Vanderbilt University Political Science - Stats I
Fall 2024 - Prof. Jim Bisbee

13 Lecture 13

13.1 Associations between and among variables

- Until now, we've focused on description and inference regarding one variable, and at times we've considered description and inference regarding comparisons of two variables drawn from different units.

- Now we turn the page to a task that political scientists spend a lot of time doing: considering relationships (or associations) between variables drawn from the *same* units.

- We'll start by considering the relationship between two variables, but quickly move on to considering relationships among many variables.

- First some very familiar terminology:

  - When we talk about a bivariate relationship, we typically refer to the two variables as X and Y.
  - We, of course, usually choose these labels with a causal model in mind: specifically, that X causes Y. But that assumption is UNNECESSARY for what we'll be discussing today. In this vein, X is known as the "independent variable," and Y is known as the "dependent variable."

- We'll proceed in several steps (list on board):

  1. DISPLAYING the relationship
     (a) Cross(tabulations)
     (b) Scatterplot
     (c) Boxplot
     (d) "Binning out" X

2. SUMMARIZING the relationship NONPARAMETRICALLY with central tendencies of Y by values of X:
    (a) TABLE: Summary statistics of Y for values of X
    (b) FIGURES (generally appropriate when X, Y or both are interval-level or higher)
        i. bar chart (more values of X, Y is interval-level)
        ii. scatterplot with smoother, indicating the central tendency of Y by values of X
3. SUMMARIZING the relationship PARAMETRICALLY, that is saying how closely the relationship approximates a perfectly linear relationship
    (a) Correlation
    (b) Bivariate Regression
4. Making INFERENCES about the nature of the relationship in a population from the relationship in a sample
    (a) Non-parametric: Pearson's chi-squared
    (b) Parametric: Correlation
    (c) Parametric: Linear regression

- (Walk through Parts 1 and 2 with handout)

- (Before part 3) To simplify this task, we will often need to resort to *models* that describe the theoretical relationship between the variables. As usual, we face a tradeoff between parsimony and precision. What these models buy us is parsimony: the assumptions we make with models allow us to summarize relationships between and among variables with just a few numbers. But what we pay for with models is that we lose some detail about the world. And if our theoretical model is incorrect, our descriptions, inferences and predictions about relationships between and among variables will be wrong.

- It's worth noting that (so far) our development of statistical tools for making inferences about univariate data has been remarkably free of assumptions. In fact, we can develop a list of these assumptions–and it's a short list:
    – in making inferences about the population mean, μ, with *large* samples,

        - **identicality** is necessary for our estimator, $\overline{Y}$, to be an unbiased estimator of μ.
        - **independence** is necessary in order for us to say that the variance of $\overline{Y}$ is equal to $\frac{\sigma^2}{n}$.
            - both of these assumptions are met when we have a **random sample**.
    - in making inferences about the population mean, μ, with *small* samples, we need an additional assumption:
        - the distribution of the underlying population is **Normal**.
            - although the tools we've learned are robust under moderate departures from this assumption.
    - finally, in making inferences about the differences between two population means, we need additional assumptions:
        - in large samples:
            - the two samples are drawn **independently**
        - in small samples:
            - the two samples are drawn **independently**, they have the same variance, and the underlying populations are Normal (although, again, these tools are robust under moderate departures from this last assumption).

- Of course, there are lots of instances when these assumptions don't hold, and statisticians spend a lot of time thinking about how to revise their tools to account for these cases. But still, it's a remarkably short list. It will get a lot longer as we move to bivariate and multivariate analysis.

- To begin thinking together about parametric ways to describe a bivariate relationship, let's revisit the notion of correlation. You'll recall the population correlation coefficient, ρ, which is equal to

$$\rho = \frac{COV(Y_1, Y_2)}{\sigma_1 \sigma_2} = \frac{E\left[(Y_1 - \mu_1)(Y_2 - \mu_2)\right]}{\sigma_1 \sigma_2}.$$

- ρ is, of course, is a theoretical quantity. Like μ or σ^2, we never actually observe it. But the maximum-likelihood estimator of ρ is the sample correlation coefficient, r :

$$r = \frac{\sum_i \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_i \left(X_i - \overline{X}\right)^2 \sum_i \left(Y_i - \overline{Y}\right)^2}}.$$