Statistics

Mathematics Department
Phillips Exeter Academy
Exeter, NH
June 2016

Statistic Activity Book

Hershey Kisses

Before we begin:

In this lab, when new vocabulary is introduced, the word will be italicized, and the definition can be found in the glossary section of your Statistics Activity book. This lab is adapted from:

What is the Probability of a Kiss? (It's Not What You Think) Mary Richardson, Susan Haller, *Journal of Statistic Education Volume 10, Number 3* (2002).

www.amstat.org/publications/jse/v10n3/haller.html

Questions:

What is the chance that a HERSHEY'S KISS will land on its base when tossed out of a cup onto a table? What is the chance that it will not land on its base? Do we all agree that there are two possible *outcomes* when the kiss is tossed onto the table?

In this Activity:

Students work in groups of three collecting data, analyzing data and gaining experience with *empirical probability* and measures of the *center* of a *sample of data*.

Materials:

Pencils, ten plain HERSHEY'S KISSES candies, a 16-ounce plastic cup, a flat table or desktop, sticky notes and lab book.

Procedure:

1. Discuss with your group of three, the *subjective probabilities* you assign to the kiss landing on its base or on its side and record them in the table below:

HERSHEY'S KISS Tosses Subjective Probabilities

probability of base-landing		percent chance of base-landing	
probability of side-landing		percent chance of side-landing	
total	1		100%

2. Assign tasks within your group, Spiller, spills the 10 candies from the cup onto the table, Counter, counts the number of candies that land on their bases, and Recorder, records the results in their data table.

3. Let Spiller spill the cup ten times, counting and recording each time, and then switch rolls so that each person does each task once.

Toss #	HERSHEY'S KISS # on base
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
Total	

4. Refine your *subjective probabilities* based on the *empirical* evidence that you now have.

HERSHEY'S KISS Tosses Subjective Probabilities Refinements

probability of base-landing		percent chance of base-landing	
probability of side-landing		percent chance of side-landing	
total	1		100%

5. Combine the results from all of the groups on the board, using each groups totals out of 100. Together, consider the best way to display the data.

6. Now, ask your instructor to help you create a *stem and leaf plot* of the combined totals of each student in the class. In later labs, you will learn about more ways to display your data.

7. The visual data is very helpful, and you may want to adjust the subjective probability you have assigned to the a base-side landing. You may also want to get some measurement of the *center* of this data. Find the *mean* and *median* of each of the three sets of the data that your group collected.

HERSHEY'S KISS Statistics

	Data set 1	Data set 2	Data set 3
median			
mean			

8. Combine the results of your calculations on the board again, and discuss your collective findings.

9. You are probably ready to come to a consensus on the *empirical probability* that a HERSHEY'S KISS will land on its base if tossed. Please record this probability, and then consider the following vocabulary regarding this candy toss. The kiss had a _____ percent chance of landing on its base, and so in the *sample space* of *events* for this *variable* L, which stands for Landing, there are two possible values, B, for base or S, for side, and the probability of landing on its base, $P(B)$, is _____ .

Almond Hershey Kisses

Before we begin:

This lab is meant to be completed after the Hershey Kisses Lab. This lab is adapted from:

What is the Probability of a Kiss? (It's Not What You Think) Mary Richardson, Susan Haller, *Journal of Statistic Education Volume 10, Number 3* (2002).

www.amstat.org/publications/jse/v10n3/haller.html

Questions:

What is the chance that an almond HERSHEY'S KISS will land on its base when tossed out of a cup onto a table? What is the chance that it will not land on its base?

In this Activity:

Students work in groups of three collecting data, analyzing data and gaining more experience with *empirical probability* and measures of the *center* of a *sample of data*. They see first hand what it means to generalize results from one *population* to another. And they learn about *five-number summaries* (*minimum*, *first quartile*, median, *third quartile* and *maximum*) and *box plots*.

Materials:

Pencils, ten plain HERSHEY'S KISSES, ten almond HERSHEY'S KISSES candies, a 16-ounce plastic cup, a flat table or desktop and sticky notes.

Procedure:

1. Look at both types of kisses, noting their differences and similarities below:

2. Estimate the probability that the almond KISS will land on its base when spilled and record your estimate here:

Estimate	

3. As before, assign the jobs of tossing, counting and recording to different people in your group and then spill a cup with 10 almond candies and 10 plain candies onto the table 10 times recording the number of each that landed on its base.

Toss #	Almond KISS # on base	Plain KISS # on base
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
Total		

4. Rotate jobs and toss again. Rotate jobs and toss again. Each person in the group should have recorded 10 separate tosses for each type of candy.

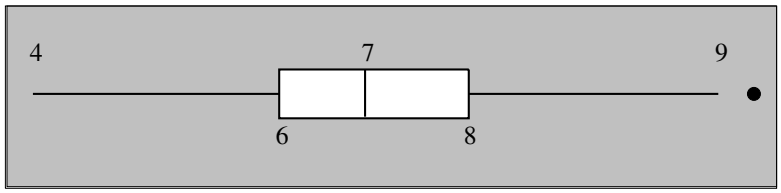
5. Record your group's data on sticky notes, one for each toss, using different colors for different candies and then collect all of the data from the class on the whiteboard. Try a *back-to-back stem plot* or two *frequency plots* with the same scale. Discuss your findings as a class.

6. Having numbers that capture what you see on the board can help you discuss the data and compare different data sets. Combine your group's data and compute the *median*, minimum and maximum of both the almond KISS numbers and the plain KISS numbers. Begin filling in the table below:

	Almond KISS	Plain KISS
minimum		
Q1		
median		
Q3		
maximum		
interquartile range		

7. Just as the *median* divides our data into a top half and a bottom half, we can additionally divide those halves into halves. $Q1$ is the "middle" value in the first half of the rank-ordered data set. $Q2$ is the median value in the set. $Q3$ is the "middle" value in the second half of the rank-ordered data set. Compute $Q1$ and $Q3$, making sure to find the mean of the two middle values if you have an even number of terms. The *interquartile range* is simply $Q3 - Q1$. Compute this as well, add all the numbers to the table above and then compare your findings with the other groups.

8. The statistics you computed above can be displayed graphically in a *box plot*. The box plot has three main features, a box whose width is the interquartile range, points graphed for the *outliers* (Something we will discuss in another lab.) and whiskers at the left and right of the box which extend to the minimum and maximum of the data. You can choose the height of your box. An example box plot is shown below:



Number out of 10 of base-landed Almond KISSES.

Construct your own box plots with two boxes, one that represents your plain KISS data and one that represents your almond KISS data.

9. Compare your findings with the class. Discuss the statistics you computed and the graphical representations of the data you created. Did they help you understand the sample space?

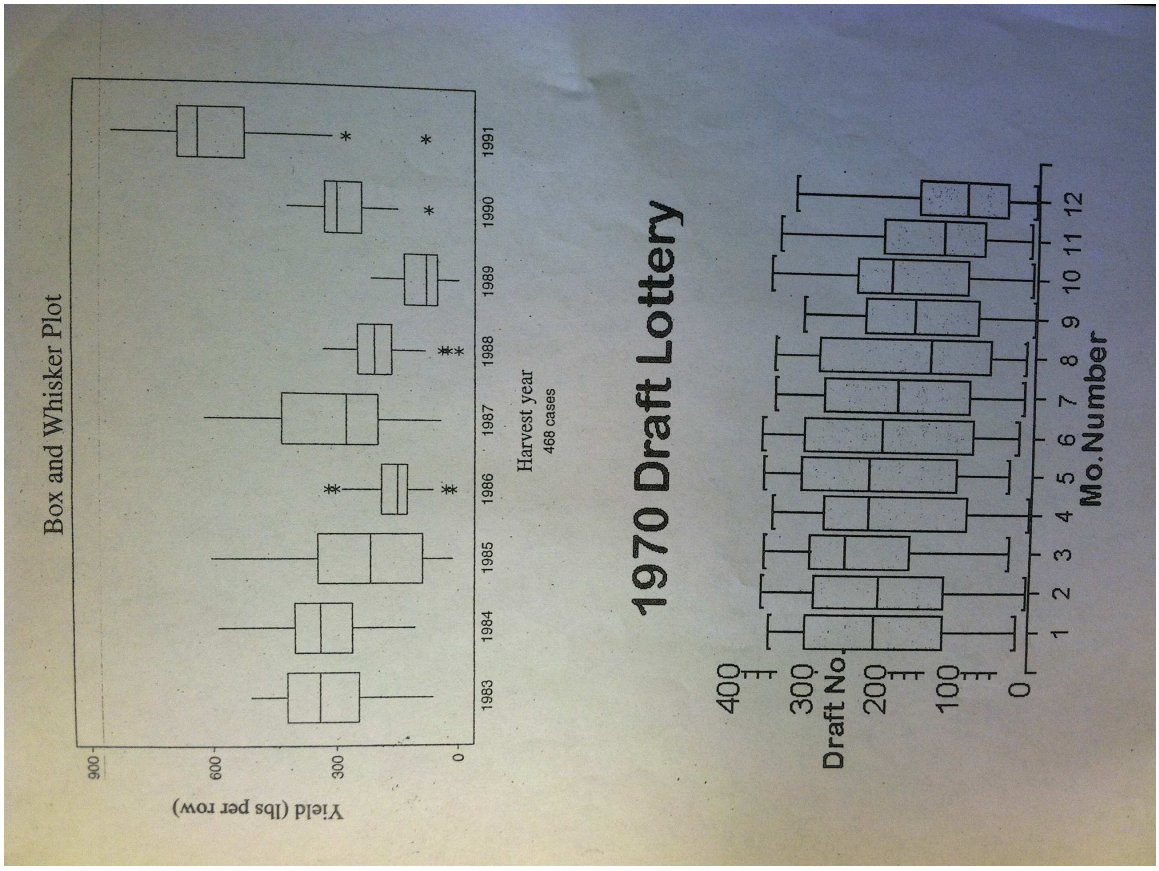
Every Graph Tells a Story

In this Activity:

We take a closer look at two stories behind two box and whisker plots.

Materials:

Graph paper and pencils.



Two Stories

Procedure:

1. The Grape Harvest The box-and whisker plot, with nine separate plots, describe the fruits of the grape harvest in the course of nine years at a small vineyard on the shores of Lake Erie. In this graph Yield (in pounds per row) is graphed against Harvest Year. The caption 468 cases means that there were 468 separate pieces of data, not 468 cases of wine. Make five observations based on this graph, and record them below:

2. **The Draft Lottery** The first lottery to select soldiers for the Vietnam War was held in 1970. The idea was to randomly match each of the 366 days in a (possibly leap) year with the integers 1 through 366. Eligible men whose birthday corresponded to 1, the first number picked, were the first to be drafted. The higher the number, the less likely you were to be drafted. To randomize the 366 possible birthdays, all the dates for January were put into small capsules, stirred vigorously, and poured into a large glass container. The capsules containing a birthday for each of the days in the subsequent months were added to the glass bowl in order, February followed by March, then April, etc. until December birthdays were added last.

Then one capsule was drawn at random by a person reaching into the glass and pulling out one capsule. This first capsule, September 14, was assigned draft number 001. The second date drawn, April 24, was assigned number 002. And so on through December, each date being matched with the order in which it was picked. This data is recorded in the *1970 Draft Lottery* box plot on the previous page. Each of the box plots represents one month.

Notice that the minimum for September looks to be very close to 1 and that Aprils minimum could well be 2.

Comment on this set of box plots which represents the outcome of an allegedly random process. Note especially the trend of the medians of each month as the year progresses.

Discuss your observations.

Student Scores

Questions:

Though they contain a great deal of information, are box plots enough?

In this Activity:

Students will learn about dot plots and compare them with box plots. They should break up into groups of 2 or 3 to work on this lab.

Materials:

You will need a pencil.

Procedure:

1. Consider the following hypothetical exam score data presented below for three classes of students.

Exam Score Data

A-period	50	50	50	63	70	70	70	71	71	72	72	79	91	91	92
B-period	50	54	59	63	65	68	69	71	73	74	76	79	83	88	92
C-period	50	61	62	63	63	64	66	71	77	77	77	79	80	80	92

2. Discuss this data with your group. Does it look like data that could have come from three of your classes? Which data represents the most successful class? Which class needs the most work?

3. By now, you are very familiar with box-plots. Fill in the 5-number summary tables for the classes below:

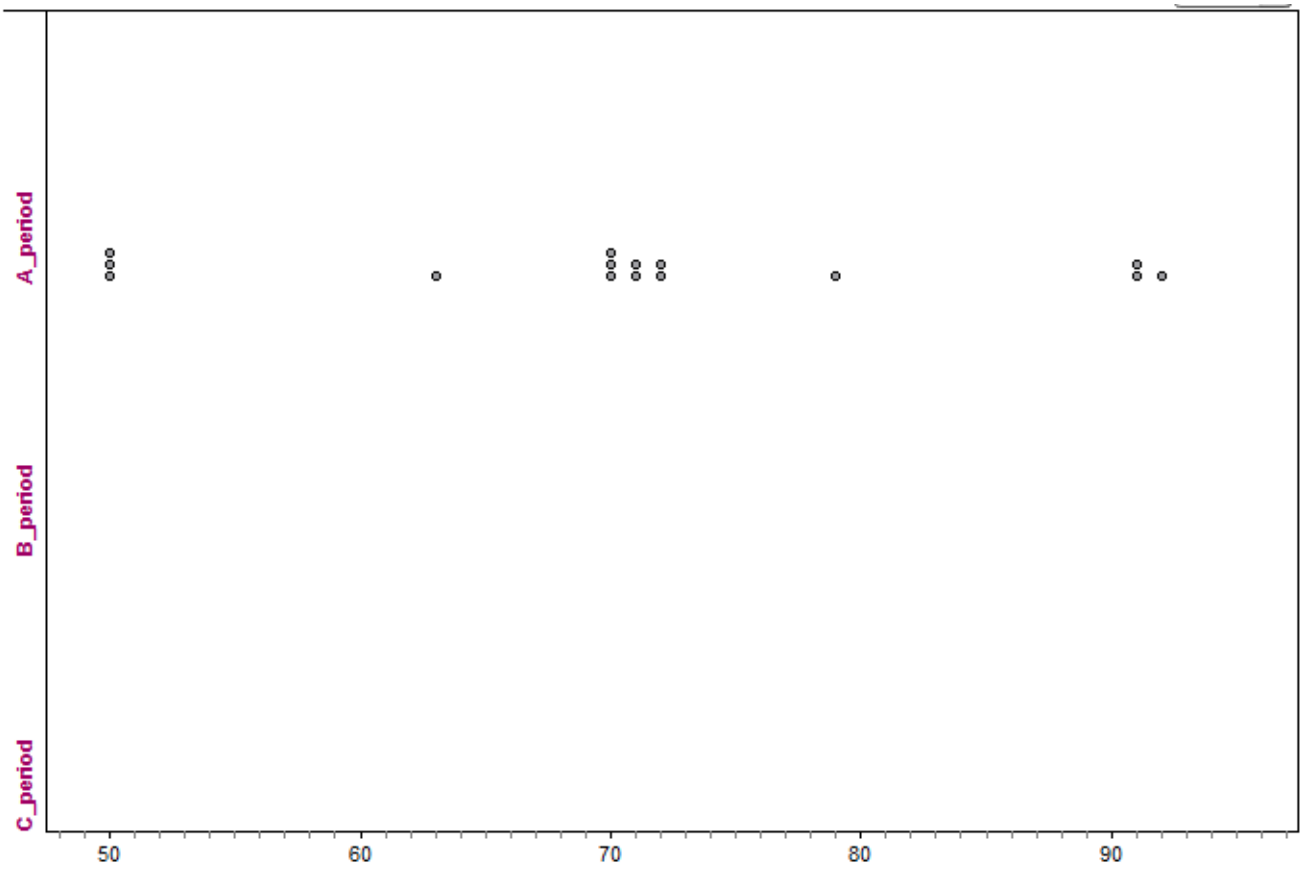
Test Scores Five-Number Summary

	A-Period	B-Period	C-Period
minimum			
Q1			
median			
Q3			
maximum			
interquartile range			

4. Consider the numbers in the table, and discuss with your group, whether this summary gives you any insight into your class scores.

5. Divide the task of creating box-plots for this data between the members of your group and compare results when you are all done. Do you have any new conclusions?

6. Another way to create a quick and helpful picture of your data is to create a *dot plot* of your data. A dot plot is simply a record of the frequency of each score. The variable that you care about is located on the horizontal axis, and the value of each data point is recorded as a dot located at its value. The dots accumulate vertically above the values. A dot plot for the A Period class is shown. Divide the task of creating dot plots for the remaining classes between the members of your group and compare results when you are done.



