Chapter 6
Investigate Patterns of Association in Bivariate Data

In 8th grade, students investigate patterns of association in numerical (or "quantitative") bivariate data by constructing and interpreting scatter plots. An emphasis is placed on informal linear association analyses. In addition to using linear models to solve problems regarding numerical data, students explore categorical (or "qualitative") bivariate data through construction and interpretation of two-way frequency tables.

It should be noted up front that the practice of statistics is fundamentally different from the practice of mathematics. Thus, the integration of statistics within a mathematics curriculum is potentially misleading, with respect to the natures of both mathematics and statistics. While mathematics draws logical deductions from a set of axioms, statistics does not. Instead, statistics endeavors to quantitatively communicate properties of and relationships among observable phenomena. To that end, statistics is more like a scientific discipline than mathematical (e.g., the use of statistics is evidence-based, as its subject matter is analysis of data; statistical conclusions change over time, as more data are collected in various ways; mathematical conclusions do not change, as propositions are put to rest after being proved or disproved). To correctly convey the powers and weaknesses of statistics, teachers need to be steadfastly aware of the language they use when communicating statistical ideas to students.

Although some of the following italicized words are not mentioned in the Utah core for the 8th grade, it may be a good idea to casually demonstrate their usage while communicating with students.

An *experiment* is an activity for which *outcomes* occur randomly (i.e., based upon chance). The *sample space* of an experiment is the set of all possible outcomes. A *random variable* is a function that maps the sample space of an experiment to the set of real numbers. *Realizations* of a random variable are the specific values that the random variable may assume. A random variable can be thought of as a quantity that can assume more than one value, based upon chance events. We discuss two kinds of random variables: *quantitative* random variables and *categorical* random variables. Quantitative random variables are those of cardinal numerical value (e.g., 3 feet; 2.73 gallons; 4 children), whereas categorical random variables are those representing some quality or name. This distinction must be made clear in practice, since a set of categories can be easily replaced by nominal real numbers.

Now that we have introduced some general statistical concepts, we turn our focus to the 8th grade, which concerns itself specifically with bivariate data. A *bivariate data set* is a set of ordered pairs (x, y), where x and y are realizations of two different random variables (X and Y), such that the specific realizations x and y correspond to each other in some way (e.g., the ordered pair describe the same individual, they describe the same time period, or they may be related through some other such rule of correspondence). The nature of the correspondence between specific realizations x and y is described in the following examples.

Example 1.

Let X be the random variable "the European shoe size of a citizen of Springville", and let Y be the random variable "the height in centimeters of a citizen of Springville". Certainly, X and Y may assume a myriad of realizations. Maria conducts an experiment by recording the shoe sizes and heights of 53 randomly-sampled Springville citizens (this is the sample space). The bivariate data collected by Maria is the set of all 53 (x, y) such that x and y correspond to the same citizen. The "relation" between each specific realization x and y in the bivariate data set is that each pair describes the same person's shoe size and height.

Example 2.

Let X be the random variable "the average cost of gasoline in Cedar City in a given year", and let Y be the random variable "the number of speeding tickets written in a given year in Iron county". Lucas conducts an experiment by looking up and recording the realizations of X and Y from the year 1972 through 2012. Lucas decides to pair each x-value with the y-value that corresponds to the same year. Thus, Lucas's bivariate dataset is the set of (x, y) pairs for each year in the given 41 year range.

Section 6.1: Construct and Interpret Scatter Plots for Bivariate Data

Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association. 8.SP.1

Here, when the Utah Core refers to "quantities", it means "quantitative random variables". People are often interested in whether or not one random variable is associated with another. For example, is there a relationship between the number of television commercial broadcasts of a certain product and the number of sales of that product? Do 8th grade students who can do many push-ups in P.E. class also tend to be able to do more pull-ups? This section discusses methods of answering such questions about quantitative random variables such as "number of push-ups".

Conventionally, X is assigned to be the *input* random variable (i.e., the *independent'* or "explanatory" variable) from which we wish to predict the *output'* variable Y (i.e., the (presumed)*dependent'* or "response" variable). Of course, one does not initially know if there exists any dependence structure between the two; quantifying the relationship between X and Y, based upon observed data collected through random sampling, is precisely the goal of our statistical analysis here. It is important to note that there is an obvious relationship between the two specific realizations making up a single datum (i.e., a specific (x, y) pair), namely that there is some rule of correspondence, such as the fact that they describe aspects (or qualities) of the same individual or time period. However, the investigator is not interested in the relationship between realizations making up a single datum. The investigator wishes to quantify the relationship between two random variables, X and Y , which describe an entire population, not between the specific values x and y for a given subject. This point is subtle, but extremely important, lying at the very heart of statistics: One cannot make inference on a population based upon an individual, nor vice versa.

To visually inspect the potential influence of one random variable on another, we depict our sample data on a scatter plot. A *scatter plot* is a graph in the coordinate plane of the set of all (x, y) ordered pairs of bivariate data. Consistent with the usual convention, we place the independent variable X on the horizontal axis and the dependent variable Y on the vertical axis.

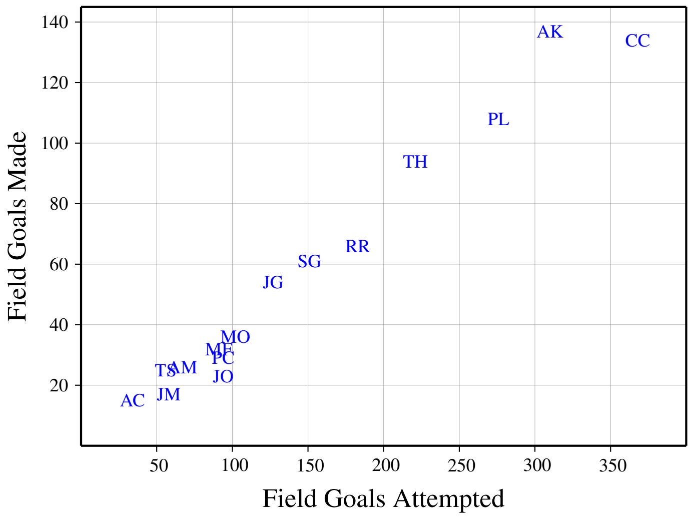
Example 3.

