The Generalizer:
A free tool to help you design sample recruitment plans

Elizabeth Tipton

Associate Professor of Statistics

Co-Director, Statistics for Evidence-Based Policy and Practice (STEPP) Center

Faculty Fellow, Institute for Policy Research

Northwestern University

Keynote at University of Nebraska-Lincoln, November 2021

Relevance

Today I’m going to talk about methods for improving the **relevance** of large-scale experiments.

Questions of **relevance** are inherently questions about **generalizability**.

Improved generalizability = improved relevance.

Assumptions

In this workshop, I’m focusing only on issues related to generalizability and **external validity**.

I’m assuming that the study has high **internal validity**, for example, through random assignment of the intervention.

Overview

Three approaches to increasing relevance:

1. Design – build a better sample

2. Assessment – Help others understand where the results may apply and where they may not

3. Estimation – Provide others with estimates of the average treatment impact for their population of interest.

Schedule

9:00-1:15	Introductions
9:15-9:45	Session 1: Background
9:45-10:30	Session 2: Planning an experiment
10:30-10:45	Break
10:45-11:30	Session 3: Assessing similarity
11:30-12:00	Wrap up

But first…

In order to make the labs move along easily, I need you to:

Go to: www.thegeneralizer.org

1. You will need to enter your email address.

2. Check your email – you will get a code (very soon).

3. Save this email for now. We will use it in the lab. (Please wait!)

Session 1: Background

How do we know if results generalize?

We all have informal ideas about generalization.

For example, we might think that results generalize if:

- The sample was selected randomly; or
- The treatment effect is constant (no interactions); or
- The contexts are similar.

Representative samples

Kruskal and Mosteller (1979) wrote a great series of papers on how the term *representative sample* is used.

They found 9 uses in the statistical literature.

The three most relevant here are:

1. **Typical** or ideal cases
2. **Coverage** of the population
3. **Miniature** of the population

Typical cases

One type of representative sample is one composed of typical cases. These are units closest to the modal type.

This kind of sample is homogenous on many covariates. It includes only "average" units.

Informal concept of generalization:

Results generalize if they are found to work in the typical unit.

Coverage

Another type of representative sample is one composed of each type of unit in the population, regardless of frequency.

For example, a study may aim to include at least one African-American, White, Asian, Hispanic, and Native American student.

Informal concept of generalization:

> *Results generalize if they are still found even with such large diversity in the sample.*

Miniature

The third type is a sample that is a miniature of the population. Like coverage, this involves including the same diversity found in the population.

In addition, however, the composition of units should be the same in both the sample and population. That is, the **frequency** of each type is the same.

Informal concept of generalization:

> *Results generalize to a population if they are found to work in sample that is "like" it.*

Miniature

The Generalizer focuses on this third type for an important reason:

Experiments estimate average treatment effects.

This begs the question: average treatment effect *for whom*?

For research to be relevant, we need this to be a policy relevant population.

Probability sampling

When the goal is for the sample to be a miniature of the population, the most obvious approach is to use **probability sampling**.

Probability sampling is great when you can do it.

However,

- It is very rarely done.
- Often there is high non-response.
- Even in big congressionally mandated studies, you can get funny probability samples.

Probability sampling

For the rest of this workshop, I’m going to assume that random sampling is infeasible.

(The methods implemented in The Generalizer can also be implemented with random sampling, however.)

Generalization

In practice, samples are rarely selected from well-defined populations. Instead, generalizations are made **post-hoc**.

For example, a policy maker visits the *What Works Clearinghouse* or reads an article and must decide whether the results of a study are relevant to their policy relevant inference population.

They want to know: Will this work in my state? For Title I schools? For schools like mine?

The problem

Policy makers have **few tools** and **limited information** to make generalizations well.

They likely *don’t know* the composition of the study population (and maybe the inference population too) if the study is not explicit.

They *don’t know* how to choose what compositional variables might matter. At best, a policy maker can judge coverage, but not frequency or joint frequency.

The role of the evaluator

Unlike policy makers, those conducting large-scale evaluations:

- Have a rich understanding of the treatment;
- Have a sense of the conditions under which it may work best;
- Understand features of the study;
- Have the relevant data to make comparisons.

For these reasons, I argue that researchers should lead the conversation about generalization.