Scores on Tests.

7. Discuss as a class the differences between the three ways of displaying data, in a table, in a box plot and in a dot plot.

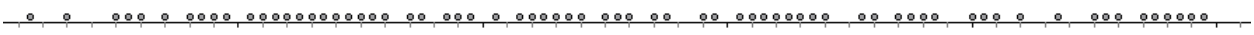
Matching Dotplots

Procedure:

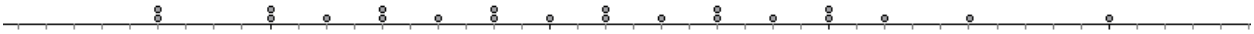
The following dotplots represent the distributions of the eight variables listed below. The scales of the plots have been omitted intentionally and the order of the plots has been scrambled. Your task is to match the variable with the plot. Provide a brief explanation of your reasoning in each case.

1. Jersey numbers from the 2014 New England Patriots
2. Annual snowfall amounts for a sample of U. S. cities
3. Margin of victory in Red Sox 2014 season games
4. Prices of properties in the Monopoly board game
5. Weights of the New England Patriots 2014 team members
6. Ages at which sample mothers had their first child
7. Weights of sample of 2014 cars
8. Scores on a Statistics exam

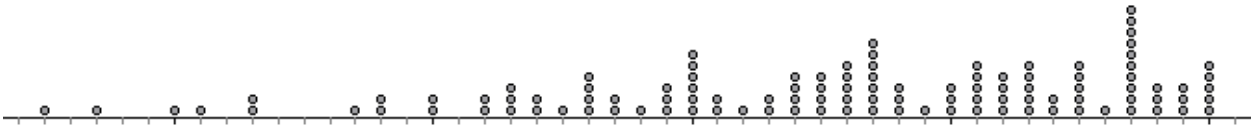
A



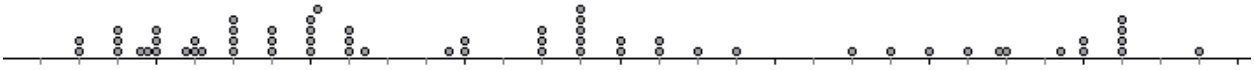
B



C



D

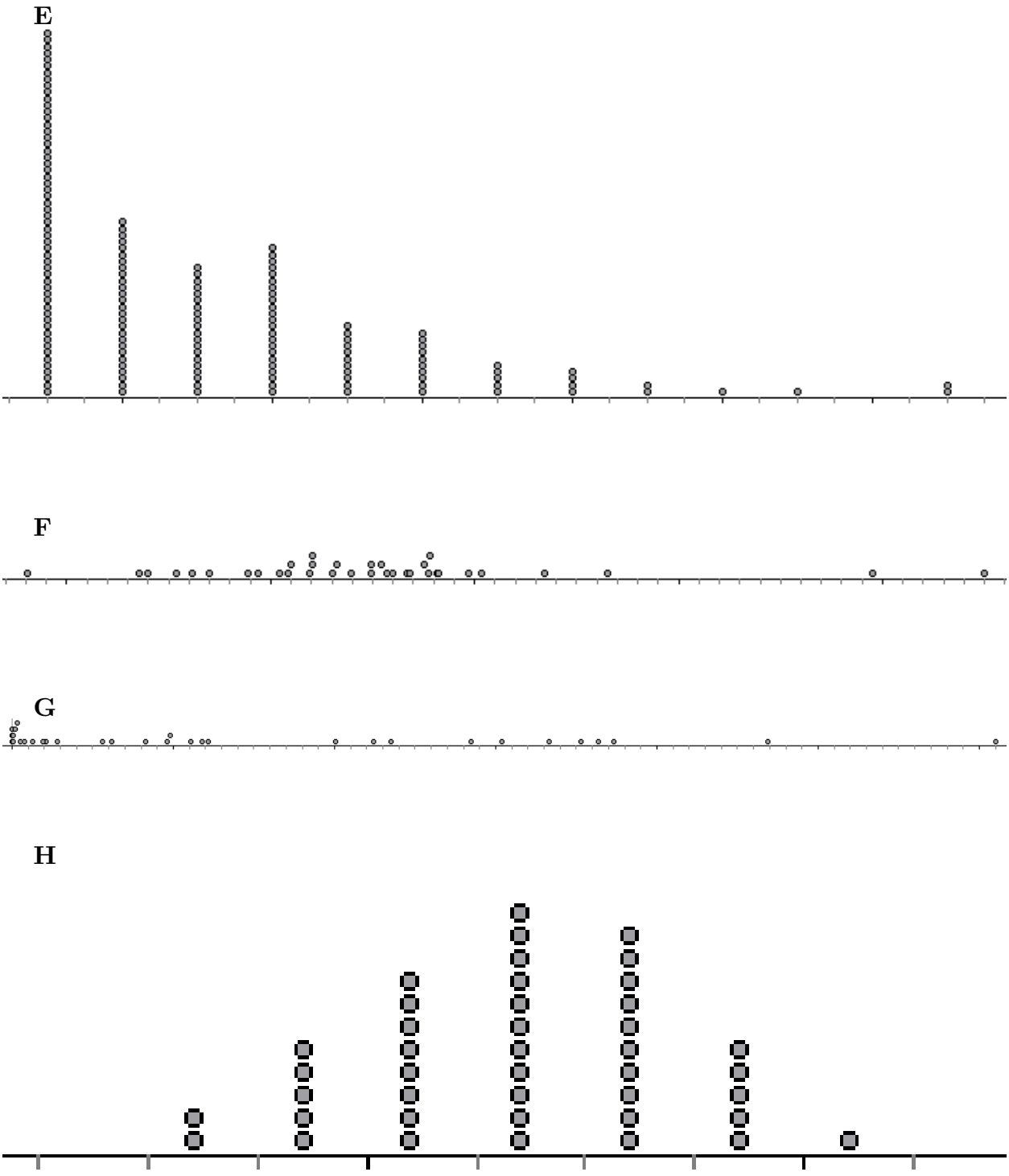


E

F

G

H



Puppies Lab

Questions:

We have learned about different ways to summarize and display your data and some of the standard vocabulary used when discussing data sets. The *shape* of the data is an important topic too.

In this Activity:

We use Labrador Retriever data to learn about *histograms* and the *shape* of a data set.

Materials:

Graph paper and pencils.

Procedure:

1. The table below shows the weight in ounces of puppies born at Meadowsweet Kennels in the last six months. Create a dot plot of the data and discuss the *shape* of the data with your group members. Look at each other's dot plots and discuss their differences and similarities.

Puppy Weight

13	14	15	15	16	16	16	16	17	17	17	17	17	17	17	18	18	18	18	18	18	18	19	20

2. A *histogram* is another way to graphically display univariate data, and this type of display can quickly communicate the shape of a data set. Whereas the primary consideration (once you have organized your data set from minimum to maximum) in creating a dot plot is how long to make the number line that contains the data, and the dot plots from your group were most likely very similar, a certain amount of design goes into creating a histogram. The summary table below was used to create the histogram shown on the next page.

Ounces	# of Puppies
13-14	2
15-16	6
17-18	15
19-20	2
total	25

Puppy Weights

3. Fill in the summary table below, and use it to create your own histogram.

Ounces	# of Puppies
13	
14	
15	
16	
17	
18	
19	
20	
total	25

4. Discuss the differences between the two histograms with your group and the class as a whole. Does one more effectively represent the data than another? Compare them with the additional histogram below.

Puppy Weights

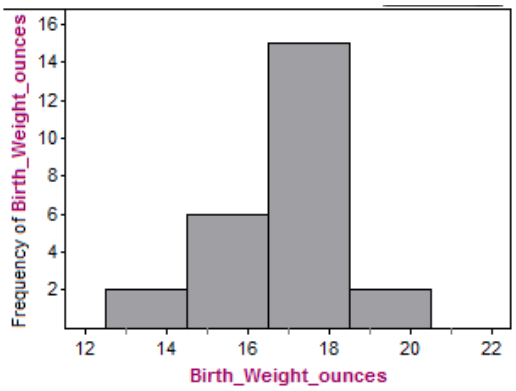
Meadowsweet Questions

In this Activity:

We take a closer look at mean, median and modal weights, first using the puppy data and then with other examples.

Materials:

Graph paper and pencils.



Puppy Weights

Procedure:

1. What are the mean, median and modal weights of the puppy data?

2. What are the range and interquartile range of the weights?

3. Graph these weights using a dot plot and a box plot and compare these graphs to the histogram above.

4. Describe the shape of the distribution of these weights, using words like skewed right or skewed left, symmetrical.

5. What is the typical weight of a Labrador Retriever puppy born in the last six months at the Meadowsweet Kennels? Explain why you chose this value.

6. Describe how variable the data are about the center.

7. The data on the Meadowsweet Labrador Retriever puppies sets are slightly skewed. The mean and the median are different from each other. Is there any relationship between the skewedness of the data and the relative size of the mean and median?

Additional Questions:

8. This example is adapted from *How to Lie With Statistics* , a classic book by Darrell Huff published in 1954. A company has 25 employees. The president earns $\$450,000$, the financial officer earns $\$150,000$,the two executives earn $\$100,000$, the bookkeeper earns $\$57,000$, three managers earn $\$50,000$, the four floor managers earn $\$37,000$, the timekeeper earns $\$30,000$ and the 12 lowly production workers earn $\$20,000$. The mean, median and modal salaries give very different answers to the question "What is the average (center) pay at this company?". Describe how these answers cast a very different light on the generosity of the company.

9. The distribution of salaries on a company is shown below. The median salary is $\$14,400$, and the mean salary is $\$17,800$.

Job	5 Executives	15 Supervisors	80 Production Workers
Salary Range	\$40,000 - \$87,500	\$ 15,800 - \$25,000	\$9,200 - \$15,700

a. If each of the supervisors is a given \$1,000 raise, how will this affect the mean, median, mode, range and interquartile range? Think for a moment before you begin calculations.

b. If each of the employees gets a \$1,000 raise how will this affect the mean, median, mode, range and interquartile range? Again, see if you can reason out the answer without calculations.

c. What is the total payroll for this company? In other words what is the sum of all the salaries? Once more, think for a moment before you compute the number.

10. In a boxplot, how much of the data lies inside the box?

11. (From *Workshop Statistics*) Using 10 integers from 0 to 100 (repeats allowed) construct three data sets as describe below, one set for a, one set for b, one set for c.

a. 90% of the data are above the mean.

b. The mean is greater than twice the mode.

c. The mean and median are different and none of the scores are between the mean and median.

12. (From *Workshop Statistics*) Are the following conclusions correct? Discuss with your neighbor.

a. A real estate agent notes that the mean housing price for an area is $225,700 and concludes that half of the houses in that area cost more than that.

b. A businesswoman calculates that the median cost of the five business trips she took last month is $750 and concludes that the total cost of the trips was $3,750.

c. A restaurant owner decides that more than half of her customers prefer chocolate ice cream because chocolate is the mode when customers are offered chocolate, vanilla and strawberry.

13. A previous president announced, truthfully, that the average net worth of an American family had risen 6%, to approximately $420,000. What he did not announce was that the median net worth was approximately $100,000, less than a quarter of this average. Comment on whether this was a misleading announcement.

Sudoku Experiment Part I

Before we begin:

Data can be gathered in many ways, in experiments, through surveys and through observational studies. This lab is explores an experiment, reinforces the use of dot plots and box plots to describe data and introduces measures of *center* and *spread*.

This lab was adapted from: Brophy and Hahn (2014) Engaging Students in a Large Lecture: An Experiment using Sudoku Puzzles. *Journal of Statistics Education* Volume 22, Number 1, www.amstat.org/publications/jse/v22n1/brophy.pdf.

Questions:

In statistics data can come from three places, *observational studies*, *experiments* and *surveys*. An *experiment* is a study in which some treatment is imposed on individuals in order to determine whether the treatment changes the outcome. How can a scientist create an experiment that will give meaningful results?

In this Activity:

Statistics class participants will have all EMI participants complete one of two 6 by 6 grid Sudoku puzzles and time how long it takes each person to complete it. Statistics class participants will collect the data. Then, using the class data, each participant will compile the data in various ways and test to see if they can draw any conclusions from their experiment.

Materials:

Enough Sudoku puzzles for all participants and a timer or clock.

Procedure:

1. Make sure that the pile of puzzles is shuffled. Make sure that participants can time themselves. The best case is to project a stopwatch on a screen so that all participants can see the same stopwatch.

2. Pass out the puzzle pages face down. Explain that each person will work to complete a different puzzle and that the puzzles vary in difficulty from easy to hard.

3. When everyone has a puzzle, tell them that when you say "begin", they should turn the paper over, read the directions, then do the puzzle. Say "begin", and start the stopwatch.

4. When finished, participants will turn their papers over and remain quiet until everyone is finished.

5. Collect and check the puzzles, recording the information on a chart on the next page.

Sudoku data

Type of puzzle	Correct	Time to finish	Experience

6. You may want to summarize your data in the following table:

Sudoku Data Summary

	Puzzle Type			Sudoku Experience		
Correct	Symbols	Numbers	Total	Yes	No	Total
NO						
YES						
Total						

7. Discuss with your group, why this activity is an experiment and not an observational study.

8. What were the 'treatments' in this experiment?

9. Find the five-number summary for each type. Record the data in the table below:

Sudoku Five-Number Summary

	Number Puzzle	Symbol Puzzle
minimum		
Q1		
median		
Q3		
maximum		
interquartile range		

10. Draw side-by-side box plots. Do you think that the differences that you see are significant?

11. There are two other numbers that are used to describe the *center* of a data set. One is the *mean* or the arithmetic average, and the other is the *mode* or the value that appears with the greatest frequency. In this case, because time is continuous it makes more sense to compute the mode if we round the time to finish to the nearest minute and compute the mode of those numbers. Compute the mean and the mode of this data and begin filling in the table below:

Sudoku Summary Statistics

	Number Puzzle	Symbol Puzzle
mean		
mode		
standard deviation		

12. The *standard deviation* is, roughly, the average deviation of each data point from the mean. Except for very small data sets, it is cumbersome to compute. Every calculator and statistics software package will compute it very quickly. The number is used to describe the variability or spread of your data. Use your calculator to compute the standard deviation for the time to complete each type of sudoku puzzle, and add this number to the table of summary statistics.

13. As you gain more experience with mean and standard deviations, you will see how these two numbers can provide a great description of a data set, giving a snapshot of both center and variability. Discuss with your group and then with the class as a whole, which information you find most helpful in representing the data, the 5-number summary or the mean and standard deviation?

Statistics Activity Book

Sudoku Experiment Part I

Instructions:

On the other side of this sheet is a six by six grid of squares broken up into six outlined boxes with Greek letters placed in a handful of the thirty-six squares. The Greek letters α, β, δ, ε, λ and μ must each appear once in each of the six outlined boxes, once in each of the six rows and once in each of the six columns. Use logic (i.e. do not guess) to determine what goes in each empty space.

Before you turn the page over and begin, please answer the following question:

Have you ever played Sudoku before today? Yes ☐ No ☐

Sudoku Experiment

Instructions:

The Greek letters α, β, δ, ε, λ and μ must each appear once in each of the six boxes, once in each of the six rows and once in each of the six columns. Use logic (i.e. do not guess) to determine what goes in each empty space.

β			μ		ε
		ε	λ		
λ	β	μ	α	ε	
	δ	α	β	λ	μ
		λ	ε		
α		β			λ

Time to completion: Minutes: _____ Seconds: _____

Sudoku Experiment Part I

Instructions:

On the other side of this sheet is a six by six grid of squares broken up into six outlined boxes with lowercase letters placed in a handful of the thirty-six squares. The lowercase letters a, b, c, d, e and f must each appear once in each of the six outlined boxes, once in each of the six rows and once in each of the six columns. Use logic (i.e. do not guess) to determine what goes in each empty space.

Before you turn the page over and begin, please answer the following question:

Have you ever played Sudoku before today? Yes ☐ No ☐

Statistics Activity Book

Sudoku Experiment

Instructions:

The lowercase letters a, b, c, d, e and f must each appear once in each of the six boxes, once in each of the six rows and once in each of the six columns. Use logic (i.e. do not guess) to determine what goes in each empty space.

b			f		d
		d	e		
e	b	f	a	d	
	c	a	b	e	f
		e	d		
a		b			e

Time to completion: Minutes: _____ Seconds: _____

Statistics Activity Book

Sudoku Experiment Part I

Instructions:

On the other side of this sheet is a six by six grid of squares broken up into six outlined boxes with numbers placed in a handful of the thirty-six squares. The numbers **1**, **2**, **3**, **4**, **5** and **6** must each appear once in each of the six outlined boxes, once in each of the six rows and once in each of the six columns. Use logic (i.e. do not guess) to determine what goes in each empty space.

