Estimands for Time-Fixed and Time-Dependent Data

Ashley I Naimi

January 2024

Contents

1 Introduction

With potential outcomes, we can begin the process of defining the effects that most suitably address the research questions of interest. There are a wide variety of estimands that we can define. In this section, we illustrate some commonly employed estimands, and discuss the distinction between estimands and the scale of the estimand.

2 Estimands, Estimators, Estimates

Before we start, it's important to distinguish between estimands, estimators, and estimates. The estimand is the mathematically defined target effect that we would like to quantify. The estimator is a function of the data that will allow us to "link" the estimand with the observed data under the necessary assumptions. Finally, the estimate the numerical value or set of values we obtain reflecting the estimate of the estimand.

2.1 Target Estimand and Target Scale: Average Treatment Effects

Causal inference starts with a clear idea of the effect of interest (the target causal parameter, or **estimand**). We use potential outcomes to do this. The **estimand** is the (mathematical) object we want to quantify. For example, a commonly used estimand is the average treatment effect, which is usually defined as:

$$E(Y^{x} - Y^{x'})$$

This estimand is interpreted as the average difference in outcomes that would be observed if X was set to x for all individuals in the population, versus if X was set to x' for all individuals in the population. If the treatment or the exposure is binary, then x is usually taken to be $x = 1$, and $x' = 0$. However, if the treatment or exposure is continuous, then one must specify values for x and x'.

Deeper Dive: Estimating the Effects of Continuous Exposures

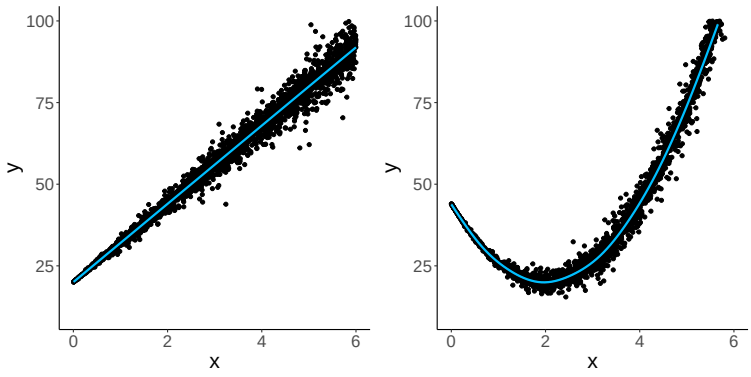
In classical settings, when conditionally adjusted regression models are used to estimate exposure effects, students are often taught to interpret the exposure coefficient for the exposure in a regression model as the **effect of a single unit change in the exposure.** For example, if we have a conditionally adjusted regression model defined as:

$$E(Y \mid X, C) = \beta_0 + \beta_1 X + \beta_2 C,$$

Then:

$$\begin{aligned} E(Y \mid X, C) &= \beta_0 + \beta_1 X + \beta_2 C \\ \implies & E(Y \mid X = 1, C) - E(Y \mid X = 0, C) \\ &= [\beta_0 + \beta_1 (X = x) + \beta_2 C] - [\beta_0 + \beta_1 (X = x') + \beta_2 C] \\ &= \beta_1, \text{ if } x' - x = 1 \text{ unit} \end{aligned}$$

However, this is really only of interest if we can either assume that the relation between the exposure X and the conditional mean of Y, $E(Y|X, C)$ is linear, or if we are interested in a *projection* of this relation (Hubbard et al., 2010). If neither of these is the case, we need to account for potential curvilinear relations between X and Y. There are many ways to do this, for example splines can be used, which allow the conditional mean of Y to depend on X curvilinearly.



Note that in the Figure on the left, the value of the difference in the mean of Y for a single unit difference between x' and x is the same no matter where we are on the x-axis of the Figure. In contrast, for the Figure on the right, the value of the difference in the mean of Y for a single unit difference between x' and x depends on where we are taking this difference on the x-axis. Additionally, there is no particular reason why we may be interested in a single unit increase in the exposure. In fact, we may be interested in a range of changes in the value of x, spanning multiple units.

Partly for these reasons, as well as because of the fact that we often do not assume that the underlying relation between x and y is linear, in a general causal inference setting, when (say), interest lies in the average treatment effect, it is important to specify which values of the exposure we wish to contrast.