Key concepts

Let’s return to this idea of the “miniature”.

There are three important concepts:

- Inference population
- Treatment effect heterogeneity
- Ignorability assumption

Inference population

The goal of a study is rarely to estimate the ATE for the *sample only*.

Typically, we think this sample represents *some* population.

Instead of being vague, this approach requires:

- A clear definition of an inference population;
- Inclusion/exclusion criteria; and
- The enumeration of units in the population.

Which population?

This population may be broad or may be narrow.

There is often a loose coupling of the intervention and population, though there may be many possible populations for the same intervention.

The inference population selected is often a combination of the ideal (which may be broad) and what is feasible (which may be more narrow).

For example

You may wish your study findings could generalize to *all elementary schools in the United States*.

But:

- You can only select sites in one state; or
- You want to only include schools in your nearby district; or
- The schools most eager to use your program are those that can’t afford to purchase it themselves.

The result: your generalizations may be more narrow.

Treatment effect heterogeneity

If treatment impacts are constant, the results of a study generalize *everywhere*.

The sample thus only needs to be a “miniature” of the population if treatment effects *vary*.

If they do, then the ATE estimated in the sample may be biased for the population.

For example

Imagine an intervention:

- works extremely well for kids who are left handed (+ 90),
- terribly for kids who are right handed (-10).

In the population, the ATE would be 90(.10) – 10(.90) = 0.

Imagine a study only including right-handers.

The study ATE would be -10. This estimate is *biased* for the population ATE.

Which covariates?

The approach developed today focuses on this idea of the “miniature.”

But “miniature” in terms of what features?

In generalization, the focus is on covariates that explain variation in treatment impacts. (For example: handedness).

How do we choose?

Unfortunately, there is not much empirical information on treatment effect variation.

This is because studies are rarely powered for detection of moderator impacts.

We thus need to select variables based on:

- Prior research (where available);
- Logic model and theory of change;
- Policy relevant subgroups; and
- Which data is actually available in the population.

Sampling ignorability

If a sample is a “miniature” of the population *all* the covariates that explain variation in treatment impacts, then the ATE estimated in the sample is *unbiased* for the ATE in the population.

This means that we need to be exhaustive in selecting variables.

It also means that we might miss some. If we do, then the estimate could still be biased (though likely *less* biased).

It is important to state the assumptions – the set of covariates included – in all statements regarding generalizability.

Goals

In Session 2:

- To develop a sample selection plan so that the achieved sample is a *miniature* of a well-defined inference population.

In Session 3:

- To *assess* how similar an achieved sample is to various inference populations. Or, put another way, is the sample a miniature of the population?

Session 2: Planning for Generalization

Basic premise

Let’s assume you are planning an intervention study.

For example, the study might be cluster randomized, in which case you’ll probably need roughly 30 – 50 schools.

Now you need to think about:

- Recruiting schools;
- Where your results will generalize.

What should/can you do?

Sample selection for generalization

Random sampling – while statistically ideal – is very rare in large-scale evaluations.

Sample selection for generalization

Random sampling – while statistically ideal – is very rare in large-scale evaluations.

Convenience sampling – wherein researchers begin with the schools or sites they know best or have previous experience with and work out from there – is much more common.

Sample selection for generalization

Random sampling – while statistically ideal – is very rare in large-scale evaluations.

Purposive sampling – what I'm going to talk about now – offers a middle ground option. Like random sampling, this starts with a well-defined population.

Convenience sampling – wherein researchers begin with the schools or sites they know best or have previous experience with and work out from there – is much more common.

Purposive sampling

The goal is to develop a recruitment strategy so that the final sample is a *miniature* of the *inference population* to whom you'd like to generalize.

Purposive sampling asks, before the experiment begins:

- What **inference population** is appropriate for this study?
- What **covariates** do we think matter? (Those that explain variation in treatment impacts).
- **How** can we select our sample so that it is, in fact, similar to this inference population?

The Generalizer

Will guide you through:

- Defining inclusion/exclusion criteria for the inference population;
- Selecting covariates that might explain variation in treatment impacts;
- Developing a recruitment plan that leads to the sample being a "miniature."

Covariates

It is common for researchers to refer to their sample as “diverse” or “generalizable” based on a few covariates:

- Free or reduced lunch %
- % Minority
- Urbanicity (urban vs rural)

Covariates

This is not enough.