Before you turn the page over and begin, please answer the following question:

Have you ever played Sudoku before today? Yes ☐ No ☐

Sudoku Experiment

Instructions:

The numbers **1**, **2**, **3**, **4**, **5** and **6** must each appear once in each of the six boxes, once in each of the six rows and once in each of the six columns. Use logic (i.e. do not guess) to determine what goes in each empty space.

2			**6**		**4**
		4	**5**		
5	**2**	**6**	**1**	**4**	
	3	**1**	**2**	**5**	**6**
		5	**4**		
1		**2**			**5**

Time to completion: Minutes: _____ Seconds: _____

Sudoku Experiment Part I

Instructions:

On the other side of this sheet is a six by six grid of squares broken up into six outlined boxes with symbols placed in a handful of the thirty-six squares. The symbols ■, △, √, ⟝, ⊖ and ♡ must each appear once in each of the six outlined boxes, once in each of the six rows and once in each of the six columns. Use logic (i.e. do not guess) to determine what goes in each empty space.

Before you turn the page over and begin, please answer the following question:

Have you ever played Sudoku before today? Yes ☐ No ☐

Sudoku Experiment

Instructions:

The symbols ■, △, √, ⟝, ⊖ and ♡ must each appear once in each of the six boxes, once in each of the six rows and once in each of the six columns. Use logic (i.e. do not guess) to determine what goes in each empty space.

△			♡		⟝
		⟝	⊖		
⊖	△	♡	■	⟝	
	√	■	△	⊖	♡
		⊖	⟝		
■		△			⊖

Time to completion: Minutes: _____ Seconds: _____

The Standard Deviation

Before we begin:

The standard deviation is a powerful and very commonly used measure of the spread or variability of a data set. It has features in common with the Mean Absolute Deviation. One feature it does not share, however, is ease of computation.

In this Activity:

This lab is designed illustrate how to find the standard deviation using a very small and arbitrary data set.

Procedure:

1. Consider the data 1, 2, 4, 6 and 9. Calculate the mean of this set. The mean is denoted $\bar{x}$ and is read "x bar".

$$\bar{x} = \_\_\_\_\_$$

2. Use the expressions given in the column headings to complete the blanks in the table below.

Score, x	Mean $\bar{x}$	Deviation from the Mean $(x - \bar{x})$	Squared Deviation $(x - \bar{x})^2$
1			
2			
4			
6			
9			

3. Next compute: S = Sum of the Deviations from the Mean and SS = Sum of the Squared Deviations from the Mean, and record them below.

$$S = \_\_\_\_\_ \text{ and } SS = \_\_\_\_\_$$

Discuss these values with your group. Are you surprised by the results?

4. This data set contains 5 points, and so the next step is to compute the following value:

$$\frac{SS}{4} = \_\_\_\_\_$$

The denominator is one less than the number of data points. This deserves an explanation, but it is a long one, and better left for another time. How do the units of the numbers you just computed compare to the units of the original data?

5. Take the square root of your answer to the previous questions, and record the number below:

$$\sqrt{\frac{SS}{4}} = \_\_\_\_\_$$

Finally! This number represents the standard deviation. The square root is necessary because if the original units are, say, pounds or dollars then without the square root, the units would be square pounds or square dollars. This value is often denoted Sx and is the standard deviation of this particular data set. Can you retrace your steps and write formulas for Sx?

$$Sx =$$

6. Now enter this small data set into your calculator to verify your answer. Compare your answers with the ones your calculator produced.

7. Without doing any calculations how would the standard deviation of a dataset change if you added 10 to each value? Explain your answer.

8. Without doing any calculations how would the standard deviation of a dataset change if you multiplied each value by 10? Explain your answer.

The Normal Distribution Lab

Before we begin:

The *normal distribution* is probably the most important distribution in all of statistics because it appears so frequently when examining univariate data. If you graph the heights of a group of 100 female basketball players, the IQs of men aged 30 to 40, the birth weights of a large group of babies, the lengths of cod caught in the North Sea, you will find that the numbers very closely approximate a normal curve.

The normal curve accords with common sense. In the context of, say, heights of human females aged 20 to 25, there are very few extremely high or low heights. As we examine heights closer to the mean, there are more females with these heights.

Many completely unrelated data sets exhibit an approximately normal distribution.

Questions:

What is *normal*?

In this Activity:

We explore the features of a normal distribution.

Materials:

Graph paper, pencils and a graphing calculator or statistical software package.

Procedure:

1. Bradford College administers a placement test to incoming freshmen to determine their appropriate math placement. One year 216 freshmen take the test, which consists of 20 multiple choice questions.The results below display the possible score on the test, that is, the total number of correctly answered questions, and the count that represents the number of students who received each score. Make a histogram of the data using 20 boxes. Compare your results with your classmates.

Frequency of Scores

Score	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Count	1	1	5	7	12	13	16	15	17	25	17	21	12	16	18	4	7	5	4	0

2. Notice that your data is symmetrical and mounded. This score data approximates a *normal*, or bell-shaped, curve. In the context of the placement test we see that relatively very few students got very low or very high scores. Compute the mean and standard deviations of the scores, and record them here:

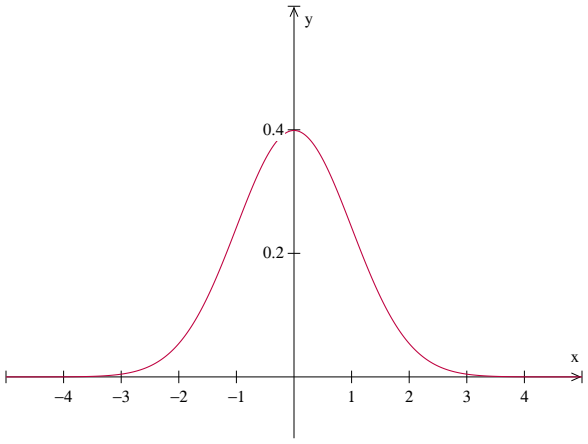
mean $=$ _____ standard deviation $=$ _____

How many of the scores are within one standard deviation of the mean?

3. What proportion of students' scores are within two standard deviations of the mean? What proportion is this?

4. How many students scored within three standard deviations of the mean? What proportion is this?

5. Your answers to the previous questions should have given you what is sometimes called the *empirical rule for normal distributions*. In data that are approximately normally distributed 68% lie within one standard deviation of the mean, 95% lie within two standard deviations of the mean and practically all of the data (99.7% in theory) lie within three standard deviations of the mean. This tendency is sometimes called the 68-95-99.7 rule. Have you ever heard of it? Do the Bradford College test scores follow this rule? This, like the Pythagorean Theorem, is something to memorize. Below is a sketch of a normal distribution. What is the mean of the distribution shown?



The Normal Distribution: $y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

6. Knowing that 99.7% of the data lies within three standard deviations of the mean, what do you guess is the standard deviation of this distribution?

7. The normal distribution in which the mean is 0 and the standard deviation is 1 is called the *standard normal distribution*. Estimate the area between the curve and the x-axis, and deduce another property of a standard normal distribution. Does the standard normal distribution accurately model the data given by Bradford College? (If not, how could you transform the given standard normal distribution so that it would?)

8. Statisticians have a name for the number of standard deviations from the mean: a z-score. A point with a z-score of 1.5 means that that value lies 1.5 standard deviations above the mean. A z-score of -0.6 means that the point lies 0.6 standard deviations below the mean. Z-scores are very useful for comparing normal distributions with different means and standard deviations. What is the z-score of a score of 4 on the Bradford College test?

9. Sophie and Pascal are applying to college. Sophie takes the SAT and Pascal takes the ACT. The scores from both of these tests are both normally distributed. The SAT has a mean of 896 and a standard deviation of 174. The ACT has a mean of 20.6 and a standard deviation of 5.2. Using the normal distribution above and the mean and standard deviation information, sketch a normal curve that would represent the scores from each of the two tests.

10. (Continued) Even without the picture, you can use the mean and standard deviation information to compare Sophie's score to Pascal's score. Sophie scores 1080 on the SAT and Pascal scores 28 on the ACT.

a. Sophies score is how many standard deviations above the mean?

b. Pascals score is how many standard deviations above the mean?

c. Which score has a higher z-score?

d. Which person do you think performed better on their respective tests?

e. Mark Pascal's and Sophie's z-scores on their respective graphs and discuss your findings.

11. By now, you may have already decided that the formula for computer a z-score of a point with value x in an approximately normal distribution is:

$$z = \frac{x - \text{mean}}{\text{StdDev}}.$$

What are the units of a z-score?

12. What would a z-score of 0 tell you about the value of a point?

13. What would a z-score of 4.2 tell you about the value of a point?

14. Using the normal curve that you drew in problem 9, decide what percentage of students who took the SAT test had a lower score than Sophie? Notice the connection between area under the curve to the left of Sophie's score and your answer.

15. Again, using the curve you drew in problem 9, decide whether or not Pascal was successful in his quest to have a score that was better than the scores of 90% of the people taking the ACT.

Minimum Wage Lab

Questions:

A scatter plot of a data set gives, as they say, a thousand words of information about *bivariate* (two-variable) data. It is very helpful to have a common vocabulary to discuss those scatter plots.

In this Activity:

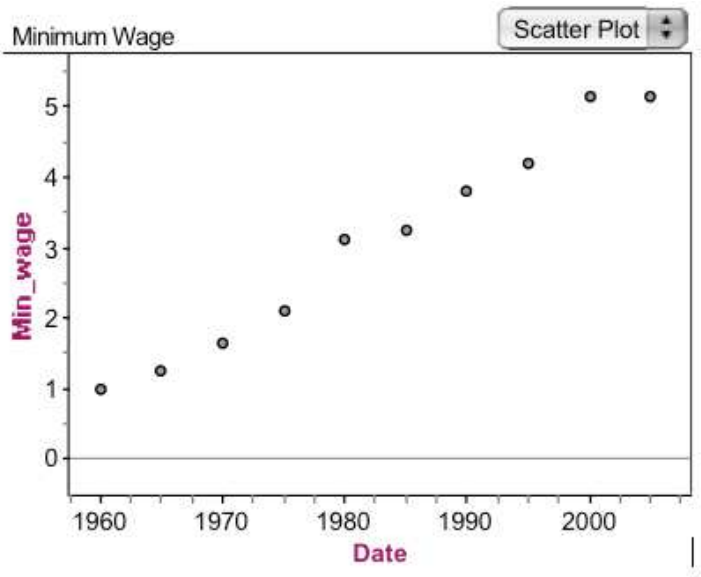
We introduce the vocabulary used to discuss scatter plots and look at many examples.

Materials:

Ruler and pencils

Procedure:

1. As a group, consider the *scatterplot* that shows the federal minimum wage at five-year increments. This is a graph of minimum wage *versus* time.



Minimum Wage Data

2. Here are some questions you might ask. What exactly does each data point represent? What shape are the data, linear, curved, clusters? Are there outliers? Do the data show a trend, positive, negative or none? How strong is the pattern, strong, weak, moderate, constant or varying? Does the pattern generalize? Is there an explanation for the pattern?

3. In your group, pick out the adjectives that best describe the minimum wage data.

4. Use a ruler to draw a line that seems to you to fit the data as well as possible. Compare with your group members.

5. Estimate the slope of the line, and then use coordinates of points on or very close to the line to compute the equation of this line. Let x stand for the number of years after 1960, and write the equation below, using $y = mx + b$ form:

6. Interpret the value of the slope m and the y-intercept b in the context of this story. Could you safely use the line to estimate the minimum wage in the current year? (What *is* the actual federal minimum wage is currently?)

7. Discuss your findings as a class.

8. Now consider the scatter plots shown on the next page and working with your group, fill in the chart given above the scatterplots.

The data on 1999 cars reported by *Consumer Reports* included not just sports cars but also classifications of small, family, large, luxury, and upscale. The following nine scatterplots display pairs of variables for these cars. The variables are:

city MPG rating	highway MPG rating
weight	% front weight
time to accelerate from 0 to 60 miles per hour	time to cover $\frac{1}{4}$ mile
page number on which the car appeared	fuel capacity

(a) Evaluate the direction and strength of the association between the variables in each graph. Do this by arranging the associations revealed in the scatterplots from those that reveal the most strongly positive association to those that reveal virtually no association to those that reveal the most strongly negative association. Arrange them by letter in the table below. (Since you are to use each letter only once, you should probably look through all nine plots first.)

letter:	strongly negative		mildly negative		virtually none		mildly positive		strongly positive

Least Squares Regression Lab

Before we begin: This is a short worksheet on the Least Squares Regression line. The goal is to summarize and condense some of the ideas about linear regression that appear in a few workshops in this book.

In this Activity: You will explore residuals with a very small data set.

Materials: You will need graph paper, a ruler and a pencil.

Procedure:

1. Graph the three points $(0, 0)$, $(2, 3)$ and $(4, 3)$, and find the centroid of the triangle that they create. In a data set of any size, this point is call the *point of averages* and denoted $(\bar{x}, \bar{y})$.

2. Add to your graph the three lines $y_1 = 2$, $y_2 = -x + 4$ and $y_3 = \frac{3}{4}x + \frac{1}{2}$. Notice that all three lines pass through $(\bar{x}, \bar{y})$.

3. The *residual* of a point $P = (x_p, y_p)$ with respect to a line $y = f(x)$ is the vertical distance between the y-value of the point and the y-value of the point of the line given by the x-value of P. It looks simpler than it sounds:

$$\text{Residual of } P \text{ with respect to } y = y_p - f(x_p).$$

The residual of $(2, 3)$ with respect to y_2 is $3 - (-2 + 4) = 1$. Statisticians describe this new data as fit. While all three lines pass through $(\bar{x}, \bar{y})$, the line that passes closest to all of the points is the one that we choose to represent our data. Fill in the table below with the three residual values for each line.

Residuals

	y_1	y_2	y_3
$(0, 0)$			
$(2, 3)$			
$(4, 3)$		3	
Sums of residuals			

2. Discuss your findings. Did you learn anything new about the three lines relative to the data?

3. The squares of the residuals give a clearer picture of the 'fitness' of a line. Fill in the table this time with the squares of the residuals.

Squares of Residuals

	y_1	y_2	y_3
$(0, 0)$			
$(2, 3)$			
$(4, 3)$		9	
Sums of squared residuals			

4. The *Least Squares Regression Line* is the line that makes the sum of the squares of the residuals as small as possible. The line will always pass through the point of averages. When, if ever, will the sum of the squares of the residuals equal zero?

Mammal Lab

Before we begin:

A table of data containing the number of days in a gestation period and the life expectancy for different mammals is located on the back page of this lab.

Questions:

How can we summarize our data with more than numbers or a description of its shape?

In this Activity:

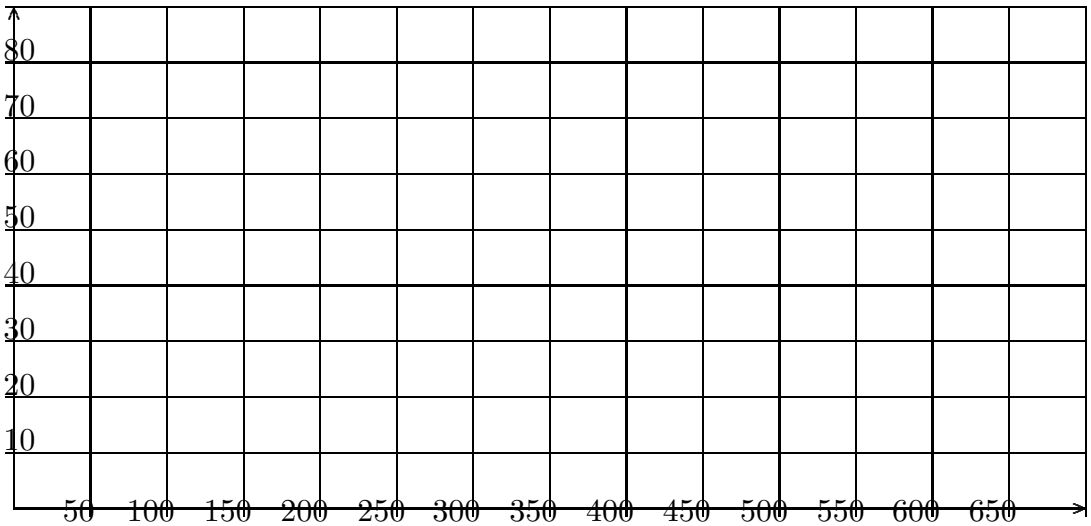
You will learn how to model your data with a line.

