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

Lecture 9

9.1 Where we are

Just to quickly review:

- We are keenly interested in identifying a good estimator for the population mean, μ, from a random sample of data from that population.
- $\overline{Y} \equiv \frac{1}{n} \sum_i Y_i$, the sample mean, is an obvious choice for such an estimator.
- We proceed by modeling the sampling process yielding n observations as a series of random variables $Y_1, Y_2, ...Y_n$. They are independent, and they are identically distributed: that is, they all have the some CDF F, the same mean μ and the same variance σ^2. With this in hand, we:
  - established that $\overline{Y}$ is an unbiased estimator of μ, i.e. that $E\left(\overline{Y}\right) = \mu$.
  - we showed that its variance is $VAR\left(\overline{Y}\right) = \sigma^2_{\overline{Y}} = \frac{\sigma^2}{n}$, and thus its standard deviation $\sqrt{VAR\left(\overline{Y}\right)} = \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$.
- That's good. Now we want to know how close, on average, the estimator Y-bar is to μ.
  - Well, the central limit theroem tells us that the sampling distribution of Y-bar is distributed Normal as n becomes large. We typically find it more useful to write this in terms of the *standardized* version of Y-bar, that is

$$U_n \equiv Z \equiv \frac{\overline{Y} - \mu}{\sigma_{\overline{Y}}} = \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}},$$

  - where the CLT tells us that this converges in probability to the *standard* Normal:

$$F\left(\frac{\overline{Y} - \mu}{\sigma / \sqrt{n}}\right) \xrightarrow{p} \Phi.$$

- This allows us to begin to quantify how close Y-bar is, on average, to μ. Since Y-bar is distributed Normal, when n is large it is generated through a process that yields intervals trapping μ in repeated sampling $1-\alpha$ percent of the time, where α and $z_{\alpha/2}$ satisfy

$$P(\overline{Y} - z_{\alpha/2}\sigma_{\overline{Y}} \leq \mu \leq \overline{Y} + z_{\alpha/2}\sigma_{\overline{Y}}) = 1 - \alpha.$$

- For any α we pick, we can find the appropriate $z_{\alpha/2}$ with statistical software or tables; it is the value at with the CDF of the standard Normal is evaluated that yields $\alpha/2$.

- And because $\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$, we know that

$$P(\overline{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha.$$

- But to quantify the distribution of Y-bar, we need one more thing. We need to contend with σ, the standard deviation of Y. To deal with this, we:
  - Identified an estimator $S_U^2 \equiv \frac{\sum_i \left(Y_i - \overline{Y}\right)^2}{n-1}$, and showed that it is unbiased for σ^2, the population variance.
  - We also showed that this estimator is *consistent* for σ^2; i.e. that $S_U^2 \xrightarrow{p} \sigma^2$.
  - We want to get to the point where we can justify substituting S_U for σ and saying

$$F\left(\frac{\overline{Y} - \mu}{S_U/\sqrt{n}}\right) \xrightarrow{p} \Phi.$$

  - This is exactly what we are about to do.

9.2 Slutzky's Theorem

- To justify our substitution of S_U for σ, we'll need one more tool: *Slutzky's Theorem* (love that name). [Put this on separate board.] This theorem tells us that:
  - if the distribution of some function is such that $F(U_n) \xrightarrow{p} \Phi$ and
  - if the distribution of some other function W_n is such that $F(W_n) \xrightarrow{p} 1$, then

– $F\left(\frac{U_n}{W_n}\right) \xrightarrow{p} \Phi$.

– In words, Slutzky's theorem tells us that the ratio of a function that converges to the Standard Normal over a function that converges to 1 itself converges to the Standard Normal.

9.3 Putting it all together

- OK, now we're ready to prove the powerful result we've been seeking:

$$F\left(\frac{\overline{Y}-\mu}{S_U/\sqrt{n}}\right) \xrightarrow{p} \Phi.$$

Proof:

– Begin by re-writing $F\left(\frac{\overline{Y}-\mu}{S_U/\sqrt{n}}\right) = F\left(\frac{\frac{\overline{Y}-\mu}{\sqrt{n}}\cdot\frac{1}{\sigma}}{S_U\cdot\frac{1}{\sigma}}\right) = F\left(\frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\frac{S_U}{\sigma}}\right)$. If we can show this final expression converges to the standard Normal, then we know that $\frac{\overline{Y}-\mu}{S_U/\sqrt{n}}$ does, too.

