Chapter 10
RANDOM VARIABILITY

Suppose an investigator conducted a randomized experiment to answer the causal question "does one's looking up to the sky make other pedestrians look up too?" She found an association between her looking up and other pedestrians' looking up. Does this association reflect a causal effect? By definition of randomized experiment, confounding bias is not expected in this study. In addition, no selection bias was expected because all pedestrians' responses—whether they did or did not look up—were recorded, and no measurement bias was expected because all variables were perfectly measured. However, there was another problem: the study included only 4 pedestrians, 2 in each treatment group. By chance, 1 of the 2 pedestrians in the "looking up" group, and neither of the 2 pedestrians in the "looking straight" group, was blind. Thus, even if the treatment (the investigator's looking up) truly had a strong average effect on the outcome (other people's looking up), half of the individuals in the treatment group happened to be immune to the treatment. The small size of the study population led to a dilution of the estimated effect of treatment on the outcome.

There are two qualitatively different reasons why causal inferences may be wrong: systematic bias and random variability. The previous three chapters described three types of systematic biases: selection bias, measurement bias—both of which may arise in observational studies and in randomized experiments—and unmeasured confounding—which is not expected in randomized experiments. So far we have disregarded the possibility of bias due to random variability by restricting our discussion to huge study populations. In other words, we have operated as if the only obstacles to identify the causal effect were confounding, selection, and measurement. It is about time to get real: the size of study populations in etiologic research rarely precludes the possibility of bias due to random variability. This chapter discusses random variability and how we deal with it.

10.1 Identification versus estimation

The first nine chapters of this book are concerned with the computation of causal effects in study populations of near infinite size. For example, when computing the causal effect of heart transplant on mortality in Chapter 2, we only had a twenty-person study population but we regarded each individual in our study as representing 1 billion identical individuals. By acting as if we could obtain an unlimited number of individuals for our studies, we could ignore random fluctuations and could focus our attention on systematic biases due to confounding, selection, and measurement. Statisticians have a name for problems in which we can assume the size of the study population is effectively infinite: identification problems.

Thus far we have reduced causal inference to an identification problem. Our only goal has been to identify (or, as we often said, to compute) the average causal effect of treatment A on the outcome Y. The concept of identifiability was first described in Section 3.1—and later discussed in Sections 7.2 and 8.4—where we also introduced some conditions generally required to identify causal effects even if the size of the study population could be made arbitrarily large. These so-called identifying conditions were exchangeability, positivity, and consistency.

Our ignoring random variability may have been pedagogically convenient to introduce systematic biases, but also extremely unrealistic. In real research

projects, the study population is not effectively infinite and hence we cannot ignore the possibility of random variability. To this end let us return to our twenty-person study of heart transplant and mortality in which 7 of the 13 treated individuals died.

Suppose our study population of 20 can be conceptualized as being a random sample from a *super-population* so large compared with the study population that we can effectively regard it as infinite. Further, suppose our goal is to make inferences about the super-population. For example, we may want to make inferences about the super-population probability (or proportion) $\Pr[Y = 1|A = a]$. We refer to the parameter of interest in the super-population, the probability $\Pr[Y = 1|A = a]$ in this case, as the *estimand*. An *estimator* is a rule that takes the data from any sample from the super-population and produces a numerical value for the estimand. This numerical value for a particular sample is the *estimate* from that sample. The sample proportion of individuals that develop the outcome among those receiving treatment level a, $\widehat{\Pr}[Y = 1 \mid A = a]$, is an estimator of the super-population probability $\Pr[Y = 1|A = a]$. The estimate from our sample is $\widehat{\Pr}[Y = 1 \mid A = a] = 7/13$. More specifically, we say that 7/13 is a *point estimate*. The value of the estimate will depend on the particular 20 individuals randomly sampled from the super-population.

As informally defined in Chapter 1, an estimator is *consistent* for a particular estimand if the estimates get (arbitrarily) closer to the parameter as the sample size increases (see Technical Point 10.1 for the formal definition). Thus the sample proportion $\widehat{\Pr}[Y = 1 \mid A = a]$ consistently estimates the super-population probability $\Pr[Y = 1|A = a]$, i.e., the larger the number n of individuals in our study population, the smaller the magnitude of $\Pr[Y = 1|A = a] - \widehat{\Pr}[Y = 1 \mid A = a]$ is expected to be. Previous chapters were exclusively concerned with identification; from now on we will be concerned with statistical estimation.

For an introduction to statistics, see the book by Wasserman (2004). For a more detailed introduction, see Casella and Berger (2002).

Even consistent estimators may result in point estimates that are far from the super-population value. Large differences between the point estimate and the super-population value of a proportion are much more likely to happen when the size of the study population is small compared with that of the super-population. Therefore it makes sense to have more confidence in estimates that originate from larger study populations. In the absence of systematic biases, statistical theory allows one to quantify this confidence in the form of a confidence interval around the point estimate. The larger the size of the study population, the narrower the confidence interval. A common way to construct a 95% confidence interval for a point estimate is to use a 95% Wald confidence interval centered at a point estimate. It is computed as follows.

First, estimate the standard error of the point estimate under the assumption that our study population is a random sample from a much larger super-population. Second, calculate the upper limit of the 95% Wald confidence interval by adding 1.96 times the estimated standard error to the point estimate, and the lower limit of the 95% confidence interval by subtracting 1.96 times the estimated standard error from the point estimate. For example, consider our estimator $\widehat{\Pr}[Y = 1 \mid A = a] = \hat{p}$ of the super-population parameter $\Pr[Y = 1|A = a] = p$. Its standard error is $\sqrt{\frac{p(1-p)}{n}}$ (the standard error of a binomial) and thus its estimated standard error is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(7/13)(6/13)}{13}} = 0.138$. Recall that the Wald 95% confidence interval for a parameter θ based on an estimator $\widehat{\theta}$ is $\widehat{\theta} \pm 1.96 \times \widehat{se}\left(\widehat{\theta}\right)$ where $\widehat{se}\left(\widehat{\theta}\right)$ is an estimate of the (exact

or large sample) standard error of $\widehat{\theta}$ and 1.96 is the upper 97.5% quantile of a standard normal distribution with mean 0 and variance 1. Therefore the 95% Wald confidence interval for our estimate is 0.27 to 0.81. The length and centering of the 95% Wald confidence interval will vary from sample to sample.

