1

Data and Case Studies

Statistics is the art and science of collecting and analyzing data and understanding the nature of variability. Mathematics, especially probability, governs the underlying theory, but statistics is driven by applications to real problems.

In this chapter, we introduce several data sets that we will encounter throughout the text in the examples and exercises.

1.1 Case Study: Flight Delays

If you have ever traveled by air, you probably have experienced the frustration of flight delays. The Bureau of Transportation Statistics maintains data on all aspects of air travel, including flight delays at departure and arrival (https://www.bts.gov/topics/airlines-and-airports/quick-links-popular-air-carrier-statistics).

LaGuardia Airport (LGA) is one of three major airports that serves the New York City metropolitan area. In 2008, over 23 million passengers and over 375 000 planes flew in or out of LGA. United Airlines and American Airlines are two major airlines that schedule services at LGA. The data set `FlightDelays` contains information on all 4029 departures of these two airlines from LGA during May and June 2009 (Tables 1.1 and 1.2).

Each row of the data set is an *observation*. Each column represents a *variable* – some characteristic that is obtained for each observation. For instance, on the first observation listed, the flight was a United Airlines plane, flight number 403, destined for Denver, and departing on Friday between 4 and 8 a.m. This data set consists of 4029 observations and 9 variables.

Questions we might ask include the following: Are flight delay times different between the two airlines? Are flight delay times different depending on the day of the week? Are flights scheduled in the morning less likely to be delayed by more than 15 min?

Mathematical Statistics with Resampling and R, Second Edition. Laura M. Chihara and Tim C. Hesterberg.

Table 1.1 Partial view of `FlightDelays` data.

Flight	Carrier	FlightNo	Destination	DepartTime	Day
1	UA	403	DEN	4–8 a.m.	Friday
2	UA	405	DEN	8–noon	Friday
3	UA	409	DEN	4–8 p.m.	Friday
4	UA	511	ORD	8–noon	Friday
		⋮			

Table 1.2 Variables in data set `FlightDelays`.

Variable	Description
Carrier	UA=United Airlines, AA=American Airlines
FlightNo	Flight number
Destination	Airport code
DepartTime	Scheduled departure time in 4 h intervals
Day	Day of week
Month	May or June
Delay	Minutes flight delayed (negative indicates early departure)
Delayed30	Departure delayed more than 30 min?
FlightLength	Length of time of flight (minutes)

1.2 Case Study: Birth Weights of Babies

The birth weight of a baby is of interest to health officials since many studies have shown possible links between this weight and conditions in later life, such as obesity or diabetes. Researchers look for possible relationships between the birth weight of a baby and the age of the mother or whether or not she smoked cigarettes or drank alcohol during her pregnancy. The Centers for Disease Control and Prevention (CDC) maintains a database on all babies born in a given year (http://wonder.cdc.gov/natality-current.html), incorporating data provided by the US Department of Health and Human Services, the National Center for Health Statistics, and the Division of Vital Statistics. We will investigate different samples taken from the CDC's database of births.

One data set that we will investigate consists of a random sample of 1009 babies born in North Carolina during 2004 (Table 1.3). The babies in the sample

Table 1.3 Variables in data set `NCBirths2004`.

Variable	Description
Age	Mother's age
Tobacco	Mother used tobacco?
Gender	Gender of baby
Weight	Weight at birth (grams)
Gestation	Gestation time (weeks)

had a gestation period of at least 37 weeks and were single births (i.e. not a twin or triplet).

In addition, we will also investigate a data set, `Girls2004`, consisting of a random sample of 40 baby girls born in Alaska and 40 baby girls born in Wyoming. These babies also had a gestation period of at least 37 weeks and were single births.

The data set `TXBirths2004` contains a random sample of 1587 babies born in Texas in 2004. In this case, the sample was not restricted to single births, nor to a gestation period of at least 37 weeks. The numeric variable `Number` indicates whether the baby was a single birth, or one of a twin, triplet, and so on. The variable `Multiple` is a factor variable indicating whether or not the baby was a multiple birth.

1.3 Case Study: Verizon Repair Times

Verizon is the primary local telephone company (incumbent local exchange carrier (ILEC)) for a large area of the Eastern United States. As such, it is responsible for providing repair service for the customers of other telephone companies known as competing local exchange carriers (CLECs) in this region. Verizon is subject to fines if the repair times (the time it takes to fix a problem) for CLEC customers are substantially worse than those for Verizon customers.

