LS 123 Data, Prediction, and Law

Spring 2022

Meeting 3

**Prediction, Randomness, Visualization**

If you haven't filled out the [student questionnaire](https://docs.google.com/forms/d/e/1FAIpQLSf5PWsxyPlN8vZA4WMlY2jLmyTgrpSMQs9hn7UApmnPQPD_9g/viewform?usp=sf_link) please do so!

Discussion questions for Meeting 3. Since we are just talking methods, you may want to enliven the discussion by starting on Lab 4 and tying examples to that. Again, threads are better than repeating material someone else has already contributed. Today is more review and some practice with empirical distributions and understanding the logic of repeated sampling. Note that Lab 4 uses the ANES 2016 Pilot Survey, rather than the main survey. For our purposes that does not make much difference but it has a different sampling strategy and so on.

1. What does a histogram show? Why is it an important first step when we are exploring a feature/variable in a dataset?
2. Describe the process known as bootstrapping. When do you think you might need to use bootstrapping to understand a statistic?
3. **Announcements**
   1. see Chancellor’s announcement: get boosted, get tested, get a good mask, stay home if you are sick
   2. Will try to do the pairing of students into teams for purposes of the Data Investigation Project by the end of next week; the project is where you get to learn from one another, which is one of Berkeley’s tremendous strengths as a university.
   3. Also, Ilya has provided a Git repo (see Meeting 5 in Unit 1 for link) with a computational text analysis toolkit—don’t be afraid to do a project where what you are analyzing is text—this is Legal Studies! You should look at the notebooks that Ilya has prepared first
   4. note that the first thing each team will do is a proposal, which is part of the overall assignment and which has been published on the assignments page—you still have plenty of time
   5. Today’s lab has fewer glitches and less actual code writing, but useful intro to visualizing data (and the problems with survey data in that regard); the bootstrapping part takes a long time to run (you can put a counter in the loop if you want to see its progress)
   6. we should publish Problem Set 1 next week
   7. **A brief note on boxplots**

Remember on Monday when I was unsure about what the whiskers on the [boxplots created by matplotlib](https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.boxplot.html) actually represented? They are from statistician and proponent of exploratory data analysis John [Tukey's formulation of a boxplot](https://www.graphpad.com/support/faq/five-ways-to-plot-whiskers-in-box-and-whisker-plots/). So the whiskers do not represent a percentile but instead represent Tukey's view of data spread (which is defined at the second link). That means the top whisker is the largest value less than the sum of the 75th percentile plus 1.5 times the interquartile range.

1. Lab from Monday introducing Pandas and talking about social science variables
   1. We are looking at intentionally generated data that researchers want to use make inferences about the world
      1. Causality
      2. Generalizations (Republicans support X policy 65-35)
   2. You should always look at the codebook or variable definitions when it is intentionally generated data like the ANES
   3. Know your variables (I guess this would go for attributes too when we are talking about prediction rather than causal infererence)
   4. Know what sort of measures they are: nominal, ordinal, interval, ratio both to understand what you can find out and to know what sorts of statistical operations you can do
   5. Know what assumptions lie behind the model that you are using (this will be made explicit next week when we talk about regression for causal inference)—e.g., independence of observations, uncorrelated error terms, etc
   6. Today’s lab is short and there is not much coding, so pay attention to issues of measurement again and what sorts of visualization are possible; also, this lab uses the **pilot study** rather than the full ANES 2016 (so the number of respondents is different, etc.—the codebook link in the lab is dead but you can find the ANES 2016 Pilot study on the top page linked in they syllabus; note that it is an internet survey conducted by YouGov, so it is methodologically different)
   7. We can talk about the differences after the lab
2. Randomness
   1. Randomness is a key concept in statistics
      1. Underlies probabilistic thinking
      2. Essential for causal inference
   2. Probabilistic Thinking
      1. Hypothesis Testing, several machine learning algorithms, etc. rely on *probability distributions*
      2. These arise from a source of randomness
   3. *Causal Inference*:
      1. In a basic sense, relies on samples being divided into “treatment” and “control” groups
      2. Compare the difference in outcome between treatment and control
      3. This works because the *randomness* eliminates *confounding*
3. Conditional Probability and Bayes’ theorem (see Aniket’s notes)
   1. *Conditional Probability*: Probability of an event occurring given that another event has occurred
      1. *P*(*A|B*)
      2. Probability of event A occurring, conditional on event B occurring
   2. Conceptually, conditional probabilities leverage the fact that a probability can change when new information is made available
      1. For instance, your evaluation of the probability of rain occurring would be different depending on whether you observe clouds or not
   3. Monty Hall Problem and Bayes’ Theorem (see slides and [Medium piece](https://towardsdatascience.com/solving-the-monty-hall-problem-with-bayes-theorem-893289953e16))
4. After the lab
   1. Tell us a little about how the ANES 2016 pilot study was carried out, and why
   2. Need to test survey questions
   3. But it is expensive, thus hiring YouGov and its constructed panel of opt-in respondents
   4. note on bootstrapping: each iteration of the confidence interval function does a bootstrap to draw a 500 case sample from the dataset—simulate a new draw from the underlying population (also was the procedure in Data 8 textbook https://inferentialthinking.com/chapters/13/2/Bootstrap.html)
   5. You can then adjust (with the weighting variable) to make sure that in aggregate the answers you get are proportional to their true population percentage
   6. visualization of the 95% intervals of the median—when would you use a viz like that?