Estimation

Introduction to Quantitative Social Science

Kosuke Imai
Harvard University / University of Tokyo

Summer 2022

What is Statistical Inference?

- Guessing what we do not observe from what we do observe
- What we want to estimate: parameter $\theta \rightsquigarrow$ unobservable
- What you do observe: data
- We use data to compute an estimate of the parameter $\hat{\theta}$

- How good is $\hat{\theta}$ as an estimate of θ?
- Ideally, we want to know estimation error $= \hat{\theta} - \theta_0$ where θ_0 is the true value of θ
- The problem: θ_0 is unknown
- Instead, we consider two hypothetical scenarios:
  1. How well would $\hat{\theta}$ perform as the sample size goes to infinity?
  2. How well would $\hat{\theta}$ perform over *repeated data generating process*?

Polling Disaster: 2016 Election

- All major preelection polls predicted Clinton's victory
- What happened?
  - FBI announcements
  - non-response bias
  - social desirability bias
  - failure to predict turnout
- We will look at polls closely in the precept this week

- For today, let's look at ABC News/Washington Post poll:
  - Nov. 3 – Nov. 6 (election was Nov. 8)
  - 2220 likely voters
  - Live phone
  - Clinton (47%), Trump (43%), Johnson (4%)
  - Margin of error: ± 2.5 percentage points
- Actual election result (national vote): Clinton (48%), Trump (47%)

Estimating Trump's Support

- Parameter θ: population proportion of likely voters who support Trump
- Estimator $\hat{\theta}$: sample proportion of respondents who support Trump
- How good is $\hat{\theta}$ as an estimate of θ?

- Assume a simple random sampling of n voters: $n = 2220$
- Define a random variable $X_i = 1$ if the ith respondent supports Trump and $X_i = 0$ otherwise for each $i = 1, 2, \ldots, n$
- Data generating process: Binomial distribution with success probability p and size n where p is the population proportion of likely voters who support Trump
- Estimator: $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- That is, $\theta = p$ and $\hat{\theta} = \overline{X}$

1. How well would $\overline{X}$ behave as the sample size increases?
   - Law of large numbers: $\overline{X} \longrightarrow p$
   - consistency
   - But, how large is large enough?
2. How would $\overline{X}$ behave over repeated data generating process?
   - hypothetical scenario: repeatedly conduct a survey under the exact same conditions many times
   - expectation = average performance: $\mathbb{E}(\overline{X}) = p$
   - unbiasedness
   - sampling distribution of $\overline{X}$: Binomial random variable divided by n
   - standard deviation of sampling distribution:

$$\sqrt{\mathbb{V}(\overline{X})} = \sqrt{\frac{p(1-p)}{n}}$$

   - standard error = estimated standard deviation

$$\sqrt{\widehat{\mathbb{V}(\overline{X})}} = \sqrt{\frac{\overline{X}(1-\overline{X})}{n}} = \sqrt{\frac{0.43 \times (1-0.43)}{2200}} \approx 0.011$$

Confidence Intervals

- Beyond standard error: characterizing the whole sampling distribution
- Central limit theorem: for a sufficiently large sample size,

$$\overline{X} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mathbb{E}(X),\ \frac{\mathbb{V}(X)}{n}\right)$$

- In the current case:

$$\overline{X} \overset{\text{approx.}}{\sim} \mathcal{N}\left(p,\ \frac{p(1-p)}{n}\right)$$

- Choose the level of confidence interval: e.g., 95%
- Compute the confidence interval, which contains the true value, e.g., 95% of time over repeated data generating process

- $(1-\alpha) \times 100\%$ (asymptotic) confidence intervals:

$$\mathrm{CI}_\alpha = [\overline{X} - z_{\alpha/2} \times \text{standard error}, \ \overline{X} + z_{\alpha/2} \times \text{standard error}]$$

  where $z_{\alpha/2}$ is called the critical value

- $P(Z > z_{\alpha/2}) = \alpha/2$ and $Z \sim \mathcal{N}(0,1)$
  1. $\alpha = 0.01$ gives $z_{\alpha/2} = 2.58$
  2. $\alpha = 0.05$ gives $z_{\alpha/2} = 1.96$
  3. $\alpha = 0.10$ gives $z_{\alpha/2} = 1.64$

- Be careful about the interpretation!!
  1. Probability that the true value is in a *particular* confidence interval is either 0 or 1
  2. Confidence intervals are *random*, while the truth is *fixed*