The data set `Verizon` contains a sample of repair times for 1664 ILEC and 23 CLEC customers (Table 1.4). The mean repair times are 8.4 h for ILEC

Table 1.4 Variables in data set `Verizon`.

Variable	Description
Time	Repair times (in hours)
Group	ILEC or CLEC

customers and 16.5 h for CLEC customers. Could a difference this large be easily explained by chance?

1.4 Case Study: Iowa Recidivism

When a person is released from prison, will he or she relapse into criminal behavior and be sent back? The state of Iowa tracks offenders over a 3-year period and records the number of days until recidivism for those who are readmitted to prison. The Department of Corrections uses this recidivism data to determine whether or not their strategies for preventing offenders from relapsing into criminal behavior are effective.

The data set `Recidivism` contains all offenders convicted of either a misdemeanor or felony who were released from an Iowa prison during the 2010 fiscal year (ending in June) (Table 1.5). There were 17 022 people released in that period, of whom 5386 were sent back to prison in the following 3 years (through the end of the 2013 fiscal year). [1]

The recidivism rate for those under the age of 25 years was 36.5% compared with 30.6% for those 25 years or older. Does this indicate a real difference in the behavior of those in these age groups, or could this be explained by chance variability?

Table 1.5 Variables in data set Iowa `Recidivism`.

Variable	Description
Gender	F, M
Race	American Indian or Alaska Native Hispanic, American Indian or Alaska Native Non-Hispanic, Asian or Pacific Islander Hispanic, Asian or Pacific Islander NonHispanic, Black, Black Hispanic, Black Non-Hispanic, White, White Hispanic, White Non-Hispanic
Age	Age at release: under 25, 25–34, 35–44, 45–54, and 55 and older
Age25	Under 25, over 25 (binary)
Offense	Original conviction: felony or misdemeanor
Recid	Recidivate? No, yes
Type	New (crime), no recidivism, tech (technical violation, such as a parole violation)
Days	Number of days to recidivism; NA if no recidivism

1 (https://data.iowa.gov/Public-Safety/3-Year-Recidivism-for-Offenders-Released-from-Pris/mw8r-vqy4).

1.5 Sampling

In analyzing data, we need to determine whether the data represent a *population* or a *sample*. A *population* represents all the individual cases, whether they are babies, fish, cars, or coin flips. The data from flight delays case study in Section 1.1 are *all* the flight departures of United Airlines and American Airlines out of LGA in May and June 2009; thus, this data set represents the population of all such flights. On the other hand, the North Carolina data set contains only a subset of 1009 births from over 100 000 births in North Carolina in 2004. In this case, we will want to know how representative statistics computed from this sample are for the entire population of North Carolina babies born in 2004.

Populations may be finite, such as births in 2004, or infinite, such as coin flips or births next year.

Throughout this book, we will talk about drawing random samples from a population. We will use capital letters (e.g. X, Y, Z, and so on) to denote random variables and lower-case letters (e.g. x_1, x_2, x_3, and so on) to denote actual values or data.

There are many kinds of random samples. Strictly speaking, a "random sample" is any sample obtained using a random procedure. However, in this book we use *random sample* to mean a sample of independent and identically distributed (i.i.d.) observations from the population, if the population is infinite.

For instance, suppose you toss a fair coin 20 times and consider each head a "success." Then your sample consists of the random variables $X_1, X_2, \ldots, X_{20}$, each a Bernoulli random variable with success probability 1/2. We use the notation $X_i \sim \text{Bern}(1/2)$, $i = 1, 2, \ldots, 20$.

If the population of interest is finite $\{x_1, x_2, \ldots, x_N\}$, we can choose a random sample as follows: Label N balls with the numbers $1, 2, \ldots, N$ and place them in an urn. Draw a ball at random, record its value $X_1 = x_{i_1}$, and then replace the ball. Draw another ball at random, record its value, $X_2 = x_{i_2}$, and replace. Continue until you have a sample $x_{i_1}, x_{i_2}, \ldots, x_{i_n}$. This is *sampling with replacement*. For instance, if $N = 5$ and $n = 2$, then there are $5 \times 5 = 25$ different samples of size 2 (where order matters). (Note: By "order matters" we do not imply that order matters in practice, rather we mean that we keep track of the order of the elements when enumerating samples. For instance, the set $\{a, b\}$ is different from $\{b, a\}$.)

