Chapter 2

# Causality

Shallow men believe in luck, believe in circumstances.
Strong men believe in cause and effect.
— Ralph Waldo Emerson, *The Conduct of Life*

In this chapter, we consider causality, one of the most central concepts of quantitative social science. Much of social science research is concerned with the causal effects of various policies and other societal factors. Do small class sizes raise students' standardized test scores? Would universal health care improve the health and finances of the poor? What makes voters turn out in elections and determines their choice of candidates? To answer these causal questions, one must infer a counterfactual outcome and compare it with what actually happens (i.e., a factual outcome). We show how careful research design and data analysis can shed light on these causal questions that shape important academic and policy debates. We begin with a study of racial discrimination in the labor market. We then introduce various research designs useful for causal inference and apply them to additional studies concerning social pressure and voter turnout, as well as the impact of minimum-wage increases on employment. We also learn how to subset data in different ways and compute basic descriptive statistics in R.

## 2.1 Racial Discrimination in the Labor Market

Does racial discrimination exist in the labor market? Or, should racial disparities in the unemployment rate be attributed to other factors such as racial gaps in educational attainment? To answer this question, two social scientists conducted the following experiment.[1] In response to newspaper ads, the researchers sent out résumés of fictitious job candidates to potential employers. They varied only the names of job applicants, while leaving the other information in the résumés unchanged.

[1] This section is based on Marianne Bertrand and Sendhil Mullainathan (2004) "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American Economic Review*, vol. 94, no. 4, pp. 991–1013.

**Table 2.1.** Résumé Experiment Data.

| *Variable* | *Description* |
|---|---|
| `firstname` | first name of the fictitious job applicant |
| `sex` | sex of applicant (`female` or `male`) |
| `race` | race of applicant (`black` or `white`) |
| `call` | whether a callback was made (`1` = yes, `0` = no) |

For some candidates, stereotypically African-American-sounding names such as Lakisha Washington or Jamal Jones were used, whereas other résumés contained stereotypically white-sounding names, such as Emily Walsh or Greg Baker. The researchers then compared the callback rates between these two groups and examined whether applicants with stereotypically black names received fewer callbacks than those with stereotypically white names. The positions to which the applications were sent were either in sales, administrative support, clerical, or customer services.

Let's examine the data from this experiment in detail. We begin by loading the CSV data file, `resume.csv`, into R as a data frame object called `resume` using the function `read.csv()`. Table 2.1 presents the names and descriptions of the variables in this data set.

```
resume <- read.csv("resume.csv")
```

Instead of using `read.csv()`, you can also import the data set using the pull-down menu `Tools > Import Dataset > From Text File...` in RStudio.

This data frame object `resume` is an example of *experimental data*. Experimental data are collected from an experimental research design, in which a *treatment variable*, or a causal variable of interest, is manipulated in order to examine its causal effects on an *outcome variable*. In this application, the treatment refers to the race of a fictitious applicant, implied by the name given on the résumé. The outcome variable is whether the applicant receives a callback. We are interested in examining whether or not the résumés with different names yield varying callback rates.

**Experimental research** examines how a treatment causally affects an outcome by assigning varying values of the treatment variable to different observations, and measuring their corresponding values of the outcome variable.

```
dim(resume)
```

```
## [1] 4870    4
```

Using the `dim()` function, we can see that `resume` consists of 4870 observations and 4 variables. Each observation represents a fictitious job applicant. The outcome variable is whether the fictitious applicant received a callback from a prospective employer. The treatment variable is the race and gender of each applicant, though

more precisely the researchers were manipulating how potential employers perceive the gender and race of applicants, rather than directly manipulating those attributes.

Once imported, the data set is displayed in a spreadsheet-like format in an RStudio window. Alternatively, we can look at the first several observations of the data set using the `head()` function.

```
head(resume)
```

```
##   firstname    sex  race call
## 1   Allison female white    0
## 2   Kristen female white    0
## 3   Lakisha female black    0
## 4   Latonya female black    0
## 5    Carrie female white    0
## 6       Jay   male white    0
```

For example, the second observation contains a résumé for Kristen, identified as a white female who did not receive a callback. In addition, we can also create a summary of the data frame via the `summary()` function.

```
summary(resume)
```

```
##    firstname        sex          race
##  Tamika : 256   female:3746   black:2435
##  Anne   : 242   male  :1124   white:2435
##  Allison: 232
##  Latonya: 230
##  Emily  : 227
##  Latoya : 226
##  (Other):3457
##       call
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.08049
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
```

The summary indicates the number of résumés for each name, gender, and race as well as the overall proportion of résumés that received a callback. For example, there were 230 résumés whose applicants had the first name of "Latonya." The summary also shows that the data set contains the same number of black and white names, while there are more female than male résumés.

We can now begin to answer whether or not the résumés with African-American-sounding names are less likely to receive callbacks. To do this, we first create a

*contingency table* (also called a *cross tabulation*) summarizing the relationship between the race of each fictitious job applicant and whether a callback was received. A two-way contingency table contains the number of observations that fall within each category, defined by its corresponding row (`race` variable) and column (`call` variable). Recall that a variable in a data frame can be accessed using the `$` operator (see section 1.3.5). For example, the syntax `resume$race` will extract the `race` variable in the `resume` data frame.

```
race.call.tab <- table(race = resume$race,  call = resume$call)
race.call.tab
```

```
##        call
## race       0    1
##   black 2278  157
##   white 2200  235
```

The table shows, for example, that among 2435 (= 2278 + 157) résumés with stereotypically black names, only 157 received a callback. It is convenient to add totals for each row and column by applying the `addmargins()` function to the output of the `table()` function.

```
addmargins(race.call.tab)
```

```
##        call
## race       0    1  Sum
##   black 2278  157 2435
##   white 2200  235 2435
##   Sum   4478  392 4870
```

Using this table, we can compute the callback rate, or the proportion of those who received a callback, for the entire sample and then separately for black and white applicants.

```
## overall callback rate: total callbacks divided by the sample size
sum(race.call.tab[, 2]) / nrow(resume)
```

```
## [1] 0.08049281
```

```
## callback rates for each race
race.call.tab[1, 2] / sum(race.call.tab[1, ]) # black
```

```
## [1] 0.06447639
```

```
race.call.tab[2, 2] / sum(race.call.tab[2, ]) # white
```

```
## [1] 0.09650924
```

Recall that the syntax `race.call.tab[1, ]`, which does not specify the column number, extracts all the elements of the first row of this matrix. Note that in the square brackets, the number before the comma identifies the row of the matrix whereas the number after the comma identifies the column (see section 1.3.5). This can be seen by simply typing the syntax into R.

```
race.call.tab[1, ]  # the first row
##    0    1
## 2278  157

race.call.tab[, 2]  # the second column
## black white
##   157   235
```

From this analysis, we observe that the callback rate for the résumés with African-American-sounding names is 0.032, or 3.2 percentage points, lower than those with white-sounding names. While we do not know whether this is the result of intentional discrimination, the lower callback rate for black applicants suggests the existence of racial discrimination in the labor market. Specifically, our analysis shows that the same résumé with a black-sounding name is substantially less likely to receive a callback than an identical résumé with a white-sounding name.

An easier way to compute callback rates is to exploit the fact that `call` is a *binary variable*, or *dummy variable*, that takes the value 1 if a potential employer makes a callback and 0 otherwise. In general, the sample mean of a binary variable equals the sample proportion of 1s. This means that the callback rate can be conveniently calculated as the *sample mean*, or *sample average*, of this variable using the `mean()` function rather than dividing the counts of 1s by the total number of observations. For example, instead of the slightly more complex syntax we used above, the overall callback rate can be calculated as follows.

```
mean(resume$call)
## [1] 0.08049281
```

What about the callback rate for each race? To compute this using the `mean()` function, we need to first subset the data for each race and then compute the mean of the `call` variable within this subset. The next section shows how to subset data in R.

## 2.2 Subsetting the Data in R

In this section, we learn how to subset a data set in various ways. We first introduce logical values and operators, which enable us to specify which observations and variables of a data set should be extracted. We also learn about factor variables, which represent categorical variables in R.

### 2.2.1 LOGICAL VALUES AND OPERATORS

To understand subsetting, we first note that R has a special representation of the two *logical values*, `TRUE` and `FALSE`, which belong to the object class logical (see section 1.3.2).

```
class(TRUE)
## [1] "logical"
```

These logical values can be converted to a binary variable in the integer class using the function `as.integer()`, where `TRUE` is recoded as `1` and `FALSE` becomes `0`.

```
as.integer(TRUE)
## [1] 1

as.integer(FALSE)
## [1] 0
```

In many cases, R will coerce logical values into a binary variable so that performing numerical operations is straightforward. For example, in order to compute the proportion of `TRUE`s in a vector, one can simply use the `mean()` function to compute the sample mean of a logical vector. Similarly, we can use the `sum()` function to sum the elements of this vector in order to compute the total number of `TRUE`s.

```
x <- c(TRUE, FALSE, TRUE) # a vector with logical values
mean(x) # proportion of TRUEs
## [1] 0.6666667

sum(x) # number of TRUEs
## [1] 2
```

The logical values are often produced with the *logical operators* `&` and `|` corresponding to *logical conjunction* ("*AND*") and *logical disjunction* ("*OR*"), respectively. The value of "AND" (`&`) is `TRUE` only when both of the objects have a value of `TRUE`.

```
FALSE & TRUE
## [1] FALSE

TRUE & TRUE
## [1] TRUE
```

**Table 2.2.** Logical Conjunction "AND" and Disjunction "OR".

| *Statement a* | *Statement b* | *a AND b* | *a OR b* |
|---|---|---|---|
| TRUE | TRUE | TRUE | TRUE |
| TRUE | FALSE | FALSE | TRUE |
| FALSE | TRUE | FALSE | TRUE |
| FALSE | FALSE | FALSE | FALSE |

The table shows the value of *a* AND *b* and that of *a* OR *b* when statements *a* and *b* are either TRUE or FALSE.

"OR" (|) is used in a similar way. However, unlike "AND", "OR" is true when at least one of the objects has the value `TRUE`.

```
TRUE | FALSE
## [1] TRUE

FALSE | FALSE
## [1] FALSE
```

We summarize these relationships in table 2.2. For example, if one statement is `FALSE` and the other is `TRUE`, then the logical conjunction of the two statements is `FALSE` but their logical disjunction is `TRUE` (the second and third rows of the table).

