Analysis of Experiments

February 25

Outline

1. Statistical conclusion validity (briefly)
2. Experimental analysis
3. Analysis-relevant practical considerations
4. Preview of next week

Threats to statistical conclusion validity

1. Power
2. Statistical assumption violations
3. Fishing
4. Measurement error
5. Restriction of range
6. Protocol violations
7. Loss of control
8. Unit heterogeneity (on DV)
9. Statistical artefacts

SSC Table 2.2 (p.45)

Measurement and operationalization

- Content validity: does it include everything it is supposed to measure
- Construct validity: does the instrument actually measure the particular dimension of interest
- Predictive validity: does it predict what it is supposed to
- Face validity: does it make sense

How do we know we manipulated what we thought we did?

- Before the study, the best way to figure out whether a measure or a treatment serves its intended purpose is to pretest it before implementing the full study
- During the study, the best way to figure out if our manipulation worked is to do manipulation checks

Outline

1. Statistical conclusion validity (briefly)
2. Experimental analysis
3. Analysis-relevant practical considerations
4. Preview of next week

Experimental inference

- How do we know if we have a statistically detectable effect?
- How do we draw inferences about effects?
- We have a SATE estimate, what does that tell us about PATE?

Estimators and inference

- Nonparametric inference: Build a randomization (permutation) distribution
- Parametric inference: Assume a sampling distribution

"Perfect Doctor"

True potential outcomes

Unit	Y(0)	Y(1)
1	13	14
2	6	0
3	4	1
4	5	2
5	6	3
6	6	1
7	8	10
8	8	9
Mean	7	5

"Perfect Doctor"

An observational study or one realization of randomization

Unit	Y(0)	Y(1)
1	?	14
2	6	?
3	4	?
4	5	?
5	6	?
6	6	?
7	?	10
8	?	9
Mean	5.4	11

Randomization

What are all of the possible treatment effect estimates we can get from our "Perfect Doctor" data?

```
# theoretical randomizations
d <- data.frame(
    y1 = c(14,0,1,2,3,1,10,9),
    y0 = c(13,6,4,5,6,6,8,8) )
onedraw <- function(eff=FALSE){
    r <- replicate(nrow(d), sample(1:2,1))
    tmp <- d
    tmp[cbind(1:nrow(d),r)] <- NA
    if(eff) {
        return(mean(tmp[,'y1'], na.rm=TRUE) -
               mean(tmp[,'y0'], na.rm=TRUE))
    } else
        return(tmp)
}

onedraw() # one randomization

onedraw(TRUE) # one effect estimate

# simulate 2000 experiments from these data
x1 <- replicate(2000, onedraw(TRUE))
hist(x1, col=rgb(1,0,0,.5), border='white')

# where is the true effect
abline(v=-2, lwd=3, col='red')
```

Randomization inference

Once we have our experimental data, let's test the following null hypothesis:

H_0: Y is independent of treatment assignment

If we swapped the treatment assignment labels on our data (ignoring the actual randomization) in every possible combination to build a distribution of treatment effects observable due to chance, would the treatment effect estimate be likely or unlikely?

```
# compare to an empirical randomization distribution
experiment <- onedraw()
effest <- mean(experiment[,'y1'], na.rm=TRUE) -
          mean(experiment[,'y0'], na.rm=TRUE)

w <- apply(experiment, 1, function(z) which(!is.na(z)))
yobs <- experiment[cbind(1:nrow(experiment), w)]

random <- function() {
    tmp <- sample(1:8, sum(!is.na(experiment[,'y1'])), FALSE)
    mean(yobs[tmp]) - mean(yobs[-tmp])
}

# build a randomization distribution from our data
x2 <- replicate(2000, onedraw(TRUE))
hist(x2, col=rgb(0,0,1,.5), border='white', add=TRUE)

abline(v=-2, lwd=3, col='red') # true effect
abline(v=effest, lwd=3, col='blue') # estimate in our `experiment`

# empirical quantiles
quantile(x2[is.finite(x2)], c(0.025, 0.975))
# compare to actual quantiles
quantile(x1[is.finite(x1)], c(0.025, 0.975))
```

Comparison to t-test

```
# two-tailed
t.test(yobs ~ w)
sum(abs(x1[is.finite(x1)]) > effest)/2000

# one-tailed (greater)
t.test(yobs ~ w, alternative='greater')
sum(x1[is.finite(x1)] > effest)/2000
```

Effects and Uncertainty

- The estimator for the SATE is the mean-difference
- The variance of this estimate is influenced by:

  1. Sample size
  2. Variance of Y
  3. Relative treatment group sizes

- We generally assume constant individual treatment effects

Formula for SE

$$\widehat{SE}_{SATE} = \sqrt{\frac{\widehat{Var}(Y_0)}{N_0} + \frac{\widehat{Var}(Y_1)}{N_1}}$$

where

$\widehat{Var}(Y_0)$ is control group variance

and

$\widehat{Var}(Y_1)$ is treatment group variance

Estimators and inference

- Difference of means (or proportions)
    - Randomization distribution
    - t-test
- ANOVA
- Regression

Protocol

1. Plan for data collection
2. Plan for analyses
3. Plan for sample size

Practical analytic advice

1. Power analysis to determine sample size
2. Don't observe outcomes until analysis plan is settled
3. If we need to use covariates:
    - Plan for their use in advance
    - Block on them, if possible
    - Measure them well
4. Balance
    - This is controversial

Mostly from Rubin (2008)

Moderation

If we have an hypothesis about moderation, what can we do?

