Biostat M235 - Causal Inference - Assignment 1 2025 Spring

Hiroyasu Ando†

Department of Biostatistics, University of California, Los Angeles, California, 90095, USA.

1. Referee Report

1.1. Summary

In the fields of social science and medicine, researchers generally have great trust in randomized controlled trials (RCTs). The authors warn against this, and state the importance of clarifying what we are trying to discover and prior knowledge in causal inference.

At best, RCTs can only help to obtain unbiased estimates within a sample, and any expansion from that point requires reasonable justification. Another disadvantage is the potential large cost. From these points of view, it sometimes makes sense to use an observational study after controlling for confounders and other factors.

The advantage of RCTs is that they require less assumption and less prior knowledge, and they are good at proving opinions that differ from those of the audience. However, this is an attitude of abandoning scientific discoveries that have been made up to now, and is far removed from the ideal of science, which is to build discoveries on top of prior knowledge.

RCTs should be used as a tool for accumulating scientific discoveries, and they can be used in combination with conceptual and theoretical methods. As a result, it is possible to find out "why things work" rather than "what works" in scientific events.

1.2. Main Comments

According to the authors (Deaton and Cartwright, 2018), randomization makes it possible to obtain unbiased ATE without covariates, confounders, or other causes. This is based on the assumption that the above mentioned factors do not change after randomization. As a result, it is expected that the trial will be balanced when the sample size is large.

In my opinion, this post-randomization difference is a point that deserves further attention. For example, it was recently discovered that post-randomization differences occur over time in survival analyses in clinical trials of infectious diseases (Kahn et al., 2018; O'Hagan et al., 2014). It has also been noted that stratification is beneficial to mitigate this problem (Kahn et al., 2018; O'Hagan et al., 2014). This statement is consistent with that of the authors (Deaton and Cartwright, 2018).

†*Address for correspondence:* Hiroyasu Ando, Department of Biostatistics, University of California, Los Angeles, California, 90095, USA. Email: hiro1999@ucla.edu

The authors (Deaton and Cartwright, 2018) also note that the results obtained by the RCTs are for a trial sample and should be applied with caution to the population as a whole. For example, caution should be exercised when performing simple extrapolation and simple generalization from results derived in artificial environments. To remedy this, they advocate, for example, performing pre-experimental stratification and understanding causal structure and prior information.

In my view, we need to pay close attention to these extrapolations and generalizations, especially in the social sciences. For example, in the field of evolutionary game theory, there are many attempts to observe and generalize human behavior in artificial environments or game environments (Nishi et al., 2015). This deduction from internal validity to external validity needs to be done carefully. This is because findings in the social sciences are expected to be later introduced into social policy and other areas. From this perspective, evolutionary game theory, in particular, requires the pursuit of consistency between the causal structure in the game and the causal structure in the real world.

1.3. Minor Comments

In the following, I will discuss what I found particularly interesting in the description of experimental designs, analytical aspects, and extrapolation of results.

1.3.1. experimental designs and analytical aspects

First, ATEs have the remarkable advantage that no model is required, no assumptions about covariates, confounders, or other causes are needed (Deaton and Cartwright, 2018). However, we must note the assumption that the mean is a linear operator (Deaton and Cartwright, 2018). This is a point that should be kept in mind by those who blindly believe in RTCs and ATEs.

Also, the statement that RTCs do not always achieve perfect balance and should be tested for the balance further should be emphasized especially in the social sciences and in the medical field (Deaton and Cartwright, 2018). From this point of view, it is meaningful to search for and widely use the prior knowledge of the other causes. Also, it is worth noting that achieving a balanced covariates requires not only the blinding of trial participants, but also the blinding of trial investigators and the handling of outliers.

1.3.2. extrapolation of results

In scaling up RCTs, or extrapolation of results, it is necessary to examine whether the assumption that the equilibrating variable is constant is valid (Deaton and Cartwright, 2018). This explains the importance of understanding causal mechanisms.