A Wald confidence interval centered at $\hat{p}$ is only guaranteed to be valid in large samples. For simplicity, here we assume that our sample size is sufficiently large for the validity of our Wald interval.

A 95% confidence interval is *calibrated* if the estimand is contained in the interval in 95% of random samples, *conservative* if the estimand is contained in more than 95% of samples, and *anticonservative* otherwise. We will say that a confidence interval is *valid* if, for any value of the true parameter, the interval is either calibrated or conservative, i.e. it covers the true parameter at least 95% of the time. We would like to choose the valid interval whose width is narrowest.

The validity of confidence intervals is defined in terms of the frequency of coverage in repeated samples from the super-population, but we only see one of those samples when we conduct a study. Why should we care about what would have happened in other samples that we did not see? One important answer is that the definition of confidence interval also implies the following. Suppose we and all of our colleagues keep conducting research studies for the rest of our lifetimes. In each new study, we construct a valid 95% confidence interval for the parameter of interest. Then, at the end of our lives, we can look back at all the studies that were conducted, and conclude that the parameters of interest were trapped in—or covered by—the confidence interval in at least 95% of the studies. Unfortunately, we will have no way of identifying the (up to) 5% of the studies in which the confidence interval failed to include the super-population quantity.

Importantly, the 95% confidence interval from a single study does not imply that there is a 95% probability that the estimand is in the interval. In our example, we cannot conclude that the probability that the estimand lies between the values 0.27 and 0.81 is 95%. The estimand is fixed, which implies that either it is or it is not included in the particular interval (0.27, 0.81). In this sense, the probability that the estimand is included in that interval is either 0 or 1. A confidence interval only has a *frequentist* interpretation. Its level (e.g., 95%) refers to the frequency with which the interval will trap the unknown super-population quantity of interest over a collection of studies (or in hypothetical repetitions of a particular study).

In contrast with a frequentist 95% confidence interval, a Bayesian 95% credible interval can be interpreted as "there is a 95% probability that the estimand is in the interval". However, for a Bayesian, probability is defined not as a frequency over hypothetical repetitions but as degree-of-belief. In this book we adopt the frequency definition of probability. See Fine Point 11.2 for more on Bayesian intervals.

Confidence intervals are often classified as either *small-sample* or *large-sample* confidence intervals. A small-sample valid (conservative or calibrated) confidence interval is one that is valid at all sample sizes for which it is defined. Small-sample calibrated confidence intervals are sometimes called exact confidence intervals. A large-sample (equivalently, asymptotic) valid confidence interval is one that is valid only in large samples. A large-sample calibrated 95% confidence interval is one whose coverage becomes arbitrarily close to 95% as the sample size increases. The Wald confidence interval for $\Pr[Y = 1|A = a] = p$ mentioned above is a large-sample calibrated confidence interval, but not a small-sample valid interval. (There do exist small-sample valid confidence intervals for p, but they are not often used in practice.) When the sample size is small, a valid large-sample confidence interval, such as the Wald 95% confidence interval of our example above, may not be valid. In this book, when we use the term 95% confidence interval, we mean a large-sample valid confidence interval, like a Wald interval, unless stated otherwise. See also Fine Point 10.1.

There are many valid large-sample confidence intervals other than the Wald interval (Casella and Berger, 2002). One of these might be preferred over the Wald interval, which can be badly anti-conservative in small samples (Brown et al, 2001).

However, not all consistent estimators can be used to center a valid Wald confidence interval, even in large samples. Most users of statistics will consider an estimator unbiased if it can center a valid Wald interval and biased if it

Fine Point 10.1

Honest confidence intervals. The smallest sample size at which a large-sample, valid 95% confidence interval covers the true parameter at least 95% of the time may depend on the unknown value of the true parameter. We say a large-sample valid 95% confidence interval is *uniform* or *honest* if there exists a sample size n at which the interval is guaranteed to cover the true parameter value at least 95% of the time, whatever be the value of the true parameter. We demand honest intervals because, in the absence of uniformity, at any finite sample size there may be data generating distributions under which the coverage of the true parameter is much less than 95%. Unfortunately, for a large-sample, honest confidence interval, the smallest such n is generally unknown and is difficult to determine even by simulation. See Robins and Ritov (1997) for technical details.

In the remainder of the text, when we refer to valid confidence intervals, we will mean large-sample honest confidence intervals. By definition, any small-sample valid confidence interval is uniform or honest for all n for which the interval is defined.

cannot (see Technical Point 10.1 for details). For now, we will equate the term *bias* with the inability to center valid Wald confidence intervals. Also, bear in mind that confidence intervals only quantify uncertainty due to random error, and thus the confidence we put on confidence intervals may be excessive in the presence of systematic biases (see Fine Point 10.2 for details).

10.2 Estimation of causal effects

Suppose our heart transplant study was a marginally randomized experiment, and that the 20 individuals were a random sample of all individuals in a nearly infinite super-population of interest. Suppose further that all individuals in the super-population were randomly assigned to either $A = 1$ or $A = 0$, and that all of them adhered to their assigned treatment. Exchangeability of the treated and the untreated would hold in the super-population, i.e., $\Pr[Y^a = 1] = \Pr[Y = 1|A = a]$, and therefore the causal risk difference $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$ equals the associational risk difference $\Pr[Y = 1|A = 1] - \Pr[Y = 1|A = 0]$ in the super-population.

Because our study population is a random sample of the super-population, the sample proportion of individuals that develop the outcome among those with observed treatment value $A = a$, $\widehat{\Pr}[Y = 1 \mid A = a]$, is an unbiased estimator of the super-population probability $\Pr[Y = 1|A = a]$. Because of exchangeability in the super-population, the sample proportion $\widehat{\Pr}[Y = 1 \mid A = a]$ is also an unbiased estimator of $\Pr[Y^a = 1]$. Thus, traditional statistical "testing"of the causal null hypothesis $\Pr[Y^{a=1} = 1] = \Pr[Y^{a=0} = 1]$ boils down to comparing the sample proportions $\widehat{\Pr}[Y = 1 \mid A = 1] = 7/13$ and $\widehat{\Pr}[Y = 1 \mid A = 0] = 3/7$. Standard statistical methods can also be used to compute 95% confidence intervals for the causal risk difference and causal risk ratio in the super-population, which are estimated by $(7/13) - (3/7)$ and $(7/13)/(3/7)$, respectively. Slightly more involved, but standard, statistical procedures are used in observational studies to obtain confidence intervals for standardized, IP weighted, or stratified association measures.

