

DATA ACTIVITIES -- INSTRUCTOR'S GUIDE

EXPLORING DATA

LIST OF ACTIVITIES:

- D1: Getting to Know You
- D1: Meet the States Data
- D1: Introduction to Tinkerplots
- D2: Technology Activity: Choosing the Bins of a Histogram
- D2: The Shape of the Data
- D2: Matching Variables and Shapes
- D3: Collecting Some Data on Cities
- D3: V is for Variation
- D3: Technology Activity: Deviations, the Mean, and Measures of Spread
- D3: Measurement Bias
- D3: Matching Statistics with Histograms
- D4: Comparing Men and Women in the Class Dataset
- D4: Matching Statistics with Boxplots
- D4: Counting Pasta
- D5: Technology Activity: Using Tinkerplots to Study Relationships
- D6: Technology Activity: Guessing Correlations
- D6: Fitting a Line to Galton's Data
- D6: Fitting a "Best Line"
- D6: Technology Activity: Exploring Some Olympics Data
- D6: The Regression Effect
- D7: Predictable Pairs

Topic D1: Statistics, data, and variables

WARM-UP ACTIVITY: GETTING TO KNOW YOU

MATERIALS NEEDED: None

This activity is very appropriate for use in the first day of class. In this first topic, we are introduced to the concept of a variable that is a piece of information that is collected from an observational unit. Here our observational unit is the student and we collect different variables from each. After this activity, the students should have a clearer idea of variables and can better distinguish variables of the two types. I have suggested ten possible questions in the activity, but feel free to use any other questions that you think might generate interesting data.

The questions to these ten questions (and others that you or the students wish to ask) will generate an interesting dataset that can be analyzed later in the course. Here are some considerations in the design of appropriate questions:

1. **Measurement type?** Since we will be graphing and summarizing quantitative and categorical data, it is useful to ask questions that will generate questions of both types. In the list of questions, number of pairs of shoes is a quantitative variable, and the preference among water, soda or milk for an evening meal is a categorical variable.
2. **Height and gender** are two good variables to collect. It is not a sensitive question to ask about a student's height (in contrast, I wouldn't ask a student his or her weight). Heights of a single gender will tend to be bell-shape, while the heights of both genders tend not to be bell-shaped – you may see two modes corresponding to the average values for the two genders.
3. It is good to ask questions that will generate **count variables**. Counts, such as the number of pairs of shoes a student owns or the number of movie DVDs a student owns tend to be right-skewed.
4. I asked about student's go-to-bed time and wakeup time so I could later compute the hours of sleep.

5. Sometimes I ask student a question that is a **measurement**. An example of a simple measurement is a guess at the instructor’s age. (I used this question when I was relatively young and even the high guesses were not insulting.)

5. Ask questions that should generate **different responses for male and female students**. Examples of questions of this type are:

“How many pairs of shoes do you own?” (Girls tend to own more shoes.)

“How much did you spend for your latest haircut?” (Girls tend to spend more money for a haircut.)

6. Encourage your students to suggest questions to ask. To help them contribute questions, you can have them first talk in small groups and then have each group contribute a couple of questions. As the instructor, you don’t have to include all of the questions suggested by students. Don’t include inappropriate questions or questions that might generate uninteresting data. (Uninteresting data would be data with very little variation.)

When all of the students have completed answering all of the questions, collect the responses. You can prepare a datasheet containing all of the responses for all students that you can pass out to use for later activities or homework exercises. This dataset will have the basic structure shown below:

```
Student height gender other variables
1      69    male
2      68    female
3      64    female
...
24     73    male
```

Alternatively, you can prepare a computer data file that can be read into statistics software (such as Minitab or Fathom) in a future technology lab.

ACTIVITY: MEET THE STATES DATA

MATERIALS NEEDED: A special pack of state cards or a pack of baseball cards.

In this activity, pairs of students will explore data on special State Cards. On each State Card, a number of variables are listed (of both quantitative and categorical types) and the students get some initial experience looking for patterns in single variables or interesting comparisons between different regions of the country. This activity can also be done using baseball cards that are relatively inexpensive to purchase.

At this point in the class, the students have little exposure to graphing or summarizing data. So it is unreasonable to expect the students to use, say dotplots, to look at the distribution for a quantitative variable. But this activity gives the students experience in formulating interesting questions about variables and trying to use appropriate graphs or calculations to answer the questions.

Suppose a particular group decides to look at a state's population density. What type of questions would they ask about population density? They might wonder which state has the largest (or smallest) density. They might be interested in the population density for their home state and compare how this density falls relative to an "average" density of a state. They might be interested in looking for states with unusually small or large density values.

Once the group has written down a reasonable question to answer, then the next task is to construct a graph or perform some computation that will help in answering the question. When I grade this activity, I don't expect the students to use the same type of graph that I might think of using. In this activity, a common student graph is an index plot where the values of the variables are plotted as bars or points as a function of the observation number (the population density of the first state is graphed first, the density of the second state is graphed next, and so on). Although this index plot may not be the most informative graph, it can be useful in identifying the extreme values or guessing at an average value.

When the students are comparing two regions of states with respect to a particular variable, the objective is to use graphs or summary statistics to help in a comparison. Usually two graphs are needed – one graph for values of the variable for the first group of states and a second graph for values for the second group of states. Likewise, if one wishes to compare the regions quantitatively, then one might compute a mean, say, for the variable values for each region and make some statement on the basis of the values.

Again when I grade this, I don't expect to find the most helpful graphical or numerical comparison. It is acceptable to construct any graph or perform a computation that is helpful in answering the comparison question.

TECHNOLOGY ACTIVITY: INTRODUCTION TO *TINKERPLOTS*

TinkerPlots is a data analysis program designed specifically for students from grades 4 through 8. Essentially, *TinkerPlots* allows a student to construct his or her own graph using a basic toolkit of commands. Most statistics packages incorporate special types of graphs such as histogram, dotplot, and scatterplot, and the user chooses one of these special types to suit his/her needs. In contrast, *TinkerPlots* only provides basic graphing tools and the student uses these tools to organize and summarize his/her data.

We illustrate some of *TinkerPlots* basic commands by use of a simple example. Suppose you purchase a pack of baseball cards. Each card consists of a picture of a ballplayer together with some data about the player. This data includes the player's height and weight, his date of birth, and some of the statistics describing his pitching or batting performance in recent baseball seasons. Suppose that all twenty cards are baseball hitters and each card contains the player's batting average for the recent 2004 season.

Suppose you spread out all 20 cards out on your carpet. You are interested in organizing the cards in some meaningful way to get a better idea of the quality of the players on the cards. You are measuring quality of a player by his 2004 batting average. Here are some basic things you can do that correspond to basic tools provided by *TinkerPlots*.

