OpenIntro Statistics - 4th Edition
Chapter 1 Exercises
December 12, 2024
ODD SOLUTIONS

1. 1.1 Migraine and acupuncture, Part I.

A migraine is a particularly painful type of headache, which patients sometimes wish to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a randomized controlled study where 89 females diagnosed with migraine headaches were randomly assigned to one of two groups: treatment or control. 43 patients in the treatment group received acupuncture that is specifically designed to treat migraines. 46 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. Results are summarized in the contingency table below.

		Pain free		
		Yes	No	Total
Group	Treatment	10	33	43
	Control	2	44	46
	Total	12	77	89

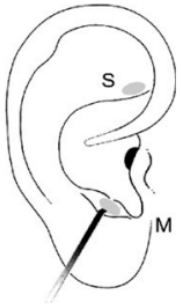


Figure from the original paper displaying the appropriate area (M) versus the inappropriate area (S) used in the treatment of migraine attacks.

(a) What percent of patients in the treatment group were pain free 24 hours after receiving acupuncture?

$10/43 \approx 0.23, \quad 23\%$

(b) What percent were pain free in the control group?

$2/46 \approx 0.04, \quad 4\%$

(c) In which group did a higher percent of patients become pain free 24 hours after receiving acupuncture?

The treatment group had a higher percent of patients becoming pain free at 23% compared to the control group's 4%.

(d) Your findings so far might suggest that acupuncture is an effective treatment for migraines for all people who suffer from migraines. However, this is not the only possible conclusion that can be drawn based on your findings so far. What is one other possible explanation for the observed difference between the percentages of patients that are pain free 24 hours after receiving acupuncture in the two groups?

Patients from the treatment group may have tried other methods for alleviating their pain outside of the experiment, within 24 hours of receiving acupuncture.

2. 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

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

What is the impact of air pollution on preterm births in Southern California?

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

The subjects in this study were people giving birth in Southern California between the years 1989 and 1993. $143,196$ **subjects were 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.

- **Length of gestation: numerical, continuous**
- **Date: categorical, ordinal**
- **Levels of carbon monoxide: numerical, continuous**
- **Levels of nitrogen dioxide: numerical, continuous**
- **Levels of ozone: numerical, continuous**
- **Levels of coarse particulate matter: numerical, continuous**

3. 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."

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

Are levels of honesty and self-control related to ones age or gender?

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

The subjects include 160 children between the ages of 5 and 15.

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

- **Age: numerical, discrete**
- **sex: categorical, nominal**
- **Only child: categorical, nominal**
- **Given explicit instructions: categorical, nominal**
- **Toss result: categorical, nominal**
- **Cheated: categorical, nominal**

4. 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?

- **Explanatory: Acupuncture insertion location**
- **Response: Migraine persistence after 24 hours**

5. 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.

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

50 **flowers from** 3 **different species creates** 150 **cases in the data.**

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

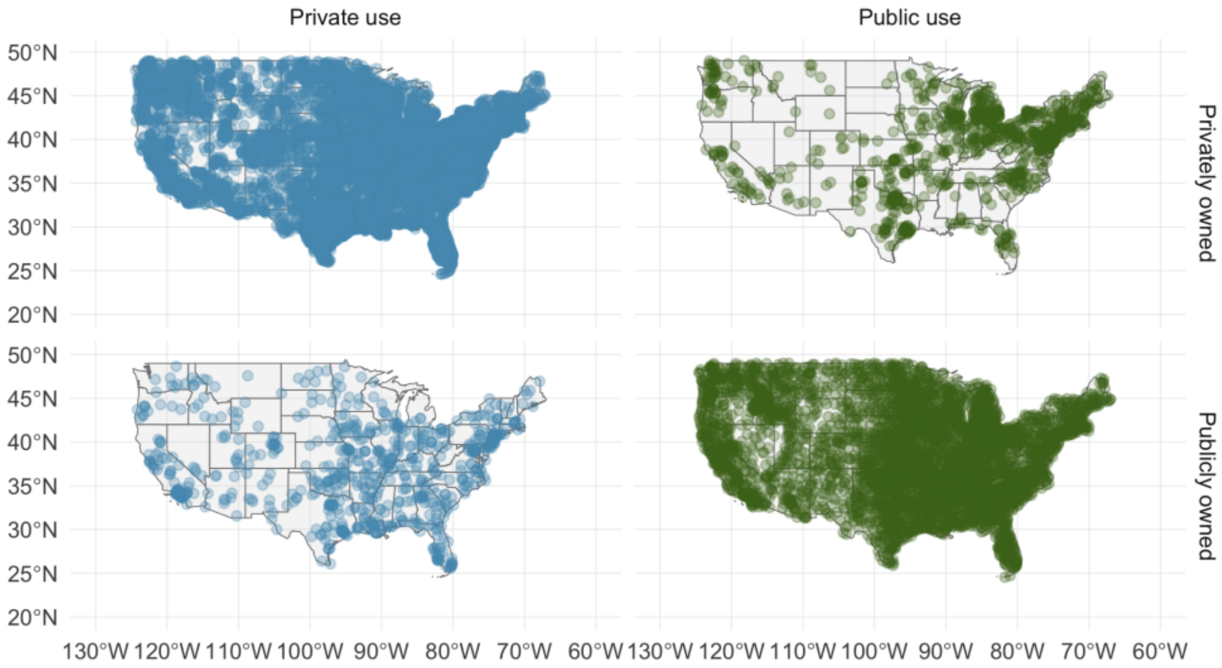
- **Sepal length: continuous**
- **Sepal width: continuous**
- **Petal length: continuous**
- **Petal width: continuous**

