Chapter 1
The Open Problem

Sherri Rose, Mark J. van der Laan

The debate over hormone replacement therapy (HRT) has been one of the biggest health discussions in recent history. Professional groups and nonprofits, such as the American College of Physicians and the American Heart Association, gave HRT their stamp of approval 15 years ago. Studies indicated that HRT was protective against osteoporosis and heart disease. HRT became big business, with millions upon millions of prescriptions filled each year. However, in 1998, the Heart and Estrogen-Progestin Replacement Study demonstrated increased risk of heart attack among women with heart disease taking HRT, and in 2002 the Women's Health Initiative showed increased risk for breast cancer, heart disease, and stroke, among other ailments, for women on HRT. Why were there inconsistencies in the study results?

Mammography gained relatively widespread acceptance as an effective tool for breast cancer screening in the 1980s. While there was still debate, several studies, including the Health Insurance Plan trial and the Swedish Two-County trial, demonstrated that mammography saved lives. This outweighed the minimal evidence against mammography. Thus, in 2009, many medical practitioners and nonprofits were surprised by the new recommendations from the U.S. Preventive Services Task Force. Among women without a family history, mammography was now only recommended for women aged 50 to 74. The previous guidelines started at age 40. Why was there a seemingly sudden paradigm shift?

A political scientist examines the effect of butterfly ballots in an election, which may in turn change local election laws. A group of economists studies the effect of microlending on the local economy in rural areas of Africa in hopes of promoting greater adoption of this practice. Public health policy decisions regarding how frequently to perform gynecological exams await the completion of several new investigations. The question then becomes, how does one translate the results from these studies, how do we take the information in the data, and draw effective conclusions?

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_1,
© Springer Science+Business Media, LLC 2011

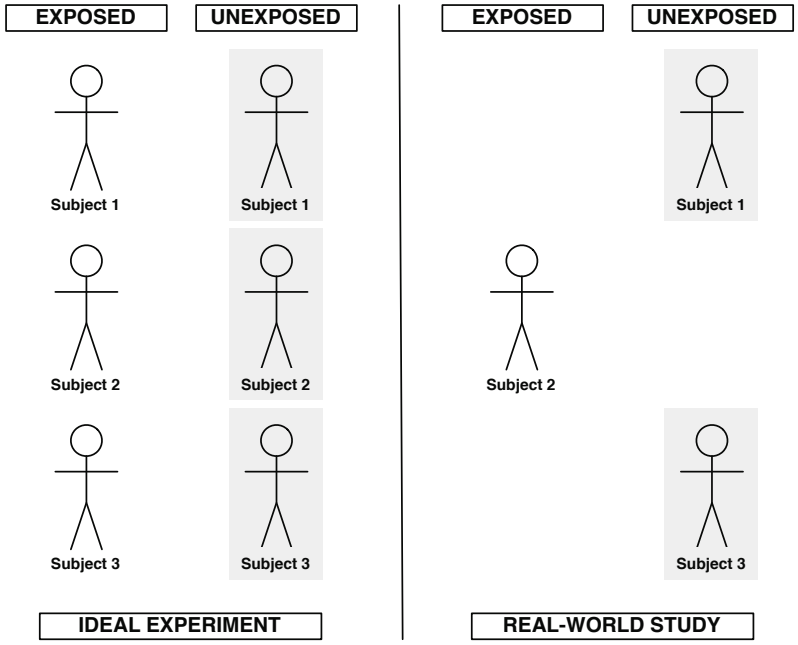

Fig. 1.1 Illustration of the "ideal experiment" vs. studies conducted in the real world

1.1 Learning from Data

One of the great open problems across many diverse fields of research has been obtaining causal effects from data. Data are typically sampled from a population of interest since collecting data on the entire population is not feasible. Frequently, the researcher is not interested in merely studying association or correlation in this sample data; she wants to know whether a treatment or exposure causes the outcome in the population of interest. If one can show that the treatment or exposure causes the outcome, we can then impact the outcome by intervening on the treatment or exposure.

Just what type of studies are we conducting? The often quoted "ideal experiment" is one that cannot be conducted in real life. Let us say we are interested in studying the causal effect of a toxin on death from cancer within 5 years. In an ideal experiment, we intervene and set the exposure to *exposed* for each subject. All subjects are followed for 5 years, where the outcome under this exposure is recorded. We then go back in time to the beginning of the study, intervene, and set all subjects to *not exposed* and *follow them under identical conditions* until the end of the study, recording the outcome under this exposure. As noted, we obviously cannot administer such a study since it is not possible to go back in time.

However, let's assume in principle there is a system where this ideal experiment could have been conducted. This experiment generates random variables. Say the experiment is that we sample a subject (i.e., draw a random variable) from a population and take several measurements on this subject. This experiment is repeated multiple times until we have sampled an a priori specified number (representing the sample size) of subjects. These random variables also have a true underlying probability distribution. Our observed data are realizations of these random variables. If we were to conduct our repeated experiment again, we would observe different realizations of these random variables.