There is an alternative way to think about sampling variability in randomized experiments. Suppose only individuals in the study population, not all individuals in the super-population, are randomly assigned to either $A = 1$

Technical Point 10.1

Bias and consistency in statistical inference. We have discussed systematic bias (due to unknown sources of confounding, selection, or measurement error) and consistent estimators in earlier chapters. Here we discuss these and other concepts of bias, and describe how they are related.

To provide a formal definition of consistent estimator for an estimand θ, suppose we observe n independent, identically distributed (i.i.d.) copies of a vector-valued random variable whose distribution P lies in a set $\mathcal{M}$ of distributions (our model). Then the estimator $\widehat{\theta}_n$ is consistent for $\theta = \theta(P)$ in model $\mathcal{M}$ if $\widehat{\theta}_n$ converges to θ in probability for every $P \in \mathcal{M}$ i.e.

$$\Pr_P\left[|\widehat{\theta}_n - \theta(P)| > \varepsilon\right] \to 0 \text{ as } n \to \infty \text{ for every } \varepsilon > 0, P \in \mathcal{M}.$$

The estimator $\widehat{\theta}_n$ is exactly unbiased in model $\mathcal{M}$ if, for every $P \in \mathcal{M}, \mathrm{E}_P\left[\widehat{\theta}_n\right] = \theta(P)$. The exact bias under P is the difference $\mathrm{E}_P\left[\widehat{\theta}_n\right] - \theta(P)$. We denote the estimator by $\widehat{\theta}_n$ rather than by simply $\widehat{\theta}$ to emphasize that the estimate depends on the sample size n. On the other hand, the parameter $\theta(P)$ is a fixed, though unknown, quantity depending on $P \in \mathcal{M}$. When P is the distribution generating the data in our study, we often suppress the P in the notation and write $\mathrm{E}\left[\widehat{\theta}_n\right] = \theta$. For many parameters θ, such as the risk ratio $\Pr[Y=1|A=1]/\Pr[Y=1|A=0]$, exactly unbiased estimators do not exist.

A systematically biased estimator is neither consistent nor exactly unbiased. Robins and Morgenstern (1987) argue that most applied researchers (e.g., epidemiologists) will declare an estimator unbiased only if it can center a valid Wald confidence interval. They show that under this definition, an estimator is only unbiased if it is uniformly asymptotic normal and unbiased (UANU), as only UANU estimators can center valid standard Wald intervals for $\theta(P)$ under the model $\mathcal{M}$. An estimator $\widehat{\theta}_n$ is UANU in model $\mathcal{M}$ if there exists sequences $\sigma_n(P)$ such that the z-statistic $\left(\widehat{\theta}_n - \theta(P)\right)/\sigma_n(P)$ converges uniformly to a standard normal random variable in the following sense: for $t \in R$,

$$\sup_{P \in \mathcal{M}} \left|\Pr_P\left[n^{1/2}\left(\widehat{\theta}_n - \theta(P)\right)/\sigma_n(P) < t\right] - \Phi(t)\right| \to 0 \text{ as } n \to \infty$$

where $\Phi(t)$ is the standard normal cumulative distribution function (Robins and Ritov,1997).

All inconsistent estimators and some consistent estimators (see Chapter 18 for examples) are biased under this definition. In this book, when we say an estimator is unbiased (without further qualification) we mean that it is UANU.

or $A = 0$. Because of the presence of random sampling variability, we do not expect that exchangeability will exactly hold in our sample. For example, suppose that only the 20 individuals in our study were randomly assigned to either heart transplant ($A = 1$) or medical treatment ($A = 0$). Suppose further that each individual can be classified as good or bad prognosis at the time of randomization. We say that the groups $A = 0$ and $A = 1$ are exchangeable if they include exactly the same proportion of individuals with bad prognosis. By chance, it is possible that 2 out of the 13 individuals assigned to $A = 1$ and 3 of the 7 individuals assigned to $A = 0$ had bad prognosis. However, if we increased the size of our sample then there is a high probability that the relative imbalance between the groups $A = 1$ and $A = 0$ would decrease.

Under this conceptualization, there are two possible targets for inference. First, investigators may be agnostic about the existence of a super-population and restrict their inference to the sample that was actually randomized. This is referred to as *randomization-based inference*, and requires taking into account some technicalities that are beyond the scope of this book. Second, investigators may still be interested in making inferences about the super-population from which the study sample was randomly drawn. From an inference stand-

See Robins (1988) and Greenland (1991) for a discussion of randomization-based inference.

Fine Point 10.2

Uncertainty from systematic biases. The width of the usual Wald-type confidence intervals is a function of the standard error of the estimator and thus reflects only uncertainty due to random error. However, the possible presence of systematic bias due to confounding, selection, or measurement is another important source of uncertainty. The larger the study population, the smaller the random error is both absolutely and as a proportion of total uncertainty, and hence the more the usual Wald confidence interval will understate the true uncertainty.

The stated 95% confidence in a 95% confidence interval becomes overconfidence as population size increases because the interval excludes uncertainty due to systematic biases, which are not diminished by increasing the sample size. As a consequence, some authors advocate referring to such intervals by a less confident name, calling them *compatibility intervals* instead. The renaming recognizes that such intervals can only show us which effect sizes are highly compatible with the data under our adjustment assumptions and methods (Amrhein et al. 2019; Greenland 2019). The compatibility concept is weaker than the confidence concept, for it does not demand complete confidence that our adjustment removes all systematic biases.

