Economics 50/SUP 135 Gregory A. Bruich, Ph.D.
Dept. of Economics, Harvard University

Lab 4: The Tennessee STAR Experiment

Part 1: Randomization Tests (or Balance Tests)

Randomization tests or tests for balance are statistical tests that we can do to verify that there really was random assignment in an experiment. The general idea is to utilize data that we might have on characteristics that were measured before the experiment. We call those predetermined characteristics. If the sample size is large enough, then the mean of any predetermined characteristic should be very close in the treatment and control groups. We know this already from the simulations we did in Lab 1.

Because these variables were measured before the experiment, they could not possibly have been changed by the treatment. But in Lab 1, recall that none of the characteristics was exactly the same in the two groups. The next question is, are the differences due to chance? Or are these differences evidence against randomization? We need to consider both the *practical significance* and the *statistical significance* of these differences.

Practical significance refers to whether the difference is large in a real-world sense. In other words, would an expert say that the difference is meaningful? A heuristic way to judge practical significance, is to benchmark our estimates to either the control group mean, or the control group standard deviation. We often have an intuitive feel for whether the difference is big when it is expressed as a percent of a benchmark.

The tool that we will use to assess *statistical significance* is a 95% confidence interval, which we will calculate as:[1]

$$\text{Estimated difference} \pm 1.96 \times \text{standard error}$$

That formula gives us the range of values that are consistent with our data, in a very specific sense. In particular, we can reject the hypothesis that the real difference is equal to any value *outside* that range, at the 5% level. If the interval does not contain zero, then we can reject the hypothesis that the real difference is zero at the 5% level. If the interval contains zero, we cannot reject the hypothesis that the real difference is zero at the 5% level.

If the confidence interval contains zero, then the estimate is not statistically significantly different from zero. If the confidence interval does not contain zero, then the estimate is statistically significantly different from zero.

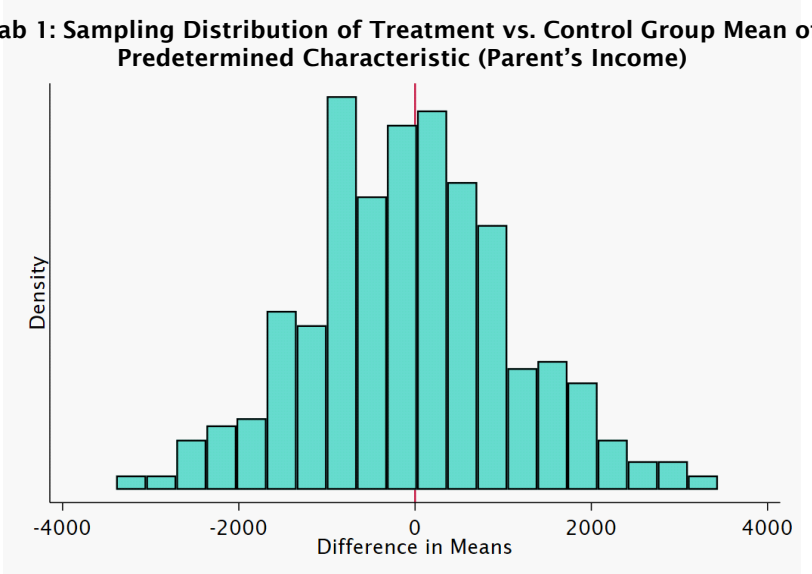
[1] The number $1.96\dot{}$ comes from the standard Normal distribution, which is the familiar bell shaped distribution that you may have seen before. The 1.96 critcial value is justified formally using the Central Limit Theorem from probability theory.