Izumi is the score keeper for her school's basketball team. Izumi's responsibilities include keeping a record of each player's total number of field goals made, the total field goals attempted during the season, the total number of assists and the total number of rebounds. For those not familiar with basketball, let us define these terms. Basketball is a game involving 5 players on each of two sides, using a ball and "goals." The playing field is a rectangle with "goals" at each short end of the rectangle. The goal is a basket set 10 feet off the floor of the playing field, and the object of the game is to put the ball in the basket. Any shot at the basket is an "attempt," and if the ball goes through the basket, this is a "goal." Goals can contribute two or three points, depending upon the distance covered. For Izumi's purpose this is not important: the important data are "goals attempted" and "goals made." An "assist" is awarded to a player who delivers the ball to someone who actually makes a goal. Finally, a "rebound" is awarded to a player who catches the ball when a goal is attempted, but not made. While the names of players have been changed, these data (for the 2012-13 season) were borrowed from actual Utah high school girls basketball players via www.maxpreps.com.

Part of Izumi's duties include helping the coach decide which players deserve awards at the end of the season. Izumi notes that Ameila Krebs was the highest-scoring player for the season, but Amelia also had a high number of failed field goal attempts. Izumi would like to further investigate the relationship between the two random variables "Field Goals Made" and "Field Goals Attempted". Izumi's data are given in the table below.

Player	Field Goals Attempted	Field Goals Made
Amber Carlson	34	15
Casey Corbin	368	134
Joan O'Connell	94	23
Monique Ortiz	102	36
Maria Ferney	91	32
Amelia Krebs	310	137
Tonya Smith	56	25
Juanita Martinez	58	17
Sara Garcia	151	61
Alicia Mortenson	67	26
Parker Chistiansen	94	29
Rachel Reagan	183	66
Paula Lyons	276	108
Thao Ho	221	94
Jessica Geffen	127	54

To better visually inspect her data, Izumi makes the following scatter plot of Table 1, where each data "point" consists of the initials of the corresponding basketball player.



2012-2013 Girls Basketball Data

Because Izumi would like to visualize each individual player, she chooses to identify them by their initials. However, if she were more interested in the relationship between her two variables (Field Goals Made and Field Goals Attempted), then she would likely make a plot with a marker for each data point (see the plot on the next page).

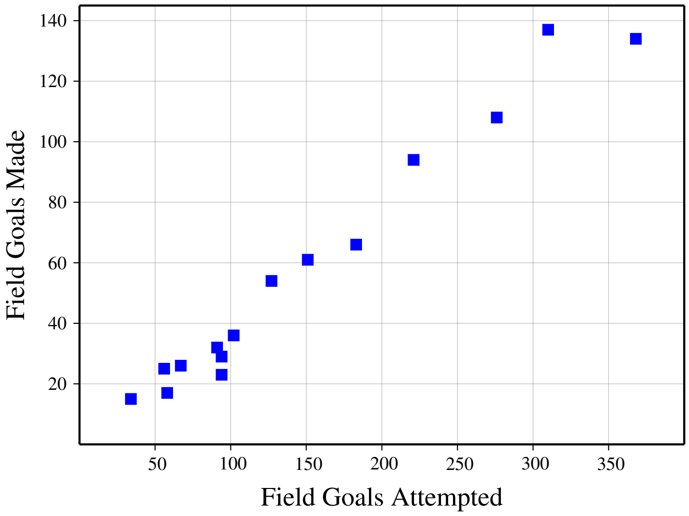


Figure 1: 2012-2013 Girls Basketball Data

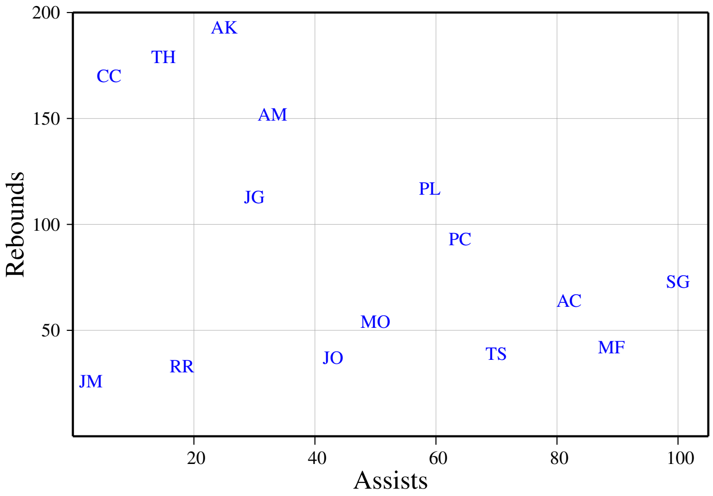
It should be noted that it is the latter plot with the points (rather than the initials) that is typically created. We include the plot with the individual subjects identified by initial for two purposes:

1. To provide an intermediate step between the data table and the typical scatter plot;
2. To be able to reference individual players during the data analysis discussion. The nature of the association between the variables Field Goals Made and Field Goals Attempted will be discussed in the next section.

In addition to data about field goals, Izumi is curious about the relationship between the number of assists a player makes and the number of rebounds a player makes in a season. She notices that the players who make the most assists tend be in positions located far away from the basket. Izumi's Assist and Rebound data are given in the table below the plot on the next page.

Player	Assists	Rebounds
Amber Carlson	82	64
Casey Corbin	6	170
Joan O'Connell	43	37
Monique Ortiz	50	54
Maria Ferney	89	42
Amelia Krebs	25	193
Tonya Smith	70	39
Juanita Martinez	3	26
Sara Garcia	100	73
Alicia Mortenson	33	152
Parker Chistiansen	64	93
Rachel Reagan	45	67
Paula Lyons	59	117
Thao Ho	15	179
Jessica Geffen	30	113

From these data, Izumi creates the scatter plot below, again using the players' initials to identify each individual.



2012-2013 Girls Basketball Data

Again, note that if Izumi were interested less in individual players and more in the general relationship between assists and rebounds, she would have made the next scatter plot.