With the same principle in mind, we can also chain multiple comparisons together where all elements must be `TRUE` in order for the syntax to return `TRUE`.

```
TRUE & FALSE & TRUE
## [1] FALSE
```

Furthermore, "AND" and "OR" can be used simultaneously, but parentheses should be used to avoid confusion.

```
(TRUE | FALSE) & FALSE # the parentheses evaluate to TRUE
## [1] FALSE

TRUE | (FALSE & FALSE) # the parentheses evaluate to FALSE
## [1] TRUE
```

We can perform the logical operations "AND" and "OR" on the entire vector all at once. In the following syntax example, each element of the `TF1` logical vector is compared against the corresponding element of the logical `TF2` vector.

```
TF1 <- c(TRUE, FALSE, FALSE)
TF2 <- c(TRUE, FALSE, TRUE)
TF1 | TF2
## [1]  TRUE FALSE  TRUE

TF1 & TF2
## [1]  TRUE FALSE FALSE
```

### 2.2.2 RELATIONAL OPERATORS

Relational operators evaluate the relationships between two values. They include "greater than" (>), "greater than or equal to" (>=), "less than" (<), "less than or equal to" (<=), "equal to" (==, which is different from =), and "not equal to" (!=). These operators return logical values.

```
4 > 3
## [1] TRUE

"Hello" == "hello"  # R is case sensitive
## [1] FALSE

"Hello" != "hello"
## [1] TRUE
```

Like the logical operators, the relational operators may be applied to vectors all at once. When applied to a vector, the operators evaluate each element of the vector.

```
x <- c(3, 2, 1, -2, -1)
x >= 2
## [1]  TRUE  TRUE FALSE FALSE FALSE

x != 1
## [1]  TRUE  TRUE FALSE  TRUE  TRUE
```

Since the relational operators produce logical values, we can combine their outputs with "AND" (&) and "OR" (|). When there are multiple instances of evaluation, it is good practice to put each evaluation within parentheses for ease of interpretation.

```
## logical conjunction of two vectors with logical values
(x > 0) & (x <= 2)
## [1] FALSE  TRUE  TRUE FALSE FALSE
```

```
## logical disjunction of two vectors with logical values
(x > 2) | (x <= -1)
## [1]  TRUE FALSE FALSE  TRUE  TRUE
```

As we saw earlier, the logical values, `TRUE` and `FALSE`, can be coerced into integers (`1` and `0` representing `TRUE` and `FALSE`, respectively). We can therefore compute the number and proportion of `TRUE` elements in a vector very easily.

```
x.int <- (x > 0) & (x <= 2) # logical vector
x.int
## [1] FALSE  TRUE  TRUE FALSE FALSE

mean(x.int) # proportion of TRUEs
## [1] 0.4

sum(x.int)  # number of TRUEs
## [1] 2
```

### 2.2.3 SUBSETTING

In sections 1.3.3 and 1.3.5, we learned how to subset vectors and data frames using indexing. Here, we show how to subset them using logical values, introduced above. At the end of section 2.1, we saw how to calculate the callback rate for the entire sample by applying the `mean()` function to the binary `call` variable. To compute the callback rate among the résumés with black-sounding names, we use the following syntax.

```
## callback rate for black-sounding names
mean(resume$call[resume$race == "black"])
## [1] 0.06447639
```

This command syntax subsets the `call` variable in the `resume` data frame for the observations whose values for the `race` variable are equal to `black`. That is, we can utilize square brackets `[ ]` to index the values in a vector by placing the logical value of each element into a vector of the same length within the square brackets. The elements whose indexing value is `TRUE` are extracted. The syntax then calculates the sample mean of this subsetted vector using the `mean()` function, which is equal to the proportion of subsetted observations whose values for the `call` variable are equal to `1`. It is instructive to print out the logical vector used inside the square brackets for

subsetting. We observe that if the value of the `race` variable equals `black` (`white`) for an observation then its corresponding element of the resulting logical vector is `TRUE` (`FALSE`).

```
## race of first 5 observations
resume$race[1:5]

## [1] white white black black white
## Levels: black white

## comparison of first 5 observations
(resume$race == "black")[1:5]

## [1] FALSE FALSE  TRUE  TRUE FALSE
```

Note that `Levels` in the above output represent the values of a *factor* or categorical variable, which will later be explained in detail (see section 2.2.5). The calculation of callback rate for black-sounding names can also be done in two steps. We first subset a data frame object so that it contains only the résumés with black-sounding names and then compute the callback rate.

```
dim(resume) # dimension of original data frame

## [1] 4870    4

## subset blacks only
resumeB <- resume[resume$race == "black", ]
dim(resumeB) # this data.frame has fewer rows than the original data.frame

## [1] 2435    4

mean(resumeB$call) # callback rate for blacks

## [1] 0.06447639
```

Here, the data frame `resumeB` contains only the information about the résumés with black-sounding names. Notice that we used square brackets `[,]` to index the rows of this original data frame. Unlike in the case of indexing vectors, we use a comma to separate row and column indexes. This comma is important and forgetting to include it will lead to an error.

Instead of indexing through the square brackets, we can alternatively use the `subset()` function to construct a data frame that contains just some of the original observations and just some of the original variables. The function's two primary arguments, other than the original data frame object, are the `subset` and `select` arguments. The `subset` argument takes a logical vector that indicates whether each individual row should be kept for the new data frame. The `select` argument takes a character vector that specifies the names of variables to be retained. For example,

the following syntax will extract the `call` and `firstname` variables for the résumés which contain female black-sounding names.

```
## keep "call" and "firstname" variables
## also keep observations with female black-sounding names
resumeBf <- subset(resume, select = c("call", "firstname"),
                   subset = (race == "black" & sex == "female"))
head(resumeBf)
```

```
##    call firstname
## 3     0   Lakisha
## 4     0   Latonya
## 8     0     Kenya
## 9     0   Latonya
## 11    0     Aisha
## 13    0     Aisha
```

When using the `subset()` function, we can eliminate the `subset` argument label. For example, `subset(resume, subset = (race == "black" & sex == "female"))` shortens to `subset(resume, race == "black" & sex == "female")`. Note that one could specify the data frame name to which the `race` and `sex` variables belong, i.e., `subset(resume, (resume$race == "black" & resume$sex == "female"))`, but this is unnecessary. By default, the variable names in this argument are assumed to come from the data frame specified in the first argument (`resume` in this case). So we can use simpler syntax: `subset(resume, (race == "black" & sex == "female"))`. It is important to pay close attention to parentheses so that each logical statement is contained within a pair of parentheses.

An identical subsetting result can be obtained using `[,]` rather than the `subset()` function, where the first element of the square brackets specifies the rows to be retained (using a logical vector) and the second element specifies the columns to be kept (using a character or integer vector).

```
## alternative syntax with the same results
resumeBf <- resume[resume$race == "black" & resume$sex == "female",
                   c("call", "firstname")]
```

We can now separately compute the racial gap in callback rate among female and male job applicants. Notice that we do not include a `select` argument to specify which variables to keep. Consequently, all variables will be retained.

```
## black male
resumeBm <- subset(resume, subset = (race == "black") & (sex == "male"))
```

```
## white female
resumeWf <- subset(resume, subset = (race == "white") & (sex == "female"))
## white male
resumeWm <- subset(resume, subset = (race == "white") & (sex == "male"))
## racial gaps
mean(resumeWf$call) - mean(resumeBf$call) # among females
```

```
## [1] 0.03264689
```

```
mean(resumeWm$call) - mean(resumeBm$call) # among males
```

```
## [1] 0.03040786
```

It appears that the racial gap exists but does not vary across gender groups. For both female and male job applicants, the callback rate is higher for whites than blacks by roughly 3 percentage points.

### 2.2.4 SIMPLE CONDITIONAL STATEMENTS

In many situations, we would like to perform different actions depending on whether a statement is true or false. These "actions" can be as complex or as simple as you need them to be. For example, we may wish to create a new variable based on the values of other variables in a data set. In chapter 4, we will learn more about *conditional statements*, but here we cover simple conditional statements that involve the `ifelse()` function.

The function `ifelse(X, Y, Z)` contains three elements. For each element in `X` that is `TRUE`, the corresponding element in `Y` is returned. In contrast, for each element in `X` that is `FALSE`, the corresponding element in `Z` is returned. For example, suppose that we want to create a new binary variable called `BlackFemale` in the `resume` data frame that equals `1` if the job applicant's name sounds black and female, and `0` otherwise. The following syntax achieves this goal.

```
resume$BlackFemale <- ifelse(resume$race == "black" &
                             resume$sex == "female", 1, 0)
```

We then use a three-way *contingency table* obtained by the `table()` function to confirm the result. As expected, the `BlackFemale` variable equals `1` only when a résumé belongs to a female African-American.

```
table(race = resume$race, sex = resume$sex,
      BlackFemale = resume$BlackFemale)
```

```
## , , BlackFemale = 0
##
##         sex
## race    female male
##   black      0  549
##   white   1860  575
```

```
##
## , , BlackFemale = 1
##
##        sex
## race    female male
##   black   1886    0
##   white      0    0
```

In the above output, the `, , BlackFemale = 0` and `, , BlackFemale = 1` headers indicate that the first two dimensions of the three-dimensional table are shown with the third variable, `BlackFemale`, equal to `0` and `1` for the first and second tables, respectively.

