Preliminaries: Multivariable thinking and Sources of variation

Preliminaries Learning Goals:

- Identify and apply basic terminology of statistical studies: observational units, response variable, explanatory variable, association, confounding variable
- Identify potential sources and measures of variation in a response variable
- Produce and describe some basic visualizations and numerical summaries to compare groups and explore relationships (e.g., bar graphs, dotplots/histograms/boxplots, scatterplots, means, medians, standard deviation)
- Explore how those comparisons and relationships can be impacted by additional variables
- Calculate a residual and relate it to typical prediction error

Introduction

How many undergraduate colleges/universities did you apply to? Did you get into all of them? Did you wonder how these schools were making their decisions for admission? Did you wonder whether maybe there were certain variables that impacted the likelihood you would be accepted? Did you wonder whether males or females would be more likely to be accepted?

Outcomes are not always the same; you probably know this from your college application results—you were accepted to some schools and not others. Even for the same school, some of your classmates were accepted and some weren't. It is this ***variability*** in outcomes that statisticians are most interested in explaining. For example, why are some people accepted into one program and others are not? Why do some workers earn higher wages than others? Does the size of a house help predict the price of a house? Does length of a pregnancy help predict a baby's weight?

One of the main goals in this course is to examine the variation in outcomes or in a ***response variable***, and to determine how much of that variation can be explained by relationships with ***explanatory*** *(or* ***predictor) variables*** (called systematic variation) versus how much variation is still left unexplained (how accurate are the predictions likely to be). You probably did this in your first statistics course. For example, response variable: acceptance or not, explanatory variable: applicant is male or female. Or response variable: price of house, and explanatory variable: size. But most real-world studies consider more, sometimes many more, than one explanatory variable. For example, in a study showing a link between higher rates of depression and poor diet, we know that other lifestyle factors such as exercise, drinking alcohol, socioeconomic status, etc. may also be responsible for the observed differences in depression, rather than diet alone. Doctors can make better decisions about the impacts of diet on depression if they understand the simultaneous roles of all these variables. By bringing more variables into the study, we are likely able to explain more variation in the response variable (reduce the "noise") and therefore will be able to make better predictions. But, we also need to keep in mind that observed relationships between two variables may change when we consider a third or fourth or fifth variable. For example, perhaps once you know the size of a lot it is no longer useful to know the size of the house.

In this Preliminaries chapter, we will review some of the basic ideas from your first course but also focus on *multivariable thinking* and how this fits into the broader statistical investigation.

- How additional variables can impact relationships between the variables of interest
- How explaining variation in the response variable can reduce prediction error

Being able to brainstorm different sources of variation in the response variable and understanding different study designs will help you become a better consumer of statistical information, by training you to ask good questions about studies and the variables to be measured.

Example P.A: Graduate School Admissions at Berkeley

In the early 1970s, the University of California at Berkeley was concerned with possible discrimination against women in its graduate admissions process. Data about the applicants for the 1972–73 school year were recorded from several programs, including their sex and whether or not they were accepted (Bickel & O'Connell, *Science,* 1975).

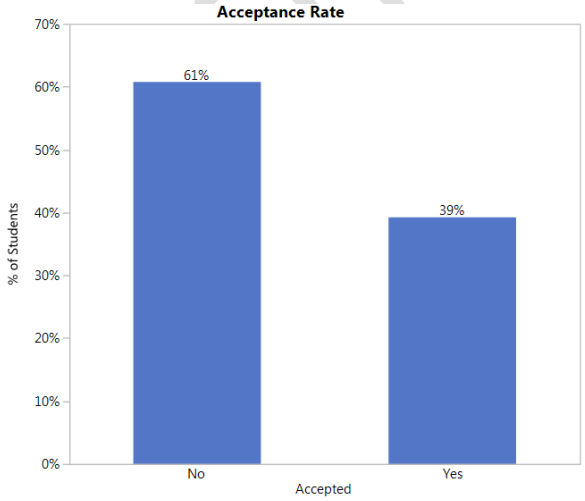
Definition: In a statistical analysis, we start by identifying

- the ***observational units*** (the people or objects we will be taking measurements on). The collection of observational units we collect data on is our ***sample.***
- the ***variables*** of study (the measurements taken on the observational units).
    - The ***response variable*** measures the outcomes of interest/what we are trying to predict.
    - We will classify the variables as ***quantitative*** (e.g., numerical) or ***categorical*** (identifying groups or categories that the observational units belong to).

Think about it: What are the observational units in this study? What is the response variable? Is the response variable quantitative or categorical?

The observational units are the applicants to Berkeley's graduate program in 1972–73. The response variable here is whether or not the applicant was accepted. We can visualize the ***distribution*** of this categorical variable using a ***bar graph*** (Figure P.1).

Figure P.1: Bar graph of application outcome for Berkeley's 1973 graduate admissions



There were different outcomes for different applicants—some students were accepted and some weren't. In fact, Berkeley's overall graduate school acceptance rate in 1973 was 39%—not an easy school to get into!

So the question becomes, can we explain some of the variation in whether or not someone is accepted? In other words, is there information we could collect that would help us predict whether or not someone would be accepted? As mentioned above, researchers became intrigued by a perceived difference in the likelihood of being accepted in Berkeley's graduate program between males and females. That is, they

wondered whether sex would explain some of the variation between being accepted (39%) and not being accepted (61%).

Definition: The variable that we believe predicts or helps explain the outcomes of the response variable is often called the ***explanatory variable***.

Think about it: Is *sex* a quantitative or a categorical variable? How might we organize and summarize data on two categorical variables?

With two categorical variables, the data can be arranged in a ***2×2 contingency table*** to show the relationship between the explanatory variable, sex of applicant, and the response variable, acceptance (Table P.1). Note: There were many academic departments investigated in the original Berkeley study, for simplicity we will focus for now on the combined data from just two of the larger departments (Freedman, Pisani, & Purves, 2007). See the HW exercises for more on this study.

Table P.1: Contingency table with Berkley graduate school acceptance counts by gender

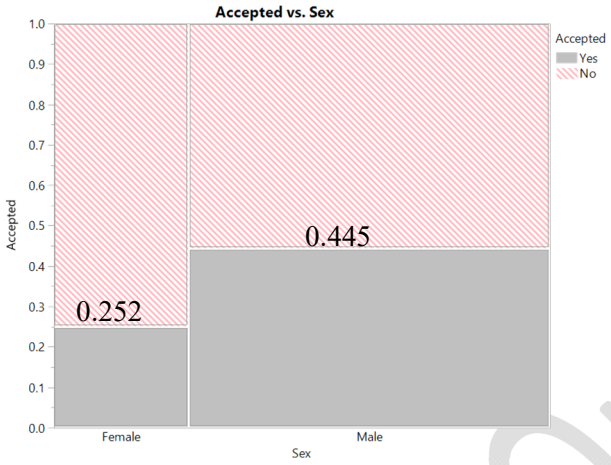
		Explanatory variable		
		Male Applicant	**Female Applicant**	**Total**
Response variable	**Accepted**	533	113	646
	Not accepted	665	336	1001
	Total	1198	449	1647

If we simply compare the ***counts***, you see that more males were accepted than females. However, more males were also denied! When the sample sizes (number of males and females) are not the same, it is much more informative to compare the ***conditional proportions*** (or percentages) accepted – the proportion of males accepted compared to the proportion of females accepted. See Figure P.2 for a ***mosaic plot*** using conditional proportions rather than counts.

Definition: A ***mosaic plot*** is a ***segmented bar graph*** (a different bar for each explanatory variable group broken down by the proportion in each response variable category) where the widths of the bars reflect the relative size of the explanatory variable categories.

