Propensity Scores in the Design of Observational Studies for Causal Effects

BY P. R. ROSENBAUM, D. B. RUBIN

Presented by Siyu Chen, Hedong Yan

Roadmap

- Recalling the early 1980s
- What is planned in planned empirical research?
- The propensity score paper
- Conclusion: design is key, and propensity scores simplify design

Recalling the early 1980s (1)

- Statisticians devoted to the design of experiments
- Ronald Fisher, Jerzey Neyman, George Box, William Cochran, David Cox, John Tukey, Paul, Erich Lehmann, Henri Scheffe
- Elaborating analytical methods were more limited in 1980's

"To find out what happens to a system when you interfere with it, you have to interfere with it."

From "The Use and Abuse of Regression" by George Box (1966)

Run an experiment, not a regression.

What is planned in planned empirical research? (2)

Outline of section 2

- The origin of planning: “The wrecks of research projects”
- First steps: The objective shapes the investigation, not the reverse
- A Study Designed in Anticipation of a Planned Primary Analysis
- Planning and Designing Observational Studies

The origin of planning: “The wrecks of research projects” (2.1)

William Cochan(1995):

- In earlier days, scientists tends to think of seeing a statistician when he has some problem in the analysis of his data.

- Now the statisticians and the scientist both had to learn, that little could be done to get these wrecks floating again.

- Some errors in data collection

- Statisticians began to study the process of collecting data

The origin of planning: “The wrecks of research projects” (2.1)

David Cox(2007)：

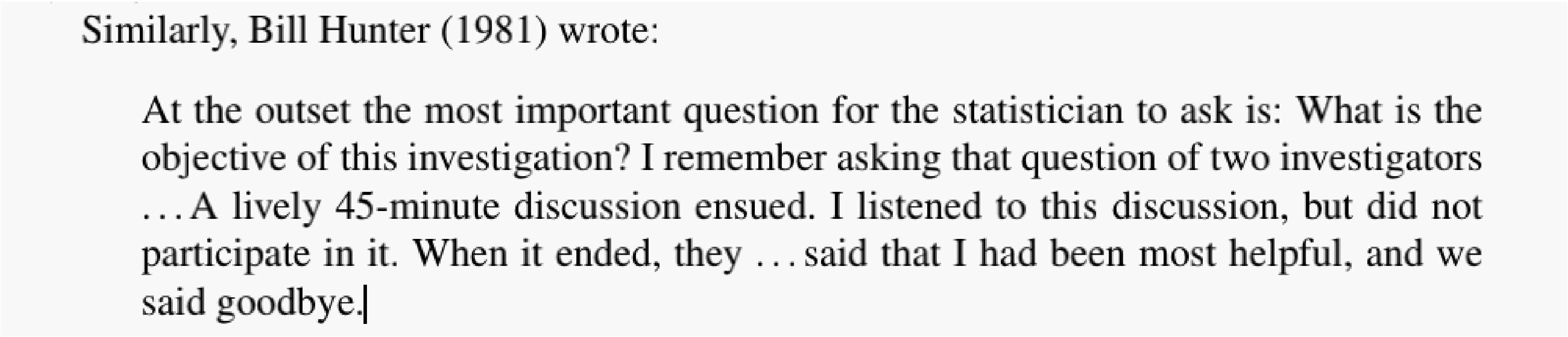
- The design of a study is crucial.

- A seriously defective design may be incapable of rescue even by the most ingenious of analyses.

- The aim in design is to achieve a secure investigation in which the analysis and its interpretation requires as few external assumptions as possible.