Materials:

Graph paper and a graphing device.

Procedure:

1. Enter the data from the Mammal Data table found on the back page of this lab into your graphing device, and create a scatter plot of your data. Add labels to the axes below and sketch your data on the grid.



2. Draw a line on your scatter plot that you think best summarizes your data. Estimate the slope and y-intercept and fill in below:

$$y = \_\_\_\_\_\, x + \_\_\_\_\_.$$

In this case, y represents the predicted life expectancy, and x represents a given gestation period. Interpret the value of the slope and the y-intercept in the context of this data.

3. Compare your summary line to the lines drawn by others in your class. Are they the same? Same slope? Same y-intercept?

4. Does your line go through all of the points on your scatter plot? Does it go through any points?

5. A residual is the error of the regression line. That is, it is the difference between the observed y value or height of the data point and the predicted y value or height on the summary line.

6. For each point on the scatter plot draw a vertical line from your data point to the point on the summary line that shares an x-value with your data point. The length of each vertical line that you drew represents the absolute value of the residual of each data point with respect to the summary line.

7. Draw squares using each residual as one side of the square. The area of each square represents the value of the squared residual. The sum of all of the areas of the squares represents the total sum of the squared residuals. Estimate the sum of the squares of your residuals:

$$\text{sum} = \_\_\_\_\_.$$

Compare your squares with the whole class. Which line produced the smallest sum of squares.

8. The *Least Squares Regression Line* is the line that produces the minimum sum of squared residuals. Use your calculator's LinReg feature to find the slope and intercept of the Least Squares Regression Line and fill them in below:

$$y = \_\_\_\_\_\, x + \_\_\_\_\_.$$

In this case, y represents the life expectancy predicted by the least squares regression line, and x represents a given gestation period.

In addition, record the value that your calculator gives you for the variable r:

$$r = \_\_\_\_\_.$$

The significance of r will be discussed in the next lab.

9. Add the line given by LinReg to your scatter plot (You can ask your calculator to do this automatically.) and compare with the lines and add it to the scatter plot on your calculator. Compare with the lines you and your classmates drew.

10. Discuss this new line as a good predictor of life expectance given gestation period. Would you accept this line as a good model for your data?

11. Homework. The three data points corresponding to the human, the hippo and the elephant are far away from the bulk of the data, and small changes in their positions will have a disproportionate effect on the equation of the least squares regression line. Statisticians call these points *influential points*. Remove these three points from your data set, and as you did in problem **2)** estimate what you think is the line that best models this data. Next have your calculator find the exact least squares regression line. (Make sure to note the value of r for this new line.) Compare your new equation to the one you obtained in class with the three influential points still in the data set.

Mammal Data

Name	Gestation Period in Days	Life Expectancy in Years
Beaver	105	5
Cat	63	12
Cow	284	15
Deer	201	8
Elephant	660	35
Fox	52	7
Gorilla	258	20
Hippopotamus	238	41
Human	266	80
Horse	330	20
Moose	240	12
Mouse	21	3
Opossum	13	1
Rabbit	31	5
Wolf	63	5

Source: *World Almanac and Book of Facts* 2001, p.237

Scrabble Letter Lab

Before we begin:

This lab refers to the scores that each letter is assigned in a standard american-english scrabble game. A table of these values is given on the back page of this lab.

Questions:

Given a data set, we can choose a line that best matches the data in many ways. What qualities would you like the line to have?

In this Activity:

You will make a scatter plot of the data, choose a line that might best match the data and also find the least squares regression line.

Materials:

You need paper, pencils, a graphing device and the table of letter scores in the text.

Procedure:

1. Enter the letter point data from the Scrabble Letter Value table into your calculator in two columns of data and create a scatter plot of your data with tiles on the x-axis and points on the y-axis. In addition, plot the line $y = -(1/3)x + 3.5$ on your graph. The plotted line has rational slopes and intercepts. How well does it match your data?

2. Add a third column of data to your table that consists of the residuals of each data point with respect to this line. Using summation notation, write an expression that you could use to compute the sum of all of these residuals.

3. It is easier to compare residuals if you consider the squares of the residuals rather than the signed residual or the absolute value of the residual. (This also favors a few small residuals rather than one large residual.) Add a new column to your data table that consists of the squared residual values. Using summation notation, write an expression that you could use to compute the sum of all of these squared residuals.

4. Your calculator can find the sum of the elements in a column of data. Compute the the sums of both the signed residuals and the squared residuals with respect to the given line.

5. The LinReg feature on your calculator finds the line that minimizes this sum of squared residuals. This line, that minimizes the sum of the squared residuals is called *the best fit line*. Use the LinReg feature on your calculator to compute the best fit line for this data. Graph the data, the line given earlier and the best fit line. Discuss your findings with your group.

6. Compute the residuals with respect to the best fit line and the squared residuals with respect to the best fit line, and discuss with the class how to compare those values with those of the original line.

Scrabble Letter Value

Letter	# of tiles	# of points
A	9	1
B	2	3
C	2	3
D	4	2
E	12	1
F	2	4
G	3	2
H	2	4
I	9	1
J	1	8
K	1	5
L	4	1
M	2	3
N	6	1
O	8	1
P	2	3
Q	1	10
R	6	1
S	4	1
T	6	1
U	4	1
V	2	4
W	2	4
X	1	8
Y	2	4
Z	1	10

Correlation Lab

Before we begin:

Statisticians summarize the characteristics of a set of data in an important number called r, the *correlation coefficient*. This lab is meant to introduce this number and serve as reference in the future.

Questions:

Wouldn't it be nice if we could quantify the notions of *strength*, *positive association* and the other words that we use to describe a data set?

In this Activity:

This lab involves reading about r and then playing a guessing game.

Materials:

Pencils.

Procedure:

1. r is a statistic that measures the direction and strength of a linear relationship. Data that is perfectly linear and has a positive slope has a measure of $r = 1$. Data that is perfectly linear but has a negative slope has a measure of $r = -1$. Data that has no discernable pattern has a correlation value of $r = 0$. Thus, $-1 \leq r \leq 1$. Have you seen other important mathematical variables that can take on this range of values?

2. There are two ways to calculate the value of r. Both use the idea that r captures the variation in the x direction, as well as the variation in the y direction. One can look at the standardized scores of each data point and think of r as the average product of these two standardized scores. Here is one version:

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right).$$

Can you guess the definitions for $\bar{x}$, $\bar{y}$, s_x and s_y? If you graph the *standardized data* and fit a least squares line to them then the value of r will be exactly the slope of the line fitted to the transformed data. Discuss with your group.

3. One can also think of r as a measure of the variability in the *response variable*, y, that is explained by the variability of the *independent variable*, x. In this case:

$$r^2 = \frac{SSTotal - SSResidual}{SSTotal}.$$

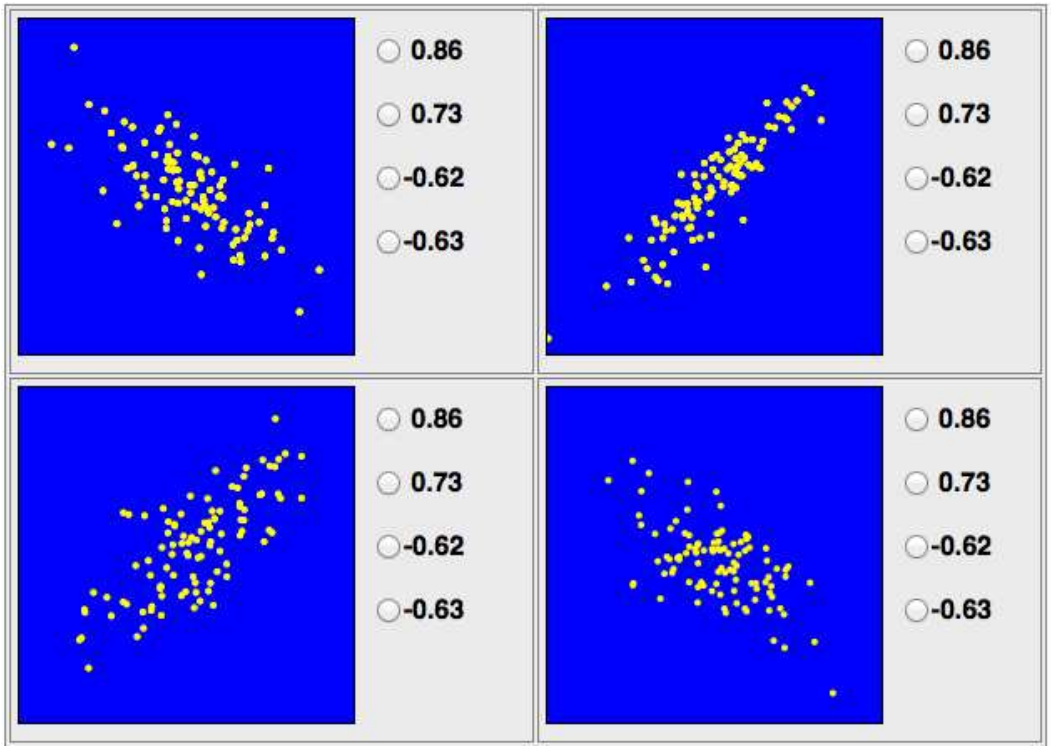
In words, r^2 is the ratio of the sums of squares of the variability that is explained by the model compared to the variability of the most basic model, $\hat{y} = \bar{y}$. Which definition resonates more with you?

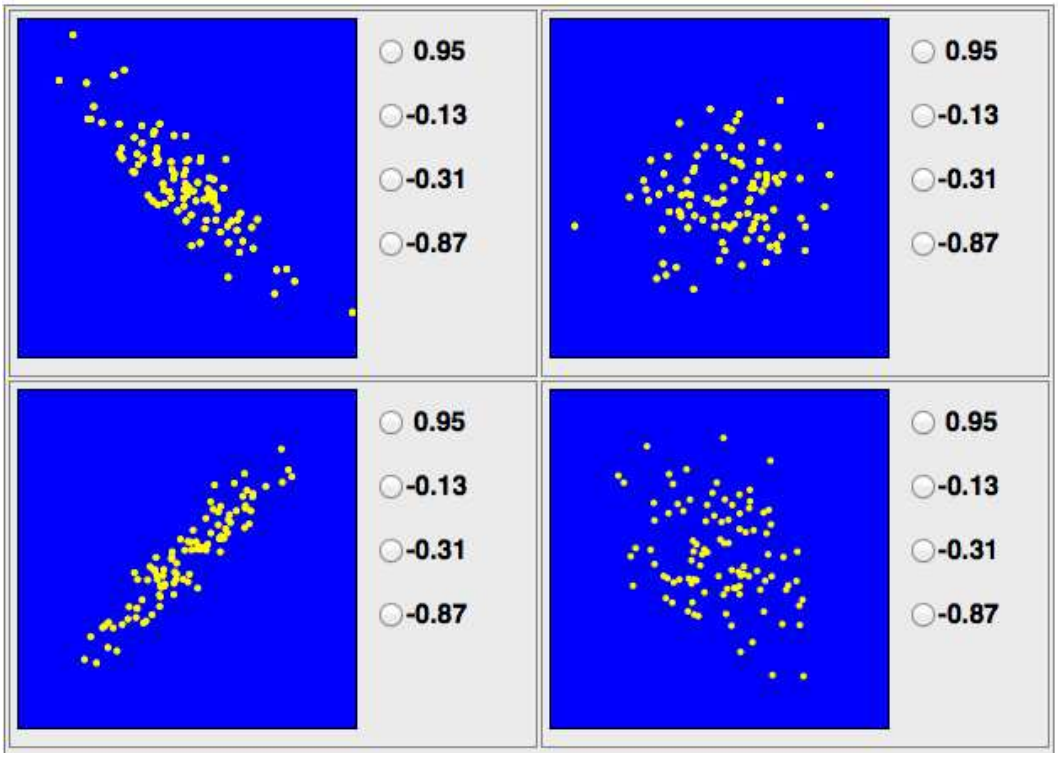
4. There are some important things to keep in mind about correlation. First, correlation is a measure of the strength of a LINEAR relationship. One can calculate r for

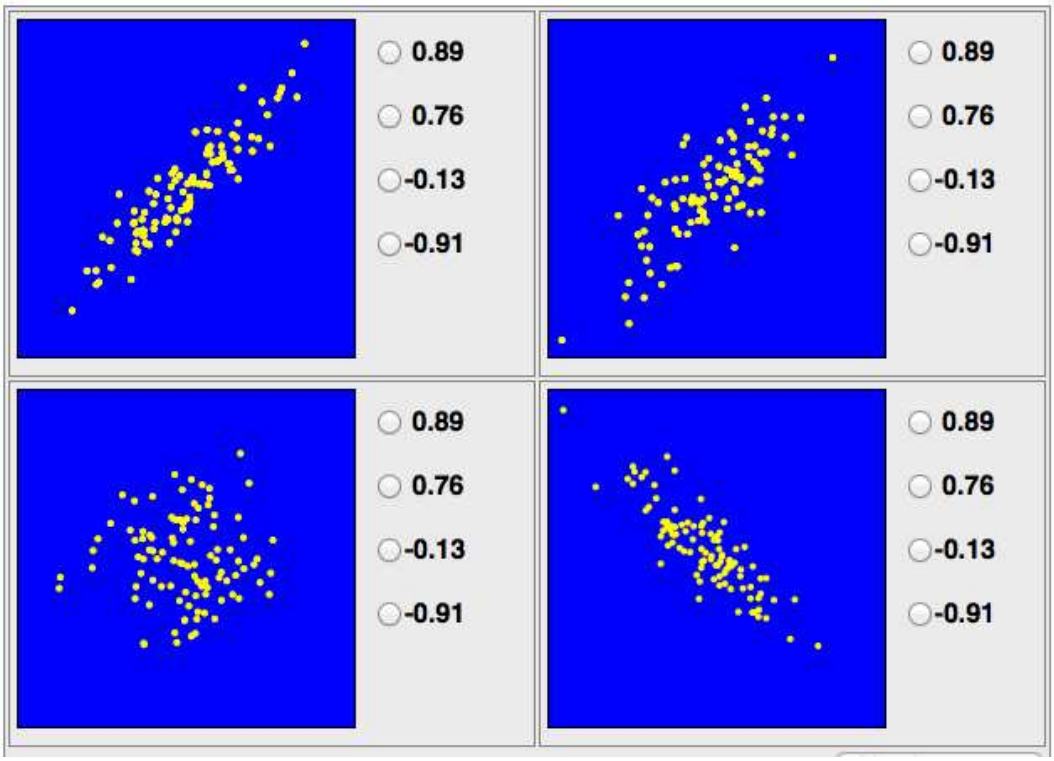
many types of non-linear relationships, but the measure is meaningless if the relationship is clearly non-linear from a visual examination of the scatterplot. Correlation is also a measure for quantitative variables only. Can you think of a time when believing something is linear when it is not can cause you to make mistakes in predicted values?

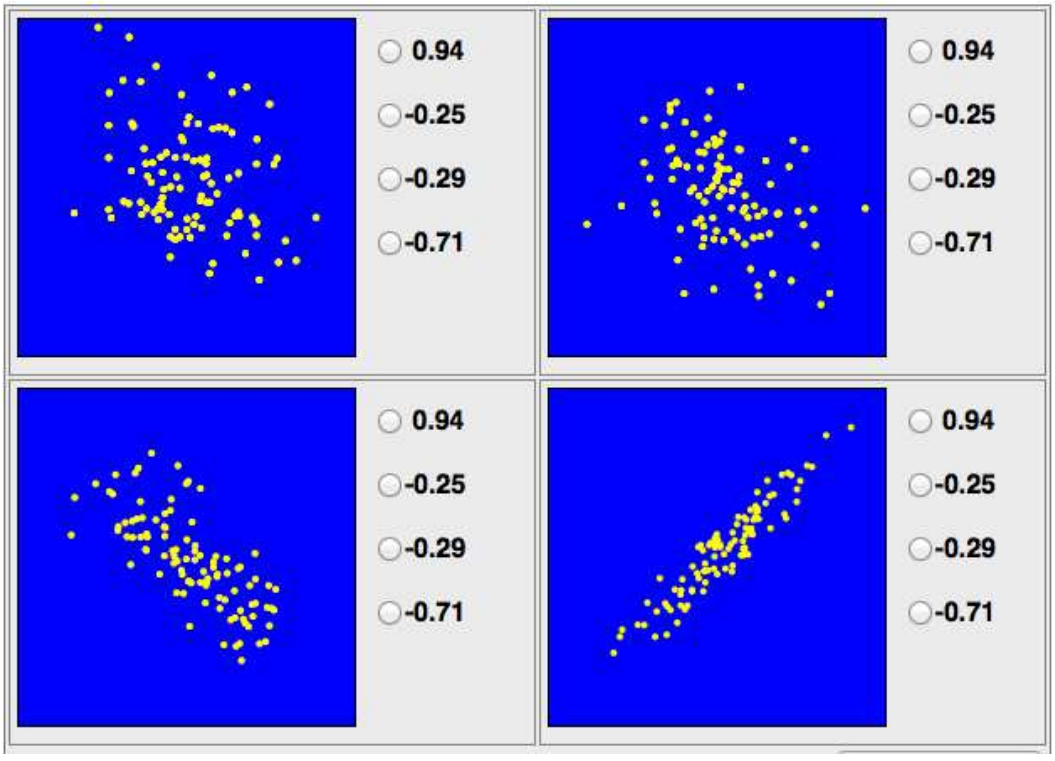
5. And second, correlation does not imply causation. Just because two quantitative variables are strongly linearly correlated, that does not mean that changes in one variable cause changes to occur in the other variable. Both variables may be responding to changes in a third variable that is not in your model. For example, in a sample of elementary school students, there is a strong positive correlation between shoe size and scores on a standardized test of arithmetic skills. Does this mean that studying arithmetic makes your feet bigger? No, shoe size and arithmetic skill are related to each other because both variables respond to a third variable, age. Can you think of an example where correlation does not mean causation?

