Uvod u statistiku

Djelo izvedeno iz
OpenIntro Statistics
Fourth Edition

Izvorni autori

David M Diez
Mine Çetinkaya-Rundel
Christopher D Barr

Prijevod

Diana Šimić
[Prijevod na hrvatski jezik]

Sadržaj

Predgovor

Udžbenik Uvod u statistiku pokriva gradivo prvog predmeta iz statistike, te pruža jasan, koncizan i pristupačan matematički utemeljen uvod u primijenjenu statistiku. Udžbenik je namijenjen preddiplomskim studentima, ali se često koristi i u nastavi na srednješkolskoj i diplomskoj razini.

Nadamo se da će čitaoci, uz to što će usvojiti temelje statističkog razmišljanja i metodologije, prihvatiti tri ideje.

- Statistika je primijenjena znanost sa širokim područjem praktične primjene.
- Ne morate biti matematički guru da biste nešto naučili iz stvarnih, zanimljivih podataka.
- Podaci su neuredni, statistički alati su nesavršeni. Ali, kada razumijete snage i slabosti ovih alata, možete ih koristiti da učite o svijetu.

Pregled sadržaja udžbenika

Udžbenik sadrži sljedeća poglavlja:

1. **Uvod u podatke.** Strukture podataka, varijable i osnovne tehnike prikupljanja podataka.
2. **Opisivanje podataka.** Opisivanje podataka, grafički prikazi (vizualizacija), i uvod u zaključivanje koristeći randomizaciju.
3. **Vjerojatnost.** Osnovna načela teorije vjerojatnosti.
4. **Razdiobe slučajnih varijabli.** Model normalne razdiobe i druge ključne razdiobe.
5. **Osnove statističkog zaključivanja.** Osnovne ideje statističkog zaključivanja u kontekstu procjene populacijske proporcije.
6. **Zaključivanje o kvalitativnim podacima.** Zaključivanje o proporcijama i tablicama primjenom normalne i hi-kvadrat razdiobe.
7. **Zaključivanje o kvantitativnim podacima.** Zaključivanje o aritmetičkim sredinama na jednom ili dva uzorka primjenom Studentove t-razdiobe, statistička snaga za usporedbu dvije grupe i usporedba više aritmetičkih sredina primjenom analize varijance.
8. **Uvod u linearnu regresiju.** Regresija za kvantitaivni ishod s jednom nezavisnom varijablom. Većinu ovog poglavlja moguće je obraditi nakon poglavlja 1.
9. **Multipla i logistička regresija.** Regresija za kvantitativne i kvalitativne podatke uz korištenje više nezavisnih varijabli.

Uvod u statistiku podržava fleksibilnost u izboru i redoslijedu tema. Ako je glavni cilj doći do multiple regresije (poglavlje **??**) što je brže moguće, ovo su idealni preduvjeti:

- Poglavlje 1, odjeljak **??**, i odjeljak **??** za solidan uvod u strukture podataka i deskriptivne statistike koje se koriste u cijelom udžbeniku.
- Odjeljak **??** za dobro razumijevanje normalne razdiobe.
- Poglavlje **??** za razumijevanje osnovnih alata za statističko zaključivanje.
- Odjeljak **??** za osnove Studentove t-razdiobe
- Poglavlje **??** za razumijevanje ideja i načela jednostavne regresije (s jednim prediktorom).

Primjeri i vježbe

Primjeri pomažu stjecanju razumijevanja kako primijeniti metode.

PRIMJER 0.1

Ovo je primjer. Kada je ovdje postavljeno pitanje, gdje se može naći odgovor?

Odgovor je ovdje, u odlomku s rješenjima primjera!

Kada mislimo da bi čitalac trebao biti spreman pokušati riješiti primjer, takav primjer navodimo kao Vođenu vježbu.

VOĐENA VJEŽBA 0.2

Čitalac može provjeriti rješenje ili naučiti kako riješiti problem iz Vođene vježbe tako da pogleda puno rješenje u fusnoti.[1]

Vježbe se nalaze i na kraju svakog poglavlja, a vježbe za ponavljanje na kraju svake glave. Rješenja neparnih vježbi nalaze se u prilogu A.

Dodatni resursi

Video snimke, prezentacije, laboratorijske vježbe sa statističkim sotverom, skupovi podataka korišteni u udžbeniku i mnogi drugi resursi (na engleskom jeziku, op.prev.) dostupni su na stranicama

openintro.org/os

Pristup podacima korištenim u ovom udžbeniku olakšava i prilog B, koji pruža dodatne informacije o svakom skupu podataka koji se koristi u osnovnom tekstu udžbenika, što je novi sadržaj u četvrtom izdanju. Za svaki od navedenih skupova podataka postoje i online smjernice (na engleskom jeziku, op.prev.) na stranici **openintro.org/data** i u okviru pratećeg R paketa .

Zahvale

Ovaj projekt ne bi bio moguć bez strasti i predanosti velikog broja osoba koje nisu navedene kao autori. Autori žele zahvaliti osoblju OpenIntro Staff na suradnji i kontinuiranom doprinosu. Zahvalni smo i stotinama studenata i nastavnika koji su nam pružili vrijedne povratne informacije od kada smo prvi puta objavili sadržaj udžbenika 2009. godine.

Želimo zahvaliti i mnogim učiteljima koji su pomagali u recenziji ovog izdanja, među kojima su Laura Acion, Matthew E. Aiello-Lammens, Jonathan Akin, Stacey C. Behrensmeyer, Juan Gomez, Jo Hardin, Nicholas Horton, Danish Khan, Peter H.M. Klaren, Jesse Mostipak, Jon C. New, Mario Orsi, Steve Phelps, i David Rockoff. Cijenimo sve njihove komentare, koji su nam pomogli bolje prilagoditi tekst i tako znatno unaprijediti ovaj udžbenik.

[1] Svrha Vođenih vježbi je da vas natjeraju na promišljanje, a točnost rješenja možete provjeriti u fusnoti Vođene vježbe.

Glava 1

Uvod u podatke

Znanstvenici traže odgovore na pitanja koristeći rigorozne metode i pažljiva opažanja. Ova opažanja – prikupljena npr. u obliku bilješki s terenskih istraživanja, popunjenih upitnika ili rezultata eksperimenata – čine okosnicu statističkih istraživanja i zovemo ih **podaci**. Statistika je znanost koja se bavi proučavanjem kako najbolje prikupiti i analizirati podatke te na temelju njih izvoditi zaključke. U ovoj prvoj glavi bavimo se svojstvima i načinima prikupljanja podataka.

Video zapise, prezentacije i druge resurse na engleskom jeziku možete naći na internetskim stranicama
www.openintro.org/os

1.1 Primjer: korištenje stenta za sprječavanje srčanog udara

Poglavlje 1.1 upoznaje nas s klasičnim izazovom za statistiku: kako vrednovati učinkovitost medicinskog postupka. Pojmove koje spominjemo u ovom poglavlju i cijeloj glavi ćemo ponavljati i kasnije u tekstu. Za sada nam je cilj steći osjećaj za ulogu statistike u primjeni.

U ovom poglavlju ćemo razmatrati eksperiment koji proučava učinkovitost stenta u liječenju pacijenata u riziku od srčanog udara. Stent je mrežica koja se stavlja u krvnu žilu kako bi pomogla u oporavku pacijenata nakon srčanog udara i smanjila rizik ponovljenog srčanog udara ili smrti. Mnogi se liječnici nadaju da bi sličnu dobrobit mogli imati i kod pacijenata u riziku za srčanu bolest. Započinjemo postavljanjem glavnog pitanja na koje istraživači žele odgovoriti:

Da li primjena stenta smanjuje rizik srčanog udara?

Istraživači koji su postavili ovo pitanje proveli su eksperiment sa 451 pacijentom u riziku. Svaki pacijent dobrovoljac bio je slučajno raspoređen u jednu od dvije grupe:

Izložena grupa. Pacijentima u izloženoj grupi ugrađen je stent i pružena medicinska obrada. Medicinska obrada uključivala je lijekove, upravljanje čimbenicima rizika i potporu za promjenu načina života.

Kontrolna grupa. Pacijentima u kontrolnoj grupi pružena je jednaka medicinska obrada, ali im nije ugrađen stent.

Istraživači su slučajno rasporedili 224 pacijenta u izloženu grupu i 227 pacijenata u kontrolnu grupu. U ovom primjeru kontrolna grupa pruža referentnu točku s kojom uspoređujemo medicinski učinak stenta u izloženoj grupi.

Istraživači su proučavali učinak stenta u dvjema vremenskim točkama: 30 dana nakon uključenja u istraživanje i 365 dana nakon uključenja u istraživanje. Rezultati 5 pacijenata prikazani su na Slici 1.1. Ishodi pacijenata zabilježeni su kao "udar" ili "bez udara", što predstavlja informaciju da li je pacijent na kraju razdoblja promatranja imao srčani udar.

Pacijent	grupa	0-30 dana	0-365 dana
1	izložena	bez udara	bez udara
2	izložena	udar	udar
3	izložena	bez udara	bez udara
⋮	⋮	⋮	
450	kontrolna	bez udara	bez udara
451	kontrolna	bez udara	bez udara

Slika 1.1: Rezultati pet pacijenata iz istraživanja stentova.

Proučavanje podataka svakog pojedinog pacijenta bilo bi dugotrajan i težak način traženja odgovora na postavljeno istraživačko pitanje. Umjesto toga, statistička analiza omogućava proučavanje svih podataka odjednom. Slika 1.2 prikazuje izvorne podatke na korisniji način. Iz ove tablice možemo brzo vidjeti što se događalo tijekom cijelog istraživanja. Npr. da bismo odredili broj pacijenata u izloženoj grupi koji su imali srčani udar unutar 30 dana, pogledamo u lijevi dio tablice na križanju izložene grupe i srčanog udara: 33.

	0-30 dana		0-365 dana	
	udar	bez udara	udar	bez udara
izložena	33	191	45	179
kontrolna	13	214	28	199
Ukupno	46	405	73	378

Slika 1.2: Deskriptivna statistika za istraživanje stentova.

VOĐENA VJEŽBA 1.1

Od 224 pacijenta u izloženoj grupi, 45 ih je imalo srčani udar do kraja prve godine. Koristeći ova dva broja, izračunajte proporciju pacijenata u izloženoj grupi koji su imali srčani udar do kraja prve godine. (Bilješka: odgovore na sve vođene vježbe možete naći u fusnotama.)[1]

Na temelju podataka u tablici možemo izračunati deskriptivne statistike. **Deskriptivna statistika** je broj koji sažeto prikazuje veliku količinu podataka. Npr. primarni rezultati istraživanja nakon 1 godine mogu se opisati s dvije deskriptivne statistike: proporcije osoba koje su doživjele srčani udar u izloženoj i kontrolnoj grupi.

Proporcija osoba koje su doživjele srčani udar u izloženoj (stent) grupi: $45/224 = 0.20 = 20\%$.

Proporcija osoba koje su doživjele srčani udar u kontrolnoj grupi: $28/227 = 0.12 = 12\%$.

Ove dvije deskriptivne statistike korisne su kada želimo opisati razliku između grupa i u ovom slučaju doživjeli smo iznenađenje: dodatnih 8% pacijenata u izloženoj grupi je doživjelo srčani udar! Ovo je važno zbog dva razloga. Prvo, suprotno je od onoga što su liječnici očekivali, a to je da će stentovi *smanjiti* stopu srčanih udara. Drugo: vodi nas do statističkog pitanja: da li podaci pokazuju "stvarnu" razliku između grupa?

Ovo drugo pitanje je istančano. Pretpostavimo da 100 puta bacamo novčić. Iako je šansa da će u bilo kojem bacanju pasti glava 50%, vjerojatno neće točno 50 puta pasti glava. Ova vrsta odstupanja dio je gotovo svakog procesa stavaranja podataka. Možda je ovih 8% razlike u istraživanju stentova rezultat takve prirodne varijacije. Međutim, što je veća opažena razlika (za istu veličinu uzorka), teže ćemo povjerovati da je razlika posljedica slučaja. To znači da je pravo pitanje da li je razlika dovoljno velika da možemo odbaciti ideju da je posljedica slučaja?

Iako još nemamo na raspolaganju statističke alate da samostalno nađemo odgovor na postavljeno pitanje, možemo razumjeti zaključak objavljene analize: postoje uvjerljivi argumenti da su stentovi u ovom istraživanju štetni za pacijente u riziku za srčani udar.

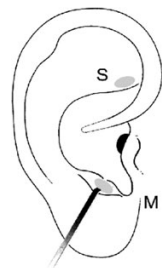
Oprez: Nemojte generalizirati rezultate ovog istraživanja na sve pacijente i sve stentove. U ovo istraživanje uključeni su pacijenti vrlo specifičnih svojstava koji su se dobrovoljno javili da sudjeluju u istraživanju i koji ne moraju biti reprezentativni za sve pacijente u riziku za srčani udar. Dodatno, postoji više vrsta stentova, a u ovom istraživanju korišteni su samo samošireći stentovi tipa Wingspan (Boston Scientific). Iz ovog smo istraživanja, međutim, naučili vrlo važnu lekciju: treba biti na oprezu i očekivati iznenađenja.

[1] Proporcija pacijenata koji su imali srčani udar unutar 365 dana među 224 pacijenta: $45/224 = 0.20$.

Vježbe

1.1 Migrena i akupunktura, DIO I. Migrena je posebno bolni oblik glavobolje protiv koje pacijenti ponekad žele koristiti akupunkturu. Kako bi ustanovili da li akupunktira olakšava bolove kod migrene, istraživači su proveli kontrolirano randomizirano istraživanje u okviru kojeg je 89 žena s dijagnozom migrene slučajno raspoređeno u dvije grupe: izloženu i kontrolnu. 43 pacijentice u izloženoj grupi primile su terapiju akupunkturom posebno osmišljenom za liječenje migrene. 46 pacijentica u kontrolnoj grupi primile su placebo terapiju (ubod iglom na mjestu koje nije akupunkturno). 24 sata nakon što su primile terapiju, istraživači su pitali pacijentice da li su bile bez bolova. Rezultati su sažeti u donjoj kontingencijskoj tablici.[2]

		Bez bolova		
		Da	Ne	Ukupno
Grupa	Izložena	10	33	43
	Kontrolna	2	44	46
	Ukupno	12	77	89



Slika iz originalnoga znanstvenoga rada koja prikazuje prikladno područje (M) u odnosu na neprikladno područje (S) koje je korišteno u terapiji protiv migrene.

(a) Koji postotak pacijentica u izloženoj grupi je bio bez bolova 24 sata nakon primanja terapije akupunkturom?

(b) Koji postotak pacijentica u kontrolnoj grupi je bio bez bolova?

(c) U kojoj grupi je postotak pacijentica bez bolova 24 sata nakon terapije bio veći?

(d) Vaši dosadašnji rezultati mogli bi sugerirati da je akupunktura učinkovita terapija protiv migrene za sve osobe koje pate od migrene. Međutim, to nije jedini mogući zaključak koji se može izvesti iz vaših rezultata. Koje je drugo moguće objašnjenje uočene razlike između grupa u postotku pacijentica koje nisu imale bolove 24 sata nakon terapije akupunkturom?

1.2 Upala sinusa i antibiotici, Dio I. Istraživači koji proučavaju učinak terapije antibioticima na akutnu upalu sinusa u usporedbi sa simptomatskim liječenjem slučajno su rasporedili 166 odraslih osoba s dijagnozom akutne upale sinusa u dvije grupe: izloženu i kontrolnu. Sudionici istraživanja dobili su ili 10-dnevnu terapiju amoxicilinom (antibiotik) ili placebom sličnog oblika i okusa. Placebo je sadržavao simptomatsku terapiju kao što je panadon, sredstvo za otčepljivanje nosa i sl. Na kraju 10-dnevnog perioda pacijente su pitali da li su osjetili poboljšanje simptoma. Razdioba odgovora prikazana je u sljedećoj tablici.[3]

		Poboljšanje simptoma (percepcija pacijenta)		
		Da	Ne	Ukupno
Grupa	Izložena	66	19	85
	Kontrolna	65	16	81
	Ukupno	131	35	166

(a) Koji je postotak pacijenata u izloženoj grupi osjetio poboljšanje simptoma?

(b) Koji postotak pacijenata u kontrolnoj grupi je osjetio poboljšanje simptoma?

(c) U kojoj grupi je postotak pacijenata koji su osjetili poboljšanje simptoma veći?

(d) Vaši dosadašnji rezultati mogu upućivati da postoji stvarna razlika u učinkovitosti antibiotika i placeba u poboljšanju simptoma upale sinusa. Međutim, to nije jedini mogući zaključak koji se može izvesti iz tih rezultata. Koje je drugo moguće objašnjenje za uočenu razliku u postotku pacijenata koji su osjetili poboljšanje simptoma između izložene i kontrolne grupe?

[2] G. Allais i dr. "Ear acupuncture in the treatment of migraine attacks: a randomized trial on the efficacy of appropriate versus inappropriate acupoints". *Neurological Sci.* 32.1 (2011.), str. 173–175.

[3] J.M. Garbutt i dr. "Amoxicillin for Acute Rhinosinusitis: A Randomized Controlled Trial". *JAMA: The Journal of the American Medical Association* 307.7 (2012.), str. 685–692.

1.2 Data basics

Effective organization and description of data is a first step in most analyses. This section introduces the *data matrix* for organizing data as well as some terminology about different forms of data that will be used throughout this book.

1.2.1 Observations, variables, and data matrices

Figure 1.3 displays rows 1, 2, 3, and 50 of a data set for 50 randomly sampled loans offered through Lending Club, which is a peer-to-peer lending company. These observations will be referred to as the `loan50` data set.

Each row in the table represents a single loan. The formal name for a row is a **case** or **observational unit**. The columns represent characteristics, called **variables**, for each of the loans. For example, the first row represents a loan of $7,500 with an interest rate of 7.34%, where the borrower is based in Maryland (MD) and has an income of $70,000.

VOĐENA VJEŽBA 1.2

What is the grade of the first loan in Figure 1.3? And what is the home ownership status of the borrower for that first loan? For these Guided Practice questions, you can check your answer in the footnote.[4]

In practice, it is especially important to ask clarifying questions to ensure important aspects of the data are understood. For instance, it is always important to be sure we know what each variable means and the units of measurement. Descriptions of the `loan50` variables are given in Figure 1.4.

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

Slika 1.3: Four rows from the `loan50` data matrix.

variable	description
loan_amount	Amount of the loan received, in US dollars.
interest_rate	Interest rate on the loan, in an annual percentage.
term	The length of the loan, which is always set as a whole number of months.
grade	Loan grade, which takes a values A through G and represents the quality of the loan and its likelihood of being repaid.
state	US state where the borrower resides.
total_income	Borrower's total income, including any second income, in US dollars.
homeownership	Indicates whether the person owns, owns but has a mortgage, or rents.

Slika 1.4: Variables and their descriptions for the `loan50` data set.

The data in Figure 1.3 represent a **data matrix**, which is a convenient and common way to organize data, especially if collecting data in a spreadsheet. Each row of a data matrix corresponds to a unique case (observational unit), and each column corresponds to a variable.

[4] The loan's grade is A, and the borrower rents their residence.

When recording data, use a data matrix unless you have a very good reason to use a different structure. This structure allows new cases to be added as rows or new variables as new columns.

VOĐENA VJEŽBA 1.3

The grades for assignments, quizzes, and exams in a course are often recorded in a gradebook that takes the form of a data matrix. How might you organize grade data using a data matrix?[5]

VOĐENA VJEŽBA 1.4

We consider data for 3,142 counties in the United States, which includes each county's name, the state where it resides, its population in 2017, how its population changed from 2010 to 2017, poverty rate, and six additional characteristics. How might these data be organized in a data matrix?[6]

The data described in Guided Practice 1.4 represents the `county` data set, which is shown as a data matrix in Figure 1.5. The variables are summarized in Figure 1.6.

[5] There are multiple strategies that can be followed. One common strategy is to have each student represented by a row, and then add a column for each assignment, quiz, or exam. Under this setup, it is easy to review a single line to understand a student's grade history. There should also be columns to include student information, such as one column to list student names.

[6] Each county may be viewed as a case, and there are eleven pieces of information recorded for each case. A table with 3,142 rows and 11 columns could hold these data, where each row represents a county and each column represents a particular piece of information.

	name	state	pop	pop_change	poverty	homeownership	multi_unit	unemp_rate	metro	median_edu	median_hh_income
1	Autauga	Alabama	55504	1.48	13.7	77.5	7.2	3.86	yes	some_college	55317
2	Baldwin	Alabama	212628	9.19	11.8	76.7	22.6	3.99	yes	some_college	52562
3	Barbour	Alabama	25270	-6.22	27.2	68.0	11.1	5.90	no	hs_diploma	33368
4	Bibb	Alabama	22668	0.73	15.2	82.9	6.6	4.39	yes	hs_diploma	43404
5	Blount	Alabama	58013	0.68	15.6	82.0	3.7	4.02	yes	hs_diploma	47412
6	Bullock	Alabama	10309	-2.28	28.5	76.9	9.9	4.93	no	hs_diploma	29655
7	Butler	Alabama	19825	-2.69	24.4	69.0	13.7	5.49	no	hs_diploma	36326
8	Calhoun	Alabama	114728	-1.51	18.6	70.7	14.3	4.93	yes	some_college	43686
9	Chambers	Alabama	33713	-1.20	18.8	71.4	8.7	4.08	no	hs_diploma	37342
10	Cherokee	Alabama	25857	-0.60	16.1	77.5	4.3	4.05	no	hs_diploma	40041
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Weston	Wyoming	6927	-2.93	14.4	77.9	6.5	3.98	no	some_college	59605

Slika 1.5: Eleven rows from the `county` data set.

variable	description
`name`	County name.
`state`	State where the county resides, or the District of Columbia.
`pop`	Population in 2017.
`pop_change`	Percent change in the population from 2010 to 2017. For example, the value `1.48` in the first row means the population for this county increased by 1.48% from 2010 to 2017.
`poverty`	Percent of the population in poverty.
`homeownership`	Percent of the population that lives in their own home or lives with the owner, e.g. children living with parents who own the home.
`multi_unit`	Percent of living units that are in multi-unit structures, e.g. apartments.
`unemp_rate`	Unemployment rate as a percent.
`metro`	Whether the county contains a metropolitan area.
`median_edu`	Median education level, which can take a value among `below_hs`, `hs_diploma`, `some_college`, and `bachelors`.
`median_hh_income`	Median household income for the county, where a household's income equals the total income of its occupants who are 15 years or older.

Slika 1.6: Variables and their descriptions for the `county` data set.

1.2.2 Types of variables

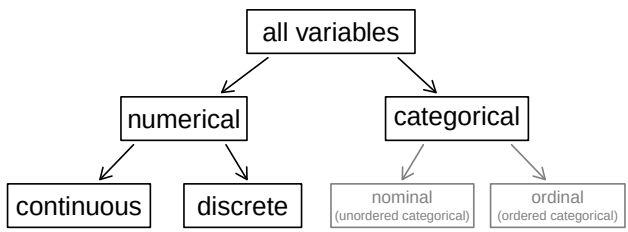
Examine the `unemp_rate`, `pop`, `state`, and `median_edu` variables in the `county` data set. Each of these variables is inherently different from the other three, yet some share certain characteristics.

First consider `unemp_rate`, which is said to be a **numerical** variable since it can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values. On the other hand, we would not classify a variable reporting telephone area codes as numerical since the average, sum, and difference of area codes doesn't have any clear meaning.

The `pop` variable is also numerical, although it seems to be a little different than `unemp_rate`. This variable of the population count can only take whole non-negative numbers (`0`, `1`, `2`, ...). For this reason, the population variable is said to be **discrete** since it can only take numerical values with jumps. On the other hand, the unemployment rate variable is said to be **continuous**.

The variable `state` can take up to 51 values after accounting for Washington, DC: `AL`, `AK`, ..., and `WY`. Because the responses themselves are categories, `state` is called a **categorical** variable, and the possible values are called the variable's **levels**.

