Chapter 15
Heterogeneity and Causality

Abstract Before R.A. Fisher introduced randomized experimentation, the literature on causal inference emphasized reduction of heterogeneity of experimental units. To what extent is heterogeneity relevant to causal claims in observational studies when random assignment of treatments is unethical or infeasible?

15.1 J.S. Mill and R.A. Fisher: Reducing Heterogeneity or Introducing Random Assignment

In his *System of Logic: Principles of Evidence and Methods of Scientific Investigation*, John Stuart Mill [11] proposed "four methods of experimental inquiry," including the "method of difference":

> 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 ... [Mill wanted] "two instances ... exactly similar in all circumstances except the one" [under study] [11, III, §8]

Notice Mill's emphasis on a complete absence of heterogeneity: "have every circumstance save one in common;" that is, on treated and control units that are identical but for the treatment. In the modern biology laboratory, nearly identical, genetically engineered mice are compared under treatment and control; this is a modern expression of Mill's 'method of difference.' It is clear from the quote that Mill believed, rightly or wrongly, that heterogeneity of experimental units is directly relevant to causal claims, and does not refer simply to reducing the standard error of an estimate.

Ronald Fisher [3, Chapter 2] took a starkly different view. In 1935, in Chapter 2 of his *Design of Experiments*, Fisher introduced randomized experimentation for the first time in book form, discussing his famous experiment of the lady tasting tea. Fisher was directly critical of the 'method of difference':

P.R. Rosenbaum, *Design of Observational Studies*, Springer Series in Statistics,
DOI 10.1007/978-1-4419-1213-8_15, © Springer Science+Business Media, LLC 2010

> 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 ... These are only examples of the differences probably present; it would be impossible to present an exhaustive list of such possible differences ... because [they] ... are always strictly innumerable. When any such cause is named, it is usually perceived that, by increased labor and expense, it could be largely eliminated. Too frequently it is assumed that such refinements constitute improvements to the experiment ... [3, page 18]

In the first omission, "...," in this quote, Fisher discussed the many ways two cups of tea may differ.

Fisher is, of course, engaged in an enormously important task: he is introducing the logic of randomized experimentation to a broad audience for the first time. Moreover, it would be reasonable to say that Fisher was correct and Mill was wrong in certain critical respects. In a randomized clinical trial conducted in a hospital, the patients are heterogeneous and not much can be done about it. There is no opportunity to replace the hospital's patients by genetically engineered, nearly identical patients. And yet it is possible to randomly assign treatments to heterogeneous patients and draw valid causal inferences in just the manner that Fisher advocated.

Despite this, one senses that there is at least something that is sensible in Mill's method of difference, something that is sensible in the fanatical efforts of the basic science laboratory to eliminate heterogeneity, in the use of nearly identical mice. One senses that Fisher, in his understandable enthusiasm for his new method, has gone just a tad too far in his criticism of Mill's drive to eliminate heterogeneity. Indeed, the issue may be particularly relevant in observational studies where random assignment of treatments is either unethical or infeasible.

Some care is required in discussing heterogeneity. Heterogeneity is itself heterogeneous; there are several kinds of heterogeneity. In the biology laboratory, it is often wise to use several different strains or species of genetically engineered, nearly identical laboratory animals, making sure that each strain or species is equally represented in treated and control groups, to verify that any conclusion is not a peculiarity of a single strain. It is uncontrolled rather than controlled heterogeneity that is the target for reduction or elimination. Controlled heterogeneity has numerous uses.[1]

See Paul Holland's [8] essay "Statistics and causal inference" for further contrasts of the views of Mill and Fisher.

[1] Recall the related discussion of Bitterman's concept of "control by systematic variation" in §5.2.2, where some factor is demonstrated to be irrelevant by systematically varying that factor.

15.2 A Larger, More Heterogeneous Study Versus a Smaller, Less Heterogeneous Study

Large I or small σ: Which is better?