ORDER. You can arrange or sort the cards by batting average with the player with the highest batting average on top.

SEPARATE. You could separate the cards into two groups corresponding to players with "high" and "low" batting averages.

STACK. Suppose you decide that a high batting average is over .300. Then after you separate the cards, it may be helpful to stack the cards so it will be easy to see the relative numbers of players in the two groups.

COUNT. You may be interested in counting the number of players with high and low batting averages.

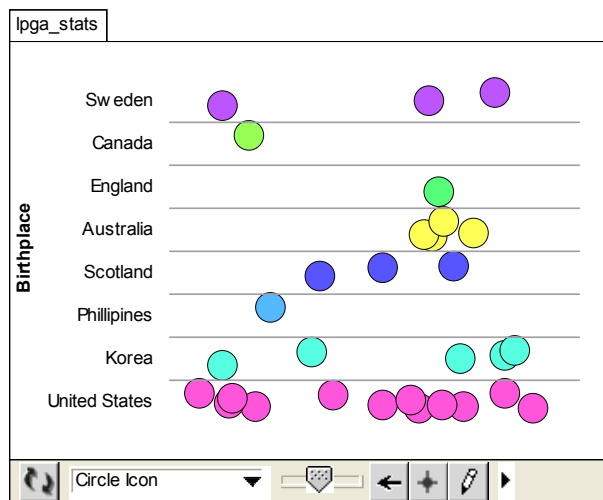
AVERAGE. You may wonder if there is any relationship between a player's batting average and the number of home runs that he hits. To check this out, you might want to compute the average number of home runs hit by the "high" batting average players and the "low" batting average players.

Learning about women professional golfers.

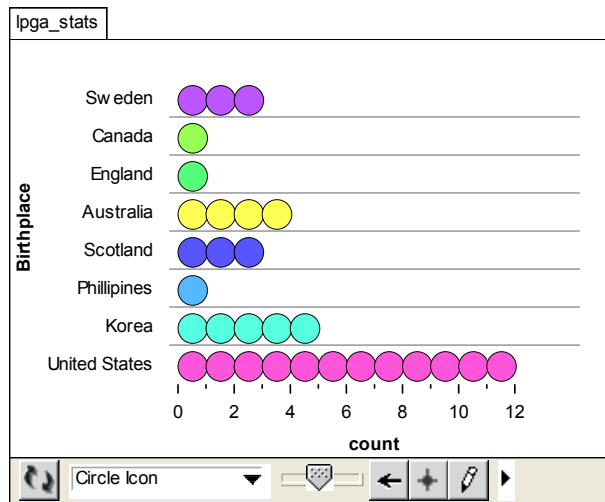
In this activity, the student is asked some directed questions and the object is to produce a *Tinkerplots* graph that will help in answering the question. There are many possible graphs for a particular problem and the instructor should be open to creative graphs that are different from the ones we typically see in a statistics class. Here I show some possible graphs that are helpful in addressing the question.

Q1: What countries are the golfers from? (Exploring the Birthplace variable.)

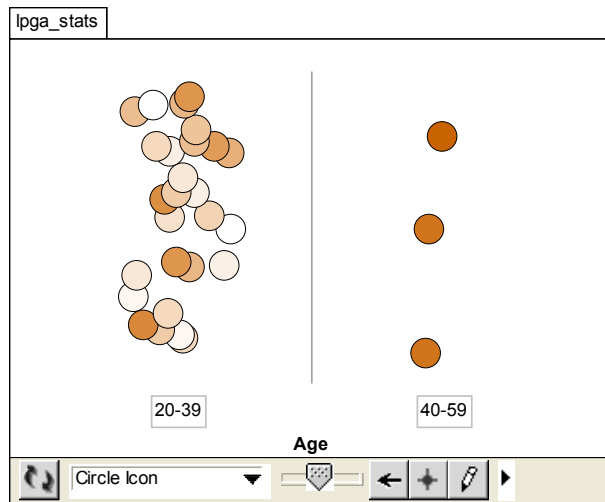
By use of the separation bar, I divide the golfers into 8 different countries.



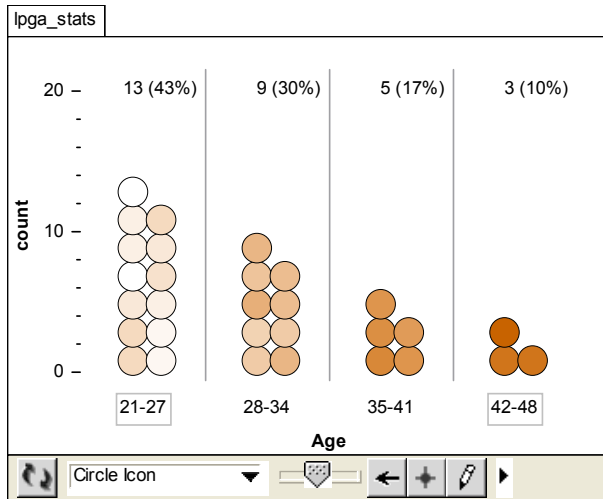
By stacking these icons, I see quickly that 12 of the golfers are American, followed next by Korea with 5, and by Australia with 4.



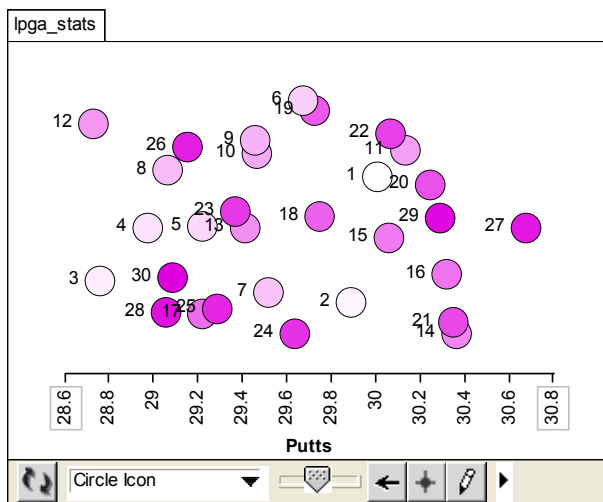
Q2: How old are these golfers? I separate the golfers into two groups by the separation tool. I see that only three of the top golfers are 40 years or older.



By dividing into four groups and having *Tinkerplots* show the frequency and percentage in each group, I see that 43% of the top golfers are 27 or younger.



Q4. Since I have heard the phrase “drive for show and putt for dough”, I might think that the number of putts per round is the variable that separates the top and bottom golfers on this list. To check this out, I graph the number of putts per round for all golfers and label each icon with the rank of the golfer. I don’t see any clear relationship between the number of putts and rank. The best putters (those with the smallest number of putts per round) have ranks 3 and 12, and the worst putter has rank 27. The best golfer Annika Sorenstam averages about 30 putts per round that is relatively high in this group.