Any knowledge we have about how these observed data were generated is referred to as a model. For example, it might be known that the data consist of observations on a number of independent and identically distributed (i.i.d.) random variables. What does i.i.d. mean? We are repeatedly drawing random variables from the same probability distribution, but each draw is mutually independent from all others. A common toy example used in statistics texts is the roll of a fair die. Say the experiment is to roll a die. We perform this experiment six times, each time rolling it following the same procedure (e.g., shaking). If we roll the die six times, we will see one set of realizations of these random variables, e.g., we observe a 1, 5, 3, 6, 2, and then a 5. Each roll of the die (i.e., each experiment) is independent from the previous roll. The observed unit in many cases may be the individual, where we sample repeatedly individual subjects from a population of interest. However, the observed unit can also be a household of individuals or a community of people.

So, our data are i.i.d. random variables, but the probability distribution of the random variable is typically completely unknown. This is also information we incorporate into our model. We will refer to this as a nonparametric model for the probability distribution of the random variable. (Do note, however, that assuming the data vector is i.i.d. in our nonparametric model is a real assumption, although one we will always make in this book.) Our model should always reflect true knowledge about the probability distribution of the data, which may often be a nonparametric model, or a semiparametric model that makes some additional assumptions. For example, perhaps it is known that the probability of death is monotonically increasing in the levels of exposure, and we want to include this information in our model.

The knowledge we have discussed thus far regarding our model pertains to our observed data and what we call the statistical model. The statistical model is, formally, the collection of possible probability distributions. The model may also contain extra information in addition to the knowledge contained in the statistical model. Now we want to relate our observed data to a causal model. We can do this with additional assumptions, and we refer to a statistical model augmented with these additional causal assumptions as the model for the observed data. These additional assumptions allow us to define the system where this ideal experiment could have been conducted. We can describe the generation of the data with nonparametric structural equations, intervene on treatment or exposure and set those values to *exposed* and *not exposed*, and then see what the (counterfactual) outcomes would

have been under both exposures. This underlying causal model allows us to define a causal effect of treatment or exposure.

One now needs to specify the relation between the observed data on a unit and the full data generated in the causal model. For example, one might assume that the observed data corresponds with observing all the variables generated by the system of structural equations that make up the causal model, up till background factors that enter as error terms in the underlying structural equations. The specification of the relation between the observed data and this underlying causal model allows one now to assess if the causal effect of interest can be identified from the probability distribution of the observed data. If that is not possible, then we state that the desired causal effect is not identifiable. If, on the other hand, our causal assumptions allow us to write the causal effect as a particular feature of the probability distribution of the observed data, then we have identified a target parameter of the probability distribution of the observed data that can be interpreted as a causal effect.

Let's assume that the causal effect is identifiable from the observed data. Our parameter of interest, here the causal effect of a toxin on death from cancer within 5 years, is now a parameter of our true probability distribution of the observed data. This definition as a parameter of the probability distribution of the observed data does not rely on the causal assumptions coded by the underlying causal model describing the ideal experiment for generating the desired full data, and the link between the observed data and the full data. Thus, if we ultimately do not believe these causal assumptions, the parameter is still an interesting statistical parameter. Our next goal becomes estimating this parameter of interest.

The open problem addressed in this book is the estimation of interesting parameters of the probability distribution of the data. This need not only be (causal) effect measures. Another problem researchers are frequently faced with is the generation of functions for the prediction of outcomes. For these problems, we do not make causal assumptions, but still define our realistic nonparametric or semiparametric statistical model based on actual knowledge. We view effect and prediction parameters of interest as features of the probability distribution of our data, well defined for each probability distribution in the nonparametric or semiparametric model. Statistical learning from data is concerned with efficient and unbiased estimation of these features and with an assessment of uncertainty of the estimator. Traditional approaches to estimation differ from this philosophy.

1.2 Traditional Approach to Estimation

We can sometimes implement one element of the ideal experiment: assigning a value for treatment or exposure in a controlled experiment. Controlled experiments are exactly what they sound like: they allow the investigator to control certain variables in the study. Randomized controlled trials (RCTs) are one type of controlled experiment where subjects are randomized to receive a specific level of treatment. For example, if each subject was assigned to one of two levels of treatment based on

the flip of a fair coin, the differences between the two groups would be solely due to treatment as all other factors would be balanced, up to random error. However, most studies are so-called observational studies where exposure or treatment is not assigned. In many cases it may not be ethical to set the exposure of interest in an RCT, or an RCT is cost prohibitive.

1.2.1 Experimental Studies

The randomization in RCTs suggests that we can estimate the causal effect of the treatment. For example, the difference of means between the treatment and control groups equals an additive causal effect. Indeed, this randomization of treatment in RCTs allows us to go from the observed data to the causal effect of interest. The difference in means can be estimated using a saturated regression of the outcome on treatment in a parametric statistical model where covariates are ignored. Since the regression is saturated (i.e., there is a parameter for each of the two observed values of treatment), this parametric statistical model is not making any unreasonable assumptions, and is thus actually nonparametric. Therefore, this parametric statistical model is not wrong, although the resulting estimator of the causal effect of the treatment is not the most efficient estimator. This so-called unadjusted estimator of the treatment effect is a nonparametric maximum likelihood estimator based on the reduced observations that only consist of the outcome and the treatment.

Suppose randomization did not occur perfectly due to chance (as is common), and there is a single covariate that is predictive of the outcome. We now have more subjects in the treatment or control group with a covariate that is predictive of the outcome, and this saturated regression ignoring the covariate will potentially contain a lot of residual error due to the exclusion of the covariate. Now, one might propose conditioning on the covariate and taking the difference in means for each stratum of the covariate. This results in a treatment effect within each stratum of the covariate. One might now estimate the causal effect of treatment as the average over all strata of these strata-specific treatment effects. This adjusted estimator of the treatment effect is a nonparametric maximum likelihood estimator based on the reduced observations that consist of the outcome, treatment, and this single covariate. This approach is generally still not efficient, since it only uses one of the measured covariates, but it is more efficient than the unadjusted treatment effect estimator. However, this strategy is not practical with multiple covariates, or even one continuous covariate, and starts to suffer in practical performance due to strata with a very small number of subjects.

