1.E: Introduction to Data (Exercises)

1.1: Case Study: Using Stents to Prevent Strokes

1.1 Migraine and acupuncture. 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 nonacupoint locations). 24 hours after patients received acupuncture, they were asked if they were pain free. Results are summarized in the contingency table below.\(^{52}\)

<table>
<thead>
<tr>
<th>Pain free: Yes</th>
<th>Pain free: No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>10</td>
<td>33</td>
</tr>
<tr>
<td>Control</td>
<td>2</td>
<td>44</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>77</td>
</tr>
</tbody>
</table>

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? What percent in the control group?
  
- b. At first glance, does acupuncture appear to be an effective treatment for migraines? Explain your reasoning.
c. Do the data provide convincing evidence that there is a real pain reduction for those patients in the treatment group? Or do you think that the observed difference might just be due to chance?

1.2 Sinusitis and antibiotics, Part I. Researchers studying the effect of antibiotic treatment for acute sinusitis compared to symptomatic treatments randomly assigned 166 adults diagnosed with acute sinusitis to one of two groups: treatment or control. Study participants received either a 10-day course of amoxicillin (an antibiotic) or a placebo similar in appearance and taste. The placebo consisted of symptomatic treatments such as acetaminophen, nasal decongestants, etc. At the end of the 10-day period patients were asked if they experienced significant improvement in symptoms. The distribution of responses are summarized below.

<table>
<thead>
<tr>
<th>Self-reported significant improvement in symptoms</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>66</td>
<td>19</td>
<td>85</td>
</tr>
<tr>
<td>Control</td>
<td>65</td>
<td>16</td>
<td>81</td>
</tr>
<tr>
<td>Total</td>
<td>131</td>
<td>35</td>
<td>166</td>
</tr>
</tbody>
</table>

a. What percent of patients in the treatment group experienced a significant improvement in symptoms? What percent in the control group?
b. At first glance, which treatment appears to be more effective for sinusitis?
c. Do the data provide convincing evidence that there is a difference in the improvement rates of sinusitis symptoms? Or do you think that the observed difference might just be due to chance?


1.2: Data Basics

1.3 Identify study components, Part I. Identify (i) the cases, (ii) the variables and their types, and (iii) the main research question in the studies described below.

a. Researchers collected data to examine the relationship between 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 $\mu g=\text{m}^3$. 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.  

b. 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 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.\footnote{J. McGowan. "Health Education: Does the Buteyko Institute Method make a difference?" In: Thorax 58 (2003).}

**1.4 Identify study components, Part II.** Identify (i) the cases, (ii) the variables and their types, and (iii) the main research question of the studies described below.

a. While obesity is measured based on body fat percentage (more than 35% body fat for women and more than 25% for men), precisely measuring body fat percentage is difficult. Body mass index (BMI), calculated as the ratio \(\frac{\text{weight}}{\text{height}^2}\), is often used as an alternative indicator for obesity. A common criticism of BMI is that it assumes the same relative body fat percentage regardless of age, sex, or ethnicity. In order to determine how useful BMI is for predicting body fat percentage across age, sex and ethnic groups, researchers studied 202 black and 504 white adults who resided in or near New York City, were ages 20-94 years old, had BMIs of 18-35 kg/m², and who volunteered to be a part of the study. Participants reported their age, sex, and ethnicity and were measured for weight and height. Body fat percentage was measured by submerging the participants in water.\footnote{Gallagher et al. "How useful is body mass index for comparison of body fatness across age, sex, and ethnic groups?" In: American Journal of Epidemiology 143.3 (1996), pp. 228-239.}

b. 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 they were for children in a nearby laboratory, but that they could take some if they wanted. Participants completed unrelated tasks and then reported the number of candies they had taken. It was found that those in the upper-class rank condition took more candy than did those in the lower-rank condition.\footnote{P.K. Pi et al. "Higher social class predicts increased unethical behavior". In: Proceedings of the National Academy of Sciences (2012).}


**1.5 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 owers (setosa, versicolor and virginica). There were 50 owers from each species in the data set.\footnote{https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Exercises_(Introductory_Statistics)/Exercises%3A_OpenIntro_…}

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).