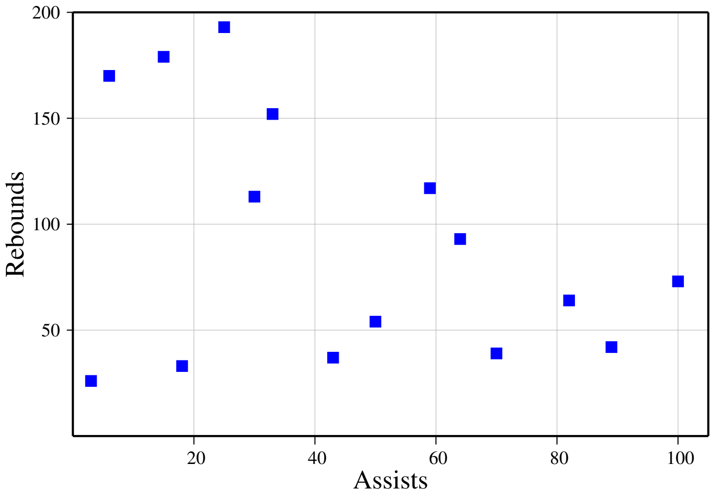


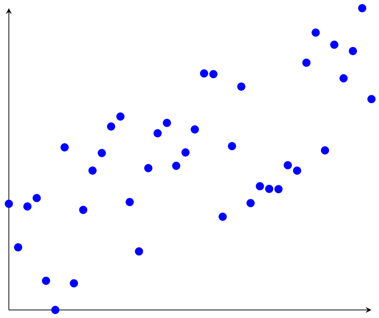
figure 2: 2012-2013 Girls Basketball Data

What is meant by *pattern of association*? It is perhaps easier to discuss what is NOT meant by this phrase. One cannot use statistics to argue whether or not change in one variable causes the other to change. As an over-used example, while there may be an association between the time of a rooster's crow and the time of sunrise, the rooster's crow certainly does not cause the sun to rise (although he might think it does). While the truth, in a particular context, may indeed be that "X causes Y", this conclusion cannot be drawn by statistical practices. The lack of ability to establish a cause-and-effect relationship between X and Y is precisely why we choose to use the word "association". More on this will be addressed later, via examples.

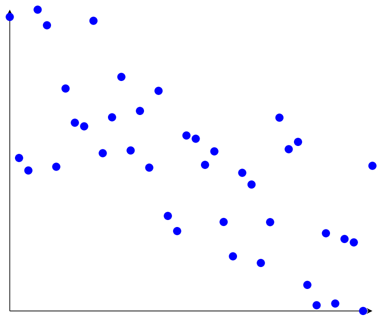
In the 8th grade, we focus on association in general, as well as linear association specifically (the latter more formally known as "correlation"). *Association* between two random variables refers to evidence of dependence, regardless of the nature of that dependence (e.g., linear, quadratic, or other). Loosely speaking, we call an association *positive* if Y generally increases as X increases, *negative* if Y generally decreases as X increases, and no (or zero) association if Y tends to remain the same regardless of changes in X. A linear association refers to an association that is well-captured by a linear relationship; exactly how that linear relationship is determined (i.e., how to sketch a "best-fit" line, and how "well" a line captures the behavior of bivariate data will be discussed in Section 6.3. A perfect linear association occurs if the data fall exactly on a line of either positive or negative slope. If the data fall perfectly on a horizontal line (zero slope), there is technically a linear (but uninteresting) association, as change in X does not influence Y .

Consider the following scatter plots.

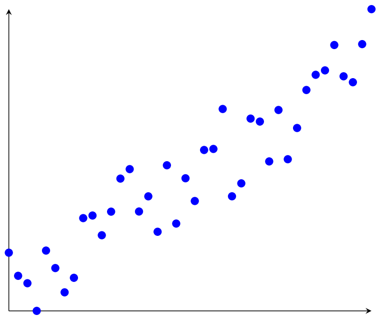
Positive Correlation



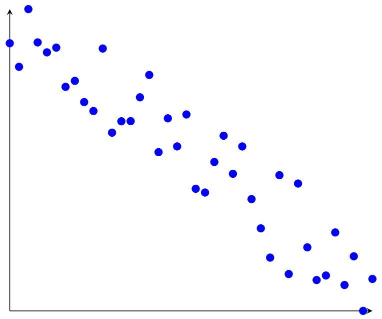
Negative Correlation



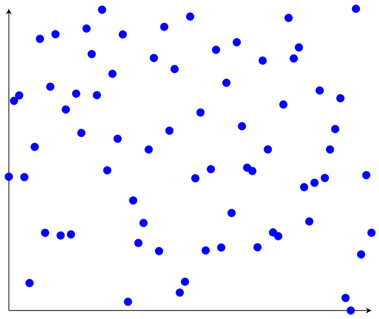
Strong Positive Correlation



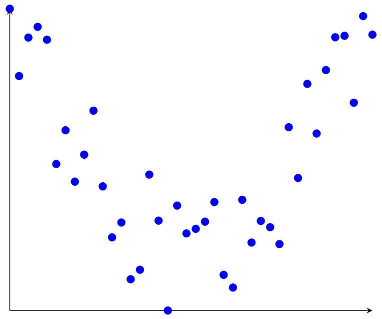
Strong Negative Correlation



No Correlation

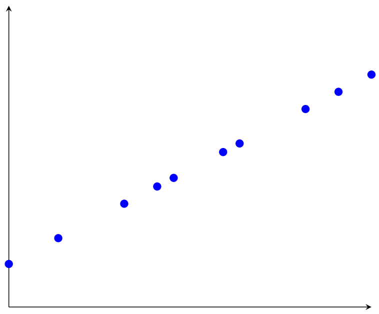


No Linear Correlation

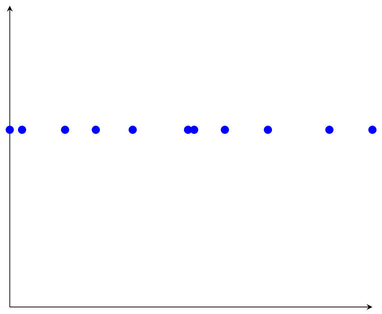


Perfect linear association occurs when all points fall upon the same line:

Perfect Linear Association



Perfect Linear Association (zero correlation)



While 8th grade students need only discuss “association”, as opposed to “correlation”, it is important for teachers and parents to know the difference, should the topic happen to arise. The above example of “perfect linear association” shows a positive correlation. However, contrast this example with the scatter plot on the right depicting points lying on a horizontal line. While the points have a perfect linear association, they have zero correlation, as Y is completely unaffected by the existence of X. Again, this needs not be discussed directly in the 8th grade, but teachers should have this distinction tucked away in the back of their minds.

Example 4.