So why not run regressions in parametric statistical models (incorporating all covariates) for RCTs? The short answer is simple: the Food and Drug Administration (FDA) does not allow it. We will explain why this is so in a few sections. For now it is sufficient to know that the FDA requires researchers to specify a priori the method of estimation, and it must rely on a statistical model that reflects true knowledge.

1.2.2 Observational Studies

Recall that observational studies do not involve randomization to treatment or exposure. In most observational studies, standard practice for effect estimation involves assuming a parametric statistical model and using maximum likelihood estimation to estimate the parameters in that statistical model. Let us be very clear again about what a statistical model is: the statistical model represents the set of possible probability distributions of the data.

In traditional practice, one assumes the actual data as observed in practice can be represented as observations of n i.i.d. random variables, and that the goal of the traditional modeling approach is to learn the true underlying probability distribution that generated the data. (This is different than the goal of causal effect estimation.) Maximum likelihood estimation uses the likelihood function to estimate the unknown parameter(s) in the statistical model. Solutions are often found by differentiating the log-likelihood with respect to these parameters, setting the resulting equation equal to zero, and solving. If the score equation has multiple solutions, the solution with the largest likelihood is selected.

This procedure is detailed in most introductory statistics books, although the pervasiveness of statistical software allows the user to implement maximum likelihood estimation without the need to understand these concepts. This also means the assumptions that come with the use of parametric statistical models are frequently not well understood or ignored.

We already acknowledged in Sect. 1.1 that we usually know very little about how our data were generated; thus the use of parametric statistical models is troublesome. We typically know that our data can be represented as a number (representing the sample size) of i.i.d. observations, which is an assumption in parametric statistical models, but we do not know the underlying probability distribution that generated the data. Parametric statistical models assume the underlying probability distribution that generated the data is known up to a finite number of parameters. It is an accepted fact within the statistical community that nonsaturated *parametric statistical models are wrong*. Thus, making an assumption known to be untrue is not the best approach. When this assumption is violated and the statistical model is misspecified, the estimate of the probability distribution can be extremely biased, and it is not even clear what the parameter estimates are even estimating. The bias resulting from statistical model misspecification cannot be overcome with a large sample size.

This brings us to another problem that arises when using misspecified parametric statistical models. The target parameter is not defined as a parameter of the true probability distribution for any possible probability distribution. The target parameter, when defined as a coefficient in a (misspecified) parametric statistical model, is only defined within that parametric statistical model, as if the statistical model were true. There is only correct inference if the parametric statistical model is correct, but we know it is wrong.

Lastly, the traditional approach does not make any explicit (untestable) causal assumptions linking the observed data to a system that generated the data. Thus, there

is no framework to make causal inference. There are also other assumptions that are part of the statistical model that are typically not addressed, such as positivity (discussed in Chaps. 2 and 10). When this (testable) assumption is violated, you may see groups of individuals where there is no experimentation in the treatment. For example, all the highly educated women received HRT, or all the wealthy women received mammograms. Since there are strata of certain covariates (e.g., level of education, socioeconomic status) where all subjects are treated, the regression will extrapolate what would have happened to these subjects had they not been treated, and this extrapolation is not based on any observed information.

To summarize, the use of parametric statistical models in observational studies is troublesome for several main reasons.

1. The statistical models are always misspecified in practice since we do not know the underlying data-generating distribution and we handle complex problems with many covariates.
2. The target parameter is not defined as a parameter of the true probability distribution that generated the data.
3. The traditional approach does not typically make causal assumptions allowing us to define the desired causal effect, and often neglects other key assumptions, such as the positivity assumption, that are part of the statistical model.

1.2.3 Regression in (Misspecified) Parametric Statistical Models

In this section we discuss briefly the traditional approach to effect estimation. Let us introduce our random variable O, which has probability distribution P_0. This is written $O \sim P_0$. Recall that a probability distribution P_0 assigns a probability to any possible event or set of possible outcomes for O. In particular, $P_0(O = o)$ for a particular value o of O can be defined as a probability if O is a discrete random variable, or we can use the concept of probability density if O is continuous. For simplicity and sake of presentation, we will often treat O as discrete so that we can refer to $P_0(O = o)$ as a probability.

We observe our random variable O n times, by repeating the same experiment n times. For a simple example, suppose our data structure is $O = (W, A, Y) \sim P_0$. We have a covariate or vector of covariates W, an exposure or treatment A, and a continuous outcome Y. These variables comprise the random variable O, which we observe repeatedly, and O has probability distribution P_0. Thus, for each possible value (w, a, y), $P_0(w, a, y)$ denotes the probability that (W, A, Y) equals (w, a, y). For example, the random variables $O_1, \ldots, O_n$ might be the result of randomly sampling n subjects from a population of patients, collecting baseline characteristics

W, assigning treatment or exposure A, and following the patients and measuring continuous outcome Y.