**1.6 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.59

<table>
<thead>
<tr>
<th>gender</th>
<th>age</th>
<th>marital</th>
<th>grossIncome</th>
<th>smoke</th>
<th>amtWeekends</th>
<th>amtWeekdays</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Female</td>
<td>42</td>
<td>Single</td>
<td>Under $2,600</td>
<td>Yes</td>
<td>12 cig/day</td>
</tr>
<tr>
<td>2</td>
<td>Male</td>
<td>44</td>
<td>Single</td>
<td>$10,400 to $15,600</td>
<td>No</td>
<td>N/A</td>
</tr>
<tr>
<td>3</td>
<td>Male</td>
<td>53</td>
<td>Married</td>
<td>Above $36,400</td>
<td>Yes</td>
<td>6 cig/day</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1691</td>
<td>Male</td>
<td>40</td>
<td>Single</td>
<td>$2,600 to $5,200</td>
<td>Yes</td>
<td>8 cig/day</td>
</tr>
</tbody>
</table>

(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.

1.3: Overview of Data Collection Principles

1.7 Generalizability and causality, Part I. Identify the population of interest and the sample in the the studies described in Exercise 1.3. Also comment on whether or not the results of the study can be generalized to the population and if the ndings of the study can be used to establish causal relationships.

1.8 Generalizability and causality, Part II. Identify the population of interest and the sample in the the studies described in Exercise 1.4. Also comment on whether or not the results of the study can be generalized to the population and if the ndings of the study can be used to establish causal relationships.


59Stats4Schools, Smoking.

1.9 GPA and study time. A survey was conducted on 218 undergraduates from Duke University who took an introductory statistics course in Spring 2012. Among many other questions, this survey asked them about their GPA and the number of hours they spent studying per week. The scatterplot below displays the relationship between these two variables.

a. (a) What is the explanatory variable and what is the response variable?
   b. (b) Describe the relationship between the two variables. Make sure to discuss unusual observations, if any.
   c. (c) Is this an experiment or an observational study?
d. (d) Can we conclude that studying longer hours leads to higher GPAs?

1.10 Income and education. 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. (a) What are the explanatory and response variables?

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

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

1.4: Observational Studies and Sampling Strategies

1.11 Propose a sampling strategy. 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. (a) What type of study is this?

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

1.12 Internet use and life expectancy. The scatterplot below shows the relationship between estimated life expectancy at birth as of 201260 and percentage of internet users in 201061 in 208 countries.

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

b. (b) What type of study is this?

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

1.13 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.

1.14 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. Three research strategies for collecting data are described below. In each, name the sampling method proposed and any bias you might expect.

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

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

1.15 Family size. Suppose we want to estimate family size, where family is defined as one or more parents living with children. If we select students at random at an elementary school and ask them what their family size is, will our average be biased? If so, will it overestimate or underestimate the true value?
1.16 Flawed reasoning. Identify the aw 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. (a) Students at an elementary school are given a questionnaire that they are required to return after their parents have completed it. One of the questions asked is, "Do you nd 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. (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. (c) A orthopedist administers a questionnaire to 30 of his patients who do not have any joint problems and nds that 20 of them regularly go running. He concludes that running decreases the risk of joint problems.

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

(a) An article called Risks: Smokers Found More Prone to Dementia states the following: 62

"Researchers analyzed the data of 23,123 health plan members who participated in a voluntary exam and health behavior survey from 1978 to 1985, when they were 50 to 60 years old. Twenty-three years later, about one-quarter of the group, or 5,367, had dementia, including 1,136 with Alzheimers disease and 416 with vascular dementia. After adjusting for other factors, the researchers concluded that pack-a-day smokers were 37 percent more likely than nonsmokers to develop dementia, and the risks went up sharply with increased smoking: 44 percent 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 called The School Bully Is Sleepy states the following: 63

"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 identi ed 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 identi ed 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 justi ed? If not, how best can you describe the conclusion that can be drawn from this study?

1.18 Shyness on Facebook. Given the anonymity a afforded to individuals in online interactions, researchers hypothesized that shy individuals would have more favorable attitudes toward Facebook and that shyness would be positively correlated with time spent on Facebook. They also hypothesized that shy individuals would have fewer Facebook \Friends" just like they have fewer friends than non-shy individuals have in the offine world. Data were collected
on 103 undergraduate students at a university in southwestern Ontario via online questionnaires. The study states "Participants were recruited through the university's psychology participation pool. After indicating an interest in the study, participants were sent an e-mail containing the study's URL as well as the necessary login credentials." Are the results of this study generalizable to the population of all Facebook users?64


1.5: Experiments

1.19 Vitamin supplements. In order 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 medication groups, and the placebo group had the shortest duration of symptoms.65

a. (a) Was this an experiment or an observational study? Why?

b. (b) What are the explanatory and response variables in this study?

c. (c) Were the patients blinded to their treatment?

d. (d) Was this study double-blind?

e. (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.