> Recall the example of Izumi and her basketball data (return to figure 1). Upon studying the scatter plot of field goal data, we see that the data suggest a strong positive linear association between the number of field goals made and the number of field goals attempted over the course of a basketball season. Izumi thinks, “I suppose this positive association makes sense, because these players are pretty good at what they do; I would expect that more field goals attempted by these skilled players would be associated with more field goals made. Of course, as a side-note, this positive association may not hold true for unskilled basketball players, such as my cat, Mittens. No matter how many field goals Mittens attempts, she’s probably not ever going to get the ball in the basket.”
>
> Izumi then turns her attention to her Rebounds vs. Assists scatter plot (Fiigure 2), noticing a general negative linear association. She thinks, “This association seems weaker than that between the field goal variables, because these data points seem to be more scattered about the plane.” After noting the negative association, Izumi thinks more deeply about what this might mean. “My scatter plot suggests that players who tend to have more assists also tend to get fewer rebounds. I guess this makes sense because players who tend to get assists are usually farther away from the basket, assisting to those players who tend to be closer to the basket. Of course! It’s those players who tend to play closely to the basket who tend to get the rebounds.”

Later, in Example 7, we discuss the important difference between association and causation. For now, note that an association between variables does not imply that changes in one variable *cause* changes in the other variable.

In addition to observing trends (i.e., linear or nonlinear associations) in bivariate data, 8th grade students are to describe other characteristics, such as clustering and outliers. *Outliers* are data points that notably deviate or “stand out” from the general behavior of the data set. In 6th grade students studied several techniques to locate such standouts; for bivariate data we make use of clustering.

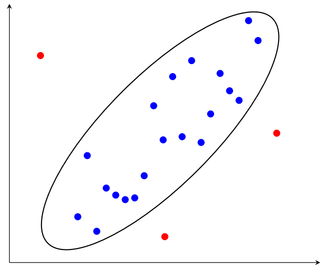


figure 3

In general, *clustering* refers to a set of data points that are in close proximity to each other. A single scatterplot may have many clusters of different sizes, and different clusters may be of different scientific interests. To exploit clustering for the purpose of identifying outliers, we consider the process of drawing a region around what seems like a self-consistent cluster of the data (see figure 3) for an example.

In this figure, even without the ellipse, it is clear that, except for three points, the data follow generally a positive linear trend. Often the data are not so convenient, and it takes some skill to identify outliers. In any case, whether easy or hard to identify, once the outliers have been identified graphically, the researcher still must give justification to treat them as outliers *in terms of the context*. Often, particularly in biomedical research, the outliers are the important data points (perhaps identifying risk factors in a new medication being tested). There are various tools that researchers can use to identify clusters and outliers (one is *projection pursuit*); but even when using those tools, the researcher must seek an explanation of outsider status (i.e., experimental error, an erratic member of the sample, origin of data suspect, etc.).

In figure 2 we see a definite negative trend in the data (rebounds decrease as assists increase), but the tendency follows a broad band rather than a line. Let’s see how Izumi treats this:

EXAMPLE 5.

Upon studying her Rebounds vs. Assists scatter plot (figure 2), Izumi concludes that the negative association would be stronger if not for the data point JM (the initials of Juanita Martinez), which lies rather strikingly outside from the main graphical trend. This particular datum is likely an outlier. Izumi thinks, "I wonder if there is some reason why Juanita's datum stands out from the rest of the team. Oh of course! Juanita transferred to my school from another one mid-season! Perhaps I should not include her in my data analysis." Here, Izumi has scientific reason to drop this outlier from her data set. Izumi will further investigate the ramifications of dropping Juanita from her analysis in the next section.

Note that Izumi has a reason, based on the circumstances, to identify Juanita as an outlier. Looking at the graph, she may also have considered Susan Garcia (SG) and Joan O'Connell (JO) as outliers, but not found a contextual reason to exclude them, so left them in the data analysis. In almost every such analysis, the identification of clusters and outliers (even if performed by a computer algorithm) has to be reinterpreted and justified in terms of the context. The important thing is that students are engaged in scientific discussions regarding various reasons for identifying various points as potential outliers, and consequently investigate dropping this particular datum. At the same time it should be stressed that there does not need to be an identifiable anomaly (e.g., some characteristic about the nature of the datum), such as Juanita being on the team for only part of the season, that logically sets it apart from the rest of the data) associated with a datum to label it an outlier. Outliers are simply data points that "stand apart from the general trend," regardless of the reason. So, the mathematics identifies "potential outliers," but the context explains why they should be excluded. Outliers should not be excluded simply because the mathematics has so identified them. An outlier may flag a confounding variable that is interfering with the relationship of interest. Or an outlier may just be an anomaly that should be ignored and dropped from the data analysis. Only an analysis in context can distinguish erroneous data points from critical pointers toward further research, and in fact, may not do so conclusively.

Section 6.2: Linear Models for Problem Solving

Construct and Assess Best Fitting Lines

Know that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line. 8.SP.2

We deepen our understanding of an association of two variables by fitting (as best as possible) a straight line to the scatter plot of collected data. There are algorithms for determining (in some measure of distance between data sets) the "best fitting line," but here we will just eyeball the data. Return to the scatterplots on page 6: the first four suggest a linear relation (the third and fourth more strongly than the first and second) and the fifth and sixth suggest that there is not a linear association. Just as a mean, median, or mode provides a single-point (zero-dimensional) description of an univariate data set (1-dimensional), a line provides a one-dimensional summary of a bivariate (2-dimensional) data set. Furthermore, just as there are various ways to "measure the center" of a one-variable data set, there are various ways to fit a line to bivariate data. Here, we explore a number of options, focusing on the "eye-balling" technique. Recommended eye-balling software include the Illuminations website from the National Council of Teachers of Mathematics:

`http://illuminations.nctm.org/ActivityDetail.aspx?ID=146`

This software has the feature that one can easily load data, and then can eyeball a best fitting line as well as ask for the calculation of "a best fitting line." You will be impressed how well the eye-ball guess is to the one created by formula. (There is a physical explanation of this: the eye reacts to the energy produced by the input, and the mathematical formula is based on a concept of energy between two sets of data).

If students want a method of fitting lines that is both accessible and consistent between individuals (as opposed to "eye-balling," wherein each student's line will slightly differ from those of others), the teacher may want to investigate the median-median line. The median-median line is accessible to 8th graders because they have previously learned the concept of median as a measure of center for univariate data (furthermore, the concept of median is reinforced). The Quantitative Literacy Series book *Exploring Data* (Dale Seymour Publications, 1986) provides an excellent explanation of this method. While "least squares" (the energy method) is the most widely-used technique by scientists, it is too computationally intensive for the 8th grade, which should be informal and exploratory. The median-median line provides a precise algorithm by which different students can obtain the same line, requiring only visual (as opposed to computational) techniques. It is a pedagogical tool, rather than a realistic tool.