To explore the issue raised in §15.1, consider the following simple situation. There are I matched pairs in an observational study with treated-minus-control differences in outcomes Y_i, $i = 1,\ldots,I$. Because it is an observational study, not a randomized experiment, we cannot assume that matching for observed covariates has removed all bias from nonrandom assignment — we cannot assume the naïve model of Chapter 3, and will report a sensitivity analysis. Although we cannot know this from the observed data, the situation is, in fact, the 'favorable situation,' in which there is a treatment effect and the matching has succeeded in removing bias, so the naïve model is correct and treatment assignment is effectively randomized within matched pairs; see §14.2. In this favorable situation, the investigator hopes to report that the treatment appears to be effective and that appearance is insensitive to small and moderate biases. Indeed, the situation is simpler still: the treatment has an additive, constant effect, $\tau = r_{Tij} - r_{Cij}$, so that $Y_i = \tau + (2Z_{i1} - 1)(r_{Ci1} - r_{Ci2})$; see §2.4.1. Moreover, the $r_{Ci1} - r_{Ci2}$ are independent and identically distributed observations drawn from a continuous distribution symmetric about zero; see §14.1. Because this is the favorable situation, $2Z_{i1} - 1 = \pm 1$, each with probability $\frac{1}{2}$ independently of $r_{Ci1} - r_{Ci2}$, so $Y_i - \tau$ itself has this same continuous distribution symmetric about zero.

The investigator faces a choice between a larger study with more heterogeneous responses or a smaller study with less heterogeneous responses, both in the 'favorable situation.' In §15.1, Mill would have advocated the smaller, less heterogeneous study. Is there any merit to Mill's claim? The heterogeneity here refers to heterogeneity that remains after matching for observed covariates, that is, heterogeneity within pairs; heterogeneity between pairs is not at issue. Specifically, the investigator faces the following admittedly stylized choice: observe either $4I$ pairs with additive effect τ and $(r_{Ci1} - r_{Ci2})/\omega \sim F(\cdot)$, where $F(\cdot)$ is a continuous distribution symmetric about zero, or alternatively observe I pairs with additive effect τ and $(r_{Ci1} - r_{Ci2})/(\omega/2) \sim F(\cdot)$. In words, the choice is between $4I$ pairs with dispersion ω or I pairs with dispersion $\omega/2$. The choice is stylized in the following sense. If $F(\cdot)$ were the standard Normal distribution, then the sample mean difference, $\overline{Y} = (1/I)\sum_{i=1}^{I} Y_i$, would be Normally distributed $\overline{Y} \sim N\left\{\tau, \omega^2/(4I)\right\}$ with expectation τ and variance $\omega^2/(4I)$ in both the larger, more heterogeneous study and the smaller less heterogeneous study. If ω were known in a randomized experiment, the larger more heterogenous study and the smaller less heterogeneous study would barely be worth distinguishing, because the sufficient statistic, $\overline{Y}$, has the same distribution in both studies.

Of course, this is not a randomized experiment. Does that matter for this choice? If so, how does it matter?

A simulated example

Figure 15.1 depicts a simulated example of the choice between a smaller, less heterogeneous study (SL) with $Y_i \sim N\left\{\tau, (\omega/2)^2\right\}$ for $i = 1, \ldots, I = 100$, and a larger, more heterogeneous study (LM) with $Y_i \sim N\left\{\tau, \omega^2\right\}$ for $i = 1, \ldots, I = 400$. In Figure 15.1, $\tau = 1/2$, $\omega = 1$. The boxplots for SL and LM have 100 and 400 pairs, respectively.

If SL and LM were analyzed as if they were randomized experiments, the inferences would be very similar. In SL, the mean difference is $\overline{Y} = 0.487$ with estimated standard error 0.054, while in LM the mean difference is $\overline{Y} = 0.485$ with estimated standard error 0.049; however, the true standard error is 0.05 in both SL and LM. Using Wilcoxon's signed rank statistic to test the null hypothesis of no effect yields a very small P-value in both SL and LM, less than 10^{-10}. The Hodges-Lehmann point estimate $\widehat{\tau}$ of τ is $\widehat{\tau} = 0.485$ for SL and $\widehat{\tau} = 0.489$ for LM. The 95% confidence interval from the randomization distribution of Wilcoxon's statistic is $[0.374, 0.600]$ from SL and $[0.390, 0.587]$ from LM. If the choice were between two randomized experiments with the distributions SL and LM yielding Figure 15.1, there would be little reason to prefer one over the other.

Suppose, however, that SL and LM came from observational studies, so the behavior of the Y_i might reflect either a treatment effect or an unmeasured bias or a combination of the two. How sensitive are the conclusions from SL and LM to departures from the naïve model (3.5)-(3.8) that underlies the inferences in the previous paragraph?