Finally, consider the `median_edu` variable, which describes the median education level of county residents and takes values `below_hs`, `hs_diploma`, `some_college`, or `bachelors` in each county. This variable seems to be a hybrid: it is a categorical variable but the levels have a natural ordering. A variable with these properties is called an **ordinal** variable, while a regular categorical variable without this type of special ordering is called a **nominal** variable. To simplify analyses, any ordinal variable in this book will be treated as a nominal (unordered) categorical variable.



Slika 1.7: Breakdown of variables into their respective types.

PRIMJER 1.5

Data were collected about students in a statistics course. Three variables were recorded for each student: number of siblings, student height, and whether the student had previously taken a statistics course. Classify each of the variables as continuous numerical, discrete numerical, or categorical.

The number of siblings and student height represent numerical variables. Because the number of siblings is a count, it is discrete. Height varies continuously, so it is a continuous numerical variable. The last variable classifies students into two categories – those who have and those who have not taken a statistics course – which makes this variable categorical.

VOĐENA VJEŽBA 1.6

An experiment is evaluating the effectiveness of a new drug in treating migraines. A `group` variable is used to indicate the experiment group for each patient: treatment or control. The `num_migraines` variable represents the number of migraines the patient experienced during a 3-month period. Classify each variable as either numerical or categorical?[7]

[7] The `group` variable can take just one of two group names, making it categorical. The `num_migraines` variable describes a count of the number of migraines, which is an outcome where basic arithmetic is sensible, which means this is numerical outcome; more specifically, since it represents a count, `num_migraines` is a discrete numerical variable.

1.2.3 Relationships between variables

Many analyses are motivated by a researcher looking for a relationship between two or more variables. A social scientist may like to answer some of the following questions:

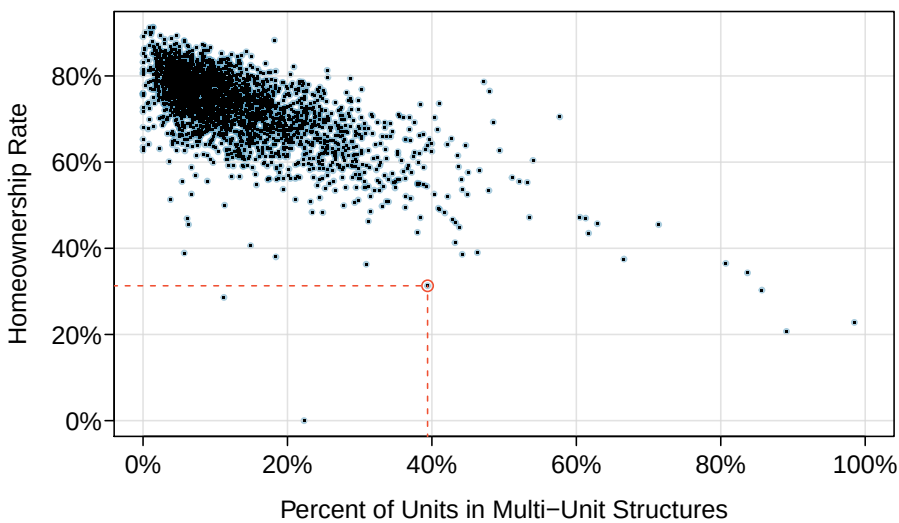
(1) If homeownership is lower than the national average in one county, will the percent of multi-unit structures in that county tend to be above or below the national average?

(2) Does a higher than average increase in county population tend to correspond to counties with higher or lower median household incomes?

(3) How useful a predictor is median education level for the median household income for US counties?

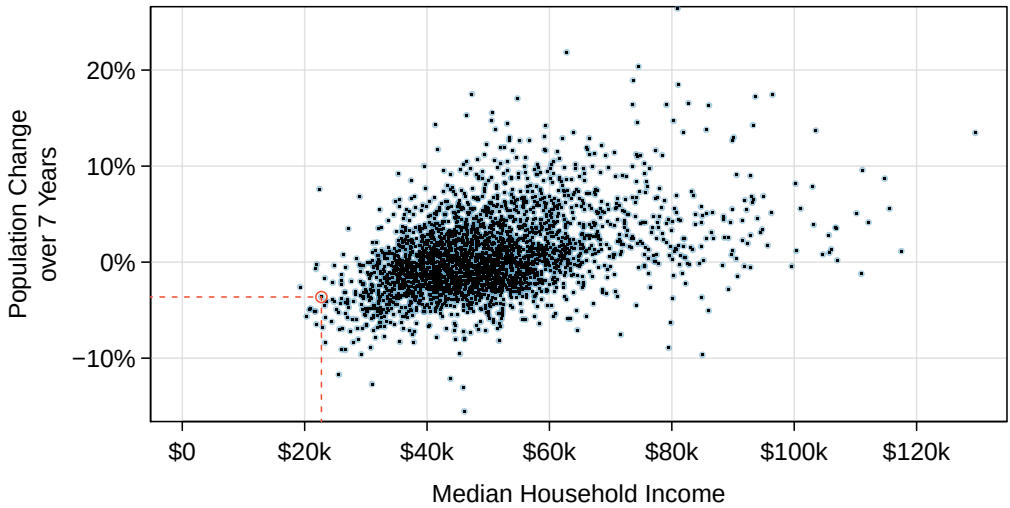
To answer these questions, data must be collected, such as the `county` data set shown in Figure 1.5. Examining summary statistics could provide insights for each of the three questions about counties. Additionally, graphs can be used to visually explore data.

Scatterplots are one type of graph used to study the relationship between two numerical variables. Figure 1.8 compares the variables `homeownership` and `multi_unit`, which is the percent of units in multi-unit structures (e.g. apartments, condos). Each point on the plot represents a single county. For instance, the highlighted dot corresponds to County 413 in the `county` data set: Chattahoochee County, Georgia, which has 39.4% of units in multi-unit structures and a homeownership rate of 31.3%. The scatterplot suggests a relationship between the two variables: counties with a higher rate of multi-units tend to have lower homeownership rates. We might brainstorm as to why this relationship exists and investigate each idea to determine which are the most reasonable explanations.



Slika 1.8: A scatterplot of homeownership versus the percent of units that are in multi-unit structures for US counties. The highlighted dot represents Chattahoochee County, Georgia, which has a multi-unit rate of 39.4% and a homeownership rate of 31.3%.

The multi-unit and homeownership rates are said to be associated because the plot shows a discernible pattern. When two variables show some connection with one another, they are called **associated** variables. Associated variables can also be called **dependent** variables and vice-versa.



Slika 1.9: A scatterplot showing `pop_change` against `median_hh_income`. Owsley County of Kentucky, is highlighted, which lost 3.63% of its population from 2010 to 2017 and had median household income of $22,736.

VOĐENA VJEŽBA 1.7

Examine the variables in the `loan50` data set, which are described in Figure 1.4 on page 11. Create two questions about possible relationships between variables in `loan50` that are of interest to you.[8]

PRIMJER 1.8

This example examines the relationship between a county's population change from 2010 to 2017 and median household income, which is visualized as a scatterplot in Figure 1.9. Are these variables associated?

The larger the median household income for a county, the higher the population growth observed for the county. While this trend isn't true for every county, the trend in the plot is evident. Since there is some relationship between the variables, they are associated.

Because there is a downward trend in Figure 1.8 – counties with more units in multi-unit structures are associated with lower homeownership – these variables are said to be **negatively associated**. A **positive association** is shown in the relationship between the `median_hh_income` and `pop_change` in Figure 1.9, where counties with higher median household income tend to have higher rates of population growth.

If two variables are not associated, then they are said to be **independent**. That is, two variables are independent if there is no evident relationship between the two.

ASSOCIATED OR INDEPENDENT, NOT BOTH

A pair of variables are either related in some way (associated) or not (independent). No pair of variables is both associated and independent.

[8] Two example questions: (1) What is the relationship between loan amount and total income? (2) If someone's income is above the average, will their interest rate tend to be above or below the average?

1.2.4 Explanatory and response variables

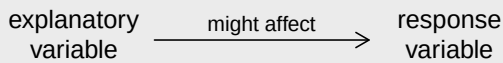
When we ask questions about the relationship between two variables, we sometimes also want to determine if the change in one variable causes a change in the other. Consider the following rephrasing of an earlier question about the `county` data set:

> *If there is an increase in the median household income in a county, does this drive an increase in its population?*

In this question, we are asking whether one variable affects another. If this is our underlying belief, then *median household income* is the **explanatory** variable and the *population change* is the **response** variable in the hypothesized relationship.[9]

EXPLANATORY AND RESPONSE VARIABLES

When we suspect one variable might causally affect another, we label the first variable the explanatory variable and the second the response variable.



For many pairs of variables, there is no hypothesized relationship, and these labels would not be applied to either variable in such cases.

Bear in mind that the act of labeling the variables in this way does nothing to guarantee that a causal relationship exists. A formal evaluation to check whether one variable causes a change in another requires an experiment.

1.2.5 Introducing observational studies and experiments

There are two primary types of data collection: observational studies and experiments.

Researchers perform an **observational study** when they collect data in a way that does not directly interfere with how the data arise. For instance, researchers may collect information via surveys, review medical or company records, or follow a **cohort** of many similar individuals to form hypotheses about why certain diseases might develop. In each of these situations, researchers merely observe the data that arise. In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.

When researchers want to investigate the possibility of a causal connection, they conduct an **experiment**. Usually there will be both an explanatory and a response variable. For instance, we may suspect administering a drug will reduce mortality in heart attack patients over the following year. To check if there really is a causal connection between the explanatory variable and the response, researchers will collect a sample of individuals and split them into groups. The individuals in each group are *assigned* a treatment. When individuals are randomly assigned to a group, the experiment is called a **randomized experiment**. For example, each heart attack patient in the drug trial could be randomly assigned, perhaps by flipping a coin, into one of two groups: the first group receives a **placebo** (fake treatment) and the second group receives the drug. See the case study in Section 1.1 for another example of an experiment, though that study did not employ a placebo.

ASSOCIATION $\neq$ CAUSATION

In general, association does not imply causation, and causation can only be inferred from a randomized experiment.

[9] Sometimes the explanatory variable is called the **independent** variable and the response variable is called the **dependent** variable. However, this becomes confusing since a *pair* of variables might be independent or dependent, so we avoid this language.

Vježbe

1.3 Air pollution and birth outcomes, study components. Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter (PM_{10}) in $\mu g/m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM_{10} and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.[10]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

1.4 Buteyko method, study components. The Buteyko method is a shallow breathing technique developed by Konstantin Buteyko, a Russian doctor, in 1952. Anecdotal evidence suggests that the Buteyko method can reduce asthma symptoms and improve quality of life. In a scientific study to determine the effectiveness of this method, researchers recruited 600 asthma patients aged 18-69 who relied on medication for asthma treatment. These patients were randomly split into two research groups: one practiced the Buteyko method and the other did not. Patients were scored on quality of life, activity, asthma symptoms, and medication reduction on a scale from 0 to 10. On average, the participants in the Buteyko group experienced a significant reduction in asthma symptoms and an improvement in quality of life.[11]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) What are the variables in the study? Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.

1.5 Cheaters, study components. Researchers studying the relationship between honesty, age and self-control conducted an experiment on 160 children between the ages of 5 and 15. Participants reported their age, sex, and whether they were an only child or not. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. The study's findings can be summarized as follows: "Half the students were explicitly told not to cheat and the others were not given any explicit instructions. In the no instruction group probability of cheating was found to be uniform across groups based on child's characteristics. In the group that was explicitly told to not cheat, girls were less likely to cheat, and while rate of cheating didn't vary by age for boys, it decreased with age for girls."[12]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) How many variables were recorded for each subject in the study in order to conclude these findings? State the variables and their types.

[10] B. Ritz i dr. "Effect of air pollution on preterm birth among children born in Southern California between 1989 and 1993". *Epidemiology* 11.5 (2000.), str. 502–511.

[11] J. McGowan. "Health Education: Does the Buteyko Institute Method make a difference?": *Thorax* 58 (2003.).

[12] Alessandro Bucciol i Marco Piovesan. "Luck or cheating? A field experiment on honesty with children". *Journal of Economic Psychology* 32.1 (2011.), str. 73–78.

1.6 Stealers, study components. In a study of the relationship between socio-economic class and unethical behavior, 129 University of California undergraduates at Berkeley were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken.[13]

(a) Identify the main research question of the study.

(b) Who are the subjects in this study, and how many are included?

(c) The study found that students who were identified as upper-class took more candy than others. How many variables were recorded for each subject in the study in order to conclude these findings? State the variables and their types.

1.7 Migraine and acupuncture, Part II. Exercise 1.1 introduced a study exploring whether acupuncture had any effect on migraines. Researchers conducted a randomized controlled study where patients were randomly assigned to one of two groups: treatment or control. The patients in the treatment group received acupuncture that was specifically designed to treat migraines. The patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free. What are the explanatory and response variables in this study?

1.8 Sinusitis and antibiotics, Part II. Exercise 1.2 introduced a study exploring the effect of antibiotic treatment for acute sinusitis. Study participants either received either a 10-day course of an antibiotic (treatment) or a placebo similar in appearance and taste (control). At the end of the 10-day period, patients were asked if they experienced improvement in symptoms. What are the explanatory and response variables in this study?

1.9 Fisher's irises. Sir Ronald Aylmer Fisher was an English statistician, evolutionary biologist, and geneticist who worked on a data set that contained sepal length and width, and petal length and width from three species of iris flowers (*setosa*, *versicolor* and *virginica*). There were 50 flowers from each species in the data set.[14]

(a) How many cases were included in the data?

(b) How many numerical variables are included in the data? Indicate what they are, and if they are continuous or discrete.

(c) How many categorical variables are included in the data, and what are they? List the corresponding levels (categories).

Photo by Ryan Claussen (http://flic.kr/p/6QTcuX) CC BY-SA 2.0 license

1.10 Smoking habits of UK residents. A survey was conducted to study the smoking habits of UK residents. Below is a data matrix displaying a portion of the data collected in this survey. Note that "£" stands for British Pounds Sterling, "cig" stands for cigarettes, and "N/A" refers to a missing component of the data.[15]

	sex	age	marital	grossIncome	smoke	amtWeekends	amtWeekdays
1	Female	42	Single	Under £2,600	Yes	12 cig/day	12 cig/day
2	Male	44	Single	£10,400 to £15,600	No	N/A	N/A
3	Male	53	Married	Above £36,400	Yes	6 cig/day	6 cig/day
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1691	Male	40	Single	£2,600 to £5,200	Yes	8 cig/day	8 cig/day

(a) What does each row of the data matrix represent?

(b) How many participants were included in the survey?

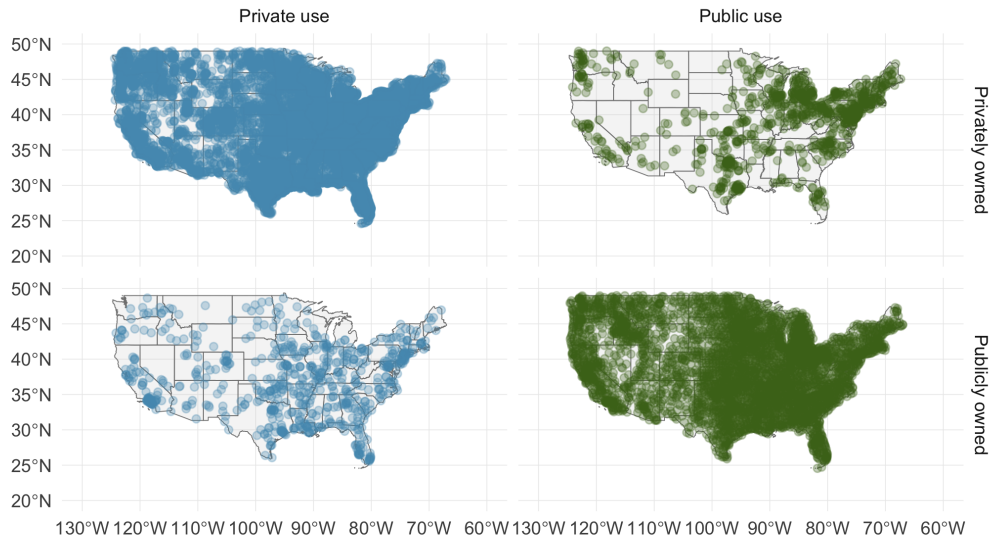
(c) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

[13] P.K. Piff i dr. "Higher social class predicts increased unethical behavior". *Proceedings of the National Academy of Sciences* (2012.).

[14] R.A Fisher. "The Use of Multiple Measurements in Taxonomic Problems". *Annals of Eugenics* 7 (1936.), str. 179–188.

[15] National STEM Centre, Large Datasets from stats4schools.

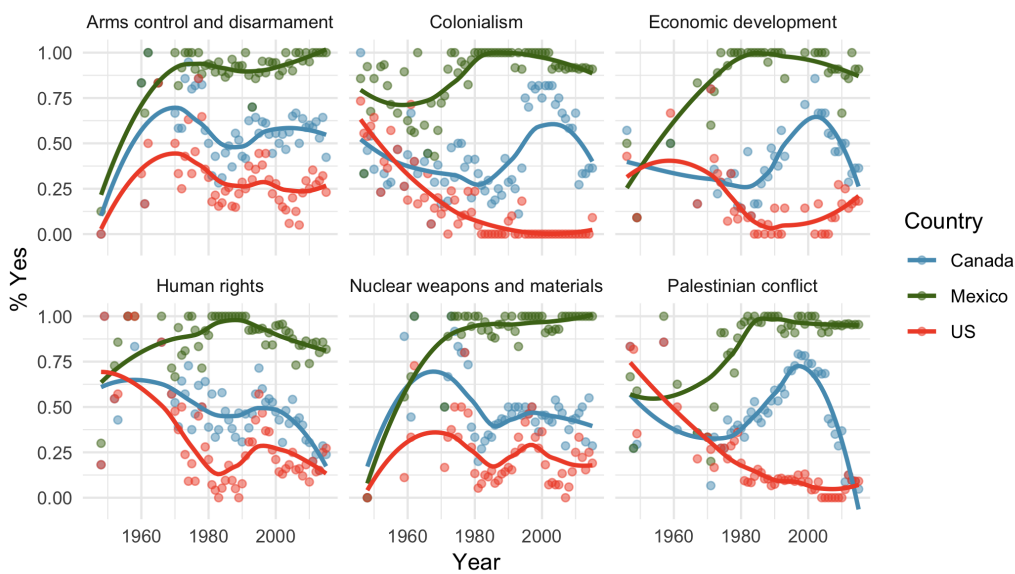
1.11 US Airports. The visualization below shows the geographical distribution of airports in the contiguous United States and Washington, DC. This visualization was constructed based on a dataset where each observation is an airport.[16]



(a) List the variables used in creating this visualization.

(b) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

1.12 UN Votes. The visualization below shows voting patterns in the United States, Canada, and Mexico in the United Nations General Assembly on a variety of issues. Specifically, for a given year between 1946 and 2015, it displays the percentage of roll calls in which the country voted yes for each issue. This visualization was constructed based on a dataset where each observation is a country/year pair.[17]



(a) List the variables used in creating this visualization.

(b) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

[16] Federal Aviation Administration, www.faa.gov/airports/airport_safety/airportdata_5010.

[17] David Robinson. *unvotes: United Nations General Assembly Voting Data*. R package version 0.2.0. 2017. URL: https://CRAN.R-project.org/package=unvotes.

1.3 Sampling principles and strategies

The first step in conducting research is to identify topics or questions that are to be investigated. A clearly laid out research question is helpful in identifying what subjects or cases should be studied and what variables are important. It is also important to consider *how* data are collected so that they are reliable and help achieve the research goals.

1.3.1 Populations and samples

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last 5 years, what is the average time to complete a degree for Duke undergrads?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each research question refers to a target **population**. In the first question, the target population is all swordfish in the Atlantic ocean, and each fish represents a case. Often times, it is too expensive to collect data for every case in a population. Instead, a sample is taken. A **sample** represents a subset of the cases and is often a small fraction of the population. For instance, 60 swordfish (or some other number) in the population might be selected, and this sample data may be used to provide an estimate of the population average and answer the research question.

VOĐENA VJEŽBA 1.9

For the second and third questions above, identify the target population and what represents an individual case.[18]

1.3.2 Anecdotal evidence

Consider the following possible responses to the three research questions:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
2. I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.
3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Each conclusion is based on data. However, there are two problems. First, the data only represent one or two cases. Second, and more importantly, it is unclear whether these cases are actually representative of the population. Data collected in this haphazard fashion are called **anecdotal evidence**.

ANECDOTAL EVIDENCE

Be careful of data collected in a haphazard fashion. Such evidence may be true and verifiable, but it may only represent extraordinary cases.

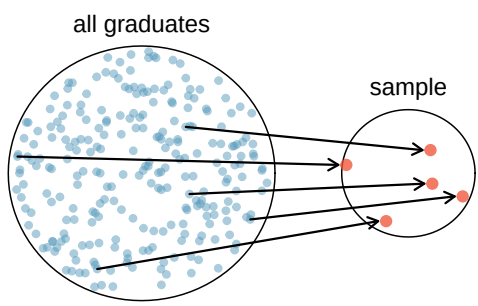
[18] (2) The first question is only relevant to students who complete their degree; the average cannot be computed using a student who never finished her degree. Thus, only Duke undergrads who graduated in the last five years represent cases in the population under consideration. Each such student is an individual case. (3) A person with severe heart disease represents a case. The population includes all people with severe heart disease.

Slika 1.10: In February 2010, some media pundits cited one large snow storm as valid evidence against global warming. As comedian Jon Stewart pointed out, "It's one storm, in one region, of one country."

Anecdotal evidence typically is composed of unusual cases that we recall based on their striking characteristics. For instance, we are more likely to remember the two people we met who took 7 years to graduate than the six others who graduated in four years. Instead of looking at the most unusual cases, we should examine a sample of many cases that represent the population.

1.3.3 Sampling from a population

We might try to estimate the time to graduation for Duke undergraduates in the last 5 years by collecting a sample of students. All graduates in the last 5 years represent the *population*, and graduates who are selected for review are collectively called the *sample*. In general, we always seek to *randomly* select a sample from a population. The most basic type of random selection is equivalent to how raffles are conducted. For example, in selecting graduates, we could write each graduate's name on a raffle ticket and draw 100 tickets. The selected names would represent a random sample of 100 graduates. We pick samples randomly to reduce the chance we introduce biases.

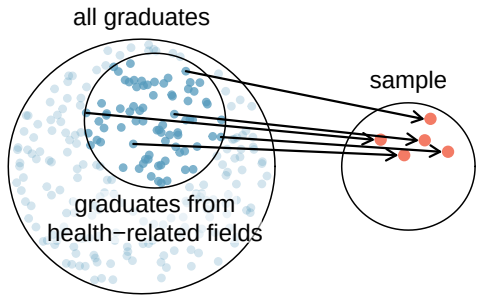


Slika 1.11: In this graphic, five graduates are randomly selected from the population to be included in the sample.

PRIMJER 1.10

Suppose we ask a student who happens to be majoring in nutrition to select several graduates for the study. What kind of students do you think she might collect? Do you think her sample would be representative of all graduates?

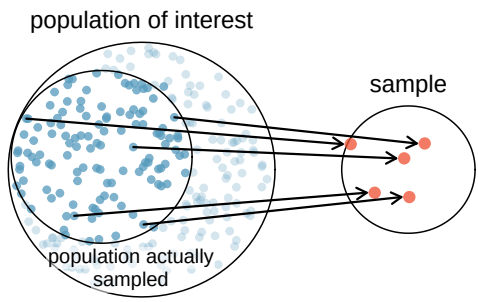
Perhaps she would pick a disproportionate number of graduates from health-related fields. Or perhaps her selection would be a good representation of the population. When selecting samples by hand, we run the risk of picking a **biased** sample, even if their bias isn't intended.



Slika 1.12: Asked to pick a sample of graduates, a nutrition major might inadvertently pick a disproportionate number of graduates from health-related majors.

