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~~ men or ~~females~~ women would be more likely to be accepted?

Outcomes of different events 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? define? or use explain?

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. make into a table? For example, in a study ~~showing a link~~ investigating the relationship 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~~ account 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 of what we might expect for a new observation (person). 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 makes 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~~ change relationships between the variables of interest
- How explaining variation in the response variable can reduce ~~prediction error~~ the errors of new predictions

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 interested in explaining or ~~trying to~~ predicting.
  - We will classify the variables as ***quantitative*** (e.g., numerical) or ***categorical*** (identifying groups or categories that the observational units belong to).

explanatory variables?

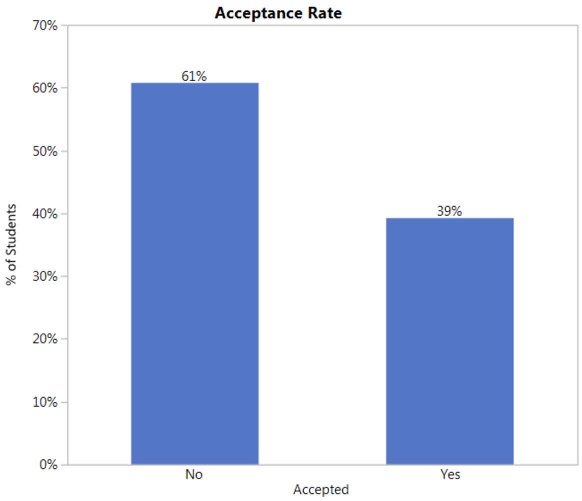
continuous

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).

which displays the frequency of each outcome of the response variable

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~~ determine 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

based on the sex of the applicant, with categories of M : F 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***.

earlier?

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

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.

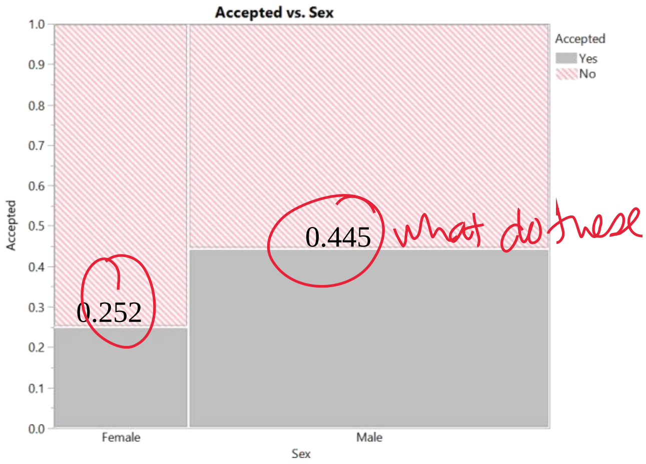
not defined

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.

include a segmented bar graph to compare differences

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



what do these correspond with?

describe what the conditional distribution is

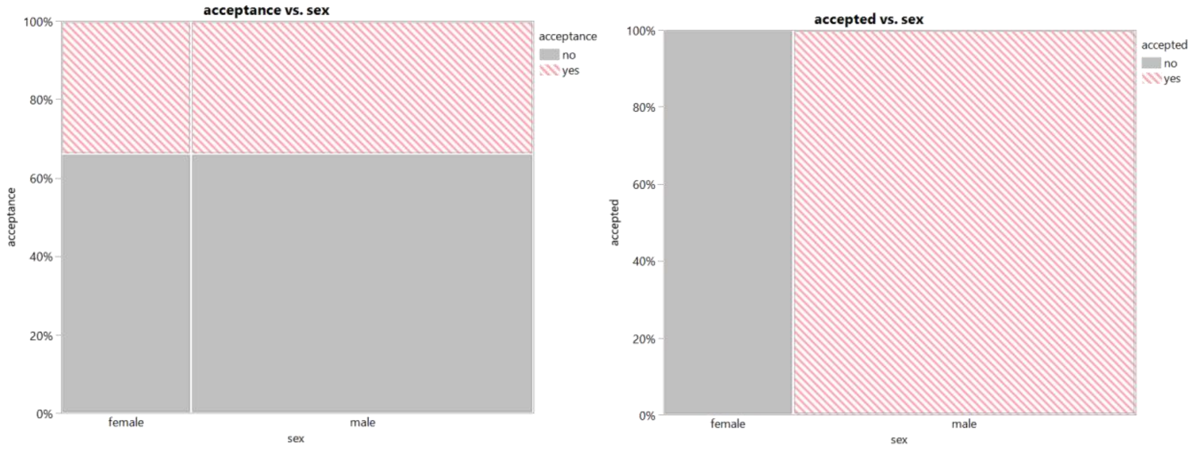
Because the (conditional) distribution of acceptance for females is not the same as the (conditional) distribution for males (the male bar and the female bar differ), we say that sex and acceptance are ***associated*** in this sample.

