LS 123 Data, Prediction, and Law

Meeting 4

Spring 2022

**Prediction vs Causation, Central Limit Theorem**

**Note: If you are sick, stay home and tell Ilya and me that you are out sick. Tune into the Zoom, which will also be recorded—unfortunately we have no course capture.**

Remember not to get hung up on the mathematical representations in “Prediction Policy Problems” and focus on the intuitions. Also get started on Lab 5, Large N, which is a long-ish lab. Here we cover the idea of classical hypothesis testing and the Central Limit Theorem, which allows us to talk about the uncertainty of our estimates when we do things like create a regression model. The lab is a preface to Lab 6, where we will try out the sort of hypothesis testing, using multiple regression, that is common in the social sciences.

1. **Prediction policy problems**
   1. Causal inference and prediction are distinct problems
      1. Knowing what brings about rain versus
      2. Knowing whether or not to carry an umbrella
      3. Social science in general worries a lot about causality but a lot of problems really require good predictions, especially if you are having to make a policy decision
   2. This means we should pay attention to method
      1. Bias-variance tradeoff in predictive techniques (see Intro to Statistical Learning Ch. 2)
         1. we can think of error, in the context of predicting values using a model, as consisting, in general of three terms
            1. variance: how much would our estimate of some value change if we made it using the same model but a different set of data
            2. bias: how much error is introduced by modeling some real world phenomenon with an algorithm like a regression equation, which is much simpler than the real world
            3. random error: error introduced from measurement, etc, which is typically assumed to cancel out (i.e., have a mean value of zero)
         2. so using expected value notation

where the Expected Test MSE (assuming repeated estimates) is the sum of the variance for that prediction, the bias squared for that prediction, and the error variance

* + - 1. a typical way to represent error is Mean Squared Error, MSE, as we will see, but it is not the only way to represent the value of the error of an estimate
      2. so we want low variance and low bias, but there are tradeoffs; if our model fits the training data very closely we will have low bias and high variance when we try it on a different set of data; if our model is not as close to the training data (higher bias) it may fit repeated draws of the data better on average (lower variance)
      3. more flexible methods (tree models) generally give lower bias and higher variance than less flexible ones (regression)
    1. For causality, we are asking a “whether or not” question, and thus OLS or other similar regression techniques do what we want; give us an idea of whether or not we can reject the hypothesis that our causal variables have no effect (and if we can reject it then we have proposed an alternative, although we have not really confirmed the alternative—that’s why we start with theory and that is why there is criticism of classical hypothesis testing); in addition we get an idea of effect size and sign (pos or neg association)
    2. Still, it is good to have an idea of whether or not, e.g., racial animus affects beliefs about other sorts of policy like health care
    3. For this we want to be able to reject the null hypothesis that our favored variable has zero effect on the outcome—we will try this next time; we assume that we have captured the variation in the model and so all the error is variance and the random error term, and we want to minimize variance (i.e., by minimizing the sum of the squared differences between the prediction and the observed value) to get the best estimator (which is why OLS is BLUE—best linear unbiased estimator)
    4. This may not be terribly sophisticated but it is a start on “why” questions, like the ones we will try out in Lab 6
  1. Sometimes we just want to make a decision based on a good prediction—should this patient get a hip replacement or not?
     1. Certainly this is often what we want in the legal context
     2. Should this person get released pending a court appearance or be held in jail for trial?
     3. For these problems we will gladly trade some bias for a prediction with less spread (i.e. variance)
     4. We want to separate defendants who are risky (although we need to define risky) from those who are not
  2. In the Kleinberg et al example, we want to know who to operate on to make it worth the pain and trouble of hip surgery, and so they ask whether the patient will die within a year of the procedure, and estimate that with a logistic regression model (which we will we get to in a couple of weeks)

