Contents

© Art Owen 2020 do not distribute or post electronically without author's permission

1

Introduction

To gain understanding from data, we must contend with noise, bias, correlation and interaction among other issues. Often there is nothing we can do about some of those things, because we are just handed data with flaws embedded. Choosing or designing the data gives us much better possibilities. By carefully designing an experiment we can gain information more efficiently, meaning lower variance for a given expenditure in time or money or subjects. More importantly, experimentation provides the most convincing empirical evidence of causality. That is, it is not just about more efficient estimation of regression coefficients and similar parameters. It is about gaining causal insight. If we think of efficiency as better handling of noise, we can think of the causal estimation as better handling of correlations among predictors as well as interactions and bias.

We all know that "correlation does not imply causation". Without a causal understanding, all we can do is predict outcomes, not confidently influence them. There are settings where prediction alone is very useful. Predicting the path of a hurricane is enough to help people get out of the way and prepare for the aftermath. Predicting stock prices is useful for an investor whose decisions are too small to move the market. However, much greater benefits are available from causal understanding. For instance, a physician who could only predict patient outcomes in the absence of treatment but not influence them would not be as effective as one who can choose a treatment that brings a causal benefit. In manufacturing, causal understanding is needed to design better products. In social science, causal understanding is needed to understand policy choices.

Our main tool will be randomizing treatment assignments. Injecting randomness into the independent variables provides the most convincing way to establish causality, though we will see that it is not perfect.

Experimental design is the science of choosing how to gather data. It has a long and continuing history spanning: agriculture, medicine and public health, education, simulation, engineering, computer experiments, A/B testing in e-commerce, philanthropy and more.

The design problem forces the statistical investigator to think carefully about the underlying domain topic, it's assumptions and costs, benefits, goals and prior history. This happens in ordinary data analysis too, but the task of choosing what data to gather amplifies the problem. More than other areas of statistics, you really need to have a some use case in mind to understand how to interpret variables models and estimators. Predictors that you can change are different from ones determined by a user or the environment set. Response variables that can be cheaply measured are different from ultimate ones. Design is often sequential, so things we learn at one stage help at the next. There are choices between safe but perhaps expensive experiments and risky but perhaps faster or cheaper experiments.

1.1 History of design

Experimental design has been systematically studied for about 100 years or more. There are many recent innovations coming from online A/B testing as well as new developments in clinical trials. We will look at new innovations but not ignore the older ones that they evolved from.

Fisher and others working on agriculture at Rothamsted starting in the 1920s. Box working on 2^{k-p} factorial experiments for industry starting in the 1960s. Computer experiments from Sacks, Welch, Wynn, Ylvisaker and others starting in the 1980s. Realization that Monte Carlo simulations are (or should be) desiged experiments starting in the 1970s. Online A/B testing in electronic commerce starting in the 2010s. (Kohavi, Tang, Xu and many others). Experiments on networks. Experiments in economics for philanthropy.

1.2 Confounding and related issues

We will begin by comparing outcomes for subjects given one of two levels. Sometimes it makes sense to call them 'treatment' versus 'control'. Control could be a default or usual or null setting. Other times the two levels are on the same footing: e.g., raspberry vs blueberry or Harvard vs Yale.

If the treatment subjects are measured in the morning and control in the afternoon, then the difference we measure is the joint effect of treatment-AM versus control-PM and any difference we measure might be partly or largely due to time of day. The treatment is ***confounded*** with time of day.

Confounding can happen easily. Maybe treatment and control were done at different labs or by different people in the same lab or on different equipment by the same person in a given lab. Treatment and control could also be confounded with some variable that we didn't measure or don't even know about.

© Art Owen 2020 do not distribute or post electronically without author's permission

A common source of confounding is the use of ***historical controls***. A physician might compare a survival rate in their clinic to what was the historical norm for that or some other clinic. The standard of care may have changed and then the new treatment is confounded with the time period of study. A better comparison would include concurrent controls. One ethical standard there is ***equipoise***. The physician should be genuinely uncertain about which treatment is better in order to justify doing an experiment. Cox (1958) has an example where historical controls would have supported the Lamarkian theory that learned effects are inherited by offspring. Each generation of rats performed better at a certain task. Because there was no control group, an alternative explanation is that something else about those animals or their condition was changing with time.