Figure P.2 show that, proportionally, males were closer to 50/50 in being accepted or denied (533/1198 ≈ 44.5%), but the majority of females were denied (1336/449 ≈ 74.8%). Also, note the male bar is wider than the female bar in the graph. This reveals that more males applied than females. In fact, almost 3 times more males than females applied to these two programs

Figure P.2: Mosaic plot for acceptance/non acceptance rates for males vs for females



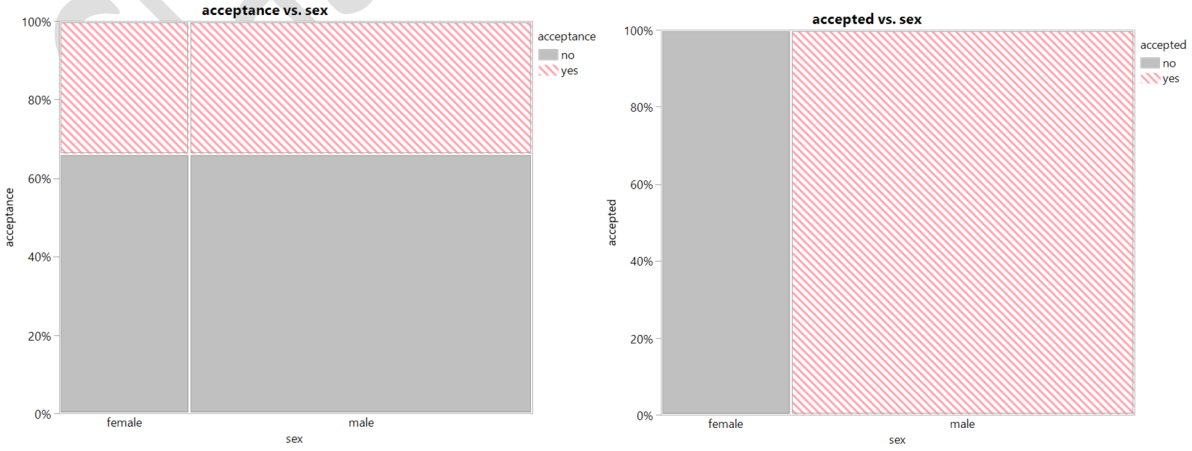
Because the (conditional) distribution of acceptance for females is not the same as the (conditional) distribution for males (the male bar and the female bar differ), we say that sex and acceptance are ***associated*** in this sample.

Definition: Two variables are ***associated*** if the conditional distribution of one variable changes depending on the explanatory variable value on which you are conditioning.

Think about it: What would the mosaic plot look like if sex was not associated with acceptance? What would the plot look like if sex was perfectly associated with acceptance?

Figure P.3 shows two hypothetical mosaic plots. The graph on the left shows no association between sex and acceptance – both males and females are denied approximately 65% of the time. When there is no association, the percentage accepted doesn't need to be 50% for each sex, there just needs to be the same percentage for males and females. The graph on the right shows perfect association – all men are accepted and all women are denied. In other words, for the graph on the right, knowing the applicant's sex would explain all of the variation in acceptance: we would be able to make perfect predictions just by knowing whether someone was male or female!

Figure P.3: Hypothetical mosaic plots shows (a) no association and (b) perfect association



Most studies will have an association somewhere between these two extremes (indeed, one of the key questions in statistics is whether the observed association could have happened by random chance alone). As we saw earlier, in the real data we have explained some of the variability by knowing the applicant's sex, but not all of it. Clearly there are some other variables at play here as well.

Think about it: What other variables might help us predict whether or not someone would be accepted?

We said earlier that the data presented here were for just two of the graduate programs. We have easy access to this variable (graduate program), so let's examine the association between acceptance and sex separately for each program. In other words, we will *condition* on the program applied to, as shown in Figure P.4. below. Note: In the original study, the administrators could condition on program but in the publically available data, university policy did not allow the individual programs to be identified, so we will call them Program A and Program F.

Think about it: Does program appear to be associated with acceptance? Does knowing the program help us further explain variation in who is accepted? What do you notice about the association between acceptance and sex within each program compared to when we combined the data altogether?

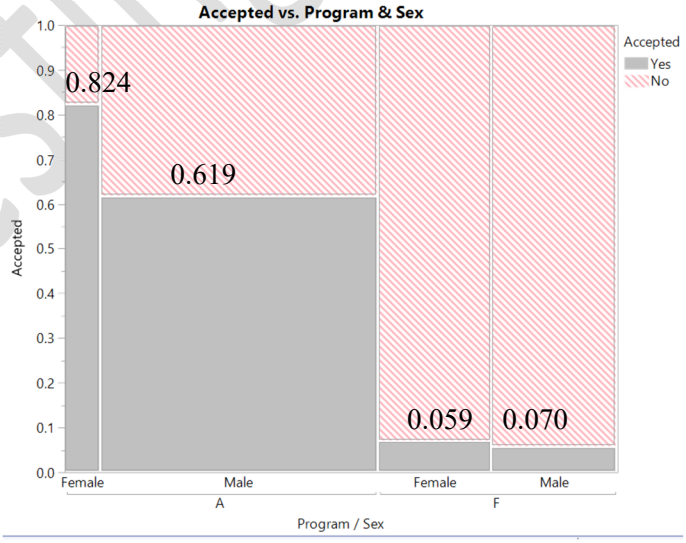
Figure P.4: Two-way tables and Mosaic plots for acceptance vs. sex separately for each program

Program A

	Male Applicant	Female Applicant	Total
Accepted	511	89	600
Not accepted	314	19	333
Total	825	108	933

Program F

	Male Applicant	Female Applicant	Total
Accepted	22	24	46
Not accepted	351	317	668
Total	373	341	714



For Program A, 89/108 or about 82% of the female applicants were accepted and 511/825 or about 62% of the male applicants were accepted. But, if we focus just on Program F, only about 7% of the females and 6% of the male applicants were accepted. So, program does appear to be associated with acceptance – Program A was much easier to get in to than Program F. Our predictions still would not be perfect, but they would certainly be better if we knew the program and the sex of the applicant.
But notice one more feature to the conditional associations. In Program A, females were actually accepted at a higher rate than males! In Program F, the conditional acceptance proportions were similar, but also slightly higher for females than for males.

Think about it: How does your conclusion about the way sex associates with acceptance change from when we looked at sex alone, when we consider program as well?

When we consider the programs separately, the conditional associations within each program are in the opposite direction from the overall association that pooled the data across the programs (recall overall, men were accepted 45% of the time and women were accepted 25% of the time)! If we hadn't looked program by program, we would have drawn the wrong conclusion!

Key Idea: The associations within each subgroup can look quite different from the overall association.

This reversal of the direction of the association between acceptance and sex is referred to as ***Simpson's Paradox*** (named after a British statistician, one of the first to write about it.) Although Simpson's Paradox is not very common in practice, it is good to be on the lookout for, and be able to explain, such a phenomenon. Later in the course, we spend more time exploring the more general case of when the association between two variables looks different when considering a third variable – the case of a statistical *interaction*.

Think about it: How can it be that females have higher (conditional) acceptance rates than males in Program A and in Program F, but when we combine the two programs together, the overall acceptance rate is noticeably smaller for the females? *Hint*: What else do you learn from the mosaic plots?

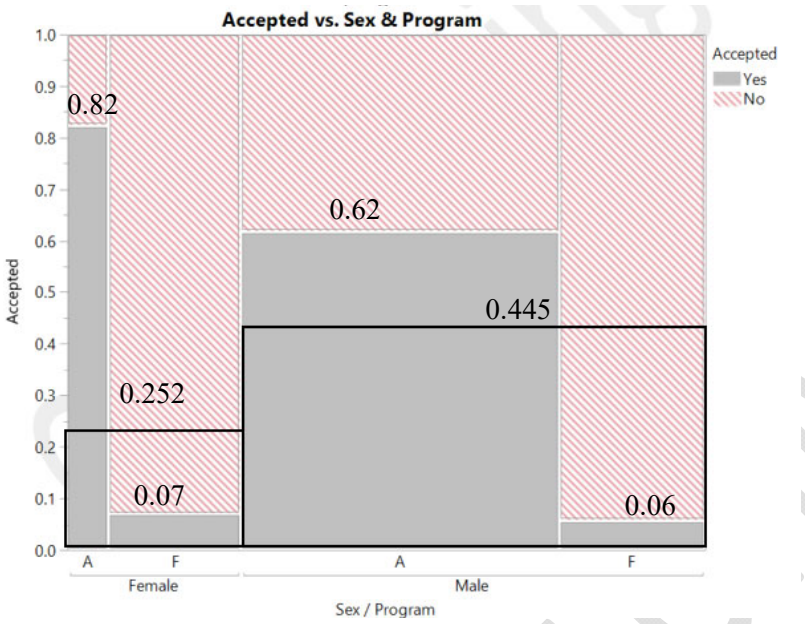
For these data, not only is program associated with acceptance, we also see an association between program and sex – In Program A (perhaps Engineering?), most applicants were male, whereas in Program F (perhaps English?), it was a little more equal between males and females. Another way to describe this association, males were much more likely to apply to Program A than F, and females were more likely to apply to Program F than A. Combine this with the higher acceptance rate to program A, and the overall acceptance rate for males is pulled larger (towards the 0.62) than the overall acceptance rate for females which is going to be closer to the 0.07 (see Figure P.5).