For exposures in general, the practice of selecting specific contrast values is an important part of defining the target estimand of interest.

In addition to selecting an estimand such as the average treatment effect, it is also essential that we select the scale of the contrast of interest. For example, if the outcome is a binary (or more generally, multinomial) variable, we can quantify the average treatment effect on the risk difference, risk ratio, or odds ratio scale:

$$E(Y^x - Y^{x'}), \quad \frac{E(Y^x)}{E(Y^{x'})}, \quad \frac{Odds(Y^x = 1)}{Odds(Y^{x'} = 1)},$$

where $Odds(Y^x = 1) = E(Y^x)/[1 - E(Y^x)]$, and where $E(.)$ is the expectation operator taken with respect to the total population. If the outcome Y is binary, then $E(Y) \equiv P(Y = 1)$. Or, the expectation of Y is equivalent to the probability that $Y = 1$. This assumes that the binary outcome variable Y is coded as $\{0, 1\}$, and not, e.g., $\{1, 2\}$.

If the outcome is a continuous or count variable, we can quantify average treatment effects on the mean difference or mean ratio scales:

$$E(Y^x - Y^{x'}), \quad \frac{E(Y^x)}{E(Y^{x'})}.$$

Technical Note:
Of course, the estimand need not always be causal (Casella and Berger, 2002). We may be interested in a statistical estimand, such as the conditional risk difference, risk ratio, or odds ratio:

$$E(Y \mid X = 1) - E(Y \mid X = 0), \quad \frac{E(Y \mid X = 1)}{E(Y \mid X = 0)}, \quad \frac{Odds(Y \mid X = 1)}{Odds(Y \mid X = 0)},$$

What's important is that one is clear about the objective. For example, in Naimi (2016) we defined counterfactual disparity measures as:

$$E(Y^m \mid X = 1) - E(Y^m \mid X = 0)$$

which is a mixed statistical and counterfactual estimand. It is a measure of disparity (statistical estimand) that would be observed if some variable M were set to a value m (counterfactual estimand).

The causal estimands presented above represent **average treatment effects**. This effect is sometimes referred to as a marginal treatment effect, because it averages (or marginalizes) the effect over the entire sample. To better understand this, consider this fictional data from Chapter 4 of the Mixtape (Cunningham, 2021). Instead this time, let's assume that we actually have all

the potential outcomes on each person in this dataset:

Name	X	Y	Y1	Y0	C1	C2
Andy	1	10	10	15	1	1
Ben	1	5	5	9	1	0
Chad	1	16	16	18	0	1
Daniel	1	3	3	5	0	0
Edith	0	5	0	5	1	1
Frank	0	7	3	7	1	0
George	0	8	6	8	0	1
Hank	0	10	8	10	0	0

For instance, if we consider the difference scale, the average treatment effect is

$$\begin{aligned} E(Y^1 - Y^0) &= \frac{1}{N}\sum_{i=1}^{N} Y_i^1 - \frac{1}{N}\sum_{i=1}^{N} Y_i^0 \\ &= 6.78 - 9.63 \\ &= -2.85 \end{aligned}$$

This is a marginal treatment effect because it is marginal with respect to the distribution of C. That is, in the fictional data above, we could have computed each individual difference in Y^1 and Y^0, and then taken the average of all of these differences. Because we are averaging (or marginalizing) over the distribution of all C variables, this average treatment effect is a marginal (as opposed to a conditional) treatment effect.

2.2 Average Treatment Effects in Randomized versus Observational Studies

In double blind placebo controlled randomized trials a commonly targeted effect is the intention-to-treat effect. To explain this effect, it is important to understand key elements of a basic randomized trial. These include a treatment assignment variable (X), a treatment compliance variable (A), and an outcome (Y). Figure 1 shows the relationships between these variables in an RCT.

The intention-to-treat effect is so named because it does not quantify the

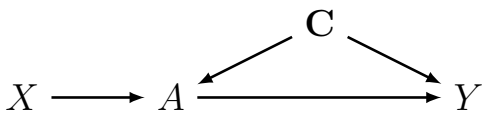


Figure 1: Simple causal diagram depicting the relationships between variables in a randomized controlled trial. These variables include a treatment assignment indicator X, a treatment compliance A, which captures whether the individual took the assigned treatment or not, and an outcome of interest Y.

effect of the treatment (A), but the effect of assigning treatment (X). That is, a physician may intend to treat someone by prescribing drug X, but whether the patient chooses to take the drug (indicated by A) is another matter. In this setting, the effect is equivalent to a contrast of potential outcomes that would be observed if everyone was assigned to treatment versus if everyone was assigned to placebo.

