

# Political Science 209 - Fall 2018

Uncertainty

---

Florian Hollenbach

18th November 2018

Goal: trying to estimate something unobservable from observable data

What we want to estimate: parameter  $\theta \rightsquigarrow$  unobservable

What you do observe: data

# Statistical Inference

Goal: trying to estimate something unobservable from observable data

What we want to estimate: parameter  $\theta \rightsquigarrow$  unobservable

What you do observe: data

We use data to compute an estimate of the parameter  $\hat{\theta}$

- **parameter**: the quantity that we are interested in

# Parameters and Estimators

- **parameter**: the quantity that we are interested in
- **estimator**: method to compute parameter of interest

Example:

- **parameter**: support for Jimbo Fisher in student population
- **estimator**: sample proportion of support as estimator

Example:

- **parameter**: average causal effect of aspirin on headache
- **estimator**: difference in mean between treatment and control

For the rest of the semester the question becomes:

How good is our estimator?



For the rest of the semester the question becomes:

How good is our estimator?

1. How close in expectation is the estimator to the truth?
2. How certain or uncertain are we about the estimate?

# Quality of estimators

How good is  $\hat{\theta}$  as an estimate of  $\theta$ ?

- Ideally, we want to know **estimation error**  $= \hat{\theta} - \theta_{truth}$

But we can never calculate this. Why?

# Quality of estimators

How good is  $\hat{\theta}$  as an estimate of  $\theta$ ?

- Ideally, we want to know **estimation error**  $= \hat{\theta} - \theta_{truth}$

But we can never calculate this. Why?

$\theta_{truth}$  is unknown

*If we knew what the truth was, we didn't need an estimate*

Instead, we consider two hypothetical scenarios:

1. How well would  $\hat{\theta}$  perform over *repeated data generating processes*? (**bias**)
2. How well would  $\hat{\theta}$  perform as the sample size goes to infinity? (**consistency**)

- Imagine the estimate being a random variable itself
- Drawing infinitely many samples of students asking about Jimbo

What is the average of the sample average? Or what is the expectation of the estimator?

$$\text{bias} = \mathbb{E}(\text{estimation error}) = \mathbb{E}(\text{estimate} - \text{truth}) = \mathbb{E}(\bar{X}) - p = p - p = 0$$

An unbiased estimator does not mean that it is always exactly correct!

An unbiased estimator does not mean that it is always exactly correct!

**To remember:** bias measures whether in expectation (on average) the estimator is giving us the truth

Essentially saying that the law of large numbers applies to the estimator, i.e.:

An estimator is said to be consistent if it converges to the parameter (truth) if  $N$  goes to  $\infty$

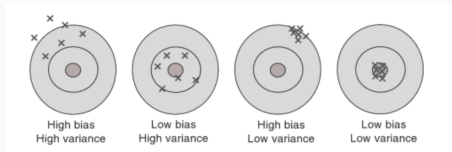


Next, we have to consider how certain we are about our results

Consider two estimators:

1. slightly *biased*, on average off by a bit, but always by the same margin
2. unbiased, but misses target left and right

# Variability



(Encyclopedia of Machine Learning)

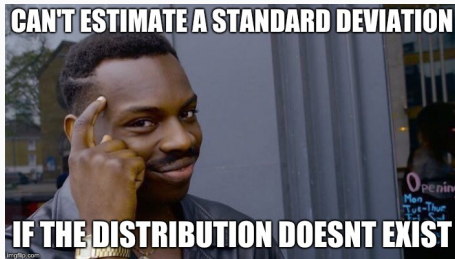
We characterize the variability of an estimator by using the standard deviation of the sampling distribution

How do we find that????

We characterize the variability of an estimator by using the standard deviation of the sampling distribution

How do we find that????

Remember, the sampling distribution is the distribution of our statistic over hypothetical infinitely many samples



# Standard Error

We estimate the standard deviation of the sampling distribution from the observed data

standard error

We estimate the standard deviation of the sampling distribution from the observed data

standard error

“*standard error* and describes the (estimated) average degree to which an estimator deviates from its expected value” (Imai 2017)

## Polling Example

Say we took a sample of 1500 students and asked whether they support Jimbo or not

Define a random variable  $X_i = 1$  if student  $i$  supports Jimbo,  
 $X_i = 0$  if not



## Polling Example

Say we took a sample of 1500 students and asked whether they support Jimbo or not

Define a random variable  $X_i = 1$  if student  $i$  supports Jimbo,  $X_i = 0$  if not

Binomial distribution with success probability  $p$  and size  $N$  where  $p$  is the proportion of *all students* who support Jimbo (population dist)

Estimator: ?