In order to provide an unbiased estimate of the population ATE, the set needs to include all covariates that explain variation in treatment impacts.

But we don’t know this!

So the goal is to be **bias-robust**. Include more to be safe!

Stratification

Question: Given a set of covariates, how can we develop a recruitment plan that leads to a sample that is a miniature of the population?

Answer: **Stratification**.

For example, if the population includes: Urban, Rural, Town, and Suburban areas, our sample should include some of each.

When there are many covariates, one method is to create strata using **k-means cluster analysis**.

Benefits

Using strata ensures that:

- An **inference population** is well-defined.
- The sample selected at the end is a **miniature** (in terms of frequency) of the population.
- A **recruitment plan** is developed that is targeted.
- Recruiters **“see”** a large pool of potential schools, not just those they are familiar with.
- **Non-response** can be tracked (allowing future analyses).

Flexibility

A sample that is a miniature of the population on a large set of covariates is best.

But second best is a sample that achieves coverage – that is, that represents a variety of schools. Post-hoc estimators can reduce bias.

The recruitment strategy can take into account:

- Regional preferences (e.g., can only sample in some states);
- District preferences (e.g., it’d be nice to find one or two districts that can help)

Strata ≠ Geography

A final aside: The way we often talk about representativeness is to focus on geography.

> “This study includes schools in 10 states …”
>
> “This study only includes schools in 1 state…”

It is unlikely that geography itself explains variation in treatment impacts.

It is possible to design a study to make *better* generalizations to the United States, for example, with schools in a single state than with schools in 10 states.

States that are similar to the US

If you must limit your study to a single state, keep in mind that some states are more diverse than others.

The following are similar to the US population of elementary schools:

- Illinois
- Virginia
- Wisconsin
- Pennsylvania
- North Carolina

Covariates: School size, %Black, %Hispanic, %Female, %FRL, Urbanicity, %TitleI, District FTE, District size; see Tipton (2014)

In action

Open your browser (Google Chrome is best) and go to:

www.thegeneralizer.org

Activity

For these activities, see the handouts provided.

Activity

You are planning a Goal 3 study evaluating an elementary school reading program.

Based on a power analysis, the study will include 40 schools.

The interest is in making broad generalizations to schools throughout the United States.

Issues & Questions

If you want to recruit in a single state, you have a few options:

1. Limit your generalizations to that state.
2. Attempt to generalize more broadly (e.g., to the US) but only including schools from a single state.

In The Generalizer, to do the latter you will need to make the broader population (e.g., the US) your inference population, but then check the resulting strata to see if there are schools in the state of interest in *every* stratum.

Issues & Questions

This may be iterative and involve playing around with different numbers of strata.

It may also result that this is impossible. For example, generalizing to the US from a study conducted in WV only may be impossible.

Issues & Questions

You get downloadable .csv files with information on schools in each stratum.

There are columns also for tracking recruitment. We recommend using these and keeping good notes.

You can then analyze the types of schools that say “yes” compared to those that say “no.” (See Tipton et al, 2016 JREE).

15 Minute Break

Session 3: Assessing Similarity

Basic premise

You completed a study and have an estimate of the sample ATE.

Is it possible to use data from this experiment to estimate the ATE in various populations?

Put another way:

- How similar is the experimental sample to different relevant inference populations?
- Is the sample a miniature of any of these populations?

Which populations?

While geography diversity isn’t essential for *sampling,* geography can matter for inferences.

At a minimum, we may want to provide ATE estimates for:

- The United States
- Each of the 50 states

These estimates correspond to decisions to roll out a program at a national or state level.

Comparisons

Since we don’t actually know treatment impacts for those in the population – and we likely also don’t know *outcomes* for those units – the goal is to assess similarity on covariates.

Which covariates? Those that might explain variation in treatment impacts.

Remember:

- Tests of treatment impact moderation are typically severely underpowered in studies.

