Generalization from Experiments

By

Larry V. Hedges

Important Note

Much of my talk is based on joint work with Colm O'Muircheartaigh

Beth Tipton has also helped with examples and other things in this

Social Experiments

The purpose of social experiments is often to inform policy choices

To do so, generalization to policy-relevant populations is necessary

The use of social experiments requires generalization, *however:*

- Probability sampling is possible, but rarely used
- The same experiment often informs different policy choices, generalization to more than one population is often desirable, and these may not be known in advance

Experimental Design

Yet the topic of sampling of units rarely occurs in work on experimental design

Sampling of treatments does occur and is treated extensively (e.g., Student, 1931; Cornfield and Tukey 1955)

Most statements about generalization are qualitative, e.g.

If we can arrange, without decreasing the accuracy of the experiment, to examine a wide range of conditions, this is desirable. This is particularly important in experiments to decide some practical course of action (Cox, 1958, p. 10)

Experimental Design

Classical experimental design has stressed the randomization distribution

RA Fisher (1935) gave the “reasoned basis” for inference in his lady tasting tea example

A lady argues that she can tell whether the tea, or the milk, is added first to the cup

Fisher proposes an experiment in which 8 cups of tea (4 of each kind) are tasted in random order

Experimental Design

It is [not sufficient] to insist that all the cups must be exactly alike in every respect except that to be tested. For this is a totally impossible requirement in our example, and equally in all other forms of experimentation.

The element in the experimental procedure which contains the essential safeguard is that the two modifications of the test beverage are to be prepared in random order. This, in fact, is the only point in the experimental procedure in which the laws of chance, which are to be in exclusive control of our frequency distribution, have been explicitly introduced. (Fisher, 1935)

Experimental Design

Fisher’s approach dominated experimental design (see Kempthorne, Cochran, Wilk, Cornfield & Tukey, etc.)

Neyman (1935) had given a sampling-based interpretation of the logic of randomized experiments, but this was largely rejected. For example

> *The difficulty inherent in this approach is that the population of experimental units and repetitions of the experiment about which the inferences are made are unspecified. For this reason we considered it desirable to examine the possibilities of making inferences about the experimental units actually used.* (Kempthorne 1952/1973, p.151)

Experimental Design

This view still dominates today

> *Experiments do not require, indeed, cannot reasonably require, that the experimental units be a sample from a population of units … For valid inferences about the effects of the treatment about the units included in the experiment, it is sufficient to require that treatments be allocated at random to experimental units …probability enters only through the random assignment of treatments* (Rosenbaum, 2002, p. 23)

Experimental Design

What role do sampling models play?

Balanced analysis of variance models give simple approximations to randomization distributions

Balance is not incidental—the statistical methods usually presented are invalid without it (although many approximations have been proposed)

Causal Inference (Rubin-Holland Model)

Assume that STUVA holds

r_{1i} = response of unit *i* under treatment 1
r_{0i} = response of unit *i* under treatment 0

Average causal effect of treatment 1 versus treatment 0 is

$$\tau = \mathrm{E}\{r_{1i} - r_{0i}\} = \mathrm{E}\{r_{1i}\} - \mathrm{E}\{r_{0i}\}$$

Note that this is a population-based definition!

Causal Inference (Rubin-Holland Model)

In a randomized experiment, treatment assignment is independent of **r** = (r_{1i}, r_{0i})

If our experiment draws a simple random sample from the population, an unbiased estimate of the average treatment effect in the population is

$$\hat{\tau} = \bar{r}_{1\bullet} - \bar{r}_{0\bullet}$$

This is well known

Three Experimental Designs

1. The completely randomized design
Treatments assigned to individuals

2. The hierarchical (nested) design
treatments assigned to blocks or clusters

3. The (generalized) randomized blocks design
Treatments assigned to individuals matched within blocks or clusters

The natural sampling design associated with 2 & 3 is two-stage cluster sampling

Probability Sampling and Experiments

Result 1: In the completely randomized design, the usual estimate of the treatment effect is unbiased for the population average under simple random sampling

Hierarchical or RB designs with two-stage sampling

Result 2: In the hierarchical or randomized block designs, the bias of the usual estimate of the treatment effect is

$$-\rho_{N\delta}\sigma_N\sigma_\delta / \bar{N} = -\rho_{N\delta}C_N C_\delta \bar{\delta}$$

where N_i is the population cluster size, δ_i is the treatment effect in the i^{th} cluster, ρ is the correlation and C is the coefficient of variation

This is exactly the bias of the group mean as an estimate of the element mean in demography