If someone was permitted to pick and choose exactly which graduates were included in the sample, it is entirely possible that the sample could be skewed to that person's interests, which may be entirely unintentional. This introduces **bias** into a sample. Sampling randomly helps resolve this problem. The most basic random sample is called a **simple random sample**, and which is equivalent to using a raffle to select cases. This means that each case in the population has an equal chance of being included and there is no implied connection between the cases in the sample.

The act of taking a simple random sample helps minimize bias. However, bias can crop up in other ways. Even when people are picked at random, e.g. for surveys, caution must be exercised if the **non-response rate** is high. For instance, if only 30% of the people randomly sampled for a survey actually respond, then it is unclear whether the results are **representative** of the entire population. This **non-response bias** can skew results.



Slika 1.13: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

Another common downfall is a **convenience sample**, where individuals who are easily accessible are more likely to be included in the sample. For instance, if a political survey is done by stopping people walking in the Bronx, this will not represent all of New York City. It is often difficult to discern what sub-population a convenience sample represents.

VOĐENA VJEŽBA 1.11

We can easily access ratings for products, sellers, and companies through websites. These ratings are based only on those people who go out of their way to provide a rating. If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product?[19]

[19] Answers will vary. From our own anecdotal experiences, we believe people tend to rant more about products that fell below expectations than rave about those that perform as expected. For this reason, we suspect there is a negative bias in product ratings on sites like Amazon. However, since our experiences may not be representative, we also keep an open mind.

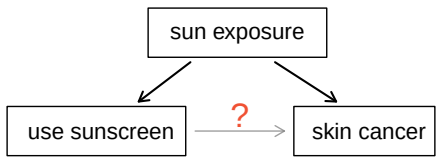
1.3.4 Observational studies

Data where no treatment has been explicitly applied (or explicitly withheld) is called **observational data**. For instance, the loan data and county data described in Section 1.2 are both examples of observational data. Making causal conclusions based on experiments is often reasonable. However, making the same causal conclusions based on observational data can be treacherous and is not recommended. Thus, observational studies are generally only sufficient to show associations or form hypotheses that we later check using experiments.

VOĐENA VJEŽBA 1.12

Suppose an observational study tracked sunscreen use and skin cancer, and it was found that the more sunscreen someone used, the more likely the person was to have skin cancer. Does this mean sunscreen *causes* skin cancer?[20]

Some previous research tells us that using sunscreen actually reduces skin cancer risk, so maybe there is another variable that can explain this hypothetical association between sunscreen usage and skin cancer. One important piece of information that is absent is sun exposure. If someone is out in the sun all day, she is more likely to use sunscreen *and* more likely to get skin cancer. Exposure to the sun is unaccounted for in the simple investigation.



Sun exposure is what is called a **confounding variable**,[21] which is a variable that is correlated with both the explanatory and response variables. While one method to justify making causal conclusions from observational studies is to exhaust the search for confounding variables, there is no guarantee that all confounding variables can be examined or measured.

VOĐENA VJEŽBA 1.13

Figure 1.8 shows a negative association between the homeownership rate and the percentage of multi-unit structures in a county. However, it is unreasonable to conclude that there is a causal relationship between the two variables. Suggest a variable that might explain the negative relationship.[22]

Observational studies come in two forms: prospective and retrospective studies. A **prospective study** identifies individuals and collects information as events unfold. For instance, medical researchers may identify and follow a group of patients over many years to assess the possible influences of behavior on cancer risk. One example of such a study is The Nurses' Health Study, started in 1976 and expanded in 1989. This prospective study recruits registered nurses and then collects data from them using questionnaires. **Retrospective studies** collect data after events have taken place, e.g. researchers may review past events in medical records. Some data sets may contain both prospectively- and retrospectively-collected variables.

1.3.5 Four sampling methods

Almost all statistical methods are based on the notion of implied randomness. If observational data are not collected in a random framework from a population, these statistical methods – the estimates and errors associated with the estimates – are not reliable. Here we consider four random

[20] No. See the paragraph following the exercise for an explanation.

[21] Also called a **lurking variable**, **confounding factor**, or a **confounder**.

[22] Answers will vary. Population density may be important. If a county is very dense, then this may require a larger fraction of residents to live in multi-unit structures. Additionally, the high density may contribute to increases in property value, making homeownership infeasible for many residents.

sampling techniques: simple, stratified, cluster, and multistage sampling. Figures 1.14 and 1.15 provide graphical representations of these techniques.

Simple random sampling is probably the most intuitive form of random sampling. Consider the salaries of Major League Baseball (MLB) players, where each player is a member of one of the league's 30 teams. To take a simple random sample of 120 baseball players and their salaries, we could write the names of that season's several hundreds of players onto slips of paper, drop the slips into a bucket, shake the bucket around until we are sure the names are all mixed up, then draw out slips until we have the sample of 120 players. In general, a sample is referred to as "simple random" if each case in the population has an equal chance of being included in the final sample *and* knowing that a case is included in a sample does not provide useful information about which other cases are included.

Stratified sampling is a divide-and-conquer sampling strategy. The population is divided into groups called **strata**. The strata are chosen so that similar cases are grouped together, then a second sampling method, usually simple random sampling, is employed within each stratum. In the baseball salary example, the teams could represent the strata, since some teams have a lot more money (up to 4 times as much!). Then we might randomly sample 4 players from each team for a total of 120 players.

Stratified sampling is especially useful when the cases in each stratum are very similar with respect to the outcome of interest. The downside is that analyzing data from a stratified sample is a more complex task than analyzing data from a simple random sample. The analysis methods introduced in this book would need to be extended to analyze data collected using stratified sampling.

PRIMJER 1.14

Why would it be good for cases within each stratum to be very similar?

We might get a more stable estimate for the subpopulation in a stratum if the cases are very similar, leading to more precise estimates within each group. When we combine these estimates into a single estimate for the full population, that population estimate will tend to be more precise since each individual group estimate is itself more precise.

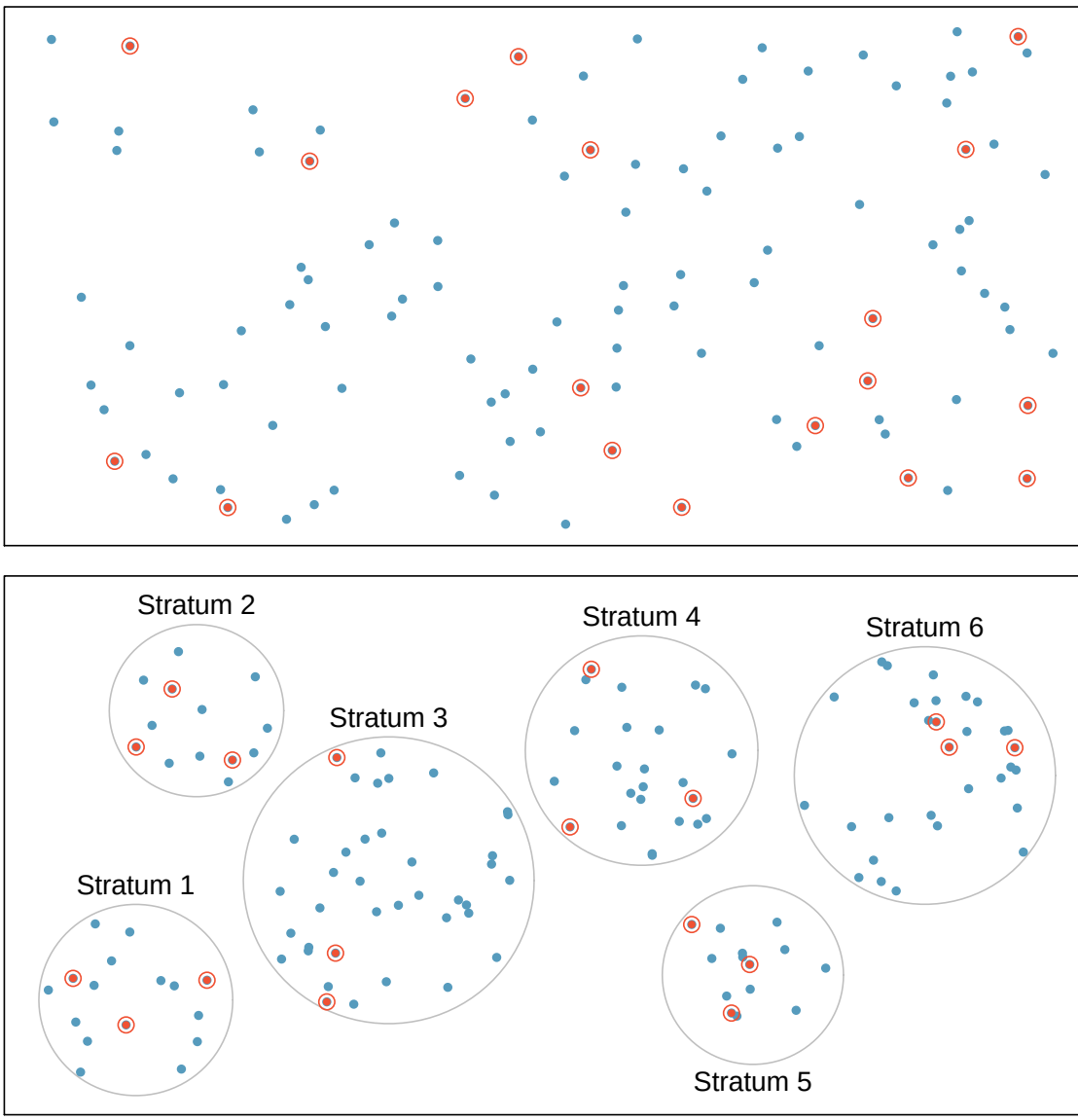
In a **cluster sample**, we break up the population into many groups, called **clusters**. Then we sample a fixed number of clusters and include all observations from each of those clusters in the sample. A **multistage sample** is like a cluster sample, but rather than keeping all observations in each cluster, we collect a random sample within each selected cluster.

Sometimes cluster or multistage sampling can be more economical than the alternative sampling techniques. Also, unlike stratified sampling, these approaches are most helpful when there is a lot of case-to-case variability within a cluster but the clusters themselves don't look very different from one another. For example, if neighborhoods represented clusters, then cluster or multistage sampling work best when the neighborhoods are very diverse. A downside of these methods is that more advanced techniques are typically required to analyze the data, though the methods in this book can be extended to handle such data.

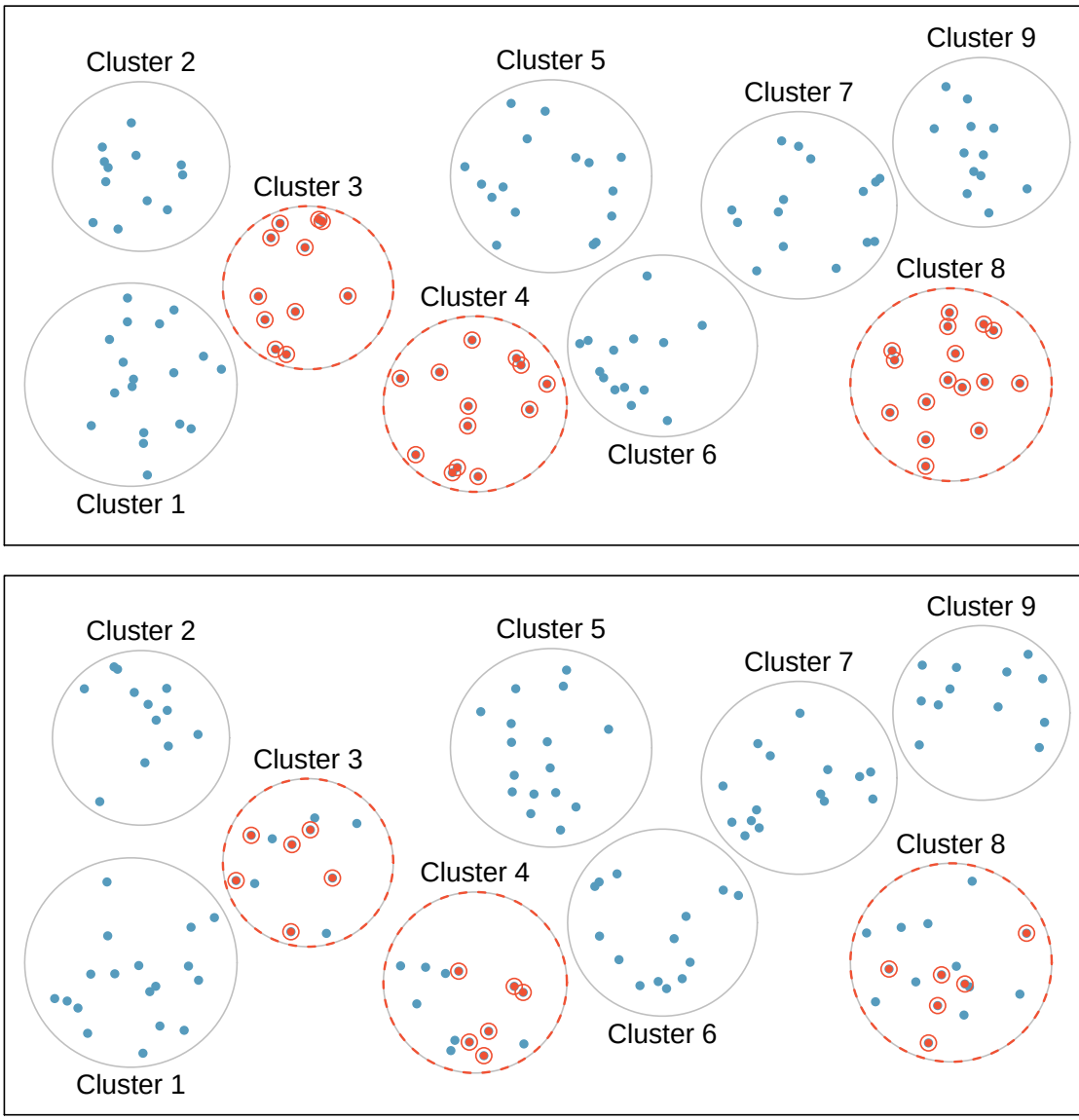
PRIMJER 1.15

Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria. What sampling method should be employed?

A simple random sample would likely draw individuals from all 30 villages, which could make data collection extremely expensive. Stratified sampling would be a challenge since it is unclear how we would build strata of similar individuals. However, cluster sampling or multistage sampling seem like very good ideas. If we decided to use multistage sampling, we might randomly select half of the villages, then randomly select 10 people from each. This would probably reduce our data collection costs substantially in comparison to a simple random sample, and the cluster sample would still give us reliable information, even if we would need to analyze the data with slightly more advanced methods than we discuss in this book.



Slika 1.14: Examples of simple random and stratified sampling. In the top panel, simple random sampling was used to randomly select the 18 cases. In the bottom panel, stratified sampling was used: cases were grouped into strata, then simple random sampling was employed within each stratum.



Slika 1.15: Examples of cluster and multistage sampling. In the top panel, cluster sampling was used: data were binned into nine clusters, three of these clusters were sampled, and all observations within these three cluster were included in the sample. In the bottom panel, multistage sampling was used, which differs from cluster sampling only in that we randomly select a subset of each cluster to be included in the sample rather than measuring every case in each sampled cluster.

Vježbe

1.13 Air pollution and birth outcomes, scope of inference. Exercise 1.3 introduces a study where researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

1.14 Cheaters, scope of inference. Exercise 1.5 introduces a study where researchers studying the relationship between honesty, age, and self-control conducted an experiment on 160 children between the ages of 5 and 15. The researchers asked each child to toss a fair coin in private and to record the outcome (white or black) on a paper sheet, and said they would only reward children who report white. Half the students were explicitly told not to cheat and the others were not given any explicit instructions. Differences were observed in the cheating rates in the instruction and no instruction groups, as well as some differences across children's characteristics within each group.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

1.15 Buteyko method, scope of inference. Exercise 1.4 introduces a study on using the Buteyko shallow breathing technique to reduce asthma symptoms and improve quality of life. As part of this study 600 asthma patients aged 18-69 who relied on medication for asthma treatment were recruited and randomly assigned to two groups: one practiced the Buteyko method and the other did not. Those in the Buteyko group experienced, on average, a significant reduction in asthma symptoms and an improvement in quality of life.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

1.16 Stealers, scope of inference. Exercise 1.6 introduces a study on the relationship between socioeconomic class and unethical behavior. As part of this study 129 University of California Berkeley undergraduates were asked to identify themselves as having low or high social-class by comparing themselves to others with the most (least) money, most (least) education, and most (least) respected jobs. They were also presented with a jar of individually wrapped candies and informed that the candies were for children in a nearby laboratory, but that they could take some if they wanted. After completing some unrelated tasks, participants reported the number of candies they had taken. It was found that those who were identified as upper-class took more candy than others.

(a) Identify the population of interest and the sample in this study.

(b) Comment on whether or not the results of the study can be generalized to the population, and if the findings of the study can be used to establish causal relationships.

1.17 Relaxing after work. The General Social Survey asked the question, "After an average work day, about how many hours do you have to relax or pursue activities that you enjoy?" to a random sample of 1,155 Americans. The average relaxing time was found to be 1.65 hours. Determine which of the following is an observation, a variable, a sample statistic (value calculated based on the observed sample), or a population parameter.

(a) An American in the sample.

(b) Number of hours spent relaxing after an average work day.

(c) 1.65.

(d) Average number of hours all Americans spend relaxing after an average work day.

1.18 Cats on YouTube. Suppose you want to estimate the percentage of videos on YouTube that are cat videos. It is impossible for you to watch all videos on YouTube so you use a random video picker to select 1000 videos for you. You find that 2% of these videos are cat videos. Determine which of the following is an observation, a variable, a sample statistic (value calculated based on the observed sample), or a population parameter.

(a) Percentage of all videos on YouTube that are cat videos.
(b) 2%.
(c) A video in your sample.
(d) Whether or not a video is a cat video.

1.19 Course satisfaction across sections. A large college class has 160 students. All 160 students attend the lectures together, but the students are divided into 4 groups, each of 40 students, for lab sections administered by different teaching assistants. The professor wants to conduct a survey about how satisfied the students are with the course, and he believes that the lab section a student is in might affect the student's overall satisfaction with the course.

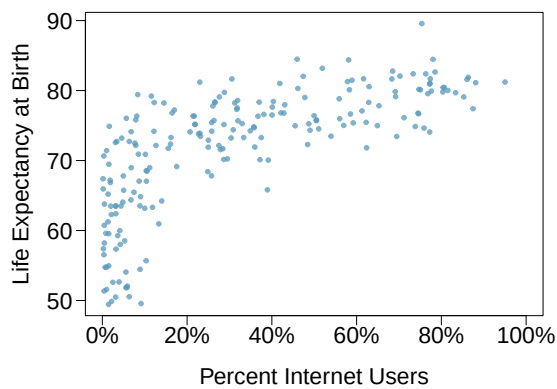
(a) What type of study is this?
(b) Suggest a sampling strategy for carrying out this study.

1.20 Housing proposal across dorms. On a large college campus first-year students and sophomores live in dorms located on the eastern part of the campus and juniors and seniors live in dorms located on the western part of the campus. Suppose you want to collect student opinions on a new housing structure the college administration is proposing and you want to make sure your survey equally represents opinions from students from all years.

(a) What type of study is this?
(b) Suggest a sampling strategy for carrying out this study.

1.21 Internet use and life expectancy. The following scatterplot was created as part of a study evaluating the relationship between estimated life expectancy at birth (as of 2014) and percentage of internet users (as of 2009) in 208 countries for which such data were available.[23]

(a) Describe the relationship between life expectancy and percentage of internet users.
(b) What type of study is this?
(c) State a possible confounding variable that might explain this relationship and describe its potential effect.



1.22 Stressed out, Part I. A study that surveyed a random sample of otherwise healthy high school students found that they are more likely to get muscle cramps when they are stressed. The study also noted that students drink more coffee and sleep less when they are stressed.

(a) What type of study is this?
(b) Can this study be used to conclude a causal relationship between increased stress and muscle cramps?
(c) State possible confounding variables that might explain the observed relationship between increased stress and muscle cramps.

1.23 Evaluate sampling methods. A university wants to determine what fraction of its undergraduate student body support a new $25 annual fee to improve the student union. For each proposed method below, indicate whether the method is reasonable or not.

(a) Survey a simple random sample of 500 students.
(b) Stratify students by their field of study, then sample 10% of students from each stratum.
(c) Cluster students by their ages (e.g. 18 years old in one cluster, 19 years old in one cluster, etc.), then randomly sample three clusters and survey all students in those clusters.

[23] CIA Factbook, Country Comparisons, 2014.

1.24 Random digit dialing. The Gallup Poll uses a procedure called random digit dialing, which creates phone numbers based on a list of all area codes in America in conjunction with the associated number of residential households in each area code. Give a possible reason the Gallup Poll chooses to use random digit dialing instead of picking phone numbers from the phone book.

1.25 Haters are gonna hate, study confirms. A study published in the *Journal of Personality and Social Psychology* asked a group of 200 randomly sampled men and women to evaluate how they felt about various subjects, such as camping, health care, architecture, taxidermy, crossword puzzles, and Japan in order to measure their attitude towards mostly independent stimuli. Then, they presented the participants with information about a new product: a microwave oven. This microwave oven does not exist, but the participants didn't know this, and were given three positive and three negative fake reviews. People who reacted positively to the subjects on the dispositional attitude measurement also tended to react positively to the microwave oven, and those who reacted negatively tended to react negatively to it. Researchers concluded that "some people tend to like things, whereas others tend to dislike things, and a more thorough understanding of this tendency will lead to a more thorough understanding of the psychology of attitudes."[24]

(a) What are the cases?
(b) What is (are) the response variable(s) in this study?
(c) What is (are) the explanatory variable(s) in this study?
(d) Does the study employ random sampling?
(e) Is this an observational study or an experiment? Explain your reasoning.
(f) Can we establish a causal link between the explanatory and response variables?
(g) Can the results of the study be generalized to the population at large?

1.26 Family size. Suppose we want to estimate household size, where a "household" is defined as people living together in the same dwelling, and sharing living accommodations. If we select students at random at an elementary school and ask them what their family size is, will this be a good measure of household size? Or will our average be biased? If so, will it overestimate or underestimate the true value?

1.27 Sampling strategies. A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey. Various research strategies for collecting data are described below. In each, name the sampling method proposed and any bias you might expect.

(a) He randomly samples 40 students from the study's population, gives them the survey, asks them to fill it out and bring it back the next day.
(b) He gives out the survey only to his friends, making sure each one of them fills out the survey.
(c) He posts a link to an online survey on Facebook and asks his friends to fill out the survey.
(d) He randomly samples 5 classes and asks a random sample of students from those classes to fill out the survey.

1.28 Reading the paper. Below are excerpts from two articles published in the *NY Times*:

(a) An article titled *Risks: Smokers Found More Prone to Dementia* states the following:[25]

> "Researchers analyzed data from 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50-60 years old. 23 years later, about 25% of the group had dementia, including 1,136 with Alzheimer's disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37% more likely than nonsmokers to develop dementia, and the risks went up with increased smoking; 44% for one to two packs a day; and twice the risk for more than two packs."

Based on this study, can we conclude that smoking causes dementia later in life? Explain your reasoning.

(b) Another article titled *The School Bully Is Sleepy* states the following:[26]

> "The University of Michigan study, collected survey data from parents on each child's sleep habits and asked both parents and teachers to assess behavioral concerns. About a third of the students studied were identified by parents or teachers as having problems with disruptive behavior or bullying. The researchers found that children who had behavioral issues and those who were identified as bullies were twice as likely to have shown symptoms of sleep disorders."

A friend of yours who read the article says, "The study shows that sleep disorders lead to bullying in school children." Is this statement justified? If not, how best can you describe the conclusion that can be drawn from this study?

[24] Justin Hepler i Dolores Albarracín. "Attitudes without objects - Evidence for a dispositional attitude, its measurement, and its consequences". *Journal of personality and social psychology* 104.6 (2013.), str. 1060.
[25] R.C. Rabin. "Risks: Smokers Found More Prone to Dementia". *New York Times* (2010.).
[26] T. Parker-Pope. "The School Bully Is Sleepy". *New York Times* (2011.).