– Note that $\frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\frac{S_U}{\sigma}}$ is a ratio of a function that converges to the Standard Normal over the function $\frac{S_U}{\sigma}$:

  * The CLT tells us that
$$F\left(\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}\right) \xrightarrow{p} \Phi.$$

  * So if we can show that $\frac{S_U}{\sigma}$ converges to 1, then Slutzky's Theorem implies that
$$F\left(\frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\frac{S_U}{\sigma}}\right) \xrightarrow{p} \Phi.$$

– To do this,

– recall that we've shown $S_U^2 \xrightarrow{p} \sigma^2$ [consistency of S_U^2 .]

– Now note that $\frac{S_U}{\sigma} = +\sqrt{\frac{S_U^2}{\sigma^2}}$. Because the function $g(x) = +\sqrt{\frac{x}{c}}$ is continuous if both x, c positive, then we can invoke the rule that if $\widehat{\theta} \xrightarrow{p} \theta$ and $g(\cdot)$ continuous at θ, then $g(\widehat{\theta}) \xrightarrow{p} g(\theta)$.

– Here $\frac{S_U^2}{\sigma^2} \xrightarrow{p} \frac{\sigma^2}{\sigma^2} = 1$, and $\sqrt{}$is clearly continuous at 1, so $\frac{S_U}{\sigma} = +\sqrt{\frac{S_U^2}{\sigma^2}} \xrightarrow{p} \sqrt{\frac{\sigma^2}{\sigma^2}} = 1$.

- Now we invoke Slutzky's Theorem to show that the distribution of this ratio, and therefore the distribution of $\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}$, converges in probability to the standard Normal.

- Whew, that was a lot of work! What does it buy us? It tells us that when n is large, $\frac{\overline{Y}-\mu}{S_U/\sqrt{n}}$ is distributed approximately standard Normal, whatever the distribution of the underlying population.

- Therefore it follows that

$$
\begin{aligned}
P\left[-z_{\alpha/2} \leq \frac{\overline{Y}-\mu}{S_U/\sqrt{n}} \leq z_{\alpha/2}\right] &\approx 1-\alpha \text{ and so} \\
P\left[\overline{Y} - z_{\alpha/2}\left(\frac{S_U}{\sqrt{n}}\right) \leq \mu \leq \overline{Y} + z_{\alpha/2}\left(\frac{S_U}{\sqrt{n}}\right)\right] &\approx 1-\alpha.
\end{aligned}
$$

- Thus $\overline{Y} \pm z_{\alpha/2}\left(\frac{S_U}{\sqrt{n}}\right)$ forms a valid **large-sample CI** for μ. And this is the challenge we originally faced. We can now substitute $\frac{S_U}{\sqrt{n}}$ for $\sigma_{\widehat{\theta}}$.

9.4 Examples of Large-Sample CIs

- Let's revisit the notion of a large-sample CI with an example.

- The American Community Study (ACS) is a program of the Census Bureau that estimates quantities of interest in the population using a large-sample survey.

- For example, the mean household income of New York State was estimated to be \$76,247 using a sample of about 350,000 households. The unbiased estimate of the population standard deviation is $S_U = 61,427$. What is the 90% CI associated with this estimate?

  - Recall that we write the $100\,(1-\alpha)$ percent CI for the population mean, μ as

$$
\overline{Y} \pm z_{\alpha/2}\left(\sigma_{\overline{Y}}\right), \text{where } z_{\alpha/2} = -\Phi^{-1}\left(\frac{\alpha}{2}\right) \text{ and } \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}.
$$

  - Let's first find $z_{\alpha/2}$.
  - What is alpha here? It's 1 minus the confidence coefficient (in this case, .90), or .10 .
  - So what is $z_{.10/2} = z_{.05}$? It's $z_{.05} = -\Phi^{-1}(.05)$. Calculate this by typing `qnorm(.05)` in `R`, obtaining -1.64. So $z_{.05} = 1.64$.

– We're almost there. Our 90% CI can be written

$$\overline{Y} \pm z_{\alpha/2}\left(\sigma_{\overline{Y}}\right) = \$76,247 \pm (1.64)\ \sigma_{\overline{Y}}.$$

– Recall that we've shown we can substitute

$$\begin{aligned} S_U &= \sqrt{\frac{\sum\left(y_i - \overline{y}\right)^2}{n-1}} \text{ for the population standard deviation,} \\ &\quad \text{and thus can rewrite our CI as} \\ \overline{Y} \pm z_{\alpha/2}\left(\sigma_{\overline{Y}}\right) &= \$76,247 \pm (1.64)\ \left(\frac{S_U}{\sqrt{n}}\right) \\ &= \$76,247 \pm (1.64)\ \left(\frac{61,427}{\sqrt{350,000}}\right) \\ &= \$76,247 \pm (1.64)\ \ (103.83) \\ &= \$76,247 \pm 170.28,\ \text{or}\ \left[\$76,077,\ \$76,417\right]. \end{aligned}$$