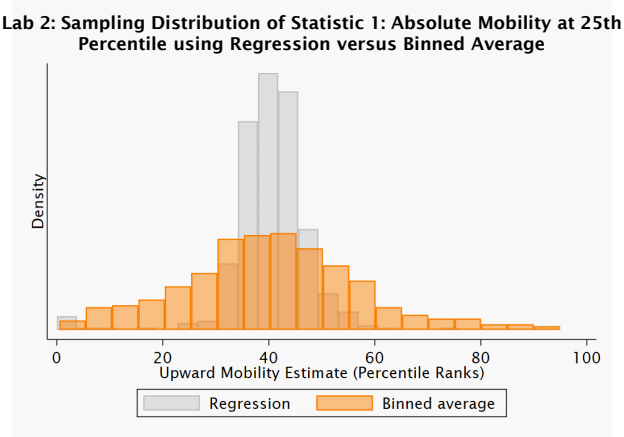
Primer on Standard Errors

To shed more light on the standard error that appears in the formula for the 95% confidence interval, recall that in Lab 1, each student in the class conducted their own simulated experiment, and there was a distribution of estimates across the 500 students in the class. This distribution of estimates is called a *sampling distribution*, and is shown in the figure below. A standard error measures the *standard deviation* of the sampling distribution. Recall the rule of thumb from Lab 1: most of the data will usually be within one standard deviation of the mean; almost all of the data will usually be within two standard deviations of the mean. This intuition is why the confidence interval uses 1.96 $\times$ standard error.

Standard errors help quantify uncertainty. If we think the difference should really be zero, and the sampling distribution is very spread out, then our estimate in our one experiment is more likely to be far away from zero just by chance. In contrast, if the sampling distribution is more concentrated, then it is not as likely. In the simulations from Lab 2 (shown on the next page), we can see that the sampling distribution for Statistic 1: Absolute Mobility at the 25th Percentile is very concentrated when estimated using a linear regression even in a sample of size 50. In contrast, the sampling distribution for the statistic based on the binned average is very spread out: there is a lot of uncertainty about what the statistic is based on any one student's estimate.

It turns out that we can *estimate* the standard error using data from our *one experiment*. In the context of an experiment, the standard error depends on three things. First is the number of subjects in the experiment. If you run a bigger experiment, you have less uncertainty and the standard error will be smaller. Second, it also depends on the fraction of subjects in the two groups. So if you only had one observation in the treatment group, even if you had a 1000 observations in the control group, we would have a lot of uncertainty, and the standard errors should be big. Third, it also depends on the variance of the variable that we are examining. If you look at an outcome with a lot of extreme values (like earnings), this will translate to bigger standard errors.





Part 2: Stratified Randomized Experiments

Stratified random assignment splits people up into specific groups, or "strata" before randomization. This modification to the treatment protocol will change how we estimate the effect of an experiment. In particular, we still estimate the difference in means across the treatment and control groups. But now we will calculate that difference using a regression that controls for strata. Because this regression has multiple independent variables, it is called a multivariable regression.

Strata often correspond to the geographic locations or sites where the experiment took place. Within each site, it is exactly like there was a separate randomized experiment. In the Moving to Opportunity Experiment, the strata were the five cities (Baltimore, Boston, Chicago, Los Angeles, and New York). In the Creating Moves to Opportunity Experiment, the two strata were the Seattle Housing Authority and King County Housing Authority. In the Tennessee STAR experiment, the strata were the 79 schools.

Let $treat_i$ be a binary indicator variable for whether person i was randomly assigned into the treatment group. We are also going to define binary indicator variables for each site, that equal 1 if individual i was in that site and zero otherwise. I often refer to the site indicators as "site fixed effects." Now we estimate a multivariable regression:

$$\widehat{Y_i} = \alpha_0 + \alpha_1 treat_i + \alpha_2 site1_i + \alpha_3 site2_i + \alpha_4 site3_i + \alpha_5 site4_i$$

The dependent variable is an outcome variable Y_i. The independent variables are: an intercept, $treat_i$, and the site fixed effects $(site1_i, ..., site4_i)$.

The coefficient α_1 estimates one overall treatment effect for the experiment. All the other coefficients (α_0, α_2, α_3, α_4, α_5) can be safely ignored, and do not have useful interpretations.

To give you some intuition for the inner workings of a multivariable regression, it turns out that the coefficient α_1 can also be expressed as a weighted average of the treatment effects estimated separately by site. The insight comes from what the weights depend on, which is the number of subjects in each site, and the fraction of them in the treatment group. The multivariable regression implicitly gives more weight to sites that had more subjects, and a closer to 50/50 split into treatment and control groups. This is the weighting scheme that will give us the most precision under a certain set of assumptions.