

8. Sampling & Estimation

Spring 2021

Matthew Blackwell

Gov 2002 (Harvard)

Where are we? Where are we going?

- Last few weeks: probability, learning how to think about r.v.s
- Now: how to estimate features of underlying distributions with data.
- How do we construct estimators? What are their properties?

1/ Point Estimation

Motivating example

- Gerber, Green, and Larimer (APSR, 2008)

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	_____
9995 JENNIFER KAY SMITH		Voted	_____
9997 RICHARD B JACKSON		Voted	_____
9999 KATHY MARIE JACKSON		Voted	_____

Motivating Example

```
load("../assets/gerber_green_larimer.RData")  
## turn turnout variable into a numeric  
social$voted <- 1 * (social$voted == "Yes")  
neigh.mean <- mean(social$voted[social$treatment == "Neighbors"])  
neigh.mean
```

```
## [1] 0.378
```

```
contr.mean <- mean(social$voted[social$treatment == "Civic Duty"])  
contr.mean
```

```
## [1] 0.315
```

```
neigh.mean - contr.mean
```

```
## [1] 0.0634
```

- Is this a “real”? Is it big?

Why study estimators?

- **Goal 1: Inference**

- What is our best guess about some quantity of interest?
- What are a set of plausible values of the quantity of interest?

- **Goal 2: Compare estimators**

- In an experiment, use simple difference in sample means $(\bar{Y} - \bar{X})$?
- Or the **post-stratification estimator**, where we estimate the estimate the difference among two subsets of the data (male and female, for instance) and then take the weighted average of the two (\bar{Z} is the share of women):

$$(\bar{Y}_f - \bar{X}_f)\bar{Z} + (\bar{Y}_m - \bar{X}_m)(1 - \bar{Z})$$

- Which (if either) is better? How would we know?

Samples from the population

- Our focus: random vectors X_1, \dots, X_n are i.i.d. draws from c.d.f. F
 - e.g.: $X_i = 1$ if citizen i votes, $X_i = 0$ otherwise.
 - n is the **sample size**
 - i.i.d. can be justified through random sampling from a population.
 - F is often called the **population distribution** or just **population**
- Two metaphors:
 - Actual/potential population of size $N \gg n$ and we randomly sample n .
 - F represents the **data generating process**, we repeat n times
- **Statistical inference** or **learning** is using data to infer F .

Point estimation

- Goal of inference: learn about the features of the population.
- **Parameter:** θ is any function of the population distribution F
 - Also called: quantities of interest, estimands.
- Examples of parameters:
 - $\mu = \mathbb{E}[X_i]$: the mean (turnout rate in the population).
 - $\sigma^2 = \mathbb{V}[X_i]$: the variance.
 - $\mu_y - \mu_x = \mathbb{E}[Y_i] - \mathbb{E}[X_i]$: the difference in mean turnout between two groups.
- **Point estimation:** providing a single “best guess” about these parameters.

Estimators

- A **statistic** is any function of the sample $\{X_1, \dots, X_n\}$.
 - Before we see the data, statistics are random and have distributions, etc.
 - After we see the data, statistic is realized and we see the specific value.

Definition

An **estimator** $\hat{\theta}_n$ for some parameter θ , is a statistic intended as a guess about θ .

- $\hat{\theta}_n$ is a r.v. because it is a function of r.v.s.
 - $\rightsquigarrow \hat{\theta}_n$ has a distribution.
- An **estimate** is one particular realization of the estimator
 - Why is the following statement wrong: “My estimate was the sample mean and my estimator was 0.38”?

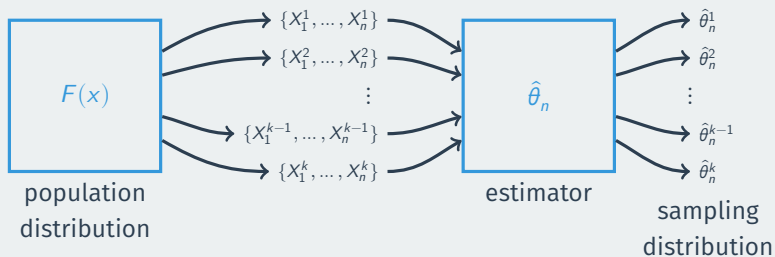
Examples of Estimators

- For the population expectation, $\mathbb{E}[X_i]$, many possible estimators:
 - $\hat{\theta}_n = \bar{X}_n$ the sample mean
 - $\hat{\theta}_n = X_1$ just use the first observation
 - $\hat{\theta}_n = \max(X_1, \dots, X_n)$
 - $\hat{\theta}_n = 3$ always guess 3

The three distributions

- **Population Distribution:** the data-generating process
 - Bernoulli in the case of the social pressure/voter turnout example)
- **Empirical distribution:** X_1, \dots, X_n
 - series of 1s and 0s in the sample
- **Sampling distribution:** distribution of the estimator over repeated samples from the population distribution
 - the 0.38 sample mean in the “Neighbors” group is one draw from this distribution

Sampling distribution, in pictures



Sampling distribution

```
## now we take the mean of one sample, which is  
## one draw from the **sampling distribution**  
my.samp <- rbinom(n = 10, size = 1, prob = 0.4)  
mean(my.samp)
```

```
## [1] 0.6
```

```
## let's take another draw from the population dist  
my.samp.2 <- rbinom(n = 10, size = 1, prob = 0.4)
```

```
## Let's feed this sample to the sample mean estimator  
## to get another estimate, which is another draw from  
## the sampling distribution  
mean(my.samp.2)
```