Suppose one poses a particular regression in a parametric statistical model, a so-called linear regression for the conditional mean of Y given A or Y given A and W. However, we leave the distributions of A and W unspecified. Linear regression in a parametric statistical model has varying levels of complexity, and what variables one includes impacts this complexity. The saturated regression for RCTs discussed in Sect. 1.2.1 includes only a treatment variable A. (This is sometimes called a crude regression.) For example, with a continuous outcome Y and a binary treatment A the regression of the conditional mean of Y given A is

$$E_0(Y \mid A) = \alpha_0 + \alpha_1 A.$$

The parameter $E_0(Y \mid A)$ is the conditional mean of Y given A, and (α_0, α_1) are the unknown regression parameters in the parametric statistical model for the conditional mean. We are estimating the regression $E_0(Y \mid A)$ based on the data $(A_1, Y_1), \ldots, (A_n, Y_n)$, ignoring the covariates W_i for subject i. Fitting this regression to the data will result in an estimate of the effect of treatment given by $\alpha_1 = E_0(Y \mid A = 1) - E_0(Y \mid A = 0)$.

In the analysis of observational studies, it is commonplace to include covariates associated with both A and Y in the regression, in an attempt to eliminate the contribution of these variables and isolate the effect of A on Y. With one covariate W, an example of such a regression in a parametric statistical model is

$$E_0(Y \mid A, W) = \alpha_0 + \alpha_1 A + \alpha_2 W.$$

The effect of A is again given by α_1, but α_1 now represents an effect of A adjusting for W, and is thus a different parameter of interest than the effect of A above. If effect modification is suspected, an interaction term between the effect modifier and A might be included:

$$E_0(Y \mid A, W) = \alpha_0 + \alpha_1 A + \alpha_2 W + \alpha_3 A \times W. \tag{1.1}$$

Effect modification between A and W occurs when the effect of A differs within strata of W. The consequence of including an interaction term in the regression is that there is now not one summary measure of the effect of A. For every level of W there is a different effect measure of A. For example, in the simple case where W is binary, such as smoking status, there will be two effect measures for A. If $W = 1$ indicates current smoker, the effect of A among current smokers is $\alpha_1 + \alpha_3$. When $W = 0$, α_3 is equal to zero thus the effect of A among current nonsmokers is α_1. As we add covariates and interaction terms to our regression, α_1 does not estimate a marginal population-level effect. In fact, each time we add a covariate or interaction the interpretation of the coefficients in the parametric statistical model changes.

In the situation where we only have one binary covariate, the regression specified in Eq. (1.1) is a saturated parametric statistical model. Let us also suppose the collection of the single covariate represents the truth, and there are no other covari-

ates that should have been measured. This parametric statistical model is therefore suitable in that it is *not misspecified*. However, we still want a marginal effect estimate of treatment. This marginal-effect could be defined as $\alpha_1 + \alpha_3 E_0(W)$, where $E_0(W)$ denotes the true marginal mean of W. A simple nonparametric maximum likelihood estimator will accomplish this for the simple case posed here. But what happens when you have a continuous covariate? Or we have an increasing number of covariates? This approach to fitting a saturated linear regression quickly becomes problematic since the number of coefficients will grow exponentially with the number of covariates.

High-dimensional data have become increasingly common, and researchers often have dozens, hundreds, or even thousands of potential covariates to include in their parametric statistical model. Not only does this provide an impossible challenge to correctly specify the parametric statistical model for the conditional mean, but the complexity of the parametric statistical model may also increase to the point that there are more unknown parameters than observations. A fully saturated parametric statistical model will usually result in a gross overfit of the data. In addition, the true functional, $(A, W) \rightarrow E_0(Y \mid A, W)$, mapping the treatment and covariates into the conditional mean, might be described by a complex function not easily approximated by main terms or simple two-way interactions.

1.2.4 The Complications of Human Art in Statistics

We now highlight further the innate challenges of parametric statistical models and the problematic human art component of data analysis. Returning to our toxin and cancer study from Sect. 1.1, where an indicator of death is the outcome, let's say that the principal investigator (PI) asserts smoking status is the only relevant covariate that we must control for in our analysis. The PI also says to use the following logistic linear regression in a parametric statistical model for the probability of death, where the α_is are the unknown regression parameters in the statistical model:

$$P_0(Y = 1 \mid A, W) = \text{expit}(\alpha_0 + \alpha_1 A + \alpha_2 W).$$

Another subject matter expert on the project enters the conversation and says that one must also control for age and gender. Smoking is now denoted W_1, with age as W_2 and gender W_3. The covariates can be represented as a vector $W = \{W_1, W_2, W_3\}$ and the logistic linear regression given by

$$P_0(Y = 1 \mid A, W) = \text{expit}(\alpha_0 + \alpha_1 A + \alpha_2 W_1 + \alpha_3 W_2 + \alpha_4 W_3).$$

A data analyst enters the picture and explains that all covariates measured at baseline, listed in Table 1.1, should be thrown into a logistic linear regression. Using the results of this regression fit, all W_is with coefficients that do not have a p-value smaller than 0.05 should be removed from the list. The regression should then be fit again in a new (different) parametric statistical model with the variables remaining

Table 1.1 Baseline covariates from a study examining the effect of a toxin on death from cancer

W_i	Covariate
W_1	Smoking status
W_2	Age
W_3	Gender
W_4	Health status
W_5	Cardiac event
W_6	Chronic illness

in the list. This continues until all coefficients in front of the W_is in the regression have a p-value of less than 0.05. The regression coefficient α_1 in front of A changes with each new regression. It is highly dependent on which variables are included.