[Topic D2: Graphing data](#)

TECHNOLOGY ACTIVITY: CHOOSING THE BINS OF A HISTOGRAM

When one constructs a stemplot or a histogram, a decision has to be made about how to group the data into bins. In this computer activity, students will get experience in constructing histograms using different bin sizes and seeing the effect of the bin choice on the appearance of this histogram. If one chooses a small number of bins or equivalently a large bin width, then one will get a box-shaped histogram that is a poor match to the underlying pattern of the measurements. (Generally, the histogram will be a biased estimate of the underlying population density.) On the other extreme, if one chooses a large number of bins or equivalently a small bin width, then it will be a closer match to the underlying pattern of measurements but it will be very bumpy since you have a lot of random variation in the height of each bar. (The histogram will display small bias but high variance.) There is a compromise choice for the number of bins that will appear to be the best fit to the underlying measurement pattern. The goal of this activity is not to find precisely the optimal number of bins, but to understand that it not good to choose a small or a large number of bins, and the choice of bins can have a big effect on the visual appearance of the histogram.

Part A: In this part, *Fathom* is used to draw a histogram from 500 test scores randomly simulated from a bell-shaped (normal) distribution. The actual population density is drawn on top of the histogram. By using the mouse to graphically adjust the number of bins, students should see that a small number of bins (with a bin width of 15) is a poor match to the population density. Also the choice of a large number of bins (with a bin width of 2) doesn't work very well – in this case there is much random variation in the bar heights. The students are asked to find a “good” choice of bin width – the answer should be a number between 2 and 15.

The choice of bin width depends on the number of data values. Generally if you observe more data, then you can use a smaller bin width. In number 4, the students are asked to find a suitable bin width for a sample of 50 test scores. Since you have less data, you would need to use a larger bin width.

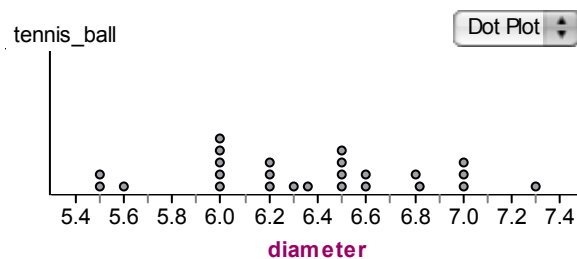
Part B and Part C: In part A, we worked with symmetric data. In Part B and Part C, the students are asked to find optimal bin widths for skewed data and data that has two humps. The exact choice of bin widths for these problems is not important, but the student should understand that bin widths chosen too small or too large will result in poor histograms that will not be good estimates at the underlying process.

ACTIVITY: THE SHAPE OF THE DATA

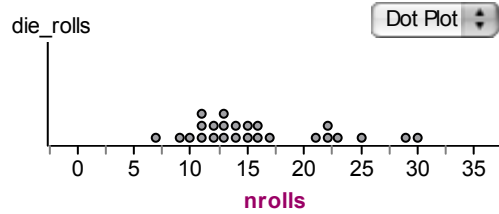
MATERIALS NEEDED: Several tennis balls, a set of dice, and a set of rulers with a centimeter scale. (Different type of spherical objects can be used instead of tennis balls.)

In this activity, students get experience in taking different measurements, graphing the data, and studying the shape of the measurement distributions. The five measurements described in the activity are chosen to demonstrate common distribution shapes. As seen below, it is possible to make substitutions if particular materials are not available for this lab.

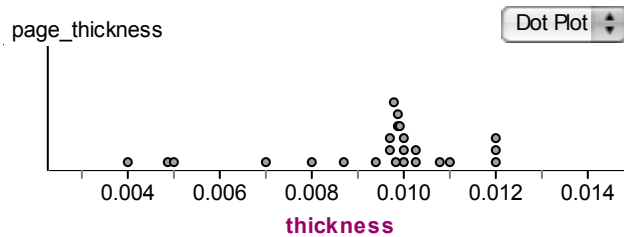
1. Diameter of a tennis ball. This is an example of a physical measurement where it is difficult to obtain the true value exactly. If you don't have a tennis ball, then you could substitute the length or perimeter of some object where it is difficult for a student to accurately obtain the true value using the measuring instrument. In this situation, the student measurements are typically bell-shaped about the true value.



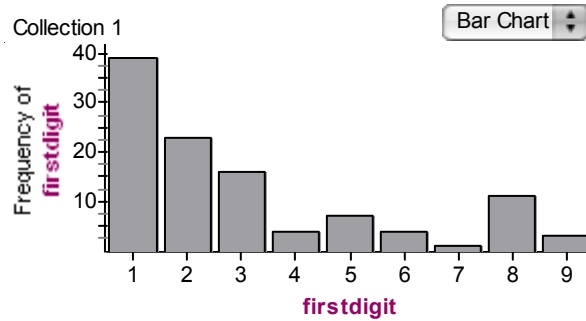
2. Number of rolls of a die until all six sides are obtained. This is an example of a count, where you are counting number of trials, seconds, etc. until a particular event occurs. Counts are typically right-skewed. Here is a dotplot of the observed counts for one class:



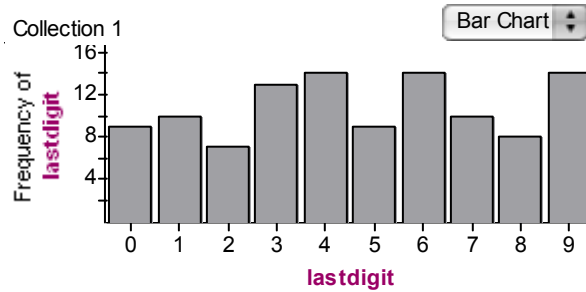
3. Thickness of a sheet of paper. This also is an example of a physical measurement where it is difficult to obtain an accurate estimate at the true value. This differs from the tennis ball measurement in that the student has to perform an arithmetic calculation – here a standard method is to measure the thickness of the entire book and divide by the number of pages. These measurements will also tend to be bell-shaped, but you may observe some outliers corresponding to students who make an error in the arithmetic calculation. In the graph below, I deleted one large value of 0.099.



4. First and last digits. If you have tables of any type of physical phenomena (city or county populations, scientific numbers, etc.) then the distribution of the first digits tends to be skewed right on the values $\{1, \dots, 9\}$ and the distribution of the last digits tends to be uniformly distributed on the values $\{0, \dots, 9\}$. Here are bar charts of the distributions using the first two columns of the county population data from the book.



count ()



count ()

Shape and measures of center. For each measurement distribution, the student is asked to compute the mean and median and relate the comparison with the shape of the data. If the data shape is symmetric, then the mean and median will be approximately equal. In contrast, if the data set is skewed right, say, then the mean will tend to be larger than the median.

