations when it’s appropriate to specify a **1-sided test**
* So far, all t-tests have been 2-sided tests.
* For instance, when we specified a 1 sample t-test for grades, the null was that the true mean was = 67.5% + the alternative = the true mean was >= 67.5%.
* Suppose we were only interested in finding out if the true mean is > 67.5% w/ no interest in testing if the true mean is < 67.5%.
* If so, our null = true mean is 67.5% or less, + the alternative = the true mean > 67.5%.
* **oneSampleTTest()** lets you do this by specifying the **one.sided** argument
* one.sided="greater" 🡪 testing to see if the true mean is larger than mu.
* one.sided="less" 🡪 testing to see if the true mean is smaller than mu.

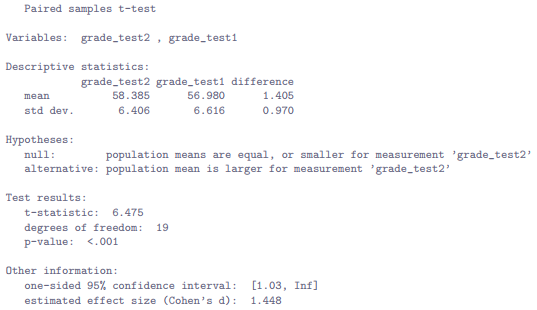


* The null + alternative have changed to reflect the different test.
* ***Note***: Although the t-statistic + dF have not changed, the p-value has.
* This is b/c the ONE-sided test has a *different rejection region* from the TWO-sided test
* ***Note***: the CI is different too + now reports a “one-sided” CI rather than a two-sided one.
* In a 2-sided CI: trying to find #’s A + B so that we’re 95% confident the true mean lies between A + B
* In a ONE-sided CI, we’re trying to find a *single* number A such that we’re 95% confident the true mean is greater than A (or less than A if you set one.sided="less").
* **All versions of the t-test can be one-sided.**
* For an independent samples t test, could have a one-sided test if only interested in testing to see if group A has higher scores than group B w/ no interest in finding out if B has higher scores than A.
* Suppose you wanted to see if A’s students had higher grades than B’s.
* **independentSamplesTTest()** lets you do this by specifying the **one.sided** argument but this time you need to specify the name of the group you’re expecting to have the higher score.

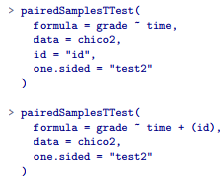
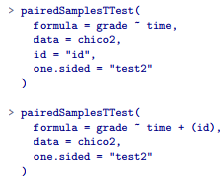


* The definition of the null + alternative have changed, p-value has changed, + it now reports a one-sided CI rather than a two-sided one.
* What about the paired samples t-test?
* Suppose we wanted to test the hypothesis that grades go up from test 1 to test 2 + are not prepared to consider the idea that grades go down.
* Again, use the one.sided argument to specify a one-sided test + in the same way as independent samples t-test, specify name of the group whose scores are expected to be larger under H1
* If your data are in **wide form**, use this command



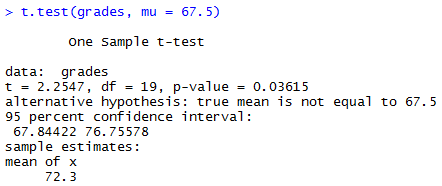


* Yet again, the output changes in a predictable way 🡪 hypotheses have changed, p-value has changed, + CI is now one-sided.
* If data are in **long form**, it still works the same way.
* Either of the following commands would work