One can quickly see, even in this simplified example, the impossible challenge involved in selecting which variables to include in the parametric statistical model, and thereby assigning the underlying probability distribution of Y, conditional on the treatment and covariates, that generated the data up to a finite number of unknown parameters. The problem that we stress again here is that we do not know the true probability distribution of the data up to a finite number of unknown parameters.

The inference made using parametric statistical models assumes that the parametric statistical model is correct and was a priori selected. If the parametric statistical model is wrong, our estimates will approximate a noninterpretable parameter, and thereby be biased for the true hypothesized target parameter one had in mind under the assumption that the parametric statistical model was true. If we run several models with the full data, the statistical inference (e.g., the p-values) is meaningless, and this statistical model should be selected before looking at the data to avoid bias.

In addition, if the parametric statistical model was not a priori specified but data-adaptively selected as the data analyst suggests, then the statistical inference is misleading, claiming a certainty that does not exist. The final parametric statistical model is reported as if it were the only one considered and evaluated. The data analyst has performed a procedure that began the moment the data were used. In other words, once you start using the data, your estimation method has also started. Therefore, our data analyst has selected an approach that, while very common, blatantly leaves us with faulty inference.

Even without the approach defined by the data analyst, the PI and the subject matter expert might run both of their regressions and then decide between them based on the results. It should not be overlooked that the process of looking at the data, examining coefficient p-values, and trying multiple statistical models is not only incredibly prevalent but is taught to students learning statistics.

This is the human art component we eluded to in Sect. 1.2. The moment we use post-hoc arbitrary criteria and human judgment to select the parametric statistical model after looking at the data, the analysis becomes prone to additional bias. This bias manifests in both the effect estimate and the assessment of uncertainty for that estimator (i.e., standard errors). One cannot even define the procedure that was used

as a function of the data so that more appropriate standard errors can be calculated (e.g., by use of bootstrapping). Statistics is not an art, it is a science.

Standard practice focuses on estimating $E_0(Y \mid A, W)$ with an assumed parametric statistical model. One then extracts the coefficient in front of A as the effect estimate, ignoring that we know that most *parametric statistical models are wrong.* This criticism extends in general to estimation procedures (e.g., prediction) using misspecified parametric regression models. There is a more natural way to think about our parameter of interest, which we introduced abstractly in Sect. 1.1. The definitions of the data, model, and parameter will allow us to target parameters that are frequently of interest, such as causal effects. These concepts will be developed more concretely in the next section, and additionally in Chap. 2, as we set aside the traditional approach to effect estimation.

1.3 Data, Model, and Target Parameter

Our discussion of the data, model, and target parameter has been relatively abstract up to this point. We formalize these concepts in this section using notation. We define O as the random variable with P_0 as the corresponding probability distribution of interest. We write $O \sim P_0$ to mean that the probability distribution of O is P_0. Our random variable, which we observe n times, could be defined in a simple case as $O = (W, A, Y) \sim P_0$ if we are without common issues such as missingness and censoring. W, A, and Y are as defined in Sect. 1.2.3.

Complex data structures. While the data structure $O = (W, A, Y) \sim P_0$ makes for effective examples, data structures found in practice are frequently more complicated. Suppose we have a right-censored data structure. Right censoring means that we do not observe a particular variable or variables to the end of the study or time period. For example, if we are following subjects for 5 years, some subjects may drop out of the study for various reasons (e.g., relocation, death, voluntarily ending participation). If we are planning to measure an outcome Y, such as developing liver cancer, within 5 years of baseline, those subjects that drop out before 5 years (i.e.,

are censored) will not have measurements across the whole time period. All subjects will be censored at year 5 if they have not already been censored, but subjects that are observed for the full 5 years provide us with the desired full-data structure and are thereby referred to as uncensored.

Censoring is always defined with respect to a desired full-data structure. This type of censoring of a desired full-data structure is referred to as right censoring since timelines are frequently numbered from left to right, and it is some portion of the right side that is censored. Now, for each subject we will observe their time of censoring, and we may observe their time to event. For example, subject 1 may develop liver cancer at year 3. Another subject may be censored at year 2 due to dropout, and we never observe whether they develop liver cancer within the 5 years. Thus our data structure now has added complexity. We have T representing time to event Y, C a censoring time, $\tilde{T} = \min(T, C)$ which represents the T or C that was observed first, and $\Delta = I(T \leq \tilde{T}) = I(C \geq T)$ an indicator that T was observed at or before C. We then define $O = (W, A, \tilde{T}, \Delta) \sim P_0$. This is another example of a possible data structure.

1.3.1 The Model

We are considering the general case that one observed n i.i.d. copies of a random variable O with probability distribution P_0. The data-generating distribution P_0 is also known to be an element of a statistical model $\mathcal{M}$, which we write $P_0 \in \mathcal{M}$. Formally, a statistical model $\mathcal{M}$ is the set of possible probability distributions for P_0; it is a collection of probability distributions. What if all we know is that we have n i.i.d. copies of O? Well, then we've stated what we know, thus this can be our statistical model, which we call a nonparametric statistical model. We don't need to assign a parametric form to the distribution of our data; it is simply known to be an element of a nonparametric statistical model $\mathcal{M}$.