Note that if program had not been associated with both gender and acceptance, then we wouldn't see Simpson's Paradox.

Definition: A ***confounding variable*** is a third variable that is related to both the explanatory variable and the response variable. You may recall from your first statistics course that we always need to be concerned about confounding variables with ***observational studies***, as they may provide an alternative explanation for an observed association between the explanatory and response variables, preventing cause-and-effect conclusions.

Thus, in this case, we can say that program is a confounding variable on the association between sex and acceptance. Program is associated with both sex and acceptance and, furthermore, when taken into account, changes the observed association between sex and acceptance.

Figure P.5: Acceptance by Sex bar graph superimposed on the "three-way" mosaic plot



Key Idea: One way to account for confounding variables is conditioning on the third variable as we have done here, but there could still be other confounding variables we don't know about!

Think about it: Does knowing the program applied to and the sex of the application explain *all* of the variation in admission decisions?

Knowing these two variables does not allow us to make perfect predictions on someone's acceptance, there are still other "sources of unexplained variation." Recall from earlier that in order to have a perfect association we would have to have one sex 100% accepted and the other 100% denied.

We can diagram what we have learned so far with a Sources of Variation diagram (Figure P.6), where we use arrows to indicate observed associations between variables.

Key Idea: A ***Sources of Variation*** diagram is a visual representation of our belief about possible sources which explain variation in the response variable.

Figure P.6: Sources of Variation diagram for Graduate Admissions at Berkeley study

<table>
<tr><td>Observed Variation in:
Acceptance (Yes or No)</td><td rowspan="2">Sources of explained variation
• Sex (male or female)
• Program</td><td rowspan="2">Sources of unexplained variation
• Quality of application
• Numerical data (e.g., test scores, grade point average)
• Unknown..</td></tr>
<tr><td>Inclusion criteria
• Class (graduate students)
• School (Berkeley)</td></tr>
</table>

This diagram also includes a box for "inclusion criterion." In this case, we only have data on graduate programs at UC Berkeley, but clearly which school someone applies to and whether someone is looking at graduate or undergraduate programs will impact whether or not someone is accepted. But which school and which level of program being applied to don't impact variability in acceptance for the data we

examined, however, they definitely limit the ***generalizability*** of our study conclusions—any associations we find here do not necessarily apply to other schools and programs. We have also listed several other possible sources which are not able to account for because the data are not available to us, but it's important to consider their possible impact and provide suggestions for variables to measure in future studies.

Definition: The observational units we collect data on are referred to as the ***sample***. Typically this is not the entire group of observational units we are interested in, but rather we would like to apply the conclusions to a larger ***population***. ***Generalizability*** refers to deciding an appropriate population to which we can generalize our conclusions. What larger group do you think these results are representative of?

The six-steps of a statistical investigation

Many people think that statistical investigations are primarily about "number crunching," but as we've seen here, there is a lot more going on than merely computing a few numbers. One way to conceptualize the overarching methods for statistically investigating research questions, is by thinking about the six-steps that most statistical investigations should follow.

• **STEP 1: Ask a research question** that can be addressed by collecting data. These questions often involve comparing groups, asking whether something affects something else, or assessing people's opinions.

• **STEP 2: Design a study and collect data.** This step involves selecting the people or objects to be studied, deciding how to gather relevant data on them, and carrying out this data collection in a careful, systematic manner.

• **STEP 3: Explore the data**, looking for patterns related to your research question as well as unexpected outcomes that might point to additional questions to pursue. It may also be possible to develop a *statistical model* of the data generating process to try to predict future observations.

• **STEP 4: Draw inferences beyond the data** by determining whether any findings in your data reflect a genuine tendency, and estimating the size of that tendency.

• **STEP 5: Formulate conclusions** that consider the scope of the inference made in Step 4. To what underlying process or larger group can these conclusions be generalized? Is a cause-and-effect conclusion warranted?

• **STEP 6: Look back and ahead** to point out limitations of the study and suggest new studies that could be performed to build on the findings of the study.

Let's apply the six-steps to this study investigating the possibility of discrimination against women in graduate school admissions.

STEP 1: Ask a research question. Is there evidence of discrimination against women in graduate school admissions at UC Berkeley? How does this evidence differ by program?

STEP 2: Design a study and collect data. Data were reported on 1647 individuals who applied to two different UC Berkeley graduate programs in 1973. For each student three variables were measured: admittance (yes/no), sex (male/female) and program (A or F).

STEP 3: Explore the data. We find that whereas overall men are accepted at a higher rate than women (45% vs. 25%), within each program the acceptance rates are higher for women than for men (Program A: 82% vs. 62%; Program F 7% vs. 6%).

STEP 4: Draw inferences beyond the data. In later chapters we will discuss how to use *tests of significance* and *confidence intervals* to draw inferences beyond the data. In this case, for example, we could determine whether acceptance rates between women and men within program are *significantly* different, larger then we might expect to see by random chance alone.

STEP 5: Formulate conclusions. We already noted that these data were from a single year of graduate admissions, at a single university for only two programs, so we need to be cautious about how far we generalize our conclusions. Furthermore, cause-effect conclusions are not possible here due to the fact that this is observational data, and not from a randomized experiment (something we will explore in more detail later in the next chapter).

STEP 6: Look back and ahead. So what do we tell the administrators who were worried about discrimination? We tell them that comparing the initial acceptance percentages for males and females (44.5% and 25.2%) is not very meaningful, but considering the percentages separately for each program (e.g., in program A of 61.9% (for males) and 82% (for females)) we get a much more meaningful comparison of the acceptance rates, one that removes the confounding explanation of program type. The issue here does not appear to be with the admissions process. Future work might explore whether these trends hold across other years, universities, programs and demographic characteristics (e.g., race) to ensure that evidence of discrimination against women or others is not present.

Exploration P.A: Salary Discrimination

In the United States, the 1963 Equal Pay Act requires that men and women be given equal pay for equal work and Title VII of the Civil Rights Act of 1964 prohibits discrimination on the basis of race, color, religion, sex, and national origin. But, how successful have these acts been?

The data file ***WageRace*** contains observations from the 1988 March U.S. Current Population Survey on 1987 weekly wages (in 1992 dollars) for a sample of 25,631 males between the age of 18 and 70 who worked full-time, along with their years of education, years of experience, race, whether they worked in a standard metropolitan statistical area, and the region of the U.S. where they worked. (Source: Ramsey and Shafer, 2002)

The primary question is whether wages for blacks differ significantly from wages for non-blacks.

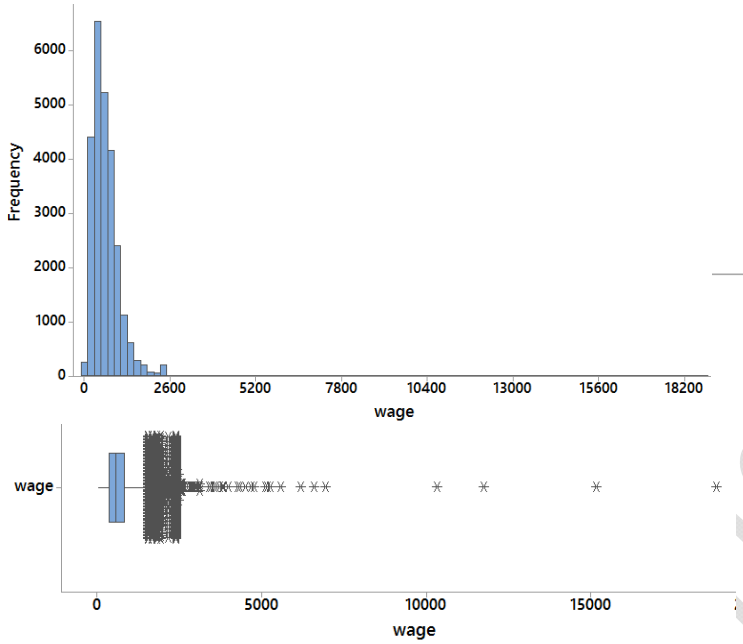
Definition: In a statistical analysis, we start by identifying the ***observational units*** (the people or objects we will be taking measurements on) and the variable of interest (the measurements taken on the observational units). Variables are usually classified as ***quantitative*** (e.g., numerical) or ***categorical*** (identifying groups or categories that the observational units belong to).