First steps: The objective shapes the investigation, not the reverse (2.2)

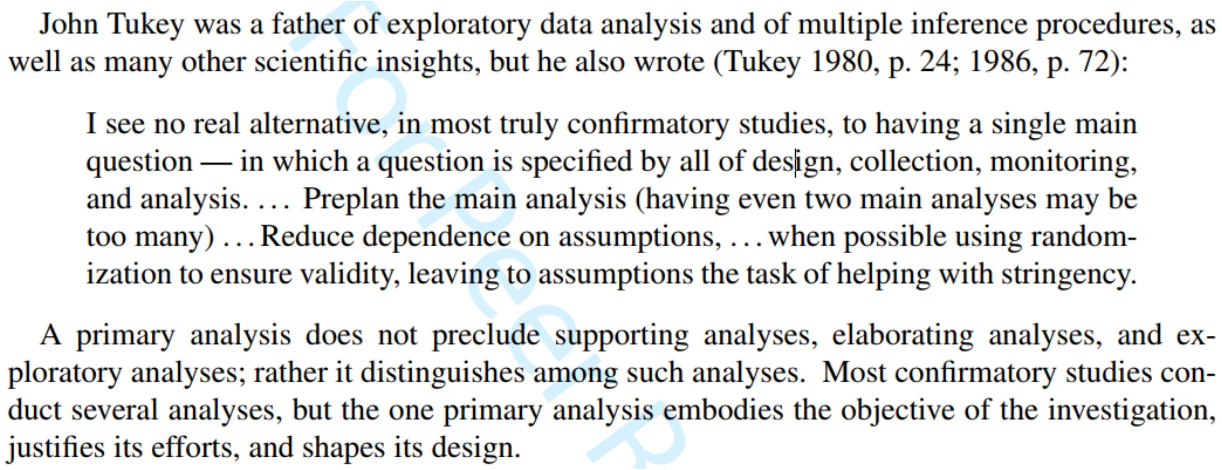
What is the objective of this investigation?



Similarly, Bill Hunter (1981) wrote:

> At the outset the most important question for the statistician to ask is: What is the objective of this investigation? I remember asking that question of two investigators …A lively 45-minute discussion ensued. I listened to this discussion, but did not participate in it. When it ended, they …said that I had been most helpful, and we said goodbye.

A Study Designed in Anticipation of a Planned Primary Analysis (2,3)



John Tukey was a father of exploratory data analysis and of multiple inference procedures, as well as many other scientific insights, but he also wrote (Tukey 1980, p. 24; 1986, p. 72):

> I see no real alternative, in most truly confirmatory studies, to having a single main question — in which a question is specified by all of design, collection, monitoring, and analysis. . . . Preplan the main analysis (having even two main analyses may be too many) . . . Reduce dependence on assumptions, . . . when possible using randomization to ensure validity, leaving to assumptions the task of helping with stringency.

A primary analysis does not preclude supporting analyses, elaborating analyses, and exploratory analyses; rather it distinguishes among such analyses. Most confirmatory studies conduct several analyses, but the one primary analysis embodies the objective of the investigation, justifies its efforts, and shapes its design.

- A single main question — in which a question is specified by all of design, collection, monitoring, and analysis. . . .
- A primary analysis does not preclude supporting analyses, elaborating analyses, and exploratory analyses; rather it distinguishes among such analyses.

Planning and Designing Observational Studies (2.4)

- The planner of an observational study always ask himself the question, ‘How would the study be conducted if it were possible to do it by controlled experimentation?’

- RCT: function of blocking or adjustment is to increase precision rather than guard against bias.

- Observational studies: blocking or adjustment has additional role of protecting against bias.

Planning and Designing Observational Studies (2.4)

- To reach conclusions about causation from association, one needs to address possible biases due to unmeasured covariates.
- The step from association to causation is central to the plan for an observational study.
- matching or blocking for observed covariates is intended to address a comparatively straightforward problem — bias from measured covariates — in such a transparent way that undivided attention can quickly shift to the key step from association to causation. (**Will it work in practice?**)

Some seeming inconsistencies (3.1)

- Point 1: “Cochran (1965, § 3) had expressed doubts about the possibility of adjusting for many measured covariates, and about adjusting for covariate distributions that are very different.”