1.20 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.

1.21 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. (a) What type of study is this?

b. (b) What are the treatment and control groups in this study?

c. (c) Does this study make use of blocking? If so, what is the blocking variable?
d. (d) Does this study make use of blinding?

e. (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. (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?


1.22 Chia seeds and weight loss. Chia Pets - those terra-cotta gurines 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 evenly 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.

a. (a) What type of study is this?

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

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

d. (d) Has blinding been used in this study?

e. (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.


1.6: Examining Numerical Data

1.23 Mammal life spans. Data were collected on life spans (in years) and gestation lengths (in days) for 62 mammals. A scatterplot of life span versus length of gestation is shown below.

a. (a) What type of an association is apparent between life span and length of gestation?

b. (b) What type of an association would you expect to see if the axes of the plot were reversed, i.e. if we plotted length of gestation versus life span?

c. (c) Are life span and length of gestation independent? Explain your reasoning.

1.24 Office productivity. Office productivity is relatively low when the employees feel no stress about their work or job security. However, high levels of stress can also lead to reduced employee productivity. Sketch a plot to represent the relationship between stress and productivity.

67 T. Allison and D.V. Cicchetti. "Sleep in mammals: ecological and constitutional correlates". In: Arch. Hydrobiol 75
1.25 Associations. Indicate which of the plots show a

a. (a) positive association
b. (b) negative association
c. (c) no association

Also determine if the positive and negative associations are linear or nonlinear. Each part may refer to more than one plot.

1.26 Parameters and statistics. Identify which value represents the sample mean and which value represents the claimed population mean.

a. (a) A recent article in a college newspaper stated that college students get an average of 5.5 hrs of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night.
b. (b) American households spent an average of about $52 in 2007 on Halloween merchandise such as costumes, decorations and candy. To see if this number had changed, researchers conducted a new survey in 2008 before industry numbers were reported. The survey included 1,500 households and found that average Halloween spending was $58 per household.
c. (c) The average GPA of students in 2001 at a private university was 3.37. A survey on a sample of 203 students from this university yielded an average GPA of 3.59 in Spring semester of 2012.

1.27 Make-up exam. In a class of 25 students, 24 of them took an exam in class and 1 student took a make-up exam the following day. The professor graded the rst batch of 24 exams and found an average score of 74 points with a standard deviation of 8.9 points. The student who took the make-up the following day scored 64 points on the exam.

a. (a) Does the new student's score increase or decrease the average score?
b. (b) What is the new average?
c. (c) Does the new student's score increase or decrease the standard deviation of the scores?

1.28 Days off at a mining plant. Workers at a particular mining site receive an average of 35 days paid vacation, which is lower than the national average. The manager of this plant is under pressure from a local union to increase the amount of paid time off. However, he does not want to give more days off to the workers because that would be costly. Instead he decides he should fire 10 employees in such a way as to raise the average number of days off that are reported by his employees. In order to achieve this goal, should he fire employees who have the most number of days off, least number of days off, or those who have about the average number of days off?

1.29 Smoking habits of UK residents, Part I. Exercise 1.6 introduces a data set on the smoking habits of UK residents. Below are histograms displaying the distributions of the number of cigarettes smoked on weekdays and weekends, excluding non-smokers. Describe the two distributions and compare them.

1.30 Stats scores. Below are the nal scores of 20 introductory statistics students.

\[79, 83, 57, 82, 94, 83, 72, 74, 73, 71, \]
Draw a histogram of these data and describe the distribution.

**1.31 Smoking habits of UK residents, Part II.** A random sample of 5 smokers from the data set discussed in Exercises 1.6 and 1.29 is provided below.