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 these two hypothetical mosaic plots. The graph on the left shows no association between sex and acceptance – both males and females are denied ~~approximately 65~~ the same amount% 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~~ perfectly predict~~ions~~ whether someone would be accepted 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 someone applied to help us further explain variation in who ~~is~~ was accepted? What do you notice about the association between acceptance and sex within each program compared to when we combined the data altogether?

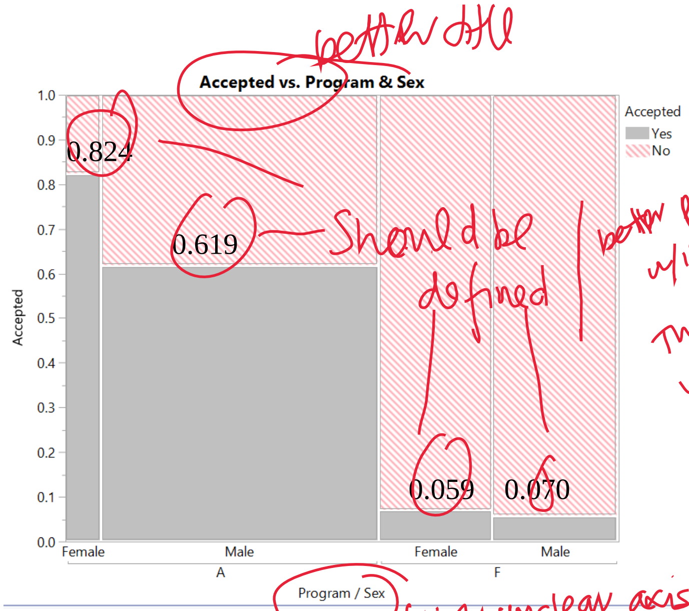
Figure P.4: Two-way tables and Mosaic plots for acceptance vs. sex separately for each program

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



better title

Should be defined

better placed w/in bar

They correspond with

fairly unclear axis labels

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 is associate~~s~~d with acceptance change from when we looked at sex alone, to when we consider the program as well? application

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~~ very 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 an ~~statistical~~ *interaction*.

really?

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 a male dominated field ~~Engineering~~?), most applicants were male, whereas in Program F (perhaps a female dominated field ~~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.

recorded or not?

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.