Confounding might not matter. Maybe there is sufficient scientific knowledge to know that the confounding variable could not affect the response. Of course that knowledge might be subject to uncertainty and debate. If we use some randomization to assign treatment versus control then we can statistically control the confounding making extreme confounding have negligible probability.

Later on in settings with k binary treatments and a budget that does not allow running all 2^k cases we will indulge in some purposeful confounding. The trick will be to control what is confounded with what.

Confounding induces a correlation of 1 (or -1) between our treatment and some other variable. When the treatment or that other variable varies continuously then we get a nonzero correlation between our treatment variable and the confounder. So perfect confounding is a special case of correlation. Randomizing the treatment will make the expected value of that correlation zero.

The confounder might be known and measured or it might be unknown (which is worse). When it is unknown, it is a missing predictor problem. A regression that is missing one or more of the predictors leads to biased estimates of the coefficient vector β. If our treatment variable is uncorrelated with that missing variable then we can put that variable into the regression error term. [Working this out might become an exercise.]

Another reason to randomly control the treatment is to put sufficient variance into its values. For a continuous variable we know that the variance of a regression coefficient is reduced by varying the corresponding predictor over a larger range.

1.3 Neyman-Rubin Causal Model

We will use the Neyman-Rubin framework to think about causality. See the book by Imbens and Rubin (2015) for full details. Begin with Lecture 1 in Wager's course notes in the class web page. For experimental unit i (e.g., a subject) we think that their response would have been Y_{i1} if treated and Y_{i0} if not treated. There is a variable $W_i \in \{0, 1\}$ with $W_i = 1$ for treated and $W_i = 0$ for control. The pair (Y_{i0}, Y_{i1}) contains the two ***potential outcomes***

© Art Owen 2020 do not distribute or post electronically without author's permission

for subject i. In this setting we get

$$Y_i = W_i Y_{i1} + (1 - W_i) Y_{i0}, \quad i = 1, \ldots, n.$$

We can arrange all of these outcomes into a ***science table*** $\mathcal{Y} \in \mathbb{R}^{n \times 2}$ as follows:

$$\mathcal{Y} = \begin{array}{c} \\ i{=}1 \\ 2 \\ 3 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cc} W_i{=}0 & W_i{=}1 \end{array} \\ \begin{bmatrix} Y_{10} & Y_{11} \\ Y_{20} & Y_{21} \\ Y_{30} & Y_{31} \\ \vdots & \vdots \\ Y_{i0} & Y_{i1} \\ \vdots & \vdots \\ Y_{n0} & Y_{n1} \end{bmatrix} \end{array} \tag{1.1}$$

Row i shows data for subject i. There is a column each for treatment and control. If we knew $\mathcal{Y}$ then we would know every subject's own personal treatment effect $\Delta_i = Y_{i1} - Y_{i0}$.

There is an important implicit assumption involved in writing the science table this way. We are assuming that the response for subject i is $Y_i = Y_i(W_i)$ and does not depend on $W_{i'}$ for any $i' \neq i$. Imagine instead the opposite where the value Y_i depends on the whole vector $\boldsymbol{W} = (W_1, \ldots, W_n) \in \{0, 1\}^n$. Then our science table would have n rows and 2^n columns, one for each possible $\boldsymbol{W}$.

We can use the small science table in (1.1) under the Stable Unit Treatment Value Assumption (***SUTVA***). Under SUTVA, Y_i does not depend on $W_{i'}$ for any $i' \neq i$. It might not even depend on W_i, but if it depends on $\boldsymbol{W}$ at all it can only be through W_i. In applications we have to consider whether SUTVA is realistic. If patients in a clinical trial swap drugs with each other, SUTVA is violated. If we are experimenting on subjects in a network we might find that the response for one subject depends on the treatment of their neighbors. That would violate SUTVA.

1.4 Random assignment and ATE

Very often the thing we are most interested in is the average treatment effect (ATE)

$$\tau = \frac{1}{n}\sum_{i=1}^{n} \Delta_i = \frac{1}{n}\sum_{i=1}^{n} (Y_{i1} - Y_{i0}) = \frac{1}{n}\sum_{i=1}^{n} Y_{i1} - \frac{1}{n}\sum_{i=1}^{n} Y_{i0}.$$