In addition, ATE is an average treatment effect and cannot be said to apply to each individual (Deaton and Cartwright, 2018). Therefore, just because the claim that ATE is non-zero is not significant does not mean that the treatment does not work for everyone (Deaton and Cartwright, 2018). In particular, it can be argued that in medicine, well-designed RCTs need to be used in an environment that is close to the popoulation of interest.

2. Computational Problem

2.1. Fisher

(***a***)

- a measurement of cholesterol level (X)

  Based on the wilcoxson's rank sum test, we can see that the p-value is 0.937, which is greater than 0.05. This indicates that there is no significant difference in the measurement of cholesterol level (X) between the control and treatment groups.

- sex at birth (S)

  Based on the fisher's exact test, we can see that the p-value is 1.000, which is greater than 0.05. This indicates that there is no significant difference in the sex at birth (S) between the control and treatment groups.

(***b***)

$$H_0 : Y_i(1) = Y_i(0) \text{ for } i = 1, 2, 3, 4, 5, 6$$

- fisher p-values for sample averages

$$\text{test statistics} = |\bar{Y(1)} - \bar{Y(0)}|$$

  The test statistic is 36.667. The p-value is 0.093. Thus, we cannot reject the null hypothesis at the 0.05 significance level.

- fisher p-values for sample median

$$\text{test statistics} = |\text{medianY}(1) - \text{medianY}(0)|$$

  The test statistic is 30.500. The p-value is 0.199. Thus, we cannot reject the null hypothesis at the 0.05 significance level.

- fisher p-values for sample variance

$$\text{test statistics} = |\text{varY}(1) - \text{varY}(0)|$$

  The test statistic is 636.267. The p-value is 0.788. Thus, we cannot reject the null hypothesis at the 0.05 significance level.

(***c***)

The p-values became smaller than those of (b). This is because the block randomization reduces the variance, leading to the precise estimates.

- fisher p-values for sample averages

  The test statistic is 73.333. The p-value is 0.065. Thus, we cannot reject the null hypothesis at the 0.05 significance level.

- fisher p-values for sample median

  The test statistic is 66.000. The p-value is 0.08. Thus, we cannot reject the null hypothesis at the 0.05 significance level.

- fisher p-values for sample variance

  The test statistic is 1216.667. The p-value is 0.690. Thus, we cannot reject the null hypothesis at the 0.05 significance level.

(***d−b***)

The gain scores were defined as $Y - X$, which is controlling for the baseline value of X. Since this method is expected to improve the efficiency, we can see that the p-values are lower than those in (b). This is what I would have expected to see.

- fisher p-values for sample averages

  The test statistic is 38.000. The p-value is 0.004. Thus, we can reject the null hypothesis at the 0.05 significance level.

- fisher p-values for sample median

  The test statistic is 43.000. The p-value is 0.013. Thus, we can reject the null hypothesis at the 0.05 significance level.

- fisher p-values for sample variance

  The test statistic is 448.800. The p-value is 0.303. Thus, we cannot reject the null hypothesis at the 0.05 significance level.

(***d−c***)

- fisher p-values for sample averages

  The test statistic is 76.000. The p-value is 0.005. Thus, we can reject the null hypothesis at the 0.05 significance level.

- fisher p-values for sample median

  The test statistic is 86.000. The p-value is 0.020. Thus, we can reject the null hypothesis at the 0.05 significance level.

- fisher p-values for sample variance

  The test statistic is 1125. The p-value is 0.33. Thus, we cannot reject the null hypothesis at the 0.05 significance level.

2.2. Neyman

(***a***)

- $\bar{Y(1)} - \bar{Y(0)} = -36.667$

(***b***)

- ATE_{FP}

  Considering treatment effect is additive, $sd(Y(1))^2 = 857.9$, $sd(Y(0))^2 = 1494.167$, and $\hat{var}(ATE_{FP}) = 392.011$. So, C.I. is $(-69.236, -4.097)$.