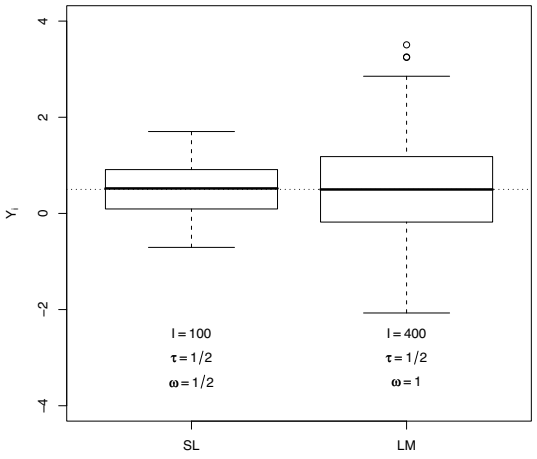

Fig. 15.1 A simulated example of the choice between a smaller, less heterogeneous study (SL) and a larger, more heterogeneous study (LM). In SL there are $I = 100$ independent matched pair differences, Y_i, that are Normal with expectation τ and standard deviation $\omega = 1/2$. In LM there are $I = 400$ independent matched pair differences, Y_i, that are Normal with expectation τ and standard deviation $\omega = 1$. The horizontal dotted line is at $1/2$.

Table 15.1 Sensitivity analysis for the larger, more heterogeneous study (LM) and the smaller, less heterogeneous study (SL). Upper bounds on the one-sided P-value from Wilcoxon's signed rank test when testing the null hypothesis of no treatment effect are given. Although the randomization inferences are similar ($\Gamma = 1$), the smaller, less heterogeneous study is much less sensitive to bias from unmeasured covariates ($\Gamma \geq 3$).

Γ	1	2	3	4	5
LM	$< 10^{-10}$	0.00046	0.39	0.97	1.00
SL	$< 10^{-10}$	0.000016	0.0021	0.022	0.083

Table 15.2 Sensitivity analysis for the larger more heterogeneous study (LM) and the smaller less heterogeneous study (SL). For $\Gamma = 1$, the table gives the value of the Hodges-Lehmann point estimate $\widehat{\tau}$ of the treatment effect, τ. For $\Gamma = 2$, the table gives the interval of possible point estimates, $[\widehat{\tau}_{\text{min}}, \widehat{\tau}_{\text{max}}]$, and the length of that interval, $\widehat{\tau}_{\text{max}} - \widehat{\tau}_{\text{min}}$. For a bias of magnitude $\Gamma = 2$, the range of possible point estimates is much longer for the larger, more heterogeneous study than for the smaller, less heterogeneous study.

	$\Gamma = 1$	$\Gamma = 2$	$\Gamma = 2$
	$\widehat{\tau}$	$[\widehat{\tau}_{\text{min}}, \widehat{\tau}_{\text{max}}]$	$\widehat{\tau}_{\text{max}} - \widehat{\tau}_{\text{min}}$
LM	0.489	[0.19, 0.79]	0.60
SL	0.485	[0.32, 0.66]	0.34

Table 15.1 displays two sensitivity analyses, one for LM, the other for SL, giving the upper bound on the one-sided P-value for testing the hypothesis of no treatment effect using Wilcoxon's signed rank statistic; see §3.4. As noted above, the randomization inferences ($\Gamma = 1$) are quite similar for LM and SL. In sharp contrast, the smaller, less heterogeneous study, SL, is much less sensitive to bias from an unmeasured covariate. A bias of $\Gamma = 3$ could produce a boxplot similar to the LM boxplot in Figure 15.1 (the upper bound on the P-value is 0.39), but a bias of $\Gamma = 3$ is very unlikely to produce the boxplot for SL (the upper bound on the P-value is 0.0021). To put this in context, SL is just slightly more sensitive to unmeasured bias than Hammond's [5] study of heavy smoking as a cause of lung cancer (see [14, §4]), one of the least sensitive observational studies, whereas LM is much more sensitive.

In parallel with Table 3.4, Table 15.2 displays two sensitivity analyses, one for LM, the other for SL, giving the interval of possible point estimates, $[\widehat{\tau}_{\text{min}}, \widehat{\tau}_{\text{max}}]$, of the treatment effect τ. For $\Gamma = 1$, the interval is a point, namely the Hodges-Lehmann point estimate, $\widehat{\tau}_{\text{min}} = \widehat{\tau}_{\text{max}} = \widehat{\tau}$, and it is about the same for LM and SL. For $\Gamma = 2$, the interval for LM, namely $[0.19, 0.79]$, is considerably longer than the interval for SL, namely $[0.32, 0.66]$.