# Polling Example

Estimator:  $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$

## Polling Example

Estimator:  $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$

In earlier notation:  $\theta_{truth} = p$  and  $\theta = \bar{X}$

# Polling Example

Estimator:  $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$

1. LLN:  $\bar{X} \rightarrow p$  (consistent)
2. Expectation:  $\mathbb{E}(\bar{X}) = p$  (unbiased)
3. standard error?

## Polling Example - standard error

$X_i$  are i.i.d Bernoulli random variables with probability =  $p$

$$\mathbb{V}(\bar{X}) = \frac{1}{N^2} \mathbb{V}(\sum_{i=1}^N X_i) = \frac{1}{N^2} \sum_{i=1}^N \mathbb{V}(X_i)$$

## Polling Example - standard error

$X_i$  are i.i.d Bernoulli random variables with probability =  $p$

$$\mathbb{V}(\bar{X}) = \frac{1}{N^2} \mathbb{V}(\sum_{i=1}^N X_i) = \frac{1}{N^2} \sum_{i=1}^N \mathbb{V}(X_i) = \frac{N}{N^2} \mathbb{V}(X)$$

## Polling Example - standard error

$X_i$  are i.i.d Bernoulli random variables with probability =  $p$

$$\mathbb{V}(\bar{X}) = \frac{1}{N^2} \mathbb{V}(\sum_{i=1}^N X_i) = \frac{1}{N^2} \sum_{i=1}^N \mathbb{V}(X_i) = \frac{N}{N^2} \mathbb{V}(X) = \frac{p \times (1-p)}{N}$$



## Polling Example - standard error

$$\mathbb{V}(\bar{X}) = \frac{p \times (1-p)}{N}$$

Standard error:  $\sqrt{\mathbb{V}(\bar{X})}$

But we don't know  $p$ ! Now what?

## Polling Example - standard error

$$\mathbb{V}(\bar{X}) = \frac{p \times (1-p)}{N}$$

Standard error:  $\sqrt{\mathbb{V}(\bar{X})}$

But we don't know  $p$ ! Now what?

We use our unbiased estimate of  $p$ :  $\bar{X}$

## Polling Example - standard error estimate

$$\sqrt{\widehat{\mathbb{V}(\bar{X})}} = \sqrt{\frac{\bar{X}(1-\bar{X})}{N}}$$

## Polling Example - standard error estimate

Assume in our sample 55% of students support Jimbo:

$$SE = \sqrt{\widehat{\mathbb{V}(\bar{X})}} = \sqrt{\frac{0.55 \times (1-0.55)}{1500}} = \sqrt{\frac{0.55 \times (0.45)}{1500}} = 0.013$$

We can expect our estimate on average to be off by 1.3 percentage points

## Polling Example - standard error estimate

Assume in our sample 55% of students support Jimbo:

$$SE = \sqrt{\widehat{\mathbb{V}(\bar{X})}} = \sqrt{\frac{0.55 \times (1-0.55)}{1500}} = \sqrt{\frac{0.55 \times (0.45)}{1500}} = 0.013$$

We can expect our estimate on average to be off by 1.3 percentage points

If  $\bar{X} = 0.8$ , then  $SE = 0.010$

If  $N = 500$ ,  $\bar{X} = 0.55$ , then  $SE = 0.022$

# Standard error estimate

Standard error is based on variance of the sampling distribution

Gives estimate of uncertainty

Each estimator/statistic has unique sampling distribution, e.g.  
difference in means

Often we don't even know the sampling distribution of our estimators

How could we approximate it?

Often we don't even know the sampling distribution of our estimators

How could we approximate it?

Central limit theorem!



Central limit theorem says:

$$\bar{X} \approx N(\mathbb{E}(X), \frac{\mathbb{V}(X)}{N})$$

regardless of distribution of  $X$

We can use the approximation to the sampling distribution,

$\bar{X} \approx N(\mathbb{E}(X), \frac{\mathbb{V}(X)}{N})$  to construct **confidence intervals**

Confidence intervals give a range of values that is likely to contain the true value

We can use the approximation to the sampling distribution,  
 $\bar{X} \approx N(\mathbb{E}(X), \frac{\mathbb{V}(X)}{N})$  to construct **confidence intervals**

Confidence intervals give a range of values that is likely to contain the true value

To start, we select a probability value for our confidence level:  
usually 95%

The 95% confidence interval specifies the range of values in which the true parameter will fall for 95% of our hypothetical samples/experiments

The 95% confidence interval specifies the range of values in which the true parameter will fall for 95% of our hypothetical samples/experiments

Put differently “Over a hypothetically repeated data generating process, confidence intervals contain the true value of parameter with the probability specified by the confidence level” (Imai 2017)

$(1-\alpha)$  large sample Confidence interval is defined as:

$$CI(\alpha) = \bar{X} - z_{\frac{\alpha}{2}} \times SE, \bar{X} + z_{\frac{\alpha}{2}} \times SE$$

$z_{\frac{\alpha}{2}}$  is the critical value which equals  $(1 - \frac{\alpha}{2})$  quantile of the standard normal distribution

Where do the critical values come from?

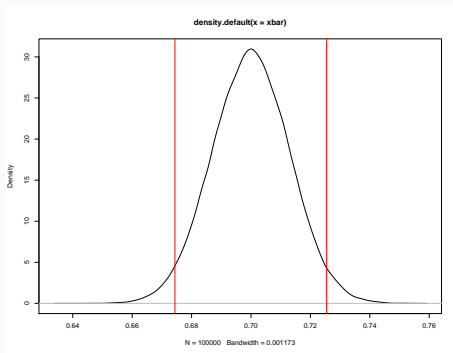
Where do the critical values come from?

