2.7 CHEATSHEETS

2.7.1 CONCEPTS AND NOTATION

concept/notation	description	example(s)
causal relationship	refers to the cause-and-effect connection between two variables in which a change in one variable systematically produces a change in the other; we represent a causal relationship with an arrow between the variables: $$X \rightarrow Y$$	in this chapter, we explore the causal relationship between attending a small class and student performance: $$\textit{small} \rightarrow \textit{performance}$$ the question we aim to answer is, does attending a small class increase, decrease, or have a zero effect on student performance, on average?
treatment variable (X)	variable whose change may produce a change in the outcome variable; variable where the change originates; in this book, the treatment variable is always binary: $$X_i = \begin{cases} 1 & \text{if individual } i \text{ receives the treatment} \\ 0 & \text{if individual } i \text{ does not receive the treatment} \end{cases}$$ treatment variables are a type of independent variable	in Project STAR, the treatment variable is *small*, which we define as: $$\textit{small}_i = \begin{cases} 1 & \text{if student } i \text{ attended a small class} \\ 0 & \text{if student } i \text{ attended a regular-size class} \end{cases}$$
outcome variable (Y)	variable that may change as a result of a change in the treatment variable; outcome variables are the same as dependent variables	in these causal relationships: $\textit{small} \rightarrow \textit{reading}$ $\textit{small} \rightarrow \textit{math}$ $\textit{small} \rightarrow \textit{graduated}$ *small* is the treatment variable, and *reading*, *math*, and *graduated* are the outcome variables
treatment condition	the condition when the treatment is present; condition when X_i=1	in Project STAR, students attending a small class were under the treatment condition
control condition	the condition when the treatment is absent; condition when X_i=0	in Project STAR, students attending a regular-size class were under the control condition
potential outcome under the treatment condition ($Y_i(X_i{=}1)$)	one of the two potential outcomes for individual i; potential outcome for individual i when the treatment is present; the value of Y_i if X_i=1	in Project STAR, the potential outcome under the treatment condition is student performance after attending a small class from kindergarten until third grade
potential outcome under the control condition ($Y_i(X_i{=}0)$)	one of the two potential outcomes for individual i; potential outcome for individual i when the treatment is absent; the value of Y_i if X_i=0	in Project STAR, the potential outcome under the control condition is student performance after attending a regular-size class from kindergarten until third grade
$\triangle$	Greek letter Delta; mathematical notation for change	$\triangle Y_i$ represents the change in Y for individual i

continues on next page...

2.7.1 CONCEPTS AND NOTATION (CONTINUED)

concept/notation	description	example(s)
individual causal effect of X on Y	change in the outcome variable Y caused by a change in the treatment variable X; if we could observe both potential outcomes for each individual, we could measure it as: $$individual\_effects_i = Y_i(X_i{=}1) - Y_i(X_i{=}0)$$	suppose that the first student in the dataset (i=1) would have scored 720 points on the reading test after attending a small class, and 700 points after attending a regular-size class; therefore: - $reading_1(\text{small}_1{=}1) = 720$ - $reading_1(\text{small}_1{=}0) = 700$ in this hypothetical case, the individual causal effect of attending a small class on this student's performance on the reading test would have been: causal effect of *small* on *reading* = $= Y_i(X_i{=}1) - Y_i(X_i{=}0)$ $= reading_1(small_1{=}1) - reading_1(small_1{=}0)$ $= 720 - 700 = 20$ attending a small class, as opposed to a regular-size one, would have increased this student's performance on the reading test by 20 points
factual outcome	potential outcome under whichever condition (treatment or control) was received in reality; we always observe the factual outcomes	if a student attended a small class, the factual outcome is this student's performance after attending a small class, which we observe
counterfactual outcome	potential outcome under whichever condition (treatment or control) was not received in reality; we never observe the counterfactual outcomes	if a student attended a small class, the counterfactual outcome is this student's performance after attending a regular-size class, which we do not observe
fundamental problem of causal inference	we never observe the counterfactual outcome; we cannot measure the individual causal effect of a treatment on an outcome because we never observe both potential outcomes; the individual causal effect is $Y_i(X_i{=}1) - Y_i(X_i{=}0)$, but we can observe only one of the two potential outcomes, $Y_i(X_i{=}1)$ or $Y_i(X_i{=}0)$, whichever occurs in reality	students attend either a small class or a regular-size class, but they cannot attend both types of classes at the same time; we can never observe each student's performance under both the treatment and control conditions, and therefore, we cannot measure the effect of attending a small class on a specific student's performance
average causal effect of X on Y or average treatment effect	effect that X has on Y at the aggregate level; average of the individual causal effects of X on Y across a group of observations: $$\overline{individual\_effects} = \frac{\sum_{i=1}^{n} individual\_effects_i}{n}$$ average change in the outcome variable Y caused by a change in the treatment variable X for a group of observations; if treatment and control groups were comparable before the treatment was administered, then we can estimate the average treatment effect using the difference-in-means estimator	(see difference-in-means estimator)

continues on next page...