Hierarchical Designs with Two-stage Sampling

In hierarchical designs, an unbiased estimate of the treatment effect is

$$\sum_{j=1}^{m_1} \frac{N_{1j}\overline{r}_{1j}}{m_1\overline{N}_{1\bullet}} - \sum_{j=1}^{m_0} \frac{N_{0j}\overline{r}_{0j}}{m_0\overline{N}_{0\bullet}}$$

where m_k is the number of clusters receiving treatment *k*, and N_{kj} is the (population) size of the *j*th cluster receiving treatment *k, and* $\overline{r}_{kj}$ is the average response to treatment *k* in cluster *j*

Randomized Block Designs with Two-stage Sampling

In generalized randomized block designs, and unbiased estimate of the population average treatment effect is

$$\sum_{j=1}^{m} \frac{N_j \left(\bar{r}_{1j} - \bar{r}_{0j} \right)}{m\bar{N}}$$

where *m* is the number of clusters, $\bar{r}_{kj}$ is the average response to treatment *k* in cluster *j*, and N_j is the (population) size of the *j*th cluster

Non-Probability Samples

Let Y be an element of a population with mean μ

Let Z be an indicator of being in the sample such that $Z = 1$ for population members who are in the sample and $Z = 0$ for those who are not

In a random sample, Y is independent of Z so that $E\{Y \mid Z = 1\} = E\{Y\} = \mu$

In a nonrandom sample Y is not independent of Z so that, in general, $E\{Y \mid Z = 1\} \neq \mu$

Non-Probability Samples

A weaker condition than independence of Z and Y is ignorability given a set of observed covariates $\mathbf{x}$

The sampling mechanism is ignorable given $\mathbf{x}$ if Y and Z are conditionally independent given $\mathbf{x}$ and if every element in the population has some chance of being included in the sample, that is

$\Pr\{Y \mid \mathbf{x}, Z\} = \Pr\{Y \mid \mathbf{x}\}$ and $0 < \Pr\{Z = 1\} < 1$

If the sampling mechanism is ignorable and if the distribution of $\mathbf{x}$ is known in the population, an unbiased estimate of μ can be obtained from the sample because

. $$\mathrm{E}_{\mathbf{x}}\left\{\mathrm{E}\left[Y \mid Z = 1, \mathbf{x}\right]\right\} = \mathrm{E}_{\mathbf{x}}\left\{\mathrm{E}\left[Y \mid \mathbf{x}\right]\right\} = \mu$$

Non-Probability Samples

If the sampling mechanism is ignorable given **x** and if the distribution of **x** is known in the population, then

$$\hat{\mu} = \sum_{\mathbf{x}} \pi(\mathbf{x}) \bar{Y}(\mathbf{x})$$

is an unbiased estimate of μ, where $\bar{Y}(\mathbf{x})$ is the mean of Y for given **x** and $\pi(\mathbf{x})$ is the probability of a particular value of **x**

This is approach is called standardization (of populations) in demography

The idea is to use an external probability distribution to reweight the data

Practical Matching Strategies

Even if **x** is of small dimension, many distinct values would need to be matched

Propensity score methods are therefore desirable to reduce the matching to a single dimension

Cochran (1968) and Rosenbaum and Rubin (1984) suggest stratification with as few as 5 strata may eliminate 90% of the bias on **x**

Stratification avoids instability in Horvitz-Thompson weighting

Generalization in Experiments

To use this idea in experiments, we need a slightly weaker ignorability assumption than that just described

Not

$\Pr\{r_{ki} \mid \mathbf{x}, Z\} = \Pr\{r_{ki} \mid \mathbf{x}\}$ *and* $0 < \Pr\{Z = 1\} < 1$

But instead

$\Pr\{r_{1i} - r_{0i} \mid \mathbf{x}, Z\} = \Pr\{r_{1i} - r_{0i} \mid \mathbf{x}\}$ *and* $0 < \Pr\{Z = 1\} < 1$

which may be considerably more plausible

Generalization in Experiments

Result 3: If this ignorability assumption is true, then

$$\sum_{\mathbf{x}} \pi(\mathbf{x})\left[\bar{r}_1(\mathbf{x}) - \bar{r}_0(\mathbf{x})\right]$$

is an unbiased estimate of the treatment effect in the population, where $\bar{r}_k(\mathbf{x})$ is the average response to treatment *k* for covariate value **x**

If **x** has 5 – 10 values (strata), this method implies computing a local average treatment effect in each stratum (value of **x**) and averaging across strata using the population weights *π*(**x**)

Implementing the Strategy

Need a well defined population frame

A set of relevant covariates measured on that population that explain the variation in treatment effects (from a census or probability sample survey)

An experiment that measures these covariates and supports the population inference