No matter what assignment vector $\boldsymbol{W}$ we choose, we never see any of the Δ_i. This is sometimes called the fundamental problem of causal inference. Let's write

$$\tau = \mathbb{E}(Y_i(1)) - \mathbb{E}(Y_i(0))$$

© Art Owen 2020 do not distribute or post electronically without author's permission

where in this instance $\mathbb{E}(\cdot)$ describes a simple average of n numbers. We will use $\mathbb{E}(\cdot)$ later for other things, so lets take

$$\tau = \mu_1 - \mu_0 \quad \text{where} \quad \mu_j \equiv \frac{1}{n}\sum_{i=1}^{n} Y_{ij}, \quad j = 0, 1.$$

Given $\boldsymbol{W}$ let $n_1 = \sum_{i=1}^{n} W_i$ and $n_0 = n - n_1$ be the number of treated and control subjects, respectively. If $\min(n_0, n_1) > 0$ then we can compute

$$\bar{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n} W_i Y_i \quad \text{and} \quad \bar{Y}_0 = \frac{1}{n_0}\sum_{i=1}^{n}(1 - W_i) Y_i$$

and estimate τ by

$$\hat{\tau} = \bar{Y}_1 - \bar{Y}_0.$$

By choosing $\boldsymbol{W}$ randomly we can get $\mathbb{E}(\hat{\tau}) = \tau$ where this $\mathbb{E}(\cdot)$ refers to randomness in $\boldsymbol{W}$.

Suppose we have a ***simple random sample*** where all $\binom{n}{n_1}$ ways of picking n_1 subjects to treat have equal probability. We saw in class that then

$$\mathbb{E}(\bar{Y}_j) = \mu_j, \quad j = 0, 1$$

assuming that $\min(n_0, n_1) > 0$. It then follows that

$$\mathbb{E}(\hat{\tau}) = \tau.$$

What if we just tossed independent coins taking $W_i = 1$ with probability $0 < p < 1$? Then we would get $n_1 \sim \mathrm{Bin}(n, p)$. Also, conditionally on this random n_1, the units getting $W_i = 1$ would be a simple random sample. So right away

$$\mathbb{E}(\hat{\tau} \,|\, 0 < n_1 < n) = \tau.$$

Ordinarily the event $0 < n_1 < n$ has overwhelming probability under independent sampling. Also, if anybody planned an experiment and got $n_1 \in \{0, n\}$ they would almost surely just re-randomize. [Exercise: how ok is that?]

If somehow it is necessary to analyze independent assignments without assuming that $n_1 \notin \{0, n\}$ then we can regard

$$\bar{Y}_1 = \frac{\sum_i W_i Y_i}{\sum_i W_i}$$

as a ratio of two random variables and work out approximate mean and variance via the delta-method. For the gory details see Rosenman et al. (2018).

Suppose that we want $\mathrm{var}(\bar{Y}_j)$ under simple random sampling. It's kind of a pain in the neck to work that out using the theory of finite population sampling (survey sampling) from Cochran (1977) or (Rice, 2007, Chapter 7). It is even worse if we toss coins because we hit the $n_1 \in \{0, n\}$ problem with probability just enough bigger than zero to be a theoretical nuisance. Both of

© Art Owen 2020 do not distribute or post electronically without author's permission

those get worse if we want $\mathrm{var}(\bar{Y}_1 - \bar{Y}_0)$. Let's just avoid it! We will see later an argument by Box et al. (1978) that will let us use plain old regression methods to get inferences. That is much simpler, and there is no reason to pick the cumbersome way to do things. There are also permutation test methods to get randomization based confidence intervals by Monte Carlo sampling. Those can work well and be straightforward to use. There is also a book in the works by Tirthankar Dasgupta and Donald Rubin on using the actual randomization in more complcated settings. I'm looking forward to seeing how that goes.

After writing the above, I saw Imbens (2019) describing a conservative estimate of $\mathrm{var}(\hat{\tau})$ due to Neyman. It is

$$\frac{1}{n_1(n_1-1)}\sum_i W_i(Y_i-\bar{Y}_1)^2 + \frac{1}{n_0(n_0-1)}\sum_i (1-W_i)(Y_i-\bar{Y}_0)^2.$$

This is just the sum of the two sample variance estimators that we might have used in regular modeling (like Box et al. advise). Let's still avoid digging into why that is conservative.

1.5 Random science tables

Maybe the ij entry of $\mathcal{Y}$ should really be a distribution, like $\mathcal{N}(\mu_{ij}, \sigma^2)$. Then we get a table of random numbers. If instead we want to account for possible correlations between Y_{ij} and $Y_{i'j'}$ then we need a more general model making a random table of numbers.

