210A Week 7 Notes

Significance test of means

For the last few weeks we've been doing significance tests against the null (either zero or some arbitrary number). This week we will be comparing means for two groups with a t-test of means. More generally, this is useful if you have a binary variable (the groups) and a continuous variable (the thing with the means). If you have two categorical variables you use χ^2 (which we'll cover next week) for two continuous variables you use OLS (which you'll learn in 210B).

For a t-test of means you need a continuous variable, like "wage" and a binary variable, like "married." The thing we're really interested in is the *diff* for the continuous variable of one group vs the other.

$diff = \bar{x_1} - \bar{x_2}$

$se_{diff} = \sqrt{se_1^2 + se_2^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

as always, $t = \frac{parameter}{se} = \frac{diff}{se}$

Unlike previous weeks, the parameter is *diff* and *se* is pooled. Note that *diff* is equivalent to a regression coefficient in an OLS model with just one predictor so we can also refer to *diff* as β. Once we plug and chug the slightly different formulas, the interpretation of t is very similar to that of last week (when we were comparing a distribution to a threshhold). Since the null is usually zero, what we usually want to know is simply "do these distributions have the same mean?" which is a two-tailed test. Of course if theoretically appropriae, we can also ask "does this group have a bigger mean than that group?" which is a one-tailed test.

Rember that $t = (\bar{x_1} - \bar{x_2})/se_{diff}$. What is the parameter that we're testing the statistical significance of here? It's not the mean for one group or the mean for the other, but *diff* or $\bar{x_1} - \bar{x_2}$. Which group is "1" and which is "2" doesn't matter for two-tailed test, but it does for one-tailed test. Note that this is one of the reasons the null is "zero." We usually, like to ask "are these two groups the same" (i.e., $H_0 : diff = 0$), not "are these two groups different by some arbitrary difference?" We can imagine some contrived examples of non-zero null, but generally we want to know simply, are they different, which implies two-tailed null of zero. This is very important because it closely parallels the interpretation of a regression coefficient, which is the main thing quants do.

The t-test of means is the basic thing (in part because it's closely analogous to regression). Most of the rest of this stuff seldom comes up in practice (for sociologists anyway). A lot of the material in the course is the sort you should get down cold because other techniques are based on it. However, the rest of this week is of the sort you should have some vague awareness of and can look up the details if you ever need to do it.

The Stata syntax for a t-test is "ttest continuous, by(binary)". For example:

```
. sysuse auto, clear  (1978 Automobile Data)
. ttest mpg, by(foreign)
Two-sample t test with equal variances
```

```
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Domestic |      52    19.82692     .657777    4.743297    18.50638    21.14747
 Foreign |      22    24.77273     1.40951    6.611187    21.84149    27.70396
---------+--------------------------------------------------------------------
combined |      74     21.2973    .6725511    5.785503     19.9569    22.63769
---------+--------------------------------------------------------------------
    diff |           -4.945804    1.362162               -7.661225   -2.230384
------------------------------------------------------------------------------
    diff = mean(Domestic) - mean(Foreign)                         t =  -3.6308
Ho: diff = 0                                     degrees of freedom =       72
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0003         Pr(|T| > |t|) = 0.0005          Pr(T > t) = 0.9997
```

Note that it gives you left-tail, two-tail, and right-tail. If you read the bottom row of the output, the first p is a one-tailed test that foreign cars are more efficient, the second p is a two-tailed test that foreign and domestic are different as to effficiency, and the third p is a one-tailed test that domestic are more efficient.

Dummy variables

The one practical issue you really need to know is how to handle it when you have more than two groups. The t-test of means only works with binary variables. To handle a categorical variable with three or more categories you need to treat it as a "dummy set." For instance, imagine we have a continuous variable (e.g., "income") and a categorical variable "race" that can take the values: white, black, other. We can't simply do a t-test for income by race. Rather we can do a t-test comparing two races (e.g., white vs. black) or one race vs. all the other races together (e.g., white vs. non-white). In the context of regression we usually have a "dummy set" which is a list of "category vs. all other categories" dummies, but leaves out one called the "reference cell" or "omitted category."

There are three ways to do this in Stata: manually, "xi" (the old way), and "factor variables" (the new way).

To do it manually just just use the commands "gen," "replace," and "recode."

```
gen race_white=1 if race=="white"
recode race_white .=0
gen race_black=1 if race=="black"
recode race_black .=0
```

Through Stata 10 we use the "xi" syntax.

```
xi i.race
xi: regress i.race
```

As of Stata 11 "xi" still works but it's highly recommended that you switch to the new more flexible and efficient "factor variables" syntax. If you add an "i." before a categorical variable it behaves like the old "xi" syntax. However it's more flexible as you can do things like "b2.var" to create a dummy for when "var==2".

```
ttest wage, by(b2.race)
reg wage i.race
```

You can also create interaction effects using the factor variables syntax and interpret them with the margin syntax. These are when you look at one variable within another. So for instance you could look not just at race and gender but race by gender. However we won't get into it more as the interpretation doesn't make sense outside of regression. You can read more about factor variables from Stata help or UCLA ATS.

two proportions

As we'll study next week, we mostly use χ^2 for this but if n_1 and n_2 are large, we can approximate with t or Z. As before, $se_{diff} = \sqrt{se_1^2 + se_2^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Since for proportions, $\sigma = \sqrt{\pi(1-\pi)}$ and $se = \sqrt{\frac{\pi(1-\pi)}{n}}$ you can plug in, to get $se = \sqrt{\frac{\pi_1-(1-\pi_1)}{n_1} + \frac{\pi_2-(1-\pi_2)}{n_2}}$

As usual, the interpretation is $t = \frac{diff}{se_{diff}}$

one-tailed test vs two-tailed

As you know, I don't like one-tailed tests. The book gives the example of prayer for the health of medical patients. It seems like one-tailed is appropriate as prayer can't hurt, right? (Especially since it's double blind and thus it's not like they don't take their meds). But what if they prayed to the wrong god? A two-tailed test would be evidence for the supernatural generally, a one-tailed test would be evidence for (intercessory conceptions of) Christianity specifically and implicitly treats a wrathful Ba'al smiting the objects of Christian prayers as part of the null.

contingency table

These are frequencies of different things eg, 2x2. We will hold for next week since they only get interesting with χ^2.

dependent samples

Just skip the dependent samples issue as sociologists don't do this. These techniques are used by people who do experiments like psychologists and physicians. Sociologists do often deal with repeated observations, but we handle them with

random-effects (cluster-level error terms) or fixed-effects (cluster dummy sets) regression models because it's easier to introduce control variables. This is covered in 210C and Stata calls these models "xt."

comparing means assuming equal sd

I've never done this myself. The only reason I might is if I didn't have the raw data, but was relying on summary data from a published paper or an almanac. In this case $se = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

This assumes that σ is reported as pooled, otherwise you'd need to create an average σ weighted by n of the subpopulations. The interpretation is the same as any t-test of means.