Under relatively simple conditions, the intention-to-treat effect is identified[1] which implies that we can simply use a difference in outcome means among those assigned to treatment versus not. In the absence of loss-to-follow-up in the trial, one can use this contrast to compute the intention-to-treat effect on (for example) the difference scale:

[1] A term we will discuss in depth soon.

$$E(Y^{x} - Y^{x'})$$

where x denotes assigned to treatment and x' denotes assigned to placebo. Consequently, the intention-to-treat effect is an average treatment effect where the exposure is an indicator of treatment assignment.

3 Conditional Average Treatment Effects

However, we may want to estimate this effect in a subset of the population. For instance, assuming X is a binary variable, $E(Y^1 - Y^0 \mid C = c)$ is the effect of $x = 1$ versus $x = 0$ among those with $C = c$. Note that C can be a single variable, or a vector. There are many different conditional (in contrast to marginal) average treatment effects (CATEs), this latter one being one of the simplest.

Research on and use of CATEs is rapidly increasing. But this area also has a long history, particularly in epidemiology. There are a number of different areas related to estimating CATEs. These include estimating statistical and causal

interactions, effect measure modification, heterogeneous treatment effect estimation, and (more recently), individual treatment effect estimation.

Much of the literature in epidemiology focuses on effect measure modification. This occurs when the effect of the exposure of interest on the outcome is modified by a third variable. For example, the effect of smoking (X) on myocardial infarction (Y) risk may be modified by the amount of dietary saturated fat consumption (C). Notationally, we could write this as:

$$E(Y^{x=1} - Y^{x=0} \mid C = c) \neq E(Y^{x=1} - Y^{x=0} \mid C = c')$$

However, some of the more recent literature on CATE estimation treats C as vector valued. Rather than focus on a single effect, CATEs are used to estimate treatment effects for individuals with certain characteristics. For example, the effect of smoking on myocardial infarction among individuals with specific levels of dietary saturated fat consumption, levels of soluble dietary fiber consumption, of a specific age, with specific weekly exercise levels, living in a specific socioeconomic status. Notationally, we would write this CATE in exactly the same way:

$$E(Y^{x=1} - Y^{x=0} \mid C = c)$$

where $C = c = c_1, c_2, \ldots c_p$ now represents a vector of specific fat consumption, fiber consumption, age, exercise levels, and socioeconomic status stratum.

Looking again at our fictional data we can compute a number of CATEs:

Name	X	Y	Y1	Y0	C1	C2
Andy	1	10	10	15	1	1
Ben	1	5	5	9	1	0
Chad	1	16	16	18	0	1
Daniel	1	3	3	5	0	0
Edith	0	5	0	5	1	1
Frank	0	7	3	7	1	0
George	0	8	6	8	0	1
Hank	0	10	8	10	0	0

For instance, we can compute the effect for those with $C_1 = 1$

$$
\begin{aligned}
E(Y^1 - Y^0 \mid C_1 = 1) &= \frac{1}{N}\sum_{i=1}^{N} Y_i^1 \mid C_1 = 1 - \frac{1}{N}\sum_{i=1}^{N} Y_i^0 \mid C_1 = 1 \\
&= 4.5 - 9 \\
&= -4.5
\end{aligned}
$$

Or we can compute the effect for those with $C_1 = 0$ and $C_2 = 1$ (which amounts to including only Chad and George in our averages):

$$
\begin{aligned}
E(Y^1 - Y^0 \mid C_1 = 0, C_2 = 1) &= \frac{1}{N}\sum_{i=1}^{N} Y_i^1 \mid C_1 = 1 - \frac{1}{N}\sum_{i=1}^{N} Y_i^0 \mid C_1 = 1 \\
&= (16 + 6)/2 - (18 + 8)/2 \\
&= 11 - 13 \\
&= -2
\end{aligned}
$$

In more complicated settings with more variables, including binary, categorical, and continuous variables in the conditioning statement, CATEs can become very interesting, but also more challenging to communicate to an audience. In a later section, we will look at some considerations for quantifying and presenting CATEs.