Suppose $\boldsymbol{W}$ satisfies SUTVA. Then for a random science table we also want W_i to be independent of (Y_{i0}, Y_{i0}). We write this as

$$W_i \perp\!\!\!\perp (Y_{i0}, Y_{i1}).$$

Think how bad it would be otherwise. If somebody purposely set $W_i = 1$ for the largest Δ_i, they would get a very biased answer.

Let $\mathcal{Y}$ be the random science table and suppose we compute $\hat{\tau}$. Is it estimating the ATE for our given science table, or is it estimating the average ATE in the underlying process that made our science table? That is a trick question. It is actually estimating both of those quantities even though they are different from each other. Let's write $\tau(\mathcal{Y})$ for the ATE when the science table is $\mathcal{Y}$ and $\hat{\tau}(\mathcal{Y}, \boldsymbol{W})$ for our estimate making explicit that it depends on the W_i. Let's assume that $\tau_0 = \mathbb{E}(\tau(\mathcal{Y}))$ exists. It is enough for all of the Y_{ij} to have a finite mean. Assume that the randomization in $\boldsymbol{W}$ never gives $\min(n_0, n_1) = 0$. Now under the randomness in $\boldsymbol{W}$,

$$\mathbb{E}(\hat{\tau}(\mathcal{Y}, \boldsymbol{W}) \,|\, \mathcal{Y}) = \tau(\mathcal{Y}),$$

so $\hat{\tau}$ estimates the actual random ATE.

Next

$$\mathbb{E}(\tau(\mathcal{Y})) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n Y_{i1} - Y_{i0}\Big) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(Y_{i1}) - \mathbb{E}(Y_{i0}) = \frac{1}{n}\sum_{i=1}^n \mu_{i1} - \mu_{i0}.$$

© Art Owen 2020 do not distribute or post electronically without author's permission

This is the average ATE over the distribution of random science tables.

We could argue whether $\tau(\mathcal{Y})$ or $\mathbb{E}(\tau(\mathcal{Y}))$ is the more important thing to estimate but in practice they may well be very close. For instance, if $Y_{ij} = \mu_{ij} + \varepsilon_{ij}$ with noise $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ then

$$\tau(\mathcal{Y}) - \mathbb{E}(\tau(\mathcal{Y})) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{i1} - \varepsilon_{i0} \sim \mathcal{N}\Big(0, \frac{2\sigma^2}{n}\Big).$$

In a large experiment the two quantities are close. If the quantities are close then we may choose to study whichever one gives the most clarity to the analysis. In a theoretical study we can model reasonable science tables by specifying a distribution for $\mathcal{Y}$ that fits the applied context.

1.6 External validity

One of the difficulties of randomized controled trials is in applying lessons from a given data set to another one, such as future uses. Think of two science tables, one right now and one that we will get later:

$$\mathcal{Y}_{\text{now}} = \begin{array}{c} \\ i{=}1 \\ 2 \\ 3 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cc} W_i{=}0 & W_i{=}1 \end{array} \\ \begin{bmatrix} Y_{10} & Y_{11} \\ Y_{20} & Y_{21} \\ Y_{30} & Y_{31} \\ \vdots & \vdots \\ Y_{i0} & Y_{i1} \\ \vdots & \vdots \\ Y_{n0} & Y_{n1} \end{bmatrix} \end{array} \qquad \mathcal{Y}_{\text{later}} = \begin{array}{c} \\ i{=}n{+}1 \\ n{+}2 \\ n{+}3 \\ \vdots \\ i \\ \vdots \\ n{+}m \end{array} \begin{array}{c} \begin{array}{cc} W_i{=}0 & W_i{=}1 \end{array} \\ \begin{bmatrix} Y_{n+1,0} & Y_{n+1,1} \\ Y_{n+2,0} & Y_{n+2,1} \\ Y_{n+3,0} & Y_{n+3,1} \\ \vdots & \vdots \\ Y_{i0} & Y_{i1} \\ \vdots & \vdots \\ Y_{n+m,0} & Y_{n_m,1} \end{bmatrix} \end{array} \tag{1.2}$$

These have ATEs τ_{now} and τ_{later} and our experiment will give us the estimate $\hat{\tau}_{\text{now}}$. If we think of $\hat{\tau}_{\text{now}}$ as the future ATE we incur an error

$$\hat{\tau}_{\text{now}} - \tau_{\text{later}} = (\hat{\tau}_{\text{now}} - \tau_{\text{now}}) + (\tau_{\text{now}} - \tau_{\text{later}}).$$

The first term can be studied using statistical inference on our data. We shied away from doing that once we saw the sampling theory issues it raises, but will look into it later. The second term is about whether the ATE might have changed. It is about external validity.

