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

***Part V – Statistical Tools***

**Chapter 13 - Comparing Two Means**

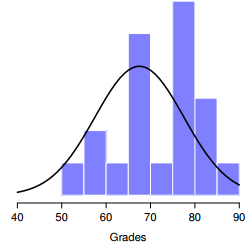
* In the previous chapter we covered the situation when the outcome variable is nominal scale + your predictor variable is *also* nominal scale.
* Lots of real world situations have that character, so you’ll find that chi-square tests in particular are quite widely used.
* However, you’re much more likely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

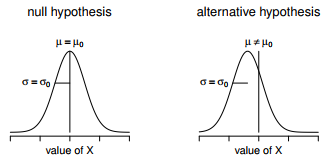
* z-test is almost never used in real life
* Only real purpose = a very convenient stepping stone along the way to the t-test, probably the most (over)used tool in all statistics.

**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:



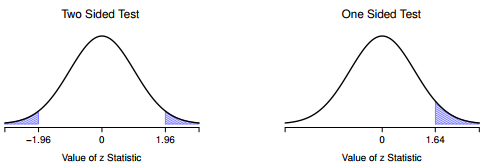
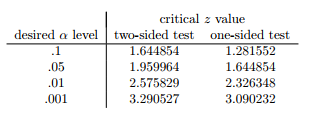
* Now comes the trick to convert the sample mean X¯ into a standard score, z, or z(X¯) = a standardized version *of a sample mean*, NOT a standardized version of a *single observation,* which is what a z-score usually refers to).
* The z-score for our sample mean is:

 OR 

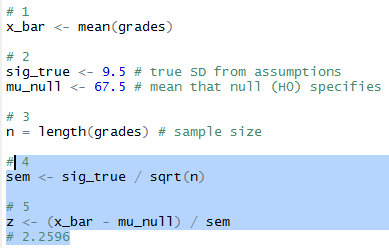
* This z-score is our test statistic + the nice thing about using it as our test statistic is, like all z-scores, *it has a standard normal distribution*:



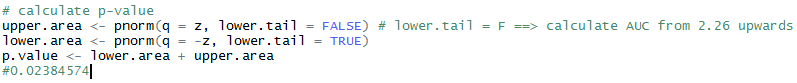
* In other words, regardless of what scale the original data are on, the **z-statistic** itself always has the same interpretation: **the # of SE’s that separate the observed sample mean X¯ from the population mean µ0 predicted by the null.**
* Better yet, regardless of what the population parameters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

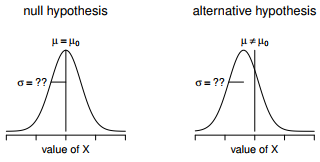
* After some thought, decide it might not be safe to assume psychology student grades necessarily have the same SD as the other students in the class.
* After all, if I’m entertaining the hypothesis that they *don’t* have the same mean, why believe they *absolutely* have the same DF?
* In view of this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



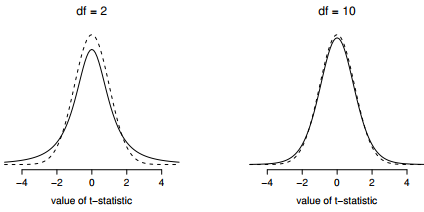
* Can’t say we *know* σ = 9.5 but *can* say σ^ = 9.52
* Running a z-test but w/ the estimated SD of 9.52 instead of relying on the assumption the true SD = 9.5 still gives a significant result, + this approach is close, but not quite correct.
* Now are relying on an *estimate* of the population SD, we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



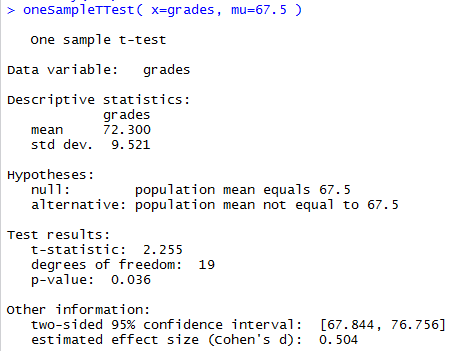
* Only thing that changed in the equation = instead of using known true value σ, use the estimate ˆσ.



* 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
* **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.