3.1 CATEs and the Scale Dependence of Effect Modification

Working within the framework of effect modification, epidemiologists have recognized for some time that the presence of effect modification depends heavily on the scale of the contrast used to evaluate effect modification. For example, consider a scenario where we have an exposure $X \in [0, 1]$, a modifier $M \in [0, 1]$, and an outcome $Y \in [0, 1]$. Let's say that we are interested in quantifying the following risks:

Risk	Abbreviation	Interpretation
$P(Y = 1 \mid X = 1, M = 1)$	R_{11}	Risk of the outcome among the doubly exposed

Risk	Abbreviation	Interpretation
$P(Y = 1 \mid X = 1, M = 0)$	R_{10}	Risk of the outcome among those exposed to X
$P(Y = 1 \mid X = 0, M = 1)$	R_{01}	Risk of the outcome among those exposed to M
$P(Y = 1 \mid X = 0, M = 0)$	R_{00}	Risk of the outcome among the doubly exposed

Let's say further we are interested in evaluating whether the association between X and Y on the difference scale depends on whether $M = 1$ or $M = 0$. If there is no modification of the effect by M, then the following relation will hold (Rothman et al., 2008, p73):

$$R_{11} - R_{01} = R_{10} - R_{00} \tag{1}$$

This equation says that the difference in risks between the doubly exposed and those exposed to only M is the same as the difference between those exposed to only X and the doubly unexposed. This can be seen visually in Figure 2.

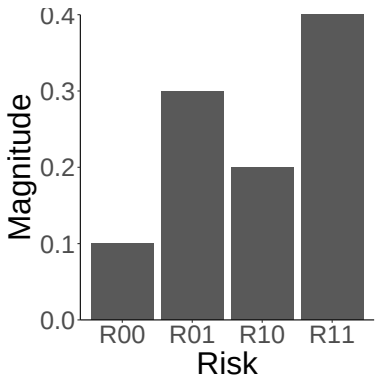


Figure 2: Doubly and singly exposed risks under no modification for the risk difference.

For the scenario demonstrated in Figure 2, there is no effect modification by M, and thus the equation above holds. However, it turns out that if this is the case, then *there must be effect modification by M on the risk ratio scale*, **unless** either X or M or both are unassociated with Y. We can demonstrate this using the values for each respective risk provided in Figure 2, which are:

Risk	Value
R_{11}	0.4
R_{10}	0.3
R_{01}	0.2
R_{00}	0.1

Thus, plugging these values into equation (1), we get:

$$R_{11} - R_{01} = R_{10} - R_{00} = 0.2$$

However, if we took the corresponding ratios, we see that

$$\frac{R_{11}}{R_{01}} \neq \frac{R_{10}}{R_{00}}$$

Rather, we obtain a value of RR = 3 for those with $M = 0$, and a value of RR = 2 for those with $M = 1$, indicating effect modification.

This scale dependence of effect modification is why Rothman et al recommend that terms such as *effect measure modification*, or preferably *risk difference modification* or *risk ratio modification* be used in the literature. Additionally, there are several different scale that we can use to measure outcome occurrences. These include odds, hazards, rates, prevalences, means, and a range of others. From this, it follows that the absence of effect modification on any particular scale implies the presence on many other scales (Rothman et al., 2008, page 74).

Finally, while these concpets have been understood for decades in epidemiology, they do not seem to be well-recognized in other areas of methodology. This is particularly true for the estimation of conditional average treatment effects. As of the present time, and to my limited evaluation of the complex and very large literature on statistical and machine learning methods for estimating conditional average treatment effects, it does not seem that the literature is aware of this feature.

Why is this important? Suppose we conduct a complex analysis looking at the relationship between some component of diet (e.g., dietary fiber intake) and some pregnancy outcome (e.g., preterm birth), and our goal is to estimate the conditional average treatment effect on the difference scale. Suppose further that we condition this effect on several physiologic, demographic, anthropomorphic, and economic variables.

Out of this analysis, we find that the **risk difference** for the relation between dietary fiber consumption and preterm birth is constant across the entire range of body mass index in the sample. That is, the effect of dietary fiber on preterm birth on the difference scale is constant across BMI. This implies that the effect of dietary fiber on preterm birth on the ratio scale must change according to different values of BMI. For example, we might find that we have

more protective risk ratios for individuals with larger BMIs.