In this activity, feel free to explore the shapes of other batches of data collected by the students. For example, you can explore some of the variables collected in the “Getting to Know You” activity.

ACTIVITY: MATCHING VARIABLES AND SHAPES

This activity should be used after the students have had some experience in exploring data distributions and understand the basic shapes. Here students are asked to match histogram shapes with the variable names. This can be a difficult task, so the activity divides the histograms into two groups (Group 1 and Group 2), and gives some hints about special characteristics of the different variables. This activity can be facilitated by discussion, either within a group of students, or between the instructor and the students.

Here are some of the main distinguishing elements of the two groups of histograms.

Group 1: Histogram A is notable in that it has spikes at equally spaced values. Histograms B and F are approximately symmetric – Histogram F is not quite as symmetric, since large values are more common than small values. Histograms C and D are both right-skewed.

Group 2: Histograms G and J both are right-skewed. Histogram H has an uneven shape with three possible peaks. Histogram I is left-skewed where small values of the variable are less common than large values.

After you make some general comments about the histogram shapes, you can talk about special characteristics of the variables described in the activity. This matching is best done by doing the easiest ones first, and making the subtle distinctions between histograms at the end.

[Topic D3: Summaries of data](#)

TECHNOLOGY ACTIVITY: COLLECTING DATA ON CITIES

This Internet activity is helpful for reviewing some of the material on graphing and summarizing a single batch in Topics D1 and D2. The student chooses 20 U.S. cities of interest and collects data on two variables from the website www.cityrating.com. For each variable, the student does a complete data analysis, including the construction of a suitable graph and computation of summary statistics. Perhaps the most important aspect of this work is a written paragraph where the student summarizes the main features of each dataset. Generally, I like the students to talk about the distributional shape, say something about the “average” and spread of the data, and mention any special features of the data such as unusually small or large values. It might be nice to ask the students to talk about any feature of the data that was surprising or unexpected to them before they started.

ACTIVITY: V IS FOR VARIATION

The concept of variation is more difficult for students to grasp than the concept of average. This is a nice in-class activity since the “V-span” reminds us that the focus is on Variation. This activity is best done in small groups of size four. Each student measures his/her v-span and the v-spans for the members of a group are entered into the table in the activity. From this table, one finds the deviation of each observation from the group mean, and these deviations are used to find the MAD (mean absolute deviation) and the standard deviation s .

The value of MAD measures a typical size of a deviation of a student’s v-span from the group mean. Groups whose members have similar v-spans will have a small value of MAD, and groups with variable v-spans will have a large MAD.

The question about the large family gathering should help clarifying the meaning of MAD. Adults will tend to have large v-spans than children – you can document this by comparing the mean v-span for adults against the mean v-span for children. But which group will have the larger value of MAD? Generally, groups with small physical measurements will tend to have smaller variability than groups with large physical measurements. So the children will tend also to have a smaller value of MAD.

TECHNOLOGY ACTIVITY: DEVIATIONS, THE MEAN, AND MEASURES OF SPREAD

This activity uses the dynamic features of *Fathom* to illustrate the concept of a deviation. After importing the data into *Fathom*, a slider is used to define a “typical value” m , and a new Attribute “deviation” is defined to be the difference between the data value and m . A Summary Table will be used to compute the sum of the deviations about m . By playing with the value of m using the slider, the student finds the value of m that makes the sum of deviations about m equal to zero. Hopefully they will know that the value of m that has this property is equal to the mean.

The second part of this activity focuses on the construction of alternative measures of spreads using the deviations about the mean. The student computes the deviation sizes and graphs them using a dotplot – by looking at the dotplot, he or she can make a statement about the typical size of a deviation. Then the student computes the

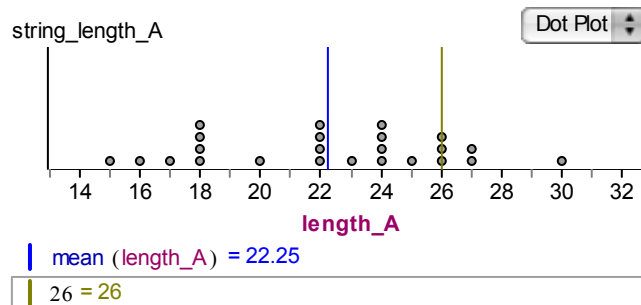
MAD (mean absolute deviation) and the standard deviations s . The student should understand that the MAD and s are simply measures of a typical size of a deviation.

ACTIVITY: MEASUREMENT BIAS

MATERIALS NEEDED: Two strings of different lengths, where the exact length of each string is known. A set of cardboard measuring instruments.

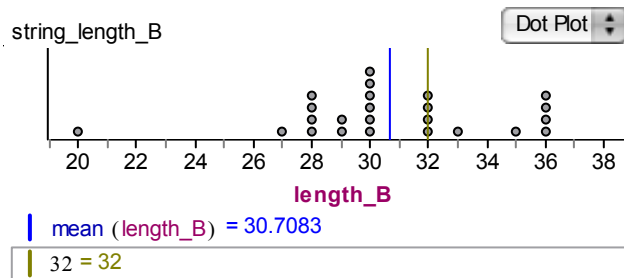
This activity introduces the idea of measurement bias. Often measurements are taken, say by sampling or some experimental procedure, where the measurement process leads to a systematic over or underestimation of the underlying true value. The difference between the average measurement and the true value is called the bias of the process. A simple example of bias is an Internet poll. Suppose an on-line poll is conducted to determine the attitude of the U.S. adult population regarding the use of the internet in education. Suppose 1000 people respond to the poll and 700 (70%) are supportive of the use of the Internet in education. This estimate 70% is a biased since the sample is not representative of the population of U.S. adults – it is likely that the true proportion of adults favoring technology in education is smaller than 70%. The difference between 70% and the true percentage (which is unknown in this case) is the bias.

In this activity, we use a measurement process that is known to have a bias. It is difficult to estimate the length of the string visually and there is a general tendency to underestimate the string length. In one class, STRING A had a true length of 26 inches and the following dotplot shows the student measurements:

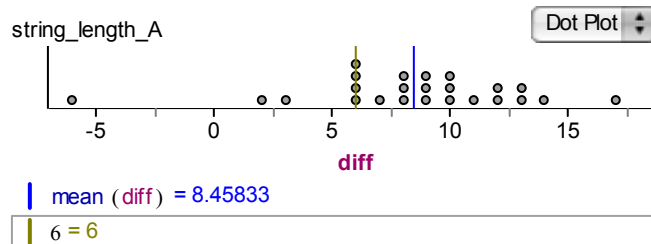


Note that the mean measurement here was 22.25 inches, so the bias would be $\text{BIAS} = 22.25 - 26 = 3.75$.

