Lecture 8

Quantitative Political Science

Prof. Bisbee

Vanderbilt University

Lecture Date: 2023/09/26

Slides Updated: 2023-09-26

Agenda

1. Bias and intuition
2. Confidence Intervals
3. σ^2

Bias

- Our gut tells us that $\bar{Y} \equiv \frac{1}{n}\sum_i^n (Y_i)$ is not biased while $\bar{Y}_B \equiv \frac{1}{n}\sum_i^n (Y_i + 1)$ is
    - $\bar{Y}$ is the mean of the sample, so intuition tells us it should be unbiased for μ
- But the **intuitive** estimator is not always the **unbiased** estimator
- Consider $S^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{n}$ as a potential estimator for the population variance σ^2
- Prove that it is biased by taking its expectation

Bias

$$
\begin{aligned}
E[S^2] &= E\left[\frac{\sum_i (Y_i - \bar{Y})^2}{n}\right] \\
&= E\left[\frac{\sum_i (Y_i^2 - 2Y_i\bar{Y} + \bar{Y}^2)}{n}\right] \\
&= E\left[\frac{\sum_i Y_i^2 - 2\bar{Y}\sum_i Y_i + \sum_i \bar{Y}^2}{n}\right]
\end{aligned}
$$

Bias

- Note that the definition of $\bar{Y} \equiv \frac{1}{n}\sum_i Y_i$
- Multiply both sides by n to yield $n\bar{Y} \equiv \sum_i Y_i$
- Thus:

$$
\begin{aligned}
E[S^2] &= E\left[\frac{\sum_i Y_i^2 - 2\bar{Y}\sum_i Y_i + \sum_i \bar{Y}^2}{n}\right] \\
&= E\left[\frac{\sum_i Y_i^2 - 2\bar{Y}n\bar{Y} + \sum_i \bar{Y}^2}{n}\right] \\
&= E\left[\frac{\sum_i Y_i^2 - 2n\bar{Y}^2 + n\bar{Y}^2}{n}\right] \\
&= E\left[\frac{\sum_i Y_i^2 - n\bar{Y}^2}{n}\right]
\end{aligned}
$$

Bias

$$
\begin{aligned}
E[S^2] &= E\left[\frac{\sum_i Y_i^2 - n\bar{Y}^2}{n}\right] \\
&= \frac{1}{n}E\left[\sum_i Y_i^2 - n\bar{Y}^2\right] \\
&= \frac{1}{n}\left(\sum_i E[Y_i^2] - nE[\bar{Y}^2]\right) \\
&= \frac{1}{n}\left(\sum_i (E[Y_i^2] - E[Y_i]^2 + E[Y_i]^2) - nE[\bar{Y}^2] - E[\bar{Y}]^2 + E[\bar{Y}]^2\right) \\
&= \frac{1}{n}\left(\sum_i (VAR(Y_i) + E[Y_i]^2) - nVAR(\bar{Y}) + E[\bar{Y}]^2\right)
\end{aligned}
$$

Bias

$$
\begin{aligned}
E[S^2] &= \frac{1}{n}\left(\sum_i(\sigma^2+\mu^2) - n(\frac{\sigma^2}{n}+\mu^2)\right) \\
&= \frac{1}{n}\left(n(\sigma^2+\mu^2) - \sigma^2 - n\mu^2\right) \\
&= \frac{1}{n}\left(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right) \\
&= \frac{1}{n}\left(n\sigma^2 - \sigma^2\right) \\
&= \frac{n-1}{n}\sigma^2 \neq \sigma^2
\end{aligned}
$$

Bias

- So it turns out the intuitive estimator is **not** the unbiased estimator!
  - Note that $S^2 \equiv \frac{1}{n}\sum_i (Y_i - \bar{Y})^2$ is the appropriate formula for calculating the variance of a sample
  - But it is **not** the appropriate estimator for calculating the variance of a population!
- Although how bad are we messing up if we make this mistake?