This begs the question: should we target individuals with higher BMI values to consume more dietary fiber as a preventative measure for preterm birth?

Unfortunately, there is no clear answer to this question. The scale dependence of effect modification has long been a problematic issue in interpreting conditional average treatment effects in epidemiology. There is some reason to believe that the additive scale is more informative from a policy perspective (e.g., Panagiotou and Wacholder, 2014). However, there does not seem to be a consensus on this issue (Knol and VanderWeele, 2012).

3.2 The Effect of Treatment on the Treated and the Untreated

One special kind of conditional average treatment effect is the effect of treatment on the treated (ETT). Defined using potential outcomes:

$$E(Y^1 - Y^0 \mid X = 1),$$

this effect compares the outcomes that would be observed if the exposure were set to 1 (Y^1) versus if the exposure were set to 0 (Y^0) among those who were observed to be or actually exposed in the sample ($X = 1$).

To illustrate the relevance of this effect, consider the following (entirely fictional) scenario: Suppose that during gestation of a high-risk pregnancy, two clinical options are available to manage the risk of fetal death: premature delivery induction versus expectant management. Suppose further a researcher is interested in quantifying the effect of inducing delivery prematurely on fetal and infant death. This researcher collects data on a cohort of high-risk pregnant women, including whether delivery was induced prematurely, fetal/infant death, and a host of confounding variables. All parties involved agree the study is designed perfectly (no confounding, measurement error, loss to follow-up). They calculate the average treatment effect of premature delivery induction on fetal and infant death on the risk difference scale:

$$E(Y^1 - Y^0) = 0.15$$

This researcher concludes that, if all high-risk pregnancies were induced prematurely ($X = 1$), 15 more out of every 100 pregnancies would end in death, relative to what would happen if all high-risk pregnancies were left to expectant

management ($X = 0$). In light of this incredibly high excess risk of death, this researcher advises abandoning the practice of premature delivery induction entirely.

Another researcher questions the relevance of the average treatment effect. They argue that physicians would never induce delivery prematurely in all versus no high-risk pregnancies. Rather, the more interesting question is: **for those women whose pregnancies were actually induced**, what would the risk of death have been had they not been induced? Underlying this more nuanced question is an understanding that physicians may be inducing pregnancy in women because of a number of reasons that make these women different from the rest. These differences, in turn, can lead to a different effect. This researcher thus calculates the effect of treatment on the treated:

$$E(Y^1 - Y^0 \mid X = 1) = -0.05$$

This other researcher concludes that, among those whose pregnancies were actually delivered prematurely, the risk of death would have been higher had they not been delivered prematurely.

This hypothetical example demonstrates a fundamental difference between the ATE and the ETT: for those high-risk pregnancies that were not induced prematurely, the act of inducing premature delivery would not be beneficial. But for those high-risk pregnancies that were induced prematurely, the act of inducing premature delivery was beneficial. The ATE averages the beneficial and non-beneficial effects in the entire population, to yield an overall non-beneficial effect. The ETT isolates the beneficial effect among those who actually received the intervention. Thus, in this hypothetical example, premature delivery actually did benefit those who received it, even though it would not benefit everybody.

3.3 Estimands for Time-Dependent Data

When the exposure and/or outcome are measured repeatedly over follow-up, we saw how we can adapt potential outcomes to account for this. Using these adaptations, we can construct a wide array of potential contrast that enable us to define our effects of interest. The most commonly targeted effect in the context of longitudinal data is, again, the average treatment effect:

$$E(Y^{\overline{x}_J=x} - Y^{\overline{x}_J=x'})$$

Note the subtleties in the above equation, which are not convention, but are sometimes used in the literature. In particular, $\overline{x}_J = \{x_0, x_1, \dots x_J\}$, where J is the last (discrete) time point on study. This estimand will be the primary focus in the current short course, but numerous others are available.

4 Estimators

The estimand is the object we want to estimate. The **estimator** is an equation that allows us to use our data to quantify the estimand. Suppose, for example, we were explicitly interested in quantifying the causal mean difference for the relation between smoking and some continuous measure of 5 year CVD risk. To do this, we **have to** start by quantifying the associational mean difference, but there are many ways to do this.