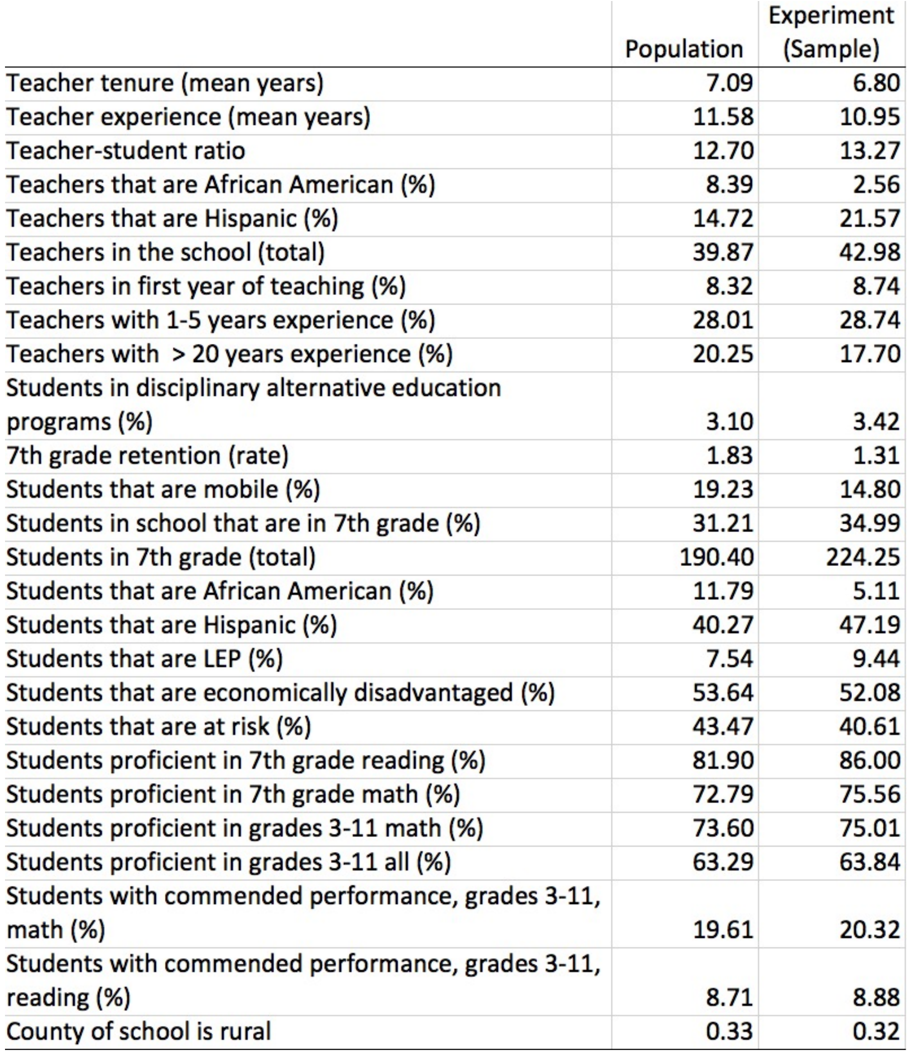
Example

SimCalc is a middle-school computer-based mathematics program that teaches proportionality and rates of change.

In 2008-9 SRI conducted a cluster-randomized evaluation of SimCalc that included 7th grade classrooms in 73 schools in Texas.

Question: Is the sample similar to the population of (relevant) schools in the United States?

Comparing: Similar?



	Population	Experiment (Sample)
Teacher tenure (mean years)	7.09	6.80
Teacher experience (mean years)	11.58	10.95
Teacher-student ratio	12.70	13.27
Teachers that are African American (%)	8.39	2.56
Teachers that are Hispanic (%)	14.72	21.57
Teachers in the school (total)	39.87	42.98
Teachers in first year of teaching (%)	8.32	8.74
Teachers with 1-5 years experience (%)	28.01	28.74
Teachers with > 20 years experience (%)	20.25	17.70
Students in disciplinary alternative education programs (%)	3.10	3.42
7th grade retention (rate)	1.83	1.31
Students that are mobile (%)	19.23	14.80
Students in school that are in 7th grade (%)	31.21	34.99
Students in 7th grade (total)	190.40	224.25
Students that are African American (%)	11.79	5.11
Students that are Hispanic (%)	40.27	47.19
Students that are LEP (%)	7.54	9.44
Students that are economically disadvantaged (%)	53.64	52.08
Students that are at risk (%)	43.47	40.61
Students proficient in 7th grade reading (%)	81.90	86.00
Students proficient in 7th grade math (%)	72.79	75.56
Students proficient in grades 3-11 math (%)	73.60	75.01
Students proficient in grades 3-11 all (%)	63.29	63.84
Students with commended performance, grades 3-11, math (%)	19.61	20.32
Students with commended performance, grades 3-11, reading (%)	8.71	8.88
County of school is rural	0.33	0.32

Problems

How do we summarize the overall similarity between the sample and population?

