DESIGN OF EXPERIMENTS FOR GENERALIZABILITY

David M. Steinberg
Tel Aviv University

TEL AVIV UNIVERSITY אוניברסיטת תל אביב

Generalization:

A general statement : a statement about a group of people or things that is based on only a few people or things in that group

From *Britannica Dictionary*

Generalization

Our data led to conclusions. Are they applicable to other:

- Subjects
- Conditions
- Time periods
- Animal strains
- Raw materials
- Laboratories, hospitals, locations
- Production facilities (scale-up)
- Tasks with common content

Related Issues

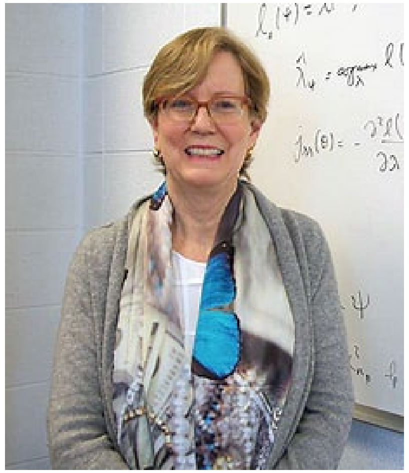
- Out of Domain Prediction
- External Validity
- Reproducibility
- Standardization
- Inclusion/Exclusion criteria
- Transfer Learning
- Key Comparison Experiments
- Robust Design Experiments
- Leveraged Gauge R&R

How can we generate data to promote the ability to generalize??

RA Fisher
The Design of Experiments

A highly standardised experiment supplies direct information only in respect of the narrow range of conditions achieved by standardisation. Standardisation, therefore, weakens rather than strengthens our ground for inferring a result, when, as is invariably the case in practice, these conditions are somewhat varied.

DR Cox & Nancy Reid
The Theory of the Design of Experiments



Sometimes … it is useful to consider the population of ultimate interest and the population of accessible individuals and to aim at conclusions that will bridge the inevitable gap between these. This is linked to the question of whether units should be chosen to be as uniform as possible or to span a range of circumstances. Where the latter is sensible it will be important to impose a clear structure on the experimental units.

My interpretation:

How should we block our experiment??

Blocking:

- Similar units comprise blocks.

- Allocate treatments to units within blocks.

Blocking for generalization:

- What blocks do we want?
- Can we "design" relevant blocks into our experiment?

Blocking

Example: *The Advice4U RCT compared AI-guided and physicians' insulin dose adjustments in 108 juveniles.*

The study included:

- *7 Treatment centers*
- *3 Age groups (10-14, 15-18, 19-21)*
- *3 Baseline levels of HbA1C (7%-8%, 8%-9%, 9%-10%).*

Nimri, R. et al. (2020), Nature Medicine, 26, 1380-1384.

Blocking for Generalization

Design variation into the blocks.

Benjamini's cross-laboratory experiment: labs are a blocking factor; they are chosen to inject inter-laboratory variability.

Key Comparison Studies in metrology follow the same cross-laboratory scheme.

Inference Framework

Denote by μ_i the true effect in block i.

This can be estimated from the data in group i by $\hat{\mu}_i$, which is approximately $N(\mu_i, \sigma_i^2)$.

We can treat μ_i as being sampled from a random effect distribution, e.g. $\mu_i \sim N(\mu, \tau^2)$.

Inference Framework

Generalization requires good knowledge of τ^2.

This comes from the *treatment by block interaction*.

Practical problem: from a small sample (e.g. just 3 groups) difficult to estimate τ^2 well.

Conceptual problem: in any real experiment, the random sampling assumption is dubious; the blocks will almost always be a convenience sample.

Solution by Design

Choose the groups to intentionally differ on key properties.

This is the idea behind *Taguchi's robust design experiments*, with properties reflecting noisy inputs to an industrial process.

Ditto for *Leveraged Gauge R&R*, with intentional choice of extreme units.

Solution by Design

There are two main benefits:

1. Estimating fixed effects is much easier than estimating variances.

2. We gain insight about the relationship of the effect to the features that were varied.

Solution by Design

Analysis now relates μ_i to x_i, the measured property of the *i*'th block: $\mu_i = \beta_0 + \beta_1 x_i$.

The inter-block variation is now a function of the measurable variable x_i.

Solution by Design

There are still major limitations:

1. The number of groups that can be generated puts a cap on the degree of learning.
2. It may be practically difficult to generate the desired blocks.
3. Requires careful thinking to choose x. There will always be a number of candidates.

Solution by Design

Richter et al. implemented this idea in an experiment on animal behavior. The goal was to use the features to mimic inter-laboratory variation.

For example, animal age was used to generate blocks.

They found that the forced blocking variation led to more consistent and reproducible results.

Richter, S.H., Garner, J.P., Auer, C., Kunert, J. & Würbel, H. (2010), Systematic variation improves reproducibility of animal experiments, *Nature Methods*, 167-168.

Summary

Study design plays an important role in achieving conclusions that can be generalized.

The statistical design research has largely ignored the issue of generalization.

Forcing in variation by “designing” the block structure has excellent potential to fill this gap.