External validity is some sort of extrapolation. It can be extremely reasonable and well supported empirically. E.g., gravity works the same here as elsewhere. It can be unreasonable. E.g., experimental findings on undergraduates who are required to participate for a grade might not generalize to people of all ages, everywhere.

Findings on present customers might not generalize perfectly to others. Findings for mice might not generalize well to humans. Findings in the US

© Art Owen 2020 do not distribute or post electronically without author's permission

might not generalize to the EU. Findings in a clinical trial with given enrolment conditions might not generalize to future patients who are sicker (or are healthier) than those in the study.

You may have heard the expression 'your mileage may vary'. This refered originally to ratings of fuel effiiciency for cars. If you're considering two cars, the advertised mileages μ_0 and μ_1 might not apply to you because you drive differently or live near different kinds of roads or differ in some other way from the test conditions. It commonly holds that the difference $\mu_1 - \mu_0$ might still apply well to you. The things that make your driving different from test conditions could affect both cars nearly equally. In that case, the ATE has reasonable external validity. This is a common occurrence and gives reason for more optimism about external validity.

External validity can be judged but not on the basis of observing a subset of $\mathcal{Y}_{\text{now}}$. The exact same $\mathcal{Y}$ could appear in one problem with external validity and another without it. External validity can be based on past experiences of similar quantities generalizing where tested. That is a form of ***meta-analysis***, a study of studies. External validity can also be based on scientific understanding that may ultimately have been gleaned by looking at what generalizes and what does not organized around an underlying theory that has successfully predicted many generalizations.

1.7 More about causality

We are anchoring our causal discusion in the Neyman-Rubin framework. We emphasize 'effects of causes' not 'causes of effects'. The first involves statements like 'if I water my plant it will grow'. The second involves statements like 'my plant grew because I watered it'. We could investigate the first problem by watering some randomly chosen plants and comparing the results to similar ones that were not watered.

The second problem is a much harder thing to disentangle. The plant could have grown for some other reason. Looking backwards for a cause, there might be 10 potential causes for why a plant grew or an accident happened or a product succeeded. Maybe changing any one of those 10 things would have given a different outcome. Or maybe some of those causes could have changed the outcome on their own while others would have only changed it if paired with some other changes. Then it is quite difficult to say what the real reason was even if you know all 2^{10} potential outcomes. This is a problem of attribution, that remains hard even when all 2^{10} causal possibilities are known without error.

For coverage of methods to infer causality from purely observational data, see Imbens and Rubin (2015), Angrist and Pischke (2008) and Angrist and Pischke (2014), and the notes by Stefan Wager in the class web page. There are several excellent courses on it here at Stanford. To make a causal conclusion requires a causal assumption. The assumption may be that a treatment was applied 'as if at random' or it may be that any important causal variables have been observed. For this course we will be relying on randomizing the treatments

© Art Owen 2020 do not distribute or post electronically without author's permission

and more generally, randomizing treatment combinations.

There is another approach to causality using directed acyclic graphs, explained in Pearl and Mackenzie (2018). Imbens (2019) remarks that it is not widely used in substantive problems and gives reasons. Roughly, it requires inputs from the user about the true underlying causal structure that are hard to come by.

© Art Owen 2020 do not distribute or post electronically without author's permission

© Art Owen 2020 do not distribute or post electronically without author's permission

Bibliography

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist, J. D. and Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect*. Princeton University Press.

Box, G. E., Hunter, W. H., and Hunter, S. (1978). *Statistics for experimenters*, volume 664. John Wiley and sons New York.

Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons.

Cox, D. R. (1958). *Planning of experiments*. Wiley.

Imbens, G. (2019). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. Technical report, National Bureau of Economic Research.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic Books.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*. Duxbury, third edition.

Rosenman, E., Owen, A. B., Baiocchi, M., and Banack, H. (2018). Propensity score methods for merging observational and experimental datasets. *arXiv preprint arXiv:1804.07863*.