In the second part, the students are asked to estimate the length of a string that is slightly longer than the first. In this class, I used a STRING B length of 32 inches and here is a dotplot of the measurements. Here note that there is still a bias, but it is smaller in magnitude than the bias in the first estimation problem. Perhaps the students are learning to improve their estimates.



The students are also used to estimate the difference in the lengths of the two strings. One can compute the difference $\text{length_B} - \text{length_A}$ for each student – here is a dotplot of the differences:



Here note that the mean difference is 8.5 inches which is about $2\frac{1}{2}$ inches longer than the actual difference of $32 - 26 = 6$ inches.

The cardboard measuring instrument is a good example of an optical illusion that will lead to a measurement bias. The lengths of lines collected from a group of students generally will tend to be different from the actual line length.

Although students may understand the concept of bias as applied to measuring string lengths or line lengths, but be unsure how this relates to real-life situations. For that reason, it is helpful for the students to do the Extension where they find an article

from the median that reports conclusions from collected data that may be subject to a bias. Since the word bias can have a different meaning especially when used in the context of the media (this newspaper reporter was biased), it is important for a student to explain how an article may display a statistical bias.

ACTIVITY: MATCHING STATISTICS WITH HISTOGRAMS

This is a good activity since the students have to connect summary statistics with the graphs of a distribution. It is a difficult task to estimate values of a mean and standard deviation from a single histogram. But here the students have several histograms and they can find the matching summary statistics by comparing graphs.

Here are some basic ideas that are helpful to give as hints to complete this activity.

1. If the dataset is symmetric, then the mean and median are approximately equal. Large differences between the mean and median indicate skewness in the distribution.
2. For a bell-shaped distribution, most of the distribution (precisely 95%) falls within two standard deviations of the mean. So the spread of the distribution is approximately $4s$, where s is the standard deviation. You can use this to estimate the value of s from a bell-shaped histogram.
3. A good way to start is to identify the summary statistics for “extreme” histograms. For example, looking at the eight histograms, histogram F stands out since the scores are high (large mean) and it has small spread (small standard deviation). This clearly has the summary statistics 5 (mean of 86.68 and standard deviation 8.42).
4. Comparison of histograms also helps. For example, histograms B and D seem to have similar spreads but B has a mean around 60 and D has a mean around 70. Looking at the summary statistics, I would match these to statistics 1 and 3, respectively.

[Topic D4: Comparing batches](#)

ACTIVITY: COMPARING MEN AND WOMEN IN THE CLASS DATASET

This activity uses the class data that was collected early in the semester. When you think of questions to ask for this dataset, it is helpful to think of variables (such as the cost of the most recent haircut and the number of pairs of shoes owned) where you expect different responses for males and females.

To make this comparison, I ask the students to first construct parallel dotplots or stemplots – this helps in ordering the data which is helpful in computing the five-number summaries. By constructing parallel boxplots, the students see visually the spreads of the two batches, and he or she can see a comparison is possible (when the spreads are approximately equal).

Students often go through the manipulations without answering the key question: How much greater is one group than the other? It is common to say that one group tends to be larger, but this may have been obvious from the start. We know men are taller than women, but how much taller is an average man than an average woman?

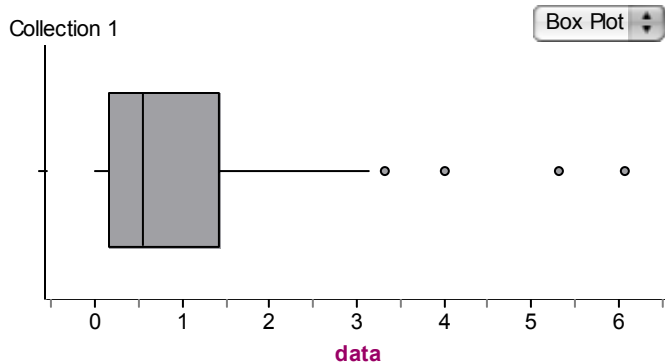
ACTIVITY: MATCHING STATISTICS WITH BOXPLOTS

A boxplot essentially is a graph of a five-number summary. But how is it related to other summary measures such as a mean and standard deviation? In this activity, students are asked to make this connection by matching boxplots with sets of statistics.

This might seem on the surface to be a difficult task. But a boxplot is informative about the general shape, location, and spread of a dataset:

1. The median line of the boxplot measures the center of the dataset and will be approximately equal to the mean for symmetric data.
2. The width of the box part of the boxplot measures the spread of the middle 50% of the data; that is, the width is the IQR. When comparing two boxplots, if one boxplot has a larger IQR, then likely it also has a larger value of s .
3. The shape of the boxplot reflects its shape. If the line for the median is approximately at the midpoint of the boxplot, then the middle 50% of the data is symmetric. In

contrast, as shown below, if the data is right-skewed, then the median line will be closer to the line for the lower quartile. Here the mean will be larger than the median.



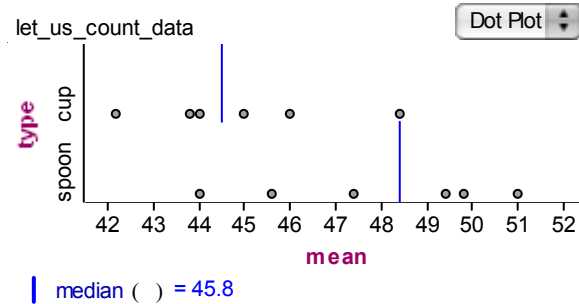
ACTIVITY: COUNTING PASTA

MATERIALS NEEDED: Several boxes of pasta shells. A set of clear one-cup measures and a set of plastic spoon measures.

In this activity, you compare two methods for measuring a half-cup of pasta. The class is broken down into pairs, where half of the pairs are measuring pasta in a plastic cup and the other half are measuring pasta using plastic spoons. The point is to use the notion of standard deviation to compare the two measurement processes and decide which one is superior.

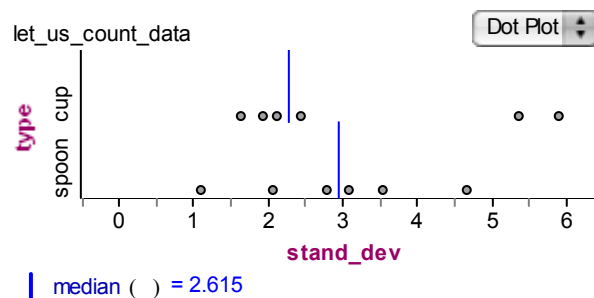
When each pair of students has made five measurements of pasta, they summarize the measurements by the computation of a mean and a standard deviation. The instructor collects all of the summary measurements and distributes the means and the standard deviations to the whole class. How can one use these summaries to decide which measurement process is better?