We might also consider a semiparametric statistical model if we have additional information about the way our data were generated that puts restrictions on the data-generating distribution P_0. For example, we may know that the effect of exposure A on the mean outcome is linear. Note, though, that semiparametric statistical models can be wrong by not containing the true P_0 if our "knowledge" is faulty. While we might have additional knowledge, we do not have enough knowledge to parameterize P_0 by a finite-dimensional parameter. These nonparametric and semiparametric statistical models should represent true knowledge about the underlying mechanism generating the data, that is, they are supposed to contain the true probability distribution P_0 of the experimental data.

We will frequently use *semiparametric* to include both nonparametric and semiparametric, such as the phrase "semiparametric estimation" referring to estimation in a nonparametric or semiparametric statistical model. When semiparametric excludes nonparametric and we make additional assumptions, this will be explicit.

Statistical model vs. model. A statistical model can be augmented with additional (causal) assumptions providing a parameterization so that $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$, where the space of θ-values, Θ, is itself infinite dimensional. Even though such a parameterization does not change the statistical model, thereby providing nontestable causal assumptions, it does allow one to enrich the interpretation of $\Psi(P_0)$ in terms of a statement of an underlying truth θ_0. We refer to the statistical model augmented with a parameterization as a model. We will return to the issue of modeling, thereby making (causal) assumptions that go beyond specifying a statistical model $\mathcal{M}$, in Chap. 2. The important take-home message for now is that the statistical model is the only relevant information for the estimation problem, while the additional (causal) assumptions will provide enriched (or misleading, if wrong) interpretations of the target parameter.

1.3.2 The Target Parameter

What are we trying to learn from our data? Often the question of interest is related to quantifying some difference in the probability distribution of an outcome of interest between the treated and untreated or the exposed and unexposed groups. We want to understand the effect of treatment or exposure on the probability distribution of the outcome of interest. This difference could be measured on an additive scale or multiplicative scale, such as a relative risk or odds ratio.

Either way, once an agreement is reached concerning what one wants to learn, we can explicitly define the target parameter of the probability distribution P_0 as some function of P_0: $\Psi(P_0)$ for some function Ψ that maps the probability distribution P_0 into the target feature. That is, we are interested in estimating a parameter $\Psi(P_0)$ of the probability distribution $P_0 \in \mathcal{M}$, which is known to be an element of a nonparamteric or semiparametric statistical model $\mathcal{M}$. The parameter $\Psi(P_0)$ is a function of the unknown probability distribution P_0. We are not interested in estimating an effect defined by a coefficient of a (misspecified) parametric statistical model. Rather, we define a parameter as a feature of the true probability distribution P_0 of the data using true knowledge we have about P_0 as embodied by the statistical model $\mathcal{M}$. Thus, we are explicitly confronted with the fact that we need to know how to define our target parameter as a feature of P_0: it does not suffice to grab a parametric statistical model and just target the coefficients in that model.

First, one needs to define the parameter of interest as a function of the data-generating distribution varying over the nonparametric or semiparametric statistical model. Many practitioners are used to thinking of their parameter in terms of a regression coefficient, but that is often not possible in realistic nonparametric and semiparametric statistical models. Instead, one has to carefully think about what feature of the distribution of the data one wishes to target. With an experimental unit-specific data structure $O = (W, A, Y) \sim P_0$, the risk difference is the following function of the distribution P_0 of O:

$$\Psi(P_0) = E_{W,0}[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)],$$

where $E_0(Y \mid A = a, W)$ is the conditional mean of Y given $A = a$ and W. Here A is binary and therefore a takes on two values, 1 and 0. E_W indicates that we take the average over the observed distribution of our covariate(s) W. Uppercase letters represent random variables and lowercase letters are a specific value for that variable. For example, if all variables are discrete, $P_0(W = w, A = a, Y = y)$ assigns a probability to any possible outcome (w, a, y) for $O = (W, A, Y)$. P_0 is like a calculator: we input (w, a, y) and it returns a probability. $\Psi(P_0)$ for the risk difference can then also be written:

$$\Psi(P_0) = \sum_w \Big[\sum_y y P_0(Y = y \mid A = 1, W = w) - \sum_y y P_0(Y = y \mid A = 0, W = w) \Big] P_0(W = w),$$

where

$$P_0(Y = y \mid A = a, W = w) = \frac{P_0(W = w, A = a, Y = y)}{\sum_y P_0(W = w, A = a, Y = y)}.$$

After obtaining an estimate of $\Psi(P_0)$ and a confidence interval, we can provide two interpretations, one as a purely statistical parameter of P_0, and one as a causal parameter under additional causal assumptions representing a causal model that goes beyond the specification of the statistical model $\mathcal{M}$. We discuss these causal assumptions in detail in Chap. 2.

1.3.3 Summary of Concepts

1. **Data.** Our data are comprised of n i.i.d. copies of a random variable $O \sim P_0$. P_0 is the true probability distribution for O.
2. **Model.** Our statistical model $\mathcal{M}$ is nonparametric or semiparametric and represents only what we know about our data-generating distribution P_0. $\mathcal{M}$ is the set of possible probability distributions for P_0. Our model includes possible additional causal assumptions, allowing an enriched interpretation of the parameter of interest.
3. **Target parameter.** Our parameter $\Psi(P_0)$ is a particular feature of the unknown probability distribution P_0. The explicit definition of this mapping Ψ on the statistical model requires that one defines $\Psi(P)$ at each P in the statistical model. The parameter typically has two interpretations, one as a parameter $\Psi(P_0)$ of a probability distribution P_0 and one as a causal parameter under additional (causal) assumptions to be discussed in Chap. 2.