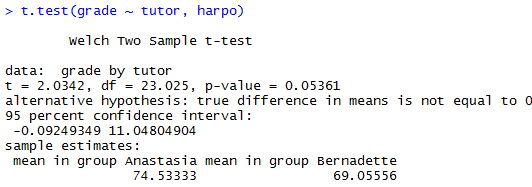
 

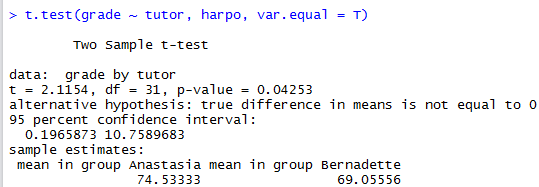
**13.7 Using the t.test() function**

* To run 3 different kinds of t-tests: **1-sample t-test**, **independent samples t-test** (Student’s + Welch’s), + **paired samples t-test.**
* In order to run these different tests, 3 different functions: **oneSampleTTest(), independentSamplesTTest(), pairedSamplesTTest().**
* Can also use **t.test()** = 1 function that can run all 4 different t-tests we’ve talked about.
* To run a one sample t-test on the grades data:



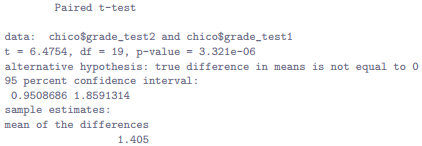
* Same input as oneSampleTTest() 🡪 specify sample data + value (μ) against which it is to be tested.
* Output is a lot more compressed: still has all the info you need 🡪 type of test ran, data tested it on, the t-statistic, the dF, + the p-value, etc.
* Once you know what you’re looking at though, it’s pretty easy to read off the relevant info.
* What about independent samples t-tests? 🡪 Specify formula, data frame, + use **var.equal** to indicate Student or Welch test.





* Paired samples t-test 🡪 Somewhat surprisingly, given most R functions for dealing w/ **repeated measures data** require data to be in **long form**, t.test() isn’t really set up to handle data in long form.
* Instead it expects to be given 2 separate variables, x + y, + you need to specify **paired = TRUE**, also make sure the 1st element of x + 1st element of y actually correspond to the same person, b/c it doesn’t ask for an “id” variable.
* So, in order to run the paired samples t-test on data from Chico’s class:





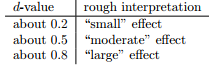
* Yet again, these are the same numbers that we saw in Section 13.5

**13.8 Effect size**

* **Cohen’s d** = most commonly used measure of **effect size** for a *t-test*
* Very simple measure in principle, w/ quite a few wrinkles when digging into details.
* Cohen defined it primarily in the context of an *independent samples t-test*, specifically Student’s test
* In that context, a natural way of defining the effect size is to divide the difference between means by an estimate of the SD.
* In other words, looking to calculate something along the lines of this:



* w/ a Rough Guide For Interpreting d



* Don’t use these blindly 🡪 d statistic has a natural interpretation in + of itself = re-describes difference in means as # of SDs that separates those means.
* So, generally a good idea to think about what that means in practical terms.
* In some contexts a “small” effect could be of *big* practical importance.
* In other situations a “large” effect may not be all that interesting
* You’d think that this would be pretty unambiguous, but it’s not; largely b/c Cohen wasn’t too specific on what he thought should be used as the measure of the SD
* There are several different versions in common usage, + each tends to adopt slightly different notation
* For sake of simplicity (as opposed to accuracy) use:
* **d** to refer to *any statistic you calculate from the sample*
* ***δ*** *to refer to a theoretical population effect.*
* Obviously, this means there are several different things all called **d**.
* **cohensD()** in **lsr** uses the **method argument** to distinguish between them
* Suspicion: only time you’d want Cohen’s d is when running a t-test
* If using oneSampleTTest, independentSamplesTTest, or pairedSamplesTTest() to run t-test 🡪 they automatically produce an estimate of Cohen’s d as part of the output.
* However, if using t.test(), you’ll need to use **cohensD()** (also in lsr) to do the calculations.
* The simplest situation to consider is the one corresponding to a one-sample t-test w/ the 1 sample mean **X¯** + 1 (hypothesised) population mean µ0 to compare it to.
* Not only that, there’s really only 1 sensible way to estimate population SD 🡪 the usual estimate σˆ
* End up w/ the following as the only way to calculate d