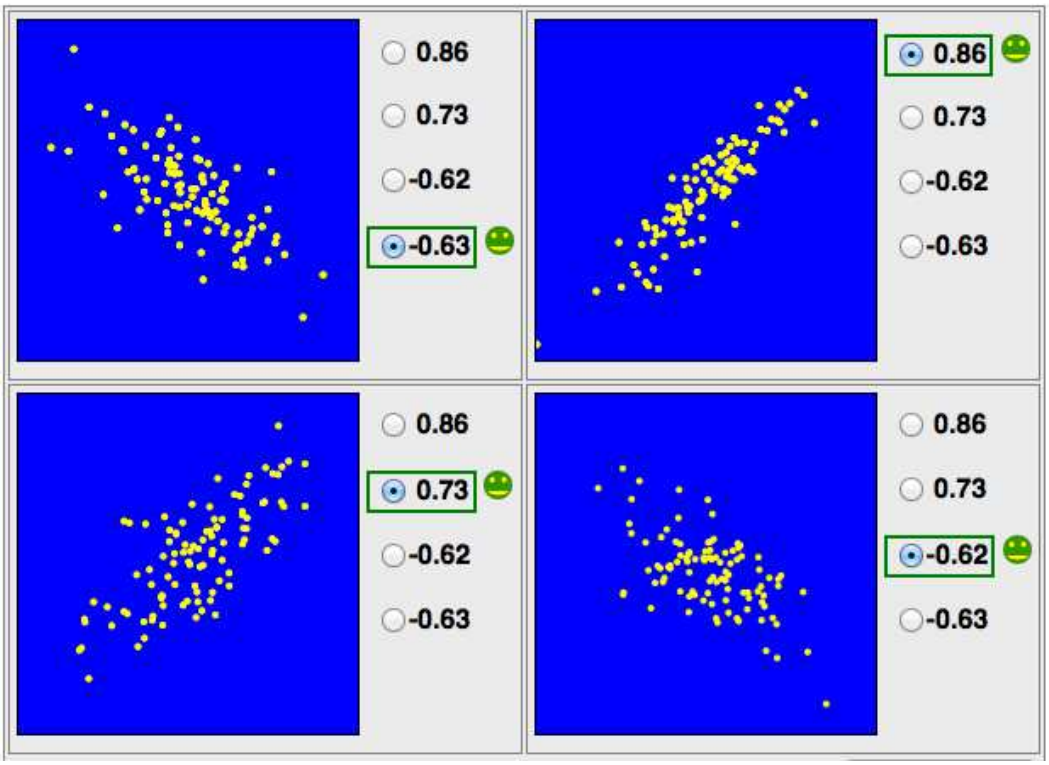
6. To get sense of the measure of correlation, try to guess the correlation r for the scatterplots on the following pages.

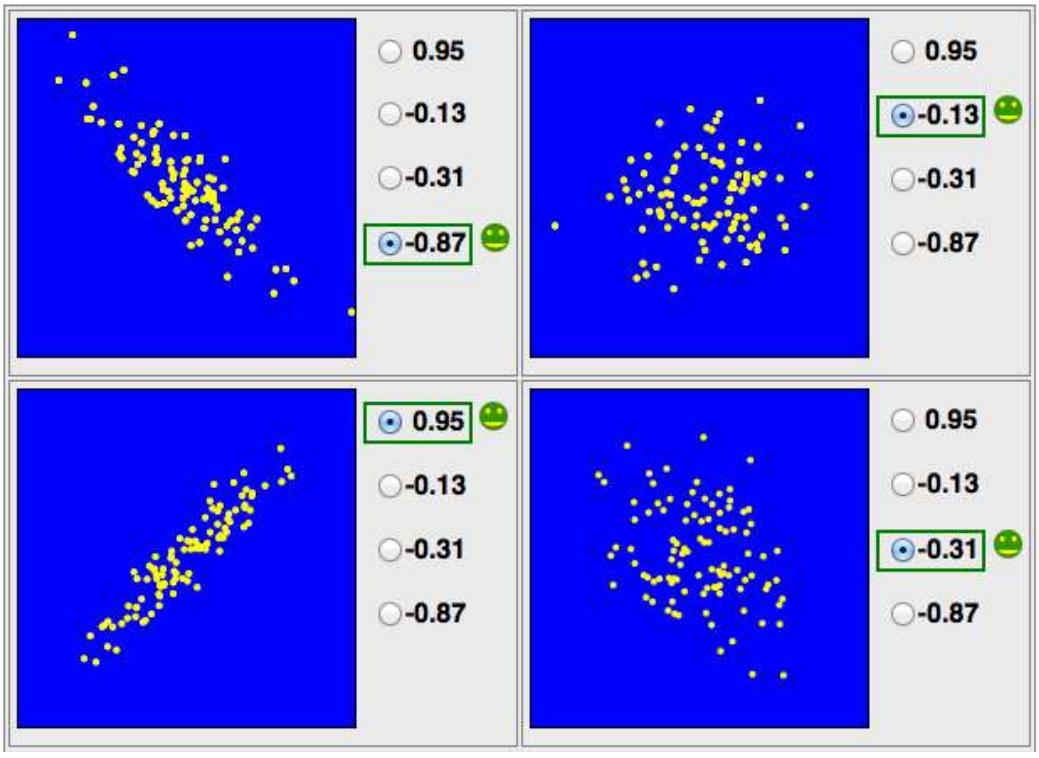

0.86
0.73
-0.62
-0.63
0.86
0.73
-0.62
-0.63
0.86
0.73
-0.62
-0.63
0.86
0.73
-0.62
-0.63

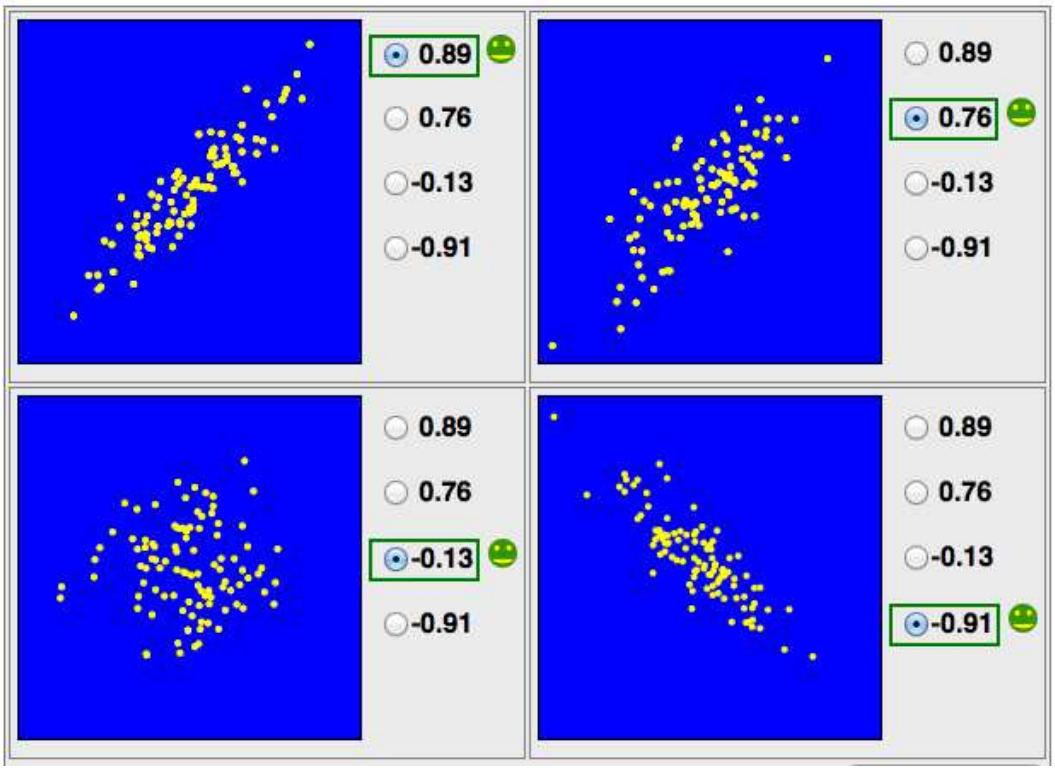

0.95
-0.13
-0.31
-0.87
0.95
-0.13
-0.31
-0.87
0.95
-0.13
-0.31
-0.87
0.95
-0.13
-0.31
-0.87

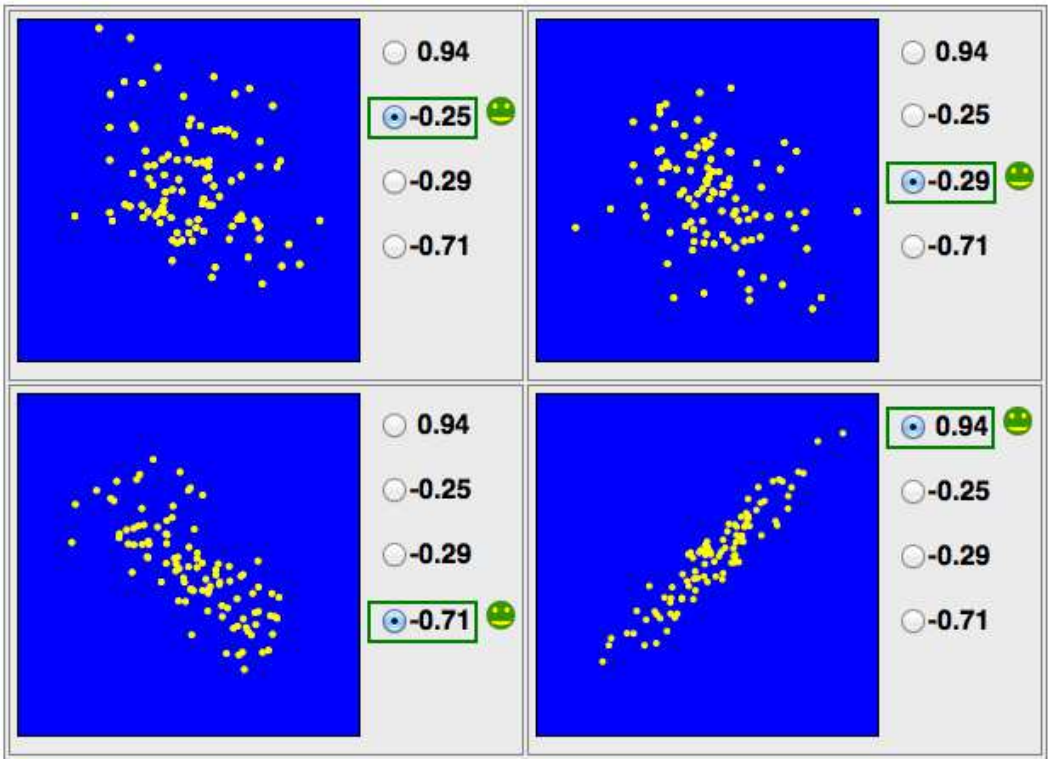

0.89
0.76
-0.13
-0.91
0.89
0.76
-0.13
-0.91
0.89
0.76
-0.13
-0.91
0.89
0.76
-0.13
-0.91


0.94
-0.25
-0.29
-0.71
0.94
-0.25
-0.29
-0.71
0.94
-0.25
-0.29
-0.71
0.94
-0.25
-0.29
-0.71


0.86
0.73
-0.62
-0.63
0.86
0.73
-0.62
-0.63
0.86
0.73
-0.62
-0.63
0.86
0.73
-0.62
-0.63


0.95
-0.13
-0.31
-0.87
0.95
-0.13
-0.31
-0.87
0.95
-0.13
-0.31
-0.87
0.95
-0.13
-0.31
-0.87





These graphics were generated at
http://www.istics.net/Correlations/

Scrabble Word Lab

Before we begin:

This lab refers to the scores that each letter is assigned in a standard american-english scrabble game. A table of these values is given in the Scrabble Letter Lab. This lab is adapted from:

*Workshop Statistics, Discover with data and Fathum.*A Rossman, B Chance, R Lock. Key Curriculum Press, Emeryville 2001, 1-930190-07-7.

In this Activity:

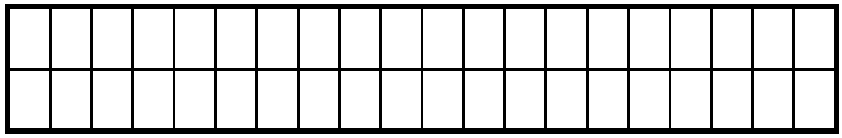
You will generate data from the names in the room, plot the data, and consider the *trend*, *form* and *strength* of data. It is always a good idea to look at your data in the form of a scatterplot before you compute a best fit line and its correlation coefficient.

Materials:

You need paper, pencils, a graphing device and the table of letter scores in the text.

Procedure:

1. Print your whole name in the top row of the table below, one letter per space.



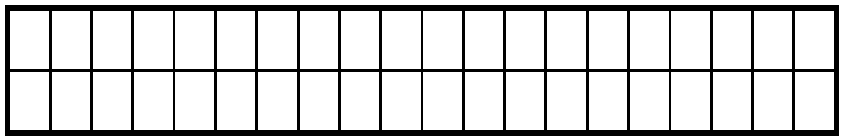
2. Count the number of letters in your name ignoring blanks and spaces.

Number of letters in my name: _____

3. Using the data given on the back page of the Scrabble Letter Lab, write the point value for each letter in your name in the space below it in the Name Score table above. Add the numbers to compute the scrabble value of your name.

Scrabble value of my name: _____

4. Repeat these steps with one or more names of your own choosing ("Double 0 Seven" for instance).



Number of letters: _____ Scrabble value: _____

5. Share your score with the class, and fill in the table given on the last page of this lab with the data.

6. Enter your data into your calculator or graphing utility and make a scatter plot of

word length vs point value for each name. Discuss with your group, the *strength*, *trend* and *form* of your data. Comment on the association between the length of names and their Scrabble value. Describe the three features with a few words below:

Strength

Trend

Form

7. How well do you think you could predict the Scrabble value of a person's name given the length of the name? Try to predict the scrabble value of these names: **Dustin Pedroia** and **Vladimir Putin**. Discuss this question as a group. or class.

8. Can you find examples of pairs of names in which the longer name has a lower value? Can you find groups of three or four such names?

Name Length and Score Data.

Name	Number of Letters	Scrabble Score
Pat Tecake	9	16

Heights Lab

Questions:

Sometimes the r value can be misleading. It is important to also use the residuals to analyze the fit of your fit.

In this Activity:

You will consider both the residuals and the r-value for the data concerning heights in inches and age in years that is located at the beginning of this lab.

Materials:

Graph paper and a graphing calculator.

Procedure:

1. Enter the data from the table below into your calculator.

Height vs. Age

Age (yrs)	2	3	4	5	6	7	8	9	10	11	12	13	14
Height (inches)	35.1	38.7	41.3	44.1	46.5	48.6	51.7	53.7	56.1	59.5	61.2	62.9	63.6

2. Use your calculator to make a scatter plot of height vs. age. Find the equation of the least squares line for predicted median height versus age and graph the line on the plot.

3. Find the value of r that goes with this line. What can you conclude about your regression line based on r. Discuss the data, the line and r as a group.

4. If $L1$ contains the ages and $L2$ contains the heights, then define $L3$ to be $Y1(L1)$. Store the residuals of the fitted line in $L4$ by defining $L4 = L2 - L3$. Plot the residuals vs age on a new graph. Discuss the proper domain and range for these values with your neighbor.