### 2.2.5 FACTOR VARIABLES

Next we show how to create a *factor variable* (or *factorial variable*) in R. A factor variable is another name for a *categorical variable* that takes a finite number of distinct values or levels. Here, we wish to create a factor variable that takes one of the four values, i.e., `BlackFemale`, `BlackMale`, `WhiteFemale`, and `WhiteMale`. To do this, we first create a new variable, `type`, which is filled with missing values `NA`. We then specify each type using the characteristics of the applicants.

```
resume$type <- NA
resume$type[resume$race == "black" & resume$sex == "female"] <- "BlackFemale"
resume$type[resume$race == "black" & resume$sex == "male"] <- "BlackMale"
resume$type[resume$race == "white" & resume$sex == "female"] <- "WhiteFemale"
resume$type[resume$race == "white" & resume$sex == "male"] <- "WhiteMale"
```

It turns out that this new variable is a character vector, and so we use the `as.factor()` function to turn this vector into a factor variable. While a factor variable looks like a character variable, the former actually has numeric values called *levels*, each of which has a character label. By default, the levels are sorted into alphabetical order based on their character labels. The levels of a factor variable can be obtained using the `levels()` function. Moreover, the `table()` function can be applied to obtain the number of observations that fall into each level.

```
## check object class
class(resume$type)
## [1] "character"

## coerce new character variable into a factor variable
resume$type <- as.factor(resume$type)
## list all levels of a factor variable
levels(resume$type)
## [1] "BlackFemale" "BlackMale" "WhiteFemale" "WhiteMale"
```

```
## obtain the number of observations for each level
table(resume$type)

##
## BlackFemale   BlackMale WhiteFemale   WhiteMale
##        1886         549        1860         575
```

The main advantage of factor objects is that R has a number of useful functionalities for them. One such example is the `tapply()` function, which applies a function repeatedly within each level of the factor variable. Suppose, for example, we want to calculate the callback rate for each of the four categories we just created. If we use the `tapply()` function this can be done in one line, rather than computing them one by one. Specifically, we use the function as in `tapply(X, INDEX, FUN)`, which applies the function indicated by argument `FUN` to the object `X` for each of the groups defined by unique values of the vector `INDEX`. Here, we apply the `mean()` function to the `call` variable separately for each category of the `type` variable using the `resume` data frame.

```
tapply(resume$call, resume$type, mean)

## BlackFemale   BlackMale WhiteFemale   WhiteMale
##  0.06627784  0.05828780  0.09892473  0.08869565
```

Recall that the order of arguments in a function matters unless the name of the argument is explicitly specified. The result indicates that black males have the lowest callback rate followed by black females, white males, and white females. We can even go one step further and compute the callback rate for each first name. Using the `sort()` function, we can sort the result into increasing order for ease of presentation.

```
## turn first name into a factor variable
resume$firstname <- as.factor(resume$firstname)
## compute callback rate for each first name
callback.name <- tapply(resume$call, resume$firstname, mean)
## sort the result into increasing order
sort(callback.name)

##      Aisha    Rasheed     Keisha   Tremayne     Kareem
## 0.02222222 0.02985075 0.03825137 0.04347826 0.04687500
##    Darnell     Tyrone      Hakim     Tamika    Lakisha
## 0.04761905 0.05333333 0.05454545 0.05468750 0.05500000
##    Tanisha       Todd      Jamal       Neil      Brett
## 0.05797101 0.05882353 0.06557377 0.06578947 0.06779661
##   Geoffrey    Brendan       Greg      Emily       Anne
## 0.06779661 0.07692308 0.07843137 0.07929515 0.08264463
```

```
##       Jill     Latoya      Kenya    Matthew    Latonya
## 0.08374384 0.08407080 0.08673469 0.08955224 0.09130435
##      Leroy    Allison      Ebony   Jermaine     Laurie
## 0.09375000 0.09482759 0.09615385 0.09615385 0.09743590
##      Sarah   Meredith     Carrie    Kristen        Jay
## 0.09844560 0.10160428 0.13095238 0.13145540 0.13432836
##       Brad
## 0.15873016
```

As expected from the above aggregate result, we find that many typical names for black males and females have low callback rates.

## 2.3 Causal Effects and the Counterfactual

In the résumé experiment, we are trying to quantify the *causal effects* of applicants' names on their likelihood of receiving a callback from a potential employer. What do we exactly mean by causal effects? How should we think about causality in general? In this section, we discuss a commonly used framework for *causal inference* in quantitative social science research.

The key to understanding causality is to think about the *counterfactual*. Causal inference is a comparison between the factual (i.e., what actually happened) and the counterfactual (i.e., what would have happened if a key condition were different). The very first observation of the résumé experiment data shows that a potential employer received a résumé with a stereotypically white female first name `Allison` but decided not to call back (the value of the `call` variable is `0` for this observation).

```
resume[1, ]
```

```
##   firstname    sex  race call BlackFemale        type
## 1   Allison female white    0           0 WhiteFemale
```

The key causal question here is whether the same employer would have called back if the applicant's name were instead a stereotypically African-American name such as `Lakisha`. Unfortunately, we would never observe this counterfactual outcome, because the researchers who conducted this experiment did not send out the same résumé to the same employer using `Lakisha` as the first name (perhaps out of fear that sending two identical résumés with different names would raise suspicion among potential employers).

Consider another example where researchers are interested in figuring out whether raising the minimum wage increases the unemployment rate. Some argue that increasing the minimum wage may not be helpful for the poor, because employers would hire fewer workers if they have to pay higher wages (or hire higher-skilled instead of low-skilled workers). Suppose that one state in a country decided to raise the minimum wage and in this state the unemployment rate increased afterwards. This does not

**Table 2.3.** Potential Outcome Framework of Causal Inference.

| *Résumé* $i$ | *Black-sounding name* $T_i$ | *Callback* $Y_i(1)$ | *Callback* $Y_i(0)$ | *Age* | *Education* |
|---|---|---|---|---|---|
| 1 | 1 | 1 | ? | 20 | college |
| 2 | 0 | ? | 0 | 55 | high school |
| 3 | 0 | ? | 1 | 40 | graduate school |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| $n$ | 1 | 0 | ? | 62 | college |

*Note:* The table illustrates the potential outcome framework of causal inference using the example of the résumé experiment. For each résumé of fictitious job applicant $i$, either the black-sounding, $T_i = 1$, or white-sounding, $T_i = 0$, name is used. The résumé contains other characteristics such as age and education, which are neither subject to nor affected by the manipulation. For a résumé with a black-sounding name, we can observe whether or not it receives a callback from the potential employer who received it, $Y_i(1)$, but will not be able to know the callback outcome if a white-sounding name was used, $Y_i(0)$. For every résumé, only one of the two potential outcomes is observed and the other is missing (indicated by "?").

necessarily imply that a higher minimum wage led to the increase in the unemployment rate. In order to know the causal effect of increasing the minimum wage, we would need to observe the unemployment rate that would have resulted if this state had not raised the minimum wage. Clearly, we would never be able to directly survey this counterfactual unemployment rate. Another example concerns the question of whether a job training program increases one's prospect of employment. Even if someone who actually had received job training secured a job afterwards, it does not necessarily follow that it was the job training program which led to the employment. The person may have become employed even in the absence of such a training program.

These examples illustrate the *fundamental problem of causal inference*, which arises because we cannot observe the counterfactual outcomes. We refer to a key causal variable of interest as a *treatment variable*, even though the variable may have nothing to do with a medical treatment. To determine whether a *treatment* variable of interest $T$, causes a change in an outcome variable $Y$, we must consider two *potential outcomes*, i.e., the potential values of $Y$ that would be realized in the presence and absence of the treatment, denoted by $Y(1)$ and $Y(0)$, respectively. In the résumé experiment, $T$ may represent the race of a fictitious applicant ($T = 1$ is a black-sounding name and $T = 0$ is a white-sounding name) while $Y$ denotes whether a potential employer who received the résumé called back. Then, $Y(1)$ and $Y(0)$ represent whether a potential employer calls back when receiving a résumé with stereotypically black and white names, respectively.

All of these variables can be defined for each observation and marked by a corresponding subscript. For example, $Y_i(1)$ represents the potential outcome under the treatment condition for the $i$th observation, and $T_i$ is the treatment variable for the same observation. Table 2.3 illustrates the potential outcome framework in the context of the résumé experiment. Each row represents an observation for which only one of the two potential outcomes is observed (the missing potential outcome is indicated by "?"). The treatment status $T_i$ determines which potential outcome is observed. Variables such as age and education are neither subject to nor affected by the manipulation of treatment.

We can now define, for each observation, the causal effect of $T_i$ on $Y_i$ as the difference between these two potential outcomes, $Y_i(1) - Y_i(0)$. The race of the applicant has a causal effect if a potential employer's decision to callback depends on it. As stated earlier, the fundamental problem of causal inference is that we are only able to observe one of the two potential outcomes even though causal inference requires comparison of both. An important implication is that for estimation of causal effects, we must find a credible way to infer these unobserved counterfactual outcomes. This requires making certain assumptions. The credibility of any causal inference, therefore, rests upon the plausibility of these identification assumptions.

For each observation $i$, we can define the **causal effect** of a binary treatment $T_i$ as the difference between two potential outcomes, $Y_i(1) - Y_i(0)$, where $Y_i(1)$ represents the outcome that would be realized under the treatment condition ($T_i = 1$) and $Y_i(0)$ denotes the outcome that would be realized under the control condition ($T_i = 0$).

The **fundamental problem of causal inference** is that we observe only one of the two potential outcomes, and which potential outcome is observed depends on the treatment status. Formally, the observed outcome $Y_i$ is equal to $Y_i(T_i)$.

This simple framework of causal inference also clarifies what is and is not an appropriate causal question. For example, consider a question of whether one's race causally affects one's employment prospects. In order to answer this question directly, it would be necessary to consider the counterfactual employment status if the applicant were to belong to a different racial group. However, this is a difficult proposition to address because one's race is not something that can be manipulated. Characteristics like gender and race are called *immutable characteristics*, and many scholars believe that causal questions about these characteristics are not answerable. In fact, there exists a mantra which states, "No causation without manipulation." It may be difficult to think about causality if the treatment variable of interest cannot be easily manipulated.

The résumé experiment, however, provides a clever way of addressing an important social science question about racial discrimination. Instead of tackling the difficult task of directly estimating the causal effect of race, the researchers of this study manipulated potential employers' *perception* of job applicants' race by changing the names on identical résumés. This research design strategy enables one to study racial discrimination in the causal inference framework by circumventing the difficulty of manipulating one's race itself. Many social scientists use similar research design strategies to study discrimination due to factors such as race, gender, and religion in various environments.

## 2.4 Randomized Controlled Trials

Now that we have provided the general definition of causal effects, how should we go about estimating them? We first consider *randomized experiments*, also referred

to as *randomized controlled trials* (RCTs), in which researchers randomly assign the receipt of treatment. An RCT is often regarded as the gold standard for establishing causality in many scientific disciplines because it enables researchers to isolate the effects of a treatment variable and quantify uncertainty. In this section, we discuss how randomization identifies the average causal effects. A discussion of how to quantify uncertainty will be given in chapter 7.

### 2.4.1 THE ROLE OF RANDOMIZATION

As explained in the previous section, the fundamental problem of causal inference states that for the estimation of causal effects, we must infer counterfactual outcomes. This problem prevents us from obtaining a valid estimate of the causal effect of treatment for each individual. However, it turns out that the randomization of treatment assignment enables the estimation of *average treatment effect*, which averages the treatment effect over a group of individuals.

Suppose that we are interested in estimating the *sample average treatment effect* (SATE), which is defined as the average of individual-level treatment effects in the sample.

The **sample average treatment effect** (SATE) is defined as the sample average of individual-level causal effects (i.e., $Y_i(1) - Y_i(0)$):

$$\text{SATE} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(1) - Y_i(0)\}, \tag{2.1}$$

where $n$ is the sample size, and $\sum_{i=1}^{n}$ denotes the summation operator from the first observation, $i = 1$, to the last, $i = n$.

