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

Lecture 7

7.1 Sampling, sample statistics and sampling distributions

- Let's think about our journey through the world of inferential statistics so far. Our goal the entire time has been to come up with the best possible way to make inferences about populations from samples. To do this, we have:
  - recast social phenomena as *experiments* that yield observations for analysis.
    - we called the outcomes of these experiments *events*
    - and defined the *sample space* of an experiment as the set of all the events that are possible
    - and considered carefully how we assign probabilities to events of interest.
  - we then defined a *random variable* as a function mapping a sample space to the real numbers
    - and then discussed the ways we describe the probability distribution of a random variable...
    - and the probability distribution of a *function* of random variables.
- In this context, observed social phenomena (election results, outbreaks of war, passage of legislation) can all be considered realizations of random variables.
  - Let's be a little more clear about this. The phenomena that social scientists study are random events.
  - This may sound odd: in common usage we say something is "random" when it cannot be anticipated.
  - Social scientists use this term differently. When we say an event is random, we mean that it is probabilistic rather than deterministic.

  - For a random event, we attempt to specify the causal processes that alter the chance that it occurs. But we cannot specified the causal process that guarantee the event will occur. If we could, we would be studying a deterministic event.
    * An election is a good example of a random event. The best we can do is develop a theory that specifies factors important to determining election winners and then hypothesize that these factors change the odds of a particular result.
  - Thus, saying that teh variables in our theories are random variables amounts to saying that we expect that the values of these variables that we observe are draws from from an associated probability distribution.

- Finally, we arrived at a very powerful result. If the probability distribution is an accurate representation of the population frequency distribution, then the *expected value* of a random variable is the population mean, μ, where we define expected value as

$$\begin{aligned} E(Y) &\equiv \sum_y yp(y) \text{ in the discrete case and} \\ E(Y) &\equiv \int_y yf(y)\,dy \text{ in the continuous case.} \end{aligned}$$

- Now it's time to put all this theory to work helping us undertake the fundamental challenge we face in statistics: making inferences from samples to populations.

- The relevance of what we've learned previously to this task is that estimates are almost always functions of the n random observations that appear in a sample. Therefore, they are:
  - outcomes of experiments that are themselves random variables with their own probability distributions.

- If we can be very specific about the process giving rise to the sample, we can develop an **estimator** to make inferences from the sample to the population.

- An **estimator** is a *rule*, often expressed as a formula, that tells us how to calculate an estimate from a sample.

- As an example, consider the following as an estimator for the population mean, μ :

- Draw a random sample of n observations, $y_1, y_2, ...y_n$, from the population and employ the observed sample mean

$$\overline{Y} \equiv \frac{y_1 + y_2 + ... + y_n}{n} = \frac{1}{n}\sum_i y_i$$

- As an estimator for μ, Y-bar seems pretty intuitive. But let's be precise about just how good this estimate is for μ.
  - To do this, we first think of each of the n draws that gave rise to the sample as a realization of a random variable.
  - An example: we want to know the mean income of the American population, μ.
  - To do so, I draw a sample of the American population and ask each person in the sample one question: what is your annual income?
  - Denote the response given by the first person in my sample, draw number 1, as y_1.
  - Now here's the key concept: y_1 is one of literally millions of responses I could have observed in my first survey participant. Therefore it is a realization of the random variable Y_1.
  - Now do this again: pick participant number 2 and ask him or her the same question. My observation y_2 is again one of millions of responses I could have observed for respondent number 2. Therefore it is a realization of the random variable Y_2.
  - So if we think about it this way, a sample of n observations is the realization of how many random variables? n: from $Y_1, Y_2, ...Y_n$.
  - OK, now let's think again about Y-bar.

$$\overline{Y} = \frac{1}{n}\sum_i Y_i,$$

  - In this context, our sample of n observations is just one possible realization of Y-bar.
  - That is, Y-bar is a *function* of random variables, and therefore is *itself* a random variable.

- Note use of capital letters here, we are talking about theoretical, not observed, quantitites. In keeping with our convention, little $\bar{y}$ is a realization of the random variable $\overline{Y}$.
- Because it is a random variable, $\overline{Y}$ has a probability distribution. To specify it, we will stipulate some simple but powerful assumptions that hold whenever we have a **random sample**. (Recall how we defined random sample: each of the $\binom{N}{n}$ different possible samples has an equal probability of being drawn.)