5. Do the new data, the residuals, seem to be randomly scattered about the x-axis? What can be said about modeling this data with a line?

Was Leonardo Correct?

Questions:

Leonardo da Vinci wrote instructions to artists about how to proportion the human body in painting and sculpture. Three of Leonardo's rules were:

- Height equals the span of the outstretched arms.
- Kneeling height is three-fourths of the standing height.
- The length of the hand is one-ninth of the height.

Discuss as a group. Do these proportions seem reasonable? Is Leonardo really suggesting that there is a linear relationship between these lengths?

In this Activity:

In this activity, you will gather data, compute the best fit line, compare it to Leonardo's predicted linear models and use the r-value to support your findings.

Materials:

You will need meter sticks, pencils and classmates to complete this activity.

Procedure:

1. Decide as a class which units of measure you will use. Then working with a partner, measure your height, kneeling height, arm span and hand length, and record it in the table below:

My Lengths

height	kneeling height	arm span	hand length

2. Make a data table that includes the measurements from everyone in the class. You can use the table on the next page to record the class data if you wish.

3. Make three scatterplots of the data, arm span vs height, kneeling height vs standing height and hand length vs height. On each scatter plot, add the lines that Leonardo predicted would model the data.

4. For the plots that have a linear trend, use your calculator to find the least squares regression line and compute the r value or correlation coefficient.

5. Discuss the meaning of the regression lines as a group. In particular, discuss the slopes and y- intercepts in the context of this activity.

6. How well do the lines fit the data? Does the value of r support Leonardo's rules?

Leonardo's Lengths

height	kneeling height	arm span	hand length

Letting Height = H, Kneeling Height = K, Arm Span = A, and Hand Length = L, Leonardo says:

$$K = \frac{3}{4} \cdot H + 0.$$

$$A = 1 \cdot H + 0.$$

$$L = \frac{1}{9} \cdot H + 0.$$

Counting F's

Before we begin:

The text on the following page should not be handed out until you have explained the activity to the students.

Questions:

Survey's are a ubiquitous part of life these days. A well-written survey is very difficult to construct. Let's say, you wanted to find out how many hours a night each participant slept of the week, how would you phrase your question?

In this Activity:

Students are going to count F's in a text and compare their results.

Materials:

Enough copies of the text on the next page.

Procedure:

1. Pass out the text on the next page face down. Tell the participants that they are going to have 1 minute to count all of the F's in the text.

2. Start your watch and let them count. Stop your watch and record the number of F's counted on the board. Did anyone count 34?

3. Let everyone know that no one found the correct number, and give them 3 more minutes to count the F's.

4. Again compare answers.

5. Discuss with the group the implications of this activity.

THE NECESSITY OF TRAINING HANDS FOR FIRST-CLASS FARMS IN THE FATHERLY HANDLING OF FRIENDLY FARM LIVESTOCK IS FOREMOST IN THE MINDS OF FARM OWNERS. SINCE THE FOREFATHERS OF THE FARM OWNERS TRAINED THE FARM HANDS FOR THE FIRST-CLASS FARMS IN THE FATHERLY HANDLING OF FARM LIVESTOCK, THE OWNERS OF THE FARMS FEEL THEY SHOULD CARRY ON WITH THE FAMILY TRADITION OF TRAINING FARM HANDS IN THE FATHERLY HANDLING OF FARM LIVESTOCK BECAUSE THEY BELIEVE IT IS THE BASIS OF GOOD FUNDAMENTAL FARM EQUIPMENT.

Jelly Blubbers Colony

Before we begin:

Make sure that you have plenty of copies of the Jelly Blubbers Colony. This lab was adapted from a Jellyblubber activity invented by Rex Boggs, a teacher in Queensland, Australia.

Questions:

Sampling is another important activity undertaken by amateur and professional statisticians. If I were curious to find out what Americans did to celebrate Memorial Day, I would probably ask my friends and family what they were doing, but this would not give me a very good sample. How could I improve my sample?

In this Activity:

This lab encourages good sampling practices and techniques. Jellyblubbers are a recently discovered marine species. Scientists have discovered a colony of jellyblubbers and they are trying to determine the width of a typical jellyblubber.

Materials:

Just the Jelly Blubbers Colony page.

Procedure:

1. You have been handed a sheet of 100 jellyblubbers. Study the sheet for 10 seconds and then record the numbers of 5 jellyblubbers that you think form a representative sample of this jellyblubber population. Use the second sheet, which lists all the blubbers and their widths, to write down the widths of your chosen blubbers.

Number					
Width					

2. The sample you chose is called a *judgment sample*. Compute the mean width of your sample and record it below:

Mean Width	

3. Share your data with the class by adding your result to the table of widths on the board. You can record the class data on the data sheet on the last page of this lab.

4. Make a dot plot of the data class collected and notice the shape, approximate center and range of the graph. Discuss with the class the shape, approximate center and range of the graph.

5. Next, use your table of random digits to select ten two-digit numbers from 00 through 99. The pair 00 represents 100, and single digit numbers, like 7 for instance,

are represented with two-digits, as 07 for instance. Find the mean of these ten numbers. Contribute your mean to the class data. Again, decide how to compare and comment on the shape, center and range of this new data set. The sample you found using random numbers is called a simple random sample, abbreviated SRS.

6. The *true mean*, the actual, computed mean, of the widths is 18.6 cm. Which of the three methods gave a center closest to 19.4? Which method do you think is the more accurate for finding the mean, a judgment sample of an SRS? Why?

7. If you shake a collection of blubbers, the larger ones tend to sink and the smaller ones rise. You have been a given a sheet divided into five strata. Notice that there is little variability among blubbers within each stratum (singular of strata), but more variability between strata (plural of stratum). Using random numbers and the table of jellyblubber widths by strata, select two blubbers from each stratum and find the mean of these ten numbers. This method is called *stratified sampling*. Share your data with the class and compare the mean of the class data from the stratified samples with the means from the SRS.

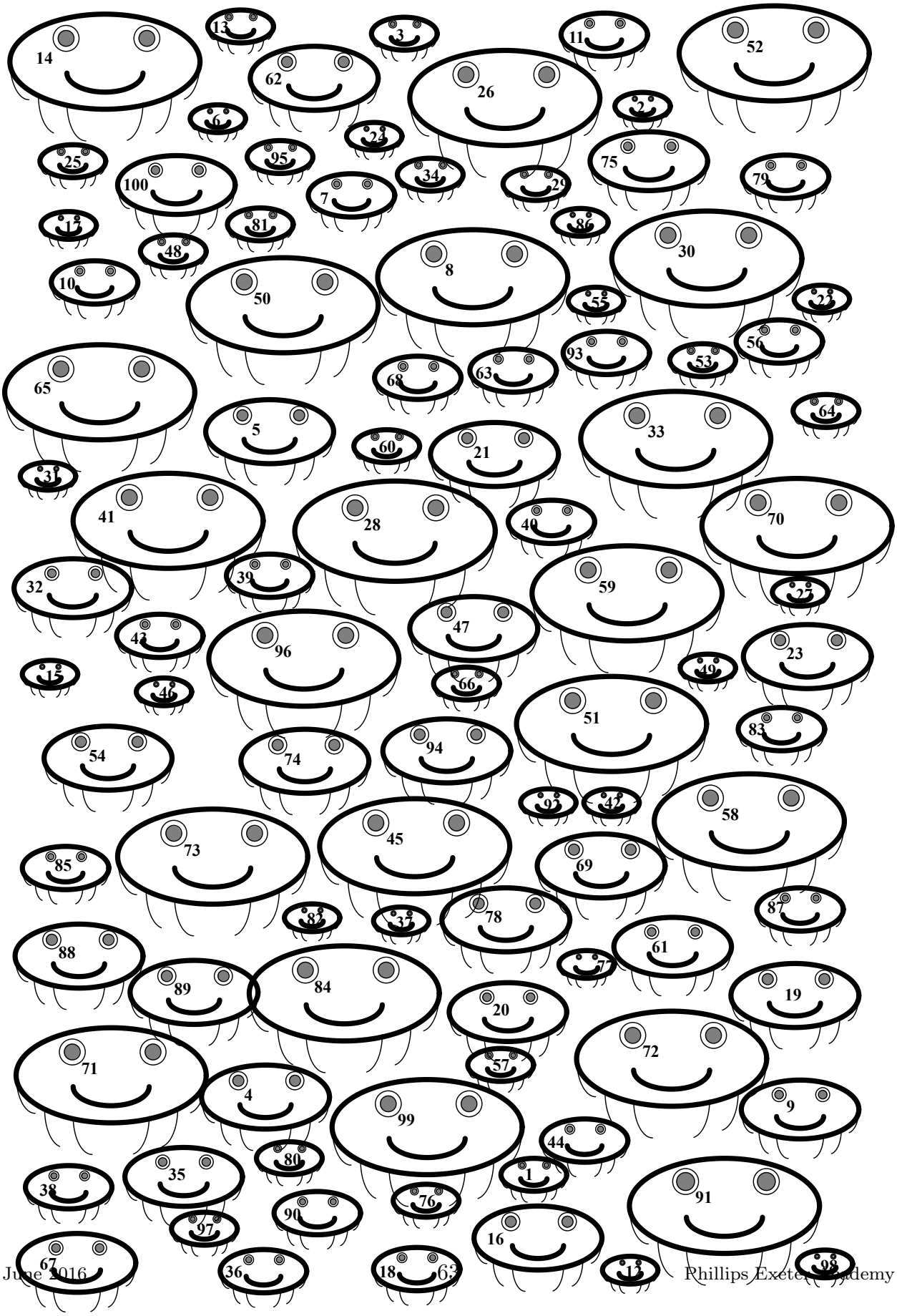
8. There are more ways to pick samples! A *cluster* of blubbers is a group of blubbers near each other in the non shaken collection. There is usually a lot of variability within each cluster, but not much variability among clusters. To select a *cluster sample* pick a random digit between 1 and 20. Call it r. Multiply this digit by 5. Your cluster will consist of that number, $5r$, and the four numbers preceding it. For example, if you pick 11 as your random digit then the cluster will consist of blubbers numbered 55, 54, 53, 52, 51. Referring to the table of jellyblubber width values, write down the 5 widths from this cluster and compute their mean. Share your cluster mean data with the class. Decide on the best way to graph the class data generates from these cluster samples.

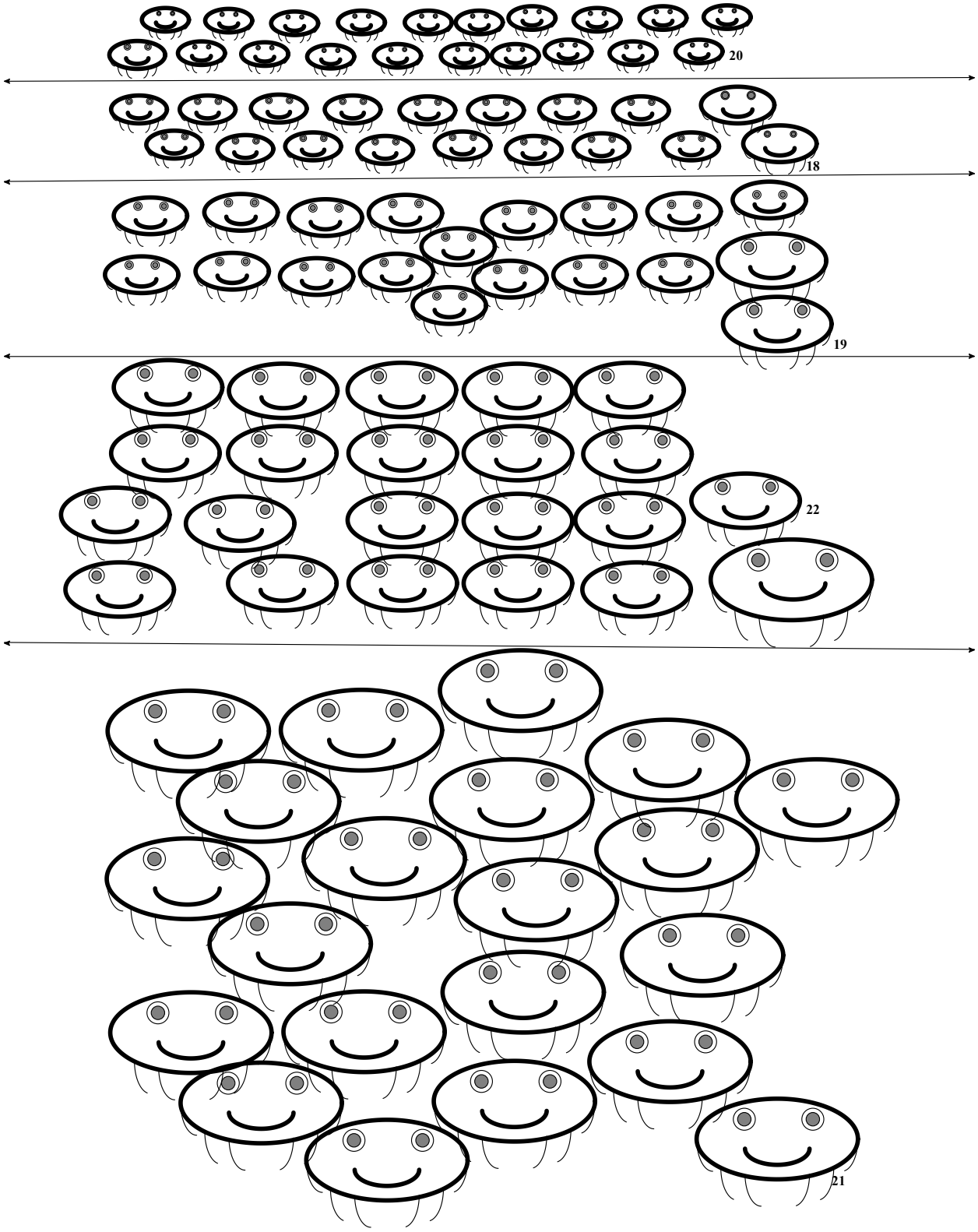
9. There is one more method of sampling using the original population, not the stratified population. Pick a random digit between 1 and 20. This number is the first blubber. Add 20 to your random number. This is your second blubber. Continue until you have five blubbers. For example, if you pick 07 as your random number then your sample will consist of blubbers 7, 27, 47, 67, 87. Compute the mean jellyblubber width of your sample. This is called a *systematic sample*. Comment on the graph generated by these means.

10. Discuss with your group and the class as a whole the advantages and disadvantages of the different sampling methods.

Jellyblubber Sampling Means

	Sample Width				
Sample Number	Judgement	SRS	Stratified	Cluster	Systematic
Mean of Samples					


14
13
3
11
52
62
26
2
6
24
25
95
75
100
34
29
79
7
17
81
86
48
30
8
10
50
55
23
56
93
53
65
68
63
64
5
33
60
21
31
41
28
40
70
32
39
59
27
42
47
96
23
12
49
46
66
51
83
54
74
94
92
42
58
73
45
85
69
87
82
37
78
88
61
77
89
84
19
20
71
72
57
4
99
9
44
80
35
1
38
91
76
90
97
16
67
36
18
12
98


20
18
19
22
21

Jellyblubber Width Values

Blubber #	Width	Blubber #	Width	Blubber #	Width	Blubber #	Width
1	9	26	40	51	35	76	7
2	5	27	5	52	37	77	5
3	9	28	49	53	9	78	25
4	33	29	9	54	25	79	17
5	22	30	41	55	5	80	8
6	5	31	5	56	10	81	8
7	10	32	20	57	9	82	5
8	40	33	43	58	45	83	13
9	20	34	7	59	40	84	40
10	10	35	20	60	8	85	10
11	12	36	10	61	20	86	5
12	5	37	5	62	25	87	10
13	8	38	14	63	10	88	27
14	41	39	15	64	8	89	30
15	5	40	10	65	37	90	10
16	32	41	41	66	8	91	40
17	5	42	5	67	20	92	6
18	10	43	17	68	13	93	10
19	21	44	15	69	34	94	25
20	20	45	40	70	42	95	7
21	34	46	5	71	40	96	40
22	5	47	30	72	40	97	8
23	32	48	8	73	40	98	5
24	5	49	5	74	30	99	40
25	9	50	40	75	20	100	20

Jellyblubber Width Values, Stratified

Blubber #	Width	Blubber #	Width	Blubber #	Width	Blubber #	Width
1	2	26	8	51	13	76	32
2	2	27	8	52	13	77	32
3	5	28	8	53	14	78	33
4	5	29	8	54	15	79	34
5	5	30	8	55	15	80	34
6	5	31	8	56	17	81	35
7	5	32	8	57	17	82	37
8	5	33	9	58	20	83	37
9	5	34	9	59	20	84	40
10	5	35	9	60	20	85	40
11	5	36	9	61	20	86	40
12	5	37	9	62	20	87	40
13	5	38	9	63	20	88	40
14	5	39	10	64	20	89	40
15	5	40	10	65	20	90	40
16	5	41	10	66	21	91	40
17	5	42	10	67	22	92	40
18	5	43	10	68	25	93	40
19	5	44	10	69	25	94	41
20	5	45	10	70	25	95	41
21	6	46	10	71	25	96	41
22	7	47	10	72	27	97	42
23	7	48	10	73	30	98	43
24	7	49	10	74	30	99	45
25	8	50	12	75	30	100	49

Discrimination or Not?