Regardless of the name of the intervals, the uncertainty due to systematic bias is usually a central part of the discussion section of scientific articles. However, most discussions revolve around informal judgments about the potential direction and magnitude of the systematic bias. Some authors argue that quantitative methods need to be used to produce intervals around the effect estimate that integrate random and systematic sources of uncertainty. These methods are referred to as quantitative bias analysis. See the book by Lash, Fox, and Fink (2009). Bayesian alternatives are discussed by Greenland and Lash (2008), and Greenland (2009a, 2009b).

point, this latter case turns out to be mathematically equivalent to the conceptualization of sampling variability described at the start of this section in which the entire super-population was randomly assigned to treatment. That is, randomization followed by random sampling is equivalent to random sampling followed by randomization.

In many cases we are not interested in the first target. To see why, consider a study that compares the effect of two first-line treatments on the mortality of cancer patients. After the study ends, we may determine that it is better to initiate one of the two treatments, but this information is now irrelevant to the actual study participants. The purpose of the study was not to guide the choice of treatment for patients in the study but rather for a group of individuals similar to—but larger than—the studied sample. Heretofore we have assumed that there is a larger group—the super-population—from which the study participants were randomly sampled. We now turn our attention to the concept of the super-population.

10.3 The myth of the super-population

As discussed in Chapter 1, there are two sources of randomness: sampling variability and nondeterministic counterfactuals. Below we discuss both.

Consider our estimate $\widehat{\Pr}[Y = 1 \mid A = 1] = \hat{p} = 7/13$ of the super-population risk $\Pr[Y = 1|A = a] = p$. Nearly all investigators would report a binomial confidence interval $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 7/13 \pm 1.96\sqrt{\frac{(7/13)(6/13)}{13}}$ for the probability p. If asked why these intervals, they would say it is to incorporate the uncertainty due to random variability. But these intervals are valid only if $\hat{p}$ has a binomial sampling distribution. So we must ask when would that happen. In fact there are two scenarios under which $\hat{p}$ has a binomial sampling

distribution.

- *Scenario 1.* The study population is sampled at random from an essentially infinite super-population, sometimes referred to as the source or target population, and our estimand is the proportion $p = \Pr[Y = 1|A = 1]$ of treated individuals who developed the outcome in the super-population. It is then mathematically true that, in repeated random samples of size 13 from the treated individuals in the super-population, the number of individuals who develop the outcome among the 13 is a binomial random variable with success probability $\Pr[Y = 1|A = 1]$. As a result, the 95% Wald confidence interval calculated in the previous section is asymptotically calibrated for $\Pr[Y = 1|A = 1]$. This is the model we have considered so far.

The term i.i.d. used in Technical Point 10.1 means that our data were a random sample of size n from a super-population.

- Scenario 2. The study population is not sampled from a hypothetical super-population. Rather (i) each individual i among the 13 treated individuals has an individual nondeterministic (stochastic) counterfactual probability $p_i^{a=1}$ (ii) the observed outcome $Y_i = Y_i^{a=1}$ for subject i occurs with probability $p_i^{a=1}$ and (iii) $p_i^{a=1}$ takes the same value, say p, for each of the 13 treated individuals. Then the number of individuals who develop the outcome among the 13 treated is a binomial random variable with success probability p. As a result, the 95% confidence interval calculated in the previous section is asymptotically calibrated for p.

Robins (1988) discussed these two scenarios in more detail.

Scenario 1 assumes a hypothetical super-population. Scenario 2 does not. However, Scenario 2 is untenable because the probability $p_i^{a=1}$ of developing the outcome when treated will almost certainly vary among the 13 treated individuals due to between-individual differences in risk. For example we would expect the probability of death $p_i^{a=1}$ to have some dependence on an individual's genetic make-up. If the $p_i^{a=1}$ are nonconstant then the estimand of interest in the actual study population would generally be the average, say p, of the 13 $p_i^{a=1}$. But in that case the number of treated who develop the outcome is not a binomial random variable with success probability p, and the 95% confidence interval for p calculated in the previous section is not asymptotically calibrated but conservative.

Therefore, any investigator who reports a binomial confidence interval for $\Pr[Y = 1|A = a]$, and who acknowledges that there exists between-individual variation in risk, must be implicitly assuming Scenario 1: the study individuals were sampled from a near-infinite super-population and that all inferences are concerned with quantities from that super-population. Under Scenario 1, the number with the outcome among the 13 treated is a binomial variable regardless of whether the underlying counterfactual is deterministic or stochastic.

An advantage of working under the hypothetical super-population scenario is that nothing hinges on whether the world is deterministic or nondeterministic. On the other hand, the super-population is generally a fiction; in most studies individuals are not randomly sampled from any near-infinite population. Why then has the myth of the super-population endured? One reason is that it leads to simple statistical methods.

A second reason has to do with generalization. As we mentioned in the previous section, investigators generally wish to generalize their findings about treatment effects from the study population (e.g., the 20 individuals in our heart transplant study) to some large target population (e.g., all immortals in the Greek pantheon). The simplest way of doing so is to assume the study population is a random sample from a large population of individuals who

are potential recipients of treatment. Since this is a fiction, a 95% confidence interval computed under Scenario 1 should be interpreted as covering the super-population parameter had, often contrary to fact, the study individuals been sampled randomly from a near infinite super-population. In other words, confidence intervals obtained under Scenario 1 should be viewed as what-if statements.

It follows from the above that an investigator might not want to entertain Scenario 1 if the size of the pool of potential recipients is not much larger than the size of the study population, or if the target population of potential recipients is believed to differ from the study population to an extent that cannot be accounted for by sampling variability. Here we will accept that individuals were randomly sampled from a super-population, and explore the consequences of random variability for causal inference in that context. We first explore this question in a simple randomized experiment.

10.4 The conditionality "principle"

Table 10.1 summarizes the data from a randomized trial to estimate the average causal effect of treatment A (1: yes, 0: no) on the 1-year risk of death Y (1: yes, 0: no). The experiment included 240 individuals, 120 in each treatment group. The associational risk difference is $\Pr[Y=1|A=1] - \Pr[Y=1|A=0] = \frac{24}{120} - \frac{42}{120} = -0.15$. Suppose the experiment had been conducted in a super-population of near-infinite size, the treated and the untreated would be exchangeable, i.e., $Y^a \perp\!\!\!\perp A$, and the associational risk difference would equal the causal risk difference $\Pr\left[Y^{a=1}=1\right] - \Pr\left[Y^{a=0}=1\right]$. Suppose the study investigators computed a 95% confidence interval $(-0.26, -0.04)$ around the point estimate -0.15 and published an article in which they concluded that treatment was beneficial because it reduced the risk of death by 15 percentage points.