- CIs for ABC/WP poll:

$$90\%\mathrm{CI}: [0.43 - 1.64 \times 0.011, \ 0.43 + 1.64 \times 0.011] = [0.412, \ 0.447]$$

$$95\%\mathrm{CI}: [0.43 - 1.96 \times 0.011, \ 0.43 + 1.96 \times 0.011] = [0.409, \ 0.451]$$

$$99\%\mathrm{CI}: [0.43 - 2.58 \times 0.011, \ 0.43 + 2.58 \times 0.011] = [0.402, \ 0.457]$$

Summary: Inference with Random Sampling

- Random sampling from a large population
- Sample analogue principle: use sample mean to infer population mean
- Asymptotic inference:
  1. Law of large Numbers:
  $$\overline{X} \rightsquigarrow \mathbb{E}(X)$$
  2. Central Limit Theorem:
  $$\overline{X} \overset{\text{approx.}}{\sim} \mathcal{N}\left(\mathbb{E}(X), \frac{\mathbb{V}(X)}{n}\right)$$
- Standard error: $\sqrt{\frac{\widehat{\mathbb{V}(X)}}{n}}$
- Confidence interval:

$$[\overline{X} - z_{\alpha/2} \times \text{standard error}, \ \overline{X} + z_{\alpha/2} \times \text{standard error}]$$

Comparison of Two Samples

- Comparison of two groups is more interesting
- Public opinion differences across groups
- Difference between treatment and control groups in experiments

- Causal inference with randomized experiments
- Back to the GOTV example
- The 2006 Michigan August primary experiment
- Treatment Group: postcards showing their own and their neighbors' voting records
- Control Group: received nothing

Social Pressure Experiment Revisited

- Turnout rate: $\overline{X}_T = 0.37$, $\overline{X}_C = 0.30$,
- Sample size: $n_T = 360$, $n_C = 1890$
- Estimated average treatment effect:

$$\widehat{\text{ATE}} \;=\; \overline{X}_T - \overline{X}_C \;=\; 0.07$$

- Standard error:

$$\sqrt{\frac{\overline{X}_T(1-\overline{X}_T)}{n_T} + \frac{\overline{X}_C(1-\overline{X}_C)}{n_C}} \;=\; 0.028$$

- 95% Confidence intervals based on CLT:

$$\begin{aligned} & [\widehat{\text{ATE}} - \text{standard error} \times z_{0.025},\; \widehat{\text{ATE}} + \text{standard error} \times z_{0.025}] \\ = \;\; & [0.016,\; 0.124] \end{aligned}$$

Minimum Wage Study Revisited

- Three identification strategies
  1. Cross-section comparison
  2. Before-and-after comparison
  3. Difference-in-differences
- How should we calculate the standard error under each strategy?
- What about confidence intervals?

```
minwage <- read.csv("data/minwage.csv")
## proportion of those fully employed before and after
## the increase in the minimum wage
minwage$fullPropBefore <- minwage$fullBefore /
    (minwage$fullBefore + minwage$partBefore)
minwage$fullPropAfter <- minwage$fullAfter /
    (minwage$fullAfter + minwage$partAfter)
## separate NJ and PA
minwageNJ <- subset(minwage, subset = (location != "PA"))
minwagePA <- subset(minwage, subset = (location == "PA"))
```

Cross-section Comparison: Assume no confounder

- Estimate: $\widehat{\text{ATE}} = \overline{X}_{\text{NJ}} - \overline{X}_{\text{PA}}$

```
est <- mean(minwageNJ$fullPropAfter) -
    mean(minwagePA$fullPropAfter)
est

## [1] 0.0481
```

- Standard error:

$$\sqrt{\frac{\widehat{\mathbb{V}(X_{\text{NJ}})}}{n_{\text{NJ}}} + \frac{\widehat{\mathbb{V}(X_{\text{PA}})}}{n_{\text{PA}}}}$$

```
nNJ <- nrow(minwageNJ)
nPA <- nrow(minwagePA)
se <- sqrt(var(minwageNJ$fullPropAfter) / nNJ +
           var(minwagePA$fullPropAfter) / nPA)
se

## [1] 0.0336
```

- Confidence intervals based on CLT:

$$[\widehat{\text{ATE}} - \text{standard error} \times z_{\alpha/2},\ \widehat{\text{ATE}} + \text{standard error} \times z_{\alpha/2}]$$

```
## 90%
c(est - se * qnorm(0.95), est + se * qnorm(0.95))
## [1] -0.00715  0.10338
## 95%
c(est - se * qnorm(0.975), est + se * qnorm(0.975))
## [1] -0.0177  0.1140
## 99%
c(est - se * qnorm(0.995), est + se * qnorm(0.995))
## [1] -0.0384  0.1347
```