The scenario:

At Main Street Bank last year 48 male bank supervisor were each given a personnel file and asked to judge whether Pat Tecake, the person represented in the file should be recommended for promotion to a branch-manager position or whether the Pat Tecake should not be recommended for promotion. The files given to each of the 48 supervisors were in identical except that half of the files (24) were labelled "male" and half of the files were labelled "female". Of the 48 files reviewed, 35 were recommendation for promotion.

In this Activity: You explore the notion of chance variation and see how to use simulations to determine whether an outcome can be explained by chance variation.

Materials: You will need a deck of cards and pencils.

Procedure:

1. Suppose that the recommendations showed no evidence of discrimination on the basis of gender. How many male Pats would you expect to be recommended for promotion? How many females? Enter these values in the table below

No discrimination by gender

	Promotion	No Promotion	Total
Male			24
Female			24
	35	13	48

2. Now suppose that the recommendations showed strong evidence of discrimination on the basis of gender. How many male Pats might you expect to be recommended for promotion? How many females? Complete the table below to show a possible example of this case.

Discrimination on the basis of gender

	Promotion	No Promotion	Total
Male			24
Female			24
	35	13	48

3. Suppose the evidence for discrimination was inconclusive, neither strongly in favor nor strongly against. Complete the following table to illustrate this situation.

Inconclusive

	Promotion	No Promotion	Total
Male			24
Female			24
	35	13	48

4. In the actual situation in the study, the results were that 21 of the 24 files labelled "male" were recommended for promotion, and 14 of the 24 files labelled "female" were recommended for promotion. Enter these data into the table below.

Actual

	Promotion	No Promotion	Total
Male			24
Female			24
	35	13	48

5. In the actual situation, what percentage of the recommended candidates were male? Female?

6. Do you think there is evidence of discrimination against the female candidates? How certain are you?

7. How likely do you think that mere chance was responsible for the smaller number of females recommended for promotion?

8. If you were the attorney retained by the female applicants how would you go about collecting evidence to decide whether the results occurred by chance or whether there really was discrimination?

Simulating the Case

If there really were no difference between the male and female candidates then you might as well roll a die or use playing cards to pick male or female completely at random. Here's how you will do it.

9. Remove two red cards and two black cards from a full deck of cards. You now have 48 cards, 24 of each color. Let black represent male and red represent female.

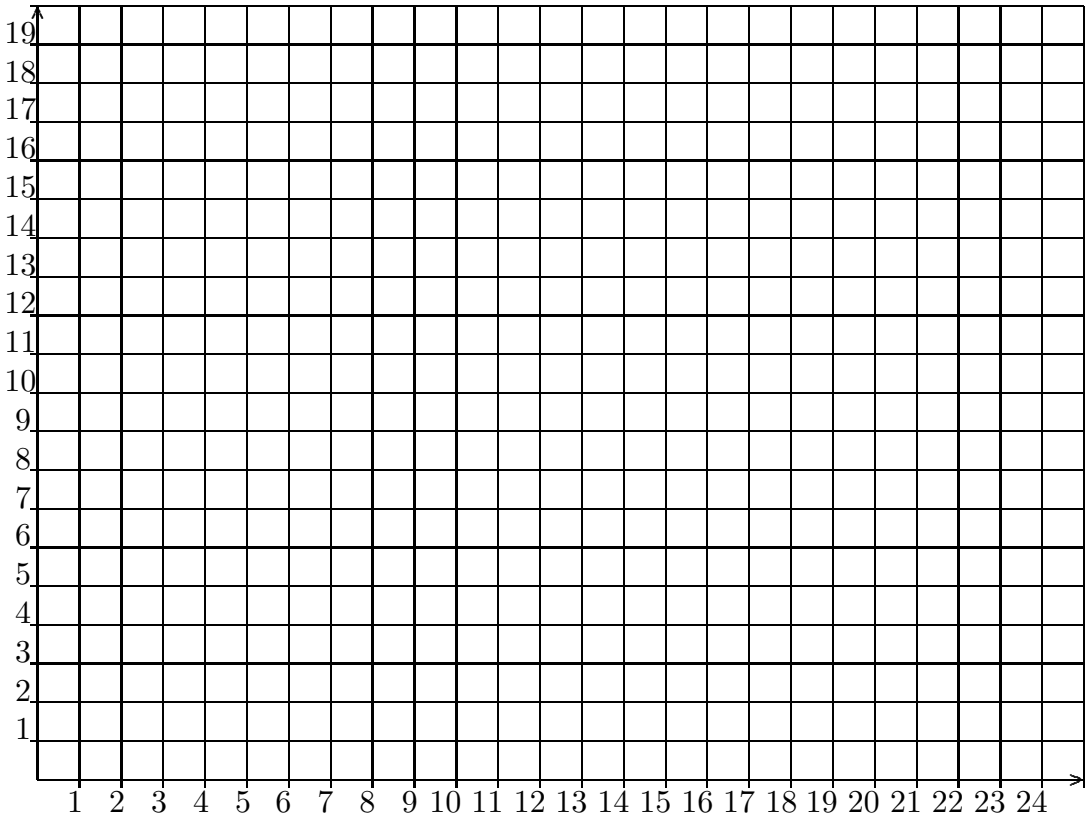
10. Shuffle the cards thoroughly and deal out 35 cards to represent the 35 candidates who were recommended for promotion. You could do this more efficiently by dealing out 13 cards to represent the candidates who were not recommended for promotion.)

11. Count the number of black cards to represent the number of men recommended for promotion. Record this number in the table below. And then repeat steps 10 and 11 nineteen more times for a total of 20 simulations.

Simulation Data

Trial #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
# of black cards																				

12. When your table is complete, record your findings on the grid given by placing a dot above the number of black cards you counted. The dots will stack up as the numbers are repeated.



Number of men recommended for promotion

13. Estimate the chances that 21 or more black cards (males) would have been selected based on your simulation.

14. Based on your simulation do you think there is any evidence to support the claim that recommending 21 males out of 35 candidates was due to discrimination rather than to chance variation? In other words, how do your simulated results compare with those of the original study?

15. How do your simulated results compare with those of your classmates?

Sudoku Experiment Part II

Before we begin:

The class will need to have complete Sudoku Experiment Part I in order to complete this lab.

Questions:

Perhaps the differences in times to complete a puzzle were not actually due to which type of puzzles were solved, but occurred only by chance. How could we check to see if the difference in completion times that we recorded were due simply to chance? What does that question mean in a statistical context?

In this Activity:

Students will explore the difference of two means to test the strength of their conclusions. Students will analyze the data again, this time using groups of times that are randomly selected.

Materials:

Students need the data collected in Sudoku Experiment Part I and note cards.

Procedure:

1. Using one notecard per Sudoku puzzle, write only the time it took to complete the puzzle on the notecard. Once you have recorded each time on a separate notecard, shuffle the note cards. Split the deck into two piles divided roughly in half. Compute the mean time for each of the two piles and find their difference. Record your numbers in the table below:

Sudoku Random Grouping Means

Simulation	1	2	3	4	5	6	7	8	Experimental Data
Group 1									
Group 2									
Difference									

2. Shuffle the deck again, and repeat the process of dividing the deck into two parts and computing the mean of each part. Record your data in the your table and then share your data with the whole class.

3. Make a dot plot of all of the differences between group means.

4. Describe the distribution of the differences in sample means that were collected.

5. Did any of the simulations result in a difference value as large as the one initially found?

6. Use your calculator to compute the standard deviation of the differences. How many standard deviations away from the mean does the original difference fall? Does this suggest that there is a significant difference in the time it takes to do each kind of puzzle, or do you think that the difference was just a matter of chance?

Titanic

Before we begin:

This data in this lab comes from:
www.encyclopedia-titanica.org/titanic-statistics.html
and some of the questions come from:
Statistics In Action, A Watkins, R Scheaffer, G Cobb. **Key Curriculum Press**, Emeryville 2008, 978 -1-55953-909-8.

Questions:

You've heard expressions like, "the chance of striking out a right-handed batter given that you are a left-handed pitcher" or "the chance of having a boy given that you already have 2 girls." How can you answer these questions effectively?

In this Activity:

In this lab, we explore these expressions of *conditional probability*. We begin the disastrous event of April 15^{th}, 1912, when the unsinkable Royal Mail Ship *Titanic* struck an iceberg in the North Atlantic and sank. Only 710 of her 2204 passengers and crew survived. The following *two-way table* records the data on the fate of her passengers.

Titanic Survival Data

	Survived	Did Not Survive	Total
First Class	201	123	324
Second Class	118	166	284
Third Class	181	528	709
Total Passengers	500	817	1317

Materials:

Pencils and paper.

Procedure:

1. Calculate the following probabilities. Leave your answers in fraction form.

A. If one passenger is randomly selected, what is the probability that the passenger was in first class?

B. If one passenger is randomly selected, what is the probability that this passenger survived?

C. If one passenger is randomly selected, what is the probability that this passenger was in first class and survived?

D. If one passenger is randomly selected, what is the probability that this passenger was either in first class or survived or possibly both?

E. If one of the passengers is randomly selected from the first class passengers, what is the probability that this passenger survived?

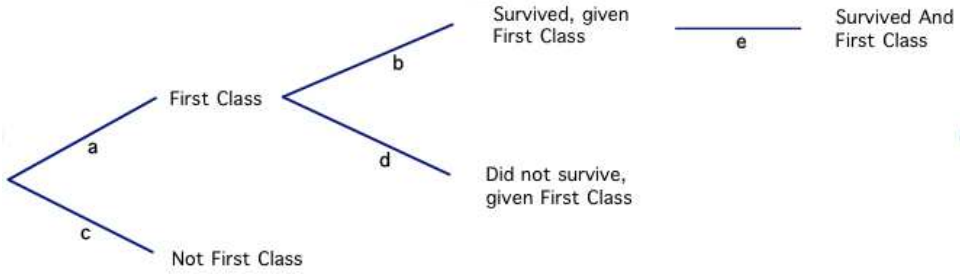
F. If one of passengers is randomly selected from the group of survivors, what is the probability that this passenger was in first class?

2. Why is the answer to part **E** above larger than the answer to part **C** above?

3. The questions asked in parts **E** and **F** are examples of *conditional probability.* In part **E** we already know that the passenger is in first class. This is the condition. We could rephrase the question as, "What is the probability that the passenger survived, given that he/she is in first class?" Rephrase the question to part **F** using this phrasing.

4. How are your answers to parts **A**, **C** and **E** above related?

5. A *tree diagram* is a great way to help answer the previous question. Below is a part of the tree which illustrates this story.



Each lowercase letter in the diagram should be replace with the probability of the event at the end of the branch. You have already computed some of the probabilities, add in all of the probabilities to the tree.

6. The conventional shorthand for writing the probability of event S is $P(S)$, so $P(\text{first class}) = 324/1317$. Use your answers to question 5 to find a formula for $P(S \,\text{given}\, T)$ in terms of $P(S \,\text{and}\, T)$ and $P(T)$.

7. One more example before we go.This problem is a classic example of a probability question with a surprising outcome. Suppose there is a rare but serious disease that affects 1 % (0.01) of the population. There is a test that correctly identifies 95% (0.95) of the time that you have the disease if you are in fact infected. The test also correctly identifies 95% of the time that you are not infected if you are in fact not infected.

Suppose that during a routine physical you take the test which turns out positive. The test says that you are infected.

(A) What is the probability that the test reports a positive result if you are infected?

(B) What is the probability that you are infected if the tests reports a positive result?

It may help to assume that the population is one million people. You might also want to use a tree diagram or a two-way table like the one shown earlier in this lab. In the case of the table the variables would be **test: positive** or **test: negative** and **infected** or **not infected**.

Probability and Independent Events

Before we begin:

This is a short worksheet on *statistical independence*. It seems reasonable to say that the price of a dozen eggs in Exeter and today's temperature in London are independent events. Knowing the value of one, gives you no further information about the other. But what about events, "X was in first class" and "X did not survive"?

In this Activity:

You will explore the notion statistical independence. The statistical definition of independence is that A and B are independent events if and only if $P(A \text{ given } B) = P(A)$. In other words, knowing the probability of B occurring sheds no light on the conditional probability of A given B.

Materials: You will graph paper, a deck of cards and a pencil.

Procedure:

1. Are the events "passenger survived" and "passenger was in first class" independent events? Support your answer with appropriate probability calculations.

2. Are the events "passenger survived" and "passenger was in third class" independent events? Support your answer with appropriate probability calculations.

3. Use appropriate probability calculations to support the statement: "Not all passengers on board the *Titanic* had the same probability of surviving.

4. From a well-shuffled regular 52-card deck, pick a card at random. Note its color and replace it. Take a second random card from the same deck. Note its color and replace it. What is the probability that both cards are the same color?

5. From a well-shuffled regular 52-card deck, pick a card at random. Note its color and do not replace it. Take a second random card from the same deck. Note its color and replace it. What is the probability that both cards are the same color?

6. Draw a tree diagram to illustrate the probabilities in questions 4 and 5.

7. Would you agree that with the following statement? A and B are independent if and only if

$$P(A \text{ and } B) = P(A) * P(B).$$

Hand Eye Coordination

Questions:

Do you know if your friends are right-handed or left-handed? Can you use this information to find out if they are right-eyed or left-eye?

In this Activity:

This lab explores the independence of two attributes and gives an example of a *two-way* table.

Materials:

People, pencils.

Procedure:

1. Determine whether each person in your group is right-eyed or left-eyed. The procedure for determining which is simple. Pick out an object that is 15 feet or so away from you and face the object. Hold your hands together, palms out, at an arms length, making a small space that you can see through. Look at the chosen object. Now close your right eye, keeping your left eye open. Can you still see the object? If yes, congratulations, you are left-eyed. Repeat, but this time closing your left eye and keeping your right eye open. Can you still see the object? If yes, congratulations, you are right-eyed. Record your personal data below:

Your Hand Eye Information

Name	Eyedness L or R	Handedness L or R

2. Share your data with the whole group by filling in a large table with the name and hand and eye data. Discuss your findings. Do you see any trends?

3. A *two-way* table is a great way to summarize data such as the hand and eye data you have collected. There are four possible combinations for handedness and eyedness, either LL, LR, RL or RR. Count the number of each type of person and record it in the *two-way* table below:

Hand and Eye Two-Way Table

	Handedness	
Eyedness	Right	Left
Right		
Left		

4. For a randomly selected member of the class are handedness and eyedness independent? Discuss with your group and the class.

Music and Sports

Questions:

How do you compare data that is not numeric like height in inches or time in seconds? What about data such as your favorite ice cream flavor or the color of your bicycle or whether you bike, walk, drive, bus or train to work?

In this Activity:

This lab explores *categorical data*, which is different than the numeric *quantitative* data that we have worked with in all of our previous labs.

Materials:

People, pencils, graph paper.

Procedure:

1. Ask your neighbor the following questions and record them below:

Do you play a sport? Yes ☐ No ☐

Do you play a musical instrument (voice included)? Yes ☐ No ☐

2. Now collect all of the data from the class on the board and record it in the table below.

Music and Sports

	Sports?		
Music?	Yes	No	Total
Yes			
No			
Total			

Discuss your observations as a group.

3. Is there an *association* between playing a sport and playing a musical instrument? If you play a sport, are you more or less likely to play a musical instrument?

4. Of those who play sports, what proportion also play a musical instrument? Of those who do not play sports, what proportion play a musical instrument? Does there seem to be an association?

5. One way to visualize this data is to create a graph, similar to a histogram with two bars of equal heights, one labeled Yes, and one labeled No, based whether the person plays sports or not. Then each bar is colored with one color to indicate Does play an instrument and with another color to indicate Does not play an instrument. Now compare areas and see if you change your conclusions about the association between playing a sport and playing an instrument.