The SATE is not directly observable. For the *treatment group* that received the treatment, we observe the average outcome under the treatment but do not know what their average outcome would have been in the absence of the treatment. The same problem exists for the *control group* because this group does not receive the treatment and as a result we do not observe the average outcome that would occur under the treatment condition.

In order to estimate the average counterfactual outcome for the treatment group, we may use the observed average outcome of the control group. Similarly, we can use the observed average outcome of the treatment group as an estimate of the average counterfactual outcome for the control group. This suggests that the SATE can be estimated by calculating the difference in the average outcome between the treatment and control groups or the *difference-in-means estimator*. The critical question is whether we can interpret this difference as a valid estimate of the average causal effect. In the résumé experiment, the treatment group consists of the potential employers who were sent résumés with black-sounding names. In contrast, the control group comprises other potential employers who received the résumés with stereotypically white names. Does the difference in callback rate between these two groups represent the average causal effect of the applicant's race?

Randomization of treatment assignment plays an essential role in enabling the interpretation of this *association* as a causal relationship. By randomly assigning each subject to either the treatment or control group, we ensure that these two groups are similar to each other in every aspect. In fact, even though they consist of different individuals, the treatment and control groups are *on average* identical to each other in terms of *all* pretreatment characteristics, both observed and unobserved. Since the only systematic difference between the two groups is the receipt of treatment, we can interpret the difference in the outcome variable as the estimated average causal effect of the treatment. In this way, the randomization of treatment assignment separates the causal effect of treatment from other possible factors that may influence the outcome. As we will see in section 2.5, we cannot guarantee that the treatment and control groups are comparable across all unobserved characteristics in the absence of random assignment.

In a **randomized controlled trial (RCT)**, each unit is randomly assigned either to the treatment or control group. The randomization of treatment assignment guarantees that the average difference in outcome between the treatment and control groups can be attributed solely to the treatment, because the two groups are on average identical to each other in all pretreatment characteristics.

RCTs, when successfully implemented, can yield valid estimates of causal effects. For this reason, RCTs are said to have a significant advantage for *internal validity*, which refers to whether the causal assumptions are satisfied in the study. However, RCTs are not without weaknesses. In particular, their strong internal validity often comes with a compromise in *external validity*. External validity is defined as the extent to which the conclusions can be generalized beyond a particular study. One common reason for a lack of external validity is that the study sample may not be representative of a population of interest. For ethical and logistical reasons, RCTs are often done using a convenient sample of subjects who are willing to be study subjects. This is an example of *sample selection bias*, making the experimental sample nonrepresentative of a target population. Another potential problem of external validity is that RCTs are often conducted in an environment (e.g., laboratory) quite different from real-world situations. In addition, RCTs may use interventions that are unrealistic in nature. As we saw in the résumé experiment, however, researchers have attempted to overcome these problems by conducting RCTs in the field and making their interventions as realistic as possible.

The main advantage of randomized controlled trials (RCTs) is their improved **internal validity**—the extent to which causal assumptions are satisfied in the study. One weakness of RCTs, however, is the potential lack of **external validity**—the extent to which the conclusions can be generalized beyond a particular study.

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach.
We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY – VOTE!

| MAPLE DR | Aug 04 | Nov 04 | Aug 06 |
|---|---|---|---|
| 9995 JOSEPH JAMES SMITH | Voted | Voted | _____ |
| 995 JENNIFER KAY SMITH | | Voted | _____ |
| 9997 RICHARD B JACKSON | | Voted | _____ |
| 9999 KATHY MARIE JACKSON | | Voted | _____ |

Figure 2.1. Naming-and-Shaming Get-out-the-Vote Message. Reprinted from Gerber, Green, and Larimer (2008).

### 2.4.2 SOCIAL PRESSURE AND VOTER TURNOUT

We consider a study of peer pressure and voter turnout,[2] another example of an RCT. Three social scientists conducted an RCT in which they investigated whether social pressure within neighborhoods increases participation. Specifically, during a primary election in the state of Michigan, they randomly assigned registered voters to receive different *get-out-the-vote* (GOTV) messages and examined whether sending postcards with these messages increased turnout. The researchers exploited the fact that the turnout of individual voters is public information in the United States.

The GOTV message of particular interest was designed to induce social pressure by telling voters that after the election their neighbors would be informed about whether they voted in the election or not. The researchers hypothesized that such a naming-and-shaming GOTV strategy would increase participation. An example of the actual naming-and-shaming message is shown in figure 2.1. In addition to the control group, which did not receive any mailing, the study also included other GOTV messages. For example, a standard "civic duty" message began with the same first two sentences of the naming-and-shaming message, but did not contain the additional information about neighbors learning about a person's electoral participation. Instead, the message continued to read as follows:

> The whole point of democracy is that citizens are active participants in government; that we have a voice in government. Your voice starts with your vote. On August 8, remember your rights and responsibilities as a citizen. Remember to vote. DO YOUR CIVIC DUTY – VOTE!

[2] This section is based on Alan S. Gerber, Donald P. Green, and Christopher W. Larimer (2008) "Social pressure and voter turnout: Evidence from a large-scale field experiment." *American Political Science Review*, vol. 102, no. 1, pp. 33–48.

Another important feature of this RCT is that the researchers attempted to separate the effect of naming-and-shaming from that of being observed. In many RCTs, there is a concern that study subjects may behave differently if they are aware of being observed by researchers. This phenomenon is called the *Hawthorne effect*, named after the factory where researchers observed an increase in workers' productivity simply because they knew that they were being monitored as part of a study. To address this issue, the study included another GOTV message, which starts with "YOU ARE BEING STUDIED!" followed by the same first two sentences as the naming-and-shaming message. The rest of the message reads,

> This year, we're trying to figure out why people do or do not vote. We'll be studying voter turnout in the August 8 primary election. Our analysis will be based on public records, so you will not be contacted again or disturbed in any way. Anything we learn about your voting or not voting will remain confidential and will not be disclosed to anyone else. DO YOUR CIVIC DUTY – VOTE!

The **Hawthorne effect** refers to the phenomenon where study subjects behave differently because they know they are being observed by researchers.

In this experiment, therefore, there are three treatment groups: voters who receive either the social pressure message, the civic duty message, or the Hawthorne effect message. The experiment also has a control group which consists of those voters receiving no message. The researchers randomly assigned each voter to one of the four groups and examined whether the voter turnout was different across the groups.

Now that we understand the design of this experiment, let us analyze the data. The data file, which is in CSV format, is named `social.csv` and can be loaded into R via the `read.csv()` function. Table 2.4 displays the names and descriptions of the variables in the social pressure experiment data.

```
social <- read.csv("social.csv") # load the data
summary(social) # summarize the data

##      sex          yearofbirth    primary2004
##  female:152702   Min.   :1900   Min.   :0.0000
##  male  :153164   1st Qu.:1947   1st Qu.:0.0000
##                  Median :1956   Median :0.0000
##                  Mean   :1956   Mean   :0.4014
##                  3rd Qu.:1965   3rd Qu.:1.0000
##                  Max.   :1986   Max.   :1.0000
##        messages       primary2006         hhsize
##  Civic Duty: 38218   Min.   :0.0000   Min.   :1.000
##  Control   :191243   1st Qu.:0.0000   1st Qu.:2.000
##  Hawthorne : 38204   Median :0.0000   Median :2.000
##  Neighbors : 38201   Mean   :0.3122   Mean   :2.184
##                      3rd Qu.:1.0000   3rd Qu.:2.000
##                      Max.   :1.0000   Max.   :8.000
```

**Table 2.4.** Social Pressure Experiment Data.

| *Variable* | *Description* |
|---|---|
| `hhsize` | household size of the voter |
| `messages` | GOTV messages the voter received (`Civic Duty`, `Control`, `Neighbors`, `Hawthorne`) |
| `sex` | sex of the voter (`female` or `male`) |
| `yearofbirth` | year of birth of the voter |
| `primary2004` | whether the voter voted in the 2004 primary election (`1`=voted, `0`=abstained) |
| `primary2006` | whether the voter turned out in the 2006 primary election (`1`=voted, `0`=abstained) |

As shown in section 2.2.5, we can use the `tapply()` function to compute the turnout for each treatment group. Subtracting the baseline turnout from the control group gives the average causal effect of each message. Note that the outcome variable of interest is the turnout in the 2006 primary election, which is coded as a binary variable `primary2006` where `1` represents turnout and `0` is abstention.

```
## turnout for each group
tapply(social$primary2006, social$messages, mean)

## Civic Duty    Control  Hawthorne  Neighbors
##  0.3145377  0.2966383  0.3223746  0.3779482

## turnout for control group
mean(social$primary2006[social$messages == "Control"])

## [1] 0.2966383

## subtract control group turnout from each group
tapply(social$primary2006, social$messages, mean) -
    mean(social$primary2006[social$messages == "Control"])

## Civic Duty    Control  Hawthorne  Neighbors
## 0.01789934 0.00000000 0.02573631 0.08130991
```

We find that the naming-and-shaming GOTV message substantially increases turnout. Compared to the control group turnout, the naming-and-shaming message increases turnout by 8.1 percentage points, whereas the civic duty message has a much smaller effect of 1.8 percentage points. It is interesting to see that the *Hawthorne effect* of being observed is somewhat greater than the effect of the civic duty message, though it is far smaller than the effect of the naming-and-shaming message.

Finally, if the randomization of treatment assignment is successful, we should not observe large differences across groups in the *pretreatment variables* such as age (indicated by `yearofbirth`), turnout in the previous primary election (`primary2004`), and household size (`hhsize`). We examine these using the same syntax.

```
social$age <- 2006 - social$yearofbirth # create age variable
tapply(social$age, social$messages, mean)

## Civic Duty    Control  Hawthorne  Neighbors
##   49.65904   49.81355   49.70480   49.85294

tapply(social$primary2004, social$messages, mean)

## Civic Duty    Control  Hawthorne  Neighbors
##  0.3994453  0.4003388  0.4032300  0.4066647

tapply(social$hhsize, social$messages, mean)

## Civic Duty    Control  Hawthorne  Neighbors
##   2.189126   2.183667   2.180138   2.187770
```

We see that the differences in these pretreatment variables are negligible across groups, confirming that the randomization of treatment assignment makes the four groups essentially identical to one another on average.

## 2.5 Observational Studies

Although RCTs can provide an internally valid estimate of causal effects, in many cases social scientists are unable to randomize treatment assignment in the real world for ethical and logistical reasons. We next consider *observational studies* in which researchers do not conduct an intervention. Instead, in observational studies, researchers simply observe naturally occurring events and collect and analyze the data. In such studies, *internal validity* is likely to be compromised because of possible selection bias, but *external validity* is often stronger than that of RCTs. The findings from observational studies are typically more generalizable because researchers can examine the treatments that are implemented among a relevant population in a real-world environment.