1.4 The Need for Targeted Estimators

Let us step back for a moment. Suppose you were handed ten textbooks and told you would be asked one question in 12 h. The question might require understanding portions of several of these books. However, you are not told what the question is going to be. How would you prepare for such a test? You do not have time to read all ten textbooks, let alone master the material contained within them. You might read the chapter abstracts from each book in order to learn basic summary information.

Now, suppose you were handed the same ten textbooks, but instead you were told the question you would be asked 12 h later. Would this change your approach to studying? Yes! Since you know what question will be asked, you can more carefully discard books that will be completely unnecessary, keeping only those books with relevant chapters. You are able to spend 12 h working through the pertinent chapters and then give a thoughtful precise answer on the one question test.

This theoretical situation has a direct parallel to nontargeted learning vs. targeted learning. Maximum likelihood estimation in misspecified parametric statistical models is nontargeted learning; one estimates all the parameters (coefficients) in a parametric statistical model. One uses an empirical criterion that is only concerned with the overall fit of the entire probability distribution of the data instead of only the parameter of interest; we are trying to master all the books, spreading error uniformly across all content, when we only care about very specific portions of each book. The overall fit of the probability distribution based on the data set is then used to evaluate the target parameter of the probability distribution, i.e., the question is answered with the nontargeted fit of the distribution of the data. For a small *true* parametric statistical model, containing the true probability distribution, one with few terms or few unknown coefficients, the performance of the maximum likelihood estimator of the target parameter with regard to mean squared error may be satisfactory. However, the bigger the statistical model, the more problematic nontargeted learning becomes. We have no problem with maximum likelihood estimation for relatively low-dimensional parametric statistical models if they are correct, but this is not the case in practice, and we wish for our statistical models to represent true knowledge. Indeed, in semiparametric statistical models, maximum likelihood estimation breaks down completely. With targeted learning, we focus on our known question of interest; we focus on the relevant information in the books, and rank the information by its relevance for the question of interest.

1.5 Road Map for Targeted Learning

The first six chapters of this textbook are meant to provide the reader with a firm grasp of the targeted learning road map and the solution to prediction and causal inference estimation problems: super learning and targeted maximum likelihood estimation (TMLE). For the sake of presentation, in these introductory chapters we will focus on the data structure $(W, A, Y) \sim P_0$, the nonparametric statistical