1. Identify the observational units in this study. How many are there?

2. Is the wages variable a *quantitative* or *categorical* variable?

Figure P.7 summarizes the wage *distribution* (the pattern of the possible values and how often they occur) for these 25,361 males.

Figure P.7: Wage *distribution* for 25,361 males in 1987



Mean	640.2
StDev	444.3
N	25631
Minimum	50.4
1st Quartile	356.1
Median	567.2
3rd Quartile	826.2
Maximum	18777.2

3. Why are we looking at histograms and boxplots rather than a bar graph (as in Example P.1)?

4. Does anything stand out to you about the boxplot that is less obvious in the histogram?

5. Which visual, the histogram or boxplot, do you like better? Why?

6. Do the mean and median wage differ? Which is larger? Why?

7. Do the wages appear to follow a normal distribution? How are you deciding?

Definition: When we have more than one variable, we often identify one as the ***response variable*** (the primary outcome of interest) and one as the ***explanatory variable*** (a variable that we believe predicts or helps explain the outcome of the response variable).

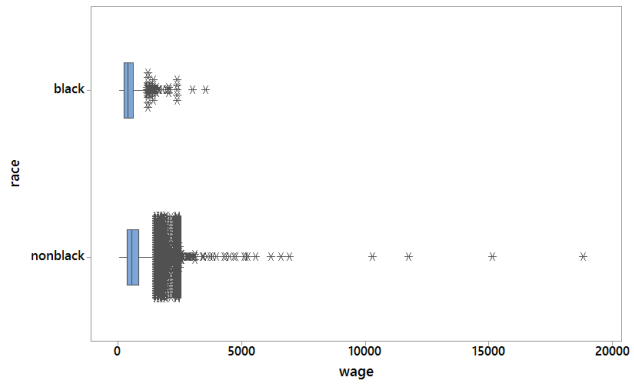
In this study, the researchers were most interested in whether the person's race explained differences in wages.

8. Which variable is the explanatory variable? Which is the response variable?

9. Make a prediction: Do you think that the explanatory variable explains some variation in the response variable? Do you think that the explanatory variable explains *all* of the variation in the response variable? Why or why not?

Figure P.8 *conditions* the wage distributions by whether the male was "black" or "nonblack."

Figure P.8: Conditional distributions of wages for the blacks and non-blacks



Variable	race	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
wage	black	1988	478.73	307.81	53.83	261.16	412.29	617.28	3527.34
	nonblack	23643	653.7	451.3	50.4	367.9	569.8	831.0	18777.2

10. Consider whether there appears to be an association between wage and race: Does the wage distribution differ substantially between blacks and non-blacks? What is the difference in the mean weekly wages for blacks and non-blacks? Can we conclude wage discrimination? Explain.

The output shows differences between the two distributions including a shift to the right for the non-blacks, and more variability for non-blacks as well. With non-blacks making about $175 more per week, on average, than blacks ($654 – $479 = $175), this is some evidence of wage discrimination based on race that the Civil Rights Act of 1964 was supposed to prohibit.

Definition: Two variables are ***associated*** if the conditional distribution of one variable changes depending on the explanatory variable value on which you are conditioning.

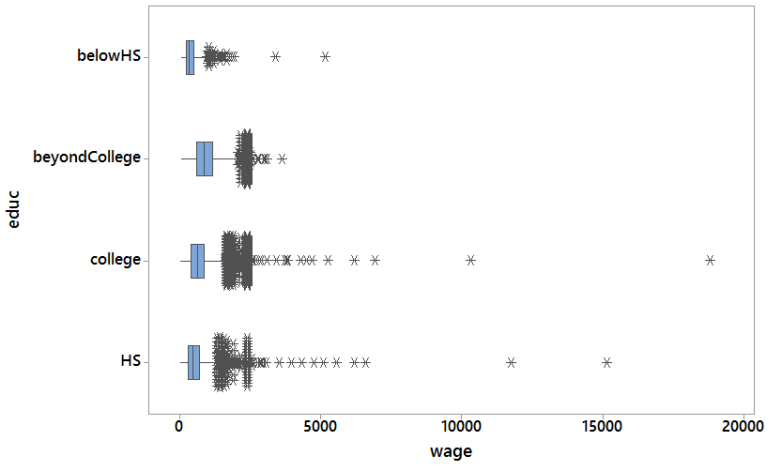
But, this story has some limitations. Although we can conclude that there is an ***association*** between the weekly wages and race, or in other words race explains some of the variation in weekly wages, we cannot draw any cause-and-effect conclusions from this *observational study*. You may recall from your first

statistics course that observational studies are prone to ***confounding variables***, which may be able to provide an alternative explanation for the observed association.

Definition: A ***confounding variable*** is a third variable that is related to both the explanatory variable and the response variable. You may recall from your first course that we always need to be concerned about confounding variables with ***observational studies***, as they may provide an alternative explanation for an observed association between the explanatory and response variables, preventing cause-and-effect conclusions.

For example, we might expect level of education to also be related to weekly wages. The boxplots in Figure P.9 condition the wage distributions on the level of education.

Figure P.9: Boxplots for wage distributions conditional on level of education



Variable	Education	Count	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
wage	belowHS	1509	413.57	281.86	61.73	237.42	356.13	522.32	5144.03
	HS	12220	544.0	366.8	50.4	316.6	480.3	712.3	15123.5
	College	8977	703.1	475.6	54.3	412.8	636.3	884.2	18777.2
	BeyondCollege	2925	965.80	499.12	59.35	617.28	878.44	1187.08	3600.82

11. Suggest an easy way to improve this graphical display to better focus on a trend of increasing salaries with increasing education.

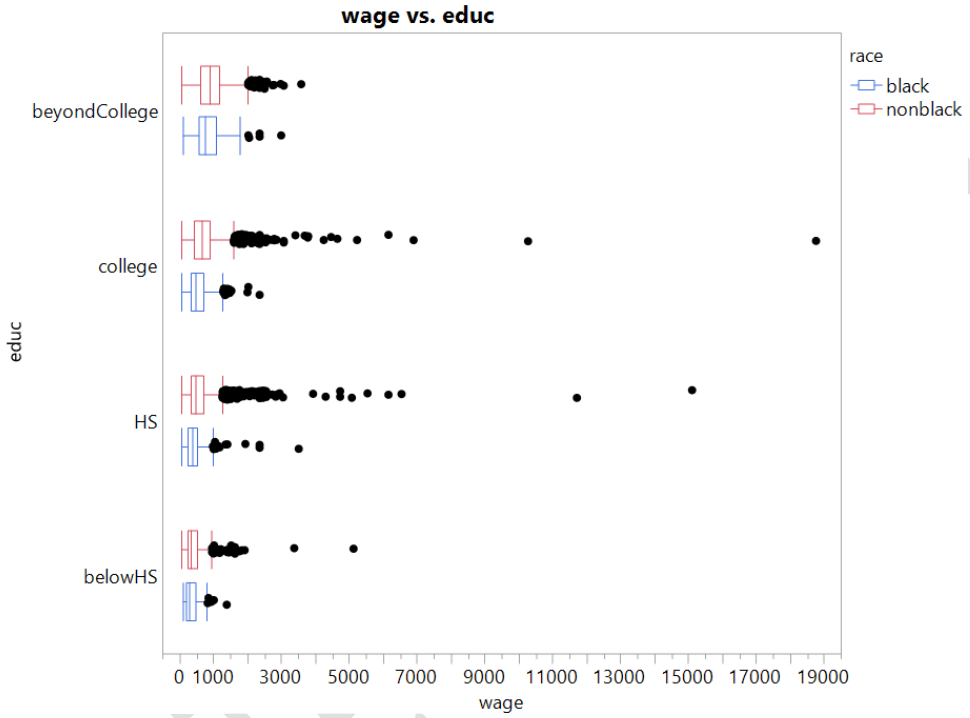
12. Describe the association between education and wage. Is it as you would have predicted? Explain.