In Tables 15.1 and 15.2, the smaller, less heterogeneous study is better than the larger, more heterogeneous study, better in the sense of being less sensitive to unmeasured biases. It is important to keep in mind that Figure 15.1 depicts two 'favorable situations,' that is, treatment effects without unmeasured biases. Because these are observational studies, the investigator does not know when she is in the 'favorable situation,' and so she cannot assert that Figure 15.1 depicts effects, not biases. The investigator can, however, assert that a bias of magnitude $\Gamma = 4$ is too small to explain away as unmeasured bias the ostensible effect in the smaller, less

Table 15.3 Power of a sensitivity analysis for a larger, more heterogeneous study (LM, $4I$ pairs, $\omega = 1$), and a smaller, less heterogeneous study, (SL, I pairs, $\omega = 1/2$), with the same additive, constant treatment effect, $\tau = 1/2$, in the favorable situation with $\{r_{Ci1} - r_{Ci2}\}/\omega \sim_{iid} F(.)$. The power is similar when $\Gamma = 1$, but is higher for SL when Γ is larger.

Study	Distribution	Number	Dispersion	Power		
	$F(\cdot)$	of Pairs	ω	$\Gamma = 1$	$\Gamma = 1.5$	$\Gamma = 2$
LM	Normal	120	1	1.00	0.96	0.60
SL	Normal	30	$\frac{1}{2}$	1.00	1.00	0.96
LM	Logistic	120	1	0.93	0.31	0.04
SL	Logistic	30	$\frac{1}{2}$	0.93	0.61	0.32
LM	Cauchy	200	1	0.98	0.32	0.02
SL	Cauchy	50	$\frac{1}{2}$	0.95	0.60	0.28

heterogeneous study, but could not assert this about the larger, more heterogeneous study.

Power comparisons with Normal, logistic and Cauchy errors

Table 15.3 contrasts the larger, more heterogeneous study (LM) and the smaller, less heterogeneous study (SL) in terms of the power of the sensitivity analysis. Table 15.3 refers to a one-sided, 0.05 level test of the null hypothesis of no treatment effect. The power is the probability that the upper bound on the one-sided P-value is at most 0.05. The power is calculated as in §14.2.

The power of the randomization test ($\Gamma = 1$) is similar for LM and SL, but for larger Γ, particularly for $\Gamma = 2$, the power is higher for the smaller, less heterogeneous study (SL). The pattern seen in Table 15.1 is not a peculiarity of one simulation; rather, it is the anticipated pattern based on a comparison of the powers of LM and SL.

Design sensitivity

Proposition 14.1 and expressions (14.15) and (14.16) may be used to determine the design sensitivity $\widetilde{\Gamma}$ of the larger, more heterogeneous study (LM) and the smaller, less heterogeneous study (SL) as

$$\widetilde{\Gamma} = \frac{\Phi\left(\sqrt{2}\frac{\tau}{\omega}\right)}{1-\Phi\left(\sqrt{2}\frac{\tau}{\omega}\right)} \tag{15.1}$$

for Normal errors and as

$$\widetilde{\Gamma} = \frac{\Upsilon\left(\frac{\tau}{\omega}\right)}{1-\Upsilon\left(\frac{\tau}{\omega}\right)} \tag{15.2}$$

for Cauchy errors. For Normal errors with $\tau = 1/2$, the designs LM ($\omega = 1$) and SL ($\omega = 1/2$) have design sensitivities $\widetilde{\Gamma}$ of 3.171 and 11.715, respectively. In light of this, the power of both designs with Normal errors tends to 1 as $I \to \infty$ for the values of Γ in Table 15.3, but for $\Gamma = 5$ the power of LM would tend to 0 while the power of SL would tend to 1. In parallel, for Cauchy errors with $\tau = 1/2$, the designs LM ($\omega = 1$) and SL ($\omega = 1/2$) have design sensitivities $\widetilde{\Gamma}$ of 1.838 and 3, respectively. In light of this, with Cauchy errors, for $\Gamma = 2$ in Table 15.3, the power of SL tends to 1 as $I \to \infty$ while the power of LM tends to 0.