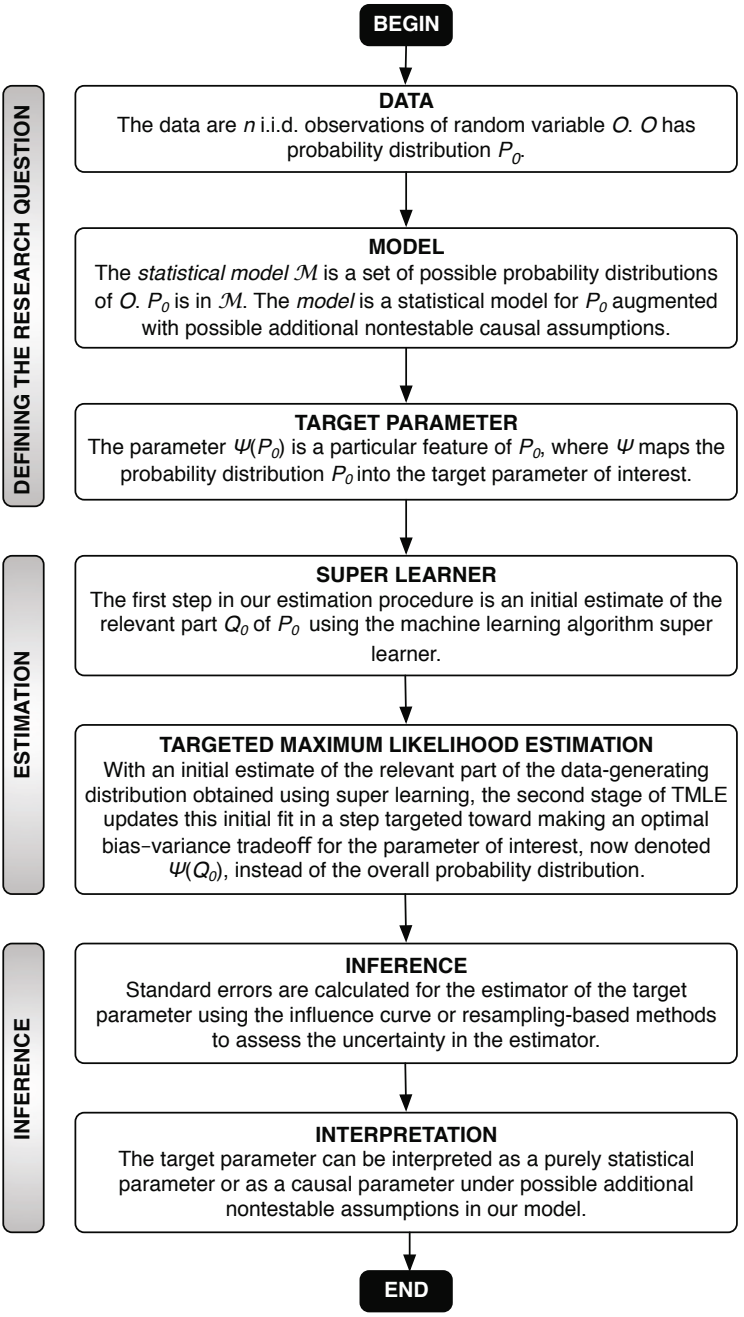

Fig. 1.2 Road map for targeted learning

model $\mathcal{M}$, and the additive causal effect target parameter $\Psi(P_0) = E_{W,0}[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)]$. Our estimator of the treatment effect will be obtained by plugging in a (targeted) estimator of (or the relevant part of) P_0 into the parameter mapping Ψ. Such an estimator is called a plug-in or substitution estimator. Substitution estimators have the advantage of fully respecting the constraints implied by the statistical model $\mathcal{M}$ and respecting that the target parameter is a very specific function of P_0. As a consequence, substitution estimators are generally robust, even in small samples.

This first chapter was intended to motivate the need for improved estimation methods, highlight the troublesome nature of the traditional approach to estimation, and introduce important concepts such as the data, model, and target parameter. We develop the following concepts, as part of the road map for targeted learning, in the remaining five introductory chapters.

Defining the model and target parameter. By defining a structural causal model (SCM), we specify a model for underlying counterfactual outcome data, representing the data one would be able to generate in an ideal experiment. This is a translation of our knowledge about the data-generating process into causal assumptions. We can define our target parameter in our SCM, i.e., as a so-called causal effect of an intervention on a variable A on an outcome Y. The SCM also generates the observed data O, and one needs to determine if the target parameter can be identified from the distribution P_0 of O alone. In particular, one needs to determine what additional assumptions are needed in order to obtain such identifiability of the causal effect from the observed data.

Super learning for prediction. The first step in our estimation procedure is an initial estimate for the part of the data-generating distribution P_0 required to evaluate the target parameter. This estimator needs to recognize that P_0 is only known to be an element of a semiparametric statistical model. That is, we need estimators that are able to truly learn from data, allowing for flexible fits with increased amounts of information. We introduce cross-validation and machine learning as essential tools and then present the method of super learning for prediction with its theoretical grounding, demonstrating that super learning provides an optimal approach to estimation of P_0 (or infinite-dimensional parameters thereof) in semiparametric statistical models. Since prediction can be a research question of interest in itself, super learning for prediction is useful as a standalone tool as well.

TMLE. With an initial estimate of the relevant part of the data-generating distribution obtained using super learning, we are prepared to present the remainder of the TMLE procedure. The second stage of TMLE updates this initial fit in a step targeted towards making an optimal bias–variance tradeoff for the parameter of interest, instead of the overall probability distribution P_0. This results in a targeted estimator of the relevant part of P_0, and thereby in a corresponding substitution estimator of $\Psi(P_0)$.

Many of the topics we have presented in this road map may be new to you. They will be explained in detail in the coming chapters. This brief road map is introduced

for the reader to see where we are going, how the pieces fit together, and why we will present material in this order.

1.6 Notes and Further Reading

We motivated this chapter with two real-world debates: HRT and screening guidelines for breast cancer. In a *New York Times* piece, Taubes (2007) discussed the merits of epidemiology using the HRT studies as an example. For those interested in reading more about this topic, it is an excellent comprehensive starting point with thorough references. For the statistician and researcher, it also raises one of the questions we seek to answer with this text. Can we estimate causal effects from observational studies? Two starting points for the mammography debate include U.S. Preventive Services Task Force (2009) for the official recommendation statement on breast cancer screenings, as well as Freedman et al. (2004) for a qualitative review of breast cancer mammography studies.

For additional background on study designs and covariate adjustment we direct readers to Rothman and Greenland (1998) and Jewell (2004). For a readable introductory statistics text on traditional regression techniques and key statistics concepts such as the central limit theorem (CLT) we refer readers to Freedman (2005).

A popular article drawing attention to false research findings, due in part to current statistical practice, is Ioannidis (2006). Ioannidis was also interviewed in journalist David H. Freedman's new book, *Wrong: Why Experts Keep Failing Us—And How to Know When to Trust Them*. This text focuses on problems in research fields, including the way data are analyzed and presented to the public (Freedman 2010).

George Box famously discussed that (parametric) statistical models are wrong, but may be useful (Box and Draper 1987). As presented in this chapter, misspecified parametric statistical models may not perform terribly for low-dimensional data structures and small sample sizes. Over 20 years after Box's statements, data sets have become increasingly high dimensional, and large studies are very common. We are also still left with the issue that the coefficients in misspecified parametric statistical models do not represent the target parameter of interest. Therefore, the usefulness of misspecified parametric statistical models is extremely limited. Note, however, that maximum likelihood estimators according to candidate parametric working statistical models can be included in the library of the super learner, discussed in Chap. 3, and can play a useful role in that manner.

The use of data-adaptive tools can be beneficial, although we discuss in this chapter (Sect. 1.2.4) a commonly used data-adaptive procedure in parametric statistical models that provides faulty inference. Data-adaptive methods, when guided by a priori benchmarks in a nonparametric or semiparametric statistical model, are advantageous for prediction and discussed in detail in Chap. 3. We use the terms *data-adaptive* and *machine learning* interchangeably in this text. Targeted estimators will be discussed in Chaps. 4–6.