Some seeming inconsistencies (3.1)

- Point 2: John Stuart Mill (1872, Book III, § 8)
  - “method of difference”: compare treated and control individuals who were identical but for the treatment
  - “If an instance in which the phenomenon . . . occurs and an instance in which it does not . . . have every circumstance save one in common . .. [then] the circumstance [in] which alone the two instances differ is the ... cause or a necessary part of the cause (III, § 8)”
  - “It is not sufficient remedy 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” *Design of Experiments by Fisher*

Some seeming inconsistencies (3.1)

- Point 3: Fisher’s (1925, ch. 8; 1935)
  - “In a randomized experiment, Fisher’s (1925, ch. 8; 1935) theory of randomization inference warranted causal inferences whether there were many covariates, just a few, or none at all. People differ in their genes, in the organization of their neurons, in their immunological histories, all of which are immensely complex, but these differences do not matter for successful causal inference in randomized experiments, where “success” means that point estimates of causal effects are consistent and interval estimates achieve their nominal coverage rates.”

Some seeming inconsistencies (3.1)

- Point 4: Rubin 1976
  - “multivariate Normal covariates with different mean vectors in treated and control groups, but the same covariance matrix”

  - “all of the bias from observed covariates lies along one dimension, the linear discriminant for these observed covariates: a treated and a control individual with the same linear discriminant have the same conditional distribution of observed covariates given the linear discriminant”

  - “Match for the linear discriminant — one observed covariate — and you expect to balance all covariates”

  - “A related situation occurs with ellipsoidal distributions”

The futile attempt to compare identical people under alternative treatments (3.2)

- “It is impossible to place 2*n* people into *n* pairs to exactly match for *K* binary covariates”

- “Saying this more precisely, with *K* binary covariates, there are 2*K* exact-match categories, so the number of exact match categories grows exponentially with *K* and cannot be exactly balanced even in enormous sample sizes, 2*n”*

- “There is no need to condition on covariates in a randomized experiment: the unadjusted treated-minus-control difference in mean outcomes is unbiased and consistent for the average treatment effect with no adjustment for covariates”

Balance as an alternative to exact matching: Comparable treated and control groups (3.3)

- “Treated and control groups can be comparable as whole groups, though not paired exactly.”

$$\Pr(\max_{1\leq k\leq K} |d_k| \geq t) \leq 2Ke^{\frac{-2t^2}{\frac{2}{n}}}$$

- where d_k is treated-minus-control difference in the sample proportion of positive values for covariate k
  - 2n independent people;
  - Binary treatment Z and K covariates
  - “the chance that the largest of more than twenty million measures of absolute covariate imbalance exceeds $t = 0.05$ or 5% is smaller than 0.00058”

Balance as an alternative to exact matching: Comparable treated and control groups (3.3)

- “Only K=2 covariates (r_T, r_C) matter in in a randomized trial when estimating the average effect of a treatment on a binary response.”

$$\Pr(|\overline{r_T} - \overline{r_c} - \tau|) \leq 2e^{\frac{-2t^2}{\frac{2}{n}}}$$

where $\tau = E(\overline{r_T} - \overline{r_c})$

“so $\Pr(|\overline{r_T} - \overline{r_c} - \tau| \geq 0.02) \leq 0.037$ which is sharpened to $\Pr(|\overline{r_T} - \overline{r_c} - \tau| \geq 0.02) \leq 0.0047$ using the Normal approximation to the binomial with maximum variance 1/4.”

Propensity score and principal unobserved covariate (3.4, 3.7, 3.8)

- Propensity score

  Definition: $0 \leq \lambda(x) = \Pr(Z = 1|x) \leq 1$

  Balancing Property: $Z \perp x \mid \{\lambda(x)\ or\ \rho(x)\}$ Dawid (1979, Theorem 3.1) two vertical line

  Other function: $\rho(x)$ (matching exactly for *x* when $\rho(x) = x$)

  Treated and control individuals with the same propensity score have the same distribution of *x*

- Principal unobserved covariate Frangakis & Rubin (2002)

  Definition: $0 \leq u = \zeta(x, r_T, r_C) = \Pr(Z = 1|x, r_T, r_C) = \Pr(Z = 1|\zeta) \leq 1$

  Property: Theorem 1 in Rosenbaum & Rubin (1983), Rosenbaum (2020b, § 6.3), Dawid's (1979, Lemma 4.2(ii))