15.3 Heterogeneity and the Sensitivity of Point Estimates

In the current chapter, the treatment has an additive constant effect, $\tau = r_{Tij} - r_{Cij}$, and in a randomized experiment, the Hodges-Lehmann estimate $\widehat{\tau}$ is a consistent estimate of τ. For a given deviation Γ from randomized treatment assignment, Tables 3.4 and 15.2 displayed the interval of possible Hodges-Lehmann point estimates $[\widehat{\tau}_{\min}, \widehat{\tau}_{\max}]$ of τ, where $\widehat{\tau}_{\min} = \widehat{\tau}_{\max} = \widehat{\tau}$ for $\Gamma = 1$; see §3.5 and [13]. As the sample size increases, $I \to \infty$, the endpoints of this interval converge in probability to the endpoints of a fixed interval, $[\tau_{\min}, \tau_{\max}]$; this interval reflects the uncertainty about τ that is due to a potential bias of magnitude Γ when there is no longer any sampling uncertainty.

In the 'favorable situation,' with errors having Normal $\Phi(\cdot)$ or Cauchy $\Upsilon(\cdot)$ cumulative distributions, the following proposition gives the form of this limiting interval. See [16, Appendix] for proof of Proposition 15.1.[2]

Proposition 15.1. *If* $(D_i - \tau)/\omega \sim_{iid} \Phi(\cdot)$ *then* $[\tau_{\min}, \tau_{\max}]$ *is* $\tau \pm \omega \Phi^{-1}(\kappa)/\sqrt{2}$, *where* $\kappa = \Gamma/(1+\Gamma)$. *If* $(D_i - \tau)/\omega \sim_{iid} \Upsilon(\cdot)$ *then* $[\tau_{\min}, \tau_{\max}]$ *is* $\tau \pm \omega \Upsilon^{-1}(\kappa)$.

Proposition 15.1 is consistent with Mill's view that heterogeneity of experimental units, ω, is directly relevant to causal claims. In Proposition 15.1, there is no sampling variability, because Proposition 15.1 refers to the limit as $I \to \infty$. The uncertainty addressed in Proposition 15.1 is quantified by the length of the limiting interval, $\tau_{\max} - \tau_{\min}$, and despite the absence of sampling variability, the length of that interval is directly proportional to ω.

Proposition 15.1 does not contradict Fisher's view, but it does emphasize that this view is applicable only when biases from nonrandom treatment assignment have been avoided by randomization. Reducing heterogeneity ω and increasing sample

[2] Although the proof has a few details, it is simple in concept. The lower endpoint $\widehat{\tau}_{\min}$ of the interval of possible point estimates is obtained by equating Wilcoxon's signed rank statistic T computed from $Y_i - \tau_0$, say T_{τ_0}, to the maximum null expectation of T, namely $E\left(\overline{\overline{T}} \middle| \mathscr{F}, \mathscr{Z}\right) = \Gamma I(I+1)/\{2(1+\Gamma)\}$ from (3.19) and solving the equation for $\widehat{\tau}_{\min}$. Dividing the equation by $I(I+1)/2$ yields the equivalent equation $2T_{\tau_0}/\{I(I+1)\} = \Gamma/(1+\Gamma)$. If $(Y_i - \tau)/\omega \sim \Phi(\cdot)$ or $(Y_i - \tau)/\omega \sim \Upsilon(\cdot)$ then as $I \to \infty$, the left side of the equation, $2T_{\tau_0}/\{I(I+1)\}$, converges in probability to a function of $(\tau - \tau_0)/\omega$, and the rest of the proof is detail devoted to showing that equation can be solved to give the solutions in the statement of Proposition 15.1.

size I compete for resources in a randomized experiment because bias is known to have been avoided, so the analysis can be conducted with $\Gamma = 1$. More precisely, in Proposition 15.1, if it were known that $\Gamma = 1$, then $\kappa = 1/2$, so $\Phi^{-1}(\kappa)/\sqrt{2} = \Phi^{-1}(1/2)/\sqrt{2} = 0$ and $\Upsilon^{-1}(\kappa) = \Upsilon^{-1}(1/2) = 0$, and the length of the limiting interval is $\tau_{\max} - \tau_{\min} = 0$ for every value of ω.

