Political Science 209 - Fall 2018

Uncertainty

Florian Hollenbach

13th November 2018

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

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 $\hat{\theta}$

Parameters and Estimators

- 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

Parameters and Estimators

Example:

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

Parameters and Estimators

Example:

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

Quality of estimators

For the rest of the semester the question becomes:

How good is our estimator?

Quality of estimators

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 θ?

- 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 θ?

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

But we can never calculate this. Why?

θ_{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)

Bias

- 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?

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

Bias - Important

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

Bias - Important

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

Consistency

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 ∞

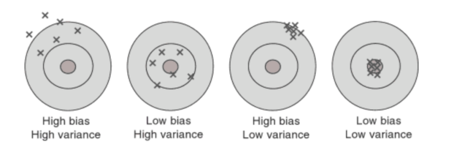
Variability

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)

Variability

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

How do we find that????

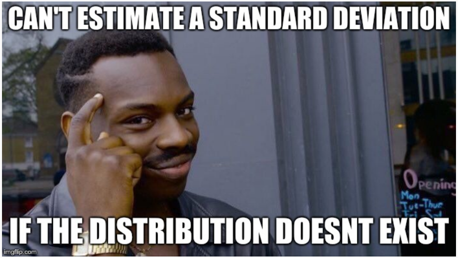
Variability

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

Variability



Standard Error

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

standard error

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 1000 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)

Polling Example

Estimator: ?

Polling Example

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

Polling Example

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

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

Polling Example

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

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

Polling Example - standard error

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

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

Polling Example - standard error

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

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

Polling Example - standard error

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

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

Polling Example - standard error

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

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

But we don't know p! Now what?

Polling Example - standard error

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

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

But we don't know p! Now what?

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

Polling Example - standard error estimate

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

Polling Example - standard error estimate

Assume in our sample 55% of students support Jimbo:

$$\text{SE} = \sqrt{\widehat{\mathbb{V}(\overline{X})}} = \sqrt{\frac{0.55 \times (1 - 0.55)}{1500}} = \sqrt{\frac{0.55 \times (0.45)}{1000}} = 0.016$$

We can expect our estimate on average to be off by 1.6 percentage points

Polling Example - standard error estimate

Assume in our sample 55% of students support Jimbo:

$$\text{SE} = \sqrt{\widehat{\mathbb{V}(\overline{X})}} = \sqrt{\frac{0.55 \times (1 - 0.55)}{1500}} = \sqrt{\frac{0.55 \times (0.45)}{1000}} = 0.016$$

We can expect our estimate on average to be off by 1.6 percentage points

If $\overline{X} = 0.8$, then SE $= 0.012$

If N $= 500$, $\overline{X} = 0.55$, 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

Confidence Intervals