13. What would need to be true for education level to provide an alternative explanation for why non-blacks in this sample tended to earn more than blacks?

One possibility is that if the blacks tended to have lower education levels and the non-blacks tended to have higher education levels, then the lower weekly wages for blacks could potentially be explained by education level rather than by racial differences.

The output in Figure P.10 compares races separately within each of the four education levels.

Figure P.10: Conditional distributions of wage/race association for different levels of education



	wage			
	Mean		Std Dev	
	race		race	
educ	black	nonblack	black	nonblack
belowHS	367.6	418.8	208.0	288.7
beyondCollege	846.8	970.5	486.7	499.1
college	544.3	714.0	314.9	482.7
HS	424.8	556.3	256.8	374.3

14. Is there a difference in the average wage between blacks and non-blacks in the "Beyond College" group? Is this difference larger or smaller than when we did not take the education level into account?

15. Do the lower average wages for blacks compared to non-blacks appear to be consistent across each of the education levels?

16. If you were to compare the average weekly wage for blacks to the average weekly wage for non-blacks *in the same education group*, roughly how large would you say that difference is?

17. How do you respond to the argument that the wage disparity between blacks and non-blacks is really an issue of education level, those with lower levels of education tend to earn less and in this sample non-blacks tended to have higher levels of education?

Key Idea: One way to account for confounding variables is conditioning on the third variable, but there could still be other confounding variables we don't know about!

18. Summarize what you have learned about the wage disparity between blacks and non-blacks within the education levels. After addressing the confounding variable of education, does the wage disparity still exist (evidence that the Civil Rights Act of 1964 had not been effective at equalizing pay between non-blacks and blacks)?

Key Idea: When you have a confounding variable that is associated with both the explanatory and response variables, ***conditional associations*** can look quite different from the overall association (ignoring the possible confounding variable).

In some cases, the conditional association can look very different (e.g., in Example P.A the direction of the association reversed when the data were disaggregated across groups) and sometimes, like here, the direction is the same, but the size of the differences is smaller. Here, even within the same education level, we still see a difference in the average weekly wages for blacks and non-blacks, it's just not as large as when we ignore education level.

19. Name some other potential confounding variables in the association between weekly wages and race. That is, what are some examples of variables that might explain variation in weekly wages, but might also be associated with race?

We can diagram what we have learned so far with a Sources of Variation diagram, where we use arrows to indicate observed associations between variables.

Key Idea: A ***Sources of Variation*** diagram is a visual representation of our belief about possible sources which explain variation in the response variable.

Figure P.11: Sources of Variation (SV) diagram for Salary Discrimination study

Observed Variation in: Monthly wages (thousands of dollars)	**Sources of explained variation** • Race • Education level	**Sources of unexplained variation** • Years of experience • Region of country • Type of job • Unknown
Inclusion criteria • Sex (males) • Time period (one year)		

We can now summarize the process of a statistical investigation through size steps.

The six steps of a statistical investigation

• **STEP 1: Ask a research question** that can be addressed by collecting data. These questions often involve comparing groups, asking whether something affects something else, or assessing people's opinions.

• **STEP 2: Design a study and collect data.** This step involves selecting the people or objects to be studied, deciding how to gather relevant data on them, and carrying out this data collection in a careful, systematic manner.

• **STEP 3: Explore the data**, looking for patterns related to your research question as well as unexpected outcomes that might point to additional questions to pursue. We can also develop a *model* of the data generating process to try to predict future observations.

• **STEP 4: Draw inferences beyond the data** by determining whether any findings in your data reflect a genuine tendency, and estimating the size of that tendency.

• **STEP 5: Formulate conclusions** that consider the scope of the inference made in Step 4. To what underlying process or larger group can these conclusions be generalized? Is a cause-and-effect conclusion warranted?

• **STEP 6: Look back and ahead** to point out limitations of the study and suggest new studies that could be performed to build on the findings of the study.

20. Map each of these steps to this particular study. Highlight any areas where you need any additional information. In particular, in Step 5, to what population are you willing to ***generalize*** these results? In Step 6, for what additional potential confounding variables would you like data?

Definition: The observational units we collect data on are referred to as the ***sample***. Typically this is not the entire group of observational units we are interested in, but rather we would like to apply the conclusions to a larger ***population***. ***Generalizability*** refers to deciding an appropriate population to which we can generalize our conclusions. What larger group do you think these results are representative of?

Note. The researchers did collect information on several other possible confounding variables. You can explore these other possible confounding variables in the homework exercises.

In the next pair of studies, you will begin to explore how we measure and quantify the amount of variation explained by different sources, and how explaining more variation in the response leads to better predictions.

Example P.B: Predicting Birthweights

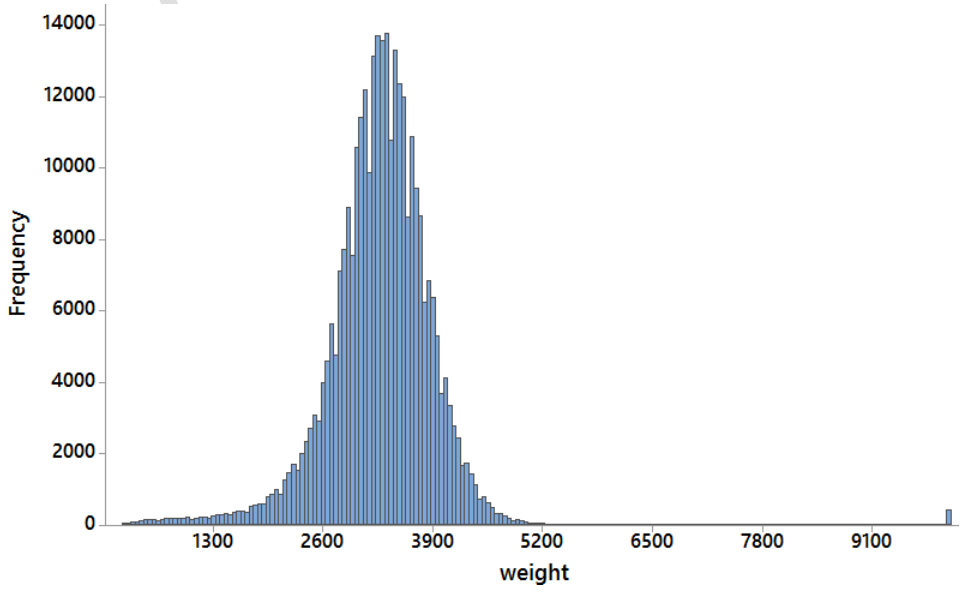
STEP 1: Ask a research question. Suppose our research question is how well are we able to anticipate, or predict, the birthweight of a newborn?

STEP 2: Design a study and collect data. One way to answer this question is to use historical data. The CDC's Vital Statistics Data allows you to download birth records for all births in the U.S. in a particular year. In fact, we downloaded the records for all 317,445 births in January, 2016 and then extracted several variables including the birth weight of the child (in grams).

STEP 3: Explore the data. The first step in exploring this variable is to look at a graph. Figure P.12 is a histogram of all birthweights for babies born in the U.S. in January, 2016.

Think about it: What are the *observational units* and *variable* in this graph? What are the most interesting features about this distribution? How would you assess the amount of variation in the distribution?

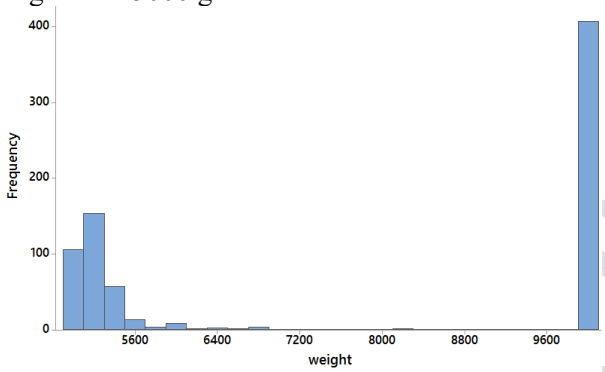
Figure P.12: Histogram of all birthweights for babies born in the U.S. in January, 2016



The distribution of birthweights for these 317,445 births is relatively symmetric and bell-shaped, but with an interesting group of very large values (above 9100 grams) and a slight skew (or bump) to the left. Let's look more closely at the very large birthweights. Figure P.13 is a histogram of birth weights larger than 5000 grams.