The line that one fits to a scatter plot is meant to capture the behavior of the bivariate data, but in a simpler form than the entire plot. Using this line, one can make estimated predictions (e.g., interpolation and extrapolation) about the random variables X and Y, as they behave together.

Example 6.

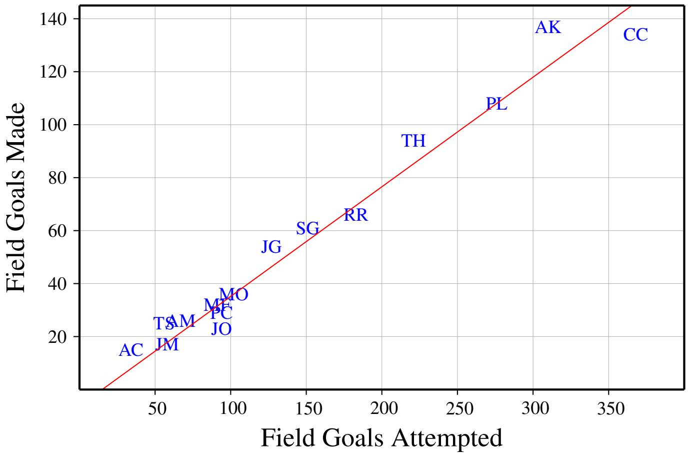


Figure 4: 2012-2013 Girls Basketball Data

To summarize the general behavior of her data, Izumi decides to draw a line to fit her scatter plots. Izumi first looks at the scatter plot of Field Goals Made vs. Field Goals Attempted (figure 1). Using her artistic skills, she adjusts her transparent ruler on the plot until she thinks she has found a line that is as close to each datum as possible. That is, Izumi tries her best to find the line that minimizes the distance between each data point and that line. After carefully changing the angle of her ruler, she decides to trace the red line shown in Figure 4.

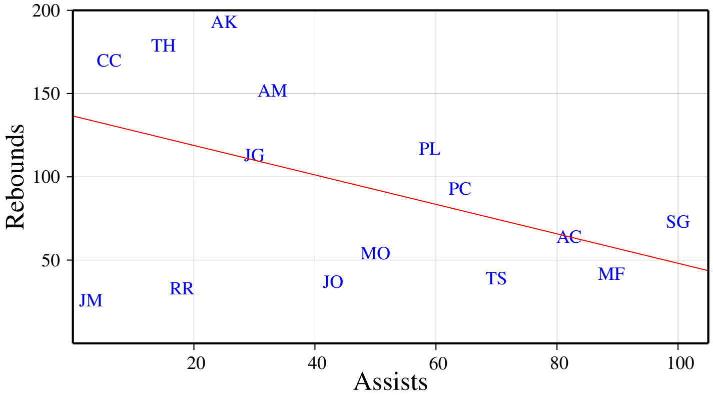


Figure 5: 2012-2013 Girls Basketball Data

Next, Izumi turns to her Rebounds vs. Assists data, using her ruler to fit the red line as shown in Figure 5. Notice that the data points in Figure 4 are more tightly clustered around Izumi's best-fit line than the points in the Rebounds vs. Assists scatter plot Figure 5). In fact, figure 5 supports the earlier decision to exclude as an outlier the point denoted JM: If we ignore that point, clearly there is a better fitting best fitting line. Since the context supports the decision to consider JM as an outlier, now Izumi eyeballs a best fitting line with that point excluded, and generates Figure 6.

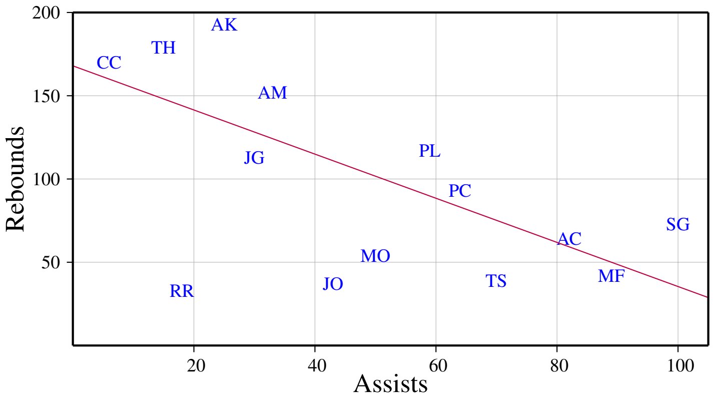


Figure 6: Girls Basketball Data, Outlier Removed

Izumi notices that the fit is better than in Figure 5; in fact, Izumi created a composite, showing both lines (Figure 7) to drive home this observation.

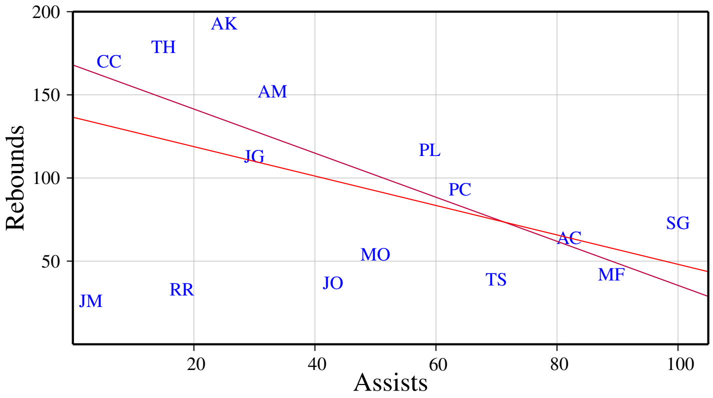


Figure 7: 2012-2013 Girls Basketball Data

Using Linear Models to Solve Problems

Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept. For example, in a linear model for a biology experiment, interpret a slope of 1.5 cm/hr as meaning that an additional hour of sunlight each day is associated with an additional 1.5 cm in mature plant height. 8.SP.3

In the preceding section we "modeled" bivariate data (X, Y) with a "best fitting line." What can our line suggest to us regarding the true relationship between random variables X and Y? First, let's be careful about this question: we don't know the true relation; we don't even know if there is a relation. So the question really is: what does the best fitting line allow us to conclude about our relation, and can we justify those conclusions with scientific arguments. Let's return to our good friend, Izumi, who is wrestling with these issues in a specific context.

Example 7.

Izumi notes that her lines of best-fit have slopes consistent with her original scatter plot assessments: The Field Goals Made vs. Field Goals Attempted plot suggested a positive linear association between her variables, and her line of best fit has a positive slope (see figure 4 above) ; likewise the negative association between the variables of her Rebound vs. Assists scatter plot is reflected by the negative slope of her best-fit line (see figure 7). Izumi realizes that she can calculate the slopes of her lines and make quantified statements about her variables.