1.4 Experiments

Studies where the researchers assign treatments to cases are called **experiments**. When this assignment includes randomization, e.g. using a coin flip to decide which treatment a patient receives, it is called a **randomized experiment**. Randomized experiments are fundamentally important when trying to show a causal connection between two variables.

1.4.1 Principles of experimental design

Randomized experiments are generally built on four principles.

Controlling. Researchers assign treatments to cases, and they do their best to **control** any other differences in the groups.[27] For example, when patients take a drug in pill form, some patients take the pill with only a sip of water while others may have it with an entire glass of water. To control for the effect of water consumption, a doctor may ask all patients to drink a 12 ounce glass of water with the pill.

Randomization. Researchers randomize patients into treatment groups to account for variables that cannot be controlled. For example, some patients may be more susceptible to a disease than others due to their dietary habits. Randomizing patients into the treatment or control group helps even out such differences, and it also prevents accidental bias from entering the study.

Replication. The more cases researchers observe, the more accurately they can estimate the effect of the explanatory variable on the response. In a single study, we **replicate** by collecting a sufficiently large sample. Additionally, a group of scientists may replicate an entire study to verify an earlier finding.

Blocking. Researchers sometimes know or suspect that variables, other than the treatment, influence the response. Under these circumstances, they may first group individuals based on this variable into **blocks** and then randomize cases within each block to the treatment groups. This strategy is often referred to as **blocking**. For instance, if we are looking at the effect of a drug on heart attacks, we might first split patients in the study into low-risk and high-risk blocks, then randomly assign half the patients from each block to the control group and the other half to the treatment group, as shown in Figure 1.16. This strategy ensures each treatment group has an equal number of low-risk and high-risk patients.

It is important to incorporate the first three experimental design principles into any study, and this book describes applicable methods for analyzing data from such experiments. Blocking is a slightly more advanced technique, and statistical methods in this book may be extended to analyze data collected using blocking.

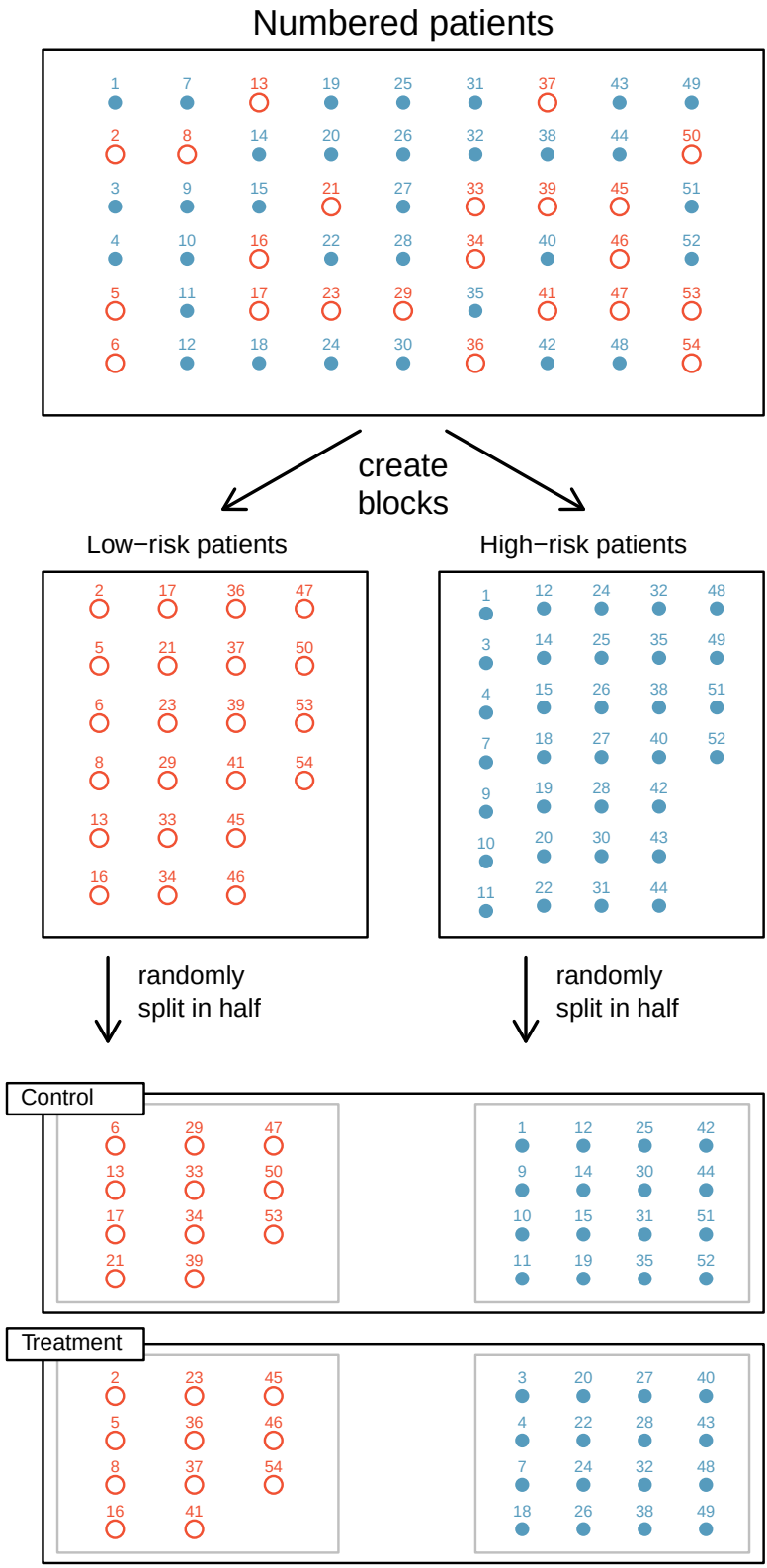
1.4.2 Reducing bias in human experiments

Randomized experiments are the gold standard for data collection, but they do not ensure an unbiased perspective into the cause and effect relationship in all cases. Human studies are perfect examples where bias can unintentionally arise. Here we reconsider a study where a new drug was used to treat heart attack patients. In particular, researchers wanted to know if the drug reduced deaths in patients.

These researchers designed a randomized experiment because they wanted to draw causal conclusions about the drug's effect. Study volunteers[28] were randomly placed into two study groups. One group, the **treatment group**, received the drug. The other group, called the **control group**, did not receive any drug treatment.

[27] This is a different concept than a *control group*, which we discuss in the second principle and in Section 1.4.2.

[28] Human subjects are often called **patients**, **volunteers**, or **study participants**.



Slika 1.16: Blocking using a variable depicting patient risk. Patients are first divided into low-risk and high-risk blocks, then each block is evenly separated into the treatment groups using randomization. This strategy ensures an equal representation of patients in each treatment group from both the low-risk and high-risk categories.

Put yourself in the place of a person in the study. If you are in the treatment group, you are given a fancy new drug that you anticipate will help you. On the other hand, a person in the other group doesn't receive the drug and sits idly, hoping her participation doesn't increase her risk of death. These perspectives suggest there are actually two effects: the one of interest is the effectiveness of the drug, and the second is an emotional effect that is difficult to quantify.

Researchers aren't usually interested in the emotional effect, which might bias the study. To circumvent this problem, researchers do not want patients to know which group they are in. When researchers keep the patients uninformed about their treatment, the study is said to be **blind**. But there is one problem: if a patient doesn't receive a treatment, she will know she is in the control group. The solution to this problem is to give fake treatments to patients in the control group. A fake treatment is called a **placebo**, and an effective placebo is the key to making a study truly blind. A classic example of a placebo is a sugar pill that is made to look like the actual treatment pill. Often times, a placebo results in a slight but real improvement in patients. This effect has been dubbed the **placebo effect**.

The patients are not the only ones who should be blinded: doctors and researchers can accidentally bias a study. When a doctor knows a patient has been given the real treatment, she might inadvertently give that patient more attention or care than a patient that she knows is on the placebo. To guard against this bias, which again has been found to have a measurable effect in some instances, most modern studies employ a **double-blind** setup where doctors or researchers who interact with patients are, just like the patients, unaware of who is or is not receiving the treatment.[29]

VOĐENA VJEŽBA 1.16

Look back to the study in Section 1.1 where researchers were testing whether stents were effective at reducing strokes in at-risk patients. Is this an experiment? Was the study blinded? Was it double-blinded?[30]

VOĐENA VJEŽBA 1.17

For the study in Section 1.1, could the researchers have employed a placebo? If so, what would that placebo have looked like?[31]

You may have many questions about the ethics of sham surgeries to create a placebo after reading Guided Practice 1.17. These questions may have even arisen in your mind when in the general experiment context, where a possibly helpful treatment was withheld from individuals in the control group; the main difference is that a sham surgery tends to create additional risk, while withholding a treatment only maintains a person's risk.

There are always multiple viewpoints of experiments and placebos, and rarely is it obvious which is ethically "correct". For instance, is it ethical to use a sham surgery when it creates a risk to the patient? However, if we don't use sham surgeries, we may promote the use of a costly treatment that has no real effect; if this happens, money and other resources will be diverted away from other treatments that are known to be helpful. Ultimately, this is a difficult situation where we cannot perfectly protect both the patients who have volunteered for the study and the patients who may benefit (or not) from the treatment in the future.

[29] There are always some researchers involved in the study who do know which patients are receiving which treatment. However, they do not interact with the study's patients and do not tell the blinded health care professionals who is receiving which treatment.

[30] The researchers assigned the patients into their treatment groups, so this study was an experiment. However, the patients could distinguish what treatment they received, so this study was not blind. The study could not be double-blind since it was not blind.

[31] Ultimately, can we make patients think they got treated from a surgery? In fact, we can, and some experiments use what's called a **sham surgery**. In a sham surgery, the patient does undergo surgery, but the patient does not receive the full treatment, though they will still get a placebo effect.

Vježbe

1.29 Light and exam performance. A study is designed to test the effect of light level on exam performance of students. The researcher believes that light levels might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The treatments are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps).

(a) What is the response variable?
(b) What is the explanatory variable? What are its levels?
(c) What is the blocking variable? What are its levels?

1.30 Vitamin supplements. To assess the effectiveness of taking large doses of vitamin C in reducing the duration of the common cold, researchers recruited 400 healthy volunteers from staff and students at a university. A quarter of the patients were assigned a placebo, and the rest were evenly divided between 1g Vitamin C, 3g Vitamin C, or 3g Vitamin C plus additives to be taken at onset of a cold for the following two days. All tablets had identical appearance and packaging. The nurses who handed the prescribed pills to the patients knew which patient received which treatment, but the researchers assessing the patients when they were sick did not. No significant differences were observed in any measure of cold duration or severity between the four groups, and the placebo group had the shortest duration of symptoms.[32]

(a) Was this an experiment or an observational study? Why?
(b) What are the explanatory and response variables in this study?
(c) Were the patients blinded to their treatment?
(d) Was this study double-blind?
(e) Participants are ultimately able to choose whether or not to use the pills prescribed to them. We might expect that not all of them will adhere and take their pills. Does this introduce a confounding variable to the study? Explain your reasoning.

1.31 Light, noise, and exam performance. A study is designed to test the effect of light level and noise level on exam performance of students. The researcher believes that light and noise levels might have different effects on males and females, so wants to make sure both are equally represented in each treatment. The light treatments considered are fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). The noise treatments considered are no noise, construction noise, and human chatter noise.

(a) What type of study is this?
(b) How many factors are considered in this study? Identify them, and describe their levels.
(c) What is the role of the sex variable in this study?

1.32 Music and learning. You would like to conduct an experiment in class to see if students learn better if they study without any music, with music that has no lyrics (instrumental), or with music that has lyrics. Briefly outline a design for this study.

1.33 Soda preference. You would like to conduct an experiment in class to see if your classmates prefer the taste of regular Coke or Diet Coke. Briefly outline a design for this study.

1.34 Exercise and mental health. A researcher is interested in the effects of exercise on mental health and he proposes the following study: Use stratified random sampling to ensure representative proportions of 18-30, 31-40 and 41- 55 year olds from the population. Next, randomly assign half the subjects from each age group to exercise twice a week, and instruct the rest not to exercise. Conduct a mental health exam at the beginning and at the end of the study, and compare the results.

(a) What type of study is this?
(b) What are the treatment and control groups in this study?
(c) Does this study make use of blocking? If so, what is the blocking variable?
(d) Does this study make use of blinding?
(e) Comment on whether or not the results of the study can be used to establish a causal relationship between exercise and mental health, and indicate whether or not the conclusions can be generalized to the population at large.
(f) Suppose you are given the task of determining if this proposed study should get funding. Would you have any reservations about the study proposal?

[32] C. Audera i dr. "Mega-dose vitamin C in treatment of the common cold: a randomised controlled trial". *Medical Journal of Australia* 175.7 (2001.), str. 359–362.

Vježbe

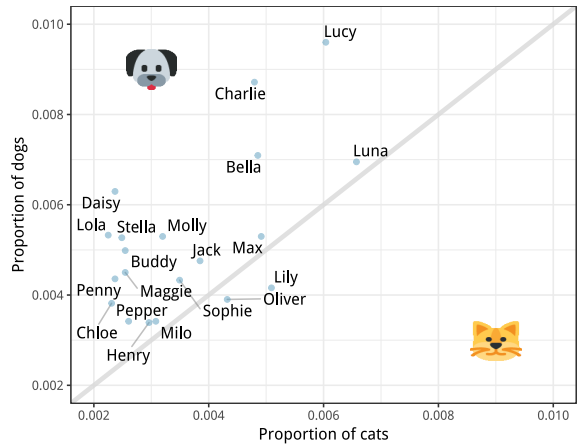
1.35 Pet names. The city of Seattle, WA has an open data portal that includes pets registered in the city. For each registered pet, we have information on the pet's name and species. The following visualization plots the proportion of dogs with a given name versus the proportion of cats with the same name. The 20 most common cat and dog names are displayed. The diagonal line on the plot is the $x = y$ line; if a name appeared on this line, the name's popularity would be exactly the same for dogs and cats.

(a) Are these data collected as part of an experiment or an observational study?

(b) What is the most common dog name? What is the most common cat name?

(c) What names are more common for cats than dogs?

(d) Is the relationship between the two variables positive or negative? What does this mean in context of the data?



1.36 Stressed out, Part II. In a study evaluating the relationship between stress and muscle cramps, half the subjects are randomly assigned to be exposed to increased stress by being placed into an elevator that falls rapidly and stops abruptly and the other half are left at no or baseline stress.

(a) What type of study is this?

(b) Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

1.37 Chia seeds and weight loss. Chia Pets – those terra-cotta figurines that sprout fuzzy green hair – made the chia plant a household name. But chia has gained an entirely new reputation as a diet supplement. In one 2009 study, a team of researchers recruited 38 men and divided them randomly into two groups: treatment or control. They also recruited 38 women, and they randomly placed half of these participants into the treatment group and the other half into the control group. One group was given 25 grams of chia seeds twice a day, and the other was given a placebo. The subjects volunteered to be a part of the study. After 12 weeks, the scientists found no significant difference between the groups in appetite or weight loss.[33]

(a) What type of study is this?

(b) What are the experimental and control treatments in this study?

(c) Has blocking been used in this study? If so, what is the blocking variable?

(d) Has blinding been used in this study?

(e) Comment on whether or not we can make a causal statement, and indicate whether or not we can generalize the conclusion to the population at large.

1.38 City council survey. A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. For each part below, identify the sampling methods described, and describe the statistical pros and cons of the method in the city's context.

(a) Randomly sample 200 households from the city.

(b) Divide the city into 20 neighborhoods, and sample 10 households from each neighborhood.

(c) Divide the city into 20 neighborhoods, randomly sample 3 neighborhoods, and then sample all households from those 3 neighborhoods.

(d) Divide the city into 20 neighborhoods, randomly sample 8 neighborhoods, and then randomly sample 50 households from those neighborhoods.

(e) Sample the 200 households closest to the city council offices.

[33] D.C. Nieman i dr. "Chia seed does not promote weight loss or alter disease risk factors in overweight adults". *Nutrition Research* 29.6 (2009.), str. 414–418.

1.39 Flawed reasoning. Identify the flaw(s) in reasoning in the following scenarios. Explain what the individuals in the study should have done differently if they wanted to make such strong conclusions.

(a) Students at an elementary school are given a questionnaire that they are asked to return after their parents have completed it. One of the questions asked is, "Do you find that your work schedule makes it difficult for you to spend time with your kids after school?" Of the parents who replied, 85% said "no". Based on these results, the school officials conclude that a great majority of the parents have no difficulty spending time with their kids after school.

(b) A survey is conducted on a simple random sample of 1,000 women who recently gave birth, asking them about whether or not they smoked during pregnancy. A follow-up survey asking if the children have respiratory problems is conducted 3 years later. However, only 567 of these women are reached at the same address. The researcher reports that these 567 women are representative of all mothers.

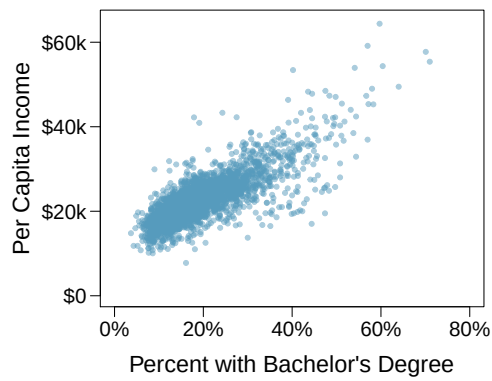
(c) An orthopedist administers a questionnaire to 30 of his patients who do not have any joint problems and finds that 20 of them regularly go running. He concludes that running decreases the risk of joint problems.

1.40 Income and education in US counties. The scatterplot below shows the relationship between per capita income (in thousands of dollars) and percent of population with a bachelor's degree in 3,143 counties in the US in 2010.

(a) What are the explanatory and response variables?

(b) Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.

(c) Can we conclude that having a bachelor's degree increases one's income?



1.41 Eat better, feel better? In a public health study on the effects of consumption of fruits and vegetables on psychological well-being in young adults, participants were randomly assigned to three groups: (1) diet-as-usual, (2) an ecological momentary intervention involving text message reminders to increase their fruits and vegetable consumption plus a voucher to purchase them, or (3) a fruit and vegetable intervention in which participants were given two additional daily servings of fresh fruits and vegetables to consume on top of their normal diet. Participants were asked to take a nightly survey on their smartphones. Participants were student volunteers at the University of Otago, New Zealand. At the end of the 14-day study, only participants in the third group showed improvements to their psychological well-being across the 14-days relative to the other groups.[34]

(a) What type of study is this?

(b) Identify the explanatory and response variables.

(c) Comment on whether the results of the study can be generalized to the population.

(d) Comment on whether the results of the study can be used to establish causal relationships.

(e) A newspaper article reporting on the study states, "The results of this study provide proof that giving young adults fresh fruits and vegetables to eat can have psychological benefits, even over a brief period of time." How would you suggest revising this statement so that it can be supported by the study?

[34] Tamlin S Conner i dr. "Let them eat fruit! The effect of fruit and vegetable consumption on psychological well-being in young adults: A randomized controlled trial". *PloS one* 12.2 (2017.), e0171206.

1.42 Screens, teens, and psychological well-being. In a study of three nationally representative large-scale data sets from Ireland, the United States, and the United Kingdom (n = 17,247), teenagers between the ages of 12 to 15 were asked to keep a diary of their screen time and answer questions about how they felt or acted. The answers to these questions were then used to compute a psychological well-being score. Additional data were collected and included in the analysis, such as each child's sex and age, and on the mother's education, ethnicity, psychological distress, and employment. The study concluded that there is little clear-cut evidence that screen time decreases adolescent well-being.[35]

(a) What type of study is this?

(b) Identify the explanatory variables.

(c) Identify the response variable.

(d) Comment on whether the results of the study can be generalized to the population, and why.

(e) Comment on whether the results of the study can be used to establish causal relationships.

1.43 Stanford Open Policing. The Stanford Open Policing project gathers, analyzes, and releases records from traffic stops by law enforcement agencies across the United States. Their goal is to help researchers, journalists, and policymakers investigate and improve interactions between police and the public.[36] The following is an excerpt from a summary table created based off of the data collected as part of this project.

County	State	Driver's race	No. of stops per year	% of stopped cars searched	% of stopped drivers arrested
Apaice County	Arizona	Black	266	0.08	0.02
Apaice County	Arizona	Hispanic	1008	0.05	0.02
Apaice County	Arizona	White	6322	0.02	0.01
Cochise County	Arizona	Black	1169	0.05	0.01
Cochise County	Arizona	Hispanic	9453	0.04	0.01
Cochise County	Arizona	White	10826	0.02	0.01
...	...	...	...	...	...
Wood County	Wisconsin	Black	16	0.24	0.10
Wood County	Wisconsin	Hispanic	27	0.04	0.03
Wood County	Wisconsin	White	1157	0.03	0.03

(a) What variables were collected on each individual traffic stop in order to create to the summary table above?

(b) State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.

(c) Suppose we wanted to evaluate whether vehicle search rates are different for drivers of different races. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

1.44 Space launches. The following summary table shows the number of space launches in the US by the type of launching agency and the outcome of the launch (success or failure).[37]

	1957 - 1999		2000 - 2018	
	Failure	Success	Failure	Success
Private	13	295	10	562
State	281	3751	33	711
Startup	-	-	5	65

(a) What variables were collected on each launch in order to create to the summary table above?

(b) State whether each variable is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is ordinal or not.

(c) Suppose we wanted to study how the success rate of launches vary between launching agencies and over time. In this analysis, which variable would be the response variable and which variable would be the explanatory variable?

[35] Amy Orben i AK Baukney-Przybylski. "Screens, Teens and Psychological Well-Being: Evidence from three time-use diary studies". *Psychological Science* (2018.).

[36] Emma Pierson i dr. "A large-scale analysis of racial disparities in police stops across the United States". *arXiv preprint arXiv:1706.05678* (2017.).

[37] JSR Launch Vehicle Database, A comprehensive list of suborbital space launches, 2019 Feb 10 Edition.

Prilog A

Exercise solutions

1 Introduction to data

1.1 (a) Treatment: $10/43 = 0.23 \rightarrow 23\%$. (b) Control: $2/46 = 0.04 \rightarrow 4\%$. (c) A higher percentage of patients in the treatment group were pain free 24 hours after receiving acupuncture. (d) It is possible that the observed difference between the two group percentages is due to chance.

1.3 (a) "Is there an association between air pollution exposure and preterm births?" (b) 143,196 births in Southern California between 1989 and 1993. (c) Measurements of carbon monoxide, nitrogen dioxide, ozone, and particulate matter less than $10\mu g/m^3$ (PM_{10}) collected at air-quality-monitoring stations as well as length of gestation. Continuous numerical variables.

1.5 (a) "Does explicitly telling children not to cheat affect their likelihood to cheat?". (b) 160 children between the ages of 5 and 15. (c) Four variables: (1) age (numerical, continuous), (2) sex (categorical), (3) whether they were an only child or not (categorical), (4) whether they cheated or not (categorical).

1.7 Explanatory: acupuncture or not. Response: if the patient was pain free or not.

1.9 (a) $50 \times 3 = 150$. (b) Four continuous numerical variables: sepal length, sepal width, petal length, and petal width. (c) One categorical variable, species, with three levels: *setosa*, *versicolor*, and *virginica*.

1.11 (a) Airport ownership status (public/private), airport usage status (public/private), latitude, and longitude. (b) Airport ownership status: categorical, not ordinal. Airport usage status: categorical, not ordinal. Latitude: numerical, continuous. Longitude: numerical, continuous.

1.13 (a) Population: all births, sample: 143,196 births between 1989 and 1993 in Southern California. (b) If births in this time span at the geography can be considered to be representative of all births, then the results are generalizable to the population of Southern California. However, since the study is observational the findings cannot be used to establish causal relationships.

1.15 (a) Population: all asthma patients aged 18-69 who rely on medication for asthma treatment. Sample: 600 such patients. (b) If the patients in this sample, who are likely not randomly sampled, can be considered to be representative of all asthma patients aged 18-69 who rely on medication for asthma treatment, then the results are generalizable to the population defined above. Additionally, since the study is experimental, the findings can be used to establish causal relationships.