- Each covariate results in a separate SMD.

How do we know if it is “similar enough”?

Now, imagine replicating this 50 times, producing separate tables for each state!

Solution

What we need is a **single number summary** that that provides the degree of similarity *across all these covariates.*

The **generalizability index** (Tipton,2014) does exactly this:

- It is simple to compute (and automatic in The Generalizer);
- It takes values between 0 and 1;
    - 1 indicates the sample is an exact miniature of the population;
    - 0 indicates the sample and population share no common features;

Index > .90

If the generalizability index > .90, the sample is as similar to the population *as a random sample of the same size*.

In this case, the ATE estimated in the sample is unbiased for the ATE in the population (assuming ignorability).

But also…

When the index < .90,

- the sample is not sufficiently similar to make generalizations directly from the ATE in the sample to that in the population.

But we could do some statistical adjustments, right?

Statistical adjustments

A growing literature focuses on methods for *reweighting* the sample to be more similar to the population:

- Inverse probability weighting (Stuart et al, 2011);
- Post-stratification (Tipton, 2013);
- GBM, regression, and doubly robust methods (Kern, Stuart, Hill, & Green, 2016).

But these methods don’t always work well.

Simple reweighting

	Urban	**Rural**	**Suburban**	**Town**
Population proportion	20%	30%	40%	10%
Proportion in experimental sample	10%	50%	20%	20%
Estimated Treatment impact	**0.29**	**-0.13**	**0.49**	**0.08**
Standard error	***.18***	***.04***	***.08***	***.10***

$$\widehat{\tau}_p = \sum_{i=1}^{k} w_{pi}\widehat{\tau}_i = .20(.29) + .30(-.13) + .40(.49) + .10(.08)$$

$$SE\left(\widehat{\tau}_p\right) = \sqrt{\sum_{i=1}^{k} w_{pi}^2 SE\left(\widehat{\tau}_i\right)^2} = \sqrt{.20^2\left(.18^2\right) + .30^2\left(.04^2\right) + .40^2\left(.08^2\right) + .10^2\left(.08^2\right)}$$

Reweighting

There are two issues:

- If there is an empty cell – i.e., under-coverage – reweighting is *impossible*.
- If the sample and population are really different, reweighting is *possible*, but *not useful.*
  - The estimator can still have bias remaining (more common in complex reweighting);
  - The estimator can have a large standard error.

Reweighting examples

	Urban	Rural	Suburban	Town
Population proportion	20%	30%	40%	10%
Schools to recruit (n=40)	40x20% = 8	12	16	4
Actual Sample #1	7	14	13	6
Actual Sample #2	10	0	20	10
Actual Sample #3	20	2	8	10

Compared to an unadjusted estimator (that is biased!):

- Sample #1: the reweighted estimator has a standard error about **2%** larger.
- Sample #2: this is an example with a coverage error. We cannot reweight.
- Sample #3: the reweighted estimator has a standard error about **64%** larger!

Link between estimation and index

The generalizability index is predictive of the:

- Amount of bias that can be reduced using reweighting;
- The increase in standard errors resulting from reweighting.

Thus, a larger index --> a more **useful** estimate.

.9 > index > .5

A lot of states will end up in this range. What can you do?

First, the ATE and standard error estimated in the study are biased.

Second, you’ll need to do more work if you want to provide estimates. You’ll need to reweight using propensity score-based methods. (This is beyond The Generalizer).

Index < .5

In these states, generalizations should not be made.

That is, the sample is *so different* from the population that reweighting won’t help:

- Under-coverage: e.g., no schools with Hispanic students, yet want to generalize to Texas.
- Large distributional differences: e.g., most schools in the study have low FRL %, whereas most in the population have high FRL%.

The sample does not represent the population of these states. Further research is needed to understand if the results would generalize there.

How does it work?