As Izumi has already learned how to compute the slope of a line and locate the line's y-intercept, she approximates the following equation for her field goal line: $Y = 0.41X - 3.4$. Izumi thinks about what this means and concludes, "My line of best fit suggests that approximately each additional shot that a basketball player at my school attempts during a game is associated with a roughly 41% increase in the number of her made field goals for the season. Of course, this doesn't say anything about a given individual player; it just describes my data in general. It gives me a good guess at what I might expect from a random player, but I can't be certain." Izumi continues to think about her equation, focusing on the y-intercept of $(0, -3.4)$. "Hmm, speaking of making guesses, my line would predict that a random player who had zero field goal attempts would have made -3.4 of them! That's ridiculous! It just goes to show that my linear model (a linear equation describing a relationship between variables, informed by observations) has limitations. While my model may make reasonable predictions for numbers near the range of my data, it may not make sense for extremes."

As Izumi continues to think more deeply about what this association might be telling her, she realizes, "While it makes sense that a higher number of field goals made necessarily means that at least that number of field goals were attempted, it is interesting that, for any given person in general, the number of attempted field goals does not necessarily cause the number of made field goals to be a certain number. Instead, all that I can say is that my data suggest an association between the two variables. For example, anybody can throw the basketball in an attempt to make a field goal, but it's the player's skill (and perhaps the defense's lack of skill?) that actually causes the ball to go into the basket."

After patting herself on the back for recognizing the difference between causation and association, Izumi turns her attention to the Rebounds vs. Assists lines of best fit. She calculates the slopes and intercepts for both the red (including the outlier, Juanita Martinez) and the blue (omitting the outlier) lines, and calculates the equations of those lines.

"Interesting," Izumi thinks. "If I omit the outlier, then my scatter plot suggests a stronger negative association between assists and rebounds than if I were to include the outlier. Specifically, if I include Juanita, then my linear model suggests that each additional assist is associated with a 92% decrease in that player's number of rebounds. However, if I omit Juanita, then my linear model suggests that each additional assist is associated with a 140% decrease in that player's rebounds!"

After thinking about the slopes of her lines in the context of her experiment, Juanita begins to think about the y-intercepts of her linear models. "If my red line were a reasonable model describing the relationship between rebounds and assists, then a reasonable expectation to have regarding a player who made zero assists for the season would be that the player made 138 rebounds. If my blue line were a reasonable model, then I might venture the guess that a player who makes zero assists would have made 172 rebounds for the season. While these numbers are predicted by my respective linear models, the usefulness is not clear to me, because I've never known a player to have zero assists for an entire season. But it makes for an interesting thought! Come to think of it, if a player has zero assists, then that player likely warmed the bench a lot more than she actually played, so it seems more reasonable to guess that she would have very few rebounds. Gee, it sure is important to continue to think critically, engage my brain, and exploit my knowledge of basketball while I analyze my data!"

Section 6.3: Analyzing Bivariate Categorical Data Using Two-way Frequency Tables

Understand that patterns of association can also be seen in bivariate categorical data by displaying frequencies and relative frequencies in a two-way table. Construct and interpret a two-way table summarizing data on two categorical variables collected from the same subjects. Use relative frequencies calculated for rows or columns to describe possible association between the two variables. For example, collect data from students in your class on whether or not they have a curfew on school nights and whether or not they have assigned chores at home. Is there evidence that those who have a curfew also tend to have chores? 8.SP.4

In the previous sections, our random variables have been *quantitative*. Scatterplots provide a natural way of visualizing bivariate quantitative data, because each real-valued realization of each quantitative random variable can be plotted on a number line (and thus a real-valued ordered pair can be plotted in the Cartesian coordinate plane). In contrast, this section investigates patterns of association between *categorical* variables, which are characterized by their qualitative nature (recall Section 6.1). Scatterplots are not useful for categorical bivariate data since categorical data cannot necessarily be ordered on a number line in any meaningful way. For example, consider the categorical variable "reptiles of Washington county, Utah". A few realizations of this categorical variable include "Desert Tortoise", "Chuckwalla", and "Western Rattlesnake". How sensible is it to plot these variables on a number line, given that there is no natural "order" affiliated with them? In short, number lines are reserved for numbers, so we need a to take a new approach to analyzing categorical data.

Two-Way Frequency Tables

Before we consider visual representations of bivariate categorical data, we first discuss a convenient way of summarizing such data: The *two-way frequency table*. The table is "two-way" because each bivariate datum is composed of an ordered pair of realizations from two categorical random variables. For example, a datum might be something like the ordered pair (female, non-smoker), or perhaps (green eyes, brown hair), or maybe (8th grader, basketball). The table is a "frequency" table because the cell entries count the number of subjects (i.e., the frequency of data points) that fall into each combination of categories. Consider the following example.

Example 8.

The Utah Fish and Wildlife Service has collected data regarding the protective status ("endangered," "threatened," "candidate," and "proposed/petitioned") of various Utah species (mammals, birds, reptiles, fishes, insects, snails, and flowering plants) with which the Utah Ecological Services is concerned. As of April 2013, the agency reported the following data in

`www.fws.gov/utahfieldoffice/endspp.html`

organized in a two-way frequency table. The first two categories are the protected categories, the category "candidate" includes those species that the Service has decided to consider for explicit protection, and the category "proposed/petitioned" consists of species brought to the attention of the Service by other groups. These data include only species that have both a protective status and are of interest to the Utah Ecological Services Field Office, disregarding all Utah species that do not have such status.

	Endangered	Threatened	Candidate	Proposed
Mammal	1	2	0	1
Bird	2	1	2	1
Reptile	0	1	0	0
Fish	7	2	1	0
Insect	0	0	0	1
Snail	1	0	0	0
Flowering PLant	11	13	6	4

Here, each datum is an ordered pair realization of the bivariate categorical random variable of the form (species type, protective status), such as (fish, threatened). Each cell of this two-way frequency table displays the frequencies (counts) of each possible combination of variables that are observed. For example, there is one mammal species with an "endangered" status, two mammal species with a status of "threatened", and six flowering plant species that are "candidates" for being granted a protective status.