### 2.5.1 MINIMUM WAGE AND UNEMPLOYMENT

Our discussion of observational studies is based on the aforementioned minimum-wage debate. Two social science researchers examined the impact of raising the minimum wage on employment in the fast-food industry.[3] In 1992, the state of New Jersey (NJ) in the United States raised the minimum wage from $4.25 to $5.05 per hour. Did such an increase in the minimum wage reduce employment as economic theory predicts? As discussed above, answering this question requires inference about the NJ employment rate in the absence of such a raise in the minimum wage. Since this

[3] This section is based on David Card and Alan Krueger (1994) "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania." *American Economic Review*, vol. 84, no. 4, pp. 772–793.

**Table 2.5.** Minimum-Wage Study Data.

| *Variable* | *Description* |
|---|---|
| `chain` | name of the fast-food restaurant chain |
| `location` | location of the restaurants (`centralNJ`, `northNJ`, `PA`, `shoreNJ`, `southNJ`) |
| `wageBefore` | wage before the minimum-wage increase |
| `wageAfter` | wage after the minimum-wage increase |
| `fullBefore` | number of full-time employees before the minimum-wage increase |
| `fullAfter` | number of full-time employees after the minimum-wage increase |
| `partBefore` | number of part-time employees before the minimum-wage increase |
| `partAfter` | number of part-time employees after the minimum-wage increase |

counterfactual outcome is not observable, we must somehow estimate it using observed data.

One possible strategy is to look at another state in which the minimum wage did not increase. For example, the researchers of this study chose the neighboring state, Pennsylvania (PA), on the grounds that NJ's economy resembles that of Pennsylvania, and hence the fast-food restaurants in the two states are comparable. Under this *cross-section comparison design*, therefore, the fast-food restaurants in NJ serve as the *treatment group* receiving the treatment (i.e., the increase in the minimum wage), whereas those in PA represent the *control group*, which did not receive such a treatment. To collect pretreatment and outcome measures, the researchers surveyed the fast-food restaurants before and after the minimum wage increase. Specifically, they gathered information about the number of full-time employees, the number of part-time employees, and their hourly wages, for each restaurant.

The CSV file `minwage.csv` contains this data set. As usual, the `read.csv()` function loads the data set, the `dim()` function gives the number of observations and the number of variables, and the `summary()` function provides a summary of each variable. Table 2.5 displays the names and descriptions of the variables in the minimum-wage study data.

```
minwage <- read.csv("minwage.csv") # load the data
dim(minwage) # dimension of data

## [1] 358   8

summary(minwage) # summary of data

##         chain          location     wageBefore
##  burgerking:149   centralNJ: 45   Min.   :4.250
##  kfc       : 75   northNJ  :146   1st Qu.:4.250
```

```
##  roys      : 88   PA       : 67   Median :4.500
##  wendys    : 46   shoreNJ  : 33   Mean   :4.618
##                   southNJ  : 67   3rd Qu.:4.987
##                                   Max.   :5.750
##    wageAfter       fullBefore       fullAfter
##  Min.   :4.250   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:5.050   1st Qu.: 2.125   1st Qu.: 2.000
##  Median :5.050   Median : 6.000   Median : 6.000
##  Mean   :4.994   Mean   : 8.475   Mean   : 8.362
##  3rd Qu.:5.050   3rd Qu.:12.000   3rd Qu.:12.000
##  Max.   :6.250   Max.   :60.000   Max.   :40.000
##    partBefore      partAfter
##  Min.   : 0.00   Min.   : 0.00
##  1st Qu.:11.00   1st Qu.:11.00
##  Median :16.25   Median :17.00
##  Mean   :18.75   Mean   :18.69
##  3rd Qu.:25.00   3rd Qu.:25.00
##  Max.   :60.00   Max.   :60.00
```

To make sure that the restaurants followed the law, we first examine whether the minimum-wage actually increased in NJ after the law was enacted. We first subset the data based on location and then calculate the proportion of restaurants in each state with hourly wages less than the new minimum wage in NJ, i.e., \$5.05. This analysis can be done using the `wageBefore` and `wageAfter` variables, which represent the wage before and after the NJ law went into effect. The `subset()` function can be used to conduct this analysis.

```
## subsetting the data into two states
minwageNJ <- subset(minwage, subset = (location != "PA"))
minwagePA <- subset(minwage, subset = (location == "PA"))
## proportion of restaurants whose wage is less than $5.05
mean(minwageNJ$wageBefore < 5.05) # NJ before
## [1] 0.9106529

mean(minwageNJ$wageAfter < 5.05)  # NJ after
## [1] 0.003436426

mean(minwagePA$wageBefore < 5.05) # PA before
## [1] 0.9402985

mean(minwagePA$wageAfter < 5.05)  # PA after
## [1] 0.9552239
```

We observe that more than 91% of NJ restaurants were paying less than $5.05 before the minimum wage was raised and yet afterwards the proportion of such restaurants dramatically declined to less than 1%. In contrast, this proportion is essentially unchanged in PA, suggesting that the NJ law had minimal impact on the wages in PA restaurants. The analysis shows that the NJ restaurants followed the law by increasing their wage above the new minimum wage $5.05 while the PA restaurants did not have to make a similar change.

We now use the PA restaurants as the control group and estimate the average causal effect of increasing the minimum wage on employment among the NJ restaurants. An economic theory would predict that raising the minimum wage will encourage employers to replace full-time employees with part-time ones to recoup the increased cost in wages. To test this theory, we examine the proportion of full-time employees as a key outcome variable by simply comparing the *sample mean* of this variable between the NJ and PA restaurants after the NJ law went into effect. Let's compute this difference-in-means estimator.

```
## create a variable for proportion of full-time employees in NJ and PA
minwageNJ$fullPropAfter <- minwageNJ$fullAfter /
      (minwageNJ$fullAfter + minwageNJ$partAfter)
minwagePA$fullPropAfter <- minwagePA$fullAfter /
    (minwagePA$fullAfter + minwagePA$partAfter)
## compute the difference-in-means
mean(minwageNJ$fullPropAfter) - mean(minwagePA$fullPropAfter)

## [1] 0.04811886
```

The result of this analysis suggests that the increase in the minimum wage had no negative impact on employment. If anything, it appears to have slightly increased the proportion of full-time employment in NJ fast-food restaurants.

### 2.5.2 CONFOUNDING BIAS

The important assumption of observational studies is that the treatment and control groups must be comparable with respect to everything related to the outcome other than the treatment. In the current example, we cannot attribute the above difference in the full-time employment rate between NJ and PA restaurants to the minimum-wage increase in NJ if, for example, there is a competing industry for low-skilled workers in NJ but such an industry does not exist in PA. If that is the case, then the restaurants in the two states are not comparable and PA restaurants cannot serve as a valid control group for NJ restaurants. Indeed, NJ restaurants may have had a relatively high full-time employment rate, even in the absence of the increased minimum wage, in order to attract low-skilled workers. More generally, any other differences that exist between the fast-food restaurants in the two states before the administration of the NJ law would bias our inference if they are also related to outcomes.

The *pretreatment variables* that are associated with both the treatment and outcome variables are known as *confounders*. They are the variables that are realized prior to the administration of treatment and hence are not causally affected by the treatment.

However, they may determine who is likely to receive the treatment and influence the outcome. The existence of such variables is said to confound the causal relationship between the treatment and outcome, making it impossible to draw causal inferences from observational data. *Confounding bias* of this type is often a serious concern for social science research because in many cases human beings self-select into treatments. The aforementioned possibility that there exists a competing industry in NJ but not in PA is an example of confounding.

A pretreatment variable that is associated with both the treatment and the outcome variables is called a **confounder** and is a source of **confounding bias** in the estimation of the treatment effect.

Confounding bias due to self-selection into the treatment group is called *selection bias*. Selection bias often arises in observational studies because researchers have no control over who receives the treatment. In the minimum-wage study, NJ politicians decided to increase the minimum wage at this particular moment in time whereas politicians in PA did not. One might suspect that there were reasons, related to the economy and employment in particular, why the minimum wage was raised in NJ but not in PA. If that is the case, then the cross-sectional comparison of NJ and PA after the minimum-wage increase in NJ is likely to yield selection bias. The lack of control over treatment assignment means that those who self-select themselves into the treatment group may differ significantly from those who do not in terms of observed and unobserved characteristics. This makes it difficult to determine whether the observed difference in outcome between the treatment and control groups is due to the difference in the treatment condition or the differences in confounders. The possible existence of confounding bias is the reason behind the existence of the popular mantra, "Association does not necessarily imply causation."

In observational studies, the possibility of confounding bias can never be ruled out. However, researchers can try to address it by means of *statistical control*, whereby the researcher adjusts for confounders using statistical procedures. We describe some basic strategies in this section. One simple way is the statistical method called *subclassification*. The idea is to make the treatment and control groups as similar to each other as possible by comparing them within a subset of observations defined by shared values in pretreatment variables or a subclass. For example, we notice that the PA sample has a larger proportion of Burger Kings than the NJ sample. This difference between the two states could confound the relationship between minimum-wage increase and employment if, for example, Burger King has an employment policy that is different from that of other fast-food chains. To address this possibility, we could conduct a comparison only among Burger King restaurants. This analysis enables us to eliminate the confounding bias due to different fast-food chains through statistical control.

To begin our analysis, we first check the proportions of different fast-food chains for each of the two samples. We use the `prop.table()` function, which takes as its main input the output from the `table()` function, i.e., a table of counts, and converts it to proportions.

```
prop.table(table(minwageNJ$chain))
```

```
##
## burgerking        kfc       roys     wendys
##  0.4054983  0.2233677  0.2508591  0.1202749
```

```
prop.table(table(minwagePA$chain))
```

```
##
## burgerking        kfc       roys     wendys
##  0.4626866  0.1492537  0.2238806  0.1641791
```

The result shows that PA has a higher proportion of Burger King restaurants than NJ. We compare the full-time employment rate between NJ and PA Burger King restaurants after the increase in the minimum wage. Though not shown here, a similar analysis can be conducted for other fast-food chain restaurants as well.

```
## subset Burger King only
minwageNJ.bk <- subset(minwageNJ, subset = (chain == "burgerking"))
minwagePA.bk <- subset(minwagePA, subset = (chain == "burgerking"))
## comparison of full-time employment rates
mean(minwageNJ.bk$fullPropAfter) - mean(minwagePA.bk$fullPropAfter)
```

```
## [1] 0.03643934
```

This finding is quite similar to the overall result presented earlier, suggesting that the fast-food chain may not be a confounding factor.