To be specific, let's simulate some hypothetical data on the relation between smoking and some continuous measure of CVD risk. Let's look at maximum likelihood, penalized GLM, the generalized method of moments, and augmented inverse probability weighting (AIPW) as estimators:

```
remotes::install_github("yqzhong7/AIPW")
library(AIPW)

install.packages("SuperLearner", repos = "https://cloud.r-project.org/",
    dependencies = TRUE)
```

```
##
## The downloaded binary packages are in
##  /var/folders/zm/rqfqp5xs0fs86qs2mcxk6q0r0000gr/T//RtmpQp6HqV/downloaded_packages
```

```
library(SuperLearner)

install.packages("glmnet", repos = "https://cloud.r-project.org/",
    dependencies = TRUE)
```

```
##
## The downloaded binary packages are in
##  /var/folders/zm/rqfqp5xs0fs86qs2mcxk6q0r0000gr/T//RtmpQp6HqV/downloaded_packages
```

```
library(glmnet)

# define the expit function
expit <- function(z) {
    1/(1 + exp(-(z)))
}
set.seed(123)

n <- 1e+06
confounder <- rbinom(n, 1, 0.5)
smoking <- rbinom(n, 1, expit(-2 + log(2) *
    confounder))
CVD <- rnorm(n, 1 + 0.05 * smoking + 0.05 *
    confounder, sd = 1)

# the data
head(data.frame(CVD, smoking, confounder))
```

```
##            CVD smoking confounder
## 1 -0.008070729       0          0
## 2  2.404939385       0          1
## 3  0.531025110       0          0
## 4  2.518193560       0          1
## 5  1.492556427       0          1
## 6  1.146203114       0          0
```

```
# MLE
round(coef(glm(CVD ~ smoking + confounder,
    family = gaussian(link = "identity")))[2],
    4)
```

```
## smoking
```

```
##  0.0467
```

```
# Penalized GLM
mod_pen <- cv.glmnet(x = as.matrix(cbind(smoking,
    confounder)), y = CVD, family = gaussian(link = "identity"),
    nlambda = 100, alpha = 0, standardize = F)
# plot(mod_pen) mod_pen$lambda.min
round(coef(mod_pen, mod_pen$lambda.min)[2,
    ], 4)
```

```
## [1] 0.0462
```

```
# GMM
round(gmm(CVD ~ smoking + confounder, x = cbind(smoking,
    confounder))$coefficients[2], 4)
```

```
## smoking
##  0.0467
```

```
# AIPW
AIPW_SL <- AIPW$new(Y = CVD, A = smoking,
    W = confounder, Q.SL.library = c("SL.ranger"),
    g.SL.library = c("SL.ranger"), k_split = 3,
    verbose = FALSE)$fit()$summary()

round(AIPW_SL$result[3, 1], 4)
```

```
## [1] 0.0462
```

In our simple setting with 1 million observations, ordinary least squares, penalized GLM, the generalized method of moments, and AIPW yield the same associational risk difference (as expected) even though they are (for some, completely) different **estimators**.

It is important to note that these estimates are not causal mean differences, but are associational. Even the results from the AIPW estimator are *associational*, even though this method is much more clearly motivated from within the

causal inference framework (Robins and Greenland, 1994). To interpret them as causal effects, we have to evaluate whether we can **identify** the effect we want to estimate, a topic that we can discuss later.

For now, we have to ask the question: which estimator should we pick? More generally, what are the conditions that we should use to determine whether a particular estimator is a "good" estimator, that can or should be selected for the purpose of our analysis?

5 Some (Frequentist) Properties of Good Estimators

There are a number of properties that we can consider when selecting an estimator. The properties that are most often deemed most relevant (or at least, that are invoked much of the time) tend to be frequentist properties. These frequentist properties are properties of a procedure (a procedure applied to, say, all analyses)[2], not the properties of a particular result, test, or inference.