However, the study population had only 240 individuals and is therefore likely that, due to chance, the treated and the untreated are not perfectly exchangeable. Random assignment of treatment does not guarantee exact exchangeability for the sample consisting of the 240 individuals in the trial; it only guarantees that any departures from exchangeability are due to random variability rather than to a systematic bias. In fact, one can view the uncertainty resulting from our ignorance of the chance correlation between unmeasured baseline risk factors and the treatment A in the study sample as contributing to the length 0.22 of the confidence interval.

A few months later the investigators learn that information on a third variable, cigarette smoking L (1: yes, 0: no), had also been collected and decide to take a look at it. The study data, stratified by L, is shown in Table 10.2. Unexpectedly, the investigators find that the proportion of individuals receiving treatment among smokers (80/120) is twice that among nonsmokers (40/120), which suggests that the treated and the untreated are not exchangeable and thus that adjustment for smoking is necessary. When the investigators adjust via stratification, the associational risk difference in smokers, $\Pr[Y=1|A=1,L=1] - \Pr[Y=1|A=0,L=1]$, is equal to 0. The associational risk difference in nonsmokers, $\Pr[Y=1|A=1,L=0] - \Pr[Y=1|A=0,L=0]$, is also equal to 0. The adjusted analysis suggests treatment has no effect in both smokers and nonsmokers, even though the marginal risk difference -0.15 suggested a net beneficial effect in the study population.

The estimated variance of the unadjusted estimator is $\frac{\frac{24}{120}\frac{96}{120}}{120} + \frac{\frac{42}{120}\frac{78}{120}}{120} = \frac{31}{9600}$. The Wald 95% confidence interval is then $-0.15 \pm \left(\frac{31}{9600}\right)^{1/2} \times 1.96 = (-0.26, -0.04)$.

Table 10.1

	$Y=1$	$Y=0$
$A=1$	24	96
$A=0$	42	78

Table 10.2

$L=1$	$Y=1$	$Y=0$
$A=1$	4	76
$A=0$	2	38

$L=0$	$Y=1$	$Y=0$
$A=1$	20	20
$A=0$	40	40

Technical Point 10.2

A formal statement of the conditionality principle. The likelihood for the observed data has three factors: the density of Y given A and L, the density of A given L, and the marginal density of L. Consider a simple example with one dichotomous L, exchangeability $Y^a \perp\!\!\!\perp A|L$, the stratum-specific risk difference $sRD = \Pr(Y = 1|L = l, A = 1) - \Pr(Y = 1|L = l, A = 0)$ known to be constant across strata of L, and in which the parameter of interest is the stratum-specific causal risk difference. Then the likelihood of the data is

$$\prod_{i=1}^{n} f(Y_i|L_i, A_i; sRD, p_0) \times f(A_i|L_i; \alpha) \times f(L_i; \rho)$$

where $p_0 = (p_{01}, p_{02})$ with $p_{0l} = \Pr(Y = 1|L = l, A = 0)$, α, and ρ are nuisance parameters associated with the conditional density of Y given A and L, the conditional density of A given L, and the marginal density of L, respectively. See, for example, Casella and Berger (2002).

The data on A and L are said to be S-ancillary for the parameter of interest when, as in this case, the distribution of the data conditional on these variables depends on the parameter of interest, but the joint density of A and L does not share parameters with $f(Y_i|L_i, A_i; sRD, p_0)$. The conditionality principle states that one should always perform inference on the parameter of interest conditional on any S-ancillary statistics. Thus one should condition on the S-ancillary statistic $\{A_i, L_i; i = 1, .., n\}$. Analogously, if the risk ratio (rather than the risk difference) were known to be constant across strata of L, $\{A_i, L_i; i = 1, .., n\}$ remains S-ancillary for the risk ratio.

An exact ancillary statistic is defined to be an S-ancillary statistic whose marginal distribution is known. In our example, this would require that α and ρ be known.

The estimated variance of the adjusted estimator is described in Technical Point 10.5. The Wald 95% confidence interval is then $(-0.076, 0.076)$.

These new findings are disturbing to the investigators. Either someone did not assign the treatment at random (malfeasance) or randomization did not result in approximate exchangeability (very very bad luck). A debate ensues among the investigators. Should they retract their article and correct the results? They all agree that the answer to this question would be affirmative if the problem were due to malfeasance. If that were the case, there would be confounding by smoking and the effect estimate should be adjusted for smoking. But they all agree that malfeasance is impossible given the study's quality assurance procedures. It is therefore clear that the association between smoking and treatment is entirely due to bad luck. Should they still retract their article and correct the results?

One investigator says that they should not retract the article. His argument goes as follows: "Okay, randomization went wrong for smoking, but why should we privilege the adjusted over the unadjusted estimator? It is likely that imbalances on other unmeasured factors U cancelled out the effect of the chance imbalance on L, so that the unadjusted estimator is still the closer to the true value in the super-population." A second investigator says that they should retract the article and report the adjusted null result. Her argument goes as follows: "We should adjust for L because the strong association between L and A introduces confounding in our effect estimate. Within levels of L, we have mini randomized trials and the confidence intervals around the corresponding point estimates will reflect the uncertainty due to the possible U-A associations conditional on L."

To determine which investigator is correct, here are the facts of the matter. Suppose, for simplicity, the true causal risk difference is constant across strata of L, and suppose we could run the randomized experiment trillions of times. We then select only (i.e., condition on) those runs in which smoking L and

Technical Point 10.3

Approximate ancillarity. Suppose that the stratum-specific risk difference (sRD_l) is known to vary over strata of L. Under our usual identifiability assumptions, the causal risk difference in the population is identified by the standardized risk difference

$$RD_{std} = \sum_l \left[\Pr\left(Y=1|L=l, A=1; \upsilon\right) - \Pr\left(Y=1|L=l, A=0; \upsilon\right)\right] f\left(l; \rho\right)$$

which depends on the parameters $\upsilon = \{sRD_l, p_{0,l}; l = 0, 1\}$ and ρ (see Technical Point 10.2). In unconditionally randomized experiments, RD_{std} equals the associational RD, $\Pr\left(Y=1|A=1\right) - \Pr\left(Y=1|A=0\right)$, because $A \perp\!\!\!\perp L$ in the super-population. Due to the dependence of RD_{std} on ρ, $\{A_i, L_i; i = 1, .., n\}$ is no longer exactly ancillary and in fact no exact ancillary exists.