Another possible confounder is the location of restaurants. In particular, it may be the case that the NJ Burger King restaurants closer to PA yield a more credible comparison to those in PA, perhaps because their local economies share similar characteristics. To address this possible confounding bias, we may further subclassify the data on the basis of restaurant location. Specifically, we focus on the Burger King restaurants located in northern and southern NJ that are near PA, while excluding those in the Jersey shore and central New Jersey, and repeat the analysis. This analysis adjusts for both the type of restaurants and their locations through statistical control.

```
minwageNJ.bk.subset <-
    subset(minwageNJ.bk, subset = ((location != "shoreNJ") &
                                   (location != "centralNJ")))
mean(minwageNJ.bk.subset$fullPropAfter) - mean(minwagePA.bk$fullPropAfter)
```

```
## [1] 0.03149853
```

The result shows that even within this smaller subset of the original data, the estimated impact of the minimum-wage increase remains similar to the overall estimate. This finding further improves our confidence in the claim that the increase in the minimum wage had little effect on full-time employment.

Confounding bias can be reduced through **statistical control**. For example, we can use the method of **subclassification** by comparing treated and control units which have an identical value of a confounding variable.

### 2.5.3 BEFORE-AND-AFTER AND DIFFERENCE-IN-DIFFERENCES DESIGNS

In observational studies, the data collected over time are a valuable source of information. Multiple measurements taken over time on the same units are called *longitudinal data* or *panel data*. Longitudinal data often yield a more credible comparison of the treatment and control groups than *cross-section data* because the former contain additional information about changes over time. In the minimum-wage study, the researchers had collected the employment and wage information from the same set of restaurants before the minimum wage was increased in NJ. This pretreatment information allows several alternative designs for estimating causal effects in observational studies.

The first possibility is comparison between pre- and posttreatment measurements, which is called the *before-and-after design*. Instead of comparing the fast-food restaurants in NJ with those in PA after the increase in the NJ minimum wage, this design compares the same set of fast-food restaurants in NJ before and after the minimum wage was raised. We compute the estimate under this design as follows.

```
## full-time employment proportion in the previous period for NJ
minwageNJ$fullPropBefore <- minwageNJ$fullBefore /
    (minwageNJ$fullBefore + minwageNJ$partBefore)
## mean difference between before and after the minimum wage increase
NJdiff <- mean(minwageNJ$fullPropAfter) - mean(minwageNJ$fullPropBefore)
NJdiff
```

```
## [1] 0.02387474
```

The before-and-after analysis gives an estimate that is similar to those obtained earlier. The advantage of this design is that any confounding factor that is specific to each state is held constant because the comparison is done within NJ. The disadvantage of the before-and-after design, however, is that time-varying confounding factors can bias the resulting inference. For example, suppose that there is an upwards *time trend* in the local economy and wages and employment are improving. If this trend is not caused by the minimum-wage increase, then we may incorrectly attribute the outcome difference between the two time periods to the raise in the minimum

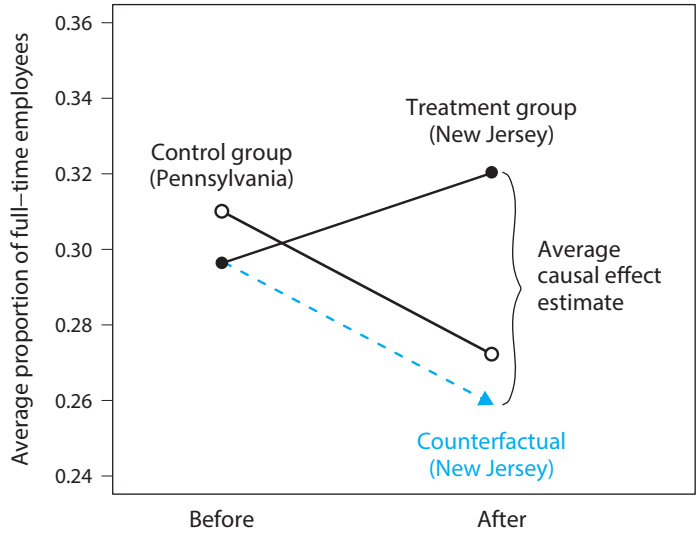


Figure 2.2. The Difference-in-Differences Design in the Minimum-Wage Study. The observed outcomes, i.e., the average proportion of full-time employees, are shown before and after the increase in the minimum wage for both the treatment group (fast-food restaurants in New Jersey; solid black circles) and the control group (restaurants in Pennsylvania; open black circles). Under the difference-in-differences design, the counterfactual outcome for the treatment group (solid blue triangle) is estimated by assuming that the time trend for the treatment group is parallel to the observed trend for the control group. The estimated average causal effect for New Jersey restaurants is indicated by the curly brace.

wage. The before-and-after design critically rests upon the nonexistence of such time trends.

The **before-and-after design** examines how the outcome variable changed from the pretreatment period to the posttreatment period for the same set of units. The design is able to adjust for any confounding factor that is specific to each unit but does not change over time. However, the design does not address possible bias due to time-varying confounders.

The *difference-in-differences* (DiD) design extends the before-and-after design to address the confounding bias due to time trends. The key assumption behind the DiD design is that the outcome variable follows a parallel trend in the absence of treatment. Figure 2.2 graphically illustrates this assumption using the minimum-wage study data. The figure shows the outcome of interest, i.e., the average proportion of full-time employees, before and after the increase in the minimum wage for both the treatment group (fast-food restaurants in NJ, indicated by the solid black circles) and the control group (restaurants in PA, represented by the open black circles). In this setting, we can estimate the counterfactual outcome for the treatment group by assuming that the time

trend for the treatment group is parallel to the observed trend for the control group. This estimate is indicated by the solid blue triangle.

Here, the counterfactual outcome of interest is the average proportion of full-time employees that we would have observed if NJ did not raise the minimum wage. We estimate this counterfactual outcome by supposing that NJ would have experienced the same economic trend as PA in the absence of the minimum-wage increase. In the figure, the blue dashed line is drawn to obtain the estimate of this counterfactual outcome and runs parallel to the observed time trend for the control group (indicated by the black solid line).

Under the DiD design, the sample average causal effect estimate for the NJ restaurants is the difference between the observed outcome after the minimum-wage increase and the counterfactual outcome derived under the parallel time trend assumption. The quantity of interest under the DiD design is called the *sample average treatment effect for the treated* (SATT). SATT differs from SATE, which is defined in equation (2.1), because it applies only to the treatment group, which consists of NJ restaurants in the current example.[4] In the figure, this estimate is indicated by the curly brace. To compute this estimate, we first calculate the difference in the outcome for the restaurants in PA after and before the minimum wage was raised in NJ. We then subtract this difference from the estimate obtained under the before-and-after design, which equals the difference in NJ after and before the minimum-wage increase. The average causal effect estimate is, therefore, given by the difference in the before-and-after differences between the treatment and control groups.

In this way, the DiD design uses the pretreatment and posttreatment measurements obtained for both the treatment and control groups. In contrast, the cross-section comparison requires only the posttreatment measurements from the two groups, and the before-and-after design utilizes the pretreatment and posttreatment measurements for the treatment group alone.

The **difference-in-differences** (DiD) design uses the following estimate of the sample average treatment effect for the treated (SATT):

$$\text{DiD estimate} = \underbrace{\left(\overline{Y}^{\text{after}}_{\text{treated}} - \overline{Y}^{\text{before}}_{\text{treated}}\right)}_{\text{difference for the treatment group}} - \underbrace{\left(\overline{Y}^{\text{after}}_{\text{control}} - \overline{Y}^{\text{before}}_{\text{control}}\right)}_{\text{difference for the control group}}.$$

The assumption is that the counterfactual outcome for the treatment group has a time trend parallel to that of the control group.

In the case of the minimum-wage study, we can compute the DiD estimate as follows.

[4] Formally, the sample average treatment effect for the treated (SATT) is the sample average of individual-level causal effect among the treated units, $\text{SATT} = \frac{1}{n_1}\sum_{i=1}^{n} T_i\{Y_i(1) - Y_i(0)\}$, where $T_i$ is the binary treatment indicator variable and $n_1 = \sum_{i=1}^{n} T_i$ is the size of the treatment group.

```
## full-time employment proportion in the previous period for PA
minwagePA$fullPropBefore <- minwagePA$fullBefore /
    (minwagePA$fullBefore + minwagePA$partBefore)
## mean difference between before and after for PA
PAdiff <- mean(minwagePA$fullPropAfter) - mean(minwagePA$fullPropBefore)
## difference-in-differences
NJdiff - PAdiff

## [1] 0.06155831
```

The result is inconsistent with the prediction of some economists that raising the minimum wage has a negative impact on employment. To the contrary, our DiD analysis suggests that, if anything, the increase in the minimum wage may have led to a small rise in the proportion of full-time employees in NJ fast-food restaurants. The DiD estimate is greater than the before-and-after estimate, which reflected a negative trend in PA.

When does the DiD design fail? The DiD design yields an invalid estimate of causal effect if the time trend of the counterfactual outcome for the treatment group is not parallel to the observed time trend for the control group. We cannot verify this assumption because the counterfactual time trend for the treatment group is unobserved. However, in some cases, we can increase the credibility of this assumption. For example, if researchers had collected employment information from the restaurants in earlier time periods, then they could have examined whether the proportion of full-time employees in NJ restaurants had changed parallel to that of PA restaurants when the minimum wage had not been raised.

## 2.6 Descriptive Statistics for a Single Variable

So far, we have been examining the average outcome as the quantity of interest, but it is also possible to consider some other statistics of outcome. As the final topic of this chapter, we discuss how to numerically summarize the distribution of a single variable using *descriptive statistics*. We have already seen some examples of descriptive statistics, including the range (i.e., minimum and maximum values), median, and mean. In this section, we introduce other commonly used univariate statistics to describe the distribution of a single variable.

### 2.6.1 QUANTILES

We begin by introducing *quantiles*, which divide a set of observations into groups based on the magnitude of the variable. An example of quantiles is the *median*, which divides the data into two groups, one with lower data values and the other with higher values. That is, the median of a variable equals the middle value if the total number of observations is odd, whereas the median is the average of two middle values if the total number of observations is even (because there is no single middle value in this case). For example, the median of {1, 3, 4, 10} is 3.5, which is the average of the middle values 3 and 4, because this example has an even number of values. Meanwhile, the mean of this vector is 4.5.