In my class I had 12 pairs of students doing this experiment and 6 pairs used the cup and 6 pairs used the spoon. Below I have drawn parallel dotplots of the means for the cup and spoon measurements.



There is quite a bit of variability in these means, but it is interesting that the spoon tended to give larger counts than the cup. This suggests that there might be a small bias in using the spoon. Perhaps it is harder to fill the spoon with a level amount of pasta and the extra roundedness in the measurements results in counting more pasta.

But actually if we are interested in obtaining consistent measurements, then we should focus on the standard deviations of the cup and spoon measurements.



This is not an easy comparison since there seems to be a lot of variability in these standard deviations. (It is possible that there are some errors in the calculations of s.) But if we ignore the two high values of s for “cup”, then it seems that the standard deviation for cup tends to be lower than the standard deviation for spoon. Thus, using this criteria, the cup is the more stable process for measuring pasta and should be preferred.

The activity talks about “special cause” and “common cause” types of variation. Common cause is the type of variation that is common for all groups making the measurement. For example, if there is some general difficulty in using the cup in measuring pasta, then variation due to this difficulty would be common cause. There may be variations in the measurements due to other causes. Perhaps one group wasn’t

following the instructions or another group was using a unique method in their measuring. Variation due to these individual causes is called special cause. Both types of causes will affect the total variation that we see in the measurements.

[Topic D5: Relationships – an introduction](#)

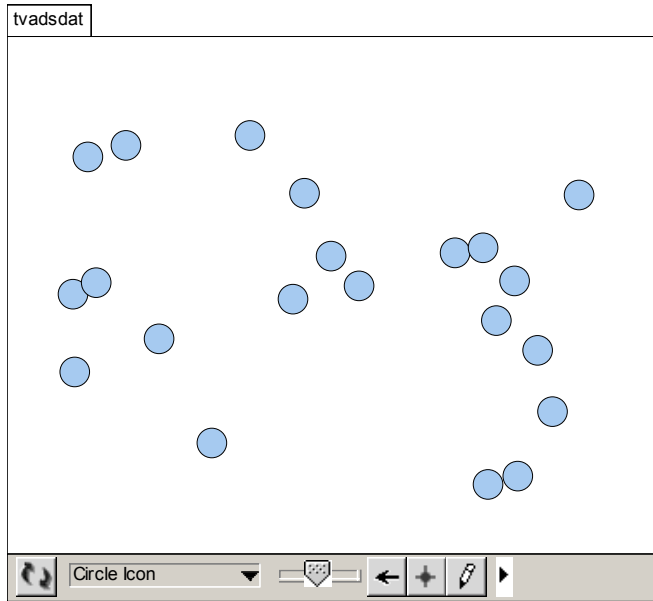
TECHNOLOGY ACTIVITY: USING *TINKERPLOTS* TO STUDY
RELATIONSHIPS

The point of this activity is for the students to select a dataset of interest among the variety of applications in the DASL library and then experiment with different *Tinkerplots* tools to graphically explore the relationship between two variables of interest. The objective is not to find the “best” tool for describing a relationship, but rather to illustrate the variety of ways that a person can use to relate two variables.

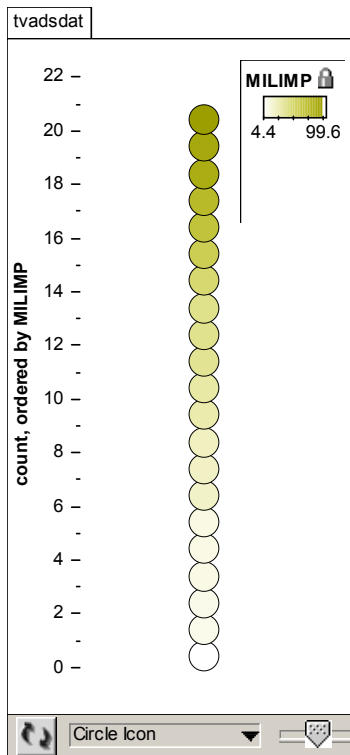
Here we illustrate different possible analyses with one of the DASL datafiles, although they certainly won’t exhaust the possible graphs that one can make.

The DASL story “TV Ad Yields” describes a survey of 4000 adults that was taken to measure the number of “retained impressions” of TV commercials that were recently broadcast. For a group of 21 companies, the dataset contains the company’s TV advertising budget (SPEND) and the estimated millions of retained impressions of the company’s commercials per week (MILIMP). A company might hope that there is a positive association between SPEND and MILIMP – if a company spends more money, it should hope to make a large impact to the TV viewers.

We import this datafile in *Tinkerplots* and start with a random scatter of 21 points corresponding to the 21 companies.

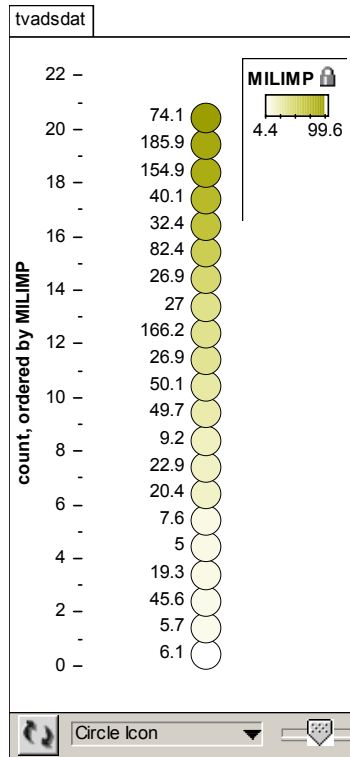


Since the company is primarily interested in the impact variable, we sort and stack the companies from the highest impact (top) to the lowest impact (bottom).