However, in most real situations, for example, in conducting surveys, we do not want to have the same person polled twice. So we would sample *without replacement*, in which case, we will not have independence. For instance, if you wish to draw a sample of size $n = 2$ from a population of $N = 10$ people, then the probability of any one person being selected is 1/10. However, after having chosen that first person, the probability of any one of the remaining people being chosen is now 1/9.

In cases where populations are very large compared to the sample size, calculations under sampling without replacement are reasonably approximated by calculations under sampling with replacement.

Example 1.1 Consider a population of 1000 people, 350 of whom are smokers, and the rest are nonsmokers. If you select 10 people at random but with replacement, then the probability that 4 are smokers is $\binom{10}{4}(350/1000)^4(650/1000)^6 \approx 0.2377$. If you select without replacement, then the probability is $\binom{350}{4}\binom{650}{6}/\binom{1000}{10} \approx 0.2388$. □

1.6 Parameters and Statistics

When discussing numeric information, we will want to distinguish between populations and samples.

Definition 1.1 A *parameter* is a (numerical) characteristic of a population or of a probability distribution.

A *statistic* is a (numerical) characteristic of data. ‖

Any function of a parameter is also a parameter; any function of a statistic is also a statistic. When the statistic is computed from a random sample, it is itself random, and hence is a random variable.

Example 1.2 μ and σ are parameters of the normal distribution with pdf $f(x) = (1/\sqrt{2\pi}\sigma)\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}$.

The variance σ^2 and *signal-to-noise ratio* μ/σ are also parameters. □

Example 1.3 If $X_1, X_2, \ldots, X_n$ are a random sample, then the mean $\overline{X} = 1/n \sum_{i=1}^{n} X_i$ is a statistic. □

Example 1.4 Consider the population of all babies born in the United States in 2017. Let μ denote the average weight of all these babies. Then μ is a parameter. The average weight of a sample of 2500 babies born in that year is a statistic. □

Example 1.5 If we consider the population of all adults in the United States today, the proportion p who approve of the president's job performance is a parameter. The fraction $\hat{p}$ who approve in any given sample is a statistic. □

Example 1.6 The average weight of 1009 babies in the North Carolina case study in Section 1.2 is 3448.26 g. This average is a statistic. □

Example 1.7 If we survey 1000 adults and find that 60% intend to vote in the next presidential election, then $\hat{p} = 0.60$ is a statistic: It estimates the parameter p, the proportion of all adults who intend to vote in the next election. □

1.7 Case Study: General Social Survey

The General Social Survey (GSS) is a major survey that has tracked American demographics, characteristics, and views on social and cultural issues since the 1970s. It is conducted by the National Opinion Research Center (NORC) at the University of Chicago. Trained interviewers meet face to face with the adults chosen for the survey and question them for about 90 min in their homes.

The GSS case study includes the responses of 2765 participants selected in 2002 to about a dozen questions, listed in Table 1.6. For example, one of the questions (`SpendEduc`) asked whether the respondent believed that the

Table 1.6 Variables in data set `GSS2002`.

Variable	Description
Region	Interview location
Gender	Gender of respondent
Race	Race of respondent: White, Black, Other
Marital	Marital status
Education	Highest level of education
Happy	General happiness
Income	Respondent's income
PolParty	Political party
Politics	Political views
Marijuana	Legalize marijuana?
DeathPenalty	Death penalty for murder?
OwnGun	Have gun at home?
GunLaw	Require permit to buy a gun?
SpendMilitary	Amount government spends on military
SpendEduc	Amount government spends on education
SpendEnv	Amount government spends on the environment
SpendSci	Amount government spends on science
Pres00	Whom did you vote for in the 2000 presidential election?
Postlife	Believe in life after death?

amount of money being spent on the nation's education system was too little, too much, or the right amount.

We will analyze the GSS data to investigate questions such as the following: Is there a relationship between the gender of an individual and whom they voted for in the 2000 presidential election? Are people who live in certain regions happier? Are there educational differences in support for the death penalty? These data are archived at the Computer-assisted Survey Methods Program at the University of California (www.sda.berkeley.edu).