1. **Large N lab**: logic of repeated sampling and convergence on the central tendency in the underlying population (note: the lab developer and then Wilson got it mixed up—the liberal and conservative ratings are feeling thermometers)
   1. in the lab for last time on Bootstrapping (which was more challenging than the earlier labs), what is the key thing about the results you got from resampling? (the distribution has a central tendency where the true population central tendency is)
   2. from the confidence intervals you generated, what does the logic of repeated sampling tell us about the confidence interval around the underlying population median? (most of the runs contained the “true population median” from ANES)
   3. these were median values (which apparently *also* converge around a central tendency but require more assumptions about the underlying distribution—that’s above my pay grade so ask a statistician!)
   4. for sample **means**, there is the Central Limit Theorem, which states that the **means** of repeated samples will converge on a Normal distribution around the population mean as the number of sampling repetitions increase, no matter what the underlying population distribution looks like—see Ch. 14 in the Adhikari and DeNero text
   5. Data 8 talks about this in its discussion of bootstrapping but it is good to remember; frequentist statistics of the sort that is in a lot of the course readings relies on the assumption that the data are produced by an underlying population that could be sampled repeatedly
   6. in Lab 5 we are working with respondent feeling thermometers toward liberals and conservatives, not self-ratings of ideology (and people don’t necessarily understand ideological labels)
   7. the Student’s t-test for independent samples—how confident are we in rejecting the null hypothesis that two group means are drawn from the same underlying population? That is, can we distinguish the two groups by their values on this variable
   8. in the jury example in the lab, what other test statistic might you use? (Chi-square test, which tests the hypothesis of drawing a certain proportion in each category given a known proportion in the underlying population); here we have the known proportion of eligible jurors so we can use this test to see if the panel draws are what you would expect of independent, random draws from the jury pool
   9. as the lab notes, the t-test expects you to be comparing the sample means (**caution! see note below**) of a large number of draws, rather than the result of single draw, ~~so it is rather cumbersome to use the Student’s t-test here in the jury example but we do it by using repeated draws~~ [turns out this is not the place to apply Student’s t-test, see note below]
   10. Note the purpose of an independent sample t-test and how it applies to your data
       1. what we want to know is whether two independently drawn samples are from different underlying populations (null is that they are not).
       2. So what would be meaningful? How about the “liberal” feeling thermometer scores of Dems versus those of Republicans
       3. **\*note that the t-test function will calculate a t statistic and a p value even if they are not meaningful, as long as it has valid arguments\***
       4. this sort of hypothesis testing is still very common, and we will see it with regression coefficients (which are another sort of average value), but if often is not all that interesting
       5. who cares if you clear the magic p value of 0.05? there is the problem (“p-hacking”) of looking through all of your results and then only discussing those that pass that significance threshold, but even so it still could be noise!

# Note: t-test is not what you would use for the kind of multinomial distribution in the jury example

Remember toward the end of class today I said something about how the Student's t-test for significance when we are talking about the mean of a given variable (like "Liberal Feeling Thermometer") did not seem to be the right hypothesis test for the proportion data in Lab 5 that had to do with selecting a jury? It is not the right test, because we are comparing the observed frequency of an event to the expected frequency of an event in each of the categories (Asian, Black, Latino, White, Other). The expected frequency is the known population frequency, and the observed frequency is what we see in the unfair draw from the jury pool. So a Chi-square test would be the test of significance to use. In general, Student's t is used for small samples where we don't know the variance of the error term. Student's curve is like a normal curve that has been flattened out and has fatter tails.

For binomial distributions with a large sample size, you can use a z-test (so a normal distribution), but not when we are comparing multiple categories. For a good explanation without mathematical heavy lifting, see Freedman, Pisani, and Purves (1998) Statistics, chs. 26-29. (David Freedman was a statistics professor here at Berkeley.)

So I need to go back and fix this lab, although I think that you can see [my lab with notes](https://github.com/ds-modules/Legalst-123/blob/master/labs/05_Large%20n/05_Large_n_solutions_jon.ipynb) where I used the Chi-square test. If you remember the [jury example from the Data 8 text Computational and Inferential Thinking](https://inferentialthinking.com/chapters/11/2/Multiple_Categories.html), the authors use a statistic called "Total Variation Distance" to measure the overall difference between population proportions and proportions in the jury pool and then show us a distribution of that statistic to demonstrate how unlikely the particular jury draw would be.

I am not sure why I didn't think about this last Spring, but maybe it **is** good to stand up in front of class and profess in order to see things clearly. Proportions for binomially distributed data may be like means (thus you can use a z-test), but the proportions in the jury problem are not.