Remember: Curve of the standard normal distribution:

- Symmetric around 0
- Total area under the curve is 100%
- Area between -1 and 1 is  $\sim 68\%$
- Area between -2 and 2 is  $\sim 95\%$
- Area between -3 and 3 is  $\sim 99.7\%$



# Confidence interval



Critical values are the exact values between which the standard normal distribution will include  $(1-\alpha)$  % of the area

# Confidence interval interpretation

Technically the CI is **not** the probability of the true parameter being between the two value.

# Confidence interval interpretation

Technically the CI is **not** the probability of the true parameter being between the two value.

Remember, in our view the true parameter is fixed

Instead: “95% confidence intervals contain the true value of the parameter 95% of the time during a hypothetically repeated data generating process” (Imai 2017)

# Confidence interval interpretation

Remember in the Jimbo example with  $\bar{X} = 0.55$  and  $N = 1500$

$$SE = \sqrt{\widehat{V(\bar{X})}} = \sqrt{\frac{0.55 \times (1 - 0.55)}{1500}} = \sqrt{\frac{0.55 \times (0.45)}{1500}} = 0.013$$

$$\text{CI}(\alpha) = \bar{X} - z_{\frac{\alpha}{2}} \times SE, \bar{X} + z_{\frac{\alpha}{2}} \times SE$$

$$CI(\alpha) = \bar{X} - z_{\frac{\alpha}{2}} \times SE, \bar{X} + z_{\frac{\alpha}{2}} \times SE$$

$$CI(0.05) = 0.55 - 1.96 \times 0.013, 0.55 + 1.96 \times 0.013 = 0.524, 0.576$$

What if we don't know the variance of the estimator?

Let's use the variance of the sample?

```
x <- rbinom(1500,1,0.7)
```

```
var <- var(x)/1500
```

```
SE <- sqrt(var)
```

$SE = 0.013$

# Confidence interval

```
xbar <- rep(NA, 10000)
for(i in 1:10000){
  x <- rbinom(1500,1,0.55)
  xbar[i] <-mean(x)
}
```

Write an R-script to test our confidence interval for Jimbo!



# Margin of Error in Surveys

- Margin of error is usually the difference from estimate to upper/lower 95\
- Margin of error:  $z_{0.025} \times \hat{SE} \approx z_{0.025} \times \sqrt{\frac{\bar{X} \times (1 - \bar{X})}{N}}$

# From Margin of Error to Sample Size

$$N \approx \frac{1.96 \times p \times (1-p)}{\text{margin of error}^2}$$

## In class Exercise

To isolate the causal effect of a criminal record for black and white applicants, Pager ran an audit experiment. In this type of experiment, researchers present two similar people that differ only according to one trait thought to be the source of discrimination.

To examine the role of a criminal record, Pager hired a pair of white men and a pair of black men and instructed them to apply for existing entry-level jobs in the city of Milwaukee. The men in each pair were matched on a number of dimensions, including physical appearance and self-presentation. As much as possible, the only difference between the two was that Pager randomly varied which individual in the pair would indicate to potential employers that he had a criminal record. Further, each week, the pair alternated which applicant would present himself as an ex-felon. To determine how incarceration and race influence employment chances, she compared callback rates among applicants with and without a criminal background and calculated how those callback rates varied by race.

## In class Exercise

Download data `criminalrecord.csv` from the class website and read into *R*

Summarize the data, what variables do you see?

# In class Exercise

Name	Description
jobid	Job ID number
callback	1 if tester received a callback, 0 if the tester did not receive a callback.
black	1 if the tester is black, 0 if the tester is white.
crimrec	1 if the tester has a criminal record, 0 if the tester does not.
interact	1 if tester interacted with employer during application, 0 if tester doesn't
city	1 if job is located in the city center, 0 if job is located in the suburbs.
distance	Job's average distance to downtown.
custserv	1 if job is in the customer service sector, 0 if it is not.
manualskill	1 if job requires manual skills, 0 if it does not.

# Question 1

How many observations are in the data? In how many cases is the tester black? In how many cases is he white?

## Question 2

Now we examine the central question of the study. Calculate the proportion of callbacks for white applicants with a criminal record, white applicants without a criminal record, black applicants with a criminal record, and black applicants without a criminal record.

## Question 3

Now consider the callback rate for white applicants with a criminal record. Construct a 95% confidence interval around this estimate. Also, construct a 99% confidence interval around this estimate.



## Question 4

Calculate the estimated effect of a criminal record for white applicants by comparing the callback rate in the treatment condition and the callback rate in the control condition. Create a 95% confidence interval around this estimate. Next, describe the estimate and confidence interval in a way that could be understood by a general audience.

## Question 5

Assuming a null hypothesis that there is no difference in callback rates between white people with a criminal record and white people without a criminal record, what is the probability that we would observe a difference as large or larger than the one that we observed in a sample of this size?