1.8 Sample Surveys

"Who do you plan to vote for in the next presidential election?" "Would you purchase our product again in the future?" "Do you smoke cigarettes? If yes, how old were you when you first started?" Questions such as these are typical of sample surveys. Researchers want to know something about a population of individuals, whether they are registered voters, online shoppers, or American teenagers, but to poll every individual in the population – that is, to take a *census* – is impractical and costly. Thus, researchers will settle for a sample from the target population. But if, say, 60% of those in your sample of 1000 adults intend to vote for candidate Wong in the next election, how close is this to the actual percentage who will vote for Wong? How can we be sure that this sample is truly representative of the population of all voters? We will learn techniques for *statistical inference*, drawing a conclusion about a population based on information about a sample.

When conducting a survey, researchers will start with a *sampling frame* – a list from which the researchers will choose their sample. For example, to survey all students at a college, the campus directory listing could be a sampling frame. For pre-election surveys, many polling organizations use a sampling frame of registered voters. Note that the choice of sampling frame could introduce the problem of *undercoverage*: omitting people from the target population in the survey. For instance, young people were missed in many pre-election surveys during the 2008 Obama–McCain presidential race because they had not yet registered to vote.

Once the researchers have a sampling frame, they will then draw a random sample from this frame. Researchers will use some type of *probability (scientific) sampling scheme*, that is, a scheme that gives everybody in the population a positive chance of being selected. For example, to obtain a sample of size 10 from a population of 100 individuals, write each person's name on a slip of paper, put the slips of paper into a basket, and then draw out 10 slips of paper. Nowadays, statistical software is used to draw random samples from a sampling frame.

Another basic survey design uses *stratified sampling*: The population is divided into nonoverlapping strata, and then random samples are drawn from each stratum. The idea is to group individuals who are similar in some characteristic into homogeneous groups, thus reducing variability. For instance, in a survey of university students, a researcher might divide the students by class: first year, sophomores, juniors, seniors, and graduate students. A market analyst for an electronics store might choose to stratify customers based on income levels.

In *cluster sampling*, the population is divided into nonoverlapping clusters, and then a random sample of clusters is drawn. Every person in a chosen cluster is then interviewed for the survey. An airport wanting to conduct a customer satisfaction survey might use a sampling frame of all flights scheduled to depart from the airport on a certain day. A random sample of flights (clusters) is chosen, and then all passengers on these flights are surveyed. A modification of this design might involve sampling in stages: For instance, the analysts might first choose a random sample of flights, and then from each flight choose a random sample of passengers.

The GSS uses a more complex sampling scheme in which the sampling frame is a list of counties and county equivalents (standard metropolitan statistical areas) in the United States. These counties are stratified by region, age, and race. Once a sample of counties is obtained, a sample of block groups and enumeration districts is selected, stratifying these by race and income. The next stage is to randomly select blocks and then interview a specific number of men and women who live within these blocks.

Indeed, all major polling organizations such as Gallup or Roper as well as the GSS use a *multistage* sampling design. In this book, we use the GSS data or polling results for examples as if the survey design used simple random sampling. Calculations for more complex sampling scheme are beyond the scope of this book, and we refer the interested reader to Lohr (1991) for details.

1.9 Case Study: Beer and Hot Wings

Carleton student Nicki Catchpole conducted a study of hot wings and beer consumption at the Williams Bar in the Uptown area of Minneapolis (N. Catchpole, private communication). She asked patrons at the bar to record their consumption of hot wings and beer over the course of several hours. She wanted to know if people who ate more hot wings would then drink more beer. In addition, she investigated whether or not gender had an impact on hot wings or beer consumption.

The data for this study are in `Beerwings` (Table 1.7). There are 30 observations and 3 variables.

Table 1.7 Variables in data set `Beerwings`.

Variable	Description
Gender	Male or female
Beer	Ounces of beer consumed
Hot Wings	Number of hot wings eaten

1.10 Case Study: Black Spruce Seedlings

Black spruce (*Picea mariana*) is a species of a slow-growing coniferous tree found across the northern part of North America. It is commonly found on wet organic soils. In a study conducted in the 1990s, a biologist interested in factors affecting the growth of the black spruce planted its seedlings on sites located in boreal peatlands in northern Manitoba, Canada (Camil et al. (2010)).

The data set `Spruce` contains a part of the data from the study (Table 1.8). Seventy-two black spruce seedlings were planted in four plots under varying conditions (fertilizer–no fertilizer, competition–no competition), and their heights and diameters were measured over the course of 5 years.