- This generates the canonical case, in which we have a function of the random variables $Y_1, Y_2, ...Y_n$ observed in a random sample from a population of interest.
  - When we have a random sample from a large enough population, we can safely assume that the RVs $Y_1, Y_2, ...Y_n$ are independent and identically distributed ("i.i.d.").
    * To recall: $Y_1, Y_2, ...Y_n$ are said to be **independent** iff

$$F(y_1, y_2, ...y_n) = F_1(y_1)F_2(y_2) \cdot ... \cdot F_n(y_n) \text{ for every } n\text{-tuple } (y_1, y_2, ...y_n), \text{ and}$$

  - Now we introduce another concept: $Y_1, Y_2, ...Y_n$ are said to be **identically distributed** iff

$$F_1(y_1) = F_2(y_2) = ... = F_n(y_n) = F(y) \text{ for } y_1, y_2, ...y_n.$$

  - In this context, $\overline{Y} = \frac{1}{n}\sum_i Y_i$ is a **sample statistic**, which we define as a function of the observable random variables in a sample and known constants.

- We know (from the handout last time) that $E(\overline{Y}) = \mu$, [note that this is due to the identicality assumption] and so we can be assured that, on average, $\overline{Y}$ should equal μ. So now let's ask, how good of an estimate is it?
  - A straightforward measure of "goodness" would be how far off we can expect any realization of $\overline{Y}$ to be from the population mean, μ.
  - And since the mean of $\overline{Y}$ is μ, this quantity is just the standard deviation of $\overline{Y}$,or $\sigma_{\overline{Y}}$. But THIS of course is just $\sqrt{VAR(\overline{Y})}$, which we showed last time to be $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$. [recall that we need BOTH identicality and independence to achieve this result.]

- Well, because sample statistics are themselves random variables, they have probability distributions (recall: table, graph, formula). We have a special term for the probability distributions of sample statistics: **sampling distributions**. The sampling distribution of a sample statistic is a theoretical model for the possible values of the statistic we would expect to observe through repeated sampling.
- The expected value and the variance of a sample statistic are important properties of the statistic's sampling distribution. For example, here are graphs depicting two sampling distributions of Y-bar: Y-bar and Y-bar-prime:

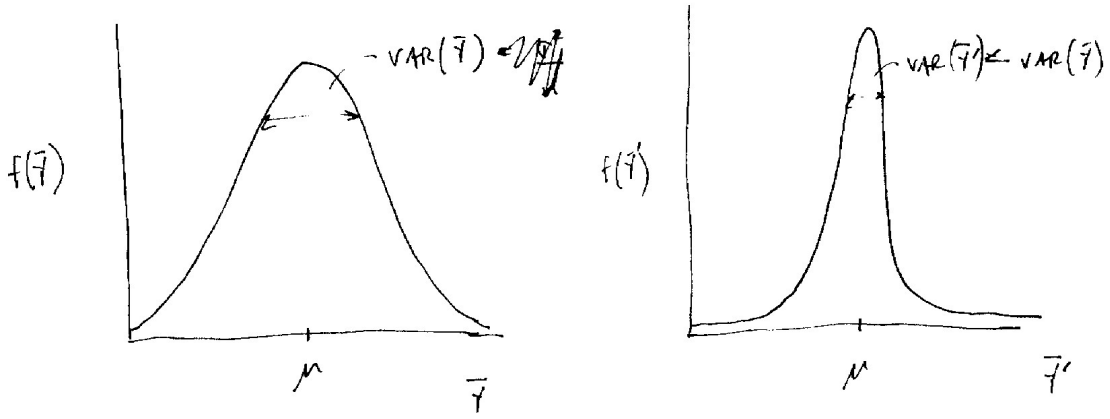

Figure 1: Variance of $\bar{y}$

- Y-bar and Y-bar-prime have the same expected value: μ. But clearly Y-bar prime, which has a smaller variance, is a better estimate of μ than Y-bar. It's closer, on average, to μ than Y-bar.
- So:
  - we've got $\overline{Y}$, an estimate for μ that, on average, equals μ.
  - we now also know how good an estimate this is: on average, it is $\frac{\sigma}{\sqrt{n}}$ units away from μ.
- But now an additional question arises: what does the distribution of the random variable Y look like? Often, we'll be drawing samples from populations about whose frequency

distribution we have no idea. E.g., I want to estimate μ, the mean number of potatoes eaten by the average American per month. Does the distribution of Y look like this? Or this? Or this?