  - $Z \perp (x, r_T, r_C) \mid \zeta$
  - $Z \perp (r_T, r_C) \mid (x, \zeta)$

  two vertical line

- Strongly ignorable Rosenbaum & Rubin (1983, Theorem 3)

  Definition: $\forall x, 0 < \Pr(Z = 1|x, r_T, r_C) = \Pr(Z = 1|x) < 1$

Checking covariate balance; checking covariate overlap (3.5)

- “First, the balance, or lack of balance, of observed covariates x is something that can be seen and checked, before examining outcomes, by comparing empirical distributions of x in treated and control groups after matching or stratifying for $\lambda\ (x)$ and perhaps also for any $\rho\ (x)$.”
- “One modern approach to this task compares the balance on many covariates achieved by matching to the balance achieved for the same individuals with the same covariates by complete randomization, where the actual matched sample is completely randomized 10,000 times to form an empirical distribution of randomization-based imbalances in x for comparison.” Pimentel et al. (2015, Table 1) and Yu (2021)

Checking covariate balance; checking covariate overlap (3.5)

- “Second, Cochran was concerned to distinguish adjustments for *x* from extrapolations.”
- “In brief, the issue is whether the distribution of *x* exhibits sufficient common support in treated and control groups to permit comparisons.”
- “The propensity score helps to provide a simple check, namely parallel boxplots of $\lambda(x)$ in treated and control groups. There is limited overlap in high dimensional *x* if and only if there is limited overlap in the scalar $\lambda(x)$.”

Checking covariate balance; checking covariate overlap (3.5)

- "Third, by examining the distribution of observed covariates *x* in matched treated and control groups, we may see that matching has succeeded in creating groups comparable in terms of *x*; then, that uncontroversial task may be seen to be completed before beginning the more challenging and controversial task of addressing bias from unmeasured covariates *u*. We may agree that bias from *x* has been controlled as we debate whether there are consequential biases due to an unmeasured covariate *u.*"

Balancing covariates that resist low-dimensional summarization (3.6)

- “Some covariates do not permit low-dimensional summarization. Sometimes a covariate has many nominal levels, conceptually L levels with $L \propto n$ as $n \rightarrow \infty$.”
- “For instance, in health services research, there are more than 70,000 sparsely populated ICD-10-PCS procedure codes, so even a study with several hundred thousand people will have difficulty balancing such a covariate using probability alone.“

Balancing covariates that resist low-dimensional summarization (3.6)

- “In this case, propensity scores can be supplemented by “fine balance,” a form of constrained optimization that forces the maximum possible balance in each of the *L* categories without constraining who is matched to whom, while also pairing closely for the propensity score and other covariates. ”

- “Fine balance is readily implemented using either network optimization, as in Pimentel et al. (2015) and Rosenbaum (2002, § 10.4.6; 2020a, ch. 11; 305 2020b, § 4.4), or mixed integer programming, as in Zubizarreta (2012).”

- “Morgan & Rubin (2012) discuss an analogous situation in the design of a randomized experiment, obtaining better balance for a covariate than is afforded by complete randomization. ”

Sensitivity analysis and study design: Planning to achieve insensitivity (3.9)

- Principal *odds* of treatment

  Definition: $o_{Z=1}(x, r_T, r_C) = \frac{\Pr(Z = 1|x, r_T, r_C)}{\Pr(Z = 0|x, r_T, r_C)} = \frac{u}{1-u} = \frac{\zeta(x, r_T, r_C)}{1-\zeta(x, r_T, r_C)}$

- Sensitivity parameter Γ Rosenbaum (1987; 2002, ch. 4; 2020a, ch. 3), Rosenbaum (2020a, § 3.6)

  Definition: $\frac{1}{\Gamma} \leq \frac{o_{Z=1}(x, r_T, r_C)}{o_{Z=1}(x, r_T', r_C')} \leq \Gamma$

  Explanation: “two parameters describing the impact of unobserved covariates on treatment assignment *Z* and on potential outcomes (r_T, r_C) given *x*”

Sensitivity analysis and study design: Planning to achieve insensitivity (3.9)