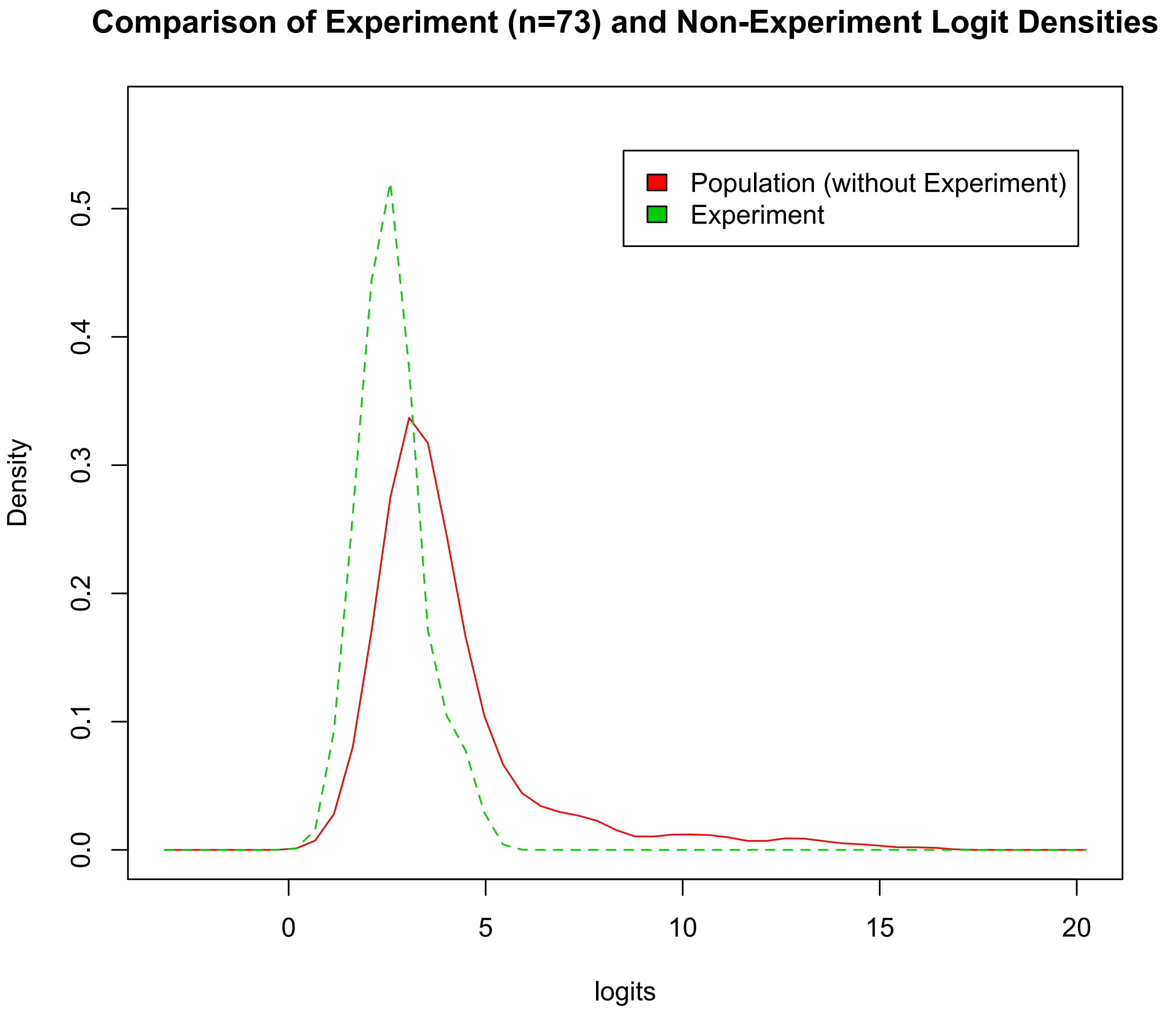
You input:

- Inference population
- Covariates

It compares the sample and population by estimating a propensity score, i.e.

$$\log\left(\frac{s(\boldsymbol{X})}{1 - s(\boldsymbol{X})}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

How does it work?



How does it work?