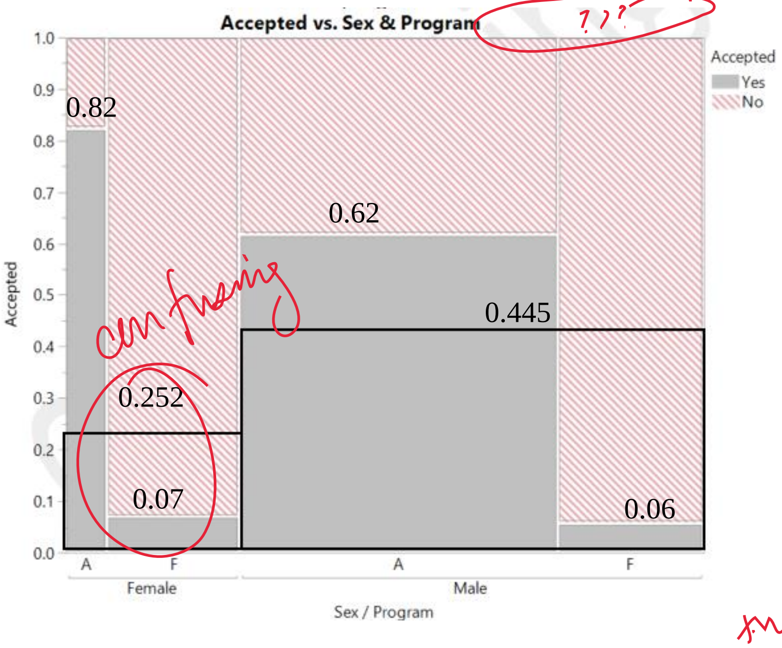
definition?

why not? what can you make these conclusions for?

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

background image?



Key Idea: One way to account for confounding variables is conditioning on ~~the third variable~~ them 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.

Where are these?

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

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

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. explain the underlying process?

- **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.
  odd choice of word

- **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*~~ substantially 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). Define what was observed?

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.

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. So ~~In fa~~ct, 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).

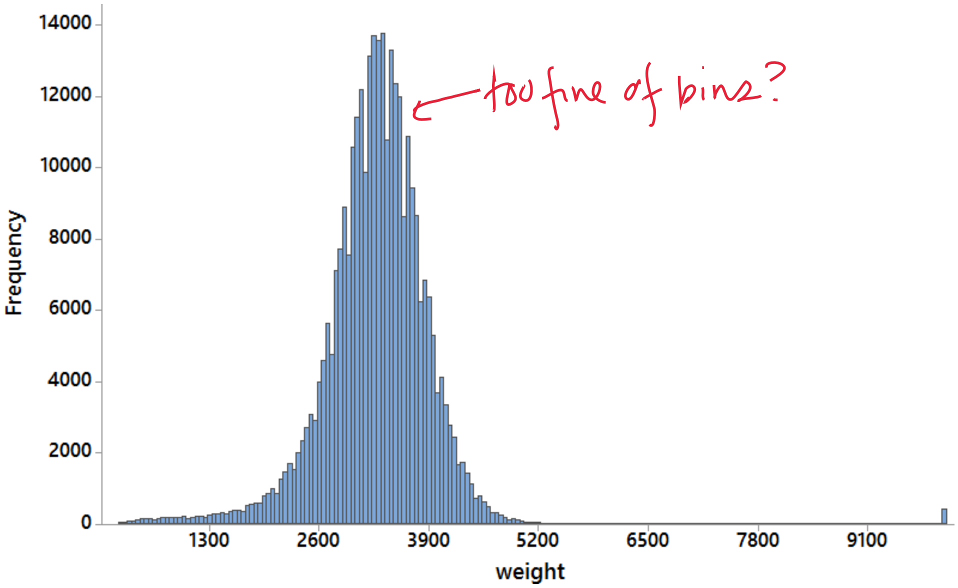
which variable

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.

what type of graph?

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

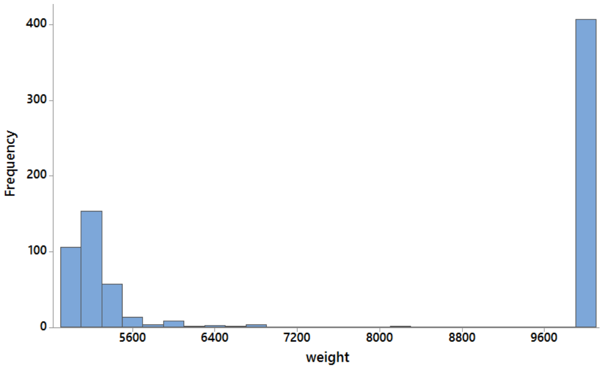


too fine of bins?