Consider the statistic $\widetilde{S} = \widehat{OR}_{AL} - OR_{AL}$ where $OR_{AL} = OR_{AL}\left(\alpha\right) = \frac{\Pr(A=1|L=1;\alpha)\Pr(A=0|L=0;\alpha)}{\Pr(A=1|L=0;\alpha)\Pr(A=0|L=1;\alpha)}$ is the A-L odds ratio in the super-population, and $\widehat{OR}_{AL}$ is OR_{AL} but with the the population proportions $\Pr\left(A=a|L=l;\alpha\right)$ replaced by the empirical sample proportions $\widehat{\Pr}\left(A=a|L=l\right)$. $\widetilde{S}$ is asymptotically normal with mean 0 conditional on the L_i and thus its distribution depends on α. Let $\widehat{S} = \widetilde{S}/\widehat{se}(\widetilde{S})$, where $\widehat{se}(\widetilde{S})$ is an estimate of the standard error of $\widetilde{S}$. The distribution of $\widehat{S}$ converges to a standard normal distribution in large samples, so that $\widehat{S}$ quantifies the A-L association in the data on a standardized scale. For example, if $\widehat{S} = 2$, then $\widehat{S}$ is two standard deviations above its (asymptotic) expected value of 0.

When the true value of OR_{AL} is known, $\widehat{S}$ is referred to as an approximate (or large sample) ancillary statistic. To see why, consider a randomized experiment with $OR_{AL} = 1$. Then $\widehat{S}$, like an exact ancillary statistic, i) can be computed from the data (i.e., $\widehat{S} = \left(\widehat{OR}_{AL} - 1\right)/\widehat{se}(\widetilde{S})$), ii) $\widehat{S}$ has an approximately known distribution, iii) the likelihood factors into a term $f\left(A|L;\alpha\right)$ that governs the distribution of $\widetilde{S}$ and a term $f\left(Y|L, A; \upsilon\right) f\left(L; \rho\right)$ that does not depend on α, and iv) conditional on $\widehat{S}$, the adjusted estimate of RD_{std} is unbiased, while the unadjusted estimate of RD_{std} is biased (Technical Point 10.4 defines and compares adjusted and unadjusted estimators). Any other statistic that quantifies the A-L association $\frac{\widehat{\Pr}(A=1|L=1)}{\widehat{\Pr}(A=1|L=0)} - 1$, can be used in place of $\widetilde{S}$.

Now consider a *continuity principle* wherein inferences about an estimand should not change discontinuously in response to an arbitrarily small known change in the data generating distribution (Buehler 1982). If one accepts both the conditionality and continuity principles, then one should condition on an approximate ancillary statistic. For example, when $OR_{AL} = 1$ is known, the continuity principle would be violated if, following the conditionality principle, we treated the unadjusted estimate of RD_{std} as biased when sRD_l was known to be a constant, but treated it as unbiased when the sRD_l were almost constant. We will say that a researcher who always conditions on both exact and approximate ancillaries follows the extended conditionality principle.

treatment A are as strongly positively associated as in the observed data. We would find that, within each level of L, the fraction of these runs in which any given pre-treatment risk factor U for Y was positively associated with A essentially equals the number of runs in which it was negatively associated. (This is true even if U and L are highly correlated in both the super-population and in the study data.)

As a consequence, the adjusted estimate of the treatment effect is unbiased but the unadjusted estimate is greatly biased when averaged over these runs. Unconditionally—over all the runs of the experiment—both the unadjusted and adjusted estimates are unbiased but the variance of the adjusted estimate is smaller than that of the unadjusted estimate. That is, the adjusted estimator is both conditionally unbiased and unconditionally more efficient. Hence either from the conditional or unconditional point of view, the Wald interval centered on the adjusted estimator is the better analysis and the article needs to be retracted. The second investigator is correct.

The unconditional efficiency of the adjusted estimator results from the adjusted estimator being the maximum likelihood estimator (MLE) of the risk difference when data on L are available.

Technical Point 10.4

Comparison between adjusted and unadjusted estimators. The adjusted estimator of RD_{std} in Technical Point 10.3 is the parametric maximum likelihood estimator $\widehat{RD}_{MLE}$, which replaces the population proportions in the RD_{std} by their sample proportions. The unadjusted estimator of RD_{std} is $\widehat{RD}_{UN} = \widehat{\Pr}(Y=1|A=1) - \widehat{\Pr}(Y=1|A=0)$. Unconditionally, both $\widehat{RD}_{MLE}$ and $\widehat{RD}_{UN}$ are asymptotically normal and unbiased for RD_{std} with asymptotic variances $aVar\left(\widehat{RD}_{MLE}\right)$ and $aVar\left(\widehat{RD}_{UN}\right)$.

In the text we stated that $\widehat{RD}_{UN}$ is both unconditionally inefficient and conditionally biased. We now explain that both properties are logically equivalent. Robins and Morgenstern (1987) prove that $\widehat{RD}_{MLE}$ has the same asymptotic distribution conditional on the approximate ancillary $\widehat{S}$ as it does unconditionally, which implies $aVar\left(\widehat{RD}_{MLE}\right) = aVar\left(\widehat{RD}_{MLE}|\widehat{S}\right)$. They also show that $aVar\left(\widehat{RD}_{MLE}\right)$ equals $aVar\left(\widehat{RD}_{UN}\right) - \left[aCov\left(\widehat{S}, \widehat{RD}_{UN}\right)\right]^2$. Hence $\widehat{RD}_{UN}$ is unconditionally inefficient if and only if $aCov\left(\widehat{S}, \widehat{RD}_{UN}\right) \neq 0$, i.e., $\widehat{S}$ and $\widehat{RD}_{UN}$ are correlated unconditionally. Further, the conditional asymptotic bias $a\,\mathrm{E}\left[\widehat{RD}_{UN}|\widehat{S}\right] - RD_{std}$ is shown to equal $aCov\left(\widehat{S}, \widehat{RD}_{UN}\right)\widehat{S}$. Hence, $\widehat{RD}_{UN}$ is conditionally biased if and only if it is unconditionally inefficient.