$$\begin{aligned} B(S^2) &= E[S^2] - \sigma^2 \\ &= \frac{n-1}{n}\sigma^2 - \sigma^2 \\ &= \left(\frac{n-1}{n} - 1\right)\sigma^2 \\ &= \frac{-\sigma^2}{n} \end{aligned}$$

- As $n \to \infty$, $B(S^2) \to 0$

Interval Estimators

- So we've covered two dimensions of the quality of a **point estimate**: bias and variance
- Recall that we also are interested in **interval estimates**: two numbers that capture the true population parameter
- Specifically, an **interval estimator** is:

  1. A rule...
  2. ...specifying how we use a sample to calculate numbers...
  3. ...that form the endpoints of an interval...
  4. ...containing the parameter of interest θ

- We again are interested in the quality of this concept
  - Want it to contain θ
  - Want it to be narrow

Confidence Intervals

- Interval estimators are commonly called **confidence intervals (CIs)**
- CIs constructed of **upper and lower confidence bounds**
- Probability that CI contains θ is **the confidence coefficient**
  - The fraction of the time...
  - ...in repeated sampling...
  - ...that the CI will contain θ
- Thus we want the confidence coefficient to be **high**
- Denoted as $1 - \alpha$
- Formally:

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_H) = (1 - \alpha)$$

Confidence Intervals

- Let's figure out how to construct a CI for a sample statistic $\hat{\theta}$ that is normally distributed with mean μ and standard error $\sigma_{\hat{\theta}}$
  - Formally, $\hat{\theta} \sim \mathcal{N}(\mu, \sigma_{\hat{\theta}})$
- Standardize the statistic as $Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$
- To construct the CI, choose two **critical values** $-z_{\alpha/2}$ and $z_{\alpha/2}$ such that

$$\begin{aligned} P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) &= \int_{-z_{\alpha/2}}^{z_{\alpha/2}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \\ &= 1 - \alpha \end{aligned}$$

Confidence Intervals

$$P(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \leq z_{\alpha/2}) = 1 - \alpha$$

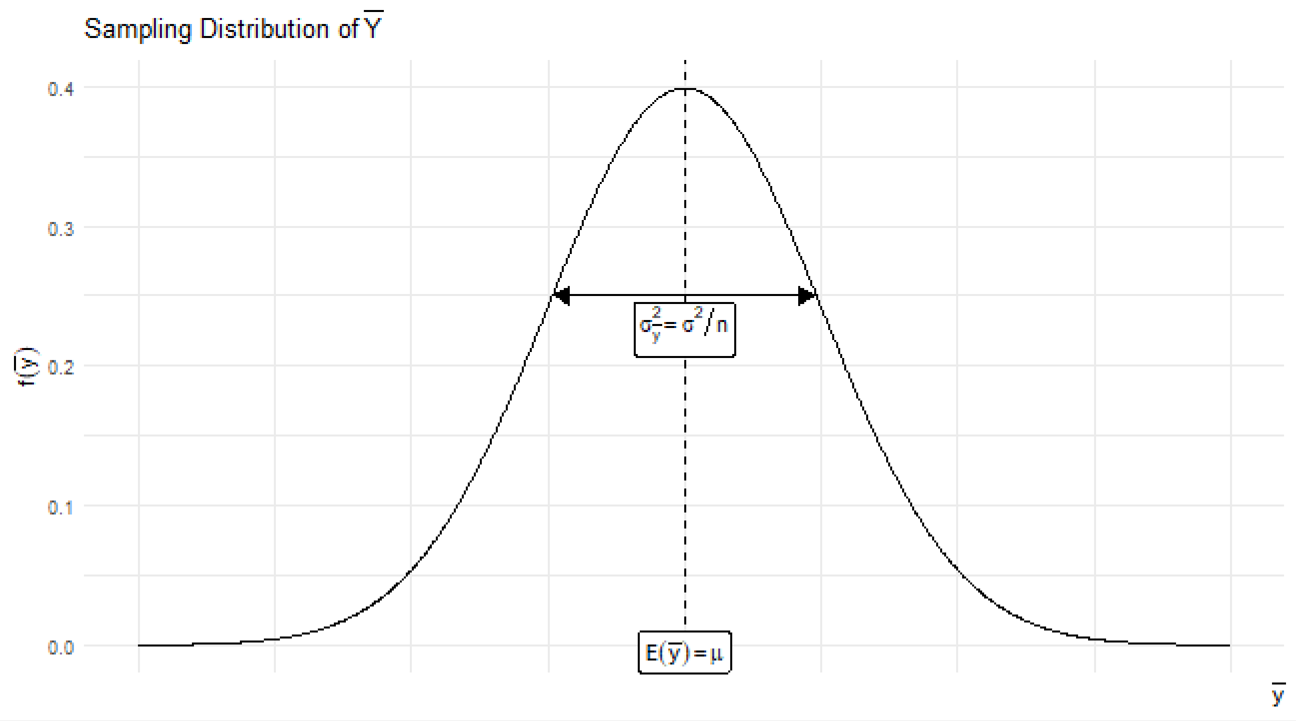
$$P(-z_{\alpha/2\sigma_{\hat{\theta}}} \leq \hat{\theta} - \theta \leq z_{\alpha/2\sigma_{\hat{\theta}}}) = 1 - \alpha$$