<table>
<thead>
<tr>
<th>gender</th>
<th>age</th>
<th>maritalStatus</th>
<th>grossIncome</th>
<th>smoke</th>
<th>amtWeekends</th>
<th>amtWeekdays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>51</td>
<td>Married</td>
<td>$2,600 to $5,200</td>
<td>Yes</td>
<td>20 cig/day</td>
<td>20 cig/day</td>
</tr>
<tr>
<td>Male</td>
<td>24</td>
<td>Single</td>
<td>$10,400 to $15,600</td>
<td>Yes</td>
<td>20 cig/day</td>
<td>15 cig/day</td>
</tr>
<tr>
<td>Female</td>
<td>33</td>
<td>Married</td>
<td>$10,400 to $15,600</td>
<td>Yes</td>
<td>20 cig/day</td>
<td>10 cig/day</td>
</tr>
<tr>
<td>Female</td>
<td>17</td>
<td>Single</td>
<td>$5,200 to $10,400</td>
<td>Yes</td>
<td>20 cig/day</td>
<td>15 cig/day</td>
</tr>
<tr>
<td>Female</td>
<td>76</td>
<td>Married</td>
<td>$5,200 to $10,400</td>
<td>Yes</td>
<td>20 cig/day</td>
<td>20 cig/day</td>
</tr>
</tbody>
</table>

a. (a) Find the mean amount of cigarettes smoked on weekdays and weekends by these 5 respondents.

b. (b) Find the standard deviation of the amount of cigarettes smoked on weekdays and on weekends by these 5 respondents. Is the variability higher on weekends or on weekdays?

**1.32 Factory defective rate.** A factory quality control manager decides to investigate the percentage of defective items produced each day. Within a given work week (Monday through Friday) the percentage of defective items produced was 2%, 1.4%, 4%, 3%, 2.2%.

a. (a) Calculate the mean for these data.

b. (b) Calculate the standard deviation for these data, showing each step in detail.

**1.33 Medians and IQRs.** For each part, compare distributions (1) and (2) based on their medians and IQRs. You do not need to calculate these statistics; simply state how the medians and IQRs compare. Make sure to explain your reasoning.

(a) (1) 3, 5, 6, 7, 9
(2) 3, 5, 6, 7, 20

(b) (1) 3, 5, 6, 7, 9
(2) 3, 5, 8, 7, 9

(c) (1) 1, 2, 3, 4, 5
(2) 6, 7, 8, 9, 10

(d) (1) 0, 10, 50, 60, 100
1.34 Means and SDs. For each part, compare distributions (1) and (2) based on their means and standard deviations. You do not need to calculate these statistics; simply state how the means and the standard deviations compare. Make sure to explain your reasoning. Hint: It may be useful to sketch dot plots of the distributions.

(a) (1) 3, 5, 5, 5, 8, 11, 11, 11, 13
(b) (1) -20, 0, 0, 0, 15, 25, 30, 30
(c) (1) 0, 2, 4, 6, 8, 10
(d) (1) 100, 200, 300, 400, 500

1.35 Box plot. Create a box plot for the data given in Exercise 1.30. The ve number summary provided below may be useful.

<table>
<thead>
<tr>
<th>Min</th>
<th>Q1</th>
<th>Q2 (Median)</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>72.5</td>
<td>78.5</td>
<td>82.5</td>
<td>94</td>
</tr>
</tbody>
</table>

1.36 Infant mortality. The infant mortality rate is defined as the number of infant deaths per 1,000 live births. This rate is often used as an indicator of the level of health in a country. The relative frequency histogram below shows the distribution of estimated infant death rates in 2012 for 222 countries.68

a. (a) Estimate Q1, the median, and Q3 from the histogram.

b. (b) Would you expect the mean of this data set to be smaller or larger than the median? Explain your reasoning.

1.37 Matching histograms and box plots. Describe the distribution in the histograms below and match them to the box plots.


1.38 Air quality. Daily air quality is measured by the air quality index (AQI) reported by the Environmental Protection Agency. This index reports the pollution level and what associated health effects might be a concern. The index is calculated for ve major air pollutants regulated by the Clean Air Act. and takes values from 0 to 300, where a higher value indicates lower air quality. AQI was reported for a sample of 91 days in 2011 in Durham, NC. The relative frequency histogram below shows the distribution of the AQI values on these days.69
a. (a) Estimate the median AQI value of this sample.

b. (b) Would you expect the mean AQI value of this sample to be higher or lower than the median? Explain your reasoning.

c. (c) Estimate Q1, Q3, and IQR for the distribution.

1.39 Histograms and box plots. Compare the two plots below. What characteristics of the distribution are apparent in the histogram and not in the box plot? What characteristics are apparent in the box plot but not in the histogram?