If we divided these densities (histograms) into k bins:

$$B = \sum_{j=1}^{k} \sqrt{w_{pj} w_{sj}}$$

It can be shown that we can also write this as

$$B = \sqrt{\theta\tau} \sum_{j=1}^{k} \sqrt{w_{p_0 j} w_{s_0 j}}$$

Where

- $\boldsymbol{\theta}$ measures the proportion of the population which overlaps with the sample;
- $\boldsymbol{\tau}$ measures the proportion of the sample that overlaps with the population; and
- $\boldsymbol{w}_{pj0}$ and $\boldsymbol{w}_{sj0}$ are the weights within the overlap region.

In action

Open your browser (Google Chrome is best) and go to:

www.thegeneralizer.org

Activity

For the activity, see the handout provided.

Activity

You completed an experiment evaluating an elementary school reading program in 28 schools.

You want to understand how useful the results might be for estimating average treatment impacts for each of the 50 states.

You want the broadest possible generalization – to all relevant schools – and you think effects may vary in relation to basic school demographics.

Issues & Questions

Sometimes you find that the results don’t generalize well to the population.

If you believe the variables you selected matter, the best strategy for improving generalization is to change the inclusion criteria for the population.

Go back to the inclusion criteria and play around.

- For example, maybe similarity increases when you focus only on small or medium sized districts. Or only on Title I schools.

Conclusion

But I want to estimate population ATEs…

We have an R package in the works.

It is called ‘`generalize`’ and will be available soon.

It has 3 functions:

- `stratify()`
- `assess()`
- `weight()`

Most importantly – it works with *any* data, not just the CCD.

Other issues in generalization

Today we’ve focused on what The Generalizer can do.

But this focuses only on generalizations over **units**.

There are also generalizations across:

- Treatments (e.g., with teacher training, without)
- Outcomes (i.e., test used)
- Settings (e.g., in school vs after-school programs)

Estimation

In Assess-mode, The Generalizer answers if the sample is like a random sample (>.90) or if generalizations are not warranted (<.50).

When the index is between .50 and .90, a less-biased estimate of the ATE can be calculated. This requires **reweighting**.

These methods use propensity scores – since they involve reweighting on several variables. Papers included in the workshop folder give an overview of these approaches.

Treatment effect heterogeneity

The approach to generalization given today focuses on **average treatment impacts**.

The flip side of this is **treatment effect heterogeneity**.

In order to understand which covariates matter, we need to know how much effects actually vary in a broad range of schools.

Treatment effect heterogeneity

The ideal study design for studying heterogeneity is not a “miniature” but instead one that maximizes diversity (‘heterogeneous irrelevencies’).

But this approach too requires starting the study with a conversation about the goals, including:

- the estimand,
- the inference population,
- covariates, and
- purpose.

Recruitment issues

In planning studies for generalization, we need to think more carefully about recruitment, including:

- How to best incentivize a diverse pool of schools to take part;

Research on this area is just starting. I encourage you to begin keeping track of what works and doesn’t work in your recruitment (i.e., study this!).

Final take home points

1. **Relevance is always about generalization.**

2. **Generalizations will be made**.
    - The question is not “do we want to generalize?” but instead “do we want to **lead** the generalizations?”

Final take home points

3. **If at all possible, plan for generalization**.
   - Even when your best efforts fail, you will be in a better situation for post-hoc statistical adjustments.
   - Aim for **bias-robustness**: attempt balance on many covariates!
4. **When not possible, help others understand where results generalize.**
   - Be sure to report your **assumptions**: Which covariates did you compare on. Why these?

Thanks!

Elizabeth Tipton

Teachers College, Columbia University

tipton@northwestern.edu

https://steppcenter.northwestern.edu

@stats_tipton

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, *47*(8), 516-524.

Also

If you use The Generalizer and run into problems, please let me know:

- Note the page (the % bar at top);
- If appropriate, send a screen shot;

And also:

- If there are features you think that would be useful, let me know!

Extra slides

Three generalization problems

1) The population may contain units that have no units in the experiment like them.

- **Example:** The population includes all public schools in the state, but the sample does not include any charter or magnet schools.