$$P(\hat{\theta} - z_{\alpha/2\sigma_{\hat{\theta}}} \leq \theta \leq \hat{\theta} + z_{\alpha/2\sigma_{\hat{\theta}}}) = 1 - \alpha$$

- Thus $\hat{\theta}_L = \hat{\theta} - z_{\alpha/2\sigma_{\hat{\theta}}}$ and $\hat{\theta}_H = \hat{\theta} + z_{\alpha/2\sigma_{\hat{\theta}}}$
- But how do we determine $z_{\alpha/2}$
- CLT to the rescue!

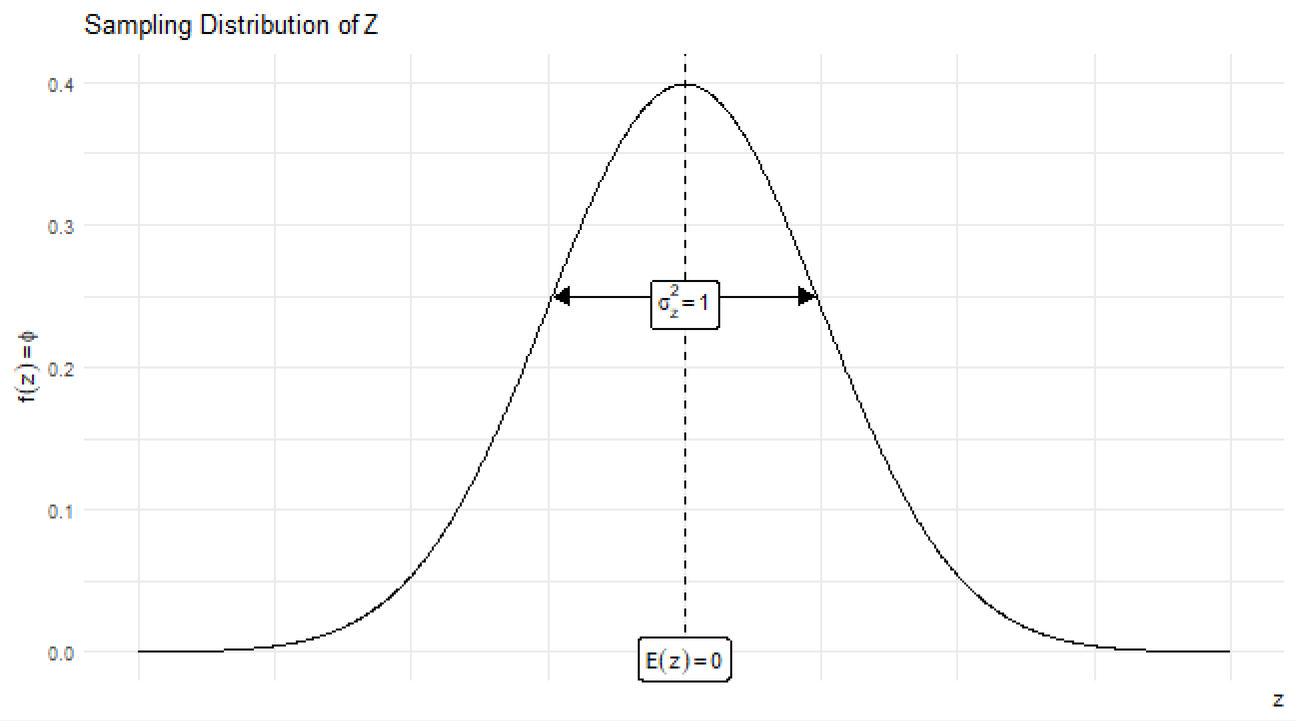
Confidence Intervals

- CLT to the rescue!
- Sampling distribution approximates the normal as $n \to \infty$



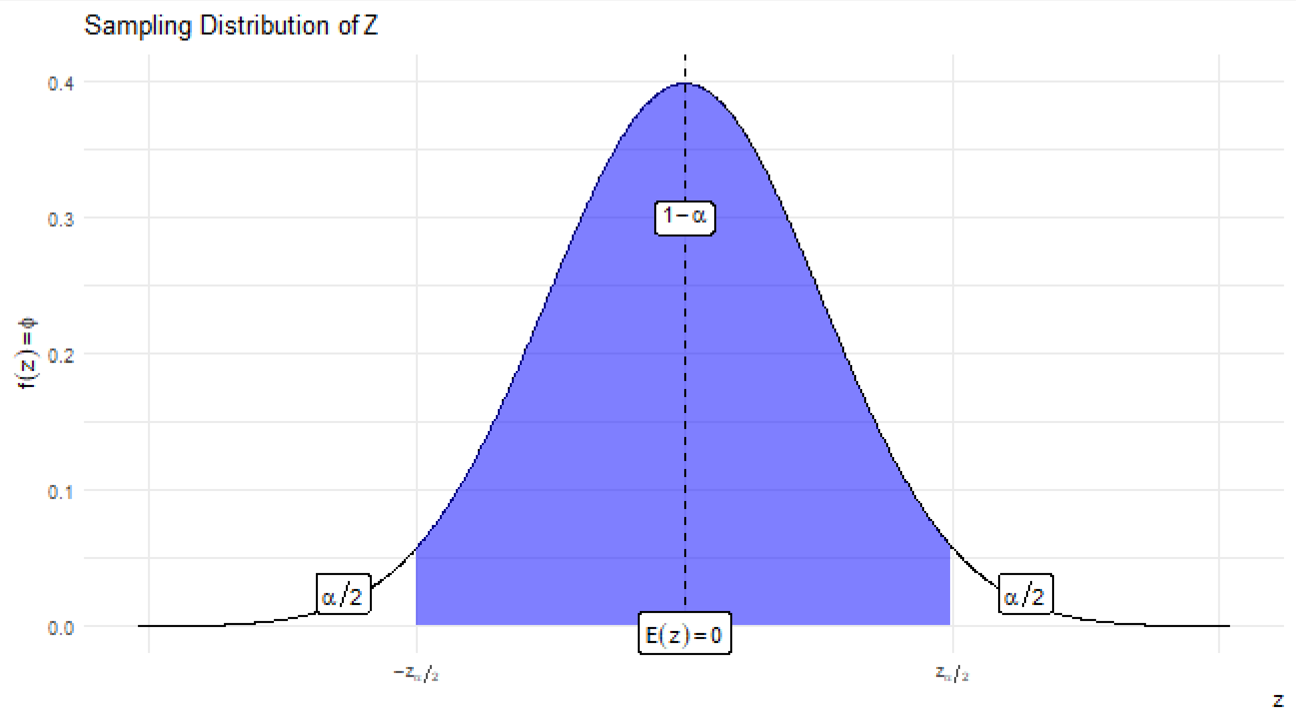
Confidence Intervals

- CLT to the rescue!
- Standardized sampling distribution simplifies!



