Good statistical methods are (1) easy to use and understand for methodologists and (2) allow nonspecialists to quickly and accurately appraise the quantitative evidence those specialists produce. Matching is simple and thus appeals to non-technical audiences. But whether or not a given match is in some senses “good” is not so clear. Clearly, we want any procedure to produce sets of “treated” and “control” units which are “similar” or “balanced” enough. But what is the standard to which we ought to compare a given measure of balance? How should we diagnose the effectiveness of a given matching in reducing bias?

In Hansen and Bowers (2008) we proposed “the randomized experiment” as such a standard and we developed statistical tests to allow quick and easy comparisons of matchings against this benchmark. In this paper we describe the \texttt{Rtools} software that we have written to make those tests broadly available to applied researchers.

The Latently Randomized Experiment and Why It Is a Useful Standard for Balance Assessment

The Potential Outcomes Approach to Causal Inference

An intervention, \( Z \), is posited to change an outcome \( Y \).

Potential outcome models stipulate that to each unit \( i \) and level of intervention \( z \) corresponds a definite value of \( y \), observed if and only if \( Z_i = z \). For the basic case that \( Z \in \{0,1\} \), \( Y_i(z = 1) \) tells us how unit \( i \) would act in the presence of a treatment, and \( Y_i(z = 0) \) tells us how unit \( i \) would act in the absence of a treatment.

The treatment effect on unit \( i \) is a difference of its potential outcomes at different levels of \( z \): for example, \( \tau_i = Y_i(1) - Y_i(0) \).

A (possibly biased) estimate of the treatment effect for \( i \) can be had by comparing \( Y_i(z = 0) \) and \( Y_i(z = 1) \). This requires a correct specification of \( Y_i(z) \), with that of some other unit \( j \) receiving a different treatment condition, \( z_i \neq z_j \), but deemed similar to \( i \) in other relevant respects.

A valid estimate of the treatment effect requires an argument in favor of the claim that the only difference between \( i \) and \( j \) which is consequential for \( Y_i \) is \( Z_i \)(Also known as an assumption about ignorability).

Balance characterizes a (simple) causal comparison when

\[ Z \perp X. \tag{1} \]

as in a simply randomized experiment.

A latently randomized experiment is an observational study accurately described by

\[ Z \perp (Y_i(z) | z \in \{0,1\}, X). \tag{2} \]

It is often practically useful to find reductions \( X \) of \( X \), particularly matchings or stratifications based on it, such that \( Z \perp |Y_i(z)|X \) should hold if \( Z \) does.

How balanced is balanced enough?

Answer #1 As balanced as possible.

Answer #2 When after matching, each adjusted baseline difference on each variable is smaller than some fixed multiple of a pooled s.d. — like \(< 2 \text{ s.d.} \) (Rosenbaum and Rubin, 1984).

Answer #3 At least as close as would be the case in a (stratified or matched) randomized experiment:

\[ Z \perp X_i. \tag{3} \]

This doesn’t preclude more balance. But establishes a minimum standard.

Each attempts in a different way to answer: “Close enough not to bias estimates of treatment effects.”

We prefer answer #2.

The Test Statistics: \( d \) for one \( X \) and \( d^2 \) for all \( X \)

Hansen and Bowers (2008) argue that the optimal test statistic for detecting imbalance on a given \( X \) is

\[ d = \hat{\beta}_X = (Z^T X)^T \hat{\beta} / \hat{\sigma}_X, \]

where \( \hat{\beta} \) and \( \hat{\sigma}_X \) are strata indicators.

Note: This regression does not assume that \( \hat{\beta} \sim N(\hat{\beta}, \sigma^2(X^T X)^{-1}) \). Nor do we assume that \( \text{E}(x) = \beta\hat{z}_i \).

Our test compares \( d \) to a null distribution that comes only from (3).

Equation 3 also implies a joint \( (\chi^2) \) distribution all of the \( \chi^2 \) across all of the relevant covariates. This test assesses balance on all of the linear combinations of all of the covariates. For most balance testing we recommend this omnibus test, \( d^2 \).

Example: The Effect of Deadlines on Drug Safety

The Prescription Drug User Fee Act (PDUFA) of 1992 required the US Food and Drug Administration to act on 90% of “standard” drugs within 12 months. This was the first time that time pressure became a part of the assessment of drug safety in the USA.


We compare submitted and approved during PDUFA I to drugs submitted and approved in the 4 years before PDUFA I to assess this claim.

Here we show the results of our balance assessment before and after full, optimal matching (Hansen, 2004) stratified by “priority” versus “standard” drug-type.

### References


*Assuming \texttt{SUTVA} (Rubin, 1986)

Polmeth 2008, Political Methodology Annual Conference, University of Michigan @ Ann Arbor, July 9-12, 2008