2.7.1 CONCEPTS AND NOTATION (CONTINUED)

concept/notation	description	example(s)
randomized experiment	also known as a randomized controlled trial (RCT); type of study design in which treatment assignment (who receives and does not receive the treatment) is randomized; the randomization of the treatment assignment ensures that treatment and control groups are, on average, identical to each other in all observed and unobserved pre-treatment characteristics	Project STAR was a randomized experiment in which students were randomly assigned to attend either a small class or a regular-size class; as a result, the students who attended a small class should have similar pre-treatment characteristics as the students who attended a regular-size class; for example, the average age of the students in both groups should be comparable
treatment group	group of individuals who received the treatment; observations for which X_i=1	in Project STAR, students attending a small class were in the treatment group
control group	group of individuals who did not receive the treatment; observations for which X_i=0	in Project STAR, students attending a regular-size class were in the control group
pre-treatment characteristics	characteristics of the individuals in a study before the treatment is administered; by definition, these characteristics cannot be affected by the treatment	in Project STAR, before students were assigned to small or regular-size classes, researchers recorded students' demographic data, such as age, gender, and race/ethnicity
difference-in-means estimator	the difference-in-means estimator is defined as the average outcome for the treatment group minus the average outcome for the control group: $\overline{Y}_{\text{treatment group}} - \overline{Y}_{\text{control group}}$ when treatment and control groups are similar with respect to all the variables that might affect the outcome other than the treatment variable itself, it produces a valid estimate of the average causal effect of X on Y; in this case, it estimates the average change in Y caused by a change in X interpret as: - an average increase in Y if positive - an average decrease in Y if negative - no average change in Y if zero unit of measurement of this estimator: - if Y is non-binary: in the same unit of measurement as Y - if Y is binary: in percentage points (after multiplying the result by 100)	in the STAR dataset, the difference-in-means estimator for the reading test scores is 632.7 points - 625.49 points = 7.21 points because Project STAR was a randomized experiment, the difference-in-means is a valid estimator of the average causal effect of attending a small class on student performance; we conclude that attending a small class, as opposed to a regular-size one, increased students' reading test scores by 7.21 points, on average
percentage point	unit of measurement of the arithmetic difference between two percentages	in the STAR dataset, the difference-in-means estimator for *graduated* is 87.35% − 86.65% = 0.7 p.p.; attending a small class is estimated to increase the proportion of students graduating from high school by about 1 percentage point, on average

continues on next page...

2.7.1 CONCEPTS AND NOTATION (CONTINUED)

concept/notation	description	example(s)
average outcome for the treatment group ($\overline{Y}_{\text{treatment group}}$)	average observed outcome for the individuals who received the treatment (after the treatment)	in the STAR dataset, the average reading score of the students who attended a small class was about 632.7 points
average outcome for the control group ($\overline{Y}_{\text{control group}}$)	average observed outcome for the individuals who did not receive the treatment (after no treatment)	in the STAR dataset, the average reading score of the students who attended a regular-size class was about 625.49 points
experimental data	data from a randomized experiment	since Project STAR was a randomized experiment, the data we analyze in this chapter are experimental data
observational data	data collected about naturally occurring events, in which treatment was received or not received without the intervention of researchers	data on class sizes and student performance from districts where the size of the classes varies as a result of factors such as school budgets, student enrollment, or the physical limitations of the school buildings
observational study	type of study that analyzes observational data	(see previous entry)

2.7.2 R SYMBOLS AND OPERATORS

code	description	example(s)
==	relational operator used to test whether the observations of a variable are equal to a particular value; values should be in quotes if text but without quotes if numbers (see "")	*data*$*variable*==1 *data*$*variable*=="yes"
$	character used to identify an element inside an object, such as a variable inside a dataframe, either to access it or to create it; to its left, we specify the name of the object where the dataframe is stored (without quotes); to its right, we specify the name of the element or variable (without quotes)	*data*$*variable* # identifies the variable named *variable* inside the dataframe stored in the object named *data*
[]	operator used to extract a selection of observations from a variable; to its left, we specify the variable we want to subset; inside the square brackets, we specify the criteria of selection; for example, we can specify a logical test using the relational operator ==; only the observations for which the logical test is true will be extracted	*data*$*var1*[*data*$*var2*==1] # extracts the observations of the variable *var1* for which the variable *var2* equals 1

2.7.3 R FUNCTIONS

function	description	required argument(s)	example(s)
ifelse()	creates the contents of a new variable based on the values of an existing one	three, separated by commas, in the following order: (1) logical test (see ==) (2) return value if test is true (3) return value if test is false values should be in quotes if text but without quotes if numbers (see "")	ifelse(*data*$*variable*=="yes", 1, 0) # returns a 1 whenever the observation of *variable* equals "yes" and a 0 otherwise, creating the contents of a binary variable using the existing character variable *variable*