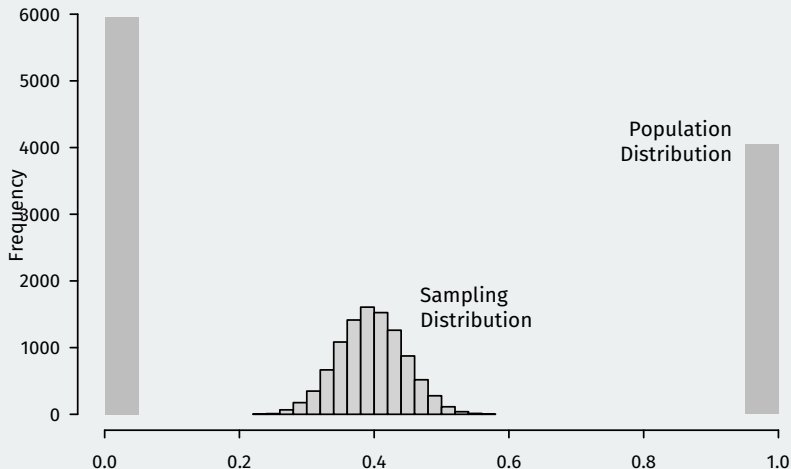
```
## [1] 0.3
```

Sampling distribution by simulation

- Let's generate 10,000 draws from the sampling distribution of the sample mean here when $n = 100$.

```
nsims <- 10000
mean.holder <- rep(NA, times = nsims)
for (i in 1:nsims) {
  my.samp <- rbinom(n = 100, size = 1, prob = 0.4)
  mean.holder[i] <- mean(my.samp) ## sample mean
  first.holder[i] <- my.samp[1] ## first obs
}
```

Sampling distribution versus population distribution



Question The sampling distribution refers to the distribution of θ , true or false.

Where do estimators come from?

- **Model-based inference:** assume $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} F$ and specify what family F is from.
 - Example: F is $\text{Pois}(\lambda)$.
 - Construct estimator $\hat{\lambda}$ using **maximum likelihood**
 - Downside: inferences are **model dependent**
- **Nonparametric inference:** make minimal assumptions on F .
- **Plug-in/analogy principle:** replace F with the empirical distribution.
 - Empirical distribution: probability $1/n$ at each observed value of X_i :

$$\hat{F}_n(x) = \frac{\sum_{i=1}^n \mathbb{I}(X_i \leq x)}{n}$$

- \rightsquigarrow if $\theta = \mathbb{E}[g(X)]$ replace \mathbb{E} sample means: $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n g(X_i)$

Plug-in estimators, examples

- Expectation:

$$\mu = \mathbb{E}[X_i] \rightsquigarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n$$

- Variance:

$$\sigma^2 = \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] \rightsquigarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

- Covariance:

$$\sigma_{xy} = \text{Cov}[X_i, Y_i] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_i - \mathbb{E}[Y_i])] \rightsquigarrow \hat{\sigma}_{xy} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

2/ Finite-Sample Properties of Estimators

Properties of estimators

- We only get one draw from the sampling distribution, $\hat{\theta}_n$.
- Want to use estimators whose distribution is “close” to the true value.
- There are two ways we evaluate estimators:
 - **Finite sample:** the properties of its sampling distribution for a fixed sample size n .
 - **Large sample:** the properties of the sampling distribution as we let $n \rightarrow \infty$.

- The **bias** of estimator $\hat{\theta}$ for parameter θ is

$$\text{bias}[\hat{\theta}] = \mathbb{E}[\hat{\theta}] - \theta.$$

- An estimator is **unbiased** if $\text{bias}[\hat{\theta}] = 0$.
- Sample mean of i.i.d. X_1, \dots, X_n with $\mathbb{E}[X_i] = \mu$

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

- Thus, \bar{X}_n is unbiased for μ if $\mathbb{E}[|X|] < \infty$
 - What about a weighted average?
- Unbiasedness is preserved under linear transformations.

Estimation variance

- **Sampling variance:** the variance of an estimator $\mathbb{V}[\hat{\theta}]$.
 - Measure of how spread the estimator it is around its mean.
- Sampling variance of the sample mean:

$$\mathbb{V}[\bar{X}_n] = \frac{1}{n^2} \sum_{i=1}^n \mathbb{V}[X_i] = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$$

- **Standard error:** standard deviation of the estimator $se(\hat{\theta}) = \sqrt{\mathbb{V}[\hat{\theta}]}$
 - Like all SDs, nice that it's on the same scale.
- Standard error of the sample mean: σ/\sqrt{n}

Mean squared error

- **Mean squared error** or **MSE** is

$$\text{MSE} = \mathbb{E}[(\hat{\theta}_n - \theta)^2]$$

- The MSE assesses the quality of an estimator.
 - How big are (squared) deviations from the true parameter?
 - Ideally, this would be as low as possible!
- Useful decomposition result:

$$\text{MSE} = \text{bias}[\hat{\theta}_n]^2 + \mathbb{V}[\hat{\theta}_n]$$

- \rightsquigarrow for unbiased estimators, MSE is the sampling variance.
- Might accept some bias for large reductions in variance for lower overall MSE.