- “Studies may be *designed* to be insensitive to larger unmeasured biases, as quantified by Γ” Rosenbaum (2004; 2020a, Part III), Stuart & Hanna (2013) and Zubizarreta et al. (2013)

- “Briefly, unmeasured bias is, of course, unmeasured; however, sensitivity to unmeasured bias is something computed from observed data, and hence it is a property of observable distributions”

- “Change the observable distributions by altering the study design and one changes the sensitivity to unmeasured bias”
  - “Dose schedules, unit heterogeneity, analytical plans that take account of either coherence among multiple outcomes or heterogeneous treatment effects”

Sensitivity analysis and study design: Planning to achieve insensitivity (3.9)

- Design sensitivity $\tilde{\Gamma}$

  "the limiting sensitivity to unmeasured bias as the sample size increases within a particular study design"

  "Different study designs have different design sensitivities, $\tilde{\Gamma}$, that may be compared when planning an observational study"

- Bahadur (1971) efficiency Rosenbaum (2015), Ertefaie et al. (2018), Karmakar et al. (2019), and Heng et al. (2020)
  - $\lim_{\Gamma \to \tilde{\Gamma}} BE = 0$
  - "For instance, weak instruments — instruments that gently encourage individuals to accept treatment — are invariably sensitive to small departures from random assignment of encouragement. In that context, the Bahadur efficiency of a sensitivity analysis guides efforts to strengthen the instrument by depicting the trade-off between instrument strength versus sample size, as discussed in Ertefaie et al. (2018)."

Conclusion: design is key, and propensity scores simplify design (4)

- “Propensity scores are one of several tools useful for ***balancing observed covariates*** when designing an observational study.”

- “When random assignment of treatments is infeasible or unethical, the focus of attention should be on the design of observational studies that support and strengthen the ***crucial step from association to causation***” Cochran (1965), Imbens & Rubin (2015, Parts V and VI), Rosenbaum (2002, ch. 4-11; 2020a, ch. 3–7, 15–21; 2021), and Stuart & Rubin (2008a,b)

Steps to causation in the design of an observational study

- “Selecting **comparisons** to increase the design sensitivity”
- “Seeking opportunities to **detect bias”**
- “Seeking **mutually** supportive evidence affected by different biases”
- “Incorporating **quasi-experimental** devices such as multiple control groups”
- “Economist’s **instruments”**

Some questions from siyu

- Under what kind of conditions, the principal unobserved covariate will be equal to the propensity score? (strong ignorable)
- What’s the basic idea and procedure to do sensitivity analysis?
- When I contact the propensity score analysis about 3000+ sample, I have to exclude about 100 samples which did not have complete cases of predictors. Will the exclusion lead to selection bias and how to attenuate the potential selection bias?

Some questions from yan

- ***Unbiasedness***: Why we do not use ***unbiasedness*** to attain the confidence of prior for a Bayesian model in the observational study with few experiments? The connection between unbiasedness and randomization?
- ***Entanglement***: Anything can be a variable for science study? How can we know the statistic of ‘age’ is not the ‘candle number on the birthday cake’?

Some questions from yan

- **Uncertainty**: Where is the randomization from in the observation study? The population/super-population is fixed or samples are fixed? Why variables do not disappear after they are fixed? Uncertainty of Sampling?
- ***Independence*** and ***manipulation***

commutative property

Consider such a process that occurred 1000 times: we toss coin A with the upside as the start side, then measure the side of coin A. Next, we use A's side as the start side of coin B and toss coin B. Then, we measure the side of coin B. If coin A was not measured, then coin B can NOT be measured. But if coin B can not be measured, coin A can STILL be measured.

A is independent of B but why does someone may think that manipulating A will influence B?

P (coin A, coin B)	A=up	A=down	A=None
B=up	0.5//0.5	0.5//0.5	0//0
B=down	0.5//0.5	0.5//0.5	0//0
B=None	0//0	0//0	0//0

Manipulation (coin A, coin B)	A=up	A=down	A=None
B=up	0.5//0.5	0.5//0.5	0//0
B=down	0.5//0.5	0.5//0.5	0//0
B=None	0//0.5	0//0.5	1//0