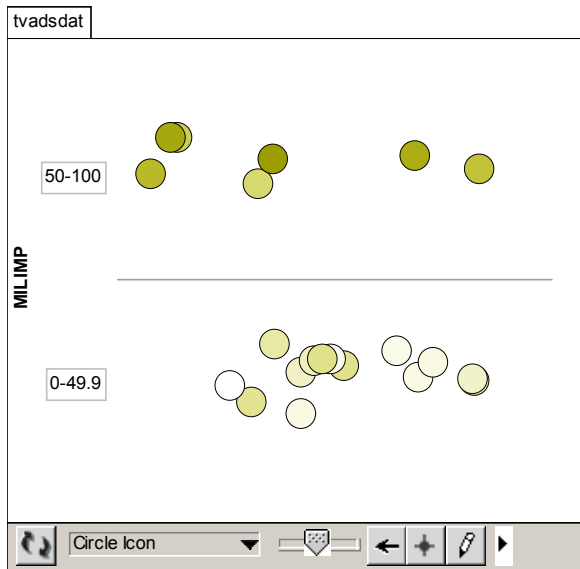


To see if there is any relationship between TV impact and spending, we label each dot with the corresponding spending. We see that some of the largest values of SPEND are

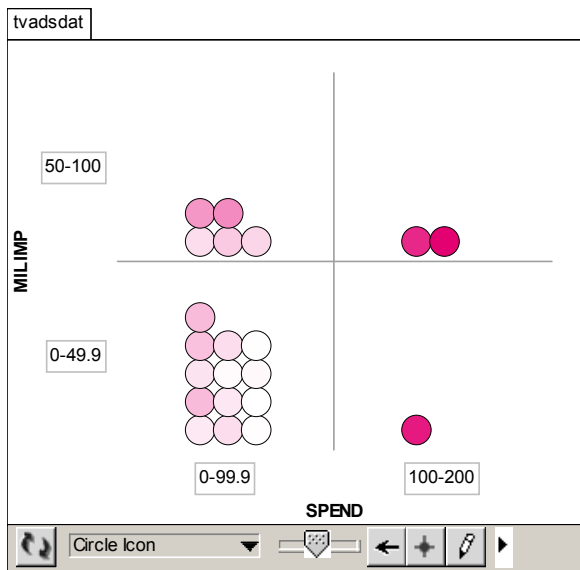
at the top of the graph, confirming that there is some relationship between the two variables.



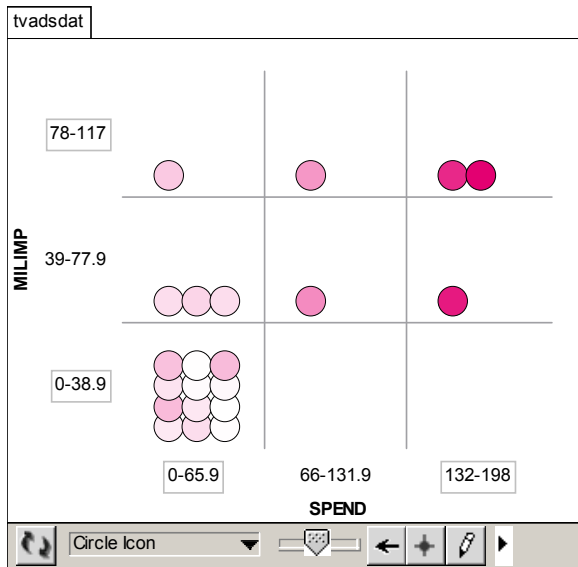
Another thing we can try is to divide the companies by impact – when we drag the variable to the vertical axis, *Tinkerplots* will divide the companies into two groups – the impact values between 50 and 100 on the top and the values between 0 and 49.9 on the bottom.



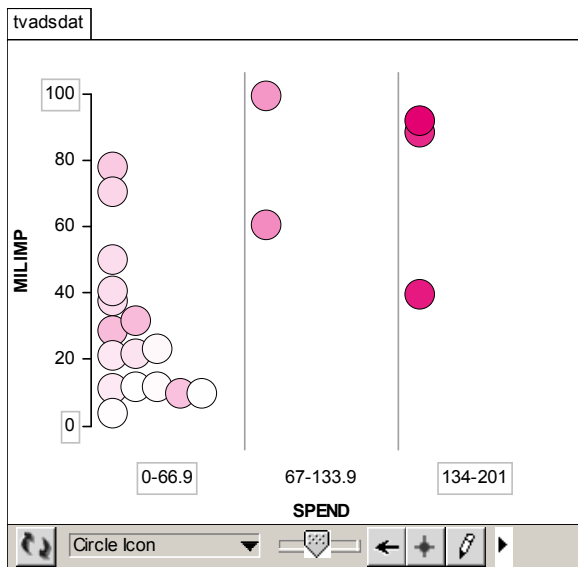
If we drag the spending variable to the horizontal axis, we get a two-way classification of the companies by spending and TV impact. We stack the dots in each cell so we can clearly see the number of companies in each cell.



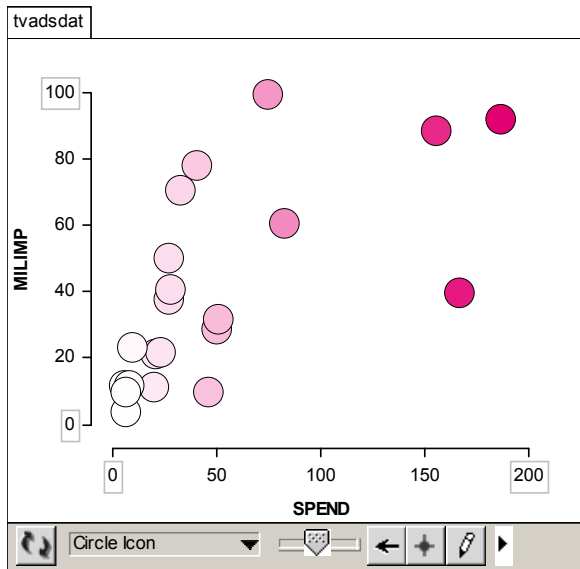
Of course, there is no reason why we need only two categories for each variable. Below we classify the two variables using three categories for each.



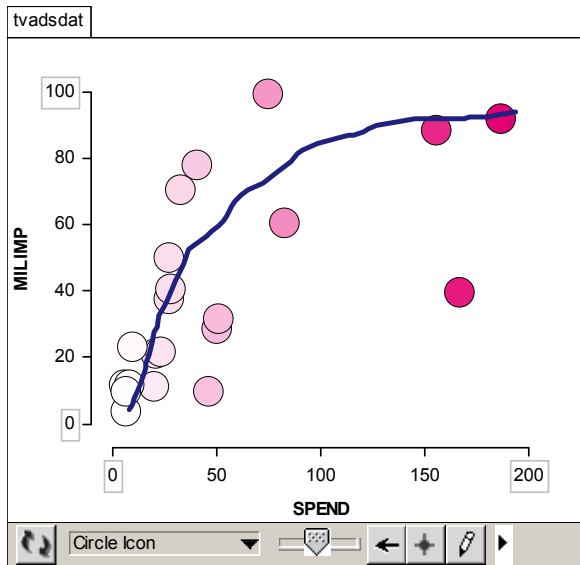
Alternatively, we can keep the impact variable as quantitative and look for differences in impact for three groups of spending.



If we treat both variables as quantitative, we get the familiar scatterplot display showing that there is a positive association between spending and impact.



Although there is a positive association, it isn't a straight-line relationship. Generally, as spending increases, impact increases until some point and it appears to level off. I indicate this by drawing a line on top of the scatterplot.