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

There is 1 nominal categorical variable in this data: species. The levels include setosa, versicolor and virginica.

6. 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.



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

- **Longitude**
- **Latitude**
- **Use by**
- **Owned by**

(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.

- **Longitude: numerical, continuous**
- **Latitude: numerical, continuous**
- **Use by: categorical, nominal**
- **Owned by: categorical, nominal**

7. 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.

   The population of interest is of all people who gave birth in Southern California between the years 1989 and 1993. The sample are the $143,196$ **births collected.**

   (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.

   The sample size of this study is large enough to conclude that the results can be generalized to the population of births in Southern California. However, since this was an observational study we could not determine any causal relationships between the two factors being analyzed.

8. 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.

   The population of interest are people who experience symptoms of asthma. The sample includes asthma patients aged 18-69 who relied on medication for asthma treatment.

   (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.

   Because of the relatively small size of the sample size, my confidence in the results representing the whole population of interest in not strong. However, this result was found through an experiment with randomly selected groups which infers a causal relationship between the Butkeyko method and asthma symptons/improve in Q.O.L.

9. 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.

   Observation

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

Variable

(c) 1.65.

Sample statistic

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

Population parameter

10. 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?

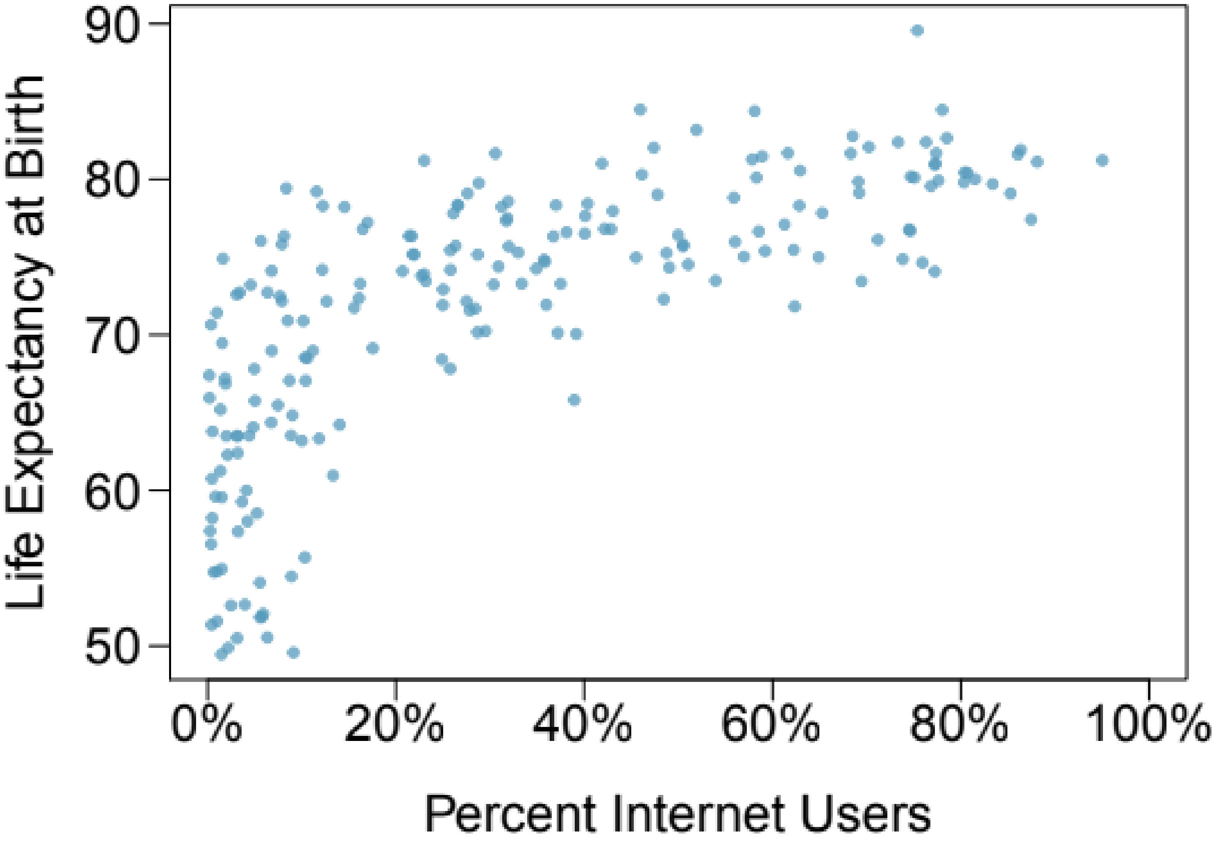
This study is an observational study since the data is being collected through the students' existing environments without interference.

