

Doubts and content errors in "CAUSAL INFERENCE for Statistics, Social, and Biomedical Sciences—An Introduction" by Imbens and Rubin, first edition (2015), ISBN 978-0-521-88588-1

PART II  
Classical Randomized Experiments  
#####

CHAPTER 5  
Fisher's Exact P-Values for Completely Randomized Experiments  
=====

Page 67:

Given the N ranks  $R_i$ ,  $i = 1, \dots, N$ , an obvious test statistic is the absolute value of the difference in average ranks for treated and control units:

$$Trank = |\bar{R}_t - \bar{R}_c| = |\sum_{i:W_i=1} R_i / N_t - \sum_{i:W_i=0} R_i / N_c|, \quad (5.7)$$

where  $\bar{R}_t$  and  $\bar{R}_c$  are the average rank in the treatment and control group respectively. In the absence of ties, the p-value for this test statistic is closely related to that based on the Wilcoxon rank sum test statistic, which is defined as  $T_{Wilcoxon} = \sum_{i=1}^N \tilde{R}_i$ , because Trank is a simple

transformation of  $T_{Wilcoxon}$ :

$$Trank = |(\underbrace{T_{Wilcoxon}} - N(N + 1)/2) / N_t - (N(N - 1)/2 - \underbrace{T_{Wilcoxon}}) / N_c|.$$

I think it should define  $T_{Wilcoxon}$  separately for the two groups.

CHAPTER 8  
Model-Based Inference for Completely Randomized Experiments  
=====

Page 145:

Note that annual earnings for these men are very low, even for those years; when we average only over those with positive earnings, average annual earnings in 1978 are on the order of only approximately \$8,000 after the program. Prior to the program, earnings are even lower, partly because low earnings in 1978 were a component for determining eligibility.

~~~~~ 1978 is post treatment, shouldn't the units have been selected based on pre-treatment earnings?  
Probably a typo

CHAPTER 9  
Stratified Randomized Experiments  
=====

Page 199, end of section:

Thus there is only limited evidence against the null hypothesis that the variation in average scores differs between small and regular-sized classes.

~~~~~ "is equal", or possibly put commas  
around "against the null hypothesis"

PART III

Regular Assignment Mechanisms: Design

#####

CHAPTER 12

Unconfounded Treatment Assignment

=====

Page 261, section 12.2.3:

A key feature of the unconfoundedness assumption is that it has no directly testable implications, even in settings with a large number of units. There is no information in the data that can tell us that unconfoundedness does not hold. Of course this does not mean that unconfoundedness actually holds, or even that it is plausible, but it implies that any assertion that it does not hold must rely on additional, substantive, information beyond

~~~~~  
the assessment of assumptions of probabilistic and individualistic assignment.

I'm not sure I understand. Is the "not" a mistake? Or is it really making the argument that, the same way you need out of sample information to assert unconfoundedness, you need it to deny it, to defend unconfoundedness?

This would be a weak argument because assuming unconfoundedness is smaller measure than not assuming it.

.....

Page 266, end of section:

Variables that are truly instrumental variables are relatively rare, and when they exist, it is even more rare that they are mistakenly used as

~~~~~  
covariates for adjustment.

I think adjusting for an IV does not give wrong results, due to the IV being randomized. So here there is an ambiguity: does "mistaken" imply that the causal interpretation is ruined, or that you could have done better by recognizing the nature of the IV variable and doing an IV analysis?

A possible interpretation I've seen people take is that by saying "IV" you refer to the whole package of IV variable + unobserved confounding, in which case adjusting just for the IV variable is not sufficient for identification.

CHAPTER 13

Estimating the Propensity Score

=====

Page 285, table 13.1: it seems weird to me that "lpbc415, First pregnancy complication index" is balanced while "lpbc420, Second pregnancy complication index" is very unbalanced, even though I don't know what they mean precisely.

.....

It shows that there are, if anything, slightly fewer large negative values and fewer large positive values than one would expect to see if the z-values were independent draws from a normal distribution.

Figure 13.2 with:

Comparing Figure 13.2 to Figure 13.1, it is clear that including some second-order terms leads to substantially better balance in the covariates, supporting the importance of doing a careful assessment of the adequacy of the propensity score specification by inspecting covariate balance.

And Figure 13.3 with:

Comparing Figure 13.3 to Figure 13.1, it appears that the lasso does not lead to as good an in-sample fit as our proposed specification, possibly due to its focus on out-of-sample prediction.

I think they switched the axis labels in all figures, because when the z-values appear more concentrated toward 0, in the text they say there's less balance.

CHAPTER 14  
Assessing Overlap in Covariate Distributions

=====

Pages 323–324, figures: maybe the QQ plot axis labels are swapped like in figures 13.1–13.3, since the z-values look more dispersed than the Normal distribution.