- Best solution: manipulate the moderator
- Next best: block on the moderator and stratify our analysis
    - Estimate Conditional Average Treatment Effects
- Least best: include a treatment-by-covariate interaction in our regression model

Mediation

If we have hypotheses about mediation, what can we do?

- Best solution: manipulate the mediator
- Next best: manipulate the mediator for some, observe for others
- Least best: observe the mediator

Experimental Power

Simple definition:

> "The probability of not making a Type II error", or "Probability of a true positive"

Formal definition:

> "The probability of rejecting the null hypothesis when a causal effect exists"

Type I and Type II Errors

	H_0 True	H_0 False
Reject H_0	Type 1 Error	True positive
Accept H_0	False negative	Type II error

True positive rate is power

False negative rate is the significance threshold, typically $\alpha = .05$

Experimental Power

What impacts power?

- As n increases, power increases
- As the true effect size increases, power increases (holding n constant)
- As $Var(Y)$ increases, power decreases
- Conventionally, 0.80 is a reasonable power level

Doing a power analysis I

Power is calculated using:

1. Treatment group mean outcomes
2. Sample size
3. Outcome variance
4. Statistical significance threshold
5. A sampling distribution

Doing a power analysis II

$$Power = \phi\left(\frac{|\mu_1 - \mu_0|\sqrt{N}}{2\sigma} - \phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right)$$

where

- μ: treatment group mean
- N: total sample size
- σ: outcome standard deviation
- α: statistical significance level
- ϕ: Normal distribution function

Minimum Detectable Effect

- Power is a difficult thing to understand
- We can instead think about what is the smallest effect we could detect given:

    1. Treatment group sizes
    2. Expected correlation between treatment and outcome
    3. Our uncertainty about the effect size
    4. Intended power of our experiment

- Sometimes non-zero effects are not detectable

Minimum Detectable Effect

"Backwards power analysis"

```
num <- (1-cor(w, yobs)^2)
den <- prod(prop.table(table(w))) * 8

# use our observed effect SE
se_effect <- summary(lm(yobs ~ w))$coef[2,2]

sigma <- sqrt((se_effect * num)/den)
sigma
sigma * 2.49 # one-sided, 80%, .05
sigma * 2.80 # two-sided, 80%, .05

# vary our guess at the effect SE
sqrt(( seq(0,3,by=.25) * num)/den) * 2.8
```

Effect sizes

- We rarely care only about statistical significance
- We want to know if effects are large or small
- We want to compare effects across studies

Effect sizes

In two-group experiments, we can use the standardized mean difference as an effect size

Two names: Cohen's d or Hedge's g

Basically the same:

$d = \frac{\bar{x}_1 - \bar{x}_0}{s}$, where

$s = \sqrt{\frac{(n_1-1)s_1^2+(n_0-1)s_0^2}{n_1+n_0-2}}$

Effect sizes

Cohen gave "rule of thumb" labels to different effect sizes:

- Small: ~0.2
- Medium: ~0.5
- Large: ~0.8

Outline

1. Statistical conclusion validity (briefly)
2. Experimental analysis
3. Analysis-relevant practical considerations
4. Preview of next week

Broken experiments

- Attrition
- Noncompliance
  - One-sided (failure to treat)
  - One-sided (control group gets treated)
  - Cross-over
- Missing data

Analysis of data with attrition

Considerations:

- Symmetric, possibly random, attrition
- One-sided or systematic attrition
- Pre-treatment/post-treatment
- Pre-measurement/post-measurement

Noncompliance analysis

Choices:

1. Intention to treat analysis
2. As-treated analysis
3. Exclude noncompliant cases
4. Estimate a Local Average Treatment Effect (LATE)
    - aka Compliance Average Treatment Effect (CATE)

One-sided noncompliance

$ITT = \overline{Y}_1 - \overline{Y}_0$

$LATE = \frac{ITT}{Pct.Compliant}$

We need to observe compliance to estimate the LATE

Two-sided noncompliance

1. This is more complex analytically
2. Stronger assumptions are required to analyze it
    - Especially monotonicity
    - e.g., no one who who go to the library if not encouraged but who won't go to the library if encouraged
3. This is a classic design trumps analysis problem

Missing Data

Problems:

- Missing data is a threat to representativeness
- Missing data increases our uncertainty

Solutions:

- Case deletion
- Imputation

Cluster random assignment

- Cluster randomization is fine if cluster means are similar
- Otherwise, clustering introduces inefficiencies
- Or we can change our unit of analysis
    - Contrast people as units versus clusters as units

Outline

1. Statistical conclusion validity (briefly)
2. Experimental analysis
3. Analysis-relevant practical considerations
4. Preview of next week

Next week

- Continue our conversation about ethics
    - Read: The Belmont Report
- Discuss practical issues about implementation
- For Shadish, Cook, and Campbell, when reading Ch.14 focus on pp.488--504 (2nd half of chapter)