9.5 Another example of a large-sample CI: proportions

- CNN poll, Oct 16-18, 2009 with sample of 1,038 American adults.
- Finding: 64 percent say they have a "favorable" opinion of Michelle Obama; 36% do not.
- Let's construct a 95% large-sample CI around this estimate.
- Before proceeding, let's think:

  – In the previous example, we wrote our CI for the population mean, μ, as

  $$\widehat{\mu}_{LB}, \widehat{\mu}_{UB} = \overline{Y} \pm z_{\alpha/2}\left(\sigma_{\overline{Y}}\right), \text{where } z_{\alpha/2} = -\Phi^{-1}\left(\frac{\alpha}{2}\right) \text{ and } \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}.$$

  – But recall that the CLT tells us we can also write this more generically for *any* estimator that is a linear combination of random variables that are i.i.d. as

  $$\widehat{\theta}_{LB}, \widehat{\theta}_{UB} = \widehat{\theta} \pm z_{\alpha/2}\left(\sigma_{\widehat{\theta}}\right).$$

  – And in this example, our parameter of interest is p: the proportion of Americans viewing Michelle Obama favorably. Our estimator is $\widehat{p} = \frac{Y}{n}$, where $Y = 0$ if Obama is viewed unfavorably and $Y = 1$ if she is viewed favorably. We've shown previously that $\widehat{p}$ is unbiased for p. So let's write $\widehat{p} = .64$.

- Now rewrite our CI of interest as

$$\widehat{p}_{LB}, \widehat{p}_{UB} = \widehat{p} \pm z_{\alpha/2}\left(\sigma_{\widehat{p}}\right)$$

- Now think:
  * We have $\widehat{p}$.
  * We'll find $z_{\alpha/2}$ the usual way. (It's equal to - `qnorm`(.025) = 1.96.)
  * What about $\sigma_{\widehat{p}}$?
- A few lectures ago we showed that

$$VAR\left(\widehat{p}\right) = VAR\left(\frac{Y}{n}\right) = \frac{1}{n^2} VAR\left(Y\right) = \frac{np\left(1-p\right)}{n^2} = \frac{p\left(1-p\right)}{n}.$$

- And so

$$\sigma_{\widehat{p}} = \sqrt{\frac{p\left(1-p\right)}{n}}.$$

- We can substitute $\widehat{p}$, our estimate of p, in the formula for $\sigma_{\widehat{p}}$,and so a large-sample CI for a population proportion p can be written

$$\widehat{p}_{LB}, \widehat{p}_{UB} = \widehat{p} \pm z_{\alpha/2}\left(\sqrt{\frac{\widehat{p}\left(1-\widehat{p}\right)}{n}}\right).$$

- To return to our example, we can write the 95% CI about our estimate of the proportion of the population having a favorable opinion of Michelle Obama as

$$\begin{aligned} & .64 \pm 1.96\left(\sqrt{\frac{.64\left(1-.64\right)}{1038}}\right) \\ = \quad & .64 \pm .029 \end{aligned}$$

- This corresponds to the poll's published "Margin of Error" of "plus or minus 3 percentage points." When you see this reported with any poll, it is shorthand for saying how big the

95% CI is around the polling result.

9.6 A large-sample CI for the difference between two proportions

- The same logic underlies the construction of a large-sample confidence interval for the difference between two proportions. Consider this example:

  - * In a Zogby Poll conducted with 1,203 likely voters nationwide between Oct 24-26, 2008, Barack Obama led John McCain, 52.5 percent to 47.5 percent, among those expressing a preference.
    * This is a tracking poll. In the previous three-day window of the poll (Oct 21-23), Obama led McCain 55.6 to 44.4 percent (N=1,203).
    * According to the poll, Obama's lead shrunk by about six points in three days. How confident are we that this change is not due to sampling error?
    * Set it up:
    * The parameter we seek is now $p_1 - p_2$, where p_1 = Obama's true support in the first poll (Oct 21-23) and p_2 = Obama's true support in the second poll.
    * The polls may be considered two binomial experiments in which Y_1 is the number of "successes" (here, the # favoring Obama) in the first poll, (no ideological agenda) and Y_2 is the number of of such "successes" in the second poll.
    * An intuitive estimator for this quantity would be $\widehat{p}_1 - \widehat{p}_2$, where the p-hats are the proportions of respondents favoring Obama in the two polls. Is it an unbiased estimator for $p_1 - p_2$?