CHAPTER 15  
Matching to Improve Balance in Covariate Distributions

=====

Page 348, first paragraph:

Then  $(\mu_t - \mu_c)\beta$  differs from zero (so the bias after matching is non-zero), whereas the bias before matching was  $(\mu_t - \mu_c)\beta = 0$ . Hence matching method B has made the bias for this linear combination infinitely worse.

The expression "infinitely worse" is technically correct in a specific sense but conveys a hyperbole. I care about the bias on some absolute scale, not so much about the ratio of the bias between two methods.

Second paragraph:

For both Mahalanobis and propensity score matching, the matching results are invariant to affine linear transformations of the covariates, so whether we match on  $X_i$  or  $Z_i$  is irrelevant.

I guess they are invariant only if the propensity score model is a GLM.

.....

Page 356:

the treated unit with the smallest value of the propensity score (0.00).

I guess this is not really 0, show more digits.

.....

Page 358:

It is interesting to note that the worst matches are not simply for the units with the largest value of the propensity score.

This refers to Table 15.7, which reports that the worst match has propensity score 0.97. Page 356 says:

[...] first, the match for the treated unit with the largest value for the propensity score (0.97); second, the match for the treated unit with the median value of the propensity score (0.36);

The propensity scores in Table 15.7 range from 0.79 to 0.97, so the statement above does not sound right in a meaningful sense.

PART IV

Regular Assignment Mechanisms: Analysis

#####

CHAPTER 17

Subclassification on the Propensity Score

=====

Pages 389–390: I think the definitions of  $\Delta$  and  $\Gamma$  have been inverted. Consider the units of measurement of the  $\beta$ - $\beta$  covariance matrix block, which ought to be  $(Y/X)^2$  since  $\beta$  has units  $Y/X$ , where here  $Y$  and  $X$  stand for the units of measurement of outcome and covariates.

The  $\beta$ - $\beta$  block is the bottom right one, so the units of  $\Delta$  are  $X^2$  and those of  $\Gamma$  are  $Y^2X^2$ . Thus  $(\Gamma\Delta^{-1}\Gamma)^{-1}$  has units  $Y^{-4}X^{-2}$ , while  $(\Delta\Gamma^{-1}\Delta)^{-1}$  has  $Y^2X^{-2}$  as expected.

CHAPTER 19

A General Method for Estimating Sampling Variances for Standard Estimators for Average Causal Effects

=====

Page 441:

However, our view is that, in general, one should focus on the sampling variance of an estimator viewed as an estimator of the sample average effect rather than viewed as an estimator of the super-population average effect. Thus we recommend focusing on the generalization of (19.5), rather than taking into account differences between the distribution of the pre-treatment variables in the sample and the analogous distribution in a somewhat vague, hypothetical, and often ill-defined, super-population.

Why do they think so? About the ill-definition of the super-population: using an uncertainty perspective on the probability distributions instead of imagining an infinite super-population solves this non-issue, the super-population effect becomes a prediction about a single unobserved unit.

Page 486, section 21.4.2:

Because condition (21.12) is strictly stronger than unconfoundedness, (21.1),  
 ~~~~~ 21.7 ~~~~~ 21.6

Below equation 21.10:

The point is that it is difficult to envision a situation where unconfoundedness based on the two comparison groups (21.6) holds, but using only one of the two comparison groups the unconfoundedness condition fails (i.e., neither (21.9) nor (21.10) holds).

This is backwards: 21.6 implies mathematically 21.9 & 21.10. The correct sentence is, I guess: "The point is that it is difficult to envision a situation where unconfoundedness holds separately for the two control groups (i.e., (21.9) and (21.10) hold), but using them together the unconfoundedness condition (21.6) fails."

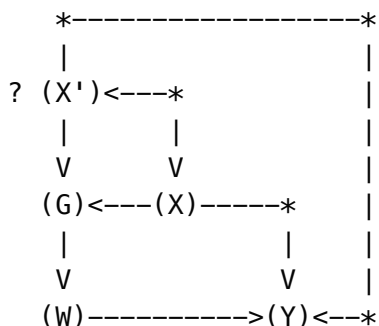
Next sentence:

So, in practice, if unconfoundedness holds, typically also the stronger  
 ~~~ Analogously  
 condition (21.6) would hold,  
 ~~~~~ 21.7

Next sentence:

Again, there is no theorem here, but an implication that when stronger  
 ~~~~~  
 conditional independence assumptions are false, weaker conditional independence assumptions are more likely also to be false.

I think a graph here helps:



Here X' is additional hypothetical confounders. By design, W depends only on G, there's not an arrow X --> W, neither X' --> W. The absence of X' --> W implies that, to make W ignorable given X, we necessarily get G ignorable given X.

.....