- ATE_{SP}

  Considering a simple random sample from a larger population of size S, which is infinitely large, $sd(Y(1))^2 = 857.9$, $sd(Y(0))^2 = 1494.167$, and $\hat{var}(ATE_{SP}) = 392.011$. So, C.I. is $(-69.236, -4.097)$.

(***c***)

- model with W

  The point estimate of the treatment effect is -36.670, which is almost the same as the Neyman's one. The variance is also similar to the Neyman's one, as the variance of the treatment effect is 392.040.

- model with W and X

  The point estimate of the treatment effect is -37.733, which is almost the same as the Neyman's one. The variance is smaller than the Neyman's one, as the variance of the treatment effect is 65.871. It is considered that controlling for the baselines led to the more precise estimates.

(***d***)

Compared to (b), the confidence interval is narrower.

- ATE_{FP}

  Considering treatment effect is additive, $sd(Y(1))^2 = 464.667$, $sd(Y(0))^2 = 15.867$, and $\hat{var}(ATE_{SP}) = 80.088$. So, C.I. is $(-52.722 - 23.279)$.

- ATE_{SP}

  Considering a simple random sample from a larger population of size S, which is infinitely large, $sd(Y(1))^2 = 464.667$, $sd(Y(0))^2 = 15.867$, and $\hat{var}(ATE_{SP}) = 80.088$. So, C.I. is $(-52.722 - 23.279)$.

2.3. Fisher VS Neyman

- The Fisher's one focuses on testing sharp null hypotheses. On the other hand, the Neyman's one focuses on the estimation of average causal effects. In this sense, the Neyman's one is more flexible than the Fisher's one. However, you have to be careful that even if the average causal is zero, it does not mean that the treatment has no effect on each individual.

- The Fisher's one does not require the asymptotic assumption and can handle flexible test statistics. On the other hand, the Neyman's one requires the asymptotic assumption and can only handle the test statistics that are asymptotically normal. Also the Fisher's one might be less powerful than the Neyman's one.
- The Fisher's one might be computationally intensive, as it requires the permutation test, compared to the Neyman's one, which is computationally less intensive.

3. Conceptual Problem

(***a***)

- bernoulli trial

(***b***)

- pros

  It's easy to understand the model mechanism and to implement it. It can be used to model various situations with binary outcomes.

- cons

  It's less informative than other models in terms of the balance between the number of treated and control units. It's less efficient than other models since it does not take into account the covariates that may affect the outcome. It might not be able to achieve the desired balance between the treated and control groups, especially when the sample size is small.

(***c***)

- fisher's exact test

(***d***)

- pros

  The Fisher's one does not require the asymptotic assumption and can handle flexible test statistics.

- cons

  The sharp null hypotheses are less informative. The Fisher's one might be computationally intensive, as it requires the permutation test. The Fisher's one might be less powerful.

4. Disclosure Statement

The authors report there are no competing interests to declare.

5. Data and Software Availability

All analysis was done in the R environment (R version 4.3.2). The code and available data to reconstruct the analyses of this paper are available at https://github.com/Ankoudon/biostat235.

References

Deaton, A. and N. Cartwright (2018). Understanding and misunderstanding randomized controlled trials. *Social Science Medicine 210*, 2–21.

Kahn, R., M. Hitchings, S. Bellan, and et al. (2018). Impact of stochastically generated heterogeneity in hazard rates on individually randomized vaccine efficacy trials. *Clinical Trials 15*(2), 207–211.

Nishi, A., H. Shirado, D. G. Rand, and N. A. Christakis (2015). Inequality and visibility of wealth in experimental social networks. *Nature 526*(7573), 426–429.

O'Hagan, J., M. Lipsitch, and M. Hernan (2014). Estimating the per-exposure effect of infectious disease interventions. *Epidemiology 25*(1), 134—-138.