In general, a two-way frequency table is designed as follows:

		Categorical Random Variable #2			
		Realization A **of Variable #2**	**Realization B** **of Variable #2**	**Realization C** **of Variable #2**	$\cdots$
Categorical Random Variable #1	**Realization A** **of Variable #1**	Frequency of (A_1, A_2)	Frequency of (A_1, B_2)	Frequency of (A_1, C_2)	$\cdots$
	Realization B **of Variable #1**	Frequency of (B_1, A_2)	Frequency of (B_1, B_2)	Frequency of (B_1, C_2)	$\cdots$
	Realization C **of Variable #1**	Frequency of (C_1, A_2)	Frequency of (C_1, B_2)	Frequency of (C_1, C_2)	$\cdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

We next give a simpler, fictional example to demonstrate the power of the two-way frequency table in the 8th grade setting.

Example 9.

Carlos enjoys spending time with his friends. He feels sad when one of his friends cannot hang out with him. Often when a friend cannot hang out, it is because the friend either cannot stay out late at night, or the friend is busy doing chores at home. Carlos notices that it tends to be the same group of friends who have curfews on school nights who also have chores to do at home. He wonders, "Do students at my school, in general, who have chores to do at home tend to also have curfews at night?"

Carlos decides to conduct an experiment to help suggest an answer to his question. He randomly surveys 52 students at his school, asking each student if s/he has a curfew and if s/he has to do household chores. To review some vocabulary, notice that Carlos's *experiment* is to record the responses of the 52 randomly selected students; the two *categorical random variables* of interest are "curfew status" (which has realizations "has curfew" and "does not have curfew") and "chores status" (which has realizations "has chores" and "does not have chores"). It is crucial that every subject in Carlos's study falls into exactly one category of each variable, that is, one cannot both have chores and not have chores.

Carlos observes that of the 52 students he surveyed, 31 have curfews and 35 have chores to do. Of the 31 students who have curfews, 26 also have chores to do. He summarizes the breakdown of his data in the following two-way frequency table:

	Has Curfew	No Curfew
Has Chores	26	9
No Chores	5	12

Notice how Carlos organizes his information: The realizations of the "curfew status" variable are the columns of the table; the realizations of the "chores status" variable are the rows of the table. Also notice that we can calculate the **marginal frequencies** (the count of the occurrence of one variable at a time).

	Has Curfew	No Curfew	
Has Chores	26	9	**35**
No Chores	5	12	**17**
	31	**21**	**52**

We see explicitly that there are 35 total students surveyed to have chores and there are 17 total who have no chores, as we take the total across the rows. Similarly, we see that there are 31 total students with curfews and 21 without curfews. Furthermore, note that the sums of each set of marginal frequencies must equal the total number of students surveyed: 35 + 17 = 52 and, likewise, 31 + 21 = 52.

It is always the case that the sum the marginal frequencies of a given variable equals the total number of subjects, so adding marginal frequencies provides a useful check for mistakes. As we will soon see, marginal frequencies help us answer important questions about our data. Let's get one more example under our belt before moving on to the interpretation of two-way frequency tables.

Example 10.

Emina loves to eat tomatoes from her garden in Salt Lake City. She asked her friend Renzo, "Don't you just love tomatoes?" Renzo crinkled his nose and replied, "Ew, tomatoes gross me out! When I see them in the grocery store, I just keep on walking." Renzo's response prompted Emina to think, "I don't buy tomatoes at the grocery store either, because I grow them in my garden. The tomatoes from my garden are delicious, whereas grocery story tomatoes look less appealing to me. I wonder if there is an association between enjoying tomatoes and having a garden at home."

Emina surveys 100 randomly-selected Salt Lake City vegetable-eating residents and asks each of them two questions: 1. Do you primarily obtain your vegetables at the grocery store (including food pantry), the farmer's market, or your home garden? 2. Do you like tomatoes? Emina summarizes her results in the following table:

	Grocery Store	Farmer's Market	Home Garden
Likes Tomatoes	50	4	12
Dislikes Tomatoes	30	1	3

Emina wonders if her data suggest an association between enjoying tomatoes and having a garden, but she's not yet sure how to use her data to investigate this question.

Making and Interpreting Two-Way Relative Frequency Tables

In this section, we transform our frequency tables into *relative* frequency tables, which often help us interpret data. A *relative frequency* refers to the ratio of the frequency of a particular realization of a bivariate categorical variable to the total number of observations. In other words, a relative frequency is a number between 0 and 1 (inclusive), commonly represented by a fraction, decimal, or percent. As a result, relative frequencies are useful in discussions of probabilities and thus interpretations of bivariate categorical data. We explain further by example, beginning with the construction of relative frequency tables, followed by their interpretation.

Example 11.

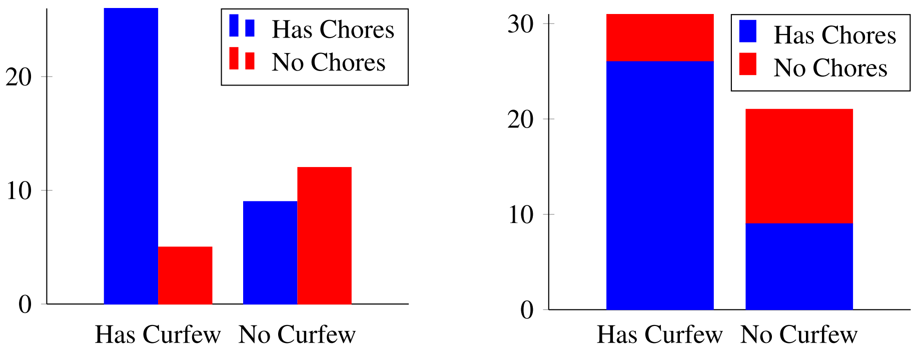
Recall Carlos's data regarding chores and curfew, specifically his two-way frequency table containing the marginal frequencies, copied below:

Frequency Table	Has Curfew	No Curfew	
Has Chores	26	9	**35**
No Chores	5	12	**17**
	31	**21**	**52**

The relative frequency table below was constructed from the table above.