The properties we will discuss here include consistency, unbiasdness, precision, and convergence rates:

5.1 Estimator Consistency

Technically, an estimator $\widehat{\psi}$ of a target estimand ψ is consistent if, for any arbitrarily small $\epsilon > 0$:

$$\lim_{n \to \infty} P(|\widehat{\psi} - \psi| > \epsilon) = 0.$$

Consistency is an important estimator property. If an estimator is not consistent, then even an infinite amount of data will not enable us to quantify the true parameter value. Consistency can be visualized in the following Figure 3:

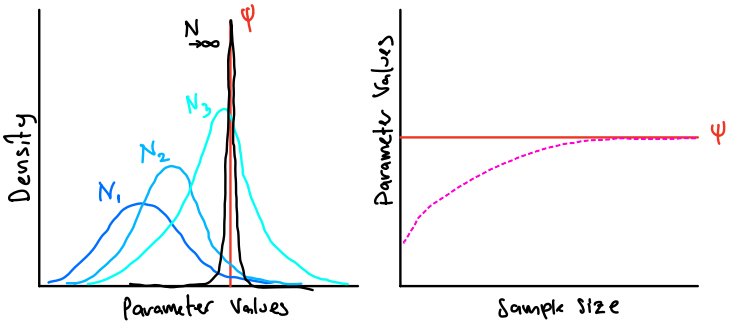


Figure 3: Demonstration of estimator consistency.

[2] Frequentist properties are often interpreted within a "repeated trial" framework. For example, the interpretation of a 95% confidence interval is the interval that, in the absence of structural biases (confounding, selection, information), would include the true parameter value 95% of the time if you were to repeat the study over an over again. However, it is not generally recognized that there is a more general interpretation: 95% confidence intervals include true parameter in 95% of all studies, even across different estimation problems, in the absence of structural biases.

This Figure provides two perspectives of the concept of consistency. The panel on the left shows the distribution of an estimator[3] for a number of different sample sizes. As the sample size increases, the center of each distribution gets closer to the true parameter value. As the sample size approaches infinity, the distribution of the estimator is centered almost fully on the true parameter value.

[3] Estimators are functions of data, or a sample from a population or data generating mechanism. As a result estimators are random variables, with central tendencies and spread.

5.2 Estimator Bias

Technically, an estimator $\widehat{\psi}$ of a target estimand ψ is unbiased if:

$$E(\widehat{\psi} - \psi) = 0$$

Bias is also an important property, but it's important to distinguish the use of the word. In applied areas such as epidemiology, we often refer to confounding bias, information bias, and selection bias. However, if this kind of bias occurs, then the true value cannot be estimated no matter the sample size. In fact, in applied areas, we often use "bias" to refer to an inconsistent estimator.

Statistical bias, on the other hand, refers to how well a particular estimator can quantify the parameter of interest *at a given sample size*.

Statistical bias can be visualized in the following Figure 4:

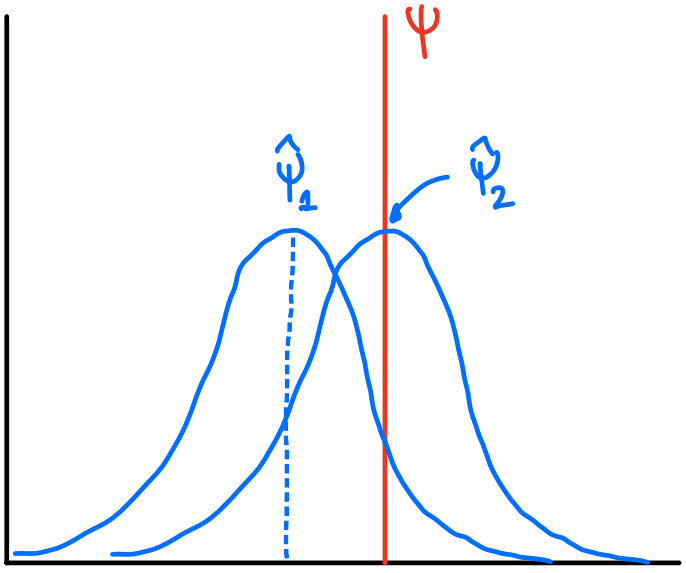


Figure 4: Demonstration of estimator bias.

This Figure shows the distribution of two different estimators for a given sample size. For example, these estimators could be a partial likelihood estimator for a hazard ratio in a Cox proportional hazards regression model ($\widehat{\psi}_1$), which is known to be biased in finite samples (Johnson et al., 1982), and a maximum likelihood estimator for a corresponding hazard ratio in a Weibull accelerated failure time model ($\widehat{\psi}_2$).