The length of the interval is also affected by the magnitude of potential bias, Γ, though $\kappa = \Gamma/(1+\Gamma)$. The two components determine the length $\tau_{\max} - \tau_{\min}$ of the limiting interval in a multiplicative manner; for the Normal, $\tau_{\max} - \tau_{\min} = 2\omega\,\Phi^{-1}(\kappa)/\sqrt{2}$. A given magnitude Γ of deviation from a randomized experiment does more harm when the units are more heterogeneous, that is, when ω is larger. If you were deceptively trying to bias a randomized trial by covertly tilting the treatment assignment probabilities by a magnitude of Γ in (3.16)-(3.18), then you could do more harm if the units were the heterogeneous patients in a clinical trial than if they were the homogeneous genetically engineered mice in a laboratory experiment.

In the 'favorable situation' in an observational study, increasing the sample size I reduces the standard error, but it does not materially reduce sensitivity to unmeasured biases. In contrast, in this situation, reducing the heterogeneity of experimental units, ω, reduces both the standard error and sensitivity to unmeasured biases. In an observational study, LM and SL of Figure 15.1 are not at all the same: SL is much better.

15.4 Examples of Efforts to Reduce Heterogeneity

Twins

What are the economic returns to additional education? You cannot compare the mid-life earnings of surgeons and high school dropouts — they differed in the middle of high school before the dropout left school. Not in every case, but typically, the child who went on to become a surgeon was receiving better grades and standardized test scores in high school, was more strongly motivated for school studies, had better educated, wealthier parents, and not inconceivably differed in some relevant genes. You would like to compare two children of the same parents with different education growing up at the same time in the same home with the same genes. Ashenfelter and Rouse [1] compared the earnings of identical twins with differing education, estimating about a 9% increase in earnings per year of additional education.

The use of twins is the canonical example of trading sample size for reduced heterogeneity. Twin pairs are quite heterogeneous between pairs and in several important respects fairly homogeneous within pairs, so the use of twins reflects the type of heterogeneity discussed in this chapter.

Road hazards

What permanent features of a road affect the risk of collisions with roadside objects? Road hazards are a fairly small part of accident risk. Also relevant are: the driver's sobriety, skill and risk tolerance; ambient light; the weather — ice, snow, rain and fog; safety equipment — use of seat belts, quality and condition of brakes, tires, air bags, traction and stability control devices. These factors are related. The risk-averse driver will drive near the legal speed limit, but will also invest in safety devices and wear seat belts. In the rain or snow, one drives on the highway to work but not on the dirt road to the picnic area, so weather and road hazards vary together. Sobriety is more common at noon than at midnight, so sobriety and ambient light vary together. You would like to compare different road hazards with the same driver, in the same car, in the same weather, with the same ambient light, in the same state of sobriety, with seat belts in the same state of use. Is this possible?

Using a simple, clever study design, Wright and Robertson [22] did just that. They compared 300 fatal roadside collisions in Georgia in 1974–1975 to 300 nonaccidents involving the same driver, in the same car, in the same ambient light, and so on. The nonaccidents occurred one mile back from the crash site, a location passed without incident by the driver just moments before the crash. At crash sites, Wright and Robertson found a substantial excess of roads that curved more than 6 degrees with downhill gradients of greater than 2%. (Technically, this is a 'case-crossover' study of the type proposed by Malcolm Maclure [10], except that it is defined by geography rather than time; see also the 'case-specular' design of Sander Greenland [4].)

The genetically engineered mice of microeconomics

Many businesses that provide products or services over large regions adopt a strategy known as 'replication' in which nearly identical outlets are reproduced at high speed in diverse locations [21]. Starbucks and Tesco are two of the many such businesses. This strategy confers various benefits to businesses that use it, but it also creates nearly identical copies of a business in locations that may have adopted different regulations, taxes or other policies. For instance, Card and Kreuger's [2] study of the minimum wage and employment compared Burger Kings in New Jersey to Burger Kings in Pennsylvania, KFCs in New Jersey to KFCs in Pennsylvania, etc., and in this way eliminated one of several sources of extraneous variation between the two states; see §4.5 and §11.3 for further discussion of this study.

Motorcycle helmets

Do helmets reduce the risk of death in motorcycle crashes? Crashes occur at different speeds with different forces, and neither speeds nor forces are likely to be measured. Motorcyclists hit different objects — pedestrians or Hummers — in dense or

light traffic, with emergency services near or far away. One would like to compare two people, one with a helmet, the other without, on the same type of motorcycle, driving at the same speed, crashing into the same object, in the same traffic, with equal proximity to medical aid. Is that possible?