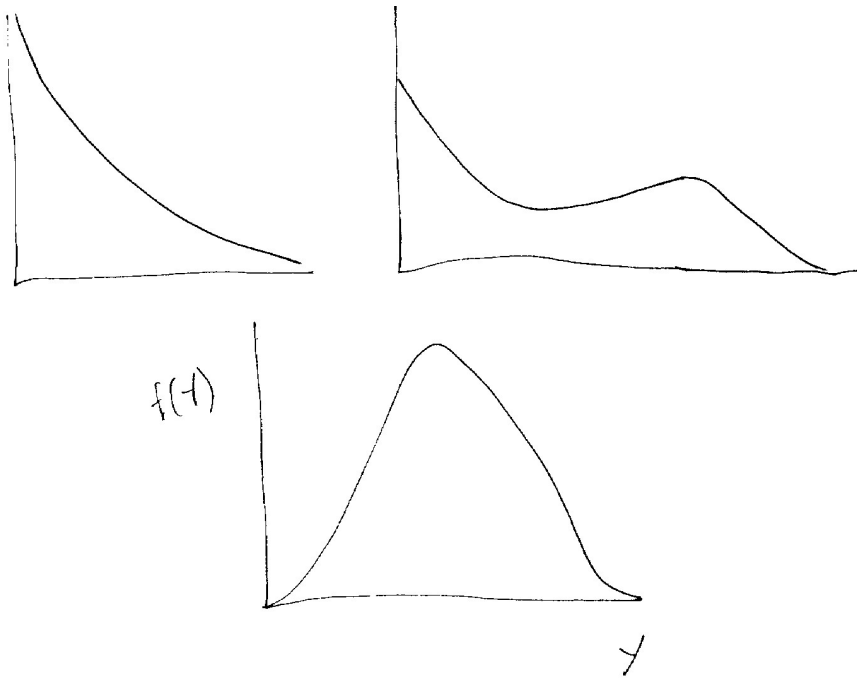

Figure 2: Distributions

- If possible, we'd really like to avoid making assumptions about the shape of the population distribution of Y in order to describe the distribution of Y-bar.
- And, it turns out, we can...

7.2 The Central Limit Theorem

- ...because if my sample size is large enough, I don't need any assumptions about the distribution of Y to describe the sampling distribution of Y-bar. For it turns out that if $Y_1 ... Y_n$ are i.i.d, then:
  - Y-bar has a sampling distribution that is approximately Normal
  - as the sample size becomes large.
  - This is the **central limit theorem**.

- Before digging into the math (and we won't dig that far), let's have a look at a few distributions that convey the intuition (go over handout: "Probability Distributions of Y and Simulated Sampling Distributions of $\overline{Y}$")

- The CLT is more formally stated as follows:

- Let $Y_1, Y_2, ...Y_n$ be i.i.d. random variables with $E(Y_i) = \mu$ and $VAR(Y_i) = \sigma^2$. Define

$$U_n \equiv \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}.$$

  That is, we standardize Y-bar by (1) subtracting its hypothesized mean (μ) and then (2) dividing this difference by Y-bar's standard deviation $(\frac{\sigma}{\sqrt{n}})$. Then the CDF of U_n converges in probability to (where "converges in probability to" means "as n becomes large it is distributed as") the standard normal CDF. That is,

$$\lim_{n\to\infty} F_{U_n}(u) = \lim_{n\to\infty} P(U_n \leq u) = P(Z \leq u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \text{ for all } u.$$

- Take a moment to appreciate the power of the CLT. It yields the sampling distribution of Y-bar *without requiring any assumptions about the probability distribution of* Y. We will not cover the proof of the CLT in this class, as it requires knowledge of moment generating functions, which itself requires knowledge of Taylor series expansions. Those equipped with these tools who want to satisfy themselves with a proof may see section 7.4 of the text.

7.3 Estimation

- With the sampling distributions yielded by the CLT in hand, we can develop **estimators** of population parameters. An estimator is a rule–often expressed as a formula–that tells us how to calculate an estimate of a population parameter.

  - Two kinds of estimates that we'll focus on here include

    * **point estimates**–in which a single value, or point, is given as the estimate of the parameter

    * **interval estimates**-in which two values are used to construct an interval that we believe contains/traps/encloses the parameter of interest.