[Topic D6: Relationships – summarizing by a correlation and a line](#)

TECHNOLOGY ACTIVITY: GUESSING CORRELATIONS

This activity is a good way of introducing the concept of a correlation. Initially the student is told simply that a correlation is a measure of the association pattern in a scatterplot. Then the student is shown a scatterplot with randomly generated data and he/she is asked to make a guess at the correlation value. After the guess is made, the student finds out the actual correlation value and the guess and actual values are recorded in a table. This procedure is repeated 20 times – at the end of this part, the student should have a reasonable idea about the properties of the correlation coefficient.

In the second part of the activity, the student applies the correlation concept to the data that was just generated. The guesses and actual values are entered into Fathom, a scatterplot is made and the student has Fathom compute a correlation. If the student was making reasonable guesses at the correlation, then there should be high positive correlation between the guesses and the actual values.

ACTIVITY: FITTING A LINE BY EYE TO GALTON'S DATA

This is a nice activity for introducing the concept of fitting a best line. Using a piece of spaghetti (this works better than a limp piece of string or yarn), the student fits a line by eye to the points of the scatterplot. Then he or she has to find the slope and intercept of this line. In my experience, the students generally do fine in computing a slope, but some will struggle with the algebra in finding the intercept. This is a good opportunity to review the method of finding the equation of a line through two points.

At the end of the activity, the slopes and intercepts are collected from all students. When the collection of slopes and collection of intercepts are graphed, the students will see that there is much variability in the lines that are found by eye. This motivates the question: is there a way of determining the “best” line through the points? This motivates the use of the least-squares algorithm to compute a best line.

TECHNOLOGY ACTIVITY: FITTING A “BEST” LINE

This activity uses two Fathom documents to illustrate the idea of a least-squares fit. In Part I, the student is shown a scatterplot of the distances from Detroit to a number of cities and the corresponding airfares. By playing with a slider, he or she finds the

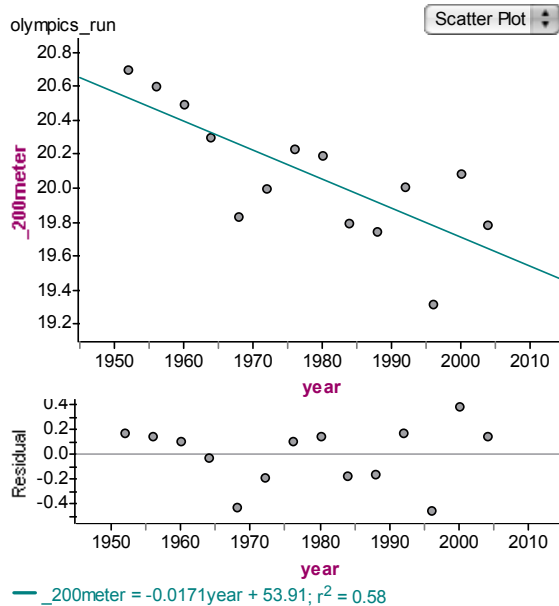
slope of a line that seems to best fit the points. A residual plot is shown below the scatterplot and can be used to select the best line. One wishes to choose a slope so that there is no tilt or up and down pattern in the residual plot. In the Fathom case table, there is an Attribute “residual” that has been defined to show the residuals for the student’s best line fit. In the lab, the student is asked to verify the computation of two residuals in the table and also use the best line to predict the airfare from Detroit to a new city.

In the second part of this activity, the student gets to experiment with the Moveable Line command in Fathom. By using the Show Squares command, he or she tries to find a line that makes the sum of squared residuals as small as possible. After the student writes down his/her final line and the sum of squares, then the least-squares line and the corresponding sum of squares is shown. If the student has chosen a good fitting line (from the least-squares criterion), then his/her sum of squares will be close (but certainly larger than) the sum of squares using the least-squares line.

TECHNOLOGY ACTIVITY: EXPLORING SOME OLYMPICS DATA

In this Fathom activity, the student will get experience interpreting a least-squares line and viewing a residual plot. He or she will choose one Olympic race and graph the winning times against the year. The least-squares line will tell us how much the winning time is decreasing each year and give us a means for predicting the winning team of this race for the next Olympics. For example, if we select the 200-meter run, the least-squares fit is $\text{TIME} = -0.017 \text{ YEAR} + 53.7$. This tells us that the winning time has decreased about .017 seconds per year, or .068 second each Olympic period of four years.

Actually, the statement that the winning race time is decreasing a certain number of seconds per year is not surprising. By looking at the residual plot, we can find particular winning times that are inconsistent with the general decreasing pattern.



Looking at the residual plot, unusually good performances are indicated by large negative values. It seems that the performances in 1968 and 1996 were unusually strong since the winning time was significantly lower than one would predict based on the least-squares fit.

ACTIVITY: REGRESSION TO THE MEAN

This activity is a demonstration of the regression effect that is not understood among the public. When you have two quantitative measurements that are positively correlated (in this example, scores on two tests for group of students), then there is a tendency for the second measurement to move back towards the mean.

An effective way of demonstrating this phenomenon is to compute the improvement from the first measurement to the second measurement. In this case, the improvement will be negatively associated with the first measurement. That is, students who do well on the first test will tend to have a negative improvement and students who do poorly on the first tend will tend to improve on the second test.

The students learn about this phenomenon by constructing graphs. In the first graph, they will see the two test grades are indeed positively correlated, and in the second graph, they'll see a negative relationship between improvement and the first test score.

This data was actually test scores from a recent class of mine. You can demonstrate this effect with many other types of data such as scores of golfers during two rounds of a tournament, statistics from athletes from two consecutive seasons, or any other pairs of measurements that are positively correlated.

[Topic D7: Relationships between categorical variables](#)

ACTIVITY: PREDICTABLE PAIRS

This is a fun activity to do in a class that contains a good number of women and men. First, find a movie that roughly half of the students have seen. Then think of a movie that is similar to the first movie. Ask students if they have seen movie 1 and movie 2, creating a two-way table. Based on computations based on this table, you are interested in seeing if there is an association between seeing the first movie and seeing the second movie.

Generally, you will find association if the movies are similar genres, such as romantic comedies, action films with some violence, etc. You could also try movies that are different genres to illustrate lack of association between two categorical variables.

After the students have performed the conditional calculations to look for association, it is important that they can express any association that they find. For example, it is not enough to say the vague statement that people who have watch “Sleepless in Seattle” have also seen “You Got Mail”. They should divide the class into two groups – those that have watched “Sleepless in Seattle” and those who have not watched “Sleepless in Seattle” and find the percentages of people in both groups that have watched “You Got Mail”. By expressing these conditional percentages, they have described the association.