Think about it: What do you notice from this graph? How might you explain this behavior?

Figure P.13: Birthweights larger than 5000 grams



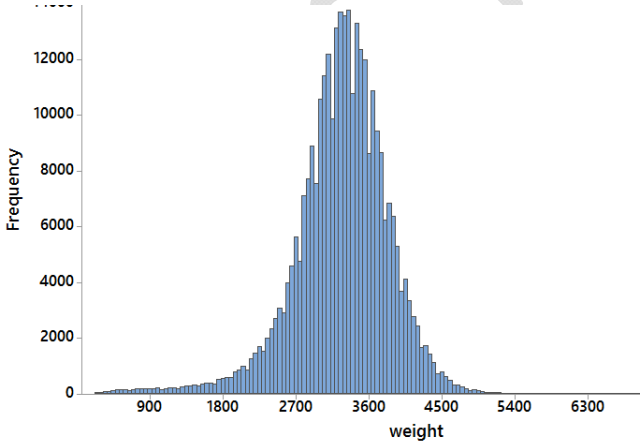
Although we might expect some larger birthweights, it is interesting that after about 6500 grams, there are no birth weights in the region until we find 407 of the 317,445 births at a weight of 9999 grams. It turns out the CDC codes any birth weight larger than 8165 grams as "not known" or "not stated" and codes them as 9999 (in order to be larger than the largest known weight).

Excerpt from CDC "Codebook"

463-466	4	DBWT	**Birth Weight – Detail in Grams**	U,R	0227-8165 Number of grams

Essentially these are "missing values" and should be removed from the dataset. Figure P.14 is the updated histogram, along with ***descriptive statistics*** (numerical summaries of the distribution).

Figure P.14: Distribution of birthweights after removing the non-responses coded as 9999



Variable	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
weight	317038	3259.1	592.2	227.0	2958.0	3297.5	3629.0	8165.0

Think about it: How could you use this distribution to predict the weight of a future newborn in the U.S.? How accurate would you say that prediction was likely to be?

You may recall:

- Both the ***mean*** and the ***median*** are effective measures of the *center* of a distribution. The median, a value with 50% of the observations on each side, is more resistant to outliers than the mean, and therefore is often interpreted as a "typical" value. With a symmetric distribution, the mean and median will be similar values.
- The ***standard deviation*** is a measure of the *variability* in the data, and is roughly interpreted as how far a typical observation lies from the mean of the distribution.

Do we detect enough of a pattern to these data that we can make predictions for births in other months?

Because the distribution of birth weights is relatively symmetric (although there is a bit of tail to the left), we can use the mean birth weight to predict another birth weight. In other words, we would predict the birth weight of a baby is about 3259.1 grams (roughly 7.2 pounds). The standard deviation of the birth weights is 592.2 grams. This means a typical birth weight in the data set is about 592.2 grams from the mean of 3259.1 grams.

Definition: We will define a ***statistical model*** as consisting of the equation that predicts the outcome of the response and a measure of the accuracy of those predictions.

For the birth weight data, this means we can use the following statistical model:

Predicted birth weight = 3259, standard deviation = 592 g

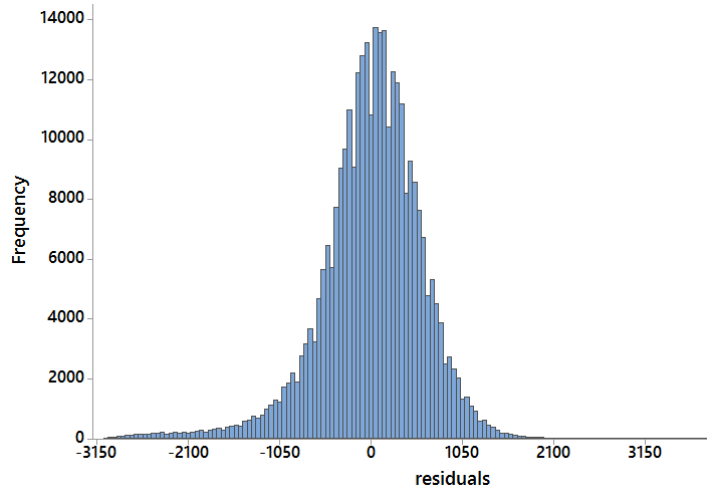
This is a pretty simplistic model, not using any of the other information we know about these births, but how accurate is it?

Definition: A ***residual*** is the difference between the response variable outcome and a predicted value of the response variable outcome, $residual = y_i - \hat{y}_i$, where i refers to the observational number, i = 1, …, n.

For example, the first baby in the data set weighed 3705 grams. The residual weight for this birth would be 3705 – 3259.1 = 445.9 grams. The residual is positive because the observed birth weight was larger than we would have predicted using the model. The residual represent what's left over after you account for a "model," here, the average birth weight.

Figure P.15 shows a histogram of the residuals for all the babies, along with the descriptive statistics.

Figure P.15: Distribution of residuals, using the mean birth weight to predict each birth weight in the sample



Mean = 0.00 g

Std Dev = 592.2 g

Think about it: How do this graph, mean, and standard deviation compare to the distribution of the birth weights? Do you notice any patterns to these residuals (perhaps a subgroup of babies that aren't well predicted by the average)?

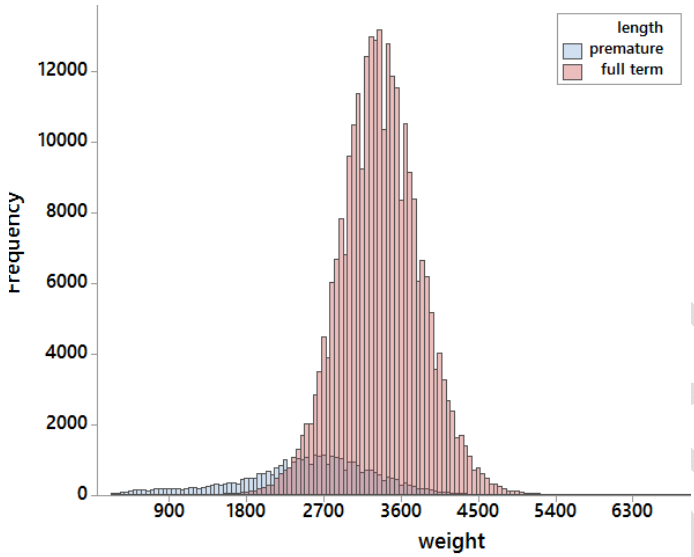
The mean of the residuals (apart from rounding discrepancies) will always be zero when using the mean of the data values to make the predictions. The standard deviation of the residuals is the same as the standard deviation of the data values and provides a measure of a typical prediction error. Subtracting the mean from each data value just shifts each observation by the same value, and does not change the shape or standard deviation of the distribution. We still see a group of babies with large negative residuals, birthweights far below the average.

Key Idea: The standard deviation of the residuals can be used to measure a typical prediction error from a statistical model. It represents the amount of "unexplained" variation in the response variable.

We see that there is still quite a bit of "unexplained" variability in these birthweights. We might wonder whether we can improve our prediction of birth weight by taking into account more information about these births. Better predictions arise by explaining more variation in the response, reducing the size of the residuals.

One variable we have access to in this data set is whether or not the baby was "full term" (37 weeks or longer). Premature babies will likely tend to have lower birth weights than full term babies, so it seems reasonable that we might improve our predictions by taking into account the length of the term. Figure P.16 shows the full term and premature birth weights separately. Note that 62 observations were removed because the length of the pregnancy was unknown.

