Tool 4.4 – Evaluation Glossary

This tool provides a glossary of terms for staff members who are learning concepts and vocabulary related to program evaluation.

Propensity Score Matching

- **Propensity score:** the probability of being in the group that is offered the program or affected by the policy change based on their observed characteristics.

- **Propensity score matching (PSM):** a method for identifying a comparison group that has observed characteristics similar to those of the program group.

Counterfactual

- What happens in the absence of a program or policy change? In an experiment, the control group provides the **counterfactual**. Without an experiment, we can use other methods (like propensity score matching) to find a counterfactual (often called a comparison group).

Study Types and Definitions

- **Outcome:** The level on a measure for an individual or a group of individuals that is used to assess program performance.

- **Impact:** The difference between the average outcomes for the program group versus the comparison or control group. Outcomes and impacts are not the same.

- **Pre-/post-test:** A research design in which the same outcome is assessed before and after a program. This is a weak research design because typically many other factors could explain pre/post differences such as maturation, other historical events, and so on.

- **Longitudinal tracking study:** Follows study participants over time and collects data to measure their outcomes. While this can be useful for generating hypotheses, it is weak for the purposes of causal inference.

- **Regression discontinuity design:** Researchers take advantage of a threshold in the program eligibility criteria (for example, a test score or income threshold). Individuals above (or below) the threshold serve as the program group and individuals below (or above) the threshold serve as the comparison group. The estimated impact is defined only for individuals very close to the threshold. The validity of the design assumes that at the threshold, the design is equivalent to a random assignment design. This is considered the second most rigorous research design. But it requires more sample than a randomized controlled trial (RCT).

- **Comparative interrupted time series:** Uses longitudinal data for a program group and a matched comparison group to estimate the effects of an intervention. The analysis compares the two groups' deviations from their baseline trends after the intervention. Because of the comparison group and the multiple observations, this method is more rigorous than pre-/post- tests.

- **Synthetic controls:** A statistical method used to evaluate the effect of an intervention that affected a place. The comparison group is similar to a propensity score created group but other design elements such as matching on time trends are included.

- **Random assignment:** Divides study participants into a "treatment group" (or "program group") that is eligible to receive program services and a "control group" that is not eligible. Comparing the outcomes of the groups over time allows us to estimate the impacts of the program. This is the most rigorous research design.

Selection Bias

- The bias introduced by the selection of individuals, groups, or data for analysis in such a way that proper randomization is not achieved, thereby failing to ensure that the sample obtained is representative of the population intended to be analyzed.

Observed and Unobserved Characteristics

- PSM matches individuals in the program and comparison groups on **observed characteristics** – so the two groups are "balanced." This may account for some of the differences in **unobserved characteristics**. And this balancing of characteristics can help reduce **selection bias**.

- **Observed characteristics** are variables for which you have measurements in your dataset.

    - Baseline characteristics, for example, age, gender, employment history

- **Unobserved characteristics** are variables for which you *don't* have measurements in your dataset.

    - Motivation, grit, decision making, for example

Comparison Pool and Comparison Group

- A **comparison pool** is identified first. Then, you match your program group to the larger pool to identify your comparison group. The **comparison group** is a subset of the comparison pool.

- When creating your comparison pool:

- The comparison pool should be larger than program group.
- You should refine your comparison pool as much as you can at this stage.
- You need good data on the comparison pool to help with selection bias and unobserved characteristics.

- Questions to consider when identifying and refining your comparison pool:
  - Can a valid comparison group be identified?
  - Are the key characteristics of individuals measurable?
- The comparison group should ideally share the following characteristics with the program group:
  - Same geographic location
  - Same time period
  - Meets the program eligibility criteria
  - Same data available

Calculating Propensity Scores

- The propensity score is the probability of being in the group that is offered the program or policy change based on their observed characteristics. This is often calculated using logistic regression. **Logistic Regression** is used when the dependent variable (target) is categorical.
- What characteristics should you use to calculate propensity scores?
  - You want to balance the program and comparison groups on many characteristics.
  - Characteristics must be measured before the program or policy change ("baseline").
  - Characteristics to consider including:
    - Baseline measures of the outcome
    - Demographics
    - Other characteristics that predict the outcome
    - Interactions and higher-order terms

Common Support

- We want a distribution of propensity scores – not everyone should have the same probability of being offered the program or policy change. In lay terms, common support means that you have enough similar individuals in your comparison pool to match. It becomes a problem at the extremes of a distribution (for example, if matching the individuals with the very lowest income, or the most serious barriers to employment)

- **Minima/maxima**: Drop observations where scores are outside the range of the other group.

- **Trimming**: Require a specific percentage of observations within minima and maxima.

Matching

- Individuals in the program group are matched to individuals in the comparison group on their calculated propensity scores. After matching, it is critical to check that the two groups are similar on observed characteristics.

- **Considerations**

  - Replacements: Can individuals in the comparison pool be matched to more than one person?

  - Oversampling: Do you want to match individuals to multiple comparison pool members?

- **Methods**
  The choice of matching method is not as important as having good data to match on!

  - **Nearest neighbor:** The most straightforward matching estimator is nearest neighbor (NN) matching. The individual from the comparison group is chosen as a matching partner for a treated individual that is closest in terms of propensity score.[1]

  - **Caliper:** Closest match within a specified boundary (consider this like a range beyond which you would not consider an individual a match).[2]

---

1. Caliendo, Marco Sabine Kopeinig (2005).
2. Caliper and radius matching: NN matching faces the risk of bad matches if the closest neighbor is far away. This can be avoided by imposing a tolerance level on the maximum propensity score distance (caliper). Imposing a caliper works in the same direction as allowing for replacement. Bad matches are avoided and hence the matching quality rises. However, if fewer matches can be performed, the variance of the estimates increases. Applying caliper matching means that the individual from the comparison group is chosen as a matching partner for a treated individual that lies within the caliper ("propensity range") and is closest in terms of propensity score.

- **Stratification:** grouping by score range.

- **Kernel matching (KM) and local linear matching (LLM):** non-parametric matching estimators that use weighted averages of all individuals in the control group to construct the counterfactual outcome.

- **Assessing the match**
  You want to confirm that the program and comparison group are similar on characteristics measured before the program or policy change.

  - Are the average characteristics of individuals similar across the two groups? (The preferred answer is "yes.")

  - Can you predict who is in the program group based on their characteristics? (The preferred answer is "no.")