The researcher wanted to see whether the addition of fertilizer or the removal of competition from other plants (by weeding) affected the growth of these seedlings.

1.11 Studies

Researchers carry out studies to understand the conditions and causes of certain outcomes: Does smoking cause lung cancer? Do teenagers who smoke marijuana tend to move on to harder drugs? Do males eat more hot wings than females? Do black spruce seedlings grow taller in fertilized plots?

Table 1.8 Variables in data set `Spruce`.

Variable	Description
Tree	Tree number
Competition	C (competition), CR (competition removed)
Fertilizer	F (fertilized), NF (not fertilized)
Height0	Height (cm) of seedling at planting
Height5	Height (cm) of seedling at year 5
Diameter0	Diameter (cm) of seedling at planting
Diameter5	Diameter (cm) of seedling at year 5
Ht.change	Change (cm) in height
Di.change	Change (cm) in diameter

The beer and hot wings case study in Section 1.9 is an example of an *observational study*, a study in which researchers observe participants but do not influence the outcome. In this case, the student just recorded the number of hot wings eaten and beer consumed by the patrons of Williams Bar.

Example 1.8 The first Nurses' Health Study is a major observational study funded by the National Institutes of Health. Over 12 0000 registered female nurses who, in 1976, were married, between the ages of 33 and 55 years, and who lived in the 11 most populous states have been responding every 2 years to written questions about their health and lifestyle, including smoking habits, hormone use, and menopause status. Many results on women's health have come out of this study, such as finding an association between taking estrogen after menopause and lowering the risk of heart disease, and determining that for nonsmokers there is no link between taking birth control pills and developing heart disease.

Because this is an observational study, no *cause-and-effect* conclusions can be drawn. For instance, we cannot state that taking estrogen after menopause will *cause* a lowering of the risk for heart disease. In an observational study, there may be many unrecorded or hidden factors that impact the outcomes. Also, because the participants in this study are registered nurses, we need to be careful about making inferences about the general female population. Nurses are more educated and more aware of health issues than the average person. □

On the other hand, the black spruce case study in Section 1.10 was an *experiment*. In an experiment, researchers will manipulate the environment in some way to observe the response of the objects of interest (people, mice, ball bearings, etc.). When the objects of interest in an experiment are people, we refer to them as *subjects*; otherwise, we call them *experimental units*. In this case, the biologist randomly assigned the experimental units – the seedlings – to plots subject to four *treatments*: fertilization with competition, fertilization without competition, no fertilization with competition, and no fertilization with no competition. He then recorded their height over a period of several years.

A key feature in this experiment was the *random assignment* of the seedlings to the treatments. The idea is to spread out the effects of unknown or uncontrollable factors that might introduce unwanted variability into the results. For instance, if the biologist had planted all the seedlings obtained from one particular nursery in the fertilized, no competition plot and subsequently recorded that these seedlings grew the least, then he would not be able to discern whether this was due to this particular treatment or due to some possible problem with seedlings from this nursery. With random assignment of treatments, the seedlings from this particular nursery would usually be spread out over the four treatments. Thus, the differences between the treatment groups should be due to the treatments (or chance).

Example 1.9 Knee osteoarthritis (OA) that results in deterioration of cartilage in the joint is a common source of pain and disability for the elderly population. In a 2008 paper, "Tai Chi is effective in treating knee osteoarthritis: A randomized controlled trial," Wang et al. (2009) at Tufts University Medical School describe an experiment they conducted to see whether practicing tai chi, a style of Chinese martial arts, could alleviate pain from OA. Forty patients over the age of 65 with confirmed knee OA but otherwise in good health were recruited from the Boston area. Twenty were randomly assigned to attend twice weekly 60 min sessions of tai chi for 12 weeks. The remaining 20 participants, the *control group*, attended twice weekly 60 min sessions of instructions on health and nutrition, as well as some stretching exercises.

At the end of the 12 weeks, those in the tai chi group reported a significant decrease in knee pain. Because the subjects were randomly assigned to the two treatments, the researchers can assert that the tai chi sessions lead to decrease in knee pain due to OA. Note that because the subjects were recruited, we need to be careful about making an inference about the general elderly population: People who voluntarily sign up to be in an experiment may be different from other people. □

Another important feature of a well-designed experiment is *blinding*: A *double-blind* experiment is one in which neither the researcher nor the subject knows who is receiving which treatment. An experiment is *single-blinded* if just the researcher or the subject (but not both) knows who is receiving which treatment. Blinding is important in reducing *bias*, the systematic favoring of one outcome over another.