* When writing cohensD(), it works similar way to t.test() + as a consequence, it can calculate effect size regardless of which type of t-test performed.
* If you want a measure of Cohen’s d to accompany a one-sample t-test, there’s only 2 arguments you need to care about:
* x = A numeric vector containing sample data
* mu = the mean against which the mean of x is compared (default value = 0).
* Don’t need to specify method to use, b/c there’s only 1 version of d that makes sense in this context
* Compute an effect size for the data from class:

* Overall, **psych students are achieving grades (mean = 72.3%) that are about .5 SDs higher than the level you’d expect (67.5%) if they were performing at the same level as other students.**
* Judged against Cohen’s rough guide, this is a moderate effect size.
* Majority of discussions of Cohen’s d focus on a situation analogous to Student’s independent samples t-test
* It’s in this context the story becomes messier b/c there’re several different versions of d you might want to use in this situation
* Can use the **method** argument to **cohensD()** to pick the one you want.
* To understand why there are multiple versions of d, helps to write down a formula that corresponds to the true population effect size δ



* As usual µ1 + µ2 = population means corresponding to group 1 + group 2 respectively + σ = SD (same for both populations).
* Obvious way to estimate δ = do exactly the same thing we did in the Student’s independent samples t-test itself 🡪 use sample means as top line + **a pooled SD estimate** for the bottom line:



* Where ˆσ(p) = exact same pooled SD measure that appears in the t-test.
* This is the most commonly used version of Cohen’s d when applied to the outcome of a Student t-test + is sometimes referred to as **Hedges’ g statistic**
* It corresponds to **method = "pooled"** in cohensD() + it’s the default.
* However, there are other possibilities
* May want to use only 1 of the 2 groups as the basis for calculating the SD.
* This approach (**Glass’ ∆**) only makes most sense when you have good reason to treat 1 of the 2 groups as a *purer reflection* of **natural variation** than the other.
* This can happen if, for instance, 1 of the 2 groups is a **control group**.
* If that’s what you want, use **method = "x.sd"** or **method = "y.sd"** in cohensD().
* Secondly, recall in the usual calculation of the pooled SD we *divide by N - 2 to correct for the bias in the sample variance*
* In 1 version of Cohen’s d this correction is omitted + instead, we divide by N (**method = "raw")** makes sense primarily when trying to calculate the effect size in the *SAMPLE* rather than estimating an effect size in the *POPULATION*.
* Finally, there is a version based on **Hedges + Olkin**, who point out a small bias in the usual (**pooled**) estimation for Cohen’s d.
* Introduces a small correction (**method = "corrected"**) by multiplying the usual value of d by:



* Look at how to calculate the default version in Harpo’s class
* Command is very similar to the relevant t.test() command but also specifies a method



* This is the version of Cohen’s d that gets reported by independentSamplesTTest() whenever it runs a Student’s t-test.
* Suppose a Welch test 🡪 still have 2 independent samples but no longer believe the corresponding populations have equal variances.
* When this happens, must redefine what we mean by the **population effect size** + refer to this new measure as **δ`** so as to keep it distinct from the measure δ defined previously.
* Cohen suggests we could define our *new* POPULATION effect size by averaging the 2 population variances:

 *where* 

* Seems quite reasonable, but notice none of the measures discussed so far are attempting to estimate this new quantity.
* Only aware of 1 version of Cohen’s d that actually estimates the *unequal*-variance effect size δ` rather than the *equal*-variance effect size δ.
* To calculate d for this version (**method = "unequal"**) = substitute sample means X¯1 + X¯2 + corrected sample SDs ˆσ1 + ˆσ2 into the equation for δ`, which gives us the following equation for d:



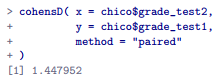
* This is our estimate of the effect size.
* There’s nothing particularly difficult about calculating this in R, since all we have to do is change the method argument:



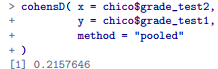
* This = version of Cohen’s d reported by independentSamplesTTest() whenever it runs Welch’s t-test
* Finally, for a Paired Samples T-Test, calculating Cohen’s D depends on what you’re trying to do
* To measure effect sizes relative to the distribution of **difference** scores, the measure of d you calculate is just (**method = "paired"**):



* where ˆσD = the estimate of the SD of the differences.