1.17 (a) Observation. (b) Variable. (c) Sample statistic (mean). (d) Population parameter (mean).

1.19 (a) Observational. (b) Use stratified sampling to randomly sample a fixed number of students, say 10, from each section for a total sample size of 40 students.

1.21 (a) Positive, non-linear, somewhat strong. Countries in which a higher percentage of the population have access to the internet also tend to have higher average life expectancies, however rise in life expectancy trails off before around 80 years old. (b) Observational. (c) Wealth: countries with individuals who can widely afford the internet can probably also afford basic medical care. (Note: Answers may vary.)

1.23 (a) Simple random sampling is okay. In fact, it's rare for simple random sampling to not be a reasonable sampling method! (b) The student opinions may vary by field of study, so the stratifying by this variable makes sense and would be reasonable. (c) Students of similar ages are probably going to have more similar opinions, and we want clusters to be diverse with respect to the outcome of interest, so this would **not** be a good approach. (Additional thought: the clusters in this case may also have very different numbers of people, which can also create unexpected sample sizes.)

1.25 (a) The cases are 200 randomly sampled men and women. (b) The response variable is attitude towards a fictional microwave oven. (c) The explanatory variable is dispositional attitude. (d) Yes, the cases are sampled randomly. (e) This is an observational study since there is no random assignment to treatments. (f) No, we cannot establish a causal link between the explanatory and response variables since the study is observational. (g) Yes, the results of the study can be generalized to the population at large since the sample is random.

1.27 (a) Simple random sample. Non-response bias, if only those people who have strong opinions about the survey responds his sample may not be representative of the population. (b) Convenience sample. Under coverage bias, his sample may not be representative of the population since it consists only of his friends. It is also possible that the study will have non-response bias if some choose to not bring back the survey. (c) Convenience sample. This will have a similar issues to handing out surveys to friends. (d) Multi-stage sampling. If the classes are similar to each other with respect to student composition this approach should not introduce bias, other than potential non-response bias.

1.29 (a) Exam performance. (b) Light level: fluorescent overhead lighting, yellow overhead lighting, no overhead lighting (only desk lamps). (c) Sex: man, woman.

1.31 (a) Experiment. (b) Light level (overhead lighting, yellow overhead lighting, no overhead lighting) and noise level (no noise, construction noise, and human chatter noise). (c) Since the researchers want to ensure equal gender representation, sex will be a blocking variable.

1.33 Need randomization and blinding. One possible outline: (1) Prepare two cups for each participant, one containing regular Coke and the other containing Diet Coke. Make sure the cups are identical and contain equal amounts of soda. Label the cups A (regular) and B (diet). (Be sure to randomize A and B for each trial!) (2) Give each participant the two cups, one cup at a time, in random order, and ask the participant to record a value that indicates how much she liked the beverage. Be sure that neither the participant nor the person handing out the cups knows the identity of the beverage to make this a double- blind experiment. (Answers may vary.)

1.35 (a) Observational study. (b) Dog: Lucy. Cat: Luna. (c) Oliver and Lily. (d) Positive, as the popularity of a name for dogs increases, so does the popularity of that name for cats.

1.37 (a) Experiment. (b) Treatment: 25 grams of chia seeds twice a day, control: placebo. (c) Yes, gender. (d) Yes, single blind since the patients were blinded to the treatment they received. (e) Since this is an experiment, we can make a causal statement. However, since the sample is not random, the causal statement cannot be generalized to the population at large.

1.39 (a) Non-responders may have a different response to this question, e.g. parents who returned the surveys likely don't have difficulty spending time with their children. (b) It is unlikely that the women who were reached at the same address 3 years later are a random sample. These missing responders are probably renters (as opposed to homeowners) which means that they might be in a lower socio- economic status than the respondents. (c) There is no control group in this study, this is an observational study, and there may be confounding variables, e.g. these people may go running because they are generally healthier and/or do other exercises.

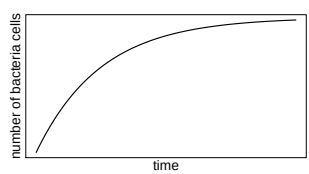
1.41 (a) Randomized controlled experiment. (b) Explanatory: treatment group (categorical, with 3 levels). Response variable: Psychological well-being. (c) No, because the participants were volunteers. (d) Yes, because it was an experiment. (e) The statement should say "evidence" instead of "proof".

1.43 (a) County, state, driver's race, whether the car was searched or not, and whether the driver was arrested or not. (b) All categorical, non-ordinal. (c) Response: whether the car was searched or not. Explanatory: race of the driver.

2 Summarizing data

?? (a) Positive association: mammals with longer gestation periods tend to live longer as well. (b) Association would still be positive. (c) No, they are not independent. See part (a).

?? The graph below shows a ramp up period. There may also be a period of exponential growth at the start before the size of the petri dish becomes a factor in slowing growth.



?? (a) Population mean, $\mu_{2007} = 52$; sample mean, $\bar{x}_{2008} = 58$. (b) Population mean, $\mu_{2001} = 3.37$; sample mean, $\bar{x}_{2012} = 3.59$.

?? Any 10 employees whose average number of days off is between the minimum and the mean number of days off for the entire workforce at this plant.

?? (a) Dist 2 has a higher mean since $20 > 13$, and a higher standard deviation since 20 is further from the rest of the data than 13. (b) Dist 1 has a higher mean since $-20 > -40$, and Dist 2 has a higher standard deviation since -40 is farther away from the rest of the data than -20. (c) Dist 2 has a higher mean since all values in this distribution are higher than those in Dist 1, but both distribution have the same standard deviation since they are equally variable around their respective means. (d) Both distributions have the same mean since they're both centered at 300, but Dist 2 has a higher standard deviation since the observations are farther from the mean than in Dist 1.

?? (a) About 30. (b) Since the distribution is right skewed the mean is higher than the median. (c) Q1: between 15 and 20, Q3: between 35 and 40, IQR: about 20. (d) Values that are considered to be unusually low or high lie more than 1.5×IQR away from the quartiles. Upper fence: Q3 + 1.5 × IQR = $37.5 + 1.5 \times 20 = 67.5$; Lower fence: Q1 - 1.5 × IQR = $17.5 - 1.5 \times 20 = -12.5$; The lowest AQI recorded is not lower than 5 and the highest AQI recorded is not higher than 65, which are both within the fences. Therefore none of the days in this sample would be considered to have an unusually low or high AQI.

?? The histogram shows that the distribution is bimodal, which is not apparent in the box plot. The box plot makes it easy to identify more precise values of observations outside of the whiskers.

?? (a) The distribution of number of pets per household is likely right skewed as there is a natural boundary at 0 and only a few people have many pets. Therefore the center would be best described by the median, and variability would be best described by the IQR. (b) The distribution of number of distance to work is likely right skewed as there is a natural boundary at 0 and only a few people live a very long distance from work. Therefore the center would be best described by the median, and variability would be best described by the IQR. (c) The distribution of heights of males is likely symmetric. Therefore the center would be best described by the mean, and variability would be best described by the standard deviation.

?? (a) The median is a much better measure of the typical amount earned by these 42 people. The mean is much higher than the income of 40 of the 42 people. This is because the mean is an arithmetic average and gets affected by the two extreme observations. The median does not get effected as much since it is robust to outliers. (b) The IQR is a much better measure of variability in the amounts earned by nearly all of the 42 people. The standard deviation gets affected greatly by the two high salaries, but the IQR is robust to these extreme observations.

?? (a) The distribution is unimodal and symmetric with a mean of about 25 minutes and a standard deviation of about 5 minutes. There does not appear to be any counties with unusually high or low mean travel times. Since the distribution is already unimodal and symmetric, a log transformation is not necessary. (b) Answers will vary. There are pockets of longer travel time around DC, Southeastern NY, Chicago, Minneapolis, Los Angeles, and many other big cities. There is also a large section of shorter average commute times that overlap with farmland in the Midwest. Many farmers' homes are adjacent to their farmland, so their commute would be brief, which may explain why the average commute time for these counties is relatively low.

?? (a) We see the order of the categories and the relative frequencies in the bar plot. (b) There are no features that are apparent in the pie chart but not in the bar plot. (c) We usually prefer to use a bar plot as we can also see the relative frequencies of the categories in this graph.

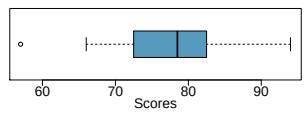
?? The vertical locations at which the ideological groups break into the Yes, No, and Not Sure categories differ, which indicates that likelihood of supporting the DREAM act varies by political ideology. This suggests that the two variables may be dependent.

?? (a) (i) False. Instead of comparing counts, we should compare percentages of people in each group who suffered cardiovascular problems. (ii) True. (iii) False. Association does not imply causation. We cannot infer a causal relationship based on an observational study. The difference from part (ii) is subtle. (iv) True.
(b) Proportion of all patients who had cardiovascular problems: $\frac{7,979}{227,571} \approx 0.035$
(c) The expected number of heart attacks in the rosiglitazone group, if having cardiovascular problems and treatment were independent, can be calculated as the number of patients in that group multiplied by the overall cardiovascular problem rate in the study: $67,593 * \frac{7,979}{227,571} \approx 2370$.
(d) (i) H_0: The treatment and cardiovascular problems are independent. They have no relationship, and the difference in incidence rates between the rosiglitazone and pioglitazone groups is due to chance. H_A: The treatment and cardiovascular problems are not independent. The difference in the incidence rates between the rosiglitazone and pioglitazone groups is not due to chance and rosiglitazone is associated with an increased risk of serious cardiovascular problems. (ii) A higher number of patients with cardiovascular problems than expected under the assumption of independence would provide support for the alternative hypothesis as this would suggest that rosiglitazone increases the risk of such problems. (iii) In the actual study, we observed 2,593 cardiovascular events in the rosiglitazone group. In the 1,000 simulations under the independence model, we observed somewhat less than 2,593 in every single simulation, which suggests that the actual results did not come from the independence model. That is, the variables do not appear to be independent, and we reject the independence model in favor of the alternative. The study's results provide convincing evidence that rosiglitazone is associated with an increased risk of cardiovascular problems.

?? (a) Decrease: the new score is smaller than the mean of the 24 previous scores. (b) Calculate a weighted mean. Use a weight of 24 for the old mean and 1 for the new mean: $(24 \times 74 + 1 \times 64)/(24 + 1) = 73.6$.
(c) The new score is more than 1 standard deviation away from the previous mean, so increase.

?? No, we would expect this distribution to be right skewed. There are two reasons for this: (1) there is a natural boundary at 0 (it is not possible to watch less than 0 hours of TV), (2) the standard deviation of the distribution is very large compared to the mean.

?? The distribution of ages of best actress winners are right skewed with a median around 30 years. The distribution of ages of best actor winners is also right skewed, though less so, with a median around 40 years. The difference between the peaks of these distributions suggest that best actress winners are typically younger than best actor winners. The ages of best actress winners are more variable than the ages of best actor winners. There are potential outliers on the higher end of both of the distributions.

?? 

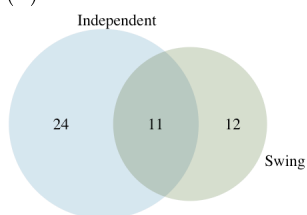
3 Probability

?? (a) False. These are independent trials. (b) False. There are red face cards. (c) True. A card cannot be both a face card and an ace.

?? (a) 10 tosses. Fewer tosses mean more variability in the sample fraction of heads, meaning there's a better chance of getting at least 60% heads. (b) 100 tosses. More flips means the observed proportion of heads would often be closer to the average, 0.50, and therefore also above 0.40. (c) 100 tosses. With more flips, the observed proportion of heads would often be closer to the average, 0.50. (d) 10 tosses. Fewer flips would increase variability in the fraction of tosses that are heads.

?? (a) $0.5^{10} = 0.00098$. (b) $0.5^{10} = 0.00098$. (c) $P(\text{at least one tails}) = 1 - P(\text{no tails}) = 1 - (0.5^{10}) \approx 1 - 0.001 = 0.999$.

?? (a) No, there are voters who are both independent and swing voters.
(b)



(c) Each Independent voter is either a swing voter or not. Since 35% of voters are Independents and 11% are both Independent and swing voters, the other 24% must not be swing voters. (d) 0.47. (e) 0.53. (f) P(Independent) $\times$ P(swing) $= 0.35 \times 0.23 = 0.08$, which does not equal P(Independent and swing) $=$ 0.11, so the events are dependent.

?? (a) If the class is not graded on a curve, they are independent. If graded on a curve, then neither independent nor disjoint – unless the instructor will only give one A, which is a situation we will ignore in parts (b) and (c). (b) They are probably not independent: if you study together, your study habits would be related, which suggests your course performances are also related. (c) No. See the answer to part (a) when the course is not graded on a curve. More generally: if two things are unrelated (independent), then one occurring does not preclude the other from occurring.

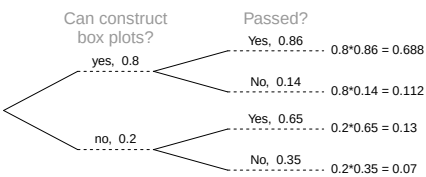
?? (a) $0.16 + 0.09 = 0.25$. (b) $0.17 + 0.09 = 0.26$. (c) Assuming that the education level of the husband and wife are independent: $0.25 \times 0.26 = 0.065$. You might also notice we actually made a second assumption: that the decision to get married is unrelated to education level. (d) The husband/wife independence assumption is probably not reasonable, because people often marry another person with a comparable level of education. We will leave it to you to think about whether the second assumption noted in part (c) is reasonable.

?? (a) No, but we could if A and B are independent. (b-i) 0.21. (b-ii) 0.79. (b-iii) 0.3. (c) No, because $0.1 \neq 0.21$, where 0.21 was the value computed under independence from part (a). (d) 0.143.

?? (a) No, 0.18 of respondents fall into this combination. (b) $0.60+0.20-0.18 = 0.62$. (c) $0.18/0.20 = 0.9$. (d) $0.11/0.33 \approx 0.33$. (e) No, otherwise the answers to (c) and (d) would be the same. (f) $0.06/0.34 \approx 0.18$.

?? (a) No. There are 6 females who like Five Guys Burgers. (b) $162/248 = 0.65$. (c) $181/252 = 0.72$. (d) Under the assumption of a dating choices being independent of hamburger preference, which on the surface seems reasonable: $0.65 \times 0.72 = 0.468$. (e) $(252 + 6 - 1)/500 = 0.514$.

?? (a)



(b) 0.84

?? 0.0714. Even when a patient tests positive for lupus, there is only a 7.14% chance that he actually has lupus. House may be right.

?? (a) 0.3. (b) 0.3. (c) 0.3. (d) $0.3 \times 0.3 = 0.09$. (e) Yes, the population that is being sampled from is identical in each draw.

?? (a) $2/9 \approx 0.22$. (b) $3/9 \approx 0.33$. (c) $\frac{3}{10} \times \frac{2}{9} \approx 0.067$. (d) No, e.g. in this exercise, removing one chip meaningfully changes the probability of what might be drawn next.

?? $P(^1\text{leggings}, {}^2\text{jeans}, {}^3\text{jeans}) = \frac{5}{24} \times \frac{7}{23} \times \frac{6}{22} = 0.0173$. However, the person with leggings could have come 2nd or 3rd, and these each have this same probability, so $3 \times 0.0173 = 0.0519$.

?? (a) 13. (b) No, these 27 students are not a random sample from the university's student population. For example, it might be argued that the proportion of smokers among students who go to the gym at 9 am on a Saturday morning would be lower than the proportion of smokers in the university as a whole.

?? (a) E(X) = 3.59. SD(X) = 9.64. (b) E(X) = -1.41. SD(X) = 9.64. (c) No, the expected net profit is negative, so on average you expect to lose money.

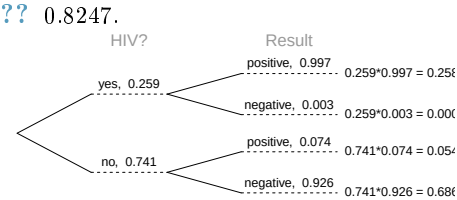
?? 5% increase in value.

?? E = -0.0526. SD = 0.9986.

?? Approximate answers are OK.
(a) $(29 + 32)/144 = 0.42$. (b) $21/144 = 0.15$. (c) $(26 + 12 + 15)/144 = 0.37$.

?? (a) Invalid. Sum is greater than 1. (b) Valid. Probabilities are between 0 and 1, and they sum to 1. In this class, every student gets a C. (c) Invalid. Sum is less than 1. (d) Invalid. There is a negative probability. (e) Valid. Probabilities are between 0 and 1, and they sum to 1. (f) Invalid. There is a negative probability.

?? 0.8247.

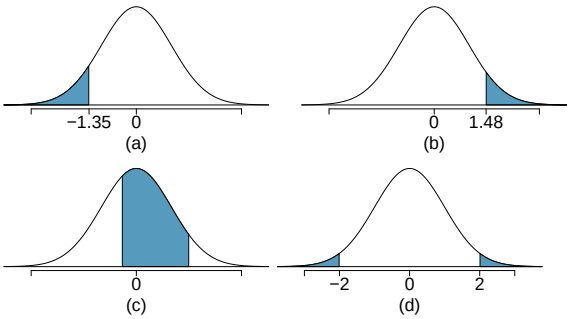


?? (a) E = \$3.90. SD = \$0.34.
(b) E = \$27.30. SD = \$0.89.

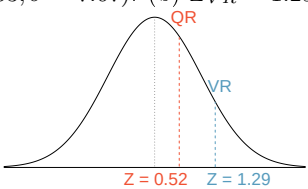
?? $Var\left(\frac{X_1+X_2}{2}\right)$
$= Var\left(\frac{X_1}{2} + \frac{X_2}{2}\right)$
$= \frac{Var(X_1)}{2^2} + \frac{Var(X_2)}{2^2}$
$= \frac{\sigma^2}{4} + \frac{\sigma^2}{4}$
$= \sigma^2/2$

?? $Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)$
$= Var\left(\frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n}\right)$
$= \frac{Var(X_1)}{n^2} + \frac{Var(X_2)}{n^2} + \cdots + \frac{Var(X_n)}{n^2}$
$= \frac{\sigma^2}{n^2} + \frac{\sigma^2}{n^2} + \cdots + \frac{\sigma^2}{n^2}$ (there are n of these terms)
$= n\frac{\sigma^2}{n^2}$
$= \sigma^2/n$

?? (a) 8.85%. (b) 6.94%. (c) 58.86%. (d) 4.56%.



?? (a) Verbal: $N(\mu = 151, \sigma = 7)$, Quant: $N(\mu = 153, \sigma = 7.67)$. (b) $Z_{VR} = 1.29$, $Z_{QR} = 0.52$.



(c) She scored 1.29 standard deviations above the mean on the Verbal Reasoning section and 0.52 standard deviations above the mean on the Quantitative Reasoning section. (d) She did better on the Verbal Reasoning section since her Z-score on that section was higher. (e) $Perc_{VR} = 0.9007 \approx 90\%$, $Perc_{QR} = 0.6990 \approx 70\%$. (f) $100\% - 90\% = 10\%$ did better than her on VR, and $100\% - 70\% = 30\%$ did better than her on QR. (g) We cannot compare the raw scores since they are on different scales. Comparing her percentile scores is more appropriate when comparing her performance to others. (h) Answer to part (b) would not change as Z-scores can be calculated for distributions that are not normal. However, we could not answer parts (d)-(f) since we cannot use the normal probability table to calculate probabilities and percentiles without a normal model.

?? (a) $Z = 0.84$, which corresponds to approximately 159 on QR. (b) $Z = -0.52$, which corresponds to approximately 147 on VR.

?? (a) $Z = 1.2$, $P(Z > 1.2) = 0.1151$.
(b) $Z = -1.28 \rightarrow 70.6°$F or colder.

?? (a) $N(25, 2.78)$. (b) $Z = 1.08$, $P(Z > 1.08) = 0.1401$. (c) The answers are very close because only the units were changed. (The only reason why they differ at all because 28°C is 82.4°F, not precisely 83°F.) (d) Since $IQR = Q3 - Q1$, we first need to find $Q3$ and $Q1$ and take the difference between the two. Remember that $Q3$ is the 75^{th} and $Q1$ is the 25^{th} percentile of a distribution. Q1 = 23.13, Q3 = 26.86, IQR = 26. 86 - 23.13 = 3.73.

?? (a) No. The cards are not independent. For example, if the first card is an ace of clubs, that implies the second card cannot be an ace of clubs. Additionally, there are many possible categories, which would need to be simplified. (b) No. There are six events under consideration. The Bernoulli distribution allows for only two events or categories. Note that rolling a die could be a Bernoulli trial if we simply to two events, e.g. rolling a 6 and not rolling a 6, though specifying such details would be necessary.

?? (a) $0.875^2 \times 0.125 = 0.096$. (b) $\mu = 8$, $\sigma = 7.48$.

?? If p is the probability of a success, then the mean of a Bernoulli random variable X is given by
$\mu = E[X] = P(X = 0) \times 0 + P(X = 1) \times 1$
$= (1 - p) \times 0 + p \times 1 = 0 + p = p$

?? (a) Binomial conditions are met: (1) Independent trials: In a random sample, whether or not one 18-20 year old has consumed alcohol does not depend on whether or not another one has. (2) Fixed number of trials: $n = 10$. (3) Only two outcomes at each trial: Consumed or did not consume alcohol. (4) Probability of a success is the same for each trial: $p = 0.697$. (b) 0.203. (c) 0.203. (d) 0.167. (e) 0.997.

?? (a) $\mu = 35$, $\sigma = 3.24$ (b) $Z = \frac{45-35}{3.24} = 3.09$. 45 is more than 3 standard deviations away from the mean, we can assume that it is an unusual observation. Therefore yes, we would be surprised. (c) Using the normal approximation, 0.0010. With 0.5 correction, 0.0017.

?? (a) $1 - 0.75^3 = 0.5781$. (b) 0.1406. (c) 0.4219. (d) $1 - 0.25^3 = 0.9844$.

?? (a) Geometric distribution: 0.109. (b) Binomial: 0.219. (c) Binomial: 0.137. (d) $1 - 0.875^6 = 0.551$. (e) Geometric: 0.084. (f) Using a binomial distribution with $n = 6$ and $p = 0.75$, we see that $\mu = 4.5$, $\sigma = 1.06$, and $Z = 2.36$. Since this is not within 2 SD, it may be considered unusual.

?? (a) $\overset{Anna}{1/5} \times \overset{Ben}{1/4} \times \overset{Carl}{1/3} \times \overset{Damian}{1/2} \times \overset{Eddy}{1/1} = 1/5! = 1/120$. (b) Since the probabilities must add to 1, there must be $5! = 120$ possible orderings. (c) $8! = 40{,}320$.

?? (a) 0.0804. (b) 0.0322. (c) 0.0193.

?? (a) Negative binomial with $n = 4$ and $p = 0.55$, where a success is defined here as a female student. The negative binomial setting is appropriate since the last trial is fixed but the order of the first 3 trials is unknown. (b) 0.1838. (c) $\binom{3}{1} = 3$. (d) In the binomial model there are no restrictions on the outcome of the last trial. In the negative binomial model the last trial is fixed. Therefore we are interested in the number of ways of orderings of the other $k - 1$ successes in the first $n - 1$ trials.

?? (a) Poisson with $\lambda = 75$. (b) $\mu = \lambda = 75$, $\sigma = \sqrt{\lambda} = 8.66$. (c) $Z = -1.73$. Since 60 is within 2 standard deviations of the mean, it would not generally be considered unusual. Note that we often use this rule of thumb even when the normal model does not apply. (d) Using Poisson with $\lambda = 75$: 0.0402.