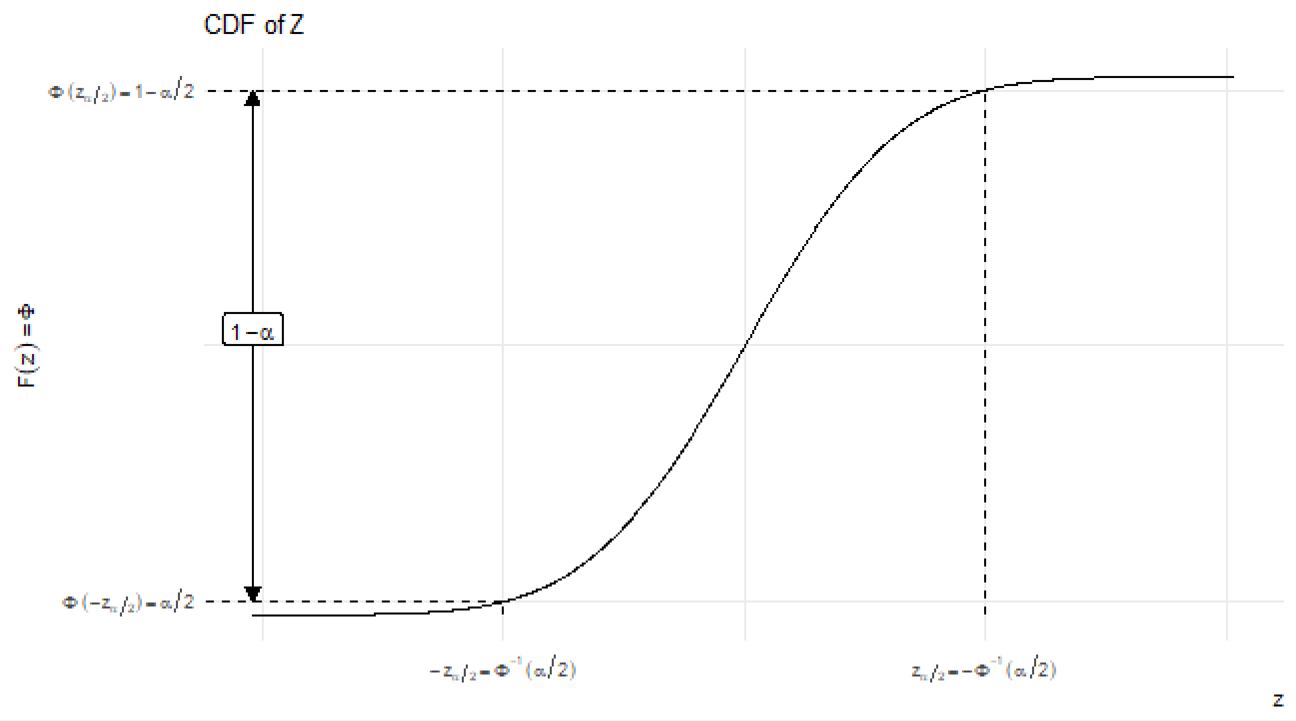
Confidence Intervals

- CLT to the rescue!
- Define our **confidence** with $1 - \alpha$ where $\alpha \in [0, 1]$



Confidence Intervals

- CLT to the rescue!
- Then use easy functions (or lookup tables in ye olden days) to get critical values!



Critical Values

- The functions to yield $z_{\alpha/2} = -\Phi^{-1}(\frac{\alpha}{2})$ and $-z_{\alpha/2} = \Phi^{-1}(\frac{\alpha}{2})$ are in `R`
- Say I want to calculate a 0.95 CI = $1 - \alpha$

```
qnorm(.025)
```

```
## [1] -1.959964
```

- Say I want to calculate a 0.90 CI = $1 - \alpha$

```
qnorm(.05)
```

```
## [1] -1.644854
```

- Say I want to calculate a 0.99 CI = $1 - \alpha$

```
qnorm(.005)
```

```
## [1] -2.575829
```

Calculating CIs

- So if we want to define the CI for an estimator $\hat{\theta}$ whose sampling distribution is normally distributed, we can write $[\hat{\theta} - 1.96\sigma_{\hat{\theta}},\ \hat{\theta} + 1.96\sigma_{\hat{\theta}}]$

- But what is $\sigma_{\hat{\theta}}$?