(b) Suggest a sampling strategy for carrying out this study.

Since all students share the same lecture and we can safely assume they are all of similar ages, using a stratified sampling technique may be useful. Each lab section can represent a strata.

11. 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.



(a) Describe the relationship between life expectancy and percentage of internet users.

There is a positive correlation between these two variables.

(b) What type of study is this?

This study is observational because it collected its data from past records from different countries.

(c) State a possible confounding variable that might explain this relationship and describe its potential effect.

One possible confounding variable might be country GDP since the use of internet requires access to devices and infrastructure supporting an internet connection.

12. 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.

This method may not be reasonable for two reasons. A simple random sample may have a convenience bias. The relatively small sample size of students also may not accurately reflect the entire population's opinion.

(b) Stratify students by their field of study, then sample 10% of students from each stratum.

This method could work well. Using field of study as the variable to create each strata is easy to do and generalizes across the entire university. One concern may be that the number of students enrolled into each field of study may be skewed.

(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.

This may not work well since there are only a small handful of skewed clusters that can be created. Using this method might lead the same to completely ignore relevant groups of students such as freshman or seniors.

13. 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."

(a) What are the cases?

Each of the 200 individuals randomly sampled are cases.

(b) What is (are) the response variable(s) in this study?

The reaction to the microwave.

(c) What is (are) the explanatory variable(s) in this study?

The reaction to the subjects on the dispositional attitude.

(d) Does the study employ random sampling?

Yes.

(e) Is this an observational study or an experiment? Explain your reasoning.

This study is an observational study because the data this study collects are observations of each individuals dispositional attitude without any interference. The response variable was not collected randomly as each individual received the same number of positive and negative reviews, so the design of this study was not experimental.

(f) Can we establish a causal link between the explanatory and response variables?

No, since this was an observational study the results cannot infer a causal link between each variable.

(g) Can the results of the study be generalized to the population at large?

No. A sample size of 200 is extremely small relative to the population of even a single city. The paragraph does not specify where the samples were collected from or what the population of interest is.

14. 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.

Simple random sampling; non-response bias

(b) He gives out the survey only to his friends, making sure each one of them fills out the survey.

Cluster sampling; convenience bias

(c) He posts a link to an online survey on Facebook and asks his friends to fill out the survey.

Multistage sampling; convenience and non-response bias

(d) He randomly samples 5 classes and asks a random sample of students from those classes to fill out the survey.

Multistage sampling; convenience bias

15. 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?

Exam performance of students

(b) What is the explanatory variable? What are its levels?

Light levels; fluorescent overhead lighting, yellow overhead lighting, and no overhead lighting (only desk lamps)

(c) What is the blocking variable? What are its levels?

The blocking variable is sex; female and male

16. 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?

This is an experimental study since the study is designed by strategically controlling explanatory variables such as noise and measuring responses from groups.

(b) How many factors are considered in this study? Identify them, and describe their levels.

- **Sex: male, female**
- **Light level: fluorescent overhead lighting, yellow overhead lighting, and no overhead lighting (only desk lamps)**
- **Noise level: no noise, construction noise, and human chatter noise**

(c) What is the role of the sex variable in this study?

Sex is the blocking variable in this study because the researcher believes the explanatory variables may affect males and females differently which cannot be controlled.

17. 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.

Since the population is classmates in my class, I would employ simple random sampling on the class to collect the sample. This is an observational study so I would simply need to ask whether the sampled classmates prefer Coke or Diet Coke and record the responses. Explanatory variables might include age, sex, and major.