$$
\begin{aligned}
E(\widehat{p}_1 - \widehat{p}_2) &= E(\widehat{p}_1) - E(\widehat{p}_2) \\
&= E\left(\frac{Y_1}{n_1}\right) - E\left(\frac{Y_2}{n_2}\right) \text{ [$\widehat{p}_1$ and $\widehat{p}_2$ are functions of the RVs Y_1, Y_2]} \\
& \quad \frac{1}{n_1}E(Y_1) - \frac{1}{n_2}E(Y_2) \\
&= \frac{1}{n_1}n_1 p_1 - \frac{1}{n_2}n_2 p_2 \text{ [$E(Y) = np$ if Y is distributed binomial]} \\
&= p_1 - p_2.
\end{aligned}
$$

* Our next step is to say how precise $\widehat{p}_1 - \widehat{p}_2$ tends to be as an estimator of $p_1 - p_2$.
* We do this by figuring out what the estimator's standard error is. It's

$$\begin{aligned}\sqrt{VAR(\widehat{p}_1 - \widehat{p}_2)} &= \sqrt{VAR(\widehat{p}_1) + VAR(\widehat{p}_2)} \text{ [assume samples drawn independently]} \\ &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\end{aligned}$$

* We make the substitution

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

* Plugging in, we have

$$\begin{aligned}&(55.6 - 52.5) \pm z_{\alpha/2}\sqrt{\frac{(55.6)\,(100-55.6)}{1,203} + \frac{(52.5)(100-52.5)}{1,203}} \\ &3.1 \pm z_{\alpha/2}(2.031).\end{aligned}$$

* Do you recall how we find $z_{\alpha/2}$? We type `qnorm(`$\frac{\alpha}{2}$`)`, substituting our chosen α. You'll remember that $z_{\alpha/2}$ associated with an $\alpha = .05$ is $z_{.025} = -1.96$. So our 95% CI is:

$$3.1 \pm 1.96(2.031) = 3.1 \pm 3.98, \text{ or } [-.9, 7.1].$$

* We are 95% confident that the true change between the two polls was between -.9 and 7.1 percentage points.

- Note that this CI includes zero. So another interpretation of this CI is that we are **not** 95% confident that there was zero change between the two polls. And this, of course, is what we really wanted to know: was there truly any movement between Oct 21-23 and Oct 24-26?
- Now, does the 90% confidence interval about our point estimate include zero?
  - Let's see: our alpha is .10.
  - typing `qnorm(.05)` gives us -1.64. So our 90% CI is:

$$3.1 \pm 1.64(2.031) = 3.1 \pm 3.33, \text{ or } [-.23,\ 6.43].$$

- Still no cigar. At what level of confidence would we be satisfied that there was movement between the two surveys?

- Think: we wish to find some α^* such that the lower bound of the $100 * (1-\alpha)$ CI is greater than zero. That is, find some α^* meeting this criterion:

$$\alpha^* : 3.1 - z_{\alpha^*/2}(2.031) > 0.$$

- To do this, manipulate the expression

$$\begin{aligned} -z_{\alpha^*/2}(2.031) &> -3.1 \\ z_{\alpha^*/2} &< \frac{3.1}{2.031} \\ z_{\alpha^*/2} &< 1.5263 \end{aligned}$$

- So for any alpha such that $z_{\alpha/2} < 1.5263$, we will be $100 * (1-\alpha)$ percent confident that the true change was greater than zero. How do we find this α? Well, if

$$\begin{aligned} z_{\frac{\alpha}{2}} &= -\Phi^{-1}\left(\frac{\alpha}{2}\right), \text{ then} \\ \Phi\left(-z_{\frac{\alpha}{2}}\right) &= \frac{\alpha}{2}, \text{ and} \\ \alpha &= 2\Phi\left(-z_{\frac{\alpha}{2}}\right). \end{aligned}$$

- So in this particular case, $\alpha = 2\Phi(-1.5263)$.
  - To find this alpha, we now type `pnorm`**(-1.5263)** in `R`, which is the CDF of the standard Normal evaluated at its argument. This returns **.063**.
  - Thus $\alpha/2 = .063$ and alpha is thus .126.
  - And thus if we are working with confidence intervals of $100 * (1 - .126) = 87.4\%$ or smaller, we will conclude that there was true movement between the two polls.

- Keep this in mind: it will connect to other concepts we'll be covering today and next lecture.