While both the mean and median measure the center of the distribution, the mean is more sensitive to *outliers*. For example, a single observation of extreme value can dramatically change the mean but it will not affect the median as much. The median of $\{1, 3, 4, 10, 82\}$ is 4, but the mean now increases to 20. In the minimum-wage data, the mean and median wages are similar. For example, the median wage before the minimum-wage increase is \$4.50, which is close to its mean of \$4.62.

The **median** of a variable $x$ is defined as:

$$\text{median} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{(n/2)} + x_{(n/2+1)}\right) & \text{if } n \text{ is even,} \end{cases} \tag{2.2}$$

where $x_{(i)}$ denotes the value of the $i$th smallest observation for variable $x$ and $n$ is the sample size. The median is less sensitive to outliers than the **mean** and hence is a more robust measure of the center of a distribution.

To examine the robustness of previous findings, we examine how the increase in the minimum wage influenced the proportion of full-time employees in terms of the median rather than the mean. The median of a variable can be computed by using the `median()` function.

```
## cross-section comparison between NJ and PA
median(minwageNJ$fullPropAfter) - median(minwagePA$fullPropAfter)
## [1] 0.07291667

## before and after comparison
NJdiff.med <- median(minwageNJ$fullPropAfter) -
    median(minwageNJ$fullPropBefore)
NJdiff.med
## [1] 0.025

## median difference-in-differences
PAdiff.med <- median(minwagePA$fullPropAfter) -
    median(minwagePA$fullPropBefore)
NJdiff.med - PAdiff.med
## [1] 0.03701923
```

These results are largely consistent with those of the previous analysis, though the DiD estimate is smaller than before. Again, there is little evidence for the hypothesis that increasing the minimum wage decreases full-time employment. If anything, it may have instead slightly increased full-time employment.

To obtain a more complete description of the distribution, we can use *quartiles*, which divide the data into four groups. The *first quartile* (or *lower quartile*) is the

value under which 25% of the observations fall, while the proportion of observations below the *third quartile* (or *upper quartile*) is 75%. The *second quartile* is equal to the median. The quartiles are a part of the output from the `summary()` function along with the minimum, mean, and maximum values. In addition, the difference between the upper and lower quartiles (i.e., 75th percentile and the 25th percentile) is called the *interquartile range* or *IQR*. That is, the IQR represents the range that contains 50% of the data, thereby measuring the spread of a distribution. This statistic can be computed by the `IQR()` function.

```
## summary shows quartiles as well as minimum, maximum, and mean
summary(minwageNJ$wageBefore)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.25    4.25    4.50    4.61    4.87    5.75

summary(minwageNJ$wageAfter)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   5.050   5.050   5.081   5.050   5.750

## interquartile range
IQR(minwageNJ$wageBefore)
## [1] 0.62

IQR(minwageNJ$wageAfter)
## [1] 0
```

This analysis shows that before the minimum-wage increase, the distribution of wages ranged from \$4.25 to \$5.75 with 75% of the fast-food restaurants in NJ having wages of \$4.87 per hour or less. However, after the minimum wage was raised to \$5.05, many restaurants raised their wages just to the new minimum wage but not any higher. As a result, both the lower and upper quartiles are equal to \$5.05, reducing the IQR from \$0.62 to \$0.

Finally, quartiles belong to a class of general statistics called *quantiles*, which divide the observations into a certain number of equally sized groups. Other quantiles include *terciles* (which divide the data into 3 groups), *quintiles* (5 groups), *deciles* (10 groups), and *percentiles* (100 groups). The `quantile()` function can generate any quantiles by specifying the `probs` argument. This argument takes a sequence of probabilities, indicating how the data should be divided up. For example, the deciles of the wage variable are obtained using the `seq()` function to create a sequence of numbers `0`, `0.1`, ..., `0.9`, `1`.

```
## deciles (10 groups)
quantile(minwageNJ$wageBefore, probs = seq(from = 0, to = 1, by = 0.1))
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
## 4.25 4.25 4.25 4.25 4.50 4.50 4.65 4.75 5.00 5.00 5.75
```

```
quantile(minwageNJ$wageAfter, probs = seq(from = 0, to = 1, by = 0.1))

##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
## 5.00 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.05 5.15 5.75
```

We find that at least 90% of the fast-food restaurants in NJ set their wages to \$5.05 or higher after the law was enacted. In contrast, before the increase in the minimum wage, there were few restaurants that offered wages of \$5.05 or higher. Thus, the law had a dramatic effect on raising the wage to the new minimum wage, but no higher than that. In fact, the highest wage stayed unchanged at \$5.75 even after the minimum wage was increased.

**Quantiles** represent a set of data values that divide observations into a certain number of equally sized groups. They include quartiles (dividing the observations into 4 groups) and percentiles (100 groups):

- 25th percentile = lower quartile;
- 50th percentile = median;
- 75th percentile = upper quartile.

The difference between the upper and lower quartiles is called the **interquartile range** and measures the spread of a distribution.

### 2.6.2 STANDARD DEVIATION

We have used the range and quantiles (including the IQR) to describe the spread of a distribution. Another commonly used measure is *standard deviation*. Before introducing standard deviation, we first describe a statistic called the *root mean square* or *RMS*. The RMS describes the magnitude of a variable and is defined as

$$\begin{aligned} \text{RMS} &= \sqrt{\text{mean of squared entries}} \\ &= \sqrt{\frac{\text{entry1}^2 + \text{entry2}^2 + \cdots}{\text{number of entries}}} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}. \end{aligned} \tag{2.3}$$

Equation (2.3) gives the formal mathematical definition. The equation exactly follows its name—square each entry, compute the mean, and then take the square root.

While the mean describes the center of the distribution, the RMS represents the average absolute magnitude of each data entry, ignoring the sign of the entry (e.g., the absolute magnitude or *absolute value* of $-2$ is 2 and is written as $|-2|$). For example, the mean of $\{-2, -1, 0, 1, 2\}$ is 0 but its RMS is $\sqrt{2}$. In the minimum-wage data, we can compute the RMS of the change in the proportion of full-time employees before and after the increase in the minimum wage, which is quite different from its mean.

```
sqrt(mean((minwageNJ$fullPropAfter - minwageNJ$fullPropBefore)^2))
## [1] 0.3014669

mean(minwageNJ$fullPropAfter - minwageNJ$fullPropBefore)
## [1] 0.02387474
```

Thus, on average, the absolute magnitude of change in the proportion of full-time employees, after the minimum wage was raised, is about 0.3. This represents a relatively large change even though the average difference is close to zero.

Using the RMS, we can define the sample *standard deviation* as the average deviation of each data entry from its mean. Therefore, the standard deviation measures the spread of a distribution by quantifying how far away data points are, on average, from their mean. Specifically, the standard deviation is defined as the RMS of deviation from the average:

$$\begin{aligned}\text{standard deviation} &= \text{RMS of deviation from average} \\ &= \sqrt{\frac{(\text{entry1} - \text{mean})^2 + (\text{entry2} - \text{mean})^2 + \cdots}{\text{number of entries}}} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \end{aligned} \tag{2.4}$$

In some cases, one uses $n - 1$ instead of $n$ in the denominator of equation (2.4) for a reason that will become clear in chapter 7, but this results in only a minor difference so long as one has enough data. We note that few data points are more than 2 or 3 standard deviations away from the mean. Hence, knowing the standard deviation helps researchers understand the approximate range of the data as well. Finally, the square of the standard deviation is called the *variance* and represents the average squared deviation from the mean. We will study variance more closely in later chapters. Variance is more difficult to interpret than standard deviation, but it has useful analytical properties, as shown in chapter 6.

The sample **standard deviation** measures the average deviation from the mean and is defined as

$$\text{standard deviation} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{or} \quad \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where $\bar{x}$ represents the sample mean, i.e., $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $n$ is the sample size. Few data points lie outside 2 or 3 standard deviations away from the mean. The square of the standard deviation is called the **variance**.

In R, we can easily compute the standard deviation using the `sd()` function (this function uses $n - 1$ in its denominator). The `var()` function returns the sample variance. The examples from the minimum-wage data are given here.

```
## standard deviation
sd(minwageNJ$fullPropBefore)
## [1] 0.2304592

sd(minwageNJ$fullPropAfter)
## [1] 0.2510016

## variance
var(minwageNJ$fullPropBefore)
## [1] 0.05311145

var(minwageNJ$fullPropAfter)
## [1] 0.0630018
```

The results indicate that, on average, the proportion of full-time employees for a NJ fast-food restaurant is approximately 0.2 away from its mean. We find that for this variable the standard deviation did not change much after the minimum wage had been increased.

## 2.7 Summary

We began this chapter with the analysis of an experimental study concerning racial discrimination in the labor market. The **fundamental problem of causal inference** is the fact that we observe only one of two potential outcomes and yet the estimation of causal effect involves comparison between counterfactual and factual outcomes. This chapter also introduced various research design strategies to infer counterfactual outcomes from observed data. It is important to understand the assumptions that underlie each research design as well as their strengths and weaknesses.

In **randomized controlled experiments (RCTs)**, a simple comparison of the treatment and control groups enables researchers to estimate the causal effects of treatment. By randomizing the treatment assignment, we can ensure that the treatment and control groups are, on average, identical to each other in all observed and unobserved characteristics except for the receipt of treatment. Consequently, any average difference between the treatment and control groups can be attributed to the treatment. While RCTs tend to yield internally valid estimates of causal effects, they often suffer from a lack of external validity, which makes it difficult to generalize empirical conclusions to a relevant population in real-world settings.

In **observational studies**, researchers do not directly conduct interventions. Since some subjects may self-select into the treatment group, the difference in outcome between the treatment and control groups can be attributed to factors other than the receipt of treatment. Thus, while observational studies often have stronger external validity, this advantage typically comes with compromises in internal validity. When the treatment assignment is not randomized, we must confront the possibility of confounding bias in observational studies using statistical control. The existence of confounders that are associated with both the treatment and outcome means that a simple comparison of the two groups yields misleading inference. We introduced various research design strategies to reduce such bias, including subclassification, before-and-after design, and difference-in-differences design.

Finally, we learned how to subset data in various ways using R. Subsetting can be done using logical values, relational operators, and conditional statements. We also introduced a number of descriptive statistics that are useful for summarizing each variable in a data set. They include the mean, median, quantiles, and standard deviation. R provides a set of functions that enable researchers to compute these and other descriptive statistics from their data sets.