?? (a) $\frac{\lambda^k \times e^{-\lambda}}{k!} = \frac{6.5^5 \times e^{-6.5}}{5!} = 0.1454$
(b) The probability will come to $0.0015 + 0.0098 + 0.0318 = 0.0431$ (0.0430 if no rounding error).
(c) The number of people per car is $11.7/6.5 = 1.8$, meaning people are coming in small clusters. That is, if one person arrives, there's a chance that they brought one or more other people in their vehicle. This means individuals (the people) are not independent, even if the car arrivals are independent, and this breaks a core assumption for the Poisson distribution. That is, the number of people visiting between 2pm and 3pm would not follow a Poisson distribution.

?? 0 wins (-\$3): 0.1458. 1 win (-\$1): 0.3936. 2 wins (+\$1): 0.3543. 3 wins (+\$3): 0.1063.

?? Want to find the probability that there will be 1,787 or more enrollees. Using the normal approximation, with $\mu = np = 2,500 \times 0.7 = 1750$ and $\sigma = \sqrt{np(1-p)} = \sqrt{2,500 \times 0.7 \times 0.3} \approx 23$, $Z = 1.61$, and $P(Z > 1.61) = 0.0537$. With a 0.5 correction: 0.0559.

?? (a) $Z = 0.67$. (b) $\mu = \$1650$, $x = \$1800$. (c) $0.67 = \frac{1800-1650}{\sigma} \rightarrow \sigma = \$223.88$.

?? (a) $(1 - 0.471)^2 \times 0.471 = 0.1318$. (b) $0.471^3 = 0.1045$. (c) $\mu = 1/0.471 = 2.12$, $\sigma = \sqrt{2.38} = 1.54$. (d) $\mu = 1/0.30 = 3.33$, $\sigma = 2.79$. (e) When p is smaller, the event is rarer, meaning the expected number of trials before a success and the standard deviation of the waiting time are higher.

?? $Z = 1.56$, $P(Z > 1.56) = 0.0594$, i.e. 6%.

?? (a) $Z = 0.73$, $P(Z > 0.73) = 0.2327$. (b) If you are bidding on only one auction and set a low maximum bid price, someone will probably outbid you. If you set a high maximum bid price, you may win the auction but pay more than is necessary. If bidding on more than one auction, and you set your maximum bid price very low, you probably won't win any of the auctions. However, if the maximum bid price is even modestly high, you are likely to win multiple auctions. (c) An answer roughly equal to the 10th percentile would be reasonable. Regrettably, no percentile cutoff point guarantees beyond any possible event that you win at least one auction. However, you may pick a higher percentile if you want to be more sure of winning an auction. (d) Answers will vary a little but should correspond to the answer in part (c). We use the 10^{th} percentile: $Z = -1.28 \rightarrow \$69.80$.

?? (a) $Z = 3.5$, upper tail is 0.0002. (More precise value: 0.000233, but we'll use 0.0002 for the calculations here.)
(b) $0.0002 \times 2000 = 0.4$. We would expect about 0.4 10 year olds who are 76 inches or taller to show up.
(c) $\binom{2000}{0}(0.0002)^0(1 - 0.0002)^{2000} = 0.67029$.
(d) $\frac{0.4^0 \times e^{-0.4}}{0!} = \frac{1 \times e^{-0.4}}{1} = 0.67032$.

5 Foundations for inference

?? (a) Mean. Each student reports a numerical value: a number of hours. (b) Mean. Each student reports a number, which is a percentage, and we can average over these percentages. (c) Proportion. Each student reports Yes or No, so this is a categorical variable and we use a proportion. (d) Mean. Each student reports a number, which is a percentage like in part (b). (e) Proportion. Each student reports whether or not s/he expects to get a job, so this is a categorical variable and we use a proportion.

?? (a) The sample is from all computer chips manufactured at the factory during the week of production. We might be tempted to generalize the population to represent all weeks, but we should exercise caution here since the rate of defects may change over time. (b) The fraction of computer chips manufactured at the factory during the week of production that had defects. (c) Estimate the parameter using the data: $\hat{p} = \frac{27}{212} = 0.127$. (d) *Standard error* (or SE). (e) Compute the SE using $\hat{p} = 0.127$ in place of p: $SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.127(1-0.127)}{212}} = 0.023$. (f) The standard error is the standard deviation of $\hat{p}$. A value of 0.10 would be about one standard error away from the observed value, which would not represent a very uncommon deviation. (Usually beyond about 2 standard errors is a good rule of thumb.) The engineer should not be surprised. (g) Recomputed standard error using $p = 0.1$: $SE = \sqrt{\frac{0.1(1-0.1)}{212}} = 0.021$. This value isn't very different, which is typical when the standard error is computed using relatively similar proportions (and even sometimes when those proportions are quite different!).

?? (a) Sampling distribution. (b) If the population proportion is in the 5-30% range, the success-failure condition would be satisfied and the sampling distribution would be symmetric. (c) We use the formula for the standard error: $SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.08(1-0.08)}{800}} = 0.0096$. (d) Standard error. (e) The distribution will tend to be more variable when we have fewer observations per sample.

?? Recall that the general formula is *point estimate* $\pm\ z^\star \times SE$. First, identify the three different values. The point estimate is 45%, $z^\star = 1.96$ for a 95% confidence level, and $SE = 1.2\%$. Then, plug the values into the formula: $45\% \pm 1.96 \times 1.2\% \quad \rightarrow \quad (42.6\%, 47.4\%)$ We are 95% confident that the proportion of US adults who live with one or more chronic conditions is between 42.6% and 47.4%.

?? (a) False. Confidence intervals provide a range of plausible values, and sometimes the truth is missed. A 95% confidence interval "misses" about 5% of the time. (b) True. Notice that the description focuses on the true population value. (c) True. If we examine the 95% confidence interval computed in Exercise **??**, we can see that 50% is not included in this interval. This means that in a hypothesis test, we would reject the null hypothesis that the proportion is 0.5. (d) False. The standard error describes the uncertainty in the overall estimate from natural fluctuations due to randomness, not the uncertainty corresponding to individuals' responses.

?? (a) False. The point estimate is always in the confidence interval, and this is a non-sensical use of a confidence interval with a point estimate (because the point estimate is, by design, listed within the confidence interval). (b) True. (c) False. The confidence interval is not about a sample mean. (d) False. To be more confident that we capture the parameter, we need a wider interval. Think about needing a bigger net to be more sure of catching a fish in a murky lake. (e) True. Optional explanation: This is true since the normal model was used to model the sample mean. The margin of error is half the width of the interval, and the sample mean is the midpoint of the interval. (f) False. In the calculation of the standard error, we divide the standard deviation by the square root of the sample size. To cut the SE (or margin of error) in half, we would need to sample $2^2 = 4$ times the number of people in the initial sample.

?? (a) The visitors are from a simple random sample, so independence is satisfied. The success-failure condition is also satisfied, with both 64 and $752 - 64 = 688$ above 10. Therefore, we can use a normal distribution to model $\hat{p}$ and construct a confidence interval. (b) The sample proportion is $\hat{p} = \frac{64}{752} = 0.085$. The standard error is

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$
$$= \sqrt{\frac{0.085(1-0.085)}{752}} = 0.010$$

(c) For a 90% confidence interval, use $z^\star = 1.65$. The confidence interval is $0.085 \pm 1.65 \times 0.010 \rightarrow (0.0685, 0.1015)$. We are 90% confident that 6.85% to 10.15% of first-time site visitors will register using the new design.

?? (a) $H_0 : p = 0.5$ (Neither a majority nor minority of students' grades improved) $H_A : p \neq 0.5$ (Either a majority or a minority of students' grades improved) (b) $H_0 : \mu = 15$ (The average amount of company time each employee spends not working is 15 minutes for March Madness.) $H_A : \mu \neq 15$ (The average amount of company time each employee spends not working is different than 15 minutes for March Madness.)

?? (1) The hypotheses should be about the population proportion (p), not the sample proportion. (2) The null hypothesis should have an equal sign. (3) The alternative hypothesis should have a not-equals sign, and (4) it should reference the null value, $p_0 = 0.6$, not the observed sample proportion. The correct way to set up these hypotheses is: $H_0 : p = 0.6$ and $H_A : p \neq 0.6$.

?? (a) This claim is reasonable, since the entire interval lies above 50%. (b) The value of 70% lies outside of the interval, so we have convincing evidence that the researcher's conjecture is wrong. (c) A 90% confidence interval will be narrower than a 95% confidence interval. Even without calculating the interval, we can tell that 70% would not fall in the interval, and we would reject the researcher's conjecture based on a 90% confidence level as well.

?? (i) Set up hypotheses. H_0: $p = 0.5$, H_A: $p \neq 0.5$. We will use a significance level of $\alpha = 0.05$. (ii) Check conditions: simple random sample gets us independence, and the success-failure conditions is satisfied since $0.5 \times 1000 = 500$ for each group is at least 10. (iii) Next, we calculate: $SE = \sqrt{0.5(1-0.5)/1000} = 0.016$. $Z = \frac{0.42-0.5}{0.016} = -5$, which has a one-tail area of about 0.0000003, so the p-value is twice this one-tail area at 0.0000006. (iv) Make a conclusion: Because the p-value is less than $\alpha = 0.05$, we reject the null hypothesis and conclude that the fraction of US adults who believe raising the minimum wage will help the economy is not 50%. Because the observed value is less than 50% and we have rejected the null hypothesis, we can conclude that this belief is held by fewer than 50% of US adults. (For reference, the survey also explores support for changing the minimum wage, which is a different question than if it will help the economy.)

?? If the p-value is 0.05, this means the test statistic would be either $Z = -1.96$ or $Z = 1.96$. We'll show the calculations for $Z = 1.96$. Standard error: $SE = \sqrt{0.3(1-0.3)/90} = 0.048$. Finally, set up the test statistic formula and solve for $\hat{p}$: $1.96 = \frac{\hat{p}-0.3}{0.048} \rightarrow \hat{p} = 0.394$ Alternatively, if $Z = -1.96$ was used: $\hat{p} = 0.206$.

?? (a) H_0: Anti-depressants do not affect the symptoms of Fibromyalgia. H_A: Anti-depressants do affect the symptoms of Fibromyalgia (either helping or harming). (b) Concluding that anti-depressants either help or worsen Fibromyalgia symptoms when they actually do neither. (c) Concluding that anti-depressants do not affect Fibromyalgia symptoms when they actually do.

?? (a) We are 95% confident that Americans spend an average of 1.38 to 1.92 hours per day relaxing or pursuing activities they enjoy. (b) Their confidence level must be higher as the width of the confidence interval increases as the confidence level increases. (c) The new margin of error will be smaller, since as the sample size increases, the standard error decreases, which will decrease the margin of error.

?? (a) H_0: The restaurant meets food safety and sanitation regulations. H_A: The restaurant does not meet food safety and sanitation regulations. (b) The food safety inspector concludes that the restaurant does not meet food safety and sanitation regulations and shuts down the restaurant when the restaurant is actually safe. (c) The food safety inspector concludes that the restaurant meets food safety and sanitation regulations and the restaurant stays open when the restaurant is actually not safe. (d) A Type 1 Error may be more problematic for the restaurant owner since his restaurant gets shut down even though it meets the food safety and sanitation regulations. (e) A Type 2 Error may be more problematic for diners since the restaurant deemed safe by the inspector is actually not. (f) Strong evidence. Diners would rather a restaurant that meet the regulations get shut down than a restaurant that doesn't meet the regulations not get shut down.

?? (a) $H_0 : p_{unemp} = p_{underemp}$: The proportions of unemployed and underemployed people who are having relationship problems are equal. $H_A : p_{unemp} \neq p_{underemp}$: The proportions of unemployed and underemployed people who are having relationship problems are different. (b) If in fact the two population proportions are equal, the probability of observing at least a 2% difference between the sample proportions is approximately 0.35. Since this is a high probability we fail to reject the null hypothesis. The data do not provide convincing evidence that the proportion of of unemployed and underemployed people who are having relationship problems are different.

?? Because 130 is inside the confidence interval, we do not have convincing evidence that the true average is any different than what the nutrition label suggests.

?? True. If the sample size gets ever larger, then the standard error will become ever smaller. Eventually, when the sample size is large enough and the standard error is tiny, we can find statistically significant yet very small differences between the null value and point estimate (assuming they are not exactly equal).

?? (a) In effect, we're checking whether men are paid more than women (or vice-versa), and we'd expect these outcomes with either chance under the null hypothesis:

$$H_0 : p = 0.5 \qquad H_A : p \neq 0.5$$

We'll use p to represent the fraction of cases where men are paid more than women.
(b) Below is the completion of the hypothesis test.

- There isn't a good way to check independence here since the jobs are not a simple random sample. However, independence doesn't seem unreasonable, since the individuals in each job are different from each other. The success-failure condition is met since we check it using the null proportion: $p_0 n = (1 - p_0)n = 10.5$ is greater than 10.
- We can compute the sample proportion, SE, and test statistic:

$$\hat{p} = 19/21 = 0.905$$
$$SE = \sqrt{\frac{0.5 \times (1 - 0.5)}{21}} = 0.109$$
$$Z = \frac{0.905 - 0.5}{0.109} = 3.72$$

  The test statistic Z corresponds to an upper tail area of about 0.0001, so the p-value is 2 times this value: 0.0002.
- Because the p-value is smaller than 0.05, we reject the notion that all these gender pay disparities are due to chance. Because we observe that men are paid more in a higher proportion of cases and we have rejected H_0, we can conclude that men are being paid higher amounts in ways not explainable by chance alone.

If you're curious for more info around this topic, including a discussion about adjusting for additional factors that affect pay, please see the following video by Healthcare Triage: youtu.be/aVhgKSULNQA.

6 Inference for categorical data

?? (a) False. Doesn't satisfy success-failure condition. (b) True. The success-failure condition is not satisfied. In most samples we would expect $\hat{p}$ to be close to 0.08, the true population proportion. While $\hat{p}$ can be much above 0.08, it is bound below by 0, suggesting it would take on a right skewed shape. Plotting the sampling distribution would confirm this suspicion. (c) False. $SE_{\hat{p}} = 0.0243$, and $\hat{p} = 0.12$ is only $\frac{0.12-0.08}{0.0243} = 1.65$ SEs away from the mean, which would not be considered unusual. (d) True. $\hat{p} = 0.12$ is 2.32 standard errors away from the mean, which is often considered unusual. (e) False. Decreases the SE by a factor of $1/\sqrt{2}$.

?? (a) True. See the reasoning of 6.1(b). (b) True. We take the square root of the sample size in the SE formula. (c) True. The independence and success-failure conditions are satisfied. (d) True. The independence and success-failure conditions are satisfied.

?? (a) False. A confidence interval is constructed to estimate the population proportion, not the sample proportion. (b) True. 95% CI: 82% ± 2%. (c) True. By the definition of the confidence level. (d) True. Quadrupling the sample size decreases the SE and ME by a factor of $1/\sqrt{4}$. (e) True. The 95% CI is entirely above 50%.

?? With a random sample, independence is satisfied. The success-failure condition is also satisfied. $ME = z^\star\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.56\times0.44}{600}} = 0.0397 \approx 4\%$

?? (a) No. The sample only represents students who took the SAT, and this was also an online survey. (b) (0.5289, 0.5711). We are 90% confident that 53% to 57% of high school seniors who took the SAT are fairly certain that they will participate in a study abroad program in college. (c) 90% of such random samples would produce a 90% confidence interval that includes the true proportion. (d) Yes. The interval lies entirely above 50%.

?? (a) We want to check for a majority (or minority), so we use the following hypotheses:

$$H_0 : p = 0.5 \qquad H_A : p \neq 0.5$$

We have a sample proportion of $\hat{p} = 0.55$ and a sample size of $n = 617$ independents.
Since this is a random sample, independence is satisfied. The success-failure condition is also satisfied: 617×0.5 and $617 \times (1 - 0.5)$ are both at least 10 (we use the null proportion $p_0 = 0.5$ for this check in a one-proportion hypothesis test).
Therefore, we can model $\hat{p}$ using a normal distribution with a standard error of

$$SE = \sqrt{\frac{p(1-p)}{n}} = 0.02$$

(We use the null proportion $p_0 = 0.5$ to compute the standard error for a one-proportion hypothesis test.)
Next, we compute the test statistic:

$$Z = \frac{0.55 - 0.5}{0.02} = 2.5$$

This yields a one-tail area of 0.0062, and a p-value of $2 \times 0.0062 = 0.0124$.
Because the p-value is smaller than 0.05, we reject the null hypothesis. We have strong evidence that the support is different from 0.5, and since the data provide a point estimate above 0.5, we have strong evidence to support this claim by the TV pundit.
(b) No. Generally we expect a hypothesis test and a confidence interval to align, so we would expect the confidence interval to show a range of plausible values entirely above 0.5. However, if the confidence level is misaligned (e.g. a 99% confidence level and a $\alpha = 0.05$ significance level), then this is no longer generally true.

?? (a) $H_0 : p = 0.5$. $H_A : p \neq 0.5$. Independence (random sample) is satisfied, as is the success-failure conditions (using $p_0 = 0.5$, we expect 40 successes and 40 failures). $Z = 2.91 \rightarrow$ the one tail area is 0.0018, so the p-value is 0.0036. Since the p-value < 0.05, we reject the null hypothesis. Since we rejected H_0 and the point estimate suggests people are better than random guessing, we can conclude the rate of correctly identifying a soda for these people is significantly better than just by random guessing. (b) If in fact people cannot tell the difference between diet and regular soda and they were randomly guessing, the probability of getting a random sample of 80 people where 53 or more identify a soda correctly (or 53 or more identify a soda incorrectly) would be 0.0036.

?? Since a sample proportion ($\hat{p} = 0.55$) is available, we use this for the sample size calculations. The margin of error for a 90% confidence interval is $1.65 \times SE = 1.65 \times \sqrt{\frac{p(1-p)}{n}}$. We want this to be less than 0.01, where we use $\hat{p}$ in place of p:

$$1.65 \times \sqrt{\frac{0.55(1-0.55)}{n}} \leq 0.01$$
$$1.65^2 \frac{0.55(1-0.55)}{0.01^2} \leq n$$

From this, we get that n must be at least 6739.

?? This is not a randomized experiment, and it is unclear whether people would be affected by the behavior of their peers. That is, independence may not hold. Additionally, there are only 5 interventions under the provocative scenario, so the success-failure condition does not hold. Even if we consider a hypothesis test where we pool the proportions, the success-failure condition will not be satisfied. Since one condition is questionable and the other is not satisfied, the difference in sample proportions will not follow a nearly normal distribution.

?? (a) False. The entire confidence interval is above 0. (b) True. (c) True. (d) True. (e) False. It is simply the negated and reordered values: (-0.06,-0.02).

?? (a) Standard error:

$$SE = \sqrt{\frac{0.79(1-0.79)}{347} + \frac{0.55(1-0.55)}{617}} = 0.03$$

Using $z^\star = 1.96$, we get:

$$0.79 - 0.55 \pm 1.96 \times 0.03 \rightarrow (0.181, 0.299)$$

We are 95% confident that the proportion of Democrats who support the plan is 18.1% to 29.9% higher than the proportion of Independents who support the plan. (b) True.

?? (a) College grads: 23.7%. Non-college grads: 33.7%. (b) Let p_{CG} and p_{NCG} represent the proportion of college graduates and non-college graduates who responded "do not know". $H_0 : p_{CG} = p_{NCG}$. $H_A : p_{CG} \neq p_{NCG}$. Independence is satisfied (random sample), and the success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 235/827 = 0.284$), is also satisfied. $Z = -3.18 \rightarrow$ p-value $= 0.0014$. Since the p-value is very small, we reject H_0. The data provide strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates. The data also indicate that fewer college grads say they "do not know" than non-college grads (i.e. the data indicate the direction after we reject H_0).

?? (a) College grads: 35.2%. Non-college grads: 33.9%. (b) Let p_{CG} and p_{NCG} represent the proportion of college graduates and non-college grads who support offshore drilling. $H_0 : p_{CG} = p_{NCG}$. $H_A : p_{CG} \neq p_{NCG}$. Independence is satisfied (random sample), and the success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 286/827 = 0.346$), is also satisfied. $Z = 0.39 \rightarrow$ p-value $= 0.6966$. Since the p-value $> \alpha$ (0.05), we fail to reject H_0. The data do not provide strong evidence of a difference between the proportions of college graduates and non-college graduates who support off-shore drilling in California.

?? Subscript $_C$ means control group. Subscript $_T$ means truck drivers. $H_0 : p_C = p_T$. $H_A : p_C \neq p_T$. Independence is satisfied (random samples), as is the success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 70/495 = 0.141$). $Z = -1.65 \rightarrow$ p-value $= 0.0989$. Since the p-value is high (default to alpha $= 0.05$), we fail to reject H_0. The data do not provide strong evidence that the rates of sleep deprivation are different for non-transportation workers and truck drivers.

?? (a) Summary of the study:

		Virol. failure		
		Yes	No	Total
Treatment	Nevaripine	26	94	120
	Lopinavir	10	110	120
	Total	36	204	240

(b) $H_0 : p_N = p_L$. There is no difference in virologic failure rates between the Nevaripine and Lopinavir groups. $H_A : p_N \neq p_L$. There is some difference in virologic failure rates between the Nevaripine and Lopinavir groups. (c) Random assignment was used, so the observations in each group are independent. If the patients in the study are representative of those in the general population (something impossible to check with the given information), then we can also confidently generalize the findings to the population. The success-failure condition, which we would check using the pooled proportion ($\hat{p}_{pool} = 36/240 = 0.15$), is satisfied. $Z = 2.89 \rightarrow$ p-value $= 0.0039$. Since the p-value is low, we reject H_0. There is strong evidence of a difference in virologic failure rates between the Nevaripine and Lopinavir groups. Treatment and virologic failure do not appear to be independent.

?? (a) False. The chi-square distribution has one parameter called degrees of freedom. (b) True. (c) True. (d) False. As the degrees of freedom increases, the shape of the chi-square distribution becomes more symmetric.

?? (a) H_0: The distribution of the format of the book used by the students follows the professor's predictions. H_A: The distribution of the format of the book used by the students does not follow the professor's predictions. (b) $E_{hard\ copy} = 126 \times 0.60 = 75.6$. $E_{print} = 126 \times 0.25 = 31.5$. $E_{online} = 126 \times 0.15 = 18.9$. (c) Independence: The sample is not random. However, if the professor has reason to believe that the proportions are stable from one term to the next and students are not affecting each other's study habits, independence is probably reasonable. Sample size: All expected counts are at least 5. (d) $\chi^2 = 2.32$, $df = 2$, p-value $= 0.313$. (e) Since the p-value is large, we fail to reject H_0. The data do not provide strong evidence indicating the professor's predictions were statistically inaccurate.

?? (a) Two-way table:

	Quit		
Treatment	Yes	No	Total
Patch + support group	40	110	150
Only patch	30	120	150
Total	70	230	300

(b-i) $E_{row_1,col_1} = \frac{(row\ 1\ total) \times (col\ 1\ total)}{table\ total} = 35$. This is lower than the observed value.
(b-ii) $E_{row_2,col_2} = \frac{(row\ 2\ total) \times (col\ 2\ total)}{table\ total} = 115$. This is lower than the observed value.

?? H_0: The opinion of college grads and non-grads is not different on the topic of drilling for oil and natural gas off the coast of California. H_A: Opinions regarding the drilling for oil and natural gas off the coast of California has an association with earning a college degree.

$$E_{row\ 1,col\ 1} = 151.5 \qquad E_{row\ 1,col\ 2} = 134.5$$
$$E_{row\ 2,col\ 1} = 162.1 \qquad E_{row\ 2,col\ 2} = 143.9$$
$$E_{row\ 3,col\ 1} = 124.5 \qquad E_{row\ 3,col\ 2} = 110.5$$

Independence: The samples are both random, unrelated, and from less than 10% of the population, so independence between observations is reasonable. Sample size: All expected counts are at least 5. $\chi^2 = 11.47$, $df = 2 \rightarrow$ p-value $= 0.003$. Since the p-value $< \alpha$, we reject H_0. There is strong evidence that there is an association between support for offshore drilling and having a college degree.

?? No. The samples at the beginning and at the end of the semester are not independent since the survey is conducted on the same students.