  - We know it is $\sqrt{VAR(\hat{\theta})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

- But then what is σ^2?

- We've been kicking this can down the road for a while now...using CLT to construct CIs for μ, but defined in terms of σ^2 -- the population standard variance

- Very unusual to know σ^2 in practice. Homeworks and exercises will tell you, but in practice we don't know

- So we need to approximate σ^2 using $S_U^2 \equiv \frac{\sum_i (Y_i - \bar{Y})^2}{n-1}$, our **unbiased** estimator for σ^2

Consistency

- But wait! Before we can plug in S_U, we need to prove it is both unbiased and **consistent**
- We already know how to prove unbiasedness
- Consistency: as the same size used to construct the estimator gets large, the probability of it being measured with error gets small
- Denote $\hat{\theta}_n$ as the estimate for a given sample size n
  - In the extreme: $\lim_{n \to \infty} P(|\hat{\theta} - \theta| > \epsilon) = 0$ where ϵ is any positive number
  - Can also express as " $\hat{\theta}_n$ converges in probability to θ ", or $\hat{\theta}_n \xrightarrow{p} \theta$
- In practice, we can evaluate this property by checking whether $VAR(\hat{\theta})$ approaches zero as n gets large (see pg. 450 for proof)
  - $\lim_{n \to \infty} VAR(\hat{\theta}) = 0$

Consistency

- Apply to $\bar{Y}$ for intuition

$$VAR(\bar{Y}) = \frac{\sigma^2}{n}$$

$$\lim_{n \to \infty} \frac{\sigma^2}{n} = 0$$

- Note that this **by itself** is insufficient to claim $\bar{Y} \xrightarrow{p} \mu$...we need to also prove unbiasedness (which we did last class)
- In other words, an estimator might be **consistent** but **biased**
- Or an estimator might be **unbiased** but not **consistent**
- Need to check both!

σ^2

- Remember what we're doing here!
  - We know that $U_n \equiv \frac{\bar{Y}-\mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(\mu, \sigma^2)$
  - But can we be sure that $\hat{\theta} \equiv \frac{\bar{Y}-\mu}{\sqrt{S_U^2/n}} \sim \mathcal{N}(\mu, \sigma^2)$?
- Note that, in the original setting, σ^2 is a **parameter** whereas in our sample setting S_U^2 is a **random variable**

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

σ^2

- So let's examine whether S_U^2 is a **consistent** estimator for σ^2

$$
\begin{aligned}
S_U^2 &= \frac{\sum_i (Y_i - \bar{Y})^2}{n-1} \\
&= \frac{1}{n-1}\left(\sum_i Y_i^2 + \sum_i \bar{Y}^2 - \sum_i 2Y_i\bar{Y}\right) \\
&= \frac{1}{n-1}\left((\sum_i Y_i^2) + n\bar{Y}^2 - 2n\bar{Y}^2\right) \\
&= \frac{1}{n-1}\left((\sum_i Y_i^2) - n\bar{Y}^2\right) \\
&= \frac{n}{n-1}\left(\frac{1}{n}\sum_i Y_i^2 - \bar{Y}^2\right)
\end{aligned}
$$

σ^2

- So let's examine whether S_U^2 is a **consistent** estimator for σ^2

$$S_U^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_i Y_i^2 - \bar{Y}^2\right)$$

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\sum_i Y_i^2 - \bar{Y}^2 &= \lim_{n\to\infty} \frac{1}{n}\sum_i Y_i^2 - \lim_{n\to\infty} \frac{1}{n}\sum_i \bar{Y}^2 \\
&= \mu_{Y^2} - \mu_Y^2 \\
&= E[Y^2] - \mu^2 \\
&= \sigma^2 \\
\text{So:}\ \ S_U^2 &= \frac{n}{n-1}(\sigma^2)
\end{aligned}$$

- But $\lim_{n\to\infty} \frac{n}{n-1} = 1$, meaning $S_U^2 \xrightarrow{p} \sigma^2$