69 US Environmental Protection Agency, AirData, 2011.

1.40 Marathon winners. The histogram and box plots below show the distribution of finishing times for male and female winners of the New York Marathon between 1980 and 1999.

a. (a) What features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?

b. (b) What may be the reason for the bimodal distribution? Explain.

c. (c) Compare the distribution of marathon times for men and women based on the box plot shown below.

d. (d) The time series plot shown below is another way to look at these data. Describe what is visible in this plot but not in the others.

1.41 Robust statistics. The first histogram below shows the distribution of the yearly incomes of 40 patrons at a college coffee shop. Suppose two new people walk into the coffee shop: one making $225,000 and the other $250,000. The second histogram shows the new income distribution. Summary statistics are also provided.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>40</td>
<td>42</td>
</tr>
<tr>
<td>Min.</td>
<td>60,680</td>
<td>60,680</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>63,620</td>
<td>63,710</td>
</tr>
<tr>
<td>Median</td>
<td>65,240</td>
<td>65,350</td>
</tr>
<tr>
<td>Mean</td>
<td>65,090</td>
<td>73,300</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>66,160</td>
<td>66,540</td>
</tr>
<tr>
<td>Max.</td>
<td>69,890</td>
<td>250,000</td>
</tr>
<tr>
<td>SD</td>
<td>2,122</td>
<td>3,7321</td>
</tr>
</tbody>
</table>

a. (a) Would the mean or the median best represent what we might think of as a typical income for the 42 patrons at this coffee shop? What does this say about the robustness of the two measures?

b. (b) Would the standard deviation or the IQR best represent the amount of variability in the incomes of the 42
patrons at this coffee shop? What does this say about the robustness of the two measures?

1.42 Distributions and appropriate statistics. For each of the following, describe whether you expect the distribution to be symmetric, right skewed, or left skewed. Also specify whether the mean or median would best represent a typical observation in the data, and whether the variability of observations would be best represented using the standard deviation or IQR.

a. (a) Housing prices in a country where 25% of the houses cost below $350,000, 50% of the houses cost below $450,000, 75% of the houses cost below $1,000,000 and there are a meaningful number of houses that cost more than $6,000,000.

b. (b) Housing prices in a country where 25% of the houses cost below $300,000, 50% of the houses cost below $600,000, 75% of the houses cost below $900,000 and very few houses that cost more than $1,200,000.

c. (c) Number of alcoholic drinks consumed by college students in a given week.

d. (d) Annual salaries of the employees at a Fortune 500 company.

1.43 Commuting times, Part I.

The histogram to the right shows the distribution of mean commuting times in 3,143 US counties in 2010. Describe the distribution and comment on whether or not a log transformation may be advisable for these data.

1.44 Hispanic population, Part I. The histogram below shows the distribution of the percentage of the population that is Hispanic in 3,143 counties in the US in 2010. Also shown is a histogram of logs of these values. Describe the distribution and comment on why we might want to use log-transformed values in analyzing or modeling these data.

1.45 Commuting times, Part II. Exercise 1.43 displays histograms of mean commuting times in 3,143 US counties in 2010. Describe the spatial distribution of commuting times using the map below.

1.46 Hispanic population, Part II. Exercise 1.44 displays histograms of the distribution of the percentage of the population that is Hispanic in 3,143 counties in the US in 2010.

a. (a) What features of this distribution are apparent in the map but not in the histogram?

b. (b) What features are apparent in the histogram but not the map?

c. (c) Is one visualization more appropriate or helpful than the other? Explain your reasoning.

1.47 Antibiotic use in children. The bar plot and the pie chart below show the distribution of pre-existing medical conditions of children involved in a study on the optimal duration of antibiotic use in treatment of tracheitis, which is an upper respiratory infection.

a. (a) What features are apparent in the bar plot but not in the pie chart?

b. (b) What features are apparent in the pie chart but not in the bar plot?

c. (c) Which graph would you prefer to use for displaying these categorical data?

1.48 Views on immigration. 910 randomly sampled registered voters from Tampa, FL were asked if they thought…
workers who have illegally entered the US should be (i) allowed to keep their jobs and apply for US citizenship, (ii) allowed to keep their jobs as temporary guest workers but not allowed to apply for US citizenship, or (iii) lose their jobs and have to leave the country. The results of the survey by political ideology are shown below.  