?? (a) H_0: The age of Los Angeles residents is independent of shipping carrier preference variable. H_A: The age of Los Angeles residents is associated with the shipping carrier preference variable. (b) The conditions are not satisfied since some expected counts are below 5.

?? (a) Independence is satisfied (random sample), as is the success-failure condition (40 smokers, 160 non-smokers). The 95% CI: (0.145, 0.255). We are 95% confident that 14.5% to 25.5% of all students at this university smoke. (b) We want $z^\star SE$ to be no larger than 0.02 for a 95% confidence level. We use $z^\star = 1.96$ and plug in the point estimate $\hat{p} = 0.2$ within the SE formula: $1.96\sqrt{0.2(1-0.2)/n} \leq 0.02$. The sample size n should be at least 1,537.

?? (a) Proportion of graduates from this university who found a job within one year of graduating. $\hat{p} = 348/400 = 0.87$. (b) This is a random sample, so the observations are independent. Success-failure condition is satisfied: 348 successes, 52 failures, both well above 10. (c) (0.8371, 0.9029). We are 95% confident that approximately 84% to 90% of graduates from this university found a job within one year of completing their undergraduate degree. (d) 95% of such random samples would produce a 95% confidence interval that includes the true proportion of students at this university who found a job within one year of graduating from college. (e) (0.8267, 0.9133). Similar interpretation as before. (f) 99% CI is wider, as we are more confident that the true proportion is within the interval and so need to cover a wider range.

?? Use a chi-squared goodness of fit test. H_0: Each option is equally likely. H_A: Some options are preferred over others. Total sample size: 99. Expected counts: (1/3) * 99 = 33 for each option. These are all above 5, so conditions are satisfied. $df = 3 - 1 = 2$ and $\chi^2 = \frac{(43-33)^2}{33} + \frac{(21-33)^2}{33} + \frac{(35-33)^2}{33} = 7.52 \rightarrow$ p-value $= 0.023$. Since the p-value is less than 5%, we reject H_0. The data provide convincing evidence that some options are preferred over others.

?? (a) $H_0 : p = 0.38$. $H_A : p \neq 0.38$. Independence (random sample) and the success-failure condition are satisfied. $Z = -20.5 \rightarrow$ p-value ≈ 0. Since the p-value is very small, we reject H_0. The data provide strong evidence that the proportion of Americans who only use their cell phones to access the internet is different than the Chinese proportion of 38%, and the data indicate that the proportion is lower in the US. (b) If in fact 38% of Americans used their cell phones as a primary access point to the internet, the probability of obtaining a random sample of 2,254 Americans where 17% or less or 59% or more use their only their cell phones to access the internet would be approximately 0. (c) (0.1545, 0.1855). We are 95% confident that approximately 15.5% to 18.6% of all Americans primarily use their cell phones to browse the internet.

7 Inference for numerical data

?? (a) $df = 6 - 1 = 5$, $t^\star_5 = 2.02$ (column with two tails of 0.10, row with $df = 5$). (b) $df = 21 - 1 = 20$, $t^\star_{20} = 2.53$ (column with two tails of 0.02, row with $df = 20$). (c) $df = 28$, $t^\star_{28} = 2.05$. (d) $df = 11$, $t^\star_{11} = 3.11$.

?? (a) 0.085, do not reject H_0. (b) 0.003, reject H_0. (c) 0.438, do not reject H_0. (d) 0.042, reject H_0.

?? The mean is the midpoint: $\bar{x} = 20$. Identify the margin of error: $ME = 1.015$, then use $t^\star_{35} = 2.03$ and $SE = s/\sqrt{n}$ in the formula for margin of error to identify $s = 3$.

?? (a) H_0: $\mu = 8$ (New Yorkers sleep 8 hrs per night on average.) H_A: $\mu \neq 8$ (New Yorkers sleep less or more than 8 hrs per night on average.) (b) Independence: The sample is random. The min/max suggest there are no concerning outliers. $T = -1.75$. $df = 25 - 1 = 24$. (c) p-value $= 0.093$. If in fact the true population mean of the amount New Yorkers sleep per night was 8 hours, the probability of getting a random sample of 25 New Yorkers where the average amount of sleep is 7.73 hours per night or less (or 8.27 hours or more) is 0.093. (d) Since p-value > 0.05, do not reject H_0. The data do not provide strong evidence that New Yorkers sleep more or less than 8 hours per night on average. (e) No, since the p-value is smaller than $1 - 0.90 = 0.10$.

?? T is either -2.09 or 2.09. Then $\bar{x}$ is one of the following:

$$-2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 56.26$$

$$2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 63.74$$

?? (a) We will conduct a 1-sample t-test. H_0: $\mu = 5$. H_A: $\mu \neq 5$. We'll use $\alpha = 0.05$. This is a random sample, so the observations are independent. To proceed, we assume the distribution of years of piano lessons is approximately normal. $SE = 2.2/\sqrt{20} = 0.4919$. The test statistic is $T = (4.6 - 5)/SE = -0.81$. $df = 20 - 1 = 19$. The one-tail area is about 0.21, so the p-value is about 0.42, which is bigger than $\alpha = 0.05$ and we do not reject H_0. That is, we do not have sufficiently strong evidence to reject the notion that the average is 5 years.
(b) Using $SE = 0.4919$ and $t^\star_{df=19} = 2.093$, the confidence interval is (3.57, 5.63). We are 95% confident that the average number of years a child takes piano lessons in this city is 3.57 to 5.63 years.
(c) They agree, since we did not reject the null hypothesis and the null value of 5 was in the t-interval.

?? If the sample is large, then the margin of error will be about $1.96 \times 100/\sqrt{n}$. We want this value to be less than 10, which leads to $n \geq 384.16$, meaning we need a sample size of at least 385 (round up for sample size calculations!).

?? Paired, data are recorded in the same cities at two different time points. The temperature in a city at one point is not independent of the temperature in the same city at another time point.

?? (a) Since it's the same students at the beginning and the end of the semester, there is a pairing between the datasets, for a given student their beginning and end of semester grades are dependent.
(b) Since the subjects were sampled randomly, each observation in the men's group does not have a special correspondence with exactly one observation in the other (women's) group. (c) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester artery thickness are dependent. (d) Since it's the same subjects at the beginning and the end of the study, there is a pairing between the datasets, for a subject student their beginning and end of semester weights are dependent.

?? (a) For each observation in one data set, there is exactly one specially corresponding observation in the other data set for the same geographic location. The data are paired. (b) $H_0 : \mu_{\text{diff}} = 0$ (There is no difference in average number of days exceeding 90°F in 1948 and 2018 for NOAA stations.) $H_A : \mu_{\text{diff}} \neq 0$ (There is a difference.) (c) Locations were randomly sampled, so independence is reasonable. The sample size is at least 30, so we're just looking for particularly extreme outliers: none are present (the observation off left in the histogram would be considered a clear outlier, but not a particularly extreme one). Therefore, the conditions are satisfied. (d) $SE = 17.2/\sqrt{197} = 1.23$. $T = \frac{2.9-0}{1.23} = 2.36$ with degrees of freedom $df = 197 - 1 = 196$. This leads to a one-tail area of 0.0096 and a p-value of about 0.019. (e) Since the p-value is less than 0.05, we reject H_0. The data provide strong evidence that NOAA stations observed more 90°F days in 2018 than in 1948. (f) Type 1 Error, since we may have incorrectly rejected H_0. This error would mean that NOAA stations did not actually observe a decrease, but the sample we took just so happened to make it appear that this was the case. (g) No, since we rejected H_0, which had a null value of 0.

?? (a) $SE = 1.23$ and $t^\star = 1.65$. $2.9 \pm 1.65 \times 1.23 \rightarrow (0.87, 4.93)$.
(b) We are 90% confident that there was an increase of 0.87 to 4.93 in the average number of days that hit 90°F in 2018 relative to 1948 for NOAA stations.
(c) Yes, since the interval lies entirely above 0.

?? (a) These data are paired. For example, the Friday the 13th in say, September 1991, would probably be more similar to the Friday the 6th in September 1991 than to Friday the 6th in another month or year.
(b) Let $\mu_{diff} = \mu_{sixth} - \mu_{thirteenth}$. $H_0 : \mu_{diff} = 0$. $H_A : \mu_{diff} \neq 0$.
(c) Independence: The months selected are not random. However, if we think these dates are roughly equivalent to a simple random sample of all such Friday 6th/13th date pairs, then independence is reasonable. To proceed, we must make this strong assumption, though we should note this assumption in any reported results. Normality: With fewer than 10 observations, we would need to see clear outliers to be concerned. There is a borderline outlier on the right of the histogram of the differences, so we would want to report this in formal analysis results.
(d) $T = 4.93$ for $df = 10 - 1 = 9 \rightarrow$ p-value $= 0.001$.
(e) Since p-value < 0.05, reject H_0. The data provide strong evidence that the average number of cars at the intersection is higher on Friday the 6^{th} than on Friday the 13^{th}. (We should exercise caution about generalizing the interpretation to all intersections or roads.)
(f) If the average number of cars passing the intersection actually was the same on Friday the 6^{th} and 13^{th}, then the probability that we would observe a test statistic so far from zero is less than 0.01.
(g) We might have made a Type 1 Error, i.e. incorrectly rejected the null hypothesis.

?? (a) $H_0 : \mu_{diff} = 0$. $H_A : \mu_{diff} \neq 0$. $T = -2.71$. $df = 5$. p-value = 0.042. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average number of traffic accident related emergency room admissions are different between Friday the 6^{th} and Friday the 13^{th}. Furthermore, the data indicate that the direction of that difference is that accidents are lower on Friday the 6^{th} relative to Friday the 13^{th}.
(b) (-6.49, -0.17).
(c) This is an observational study, not an experiment, so we cannot so easily infer a causal intervention implied by this statement. It is true that there is a difference. However, for example, this does not mean that a responsible adult going out on Friday the 13^{th} has a higher chance of harm than on any other night.

?? (a) Chicken fed linseed weighed an average of 218.75 grams while those fed horsebean weighed an average of 160.20 grams. Both distributions are relatively symmetric with no apparent outliers. There is more variability in the weights of chicken fed linseed.
(b) $H_0 : \mu_{ls} = \mu_{hb}$. $H_A : \mu_{ls} \neq \mu_{hb}$.
We leave the conditions to you to consider.
$T = 3.02$, $df = min(11, 9) = 9 \rightarrow$ p-value = 0.014. Since p-value < 0.05, reject H_0. The data provide strong evidence that there is a significant difference between the average weights of chickens that were fed linseed and horsebean.
(c) Type 1 Error, since we rejected H_0.
(d) Yes, since p-value > 0.01, we would not have rejected H_0.

?? $H_0 : \mu_C = \mu_S$. $H_A : \mu_C \neq \mu_S$. $T = 3.27$, $df = 11 \rightarrow$ p-value = 0.007. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average weight of chickens that were fed casein is different than the average weight of chickens that were fed soybean (with weights from casein being higher). Since this is a randomized experiment, the observed difference can be attributed to the diet.

?? Let $\mu_{diff} = \mu_{pre} - \mu_{post}$. $H_0 : \mu_{diff} = 0$: Treatment has no effect. $H_A : \mu_{diff} \neq 0$: Treatment has an effect on P.D.T. scores, either positive or negative. Conditions: The subjects are randomly assigned to treatments, so independence within and between groups is satisfied. All three sample sizes are smaller than 30, so we look for clear outliers. There is a borderline outlier in the first treatment group. Since it is borderline, we will proceed, but we should report this caveat with any results. For all three groups: $df = 13$. $T_1 = 1.89 \rightarrow$ p-value = 0.081, $T_2 = 1.35 \rightarrow$ p-value = 0.200), $T_3 = -1.40 \rightarrow$ (p-value = 0.185). We do not reject the null hypothesis for any of these groups. As earlier noted, there is some uncertainty about if the method applied is reasonable for the first group.

?? Difference we care about: 40. Single tail of 90%: $1.28 \times SE$. Rejection region bounds: $\pm 1.96 \times SE$ (if 5% significance level). Setting $3.24 \times SE = 40$, subbing in $SE = \sqrt{\frac{94^2}{n} + \frac{94^2}{n}}$, and solving for the sample size n gives 116 plots of land for each fertilizer.

?? Alternative.

?? H_0: $\mu_1 = \mu_2 = \cdots = \mu_6$. H_A: The average weight varies across some (or all) groups. Independence: Chicks are randomly assigned to feed types (presumably kept separate from one another), therefore independence of observations is reasonable. Approx. normal: the distributions of weights within each feed type appear to be fairly symmetric. Constant variance: Based on the side-by-side box plots, the constant variance assumption appears to be reasonable. There are differences in the actual computed standard deviations, but these might be due to chance as these are quite small samples. $F_{5,65} = 15.36$ and the p-value is approximately 0. With such a small p-value, we reject H_0. The data provide convincing evidence that the average weight of chicks varies across some (or all) feed supplement groups.

?? (a) H_0: The population mean of MET for each group is equal to the others. H_A: At least one pair of means is different. (b) Independence: We don't have any information on how the data were collected, so we cannot assess independence. To proceed, we must assume the subjects in each group are independent. In practice, we would inquire for more details. Normality: The data are bound below by zero and the standard deviations are larger than the means, indicating very strong skew. However, since the sample sizes are extremely large, even extreme skew is acceptable. Constant variance: This condition is sufficiently met, as the standard deviations are reasonably consistent across groups. (c) See below, with the last column omitted:

	Df	Sum Sq	Mean Sq	F value
coffee	4	10508	2627	5.2
Residuals	50734	25564819	504	
Total	50738	25575327		

(d) Since p-value is very small, reject H_0. The data provide convincing evidence that the average MET differs between at least one pair of groups.

?? (a) H_0: Average GPA is the same for all majors. H_A: At least one pair of means are different. (b) Since p-value > 0.05, fail to reject H_0. The data do not provide convincing evidence of a difference between the average GPAs across three groups of majors. (c) The total degrees of freedom is 195+2 = 197, so the sample size is $197 + 1 = 198$.

?? (a) False. As the number of groups increases, so does the number of comparisons and hence the modified significance level decreases. (b) True. (c) True. (d) False. We need observations to be independent regardless of sample size.

?? (a) H_0: Average score difference is the same for all treatments. H_A: At least one pair of means are different. (b) We should check conditions. If we look back to the earlier exercise, we will see that the patients were randomized, so independence is satisfied. There are some minor concerns about skew, especially with the third group, though this may be acceptable. The standard deviations across the groups are reasonably similar. Since the p-value is less than 0.05, reject H_0. The data provide convincing evidence of a difference between the average reduction in score among treatments. (c) We determined that at least two means are different in part (b), so we now conduct $K = 3 \times 2/2 = 3$ pairwise t-tests that each use $\alpha = 0.05/3 = 0.0167$ for a significance level. Use the following hypotheses for each pairwise test. H_0: The two means are equal. H_A: The two means are different. The sample sizes are equal and we use the pooled SD, so we can compute $SE = 3.7$ with the pooled $df = 39$. The p-value for Trmt 1 vs. Trmt 3 is the only one under 0.05: p-value = 0.035 (or 0.024 if using s_{pooled} in place of s_1 and s_3, though this won't affect the final conclusion). The p-value is larger than $0.05/3 = 1.67$, so we do not have strong evidence to conclude that it is this particular pair of groups that are different. That is, we cannot identify if which particular pair of groups are actually different, even though we've rejected the notion that they are all the same!

?? $H_0 : \mu_T = \mu_C$. $H_A : \mu_T \neq \mu_C$. $T = 2.24$, $df = 21 \rightarrow$ p-value $= 0.036$. Since p-value < 0.05, reject H_0. The data provide strong evidence that the average food consumption by the patients in the treatment and control groups are different. Furthermore, the data indicate patients in the distracted eating (treatment) group consume more food than patients in the control group.

?? False. While it is true that paired analysis requires equal sample sizes, only having the equal sample sizes isn't, on its own, sufficient for doing a paired test. Paired tests require that there be a special correspondence between each pair of observations in the two groups.

?? (a) We are building a distribution of sample statistics, in this case the sample mean. Such a distribution is called a sampling distribution. (b) Because we are dealing with the distribution of sample means, we need to check to see if the Central Limit Theorem applies. Our sample size is greater than 30, and we are told that random sampling is employed. With these conditions met, we expect that the distribution of the sample mean will be nearly normal and therefore symmetric. (c) Because we are dealing with a sampling distribution, we measure its variability with the standard error. $SE = 18.2/\sqrt{45} = 2.713$. (d) The sample means will be more variable with the smaller sample size.

?? (a) We should set 1.0% equal to 2.8 standard errors: $2.8 \times SE_{desired} = 1.0\%$ (see Example ?? on page ?? for details). This means the standard error should be about $SE = 0.36\%$ to achieve the desired statistical power.
(b) The margin of error was $0.5 \times (2.6\% - (-0.2\%)) = 1.4\%$, so the standard error in the experiment must have been $1.96 \times SE_{original} = 1.4\% \rightarrow SE_{original} = 0.71\%$.
(c) The standard error decreases with the square root of the sample size, so we should increase the sample size by a factor of $1.97^2 = 3.88$.
(d) The team should run an experiment 3.88 times larger, so they should have a random sample of 3.88% of their users in each of the experiment arms in the new experiment.

?? Independence: it is a random sample, so we can assume that the students in this sample are independent of each other with respect to number of exclusive relationships they have been in. Notice that there are no students who have had no exclusive relationships in the sample, which suggests some student responses are likely missing (perhaps only positive values were reported). The sample size is at least 30, and there are no particularly extreme outliers, so the normality condition is reasonable. 90% CI: (2.97, 3.43). We are 90% confident that undergraduate students have been in 2.97 to 3.43 exclusive relationships, on average.

?? The hypotheses should be about the population mean (μ), not the sample mean. The null hypothesis should have an equal sign and the alternative hypothesis should be about the null hypothesized value, not the observed sample mean. Correction:

$$H_0 : \mu = 10 \; hours$$
$$H_A : \mu \neq 10 \; hours$$

A two-sided test allows us to consider the possibility that the data show us something that we would find surprising.

?? (a) The residual plot will show randomly distributed residuals around 0. The variance is also approximately constant. (b) The residuals will show a fan shape, with higher variability for smaller x. There will also be many points on the right above the line. There is trouble with the model being fit here.

?? (a) Strong relationship, but a straight line would not fit the data. (b) Strong relationship, and a linear fit would be reasonable. (c) Weak relationship, and trying a linear fit would be reasonable. (d) Moderate relationship, but a straight line would not fit the data. (e) Strong relationship, and a linear fit would be reasonable. (f) Weak relationship, and trying a linear fit would be reasonable.

?? (a) Exam 2 since there is less of a scatter in the plot of final exam grade versus exam 2. Notice that the relationship between Exam 1 and the Final Exam appears to be slightly nonlinear. (b) Exam 2 and the final are relatively close to each other chronologically, or Exam 2 may be cumulative so has greater similarities in material to the final exam. Answers may vary.

?? (a) $r = -0.7 \rightarrow$ (4). (b) $r = 0.45 \rightarrow$ (3). (c) $r = 0.06 \rightarrow$ (1). (d) $r = 0.92 \rightarrow$ (2).

?? (a) The relationship is positive, weak, and possibly linear. However, there do appear to be some anomalous observations along the left where several students have the same height that is notably far from the cloud of the other points. Additionally, there are many students who appear not to have driven a car, and they are represented by a set of points along the bottom of the scatterplot. (b) There is no obvious explanation why simply being tall should lead a person to drive faster. However, one confounding factor is gender. Males tend to be taller than females on average, and personal experiences (anecdotal) may suggest they drive faster. If we were to follow-up on this suspicion, we would find that sociological studies confirm this suspicion. (c) Males are taller on average and they drive faster. The gender variable is indeed an important confounding variable.

?? (a) There is a somewhat weak, positive, possibly linear relationship between the distance traveled and travel time. There is clustering near the lower left corner that we should take special note of. (b) Changing the units will not change the form, direction or strength of the relationship between the two variables. If longer distances measured in miles are associated with longer travel time measured in minutes, longer distances measured in kilometers will be associated with longer travel time measured in hours. (c) Changing units doesn't affect correlation: $r = 0.636$.

?? (a) There is a moderate, positive, and linear relationship between shoulder girth and height. (b) Changing the units, even if just for one of the variables, will not change the form, direction or strength of the relationship between the two variables.

?? In each part, we can write the husband ages as a linear function of the wife ages.
(a) $age_H = age_W + 3$.
(b) $age_H = age_W - 2$.
(c) $age_H = 2 \times age_W$.
Since the slopes are positive and these are perfect linear relationships, the correlation will be exactly 1 in all three parts. An alternative way to gain insight into this solution is to create a mock data set, e.g. 5 women aged 26, 27, 28, 29, and 30, then find the husband ages for each wife in each part and create a scatterplot.

?? Correlation: no units. Intercept: kg. Slope: kg/cm.

?? Over-estimate. Since the residual is calculated as *observed* $-$ *predicted*, a negative residual means that the predicted value is higher than the observed value.

?? (a) There is a positive, very strong, linear association between the number of tourists and spending. (b) Explanatory: number of tourists (in thousands). Response: spending (in millions of US dollars). (c) We can predict spending for a given number of tourists using a regression line. This may be useful information for determining how much the country may want to spend in advertising abroad, or to forecast expected revenues from tourism. (d) Even though the relationship appears linear in the scatterplot, the residual plot actually shows a nonlinear relationship. This is not a contradiction: residual plots can show divergences from linearity that can be difficult to see in a scatterplot. A simple linear model is inadequate for modeling these data. It is also important to consider that these data are observed sequentially, which means there may be a hidden structure not evident in the current plots but that is important to consider.

?? (a) First calculate the slope: $b_1 = R \times s_y/s_x = 0.636 \times 113/99 = 0.726$. Next, make use of the fact that the regression line passes through the point $(\bar{x}, \bar{y})$: $\bar{y} = b_0 + b_1 \times \bar{x}$. Plug in $\bar{x}$, $\bar{y}$, and b_1, and solve for b_0: 51. Solution: $\widehat{travel\ time} = 51 + 0.726 \times distance$. (b) b_1: For each additional mile in distance, the model predicts an additional 0.726 minutes in travel time. b_0: When the distance traveled is 0 miles, the travel time is expected to be 51 minutes. It does not make sense to have a travel distance of 0 miles in this context. Here, the y-intercept serves only to adjust the height of the line and is meaningless by itself. (c) $R^2 = 0.636^2 = 0.40$. About 40% of the variability in travel time is accounted for by the model, i.e. explained by the distance traveled. (d) $\widehat{travel\ time} = 51 + 0.726 \times distance = 51 + 0.726 \times 103 \approx 126$ minutes. (Note: we should be cautious in our predictions with this model since we have not yet evaluated whether it is a well-fit model.) (e) $e_i = y_i - \hat{y}_i = 168 - 126 = 42$ minutes. A positive residual means that the model underestimates the travel time. (f) No, this calculation would require extrapolation.

?? (a) $\widehat{murder} = -29.901 + 2.559 \times poverty\%$. (b) Expected murder rate in metropolitan areas with no poverty is -29. 901 per million. This is obviously not a meaningful value, it just serves to adjust the height of the regression line. (c) For each additional percentage increase in poverty, we expect murders per million to be higher on average by 2.559. (d) Poverty level explains 70.52% of the variability in murder rates in metropolitan areas. (e) $\sqrt{0.7052} = 0.8398$.

?? (a) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. It is also an influential point since, without that observation, the regression line would have a very different slope.
(b) There is an outlier in the bottom right. Since it is far from the center of the data, it is a point with high leverage. However, it does not appear to be affecting the line much, so it is not an influential point.
(c) The observation is in the center of the data (in the x-axis direction), so this point does *not* have high leverage. This means the point won't have much effect on the slope of the line and so is not an influential point.

?? (a) There is a negative, moderate-to-strong, somewhat linear relationship between percent of families who own their home and the percent of the population living in urban areas in 2010. There is one outlier: a state where 100% of the population is urban. The variability in the percent of homeownership also increases as we move from left to right in the plot.
(b) The outlier is located in the bottom right corner, horizontally far from the center of the other points, so it is a point with high leverage. It is an influential point since excluding this point from the analysis would greatly affect the slope of the regression line.