It can be shown that $aCov\left(\widehat{S}, \widehat{RD}_{UN}\right) = 0$ if and only if $L \perp\!\!\!\perp Y|A$. Therefore, when data on a measured risk factor for Y are available, $\widehat{RD}_{MLE}$ is preferred over $\widehat{RD}_{UN}$. The estimator $\widehat{RD}_{UN} - aCov\left(\widehat{S}, \widehat{RD}_{UN}\right)\widehat{S}$ corrects the bias of $\widehat{RD}_{UN}$, and thus has the same asymptotic distribution as $\widehat{RD}_{MLE}$ given the approximate ancillary $\widehat{S}$.

The idea that one should condition on the observed L-A association is an example of what is referred to in the statistical literature as *the conditionality principle*. In statistics, the observed L-A association is said to be an ancillary statistic for the causal risk difference. The conditionality principle states that inference on a parameter should be performed conditional on ancillary statistics (see Technical Points 10.2 and 10.3 for details).

In the above discussion about the findings of the randomized experiment, some of the investigators intuitively followed the conditionality principle because they considered an estimator to be biased when it cannot center a valid Wald confidence interval conditional on any ancillary statistics. For such researchers, our previous definition of bias was not sufficiently restrictive. They would say that an estimator is unbiased if and only if it can center a valid Wald interval conditional on ancillary statistics. Technical Point 10.5 argues that most researchers implicitly follow the conditionality principle.

When confronted with the frequentist argument that "Adjustment for L is unnecessary because unconditionally—over all the runs of the experiment—the unadjusted estimate is unbiased," investigators that intuitively apply the conditionality principle would aptly respond "Why should the various L-A associations in other hypothetical studies affect what I do in my study? In my study L acts as a confounder and adjustment is needed to eliminate bias." This is a convincing argument for both randomized experiments and observational studies as long as, like in the randomized experiment of our example, the number of measured confounders is not large. However, when the number of measured confounders is large, strictly following the conditionality principle is no longer a wise strategy.

Technical Point 10.5

Most researchers intuitively follow the extended conditionality principle. Consider again the randomized trial data in Table 10.2. Assuming without loss of generality that the sRD is constant over the strata of a dichotomous L, the estimated variance of the MLE of sRD is $\widehat{V}_0\widehat{V}_1/\left(\widehat{V}_0+\widehat{V}_1\right)$ where $\widehat{V}_l$ is the estimated variance of $\widehat{RD}_l$.

Two possible choices for $\widehat{V}_1$ are $\widehat{V}_1^{obs}=\frac{\frac{4}{80}\frac{76}{80}}{80}+\frac{\frac{2}{40}\frac{38}{40}}{40}=1.78\times10^{-3}=$ and $\widehat{V}_1^{exp}=\frac{\frac{4}{80}\frac{76}{80}}{60}+\frac{\frac{2}{40}\frac{38}{40}}{60}=1.58\times10^{-3}$ that differ only in that $\widehat{V}_1^{obs}$ divides by the observed number of individuals in stratum $L=1$ with $A=1$ and $A=0$ (80 and 40, respectively) while $\widehat{V}_1^{exp}$divides by the expected number of subjects (60) given that $A\perp\!\!\!\perp L$. Mathematically, $\widehat{V}_1^{obs}$is the variance estimator based on the observed information and $\widehat{V}_1^{exp}$ is the estimator based on the expected information.

In our experience, nearly all researchers would choose $\widehat{V}_1^{obs}$ over $\widehat{V}_1^{exp}$as the appropriate variance estimator. Results of Efron and Hinkley (1978) and Robins and Morgenstern (1987) imply that such researchers are implicitly conditioning on an approximate ancillary $\widehat{S}$ and thus, whether aware of this fact or not, are following the extended conditionality principle. Specifically, these authors proved that the variance of $\widehat{RD}_l$, and thus of the MLE, conditioned on an approximate ancillary $\widehat{S}$ differs from the unconditional variance by order $n^{-3/2}$. (As noted in Technical Point 10.4, the conditional and unconditional asymptotic variance of an MLE are equal, as equality of asymptotic variances implies equality only up to order n^{-1}.) Further, they showed that the variance estimator based on the observed information differs from the conditional variance by less than order $n^{-3/2}$, while an estimator based on the expected information differs from the unconditional variance by less than $n^{-3/2}$. Thus, a preference for $\widehat{V}_1^{obs}$ over $\widehat{V}_1^{exp}$ implies a preference for conditional over unconditional inference.

10.5 The curse of dimensionality

The derivations in previous sections above are based on an asymptotic theory that assumed the number of strata of L was small compared with the sample size. In this section, we study the cases in which the number of strata of a vector L can be very large, even much larger than the sample size.

Suppose the investigators had measured 100 pre-treatment binary variables rather than only one, then the pre-treatment variable L formed by combining the 100 variables $L=(L_1,...,L_{100})$ has 2^{100} strata. When, as in this case, there are many possible combinations of values of the pre-treatment variables, we say that the data is of *high dimensionality*. For simplicity, suppose that there is no additive effect modification by L, i.e., the super-population risk difference $\Pr[Y=1|A=1,L=l]-\Pr[Y=1|A=0,L=l]$ is constant across the 2^{100} strata. In particular, suppose that the constant stratum-specific risk difference is 0.

The investigators debate again whether to retract the article and report their estimate of the stratified risk difference. They have by now agreed that they should follow the conditionality principle because the unadjusted risk difference -0.15 is conditionally biased. However, they notice that, when there are 2^{100} strata, a 95% confidence interval for the risk difference based on the adjusted estimator is much wider than that based on the unadjusted estimator. This is exactly the opposite of what was found when L had only two strata. In fact, the 95% confidence interval based on the adjusted estimator may be so wide as to be completely uninformative.