- Conservative inference based on Student's t-distribution is possible
- Comparison of two sample means from Normal distributions
- No exact distribution exists $\rightsquigarrow$ approximation

```
t.test(minwageNJ$fullPropAfter, minwagePA$fullPropAfter)

##
##  Welch Two Sample t-test
##
## data:  minwageNJ$fullPropAfter and minwagePA$fullPropAfter
## t = 1, df = 100, p-value = 0.2
## alternative hypothesis: true difference in means is not equa
## 95 percent confidence interval:
##  -0.0185  0.1148
## sample estimates:
## mean of x mean of y
##     0.320     0.272
```

Before-and-After Comparison

- Assumption: only change is the treatment
- Estimate: $\widehat{\text{ATE}} = \overline{X}_{\text{NJ,after}} - \overline{X}_{\text{NJ,before}}$

```
est <- mean(minwageNJ$fullPropAfter) -
    mean(minwageNJ$fullPropBefore)
est
## [1] 0.0239
```

- Standard error: $\sqrt{\widehat{\mathbb{V}(\widehat{\text{ATE}})}}$
- Variance of the sum of random variables:

$$
\begin{aligned}
\mathbb{V}(X+Y) &= \mathbb{V}(X) + \mathbb{V}(Y) + 2\text{Cov}(X,Y) \\
\mathbb{V}(aX+bY) &= a^2\mathbb{V}(X) + b^2\mathbb{V}(Y) + 2ab\text{Cov}(X,Y)
\end{aligned}
$$

- Variance of $\widehat{\text{ATE}}$:

$$\begin{aligned}\mathbb{V}(\widehat{\text{ATE}}) &= \mathbb{V}(\overline{X}_{\text{NJ,after}}) + \mathbb{V}(\overline{X}_{\text{NJ,before}}) - 2\text{Cov}(\overline{X}_{\text{NJ,after}}, \overline{X}_{\text{NJ,before}}) \\ &= \frac{\mathbb{V}(X_{\text{NJ,after}})}{n_{\text{NJ}}} + \frac{\mathbb{V}(X_{\text{NJ,before}})}{n_{\text{NJ}}} - \frac{2\text{Cov}(X_{\text{NJ,after}}, X_{\text{NJ,before}})}{n_{\text{NJ}}}\end{aligned}$$

- Standard error:

```
se <- sqrt((var(minwageNJ$fullPropAfter) +
            var(minwageNJ$fullPropBefore) -
            2 * cov(minwageNJ$fullPropAfter,
                    minwageNJ$fullPropBefore)) / nNJ)
se
## [1] 0.0176
```

- 95% confidence interval:

```
c(est - se * qnorm(0.975), est + se * qnorm(0.975))
## [1] -0.0107  0.0585
```

Difference-in-Differences

- Assumption: parallel trend assumption
- Estimate:

$$\widehat{\text{ATE}} = (\overline{X}_{\text{NJ,after}} - \overline{X}_{\text{NJ,before}}) - (\overline{X}_{\text{PA,after}} - \overline{X}_{\text{PA,before}})$$

```
est <- (mean(minwageNJ$fullPropAfter) -
          mean(minwageNJ$fullPropBefore)) -
      (mean(minwagePA$fullPropAfter) -
       mean(minwagePA$fullPropBefore))
est
## [1] 0.0616
```

- Variance:

$$\mathbb{V}(\widehat{\text{ATE}}) \;=\; \mathbb{V}(\overline{X}_{\text{NJ,after}} - \overline{X}_{\text{NJ,before}}) + \mathbb{V}(\overline{X}_{\text{PA,after}} - \overline{X}_{\text{PA,before}})$$

- Standard error:

```
se <- sqrt(var(minwageNJ$fullPropAfter) / nNJ +
           var(minwageNJ$fullPropBefore) / nNJ -
           2 * cov(minwageNJ$fullPropAfter,
                   minwageNJ$fullPropBefore) / nNJ +
          var(minwagePA$fullPropAfter) / nPA +
           var(minwagePA$fullPropBefore) / nPA -
           2 * cov(minwagePA$fullPropAfter,
                   minwagePA$fullPropBefore) / nPA)
se
## [1] 0.0455
```

- 95% confidence interval:

```
c(est - se * qnorm(0.975), est + se * qnorm(0.975))
## [1] -0.0276  0.1508
```