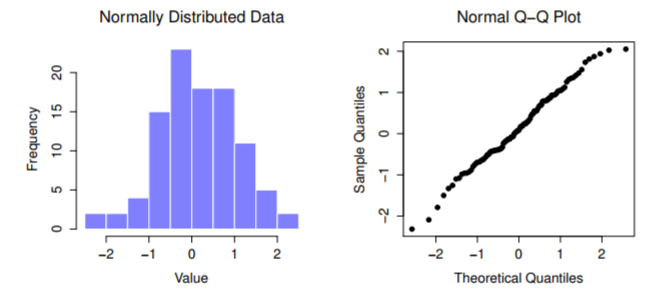
* This is the version of Cohen’s d gets reported by pairedSamplesTTest()
* The only wrinkle = figuring out whether this is the measure you want or not.
* To the extent you care about practical consequences of your research, you often want to measure effect size *relative to the original variables*, NOT to the difference scores (e.g., 1% improvement in Chico’s class = pretty small when measured against the amount of between-student variation in grades), in which case you use the same versions of Cohen’s d you’d use for a Student or Welch test.
* For instance, when we do that for Chico’s class::



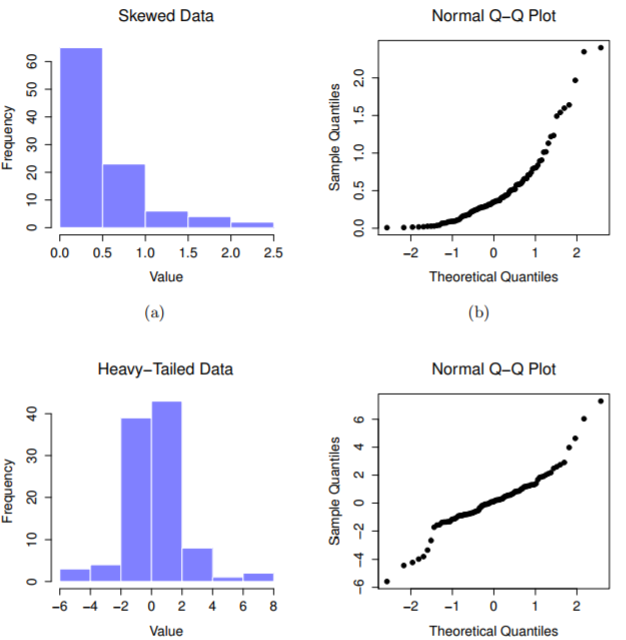
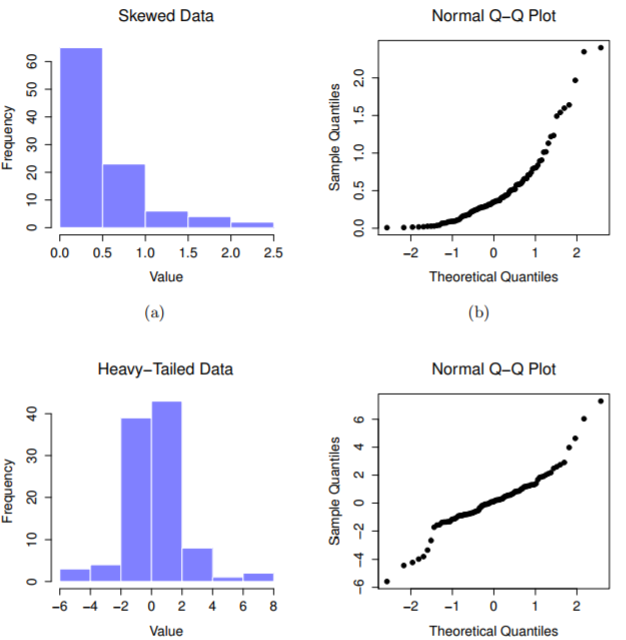
* See the overall effect size is quite small when assessed on the scale of the original variables