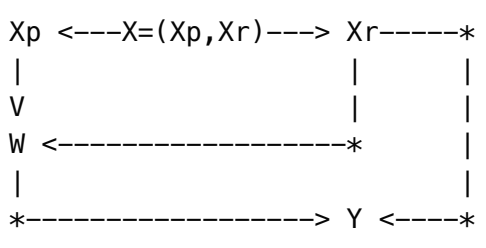
Page 489, first two equations: it seems to me that, to go from the first to the second equation, the following equality is used:

$$f_{Xi|Wi=w, Xir=xr}(x|w, xr) = f_{Xi|Xir=xr}(x|xr)$$

i.e.,

$$W \perp\!\!\!\perp X \mid X_r$$

which, however, is not implied by the assumptions of unconfoundedness and subset unconfoundedness. As a counterexample, consider the following causal graph:



This satisfies both  $W \perp\!\!\!\perp Y(w) \mid X$  and  $W \perp\!\!\!\perp Y(w) \mid X_r$ , but not  $W \perp\!\!\!\perp X \mid X_r$ .

### PART VI

#### Regular Assignment Mechanisms with Noncompliance: Analysis

#####

### CHAPTER 23

#### Instrumental Variables Analysis of Randomized Experiments with One-Sided Noncompliance

=====

Page 527, near the bottom:

Specifically, suppose a drug company estimates the efficacy of a new drug in a randomized trial. Implicitly the assumption is that, had, at the start of the trial, all individuals been told that they would receive the new active drug and that no one would receive the control treatment, the typical outcome would have been approximately the same as the typical outcome observed in the subsample actually assigned to the treatment.

This is backwards: it says that the drug is assumed to be placebo; assuming "told" means "only tell them so, without actually changing the assignments."

.....

Page 530:

For noncompliers, there would be no information in the data regarding the effect of the primary treatment on the outcome, because no noncomplier ever receives the active treatment. The data from noncompliers would therefore be discarded because of the absence of units who received the active treatment.

It would not be discarded if the analysis is done with a model, e.g., with a fully Bayesian inference.

.....

Page 538, first line:

this expression is equal to  $E[Y_i(0) | G_i = co] \cdot \pi_{co} - E[Y_i(0) | G_i = nc] \cdot \pi_{nc}$ ;  
~~~~~ +

First equation, first line:

$$\tau_{pp} = E[Y_i(1) - Y_i(0) | G_i = co] \cdot \pi_{co} + \underbrace{(E[Y_i(0) | G_i = co] - E[Y_i(0) | G_i = nc])}_{\approx 1} \cdot \pi_{nc}$$

First equation, second line:

$$= ITTY_{,co} + \pi_{nc} \cdot \underbrace{\Delta_{co,nc}}_{(E[Y_i(1) | G_i = co] - E[Y_i(0) | G_i = nc])}$$

( $\Delta_{co,nc} = E_{sp}[Y_i(0) | G_i = co] - E_{sp}[Y_i(0) | G_i = nc]$ , with  $Y(0)$  instead of  $Y(1)$  in the first expected value.)

.....

Page 539, end of section:

Survival rates for compliers assigned to the control treatment are substantially higher than for noncompliers assigned the active treatment, despite the fact that neither group took any active treatment.

I would add a sentence saying that  $\tau_{co} = \tau_{cu} + \Delta_{co,nc}$ , and recall the value of  $\tau_{co}$  (0.0033).

#### CHAPTER 24

#### Instrumental Variables Analysis of Randomized Experiments with Two-Sided Noncompliance

=====

Page 555, end of section:

Thus, earnings for compliers who do not serve appear to be substantially higher than earnings for nevertakers, but compliers who serve in the military appear to have earnings comparable to those of alwaystakers.

I wouldn't do comparisons before estimating uncertainties.

.....

Page 559: here I would mention Pearl's chapter 8.

#### CHAPTER 25

#### Model-Based Analysis in Instrumental Variable Settings: Randomized Experiments with Two-Sided Noncompliance

=====

Page 571: the conditional distribution of  $G_i$  is written in a way that makes it difficult to write a prior that is symmetric under permutation of  $G$ . I would put  $\exp(X_i \cdot \gamma_{co})$  in the first numerator. The "artificial observations" prior (defined in section 25.6.3, page 579) does not induce automatically this property, it depends on the parametrization.

.....

Page 575:

(For nevertakers and alwaystakers the distribution of  $Y_i(0)$  is identical to

that of  $Y_i(1)$  by the two exclusion restrictions, so we do not index  $\beta_{nt}$  and  $\beta_{at}$  by the treatment received.)

Fine choice, wrong motivation: for NT and AT, one of the potential outcomes is undefined (see section 25.4.1), so without loss of generality we can give it whatever distribution we find convenient to write the model, in particular the same as the defined one.

.....

Page 583:

Interestingly, the posterior probability that  $\tau_{late}$  is exactly equal to zero is 0.08,  
~~~~~ too large

See the histogram in Figure 25.1: the area of the bin containing 0 is an upper bound on  $P(\tau_{late}=0)$ , and is much smaller than 0.08. Maybe its 0.08 percent?