To see why, note that, because 2^{100} is much larger than the number of individuals (240), there will at most be only a few strata of L that will contain both a treated and an untreated individual. Suppose only one of 2^{100} strata contains a single treated individual and a single untreated individual, and no other stratum contains both a treated and untreated individual. Then the

Technical Point 10.6

Can the curse of dimensionality be reversed? In high-dimensional settings with many strata of L, informative conditional inference for the common risk difference given the exact ancillary statistic $\{A_i, L_i; i = 1, ..n\}$ is not possible regardless of the estimator used. This is not true for unconditional inference in marginally randomized experiments. For example, the unconditional statistical behavior of the unadjusted estimator $\widehat{RD}_{UN}$ is unaffected by the dimension of L. In particular, it remains unbiased with the width of the associated Wald 95% confidence interval proportional to $1/n^{1/2}$. Because $\widehat{RD}_{UN}$ relies on prior information not used by the MLE, it is an unbiased estimator of the common risk difference only if it is known that $A \perp\!\!\!\perp L$ in the super-population.

However, even unconditionally, the confidence intervals associated with the MLE, i.e., the adjusted estimator, remain uninformative. This raises the question of whether data on L can be used to construct an estimator that is also unconditionally unbiased but that is more efficient that the unadjusted estimator. In Chapter 18 we show that this is sometimes possible.

95% confidence interval for the common risk difference based on the adjusted estimator is $(-1, 1)$, and therefore completely uninformative, because in the single stratum with both a treated and an untreated individual, the empirical risk difference could be -1, 0, or 1 depending on the value of Y for each individual. In contrast, the 95% confidence interval for the common risk difference based on the unadjusted estimator remains $(-0.26, -0.04)$ as above because its width is unaffected by the fact that more covariates were measured. These results reflect the fact that the adjusted estimator is only guaranteed to be more efficient than the unadjusted estimator when the ratio of number of individuals to the number of unknown parameters is large (a frequently used rule of thumb is a minimum ratio of 10, though the minimum ratio depends on the characteristics of the data).

What should the investigators do? By trying to do the right thing—following the conditionality principle—in the simple setting with one dichotomous variable, they put themselves in a corner for the high-dimensional setting. This is the *curse of dimensionality*: conditional on all 100 covariates the marginal estimator is still biased, but now the conditional estimator is uninformative. This shows that, just because conditionality is compelling in simple examples, it should not be raised to a principle since it cannot be carried through for high-dimensional models. Though we have discussed this issue in the context of a randomized experiment, our discussion applies equally to observational studies. See Technical Point 10.6.

Robins and Ritov (1997) provide a technical description of the curse of dimensionality.

Finding a solution to the curse of dimensionality is a difficult problem and an active area of research. In Chapter 18 we review this research and offer some practical guidance. Chapters 11 through 17 provide necessary background information on the use of models for causal inference.

Technical Point 10.7

Implications of random variability for causal discovery. In Fine Point 6.3 we explained that, under faithfulness, we could sometimes learn the causal structure if we had an infinite amount of data. After the concepts introduced in this chapter, we are now ready to consider the implications for causal discovery of only having a finite sample.

Suppose we have data on 3 variables Z, A, Y and we know that their time sequence is Z first, A second, and Y last. Our data analysis finds that the empirical odds ratio of Y and Z equal to 1 at every level of A. All other odds ratios, marginal and conditional, are far from 1. In Fine Point 6.3 we showed that, if $Z \perp\!\!\!\perp Y|A$ in the super-population (which would require an infinite sample size) then, under faithfulness, the only possible causal diagram is $Z \rightarrow A \rightarrow Y$ with perhaps a common cause U of Z and A in addition to (or in place of) the arrow from Z to A. It follows that the risk difference $\mathrm{E}[Y|A=1] - \mathrm{E}[Y|A=0]$ is the average causal effect of A on Y. But, in practice, evidence of conditional or unconditional independence must be based on a finite sample size.

Robins et al. (2003) showed that, even if one is willing to assume faithfulness, inferences based on faithfulness are non-uniform, i.e., no matter how big the sample size n, even if the empirical odds ratio of Y and Z were equal to 1 at every level of A, there exist faithful distributions with the following properties: a) due to sampling variability, the true odds ratio of Y and Z at each level of A, although not equal to 1, is so close to 1 that empirical conditional odds ratios of 1 are unsurprising, and yet b) the average causal effect of A on Y is zero. As a consequence, no honest 95% frequentist confidence interval for the average causal effect of A on Y can ever exclude the value 0 even when the empirical risk difference estimate of $\mathrm{E}[Y|A=1] - \mathrm{E}[Y|A=0]$ is quite large (say, 0.2) and is many (say 30) times greater than its standard error.

Even so, advocates of causal discovery may cogently argue that, given the empirical data above, a Bayesian (with priors not depending on sample size) who believes in faithfulness will generally have a (highest posterior density) 95% credible interval for the average causal effect of A on Y that is nearly centered on the empirical risk difference, with width not much greater than the standard error of the empirical risk difference. Thus, this credible interval easily excludes zero whenever graphs with Z and Y d-separated by A are given a non-negligible prior probability.

The striking difference between the honest frequentist confidence intervals and these credible intervals is a consequence of the fact that Bayesian inference for causal effects can be very sensitive to choice of prior in the causal discovery setting. For example, many epidemiologists, including the authors, would argue that, in an observational study, the prior probability given to any causal diagram that lacks a common cause of A and Y (such as the graph $Z \rightarrow A \rightarrow Y$) should be essentially zero. To believe otherwise, A and Y must have had no common cause from the big bang till now. A Bayesian who shares our prior belief may have (depending on other aspects of the prior) a 95% credible interval much wider and with a center much closer to 0 than the credible interval described above.

In summary, in finite samples and even under faithfulness, data alone cannot distinguish the causal diagram $Z \rightarrow A \rightarrow Y$ under which $Z \perp\!\!\!\perp Y|A$ in the super-population from another causal diagram under which Z is almost independent of Y given A in the super-population. Therefore the validity of causal discovery from observational data relies heavily on a priori subject-matter knowledge about the plausibility of various causal diagrams.

Part II

Causal inference with models