Propensity score strata for membership in the experiment provide strata

The standard error of the estimate will depend on how well the distribution of observations in the experiment match the population

Implementing the Strategy

This strategy involves stratifying the experimental design (strata are fixed effects)

Estimation of the local average treatment effect within each stratum involves the same analysis that would be conducted without stratification

But

Clusters may be split between strata, so there are nonzero covariances between local average treatment effects of different strata

Implementing the Strategy

The variance of the estimate quantifies the quality of the generalization

Except for these covariances, the variance of the estimated population average treatment effect is

$$\sum_{\mathbf{x}} [\pi(\mathbf{x})]^2 V\{\bar{r}_1(\mathbf{x}) - \bar{r}_0(\mathbf{x})\}$$

This shows that the variance is large (generalization is poor) when the local average treatment effect in a large stratum is poorly estimated in the experiment

In out experience, covariances have only a small effect

Example: Project STAR

Project STAR: The Tennessee class size experiment (e.g. Nye, Hedges, Konstantopoulos, 2000)

The study involved about 6,500 students in 79 schools in 42 school districts (of the 141) in TN in 1985-1990

The treatment was reduction of class size from about 24 to about 15

This experiment has informed education policy in many parts of the US and beyond and has been called

> "One of the great experiments in education in US history" (Mosteller, et al., 1996)

But, how generalizable are these results?

Example: Project STAR

Reference Population:

US school children the same age as the STAR sample

The 1990 US Census 1% microdata (from IPUMS)

Children in “1st to 4th grade” who were 9 or 10 years of age and who were in the U.S. in 1985

We analyzed about 6,500 students from the TN experiment and about 65,000 children from the Census

Example: Project STAR

Matching with 5 covariates led to 7 propensity score strata

- Race/ethnicity
- SES
- urbanicity
- Gender
- Interactions

We evaluated the match between Census and STAR by testing for differences on main effects interactions, yielding a total of 763 tests

Only 3 of these 763 tests were significant at the 5% significance level

Example: Project STAR

The multilevel mixed model with 7 strata (*S1* to *S7*) and interactions with treatment *T*)

Level 1: Individuals (*j*) within clusters (*i*)

$$Y_{ij} = \beta_{i1}\ S1_{ij} + \cdots + \beta_{i7}\ S7_{ij} + \beta_{i1} S1xT_{ij} + \cdots + \beta_{i14}\ S7xT_{ij} + \varepsilon_{ij},$$

Level 2: Coefficients (*k*) and clusters (*i*)

$$\beta_{ik} = \gamma_{0k} + \eta_{\mathrm{ik}} \qquad \mathrm{V}\{\eta_{\mathrm{ik}}\} = \mathbf{\Sigma}$$

Example: Project STAR

Results of the multilevel analysis

$$\gamma = \begin{pmatrix} -6.28 \\ 13.54 \\ 10.86 \\ 8.25 \\ 10.86 \\ 11.30 \\ 14.76 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 36.67 & 4.63 & 4.54 & 0.12 & 0.13 & -0.01 & 1.91 \\ 4.63 & 32.96 & 3.04 & 0.07 & 0.00 & -0.01 & -1.77 \\ 4.54 & 3.04 & 46.15 & 0.78 & 0.01 & -0.05 & -2.87 \\ 0.12 & 0.07 & 0.78 & 21.75 & 1.15 & -1.21 & -0.12 \\ 0.13 & 0.00 & 0.01 & 1.15 & 13.25 & 3.51 & 0.00 \\ -0.01 & -0.01 & -0.05 & -1.21 & 3.51 & 13.46 & -0.01 \\ 1.91 & -1.77 & -2.87 & -0.12 & 0.00 & -0.01 & 20.31 \end{pmatrix}$$

OLS with cluster robust *SE*s gave similar results

Example: Project STAR

Estimated average treatment effects mathematics achievement grade 1

Standard analysis of the experiment: 12.32(2.35)

Estimates for specific populations:

Entire US: 10.00(2.716)	{19% lower}
CA: 11.94(3.96)	{3% lower}
LA: 6.31(13.58)	{49% lower}
Newark, NJ: 11.16(5.71)	{9% lower}

Implications

Generalizations from experiments can be limited, sampling matters

Therefore we need:

- Data collections in experiments that include relevant covariate sets that will facilitate matching to specific reference populations
- Surveys that measure relevant covariates in reference populations

The two must be designed to complement each other

Further Research

How well does this whole thing work when we actually have a gold standard we can check

Design of experiments to facilitate generalization (efficient estimates of population average effects)

When is it sufficient to match at the cluster (e.g., school) level [many states have relatively good school level datasets]

Understanding how to match with least hidden bias (the selection process for schools or classrooms)