For instance, suppose in a clinical trial to test the efficacy of a new drug for a disease, the subjects know whether they are receiving the drug or a placebo (a pill or drug with no therapeutic effect). Those on the placebo might feel that the trial is a waste of time and drop out, or perhaps seek additional treatment elsewhere. On the other hand, if the researcher knows that a subject received the drug, he or she might behave differently toward the subject, perhaps by asking leading questions that result in responses that appear to suggest relief from the disease.

1.12 Google Interview Question: Mobile Ads Optimization

The following question was posted on an internal Google statistics email list:

> I have a pre v post comparison I'm trying to make where alternative hypothesis is pre.mean.error > post.mean.error. My distribution for these samples are both right skewed as shown below. Anyone know what test method would be best suited for this type of situation?

When I [Tim] interview candidates for a quantitative analyst position at Google, I frequently ask applicants to "Imagine that you are consulting with this person. What would you ask or tell this person?"

Many applicants start off on the wrong track by proposing various statistical tests without first understanding the problem or data. Data analysis and statistics are ultimately about solving problems, not just applying techniques, so we need to begin by asking questions to gain insight about the data, the context, and the problem itself.

Important questions include: What is CPR error? What does pre versus post mean? How many observations are there? How were the data collected? How are the data related? Also, make sure to understand the Figure 1.1.

In this data set, the pre and post variables are paired, not independent – and an answer that does not take that into account is wrong.

It is the duty of the consultant to not only answer the questions that the client posed but also to think about whether those are the right questions to ask. In this case the client asked how to compare means (averages), but there are outliers, so comparing averages is inaccurate; we should consider other comparisons.[2] However, once we learn more about the data, it turns out that answering a slightly different question is probably better in both practical terms (does a better job of measuring what really matters) and statistical terms (more accurate, better signal-to-noise ratio). For this modified problem the outliers are much less of an issue; we discuss this in Section 8.7.

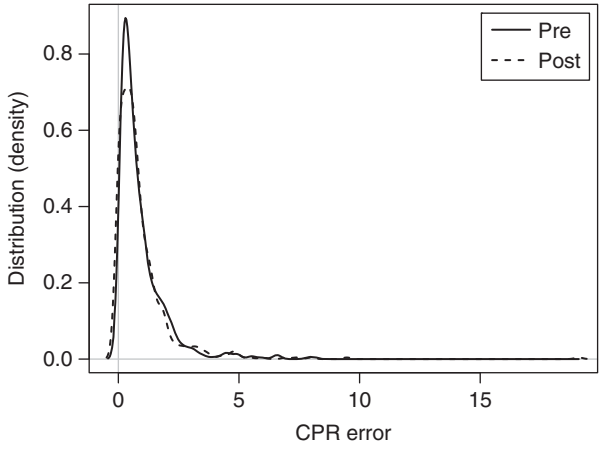

Figure 1.1 Density plot for CPR error.

2 Do not worry if you do not understand the statistical terms in this section. We will discuss them in the chapters ahead!

Once a consultant understands the data, he/she is in a position to suggest appropriate methods. In this book, we will provide guidelines for determining methods to handle different data scenarios.

Here is some more background.[3] *Google Adwords*™ lets advertisers bid for ads to appear when people search. An auction determines which ads are shown, based on a combination of the bid and Google's estimate of how interested the searcher would be in an ad. If an ad places high in the auction, it is shown to the user; if the user clicks on the ad, then the advertiser pays Google.

Within Adwords, *Enhanced Campaigns* offered advertisers the ability to customize their bidding, e.g. to show one ad to someone searching for "pizza" at 1 p.m. on their PC at work (perhaps a link to an online order form or menu) and a different ad to someone searching for pizza at 8 p.m. on a smartphone a half-mile from the restaurant (perhaps a click-to-call phone number and restaurant locator). When this was launched, many advertisers did not understand how to bid appropriately on mobile phones. Google crafted an experiment to help advertisers bid appropriately. The experiment was based on equalizing return on ad investment between desktop and mobile, because $1 revenue on mobile is the same benefit to advertisers as $1 revenue on desktop. If the return on investment (ratio of value from user purchases to the cost of advertising to the users) was higher on mobile than desktop in the "preperiod" (before the experiment), Google would recommend increasing the "mobile multiplier," the ratio between the mobile and desktop bid; this would result in more ads to mobile customers with an increased average cost, and fewer to desktop customers with a decreased average cost, and tend to make the return on investment more equal. For example, if the mobile multiplier for a particular campaign was 1.2 before the recommendation, Google might recommend increasing it to 1.4.