- So the sample mean, $\overline{Y} = \frac{1}{n}\sum_i Y_i$, is one possible point estimator of the population mean μ.
- But there are lots of others. For example, consider the estimator $\overline{Y_B} = \frac{1}{n}\sum_i (Y_i + 1)$.
- Intuitively, we know this is a worse estimator than Y-bar. But can we put some meat on this intuition? We do this by specifying two desirable criteria for evaluating a potential estimator:
  - 1. unbiasedness
  - 2. (relatively) small variance, which is also known as **efficiency** or **precision**.
- First, some terminology. We typically write the population parameter for which we seek a point estimate as θ, and a proposed estimator for this parameter as $\widehat{\theta}$.
- Now, here is why we've spent a fair amount of time learning about the math of expectations: we can use these tools to show whether an estimator is unbiased and to determine how precise it is.
- In fact, we define an estimator as unbiased if–in expectation–it is equal to the parameter it claims to estimate. That is, an estimator is unbiased if the expected value of its distribution is the parameter. Formally,
  - $\widehat{\theta}$ is an **unbiased estimator** for θ if $E(\widehat{\theta}) = \theta$.
  - If $E(\widehat{\theta}) \neq \theta$, then we say $\widehat{\theta}$ is **biased**.
  - The **bias** of a point estimator is given by $B(\widehat{\theta}) = E(\widehat{\theta}) - \theta$.
- E.g. So one way to say that the estimator $\overline{Y_B}$ isn't a good estimator is to show that it is a

biased estimator for μ. We do that by showing that $E(\overline{Y_B}) \neq \mu$, and thus that $B(\overline{Y_B}) \neq 0$.

$$\begin{aligned} E(\overline{Y_B}) &= E\left[\frac{1}{n}\sum_i (Y_i + 1)\right] \\ &= \frac{1}{n}\left\{\left[\sum_i E(Y_i)\right] + nE(1)\right\} \\ &= \frac{1}{n}\{[n\mu] + n\} \\ &= \mu + 1 \neq \mu, \text{ and } B(\overline{Y_B}) = 1. \end{aligned}$$

- So much for the first criterion. Our second criterion is that we'd like our estimator to be as close as possible to θ in repeated sampling. In mathematical terms, we of course want the variance of the sampling distribution of our estimator to be as small as possible. Recalling the definition of the variance of a random variable, we write the variance of an estimator $\widehat{\theta}$ as $VAR(\widehat{\theta}) = E\left\{[\widehat{\theta} - E(\widehat{\theta})]^2\right\}$. In an ideal world, we wish for this to be as small as possible.
- Sometimes we face a tradeoff between reducing an estimator's bias and reducing its variance. One way to evaluate the tradeoff is to minimize an estimator's **mean square error (MSE)**, which is defined as the expected value of the square of the distance between the estimator and the parameter
$$MSE(\widehat{\theta}) = E\left[\left(\widehat{\theta} - \theta\right)^2\right].$$
- It can be shown that the MSE is equal to the sum of an estimator's variance plus the square of its bias:
$$MSE(\widehat{\theta}) = VAR(\widehat{\theta}) + \left[B(\widehat{\theta})\right]^2.$$
  - Ask: Why does this measure include the square of its bias?

7.4 An example

- For example, let us say that we wish estimate the parameter $\mu_1 - \mu_2$, the difference in means of two different populations drawn from independent samples. Is the intuitive estimator,

$\overline{Y_1} - \overline{Y_2}$, unbiased? Let's see:

$$\begin{aligned} E(\overline{Y_1} - \overline{Y_2}) &= E(\overline{Y_1}) - E(\overline{Y_2}) \\ &= \mu_1 - \mu_2. \end{aligned}$$

- Yes, $\overline{Y_1} - \overline{Y_2}$ is an unbiased estimator for $\mu_1 - \mu_2$. Now what is its variance? Well,

$$\begin{aligned} VAR(\overline{Y_1} - \overline{Y_2}) &= VAR(\overline{Y_1}) + VAR(\overline{Y_2}) + 2COV(\overline{Y_1}, \overline{Y_2}) \\ &= VAR(\overline{Y_1}) + VAR(\overline{Y_2}) \ \ [\overline{Y_1}, \overline{Y_2} \text{ independent}] \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}, \ \ [\text{variance of Y-bar}] \end{aligned}$$

  where σ_1^2 and σ_2^2 are the variances of populations 1 and 2 respectively.

- Note that we can talk about the standard error of an estimator $\widehat{\theta}$, which we write $\sigma_{\widehat{\theta}}$. It is just the square root of the variance of the estimator. The standard error of the estimator $\overline{Y_1} - \overline{Y_2}$ is thus $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

- Another variant of the CLT (which does not require identicality, but instead requires some other conditions) tells us that the sampling distribution of a sum of independent random variables (like $\overline{Y_1}$ and $\overline{Y_2}$) approximates the Normal as n becomes large. And so

$$\overline{Y_1} - \overline{Y_2} \tilde{} N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \text{ as } n_1, n_2 \longrightarrow \infty.$$