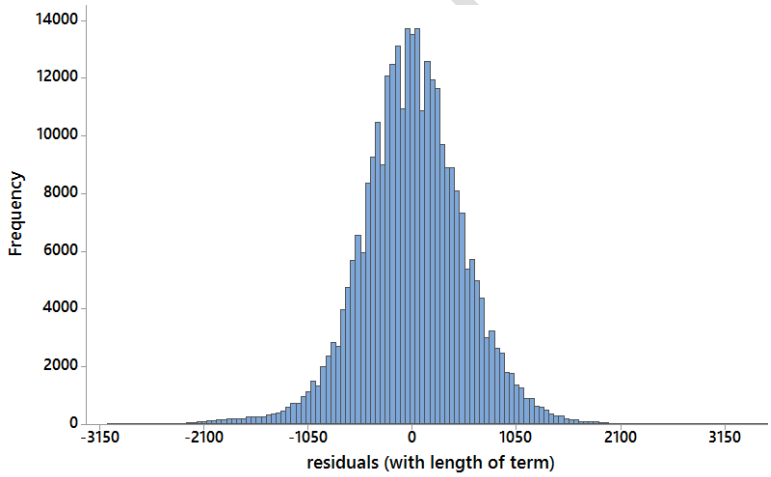
Figure P.16: Birthweights for full term pregnancies and premature pregnancies



Variable	full term	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
weight	no	36963	2493.8	795.7	227.0	2050.0	2571.0	3033.0	5670.0
	yes	280075	3360.1	475.3	320.0	3056.0	3350.0	3657.0	8165.0

Because there is a shift in the centers of these two distributions, we see that the duration of the pregnancy is associated with newborn weights in this sample. So we can predict a weight of 3360.1g for full term pregnancies and 2493.8g for premature babies. Figure P.17 shows the residuals from these predictions.

Figure P.17: Residuals from using 3360.1g as the predicted weight for full term babies and 2493.8 as the predicted weight for premature babies



Mean = 0.00 g

SE = 522.9 g

So our statistical model is now

$$predicted\ weight = \begin{cases} 3360.1, if\ full\ term \\ 2493.8, if\ prematrue \end{cases},\ SE\ of\ model\ residuals = 523.9\ g$$

The ***standard error*** of these residuals, 522.9, is slightly smaller than the standard deviation of the residuals when we didn't take the length of the pregnancy into account (592.2), indicating that we have explained some (but not the majority) of the variation in birth weights by knowing whether or not it was a premature birth. Notice how we have "explained" the pattern we saw in the left tail of the distribution.

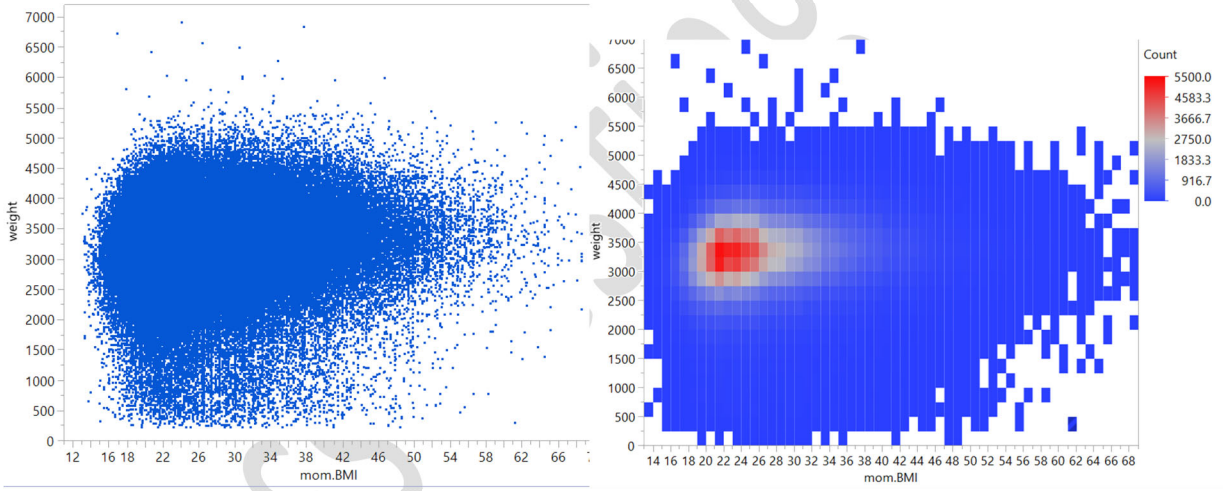
Frequently asked question: *What is the difference between a standard error and a standard deviation?* We will illustrate how the standard error of the residuals is calculated (and why) in Chapter 1. For now, you can interpret the standard error exactly as you did the standard deviation.

Is there another variable we could add to our analysis to explain some of this remaining variation?

The mother's pre-pregnancy body mass index (BMI = weight(lb)/height(in)^2 × 703) provides an indication of the mother's body fat based on her height and pre-pregnancy weight. Figure P.18 is a scatterplot of *mother's BMI* and *birth weight* (9,013 mothers who do not have a recorded BMI have now also been removed from the dataset).

Think about it: Is there evidence of an association? Positive or negative? Is this what you expected?

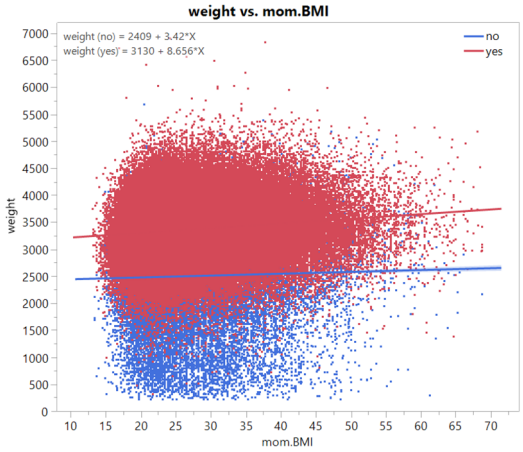
Figure P.18: Scatterplot and heat map of mother's pre-pregnancy BMI versus child's birthweight

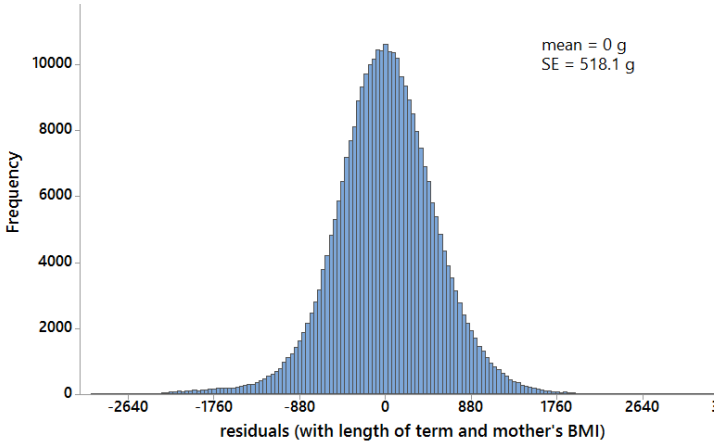


We see a weak, positive association as mother's with larger BMI values tend to have heavier babies. We also see the bulge in the lower left corner for the premature babies and less variation in the birth weights for the mother's with the largest BMI values. We can also examine the *conditional* associations between baby weight and mother BMI by fitting regression lines separately for the full term and premature births.

We can fit separate regression lines between baby birth weight and mother's BMI, for the full term and premature babies, and find a histogram of the corresponding residuals.

Figure P.19 Models of the conditional associations after adjusting for whether or not the birth was full term





So our statistical model would be:

$$predicted\ weight = \begin{cases} 3130 + 8.66\ mom\ BMI\ if\ full\ term \\ 2409 + 3.42\ mom\ BMI\ if\ premature \end{cases},\ se\ of\ residuals = 518.1\ \text{g}$$

We do see a (slight) reduction in the variability of the residuals from the model that did not consider mother BMI, where the standard error of the residuals was 523.

Figure P.20 is a Sources of Variation diagram to summarize this statistical model.

Figure P.20: Sources of Variation diagram for birth weight data

Observed Variation in: Birth weight (grams)	**Sources of explained variation**	**Sources of unexplained variation**
Inclusion criteria • Country (U.S.) • Birth month (January)	• Length of pregnancy • Mother BMI	• Baby's Sex • Mother's weight gain during pregnancy • Father's BMI • Unknown

STEP 4: Draw inferences beyond the data.

In later chapters, we will discuss confidence intervals and prediction intervals for improving statements of our predictions, along with their accuracy and reliability.

STEP 5: Formulate conclusions.

The model that uses information about the length of pregnancy and mother BMI appears to be the best model so far, producing the smallest standard error of the residuals. This model was based on all 317,038 births for which we had complete data in the CDC database for January, 2016. Even with these two predictors, there is quite a bit of variation left (residual se of 518.1 grams, compared to an overall mean of 3259 grams), suggesting that there may be other variables would explain some of this remaining variation in birthweight. In this case, we can treat (reported) US births in January as our population, but we might also want to explore whether these trends appear to apply to other months of the year and to other countries. We are also not willing to draw any cause and effect conclusions from this observational study.