As a result of Google's recommendations, advertisers raised their bids in some cases and lowered them in others, resulting in both a lower cost for advertising and greater return.

Why would Google do this, if it made less money? Two reasons – advertisers and users. In the long run, advertisers will use Google ads more if they get more bang for their buck. And it is better for Google's users if advertisers can better advertise to people interested in their product and avoid advertising to people who are not.

However, this was not a pure randomized experiment where some advertisers were given recommendations and others not. Advertisers had to agree to participate and report the number of "conversions" (purchases) and the value of those conversions. The comparison is between before-and-after results, but some advertisers might have adjusted their bids even without the recommendation. They also were not obligated to follow the recommendations.

3 I give a simpler explanation during interviews – once people ask!

The experiment was designed as a pre versus post paired t test to compare results before and after the recommendations. However, the distributions of the data shown in the figure above are very long tailed making t tests questionable.

The data `MobileAds` are a subset of the experimental data, for one advertiser. Each row corresponds to a single combination of campaign and ad group: These could be for different products, a different set of ads, target a different population, be shown for different searches, etc. Important variables are given in Table 1.9.

Most variables have two versions:

- Before the experiment (with a `_pre` suffix).
- During the experiment (with a `_post` suffix).

And two platforms:

- Mobile (with an `m.` prefix).
- Desktop+tablet (with a `d.` prefix).

Table 1.9 Variables in data set `MobileAds`.

Variable	Description
impr	Number of ad impressions (ads shown)
click	Number of clicks
cost	What advertisers paid
conv	Number of conversions (purchases)
value	Value of conversions as reported by advertisers
cpm	Cost per impression (cost/impr)
cpc	Cost per click (cost/click)
cpa	Cost per conversion (cost/conv)
	(or 0, if conv is 0)
cpr	Cost per return (cost/value)
	(or 0 if value is 0)
	Prefix indicates platform
m.*	Mobile, e.g. m.impr
d.*	Desktop/tablet, e.g. d.impr
	Suffix indicates when
*_pre	Before experiment, e.g. m.impr_pre
*_post	In experiment, e.g. m.impr_post
error.cpr*	m.cpr− d.cpr (pre, post)
mult.change	Change in mobile multiplier

For example, in the first row of the data set, there are 155 impressions on mobile in the pre period and 255 in the post period, and 1430 and 1466 on desktop or tablet during the two periods.

`error` is the difference in `cpr` (the reciprocal of return on investment) between mobile and desktop; small values indicate parity and suggest efficient bidding. The analyst was interested in whether the experiment would result in reductions in `error`.

`mult.change` is the change in mobile multiplier; a negative number indicates a lower mobile multiplier in the post period. (This is what the advertiser actually did, not what Google recommended.)

Exercises

1.1 For each of the following, describe the population and, if relevant, the sample. For each number presented, determine if it is a parameter or a statistic (or something else).

a) A survey of 1500 high school students finds that 47% watch the cable show "Game of Thrones."
b) The 2010 US Census reports that 9.6% of the US population was between the ages of 18 and 24 years.
c) Based on the rosters of all National Basketball Association teams for the 2006–2007 season, the average height of the players was 78.93 in.
d) A March 2016 Harris poll consisting of 2106 national adults, age 18 years or older, found that 19% strongly or somewhat disagree with the statement that the US has come a long way toward reaching gender equality.

1.2 Review the description of the Iowa recidivism case study in Section 1.4.

a) Does this data represent a population or a sample?
b) In this data set, 19.4% of the offenders were originally convicted of a misdemeanor. Does this number represent a parameter or a statistic?