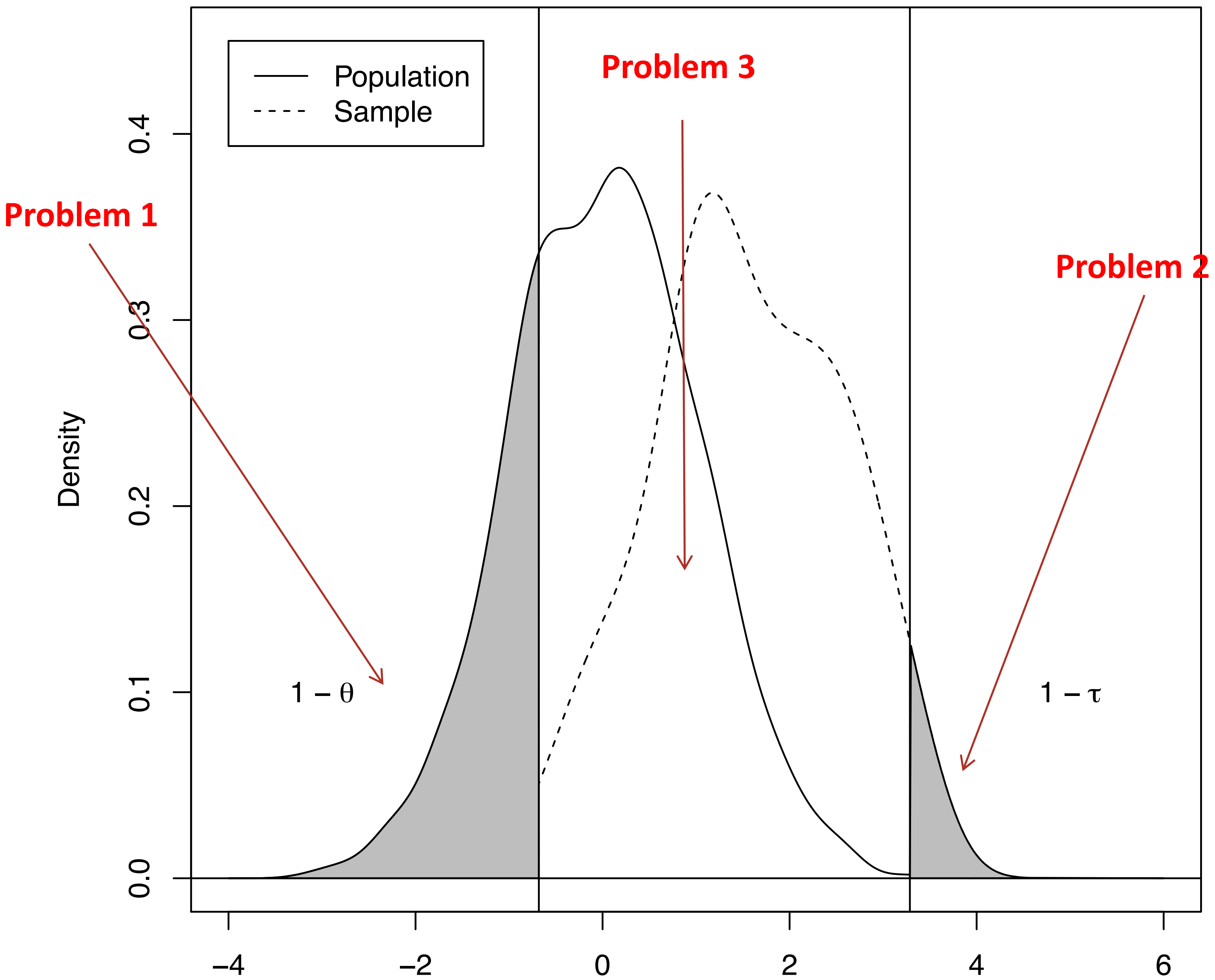
2) The sample may contain unit that have no units in the population like them.

- **Example:** The experiment includes schools with a large proportion of ELL students, while the population doesn't have any schools like this.

3) The sample and population may have large distributional differences on key covariates.

- **Example:** In the experiment, most schools are Title I schools, while the population includes both Title I and non-Title I schools.

Figure 1: Three regions when comparing densities



Abstract

The Generalizer is a free web tool (www.thegeneralizer.org) that can be used to design sample recruitment plans in education studies, including both K-12 and higher education. The tool walks researchers through the process of identifying a target population, considering potential treatment effect moderators and designing a stratified recruitment plan. In this workshop, I provide an overview of the methods behind the tool and an introduction to the tool itself. For those conducting studies in environments other than schools, there is also an R package with more flexibility, which will also be introduced.