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.

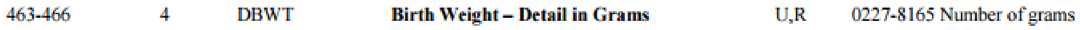
Think about it: What do you notice from this graph? How might you explain this behavior?

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"

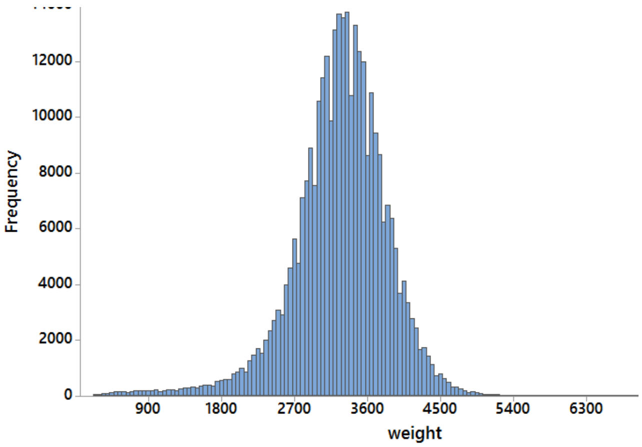


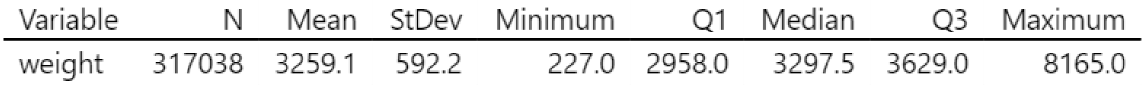
463-466	4	DBWT	**Birth Weight – Detail in Grams**	U,R	0227-8165 Number of grams

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).

sensored data: why not report as 8165?

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. why those values?

Definition: We will define a ***statistical model*** as consisting of ~~the~~ an equation that ~~predicts~~ explains or predicts the outcome of the response and a measure of the accuracy of ~~those predictions.~~ the model in explaining the outcomes

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

Predicted birth weight = 3259, standard deviation = 592 g

put in yi notation

to connect w/ this

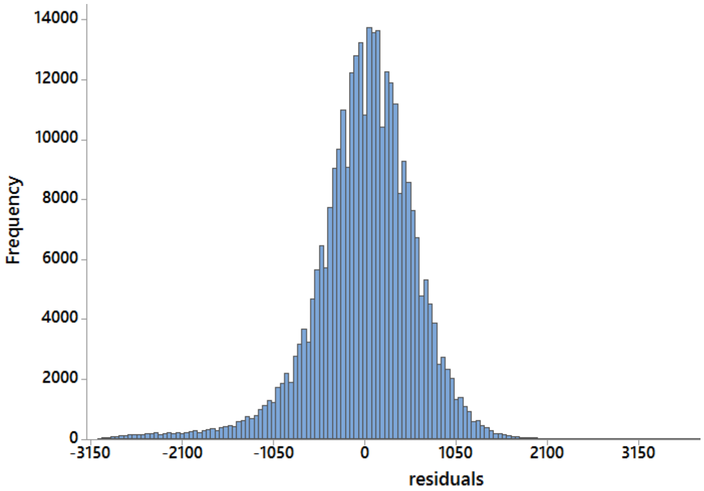
This is a pretty simplistic model, not using any of the other information we know about these births, but how accurate is it?

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 doe this graph~~,~~'s 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)?

Why???

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.