1.3 Researchers reported that strict rest after a concussion did not improve the outcome of patients between the age of 11 and 22 years (Thomas et al. (2015)). Eight-eight patients who went to a pediatric emergency department in Wisconsin within 24 h of a concussion were recruited for this study. They were randomly assigned to either strict rest for 5 days or the usual care of 1–2 days of rest followed by stepwise return to activity. Participants self-reported post-concussion symptoms in a diary. The patients who were assigned to strict rest for 5 days reported more daily symptoms and slower symptom resolution than those assigned to usual care.

a) In this experiment, what were the treatments?

b) Was this a double-blind study?
c) Can the researchers conclude that strict rest for 5 days causes more daily symptoms?
d) Can we generalize these results to a population?

1.4 The journal *Molecular Psychiatry* reported on a study claiming that playing the video game Tetris reduces the formation of bad memories after a traumatic event (Iyadurai et al. (2017)). Seventy-one patients who were involved in a motor vehicle accident and admitted to a British emergency room were recruited. After completing some baseline assessments, they were randomly assigned to either play Tetris for at least 10 uninterrupted minutes or fill out a simple log detailing their activities while waiting in the emergency room. The patients who played Tetris reported having fewer intrusive memories about their accident than the patients who had completed a log.
a) In this experiment, what were the treatments?
b) Was this a double-blind study?
c) Can the researchers conclude that playing Tetris causes the reduction of painful memories?
d) Can we generalize these results to a population?

1.5 Researchers reported that moderate drinking of alcohol was associated with a lower risk of dementia (Mukamal et al. (2003)). Their sample consisted of 373 people with dementia and 373 people without dementia. Participants were asked how much beer, wine, or shot of liquor they consumed. Those who consumed 1–6 drinks a week had a lower incidence of dementia than those who abstained from alcohol.
a) Was this study an observational study or an experiment?
b) Can the researchers conclude that drinking alcohol causes a lower risk of dementia? Why or why not?

1.6 Researchers surveyed 959 ninth graders who attended 3 large US urban high schools and found that those who listened to music that had references to marijuana were almost twice as likely to have used marijuana as those who did not listen to music with references to marijuana (Primack et al. (2010)).
a) Was this an observational study or an experiment?
b) Can the researchers conclude that listening to music with references to marijuana causes students to use drugs?
c) Can the researchers extend their results to all urban American adolescents?

1.7 Duke University researchers found that diets low in carbohydrates are effective in controlling blood sugar levels (Westman et al. (2008)). Eighty-four volunteers with obesity and type 2 diabetes were randomly assigned to either a diet of less than 20 g of carbohydrates/day or a low-glycemic, reduced calorie diet (500 calories/day). 95% of those on the low-carbohydrate diet were able to reduce or eliminate their diabetes medications compared to 62% on the low-glycemic diet.

a) Was this study an observational study or an experiment?
b) Can researchers conclude that a low-carbohydrate diet causes an improvement in type 2 diabetes?
c) Can researchers extend their results to a more general population? Explain.

1.8 In the Google mobile ads case study (Section 1.12),

a) Why is this study described as an experiment and not an observational study?
b) Can Google claim that their recommendations "caused" the outcome of the study?
c) Can Google generalize their results to all advertisers who advertise on Google?

1.9 In a population of size N, the probability of any subset of size n being chosen is $1/\binom{N}{n}$. Show this implies that any one person in the population has a n/N probability of being chosen in a sample. Then, in particular, every person in the population has the same probability of being chosen.

1.10 A typical Gallup poll surveys about $n = 1000$ adults. Suppose the sampling frame contains 100 million adults (including you). Now, select a random sample of 1000 adults.

a) What is the probability that you will be in this sample?
b) Now suppose that 2000 such samples are selected, each independently of the others. What is the probability that you will *not* be in any of the samples?
c) How many samples must be selected for you to have a 0.5 probability of being in at least one sample?

1.11 In the mobile ads case study (Section 1.12), the variables `m.cpr` and `d.cpr`, which measure cost/value (how much it costs a company to advertise per how much they make), are recorded as 0 if value is 0. The error is defined by `error = m.cpr-d.cpr`. If `m.cpr=10` and `d.cpr=1`, then `error` is 9. However, if on the desktop mobile, `d.value` is 0, that is, the company did not make any money, then

`d.cpr` is defined to be 0. So the error is −1 which is smaller in absolute value than the first case.

a) Do you think that this accurately reflects the magnitude of the difference in these two scenarios?
b) If you were a consultant for Google, can you recommend other ways of defining `cpr` when the denominator `value` is 0?