5.3 Estimator Convergence Rate

A final concept we will cover is the idea of the convergence rate of an estimator. This concept is very much related to the idea of the consistency of an estimator. We may have several consistent estimators available to us, and one criterion we can use to choose between them is how fast they converge to the true value.

The convergence rate of an estimator can be visualized in Figure 5:

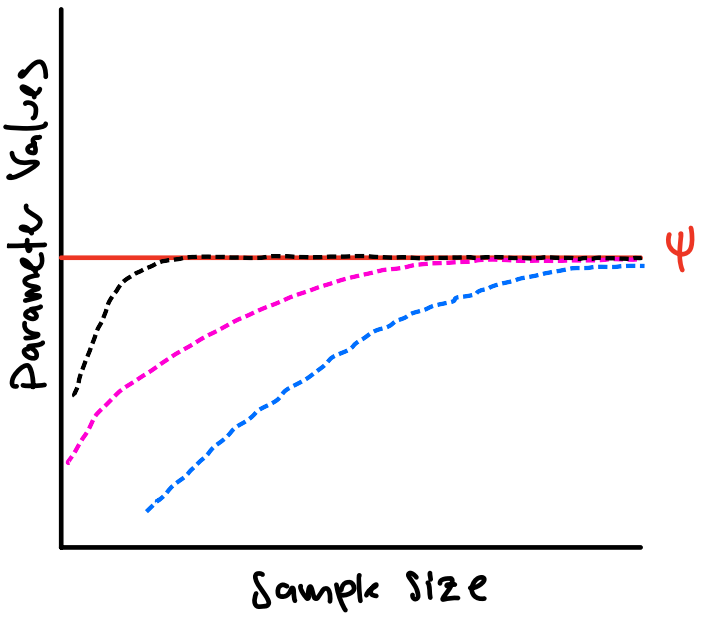


Figure 5: Demonstration of estimator convergence rate

This Figure shows three different estimators converging to the true parameter value ψ. The estimator indexed with a black dashed line converges the fastest, reaching the true parameter value at the smallest sample sizes. The estimator indexed by the magenta dashed line is the next fastest to converged, followed by the blue dashed line.

Overall, we would prefer the estimator indexed by the black dashed line, since it converges quickest to the truth. That means that we won't need as much data for optimal performance for the estimator indexed by the black dashed line versus the other two.

Statisticians typically describe estimator convergence rates as a function of the sample size n. For example, for a correctly specified parametric model in a low dimensional setting, the maximum likelihood estimator converges to the true value at an optimal root-n rate. This rate as often taken as the benchmark. The convergence rates of many other approaches are often compared to the root-n standard.

6 Estimates

Finally, the values obtained from each estimation approach (~0.05) are our **estimates**. The one important takeaway for estimates, particularly in light of the above discussion, is that these do not possess the properties we have discussed.

References

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, Pacific Grove, CA, 2nd edition, 2002.

Scott Cunningham. *Causal Inference: The Mixtape*. Yale University Press, New Haven, CT, 2021.

Alan E Hubbard, Jennifer Ahern, Nancy L Fleischer, Mark J van der Laan, Sheri A Lippman, Nicholas Jewell, Tim Bruckner, and William A Satariano. To gee or not to gee: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiol*, 21(4):467–474, 2010.

M E Johnson, H D Tolley, M C Bryson, and A S Goldman. Covariate analysis of survival data: a small-sample study of cox's model. *Biometrics*, 38(3): 685–698, Sep 1982.

Mirjam J Knol and Tyler J VanderWeele. Recommendations for presenting

analyses of effect modification and interaction. *International Journal of Epidemiology*, 41(2), 2012.

A I Naimi. The Counterfactual Implications of Fundamental Cause Theory. *Curr Epidemiol Reports*, In Press, 2016.

Orestis A Panagiotou and Sholom Wacholder. Invited commentary: How big is that interaction (in my community)–and in which direction? *Am J Epidemiol*, 180(12):1150–1158, Dec 2014.

James M. Robins and Sander Greenland. Adjusting for differential rates of prophylaxis therapy for pcp in high-versus low-dose azt treatment arms in an aids randomized trial. *J Am Stat Assoc*, 89(427):737–749, 1994.

K. J. Rothman, S. Greenland, and T. L. Lash. *Modern Epidemiology*. Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008.