**13.9 Checking the normality of a sample**

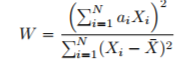
* All tests discussed so far have assumed data are normally distributed 🡪 often quite reasonable, b/c the CLT tends to ensure many real-world quantities ARE normally distributed
* Any time you suspect a variable is an average of lots of different things = pretty good chance it will be normally distributed or at least close enough to normal you can get away w/ using t-tests.
* However, life doesn’t come w/ guarantees + there are lots of ways you can end up w/ variables that are highly non-normal.
* Any time you think a variable is actually the minimum of lots of different things = very good chance it will end up quite skewed.
* In psychology, **response time** (RT) data is a good example of this.
* If you suppose there are lots of things that could trigger a response from a human participant, the actual response will occur the 1st time 1 of these trigger events occurs (massive oversimplification)
* This means RT data are systematically non-normal.
* if normality is assumed by all tests + is mostly *but not always satisfied* (at least approximately) by real world data, how can we check **normality** of a sample?
* 2 methods: **QQ plots**, + **the Shapiro-Wilk test**.
* **Quantile-quantile (QQ) plots** = 1 way to check if a sample violates the normality assumption is
* Allows you to visually check whether you’re seeing any systematic violations.
* In a QQ plot, each observation is plotted as a single dot:
* w/ x = the theoretical quantile the observation should fall in if data were normally distributed (w/ mean + variance estimated from the sample)
* w/ y = the actual quantile of the data w/in the sample.
* If data are normal = dots should form a straight line.
* Generate data by sampling from a normal distribution + drawing a QQ plot



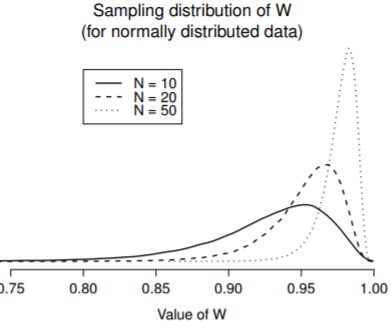
* The **Shapiro-Wilk statistic** associated w/ these data is W = .99, indicating no significant departures from normality were detected
* These data form a pretty straight line = no surprise given we sampled from a normal distribution
* In contrast:

* 1) Skewness = 1.94, Shapiro-Wilk statistic W = .80, reflecting significant departure from normality (p < .001).
* 2) Heavy tails produce a high kurtosis (2.80), resulting Shapiro-Wilk statistic W = .93, again reflecting significant non-normality (p < .001).
* Highly skewed = QQ plot curves upwards
* Heavy tailed (high kurtosis) = QQ plot flattens in middle + curves sharply at either end
* **Shapiro-Wilk tests** = more formal way to check normality of data
* Either that, or **Kolmogorov-Smirnov test =** probably more traditional, though most suggest Shapiro-Wilk = better test of normality
* **Kolomogorov-Smirnov** = general purpose test of distributional equivalence = can be adapted to handle *other* kinds of distribution tests w/ **ks.test()**
* The null being tested = *a set of N observations is normally distributed*.
* The test statistic it calculates is conventionally denoted as **W**
* First, sort observations in order of increasing size + let X1 = smallest value in the sample, X2 = 2nd smallest + so on.
* Then W is given by:



* where X¯ = mean of observations + the a(i) something complicated beyond scope of an introductory text.
* B/c it’s a little hard to explain math behind W, a better idea = give a broad brush description of how it behaves.
* Unlike most test statistics, it’s actually SMALL values of W that indicate departure from normality.
* W statistic has a maximum value = 1, which arises when data look perfectly normal.
* Smaller value of W = less normal data are.
* However, the sampling distribution for W (not one of the standard ones + is in fact a complete pain in the arse to work w/) DOES *depend on sample size N*.



* Sampling distribution of Shapiro-Wilk W statistic, under the null that data are normally distributed for samples of size 10, 20 + 50.
* Note small values of W indicate departure from normality.
* When reporting results for a Shapiro-Wilk test, should (as usual) make sure to include the test statistic W + the p-value, though given that the sampling distribution depends so heavily on N it would probably be a politeness to include N as well