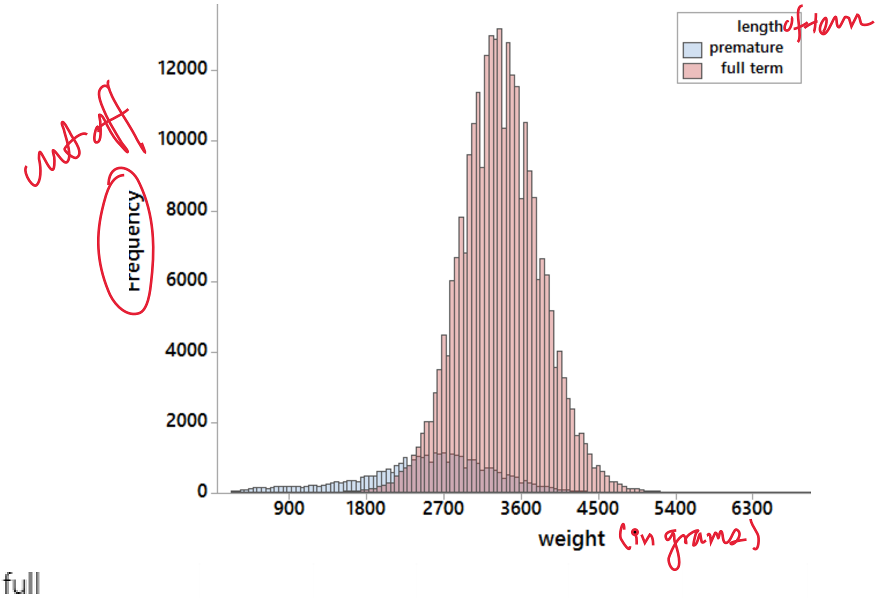
describe this process

Key Idea: The standard deviation of the residuals can be used ~~to~~ as a measure ~~a typical~~ of the 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, which then reduc~~ing~~es 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



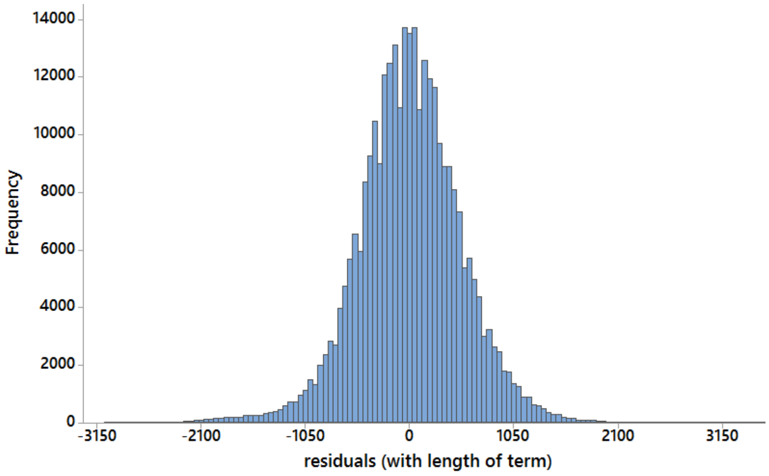
Variable	full term	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
weight	no	36963	2493.8	795.7	227.0	2050.0	2571.0	3033.0	5670.0
	yes	280075	3360.1	475.3	320.0	3056.0	3350.0	3657.0	8165.0

Because there is a shift in the centers of these two distributions, we ~~see~~ know 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.

b/c the distributions differ.

write out this stat model?

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

pregnancy was

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

something like this should go earlier

This is a confusing format w/out explanation. Why are you listing the SE? How does that relate to the model?

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 additional ~~(but not the majority) of the~~ variation in birth weights by knowing whether or not it was a premature birth, by including the length of the term, Notice how we have "explained" the pattern we saw in the left tail of the distribution.