<table>
<thead>
<tr>
<th>Political ideology</th>
<th>Conservative</th>
<th>Moderate</th>
<th>Liberal</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Apply for citizenship</td>
<td>57</td>
<td>120</td>
<td>101</td>
<td>278</td>
</tr>
<tr>
<td>(ii) Guest worker</td>
<td>121</td>
<td>113</td>
<td>28</td>
<td>262</td>
</tr>
<tr>
<td>(iii) Leave the country</td>
<td>179</td>
<td>126</td>
<td>45</td>
<td>350</td>
</tr>
<tr>
<td>(iv) Not sure</td>
<td>15</td>
<td>4</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>372</td>
<td>363</td>
<td>175</td>
<td>910</td>
</tr>
</tbody>
</table>

a. (a) What percent of these Tampa, FL voters identify themselves as conservatives?
b. (b) What percent of these Tampa, FL voters are in favor of the citizenship option?
c. (c) What percent of these Tampa, FL voters identify themselves as conservatives and are in favor of the citizenship option?
d. (d) What percent of these Tampa, FL voters who identify themselves as conservatives are also in favor of the citizenship option? What percent of moderates and liberal share this view?
e. (e) Do political ideology and views on immigration appear to be independent? Explain your reasoning.

1.49 Views on the DREAM Act. The same survey from Exercise 1.48 also asked respondents if they support the DREAM Act, a proposed law which would provide a path to citizenship for people brought illegally to the US as children. Based on the mosaic plot shown on the right, are views on the DREAM Act and political ideology independent?

1.50 Heart transplants, Part I. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. Each patient entering the program was designated an official heart transplant candidate, meaning that he was gravely ill and would most likely benefit from a new heart. Some patients got a transplant and some did not. The variable transplant indicates which group the patients were in; patients in the treatment group got a transplant and those in the control group did not. Another variable called survived was used to indicate whether or not the patient was alive at the end of the study. Figures may be found on the next page.

a. (a) Based on the mosaic plot, is survival independent of whether or not the patient got a transplant? Explain your reasoning.
b. (b) What do the box plots suggest about the efficacy (effectiveness) of transplants?

70 SurveyUSA, News Poll #18927, data collected Jan 27-29, 2012.

1.8: Case Study: Gender Discrimination (Special Topic)

1.51 Side effects of Avandia, Part I. Rosiglitazone is the active ingredient in the controversial type 2 diabetes medicine Avandia and has been linked to an increased risk of serious cardiovascular problems such as stroke, heart failure, and death. A common alternative treatment is pioglitazone, the active ingredient in a diabetes medicine called Actos. In a nationwide retrospective observational study of 227,571 Medicare beneficiaries aged 65 years or older, it was found that 2,593 of the 67,593 patients using rosiglitazone and 5,386 of the 159,978 using pioglitazone had serious cardiovascular problems. These data are summarized in the contingency table below.\textsuperscript{72}

\begin{tabular}{|c|c|c|}
\hline
& Cardiovascular problems & \\
Yes & No & Total \\
\hline
Rosiglitazone & 2,593 & 65,000 & 67,593 \\
Pioglitazone & 5,386 & 154,592 & 159,978 \\
Total & 7,979 & 219,592 & 227,571 \\
\hline
\end{tabular}

Determine if each of the following statements is true or false. If false, explain why. Be careful: The reasoning may be wrong even if the statement's conclusion is correct. In such cases, the statement should be considered false.

a. (a) Since more patients on pioglitazone had cardiovascular problems (5,386 vs. 2,593), we can conclude that the rate of cardiovascular problems for those on a pioglitazone treatment is higher.

b. (b) The data suggest that diabetic patients who are taking rosiglitazone are more likely to have cardiovascular problems since the rate of incidence was \( \frac{2,593}{67,593} = 0.038 \) 3.8% for patients on this treatment, while it was only \( \frac{5,386}{159,978} = 0.034 \) 3.4% for patients on pioglitazone.

c. (c) The fact that the rate of incidence is higher for the rosiglitazone group proves that rosiglitazone causes serious cardiovascular problems.

d. (d) Based on the information provided so far, we cannot tell if the difference between the rates of incidences is due to a relationship between the two variables or due to chance.


1.52 Heart transplants, Part II. Exercise 1.50 introduces the Stanford Heart Transplant Study. Of the 34 patients in the control group, 4 were alive at the end of the study. Of the 69 patients in the treatment group, 24 were alive. The contingency table below summarizes these results.
(a) What proportion of patients in the treatment group and what proportion of patients in the control group died?