STEP 6: Look back and ahead.

Some other variables we might want to examine to predict birth weight include the sex of the newborn, mother's weight gain during pregnancy, etc. Later in the course we will explore models which are able to account for more than two potential sources of explained variation.

Exploration P.B: Housing Prices in Michigan

STEP 1. Ask a research question.

1. I want to predict how much should I expect to pay for a home. What explanatory variables might explain variation in home prices?

STEP 2: Design a study and collect data.

In 2015, we sampled 13 homes that were for sale just north of Lake Macatawa in Michigan, and recorded the sale price (in thousands of dollars) as listed on Zillow (a real estate website).

2. Identify the observational units and the response variable of interest. Is the response variable quantitative or categorical? What kinds of graphs and numerical summaries can you use to explore these data?

STEP 3: Explore the data.

Open the **Multiple Variables** applet and the data file (***homeprices***). From the data file, select the three columns of data and copy to the clipboard. In the applet, press **Clear** and then click in the Sample data box and paste. Then press the **Use Data** button. Now drag the *price* variable from the Variables list to the Response box. Check the box for **Show descriptive** statistics.

3. Describe the shape of the home prices in this sample.

4. Suppose we want to estimate or "predict" a typical home price based on these 13 homes. Do you think the mean would be a reasonable value to use? Explain why or why not.

5. Do you think the mean is larger or smaller than the median? Explain.

To measure variation we could focus on the interquartile range ($584,900 – $204,805 = $379,950) (the width of the middle 50% of the observations) or the standard deviation, which we can loosely interpret as a typical distance from the mean.

Definition: The formula for the ***standard deviation*** of a sample of n observations is given by:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

where $y_1, \ldots, y_n$ are the individual observations and $\bar{y}$ is the mean of the response variable. Squaring this value gives us s^2, the *variance* of the distribution, but the standard deviation is usually the summary that is reported rather than the variance because the standard deviation has the same measurement units as the response variable. We can loosely interpret the standard deviation as a "typical distance from the mean."

6. What is the standard deviation of these home prices? (What are the units?) Provide a one-sentence interpretation of this value. (*Hint*: How accurately can we predict a home price?) What are some possible explanations for this variation in home prices?

Of course, we also need to be a little wary in using the standard deviation with skewed data like this, as the deviations from the mean tend to be smaller for homes below the mean compared to homes above the mean.

Key Idea: In this book, we will frequently utilize the mean and standard deviation as our statistics, but they are best applied to symmetric distributions.

7. The first home in the data set sold for $639,000. If we had used the mean to predict that sale price, how far off would we have been?

Definition: The ***residual*** is the difference between the response variable outcome and a predicted value of the response variable outcome, $residual = y_i - \hat{y}_i$, where i refers to the observation number, $i = 1, \ldots, n$. (n is the number of observations in the sample.)

8. Using the mean as our prediction, did we over-predict or under-predict the price of the first house. (*Hint:* Is the residual positive or negative?)

9. In the applet, check the box to **Show residuals**. Roughly how often do you find a positive residual (under-prediction) and how often do you find a negative residual (over-prediction)?

In fact, when the mean is used to predict the observed response values, the average of the residuals will always be zero (within rounding).

Definition: We will define a ***statistical model*** as consisting of the equation that predicts the outcome of the response and a measure of the precision of those predictions.

For example, here we could use the following ***statistical model***:

Predicted price = 408.062 thousand dollars,
standard error of residuals = 240.923 thousand dollars

The ***standard error (SE) of the residuals*** is a measure of a typical prediction error from the model. Because the above model uses the overall mean as the prediction, the residual standard error is the same as the standard deviation of the home prices. Using more complex models will change the residual SE. We'll discuss the computation of the residual SE more in Chapter 1.

The residual standard error for this statistical model is still pretty large considering the values of these home prices (the standard deviation is about half the size of the mean). With so much unexplained variation remaining, this is a pretty uninformative model; can we improve our model using more information about these homes? Let's use information about the *size of the house* in our predictions to see whether we can reduce the sizes of the residuals overall.

In the applet, drag the square footage variable (*sqft*) to the Explanatory box and examine the scatterplot (price vs. sqft). Also check the **Show Equation** box.

10. Describe the association between home price and square footage (strength, direction, linearity?). Does the association behave as you would have predicted? Explain.

11. Our new statistical model is:
a. *Predicted price* = , standard error of residuals =

b. Provide interpretations, in context, of the slope and intercept of the line.

c. What residual does this give you for the first home in the dataset? (Show your work.) Is this residual smaller or larger than before?

d. From the applet output, is the standard error of the residuals larger or smaller than before? Explain what this tells you.

Now consider the location of the home. In this dataset, the homes have been classified as lake front or not. Use the applet to examine separate regression lines for lake front and non-lake front homes.

Drag the *Lake* variable into the **Subgroup By** box. This fits a separate line to the two types of home.

12. How do the two lines compare (think about both the slopes and the intercepts). Do these comparisons make sense in this context?

13. What is our new statistical model?

Predicted price = *, standard error of residuals =*

14. Use this model to predict the price of the first home in the data set. (Show your work.)

15. Does including this additional location variable further reduce the standard error of the residuals?

16. Based on the scatterplot, is *location* (lake front or not) confounded with *size* of the home? Remember to consider both whether location appears to be related to price and whether location appears to be related to size. (Include any additional output you use to help you answer this question.)

Complete the Sources of Variation diagram to summarize what we have learned so far.

Figure P.21: Sources of Variation diagram for Housing Prices data

Observed Variation in:	**Sources of explained variation**	**Sources of unexplained variation**
Inclusion criteria •	•	•

STEP 4: Draw inferences beyond the data.

In later chapters, we will discuss how to use confidence intervals and prediction intervals to improve statements of our predictions, their accuracy, and the reliability of the methods.

STEP 5: Formulate conclusions.

17. Summarize the conclusions you would draw from this study, including which statistical model you would recommend and why. Is there a larger population of homes you are willing to generalize your conclusions to? Is it reasonable to conclude that either the size of the home and/or the location of the property is causing variation in home prices? Explain.

STEP 6: Look back and ahead.

18. Identify any limitations you see to this study. What additional data would you like to collect to answer this research question?

Preliminaries Summary

In the Berkeley and Wage discrimination studies you learned how to identify a potential confounding variable. To be a confounding variable, a variable must explain variation in the observed responses AND also be related to the explanatory variable. In some cases, the confounding variable can offer a plausible, alternative explanation for the observed association between the explanatory and response variables (e.g., females are more likely to apply to the programs that are harder to get into). As you recall from your first statistics course, confounding variables are the main problem with drawing causal conclusions in observational studies.

In both the Graduate Admissions at Berkeley study and the Salary Discrimination study, you saw how to account for a confounding variable through a *conditional* analysis. Considering how acceptance relates to sex within each program, or how weekly wages are associated with race within each education level gave us a clearer picture from which we can make stronger decisions. In the Berkeley study, accounting for the type of program actually changed the direction of the association between acceptance and sex, providing an example of Simpson's Paradox. In the Salary study, accounting for education level did not change the direction of the association between wage and race, as non-blacks still made on average more than blacks, however the magnitude of the difference was smaller within each of the education levels, than what we saw overall.

In the birth weight and housing datasets, you learned about residuals and how the standard error of the residuals provides an indication of a "typical" prediction error. Adding more variables to the statistical model generally reduces these prediction errors, but sometimes at a cost, which will be discussed in more detail in later chapters. In the birth weight data, we saw that including more variables reduced the standard error of the residuals but not by a substantial amount.

How relationships change or don't change in the presence of other variables, and how to control for and/or explain that variation, will be a key idea in this book. When we consider a third variable, we may find that the original relationship is weaker than we thought or even in the opposite direction! Our overarching goal in this course will be explaining variation in a response variable, and identifying what combination(s) of predictor variables are most effective in doing so. Keep in mind that the starting place is to define the observational units, the variables of interest, and other possible sources of variation in the response variable, some of which can be controlled at the start of the study, some of which can be measured and accounted for in the statistical model, and some which cannot be measured and remain as unexplained sources of variation.