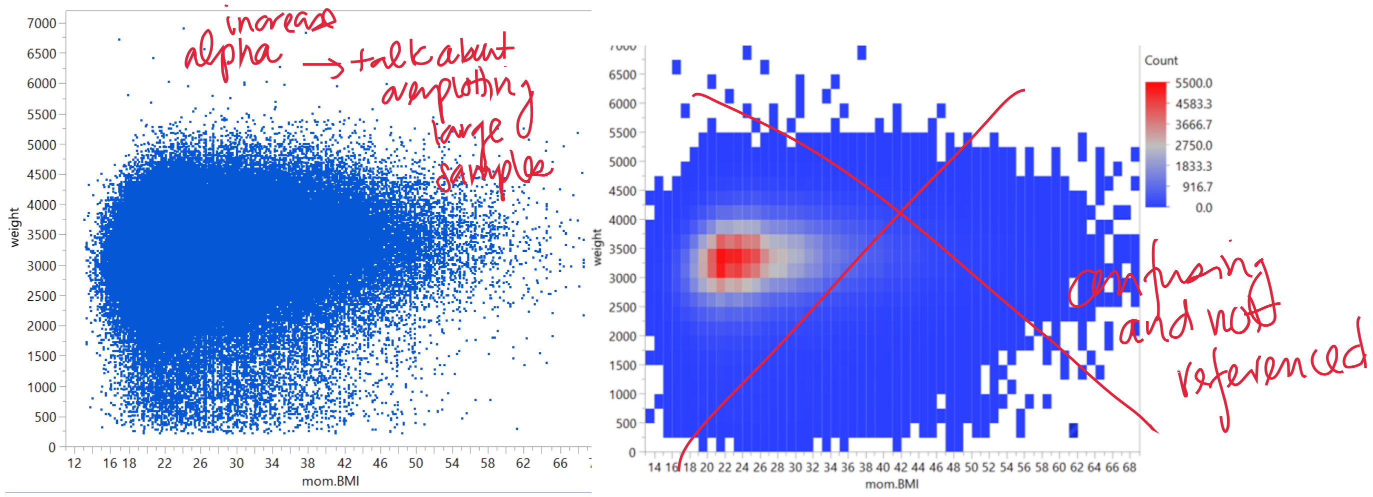
Frequently asked question: *What is the difference between a standard error and a standard deviation?* We will illustrate how the standard error of the residuals is calculated (and why) in Chapter 1. For now, you can interpret the standard error exactly as you did the standard deviation.

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 between the baby's weight and the mother's BMI? Is the relationship? 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



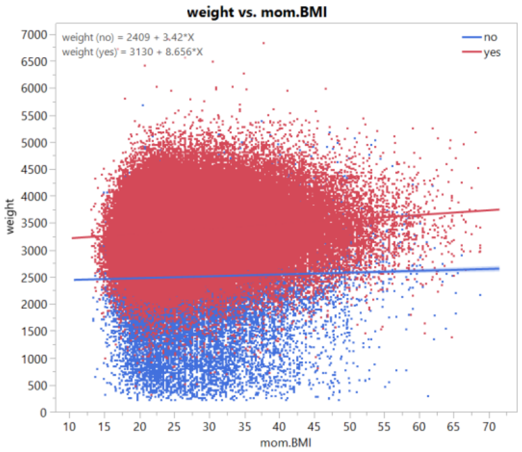
increase alpha → talk about overplotting large sample

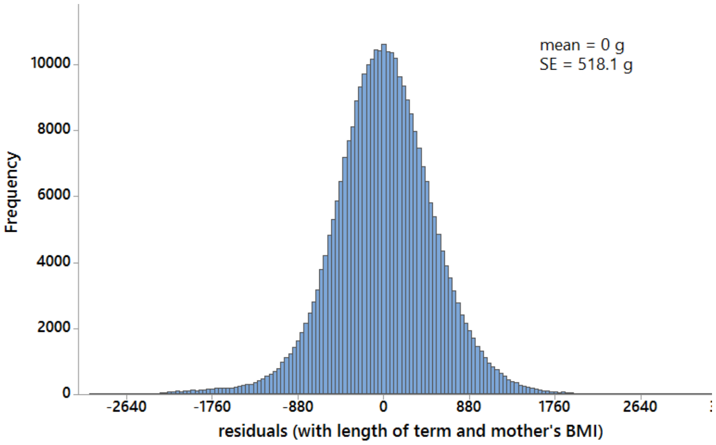
confusing and not referenced

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

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.