(b) One approach for investigating whether or not the treatment is effective is to use a randomization technique.

i. What are the claims being tested?

ii. The paragraph below describes the set up for such approach, if we were to do it without using statistical software. Fill in the blanks with a number or phrase, whichever is appropriate.

We write alive on __________ cards representing patients who were alive at the end of the study, and dead on __________ cards representing patients who were not. Then, we shuffle these cards and split them into two groups: one group of size __________-representing treatment, and another group of size __________- representing control. We calculate the difference between the proportion of dead cards in the treatment and control groups (treatment - control) and record this value. We repeat this many times to build a distribution centered at __________. Lastly, we calculate the fraction of simulations where the simulated differences in proportions are __________. If this fraction is low, we conclude that it is unlikely to have observed such an outcome by chance and that the null hypothesis (independence model) should be rejected in favor of the alternative.

iii. What do the simulation results shown below suggest about the effectiveness of the transplant program?

1.53 Side effects of Avandia, Part II. Exercise 1.51 introduces a study that compares the rates of serious cardiovascular problems for diabetic patients on rosiglitazone and pioglitazone treatments. The table below summarizes the results of the study.

<table>
<thead>
<tr>
<th></th>
<th>Cardiovascular problems</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Rosiglitazone</td>
<td>2,593</td>
<td>65,000</td>
</tr>
<tr>
<td>Pioglitazone</td>
<td>5,386</td>
<td>154,592</td>
</tr>
<tr>
<td>Total</td>
<td>7,979</td>
<td>219,592</td>
</tr>
</tbody>
</table>

a. (a) What proportion of all patients had cardiovascular problems?

b. (b) If the type of treatment and having cardiovascular problems were independent, about how many patients in the rosiglitazone group would we expect to have had cardiovascular problems?
c. (c) We can investigate the relationship between outcome and treatment in this study using a randomization technique. While in reality we would carry out the simulations required for randomization using statistical software, suppose we actually simulate using index cards. In order to simulate from the independence model, which states that the outcomes were independent of the treatment, we write whether or not each patient had a cardiovascular problem on cards, shuffled all the cards together, then deal them into two groups of size 67,593 and 159,978. We repeat this simulation 1,000 times and each time record the number of people in the rosiglitazone group who had cardiovascular problems. Below is a relative frequency histogram of these counts.

a. i. What are the claims being tested?

b. ii. Compared to the number calculated in part (b), which would provide more support for the alternative hypothesis, more or fewer patients with cardiovascular problems in the rosiglitazone group?

c. iii. What do the simulation results suggest about the relationship between taking rosiglitazone and having cardiovascular problems in diabetic patients?

1.54 Sinusitis and antibiotics, Part II. Researchers studying the effect of antibiotic treatment compared to symptomatic treatment for acute sinusitis randomly assigned 166 adults diagnosed with sinusitis into two groups (as discussed in Exercise 1.2). Participants in the antibiotic group received a 10-day course of an antibiotic, and the rest received symptomatic treatments as a placebo. These pills had the same taste and packaging as the antibiotic. At the end of the 10-day period patients were asked if they experienced improvement in symptoms since the beginning of the study. The distribution of responses is summarized below.\(^{73}\)

<table>
<thead>
<tr>
<th>Self reported improvement in symptoms</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antibiotic</td>
<td>66</td>
<td>19</td>
<td>85</td>
</tr>
<tr>
<td>Placebo</td>
<td>65</td>
<td>16</td>
<td>81</td>
</tr>
<tr>
<td>Total</td>
<td>131</td>
<td>35</td>
<td>166</td>
</tr>
</tbody>
</table>

a. (a) What type of a study is this?

b. (b) Does this study make use of blinding?

c. (c) At first glance, does antibiotic or placebo appear to be more effective for the treatment of sinusitis? Explain your reasoning using appropriate statistics.

d. (d) There are two competing claims that this study is used to compare: the independence model and the alternative model. Write out these competing claims in easy-to-understand language and in the context of the application. Hint: The researchers are studying the effectiveness of antibiotic treatment.

e. (e) Based on your nding in (c), does the evidence favor the alternative model? If not, then explain why. If so, what would you do to check if whether this is strong evidence?

Contributors and Attributions

David M Diez (Google/YouTube), Christopher D Barr (Harvard School of Public Health), Mine Çetinkaya-Rundel (Duke University)