?? (a) The relationship is positive, moderate-to-strong, and linear. There are a few outliers but no points that appear to be influential.
(b) $\widehat{weight} = -105.0113 + 1.0176 \times height$.
Slope: For each additional centimeter in height, the model predicts the average weight to be 1.0176 additional kilograms (about 2.2 pounds).
Intercept: People who are 0 centimeters tall are expected to weigh - 105.0113 kilograms. This is obviously not possible. Here, the y- intercept serves only to adjust the height of the line and is meaningless by itself.
(c) H_0: The true slope coefficient of height is zero ($\beta_1 = 0$).
H_A: The true slope coefficient of height is different than zero ($\beta_1 \neq 0$).
The p-value for the two-sided alternative hypothesis ($\beta_1 \neq 0$) is incredibly small, so we reject H_0. The data provide convincing evidence that height and weight are positively correlated. The true slope parameter is indeed greater than 0.
(d) $R^2 = 0.72^2 = 0.52$. Approximately 52% of the variability in weight can be explained by the height of individuals.

?? (a) H_0: $\beta_1 = 0$. H_A: $\beta_1 \neq 0$. The p-value, as reported in the table, is incredibly small and is smaller than 0.05, so we reject H_0. The data provide convincing evidence that wives' and husbands' heights are positively correlated.
(b) $\widehat{height}_W = 43.5755 + 0.2863 \times height_H$.
(c) Slope: For each additional inch in husband's height, the average wife's height is expected to be an additional 0.2863 inches on average. Intercept: Men who are 0 inches tall are expected to have wives who are, on average, 43.5755 inches tall. The intercept here is meaningless, and it serves only to adjust the height of the line.
(d) The slope is positive, so r must also be positive. $r = \sqrt{0.09} = 0.30$.
(e) 63.33. Since R^2 is low, the prediction based on this regression model is not very reliable.
(f) No, we should avoid extrapolating.

?? (a) $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$ (b) The p-value for this test is approximately 0, therefore we reject H_0. The data provide convincing evidence that poverty percentage is a significant predictor of murder rate. (c) $n = 20, df = 18, T^*_{18} = 2.10$; $2.559 \pm 2.10 \times 0.390 = (1.74, 3.378)$; For each percentage point poverty is higher, murder rate is expected to be higher on average by 1.74 to 3.378 per million.
(d) Yes, we rejected H_0 and the confidence interval does not include 0.

?? (a) True. (b) False, correlation is a measure of the linear association between any two numerical variables.

?? (a) The point estimate and standard error are $b_1 = 0.9112$ and $SE = 0.0259$. We can compute a T-score: $T = (0.9112 - 1)/0.0259 = -3.43$. Using $df = 168$, the p-value is about 0.001, which is less than $\alpha = 0.05$. That is, the data provide strong evidence that the average difference between husbands' and wives' ages has actually changed over time. (b) $\widehat{age}_W = 1.5740 + 0.9112 \times age_H$. (c) Slope: For each additional year in husband's age, the model predicts an additional 0.9112 years in wife's age. This means that wives' ages tend to be lower for later ages, suggesting the average gap of husband and wife age is larger for older people. Intercept: Men who are 0 years old are expected to have wives who are on average 1.5740 years old. The intercept here is meaningless and serves only to adjust the height of the line. (d) $R = \sqrt{0.88} = 0.94$. The regression of wives' ages on husbands' ages has a positive slope, so the correlation coefficient will be positive. (e) $\widehat{age}_W = 1.5740 + 0.9112 \times 55 = 51.69$. Since R^2 is pretty high, the prediction based on this regression model is reliable. (f) No, we shouldn't use the same model to predict an 85 year old man's wife's age. This would require extrapolation. The scatterplot from an earlier exercise shows that husbands in this data set are approximately 20 to 65 years old. The regression model may not be reasonable outside of this range.

?? There is an upwards trend. However, the variability is higher for higher calorie counts, and it looks like there might be two clusters of observations above and below the line on the right, so we should be cautious about fitting a linear model to these data.

?? (a) $r = -0.72 \rightarrow$ (2) (b) $r = 0.07 \rightarrow$ (4) (c) $r = 0.86 \rightarrow$ (1) (d) $r = 0.99 \rightarrow$ (3)

9 Multiple and logistic regression

?? (a) $\widehat{baby\_weight} = 123.05 - 8.94 \times smoke$ (b) The estimated body weight of babies born to smoking mothers is 8.94 ounces lower than babies born to non-smoking mothers. Smoker: $123.05 - 8.94 \times 1 = 114.11$ ounces. Non-smoker: $123.05 - 8.94 \times 0 = 123.05$ ounces. (c) H_0: $\beta_1 = 0$. H_A: $\beta_1 \neq 0$. $T = -8.65$, and the p-value is approximately 0. Since the p-value is very small, we reject H_0. The data provide strong evidence that the true slope parameter is different than 0 and that there is an association between birth weight and smoking. Furthermore, having rejected H_0, we can conclude that smoking is associated with lower birth weights.

?? (a) $\widehat{baby\_weight} = -80.41 + 0.44 \times gestation - 3.33 \times parity - 0.01 \times age + 1.15 \times height + 0.05 \times weight - 8.40 \times smoke$. (b) $\beta_{gestation}$: The model predicts a 0.44 ounce increase in the birth weight of the baby for each additional day of pregnancy, all else held constant. β_{age}: The model predicts a 0.01 ounce decrease in the birth weight of the baby for each additional year in mother's age, all else held constant. (c) Parity might be correlated with one of the other variables in the model, which complicates model estimation. (d) $\widehat{baby\_weight} = 120.58$. $e = 120 - 120.58 = -0.58$. The model over-predicts this baby's birth weight. (e) $R^2 = 0.2504$. $R^2_{adj} = 0.2468$.

?? (a) (-0.32, 0.16). We are 95% confident that male students on average have GPAs 0.32 points lower to 0.16 points higher than females when controlling for the other variables in the model. (b) Yes, since the p-value is larger than 0.05 in all cases (not including the intercept).

?? Remove age.

?? Based on the p-value alone, either gestation or smoke should be added to the model first. However, since the adjusted R^2 for the model with gestation is higher, it would be preferable to add gestation in the first step of the forward- selection algorithm. (Other explanations are possible. For instance, it would be reasonable to only use the adjusted R^2.)

?? She should use p-value selection since she is interested in finding out about significant predictors, not just optimizing predictions.

?? Nearly normal residuals: With so many observations in the data set, we look for particularly extreme outliers in the histogram and do not see any.
variability of residuals: The scatterplot of the residuals versus the fitted values does not show any overall structure. However, values that have very low or very high fitted values appear to also have somewhat larger outliers. In addition, the residuals do appear to have constant variability between the two parity and smoking status groups, though these items are relatively minor.
Independent residuals: The scatterplot of residuals versus the order of data collection shows a random scatter, suggesting that there is no apparent structures related to the order the data were collected.
Linear relationships between the response variable and numerical explanatory variables: The residuals vs. height and weight of mother are randomly distributed around 0. The residuals vs. length of gestation plot also does not show any clear or strong remaining structures, with the possible exception of very short or long gestations. The rest of the residuals do appear to be randomly distributed around 0.
All concerns raised here are relatively mild. There are some outliers, but there is so much data that the influence of such observations will be minor.

?? (a) There are a few potential outliers, e.g. on the left in the `total_length` variable, but nothing that will be of serious concern in a data set this large. (b) When coefficient estimates are sensitive to which variables are included in the model, this typically indicates that some variables are collinear. For example, a possum's gender may be related to its head length, which would explain why the coefficient (and p-value) for `sex_male` changed when we removed the `head_length` variable. Likewise, a possum's skull width is likely to be related to its head length, probably even much more closely related than the head length was to gender.

?? (a) The logistic model relating $\hat{p}_i$ to the predictors may be written as $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = 33.5095 - 1.4207 \times sex\_male_i - 0.2787 \times skull\_width_i + 0.5687 \times total\_length_i - 1.8057 \times tail\_length_i$. Only `total_length` has a positive association with a possum being from Victoria. (b) $\hat{p} = 0.0062$. While the probability is very near zero, we have not run diagnostics on the model. We might also be a little skeptical that the model will remain accurate for a possum found in a US zoo. For example, perhaps the zoo selected a possum with specific characteristics but only looked in one region. On the other hand, it is encouraging that the possum was caught in the wild. (Answers regarding the reliability of the model probability will vary.)

?? (a) False. When predictors are collinear, it means they are correlated, and the inclusion of one variable can have a substantial influence on the point estimate (and standard error) of another. (b) True. (c) False. This would only be the case if the data was from an experiment and x_1 was one of the variables set by the researchers. (Multiple regression can be useful for forming hypotheses about causal relationships, but it offers zero guarantees.) (d) False. We should check normality like we would for inference for a single mean: we look for particularly extreme outliers if $n \geq 30$ or for clear outliers if $n < 30$.

?? (a) `exclaim_subj` should be removed, since it's removal reduces AIC the most (and the resulting model has lower AIC than the None Dropped model). (b) Removing any variable will increase AIC, so we should not remove any variables from this set.

?? (a) The equation is:

$$\begin{aligned}\log\left(\frac{p_i}{1-p_i}\right) = &-0.8124 \\ &- 2.6351 \times \texttt{to\_multiple} \\ &+ 1.6272 \times \texttt{winner} \\ &- 1.5881 \times \texttt{format} \\ &- 3.0467 \times \texttt{re\_subj}\end{aligned}$$

(b) First find $\log\left(\frac{p}{1-p}\right)$, then solve for p:

$$\begin{aligned}&\log\left(\frac{p}{1-p}\right) \\ &\quad = -0.8124 - 2.6351 \times 0 + 1.6272 \times 1 \\ &\qquad - 1.5881 \times 0 - 3.0467 \times 0 \\ &\quad = 0.8148 \\ &\frac{p}{1-p} = e^{0.8148} \quad \rightarrow \quad p = 0.693\end{aligned}$$

(c) It should probably be pretty high, since it could be very disruptive to the person using the email service if they are missing emails that aren't spam. Even only a 90% chance that a message is spam is probably enough to warrant keeping it in the inbox. Maybe a probability of 99% would be a reasonable cutoff. As for other ideas to make it even better, it may be worth building a second model that tries to classify the importance of an email message. If we have both the spam model and the importance model, we now have a better way to think about cost-benefit tradeoffs. For instance, perhaps we would be willing to have a lower probability-of-spam threshold for messages we were confident were not important, and perhaps we want an even higher probability threshold (e.g. 99.99%) for emails we are pretty sure are important.

Prilog B

Skupovi podataka u ovom tekstu

Svaki skup podataka korišten u ovom udžbeniku opisan je u ovom prilogu, a za svaki skup podataka postoji odgovarajuća stranica na **openintro.org/data**. Na toj stranici nalaze se i dodatni skupovi podataka koje možete koristiti da usavršite svoje vještine. Svaki skup podataka ima vlastitu stranicu sa sljedećim informacijama:

- Popis varijabli u skupu podataka.
- datoteka za preuzimanje u CSV formatu.
- datoteka za preuzimanje u nativnom formatu R objekta.

B.1 Uvod u podatke

1.1 `stent30`, `stent365` → Podaci o stentovima podijeljeni su u dva skupa podataka, jedan o rezultatima u danima 0-30 i jedan o rezultatima u danima 0-365.
Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis. New England Journal of Medicine 365:993-1003. www.nejm.org/doi/full/10.1056/NEJMoa1105335.
Članak u NY Times: www.nytimes.com/2011/09/08/health/research/08stent.html.

1.2 `loan50`, `loans_full_schema` → Ovi podaci skupljeni u okviru kase uzajamne pomoći Lending Club (lendingclub.com), pružaju veliki skup podataka o ljudima koji su dobili kredit putem njihove platforme. Podaci koji su korišteni u udžbeniku su uzorak kredita iz prvog kvartala Q1 (siječanj, veljača, ožujak) 2018.

1.2 `county`, `county_complete` → Ovi su podaci prikupljeni iz nekoliko izvora Vlade SAD-a. Za varijable uključene u skup podataka o općinama (eng. county), sadržani su samo najnoviji podaci, koji su bili dostupni krajem 2018. godine. Podaci iz razdoblja prije 2011. godine su sa stranice census.gov, na kojoj specifična stranica Quick Facts sa koje su preuzeti podaci više nije dostupna. Noviji podaci su sa stranica USDA (ers.usda.gov), Bureau of Labor Statistics (bls.gov/lau), SAIPE (census.gov/did/www/saipe), i American Community Survey (census.gov/programs-surveys/acs).

1.3 Nurses' Health Study → Više informacija o ovom skupu podataka možete naći na stranici www.channing.harvard.edu/nhs

1.4 Studija na koju smo mislili kad smo raspravljali jednostavnu randomizaciju (bez blokova) je Anturane Reinfarction Trial Research Group. 1980. *Sulfinpyrazone in the prevention of sudden death after myocardial infarction.* New England Journal of Medicine 302(5):250-256.

B.2 ??

?? `loan50`, `county` → Ovi skupovi podataka opisani su u prilogu B.1.

?? `loan50`, `county` → Ovi skupovi podataka opisani su u prilogu B.1.

?? `malaria` → Lyke et al. 2017. PfSPZ vaccine induces strain-transcending T cells and durable protection against heterologous controlled human malaria infection. PNAS 114(10):2711-2716. www.pnas.org/content/114/10/2711

B.3 ??

?? `loan50`, `county` → Ovi skupovi podataka opisani su u prilogu B.1.

?? `playing_cards` → Data set describing the 52 cards in a standard deck.

?? `family_college` → Simulated data based on real population summaries at nces.ed.gov/pubs2001/2001126.pdf.

?? `smallpox` → Fenner F. 1988. Smallpox and Its Eradication (History of International Public Health, No. 6). Geneva: World Health Organization. ISBN 92-4-156110-6.

?? Mammogram screening, probabilities → The probabilities reported were obtained using studies reported at www.breastcancer.org and www.ncbi.nlm.nih.gov/pmc/articles/PMC1173421.

?? Jose campus visits, probabilities → Ovaj primjer je izmišljen.

?? No data sets were described in this section.

?? Course material purchases and probabilities → Ovaj primjer je izmišljen.

?? Auctions for TV and toaster → Ovaj primjer je izmišljen.

?? `stocks_18` → Monthly returns for Caterpillar, Exxon Mobil Corp, and Google for November 2015 to October 2018.

?? `fcid` → This sample can be considered a simple random sample from the US population. It relies on the USDA Food Commodity Intake Database.

B.4 ??

?? SAT and ACT score distributions → The SAT score data comes from the 2018 distribution, which is provided at
reports.collegeboard.org/pdf/2018-total-group-sat-suite-assessments-annual-report.pdf
The ACT score data is available at
act.org/content/dam/act/unsecured/documents/cccr2018/P_99_999999_N_S_N00_ACT-GCPR_National.p
We also acknowledge that the actual ACT score distribution is *not* nearly normal. However, since the topic is very accessible, we decided to keep the context and examples.

?? Male heights → The distribution is based on the USDA Food Commodity Intake Database.

?? `possum` → The distribution parameters are based on a sample of possums from Australia and New Guinea. The original source of this data is as follows. Lindenmayer DB, et al. 1995. *Morphological variation among columns of the mountain brushtail possum, Trichosurus caninus Ogilby (Phalangeridae: Marsupiala).* Australian Journal of Zoology 43: 449-458.

?? Exceeding insurance deductible → These statistics were made up but are possible values one might observe for low-deductible plans.

?? Exceeding insurance deductible → These statistics were made up but are possible values one might observe for low-deductible plans.

?? Smoking friends → Unfortunately, we don't currently have additional information on the source for the 30% statistic, so don't consider this one as fact since we cannot verify it was from a reputable source.

?? US smoking rate → The 15% smoking rate in the US figure is close to the value from the Centers for Disease Control and Prevention website, which reports a value of 14% as of the 2017 estimate:
cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm

?? Football kicker → Ovaj primjer je izmišljen.

?? Heart attack admissions → This example was made up, though the heart attack admissions are realistic for some hospitals.

?? ami_occurrences → This is a simulated data set but resembles actual AMI data for New York City based on typical AMI incidence rates.

B.5 ??

?? pew_energy_2018 → The actual data has more observations than were referenced in this chapter. That is, we used a subsample since it helped smooth some of the examples to have a bit more variability. The pew_energy_2018 data set represents the full data set for each of the different energy source questions, which covers solar, wind, offshore drilling, hydrolic fracturing, and nuclear energy. The statistics used to construct the data are from the following page:

www.pewinternet.org/2018/05/14/majorities-see-government-efforts-to-protect-the-environment-as-insufficient/

?? pew_energy_2018 → See the details for this data set above in the Section ?? data section.

?? ebola_survey → In New York City on October 23rd, 2014, a doctor who had recently been treating Ebola patients in Guinea went to the hospital with a slight fever and was subsequently diagnosed with Ebola. Soon thereafter, an NBC 4 New York/The Wall Street Journal/Marist Poll found that 82% of New Yorkers favored a "mandatory 21-day quarantine for anyone who has come in contact with an Ebola patient". This poll included responses of 1,042 New York adults between Oct 26th and 28th, 2014. Poll ID NY141026 on maristpoll.marist.edu.

?? pew_energy_2018 → See the details for this data set above in the Section ?? data section.

?? Rosling questions → We noted much smaller samples than the Roslings' describe in their book, Factfulness, The samples we describe are similar but not the same as the actual rates. The approximate rates for the correct answers for the two questions for (sometimes different) populations discussed in the book, as reported in Factfulness, are

- 80% of the world's 1 year olds have been vaccinated against some disease: 13% get this correct (17% in the US). gapm.io/q9
- Number of children in the world in 2100: 9% correct. gapm.io/q5

Here are a few more questions and a rough percent of people who get them correct:

- In all low-income countries across the world today, how many girls finish primary school: 20%, 40%, or 60%? Answer: 60%. About 7% of people get this question correct. gapm.io/q1
- What is the life expectancy of the world today: 50 years, 60 years, or 70 years? Answer: 70 years. In the US, about 43% of people get this question correct. gapm.io/q4
- In 1996, tigers, giant pandas, and black rhinos were all listed as endangered. How many of these three species are more critically endangered today: two of them, one of them, none of them? Answer: none of them. About 7% of people get this question correct. gapm.io/q11
- How many people in the world have some access to electricity? 20%, 50%, 80%. Answer: 80%. About 22% of people get this correct. gapm.io/q12

For more information, check out the book, Factfulness.

?? pew_energy_2018 → See the details for this data set above in the Section ?? data section.

?? nuclear_survey → A simple random sample of 1,028 US adults in March 2013 found that 56% of US adults support nuclear arms reduction.
www.gallup.com/poll/161198/favor-russian-nuclear-arms-reductions.aspx

?? Car manufacturing → Ovaj primjer je izmišljen.

?? stent30, stent365 → Ovi skupovi podataka opisani su u prilogu B.1.

B.6 ??

?? Payday loans → The statistics come from the following source: pewtrusts.org/-/media/assets/2017/04/payday-loan-customers-want-more-protections-methodology.pdf

?? Tire factory → Ovaj primjer je izmišljen.

?? `cpr` → Böttiger et al. *Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial.* The Lancet, 2001.

?? `fish_oil_18` → Manson JE, et al. 2018. *Marine n-3 Fatty Acids and Prevention of Cardiovascular Disease and Cancer.* NEJMoa1811403.

?? `mammogram` → Miller AB. 2014. *Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial.* BMJ 2014;348:g366.

?? `drone_blades` → The quality control data set for quadcopter drone blades is a made-up data set for an example. We provide the simulated data in the `drone_blades` data set.

?? `jury` → The jury data set for examining discrimination is a made-up data set an example. We provide the simulated data in the `jury` data set.

?? `sp500_1950_2018` → Data is sourced from finance.yahoo.com.

?? `ask` → Minson JA, Ruedy NE, Schweitzer ME. *There is such a thing as a stupid question: Question disclosure in strategic communication.* opim.wharton.upenn.edu/DPlab/papers/workingPapers/Minson_working_Ask%20(the%20Right%20Way)%20and%20You%20Shall%20Receive.pdf

?? `diabetes2` → Zeitler P, et al. 2012. *A Clinical Trial to Maintain Glycemic Control in Youth with Type 2 Diabetes.* N Engl J Med.

B.7 ??

?? Risso's dolphins → Endo T and Haraguchi K. 2009. *High mercury levels in hair samples from residents of Taiji, a Japanese whaling town.* Marine Pollution Bulletin 60(5):743-747.

Taiji was featured in the movie *The Cove*, and it is a significant source of dolphin and whale meat in Japan. Thousands of dolphins pass through the Taiji area annually, and we assumes these 19 dolphins reasonably represent a simple random sample from those dolphins.

?? Croaker white fish → fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

?? `run17` → www.cherryblossom.org

?? `textbooks`, `ucla_textbooks_f18` → Data were collected by OpenIntro staff in 2010 and again in 2018. For the 2018 sample, we sampled 201 UCLA courses. Of those, 68 required books that could be found on Amazon. The websites where information was retrieved: sa.ucla.edu/ro/public/soc, ucla.verbacompare.com, and amazon.com.

?? `stem_cells` → Menard C, et al. 2005. Transplantation of cardiac-committed mouse embryonic stem cells to infarcted sheep myocardium: a preclinical study. The Lancet: 366:9490, p1005-1012.

?? `ncbirths` → Birth records released by North Carolina in 2004. Unfortunately, we don't currently have additional information on the source for this data set.

?? Exam versions → Ovaj primjer je izmišljen.

?? Blood pressure statistics → The blood pressure standard deviation for patients with blood pressure ranging from 140 to 180 mmHg is guessed and may be a little (but likely not dramatically) imprecise from what we'd observe in actual data.

?? `toy_anova` → Data used for Figure ??, where this data was made up.

?? `mlb_players_18` → Data were retrieved from mlb.mlb.com/stats. Only players with at least 100 at bats were considered during the analysis.

?? `classdata` → Ovaj primjer je izmišljen.

B.8 ??

?? `simulated_scatter` → Fake data used for the first three plots. The perfect linear plot uses group 4 data, where `group` variable in the data set (Figure ??). The group of 3 imperfect linear plots use groups 1-3 (Figure ??). The sinusoidal curve uses group 5 data (Figure ??). The group of 3 scatterplots with residual plots use groups 6-8 (Figure ??). The correlation plots uses groups 9-19 data (Figures ?? and ??).

?? `possum` → Ovaj skup podataka opisan je u prilogu B.4.

?? `elmhurst` → These data were sampled from a table of data for all freshman from the 2011 class at Elmhurst College that accompanied an article titled *What Students Really Pay to Go to College* published online by *The Chronicle of Higher Education*: chronicle.com/article/What-Students-Really-Pay-to-Go/131435.

?? `simulated_scatter` → The plots for things that can go wrong uses groups 20-23 (Figure ??).

?? `mariokart` → Auction data from Ebay (ebay.com) for the game Mario Kart for the Nintendo Wii. This data set was collected in early October, 2009.

?? `simulated_scatter` → The plots for types of outliers uses groups 24-29 (Figure ??).

?? `midterms_house` → Data was retrieved from Wikipedia.

B.9 ??

?? `loans_full_schema` → Ovaj skup podataka opisan je u prilogu B.1.

?? `loans_full_schema` → Ovaj skup podataka opisan je u prilogu B.1.

?? `loans_full_schema` → Ovaj skup podataka opisan je u prilogu B.1.

?? `mariokart` → Ovaj skup podataka opisan je u prilogu B.8.

?? `resume` → Bertrand M, Mullainathan S. 2004. *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.* The American Economic Review 94:4 (991-1013). www.nber.org/papers/w9873

We did omit discussion of some structure in the data for the analysis presented: the experiment design included blocking, where typically four resumes were sent to each job: one for each inferred race/sex combination (as inferred based on the first name). We did not worry about this blocking aspect, since accounting for the blocking would *reduce* the standard error without notably changing the point estimates for the `race` and `sex` variables versus the analysis performed in the section. That is, the most interesting conclusions in the study are unaffected even when completing a more sophisticated analysis.