## 2.8 Exercises

### 2.8.1 EFFICACY OF SMALL CLASS SIZE IN EARLY EDUCATION

The STAR (Student–Teacher Achievement Ratio) Project is a four-year *longitudinal study* examining the effect of class size in early grade levels on educational performance and personal development.[5] A longitudinal study is one in which the same participants are followed over time. This particular study lasted from 1985 to 1989 and involved 11,601 students. During the four years of the study, students were randomly assigned to small classes, regular-sized classes, or regular-sized classes with an aid. In all, the experiment cost around $12 million. Even though the program stopped in 1989 after the first kindergarten class in the program finished third grade, the collection of various measurements (e.g., performance on tests in eighth grade, overall high-school GPA) continued through to the end of participants' high-school attendance.

We will analyze just a portion of this data to investigate whether the small class sizes improved educational performance or not. The data file name is `STAR.csv`, which is

[5] This exercise is in part based on Frederick Mosteller (1995) "The Tennessee study of class size in the early school grades." *The Future of Children*, vol. 5, no. 2, pp. 113–127.

**Table 2.6.** STAR Project Data.

| *Variable* | *Description* |
|---|---|
| `race` | student's race (white = `1`, black = `2`, Asian = `3`, Hispanic = `4`, Native American = `5`, others = `6`) |
| `classtype` | type of kindergarten class (small = `1`, regular = `2`, regular with aid = `3`) |
| `g4math` | total scaled score for the math portion of the fourth-grade standardized test |
| `g4reading` | total scaled score for the reading portion of the fourth-grade standardized test |
| `yearssmall` | number of years in small classes |
| `hsgrad` | high-school graduation (did graduate = `1`, did not graduate = `0`) |

in CSV format. The names and descriptions of variables in this data set are displayed in table 2.6. Note that there are a fair amount of missing values in this data set, which arise, for example, because some students left a STAR school before third grade, or did not enter a STAR school until first grade.

1. Create a new factor variable called `kinder` in the data frame. This variable should recode `classtype` by changing integer values to their corresponding informative labels (e.g., change `1` to `small` etc.). Similarly, recode the `race` variable into a factor variable with four levels (`white`, `black`, `hispanic`, `others`) by combining the Asian and Native American categories with the others category. For the `race` variable, overwrite the original variable in the data frame rather than creating a new one. Recall that `na.rm = TRUE` can be added to functions in order to remove missing data (see section 1.3.5).

2. How does performance on fourth-grade reading and math tests for those students assigned to a small class in kindergarten compare with those assigned to a regular-sized class? Do students in the smaller classes perform better? Use means to make this comparison while removing missing values. Give a brief substantive interpretation of the results. To understand the size of the estimated effects, compare them with the standard deviation of the test scores.

3. Instead of just comparing average scores of reading and math tests between those students assigned to small classes and those assigned to regular-sized classes, look at the entire range of possible scores. To do so, compare a high score, defined as the 66th percentile, and a low score (the 33rd percentile) for small classes with the corresponding score for regular classes. These are examples of *quantile treatment effects*. Does this analysis add anything to the analysis based on mean in the previous question?

4. Some students were in small classes for all four years that the STAR program ran. Others were assigned to small classes for only one year and had either regular-sized classes or regular-sized classes with an aid for the rest. How many students

**Table 2.7.** Gay Marriage Data.

| *Variable* | *Description* |
|---|---|
| `study` | source of the data (`1` = study 1, `2` = study 2) |
| `treatment` | five possible treatment assignment options |
| `wave` | survey wave (a total of seven waves) |
| `ssm` | five-point scale on same-sex marriage, higher scores indicate support. |

of each type are in the data set? Create a contingency table of proportions using the `kinder` and `yearssmall` variables. Does participation in more years of small classes make a greater difference in test scores? Compare the average and median reading and math test scores across students who spent different numbers of years in small classes.

5. Examine whether the STAR program reduced achievement gaps across different racial groups. Begin by comparing the average reading and math test scores between white and minority students (i.e., blacks and Hispanics) among those students who were assigned to regular-sized classes with no aid. Conduct the same comparison among those students who were assigned to small classes. Give a brief substantive interpretation of the results of your analysis.

6. Consider the long-term effects of kindergarten class size. Compare high-school graduation rates across students assigned to different class types. Also, examine whether graduation rates differ depending on the number of years spent in small classes. Finally, as in the previous question, investigate whether the STAR program has reduced the racial gap between white and minority students' graduation rates. Briefly discuss the results.

### 2.8.2 CHANGING MINDS ON GAY MARRIAGE

In this exercise, we analyze the data from two experiments in which households were canvassed for support on gay marriage.[6] Note that the original study was later retracted due to allegations of fabricated data; we will revisit this issue in a follow-up exercise (see section 3.9.1). In this exercise, however, we analyze the original data while ignoring the allegations.

Canvassers were given a script leading to conversations that averaged about twenty minutes. A distinctive feature of this study is that gay and straight canvassers were randomly assigned to households, and canvassers revealed whether they were straight or gay in the course of the conversation. The experiment aims to test the "contact hypothesis," which contends that out-group hostility (towards gay people in this case) diminishes when people from different groups interact with one another. The data file is `gay.csv`, which is a CSV file. Table 2.7 presents the names and descriptions

[6] This exercise is based on the following article: Michael J. LaCour and Donald P. Green (2015) "When contact changes minds: An experiment on transmission of support for gay equality." *Science*, vol. 346, no. 6215, pp. 1366–1369.

of the variables in this data set. Each observation of this data set is a respondent giving a response to a four-point survey item on same-sex marriage. There are two different studies in this data set, involving interviews during seven different time periods (i.e., seven waves). In both studies, the first wave consists of the interview before the canvassing treatment occurs.

1. Using the baseline interview wave before the treatment is administered, examine whether randomization was properly conducted. Base your analysis on the three groups of study 1: "same-sex marriage script by gay canvasser," "same-sex marriage script by straight canvasser" and "no contact." Briefly comment on the results.

2. The second wave of the survey was implemented two months after canvassing. Using study 1, estimate the average treatment effects of gay and straight canvassers on support for same-sex marriage, separately. Give a brief interpretation of the results.

3. The study contained another treatment that involves contact, but does not involve using the gay marriage script. Specifically, the authors used a script to encourage people to recycle. What is the purpose of this treatment? Using study 1 and wave 2, compare outcomes from the treatment "same-sex marriage script by gay canvasser" to "recycling script by gay canvasser." Repeat the same for straight canvassers, comparing the treatment "same-sex marriage script by straight canvasser" to "recycling script by straight canvasser." What do these comparisons reveal? Give a substantive interpretation of the results.

4. In study 1, the authors reinterviewed the respondents six different times (in waves 2 to 7) after treatment, at two-month intervals. The last interview, in wave 7, occurs one year after treatment. Do the effects of canvassing last? If so, under what conditions? Answer these questions by separately computing the average effects of straight and gay canvassers with the same-sex marriage script for each of the subsequent waves (relative to the control condition).

5. The researchers conducted a second study to replicate the core results of the first study. In this study, same-sex marriage scripts are given only by gay canvassers. For study 2, use the treatments "same-sex marriage script by gay canvasser" and "no contact" to examine whether randomization was appropriately conducted. Use the baseline support from wave 1 for this analysis.

6. For study 2, estimate the treatment effects of gay canvassing using data from wave 2. Are the results consistent with those of study 1?

7. Using study 2, estimate the average effect of gay canvassing at each subsequent wave and observe how it changes over time. Note that study 2 did not have a fifth or sixth wave, but the seventh wave occurred one year after treatment, as in study 1. Draw an overall conclusion from both study 1 and study 2.

**Table 2.8.** Leader Assassination Data.

| *Variable* | *Description* |
|---|---|
| `country` | country |
| `year` | year |
| `leadername` | name of the leader who was targeted |
| `age` | age of the targeted leader |
| `politybefore` | average polity score of the country during the three-year period prior to the attempt |
| `polityafter` | average polity score of the country during the three-year period after the attempt |
| `civilwarbefore` | `1` if the country was in civil war during the three-year period prior to the attempt, `0` otherwise |
| `civilwarafter` | `1` if the country was in civil war during the three-year period after the attempt, `0` otherwise |
| `interwarbefore` | `1` if the country was in international war during the three-year period prior to the attempt, `0` otherwise |
| `interwarafter` | `1` if the country was in international war during the three-year period after the attempt, `0` otherwise |
| `result` | result of the assassination attempt |

### 2.8.3 SUCCESS OF LEADER ASSASSINATION AS A NATURAL EXPERIMENT

One longstanding debate in the study of international relations concerns the question of whether individual political leaders can make a difference. Some emphasize that leaders with different ideologies and personalities can significantly affect the course of a nation. Others argue that political leaders are severely constrained by historical and institutional forces. Did individuals like Hitler, Mao, Roosevelt, and Churchill make a big difference? The difficulty of empirically testing these arguments stems from the fact that the change of leadership is not random and there are many confounding factors to be adjusted for.

In this exercise, we consider a *natural experiment* in which the success or failure of assassination attempts is assumed to be essentially random.[7] Each observation of the CSV data set `leaders.csv` contains information about an assassination attempt. Table 2.8 presents the names and descriptions of variables in this leader assassination data set. The `polity` variable represents the so-called *polity score* from the Polity Project. The Polity Project systematically documents and quantifies the regime types of all countries in the world from 1800. The polity score is a 21-point scale ranging from −10 (hereditary monarchy) to 10 (consolidated democracy). The `result` variable is a 10-category factor variable describing the result of each assassination attempt.

1. How many assassination attempts are recorded in the data? How many countries experience at least one leader assassination attempt? (The `unique()` function,

[7] This exercise is based on the following article: Benjamin F. Jones and Benjamin A. Olken (2009) "Hit or miss? The effect of assassinations on institutions and war." *American Economic Journal: Macroeconomics*, vol. 1, no. 2, pp. 55–87.

which returns a set of unique values from the input vector, may be useful here.) What is the average number of such attempts (per year) among these countries?

2. Create a new binary variable named `success` that is equal to `1` if a leader dies from the attack and `0` if the leader survives. Store this new variable as part of the original data frame. What is the overall success rate of leader assassination? Does the result speak to the validity of the assumption that the success of assassination attempts is randomly determined?

3. Investigate whether the average polity score over three years prior to an assassination attempt differs on average between successful and failed attempts. Also, examine whether there is any difference in the age of targeted leaders between successful and failed attempts. Briefly interpret the results in light of the validity of the aforementioned assumption.

4. Repeat the same analysis as in the previous question, but this time using the country's experience of civil and international war. Create a new binary variable in the data frame called `warbefore`. Code the variable such that it is equal to `1` if a country is in either civil or international war during the three years prior to an assassination attempt. Provide a brief interpretation of the result.

5. Does successful leader assassination cause democratization? Does successful leader assassination lead countries to war? When analyzing these data, be sure to state your assumptions and provide a brief interpretation of the results.