Frequency Table	Has Curfew	No Curfew	
Has Chores	$\frac{26}{52} = 0.50$	$\frac{9}{52} \approx 0.17$	$0.05 + 0.17 \approx 0.67$
No Chores	$\frac{5}{52} \approx 0.096$	$\frac{12}{52} \approx 0.23$	$0.096 + 0.23 \approx 0.33$
	$0.50 + 0.096 \approx 0.6$	$0.17 + 0.23 = 0.40$	**1**

Note further that each pair of marginal relative frequencies necessarily have a sum of 1. Carlos can use his relative frequency table to draw conclusions such as, "Of the 52 randomly-selected students I surveyed, 67% of them have chores assigned to them at home, and about 60% of the students surveyed have a curfew." Carlos continues, "Perhaps the most striking observation to be made is that the bulk of students I surveyed (50%) fall into the category of both "Has Curfew" and "Has Chores"; the second-most-popular category is both "No Curfew" and "No Chores." This is interesting because it suggests an association between having a curfew and also having chores to do at home. That is, my survey suggests that students who have curfews also tend to have chores assigned to them." Carlos then used his relative frequency table to construct visual representations of his data, shown below (one with categories side-by-side, the other stacked).



Carlos's Data

Carlos constructed these graphs so that 50% of the cumulative bar area would indicate data falling under the "Has-chores-and-Has-Curfew" category, 23% would fall under the "No-Chores-and-No- Curfew" category, 17% would fall under the "Has-Chores-but-No-Curfew" category, and about 10% would fall under the "No-Chores-but-Has-Curfew" category. Such graphical representations often make it easy to visually inspect associations between variables. Since the vast majority of the "Has Curfew" bar is darkly shaded (indicating these subject also have chores), while the majority of the "No Curfew" bar is lightly shaded (indicating subjects who do not also have chores), the association is visually depicted.

Example 12.

Recall the Utah Fish and Wildlife Service data from Example 8. To help us create a two-way relative frequency table, we again first include the marginal frequencies to our original frequency table.

Frequency Table	Endangered	Threatened	Candidate	Proposed	
Mammal	1	2	0	1	**4**
Bird	2	1	2	1	**6**
Reptile	0	1	0	0	**1**
Fish	7	2	1	0	**10**
Insect	0	0	0	1	**1**
Snail	1	0	0	0	**1**
Flowering Plant	11	13	6	4	**34**
	22	**19**	**9**	**7**	**57**

Check that the following relative frequency table can be constructed from the above two-way frequency table. Note that the relative frequencies can be expressed in fraction, decimal, or percent form, which provides an opportunity for students to review and practice such concepts.

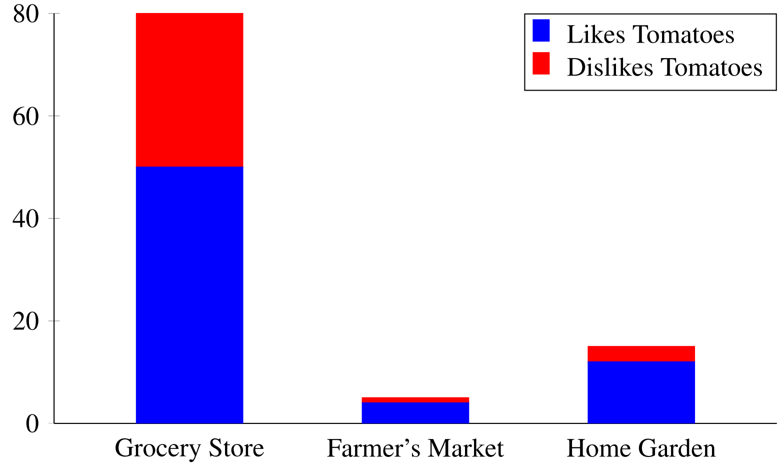
Realative Frequency Table	Endangered	Threatened	Candidate	Proposed	
Mammal	1/67	2/57	0	1/57	**4/57**
Bird	2/57	1/57	2/57	1/57	**6/57**
Reptile	0	1/57	0	0	**1/57**
Fish	7/57	2/57	1/57	0	**10/57**
Insect	0	0	0	1/57	**1/57**
Snail	1/57	0	0	0	**1/57**
Flowering Plant	11/57	13/57	6/57	4/57	**34/57**
	22/57	**19/57**	**9/57**	**7/57**	**57/57=1**

From the above chart, we can easily answer questions such as, “What percent of species with protective status in Utah are mammals?” Here, the marginal relative frequency of 4/57 tells us that only about 7% are mammals. One may ask why there are relatively few animals (mammals through snails) given protective status (about 40% of those with status) than flowering plants (about 60%). Perhaps the Fish and Wildlife Service has more of an incentive to classify plants than animals? Perhaps it is plants about which it is easier to collect data than animals which scurry about? Other questions one may be inspired to ask about these data may come from noting that, of the animals with protective status, the fish and birds far outnumber the snails and insects (16/57 and 2/57, respectively). Perhaps the Fish and Wildlife Service is more concerned with species of recreational interest (e.g., fishing and hunting)? Or perhaps there are biological or ecological reasons for the discrepancies?

Example 13.

Recall Emina and her tomato garden (Example 6.5.1c). Emina summarizes her data in the following relative frequency table and stacked bar graphs.

Frequency Table	Grocery Store	Farmer’s Market	Home Garden	
Likes Tomatoes	0.50	0.04	0.12	**0.66**
Dislikes Tomatoes	0.30	0.01	0.03	**0.34**
	0.80	**0.05**	**0.15**	**1.00**



Eminia's Data

Emina quickly sees from her relative frequency table that the majority (80%) of the vegetable-eating people she surveyed purchase most of their veggies at a grocery store, and that only 15% of those surveyed mostly eat veggies from their gardens. "What's most interesting to me," thinks Emina, "is that even though a small percentage of people surveyed use their gardens as their main vegetable source, of those 15%, a whopping 12 out of 15 people like tomatoes! That is, of those who have a home garden as their main veggie source, 80% (12/15) of them like tomatoes. This is a stark contrast with the grocery-shoppers: Of the 80% of people surveyed who buy most of their veggies at the grocery store, only 50 out of 80 like tomatoes, or just 62.5%. So it looks like there could be a positive association between having a home garden and liking the taste of tomatoes. I wonder if this means tomatoes are tastier out of a home garden than the store. Maybe I should offer Renzo a tomato from my garden...." Emina continues to think about her study results, and notices that 4 out of 5 people (also 80%) who obtain most of their veggies from the farmer's market also enjoy tomatoes. "Hmm. Eating tomatoes from a farmer's market is very similar to eating tomatoes out from a home garden, since the farmer's market produce is grown locally. Maybe I should pool these data together, since they're arguably telling me the same information about locally grown food." Emina continues to think deeply about her data, and after making the following graphs, concludes "Regardless, the ratios of darkly shaded (likes tomatoes) to lightly shaded (dislikes tomatoes) areas of individual bars on my stacked bar graph indicates that there is a positive association between locally grown produce and the enjoyment of tomatoes."