It is when two people ride one motorcycle, one with a helmet, the other without. Norvell and Cummings [12] looked at such crashes, finding about 40% lower risk associated with helmet use.

15.5 Summary

In a randomized experiment, an unbiased estimate of treatment effect is available, so increasing the sample size, I, or reducing the unit heterogeneity, ω, both serve to reduce the standard error of an unbiased estimate. The situation is strikingly different in an observational study. In the 'favorable situation' in an observational study, the treatment is effective and there are no unmeasured biases. If the favorable situation arose, the investigator would not know it, and at best would hope to report that the treatment appears to be effective and that appearance is insensitive to small and moderate biases. In this situation, reducing heterogeneity, even purely random heterogeneity, ω, confers benefits that cannot be obtained by increasing the sample size, I. Specifically, reducing heterogeneity reduces sensitivity to unmeasured biases. Several cleverly designed studies have illustrated efforts to reduce heterogeneity.

15.6 Further Reading

This chapter is based on [16], where additional discussion may be found.

References

1. Ashenfelter, O., Rouse, C.: Income, schooling and ability: Evidence from a new sample of identical twins. Q J Econ **113**, 253–284 (1998)
2. Card, D., Krueger, A.: Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. Am Econ Rev **84**, 772–793 (1994)
3. Fisher, R.A.: Design of Experiments. Edinburgh: Oliver and Boyd (1935)
4. Greenland, S.: A unified approach to the analysis of case-distribution (case-only) studies. Statist Med **18**, 1–15 (1999)
5. Hammond, E.C.: Smoking in relation to mortality and morbidity. J Natl Cancer Inst **32**, 1161–1188 (1964)
6. Heller, R., Rosenbaum, P.R., Small, D.: Split samples and design sensitivity in observational studies. J Am Statist Assoc **104**, to appear (2009)
7. Hodges, J.L., Lehmann, E.L.: Estimates of location based on ranks, Ann Math Statist **34**, 598–611 (1963)

8. Holland, P.W.: Statistics and causal inference. J Am Statist Assoc **81**, 945–960 (1986)
9. Lehmann, E.L.: Nonparametrics, San Francisco: Holden Day (1975) Reprinted New York: Springer (2006)
10. Maclure, M.: The case-crossover design: A method for studying transient effects on the risk of acute events. Am J Epidemiol **133**, 144–152 (1991)
11. Mill, J.S.: A System of Logic: The Principles of Evidence and the Methods of Scientific Investigation. Indianapolis: Liberty Fund (1867)
12. Norvell, D.C., Cummings, P.: Association of helmet use with death in motorcycle crashes: A matched-pair cohort study. Am J Epidemiol **156**, 483–487 (2002)
13. Rosenbaum, P.R.: Hodges-Lehmann point estimates of treatment effect in observational studies. J Am Statist Assoc **88**, 1250–1253 (1993)
14. Rosenbaum, P.R.: Observational Studies (2nd ed.). New York: Springer (2002)
15. Rosenbaum, P. R.: Design sensitivity in observational studies. Biometrika **91**, 153–164 (2004)
16. Rosenbaum, P. R.: Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. Am Statistician **59**, 147–152 (2005)
17. Rosenbaum, P.R.: What aspects of the design of an observational study affect its sensitivity to bias from covariates that were not observed? In: Festshrift for Paul W. Holland. Princeton, NJ: ETS (2009)
18. Salsburg, D.: The Lady Tasting Tea. San Francisco: Freeman (2001)
19. Small, D., Rosenbaum, P.R.: War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. J Am Statist Assoc **103**, 924–933 (2008)
20. Werfel, U., Langen, V., Eickhoff, I., Schoonbrood, J., Vahrenholz, C., Brauksiepe, A., Popp, W., Norpoth, K.: Elevated DNA single-strand breakage frequencies in lymphocytes. Carcinogenesis **19**, 413–418 (1998)
21. Winter, S.G., Szulanski, G.: Replication as strategy. Organizat Sci **12**, 730–743 (2001)
22. Wright, P.H., Robertson, L.S.: Priorities for roadside hazard modification. Traffic Eng **46**, 24–30 (1976)