Expected Value Lab

Questions:

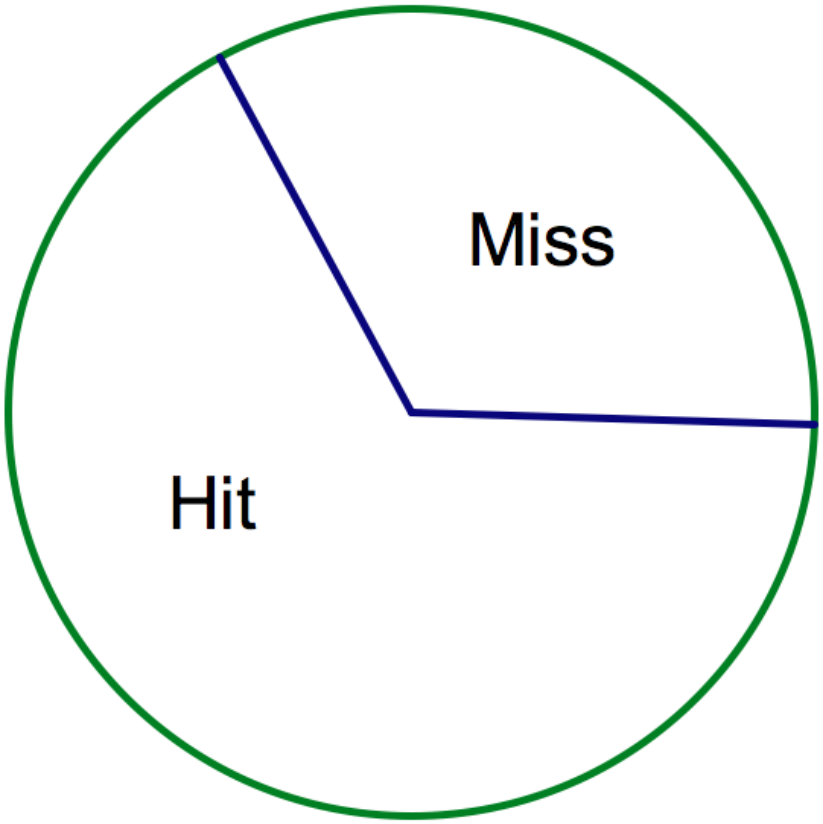
The school basketball team is in a one-and-one situation. This means that each time the team goes up for a free throw, they will get a second free throw if they make the first. Shura is a 60% shooter. This means that in the long run, Shura makes a basket 60% of the time. On any given trip to the free throw line, how many points do you expect Shura to make, 0, if she misses the first shot, 1, if she makes the first shot and misses the second, or 2, if she makes both shots?

In this Activity:

Students will simulate Shura's free throw attempts and learn about *expected value*.

Materials:

You will need a pencil or pen, a bobby pin and the spinner provided below.



Procedure:

1. Discuss with your classmates how many points you expect Shura to score on any trip to the free throw line.

2. In 100 trips to the line, about how many points will Shura make in total? Discuss again.

3. Over the long run, what do you think will be Shura's average number of points per trip to the line? Discuss how to compute this number with your classmates.

4. Place a pen or pencil through the bobby pin you have been given at the center of the circle to make a spinner. Try flicking the bobby pin to see if you can simulate Shura's hitting or missing the hoop.

5. Divide into pairs, and assign one person to record and one person to 'throw'. Use the spinner to simulate 20 trips to the free throw line, and record your results in the spinner's table. Switch jobs, and simulate 20 more trips, recording your results in the spinner's table.

Shura's Simulation

Points	Frequency	Total	Approximate Probability
0			/20
1			/20
2			/20
		20	

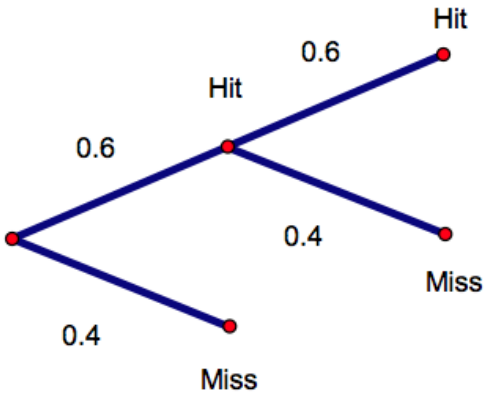
6. Combine the data from the whole class and enter the data in the table below.

Shura's Simulation, Class Data

Points	Total Frequency	Total Frequency/Total Number of Trials
0		
1		
2		

7. Based on the class data, what is Shura's most likely score, 0, 1 or 2?

8. Below is a tree diagram that helps analyze the theoretical probabilities of Shura's outcomes. Discuss the graph with your partner. Can you explain the notation $P(A, B)$? What is the sum of the three probabilities of the three different outcomes, 0 points, 1 point or 2 points?



2 Points: P(Hit, Hit) = (0.6)(0.6) = 0.36

1 Point: P(Hit, Miss) = (0.6)(0.4) = 0.24

0 Points: P(Miss) = 0.4

9. Theoretically, in 100 trips to the free throw line, how many 0's, 1's and 2's can we expect Shura to score?

10. What is Shura's average score per trip to the line? This number is called the *expected value* of Shura's score.

M&M Concentration

In this Activity:

We will use a sample to estimate the percentage of green M&Ms in a population. We will also look at sample variation to explore how confident we are in using a single sample *statistic* to estimate a *parameter*.

Materials:

A large bin filled with standard M&Ms (at least 500), napkins and Dixie cups for each student. You will also want to draw a large table on the board with the headings, Sample Number and Percent Green M&M's.

Procedure:

1. Scoop a cup full of M&Ms from the bowl. Pour them out onto a napkin and make sure you have a sample of exactly 30 M&Ms. You may have to add a few or take some away. Try to do this without sorting out or adding a particular color. You may want to close your eyes to select which you will add or take away.

2. Count the green M&Ms present in your sample. Calculate the percentage of your sample that is green. Round to the nearest percentage point.

3. Record your value on the board with the rest of the class data.

4. Repeat the sampling process until your class has 100 sample statistics. Make a distribution graph of the results.

5. Based on your class's data, what do you think is the true percentage of green M&Ms? If we were to use the first sample statistic to estimate the parameter, the percentage of green M&Ms in the entire bowl, what would our estimate be? Is another sample a better estimate?

6. Calculate the range of percentage values in the sample statistics. How confident are you that you and your classmates have found the value of this parameter?

7. Using the class data, find a range of values that would include 90% to 95% of the sample statistics from the class. Are you convinced that the parameter falls into this range?

8. As a class, repeat the process of collecting sample statistics, but this time, use samples of size 10.

9. Make a distribution of the class means for the samples of size 10. Discuss any differences that you see as a class. Did a smaller sample size produce different sampling variation? Could you have predicted this outcome?

10. Suppose you are a quality control inspector at the M& M plant. On March 1 you take a sample of 30 m &m's and determine that 10 are green. Should you worry that there is a problem with the mixing station? On April 1 you pull a sample of 30 and find that 13 of them are green. Should you report a problem now?

11. In 2008, Mars stated that their candy mix includes 16% green candies. How does this value compare to yours?

12. Topics for further discussion include: Were your samples biased? How many samples do you need to make a reasonable guess at the parameter's value? What have other interested consumers found to be the value of the parameter?

Simulation Lab

Questions:

The Blood Bank of the Redwoods in Santa Roa, California is running low on its pints of type A blood. Long-term statistics kept by the blood bank indicate that 40% of donors have type A blood. In a recent blood drive, the blood type of the donors was tested as they entered the Santa Rosa High School gymnasium to donate. What is the probability that it will take at most 4 donors to find one with type A blood?

In this Activity:

Students will use a table of random digits to simulate donor-testing and explore the empirical probability of finding a type A donor within 4 tests.

Materials:

You will need the table at the back of this book and pencil and paper.

Procedure:

1. Look at your table of random digits and notice that only the ten numbers $0, 1, \ldots, 9$ appear in the strings of digits. These digits will represent the population of all blood donors. Each digit represents one person. Assign 4 numbers to represent type those donors who have blood type A. Thus 40% of the digits, or people, in your population will have type A blood. The other 60%, represented by the remaining 6 digits, have another blood type.

2. Ask your teacher for a row assignment, and start reading the string of numbers on the left end of the row. (The gaps in the list are only for readability and have no other significance.) Read until you find a type A blood donor. When you find a type A blood donor, stop and count how many digits you have read, or donors you have tested since the last type A donor appeared. If your type A appears within the first 4 numbers, record that in the table below as a successful trial and if not, record that as a failed trial. Begin a new trial after each occurrence of a type A blood donor.

Trials

Successful	Failed	Total
		40

3. Continue this process until you have completed 40 trials. In how many trials did it take at most 4 donors to find one with type A blood?

4. What is your empirical probability of the blood bank finding a type A donor in at most 4 attempts?

5. Combine the class results and discuss your findings. Discuss the best way to visualize the results. Discuss the distribution of results.

6. The theoretical probability is 0.8704. How close is this number to the mean of the class's results?

7. How would you change your process if 60% of donors were type A?

8. How would your process change if you considered success being the occurrence within 4 donors any of type A or type B^+ or AB^- which long term statistics indicate occur at 40%, 9% and 1% respectively?

9. How would you change your process if 45% of donors were type A?

10. One way to determine theoretical probability above is:

$$1 - P(x \ni \{1, 2, 3, 4\}) = 1 - (0.6)^4 = 0.8704.$$

Another way is:

$$P(x \leq 4) = P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4)$$

$$P(x \leq 4) = (0.4) + (0.6)(0.4) + (0.6^2)(0.4) + (0.6^3)(0.4) = 0.8704.$$

Can you think of another?

Crop Sampling

Before we begin:

This lab was adapted from a lab designed by Carolyn Doetsch, Peter Flanagan-Hyde, Mary Harrison, Josh Tabor and Chuck Tiberio for the North Carolina School of Science and Mathematics Statistics Leadership Institute and found at:
courses.ncssm.edu/math/Stat_inst01/PDFS/river.pdf

Questions:

At the beginning of the spring a farmer cleared a new field and planted a first crop of corn. The new field is a unique plot of land in that a river runs along one side. The corn looks good in some areas of the field but not in others. The farmer is not sure that harvesting the field is worth the expense. He has decided to harvest 10 plots and use this information to estimate the total yield. Based on this estimate, he will decide whether to harvest the remaining plots and also whether to use this field next year.

In this Activity:

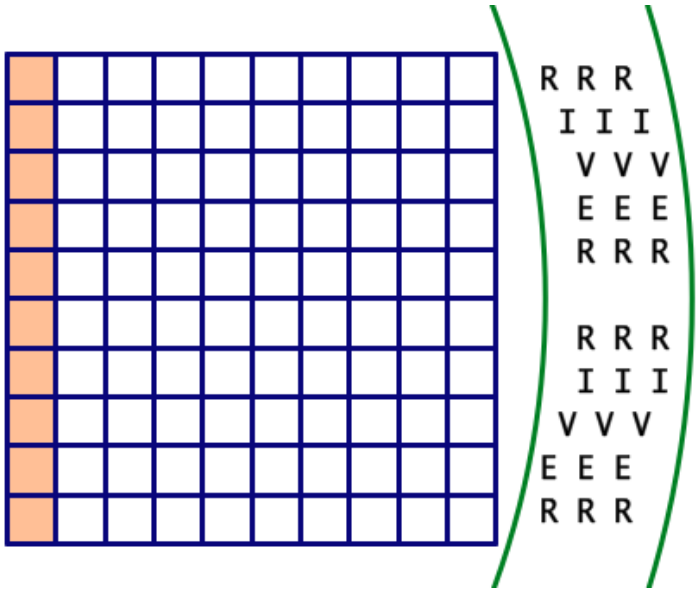
You will explore four different sampling methods: convenience samples, simple random samples, vertical strata and horizontal strata and discuss their merits and pitfalls.

Materials:

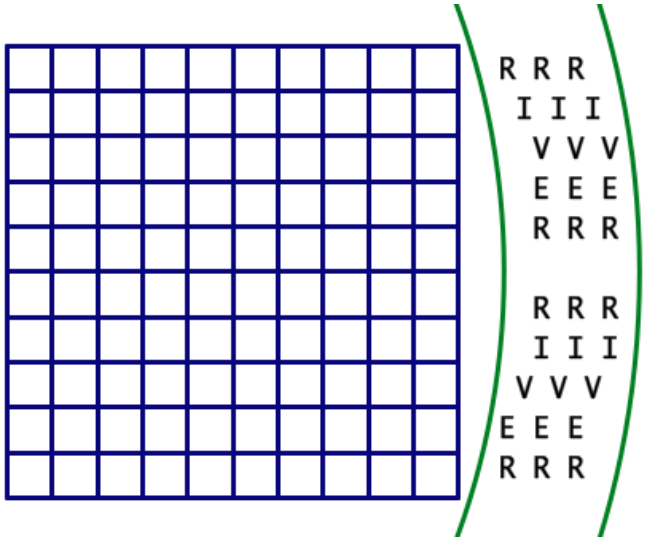
A calculator, pencils and paper.

Procedure:

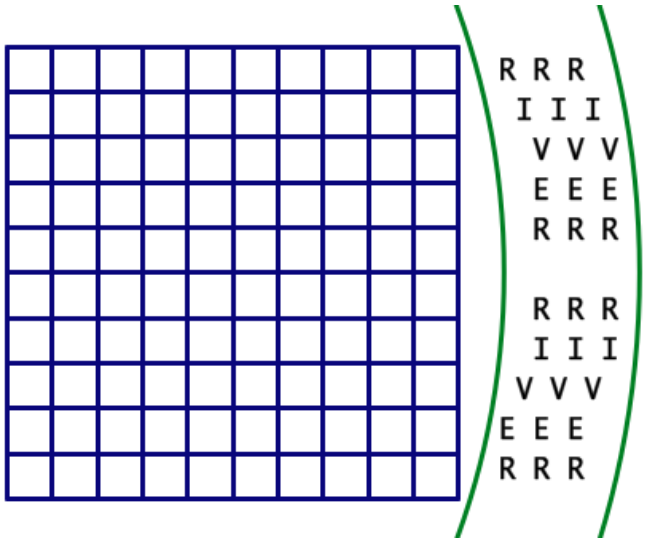
1. *Convenience Sample.* The farmer began by choosing 10 plots that would be easy to harvest. They are filled on the grid below:



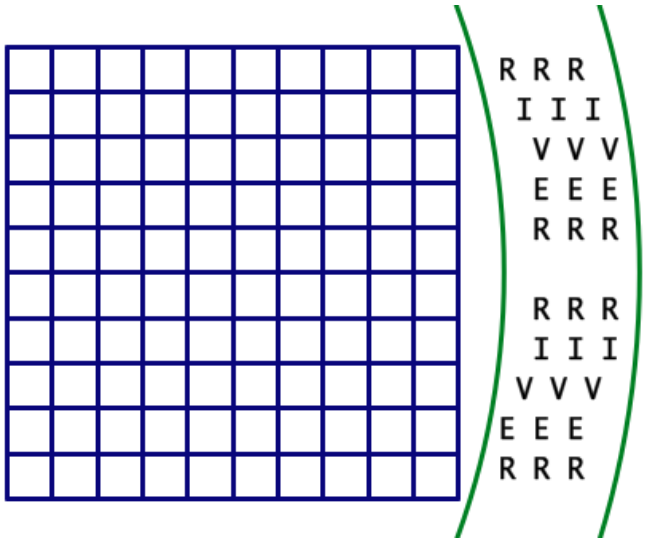
2. *Simple Random Sample.* Feeling uneasy about his plot selection, the farmer talks to his daughter who is taking statistics at Wheatridge High School, and asks if she could suggest a better choice of 10 plots to harvest early and use to estimate his total yield at the end of the growing season. His daughter suggests three other methods. The first is a simple random sample. Use your calculator or a random number table to choose 10 random plots, and mark them on the grid below:



3. *Stratified Sample, Vertical.* For this method, consider the field as grouped in vertical columns (called strata). Using your calculator or a random number table, randomly choose one plot from each vertical strata and mark these plots on the grid.



4. *Stratified Sample, Horizontal.* For this method, consider the field as grouped in horizontal rows (also called strata). Using your calculator or a random number table, randomly choose one plot from each horizontal strata and mark these plots on the grid.



5. The actual harvest yields per plot are shown in the table below (Of course, the farmer does not have access to this information ahead of time, or there would be no need use sample plots.):

6	17	20	38	47	55	69	76	82	97
7	14	23	34	45	56	63	75	81	92
2	14	28	30	50	50	62	80	85	96
9	15	27	34	43	51	65	72	88	91
4	15	28	32	44	50	64	76	82	97
5	16	27	31	48	59	69	72	86	99
5	18	28	34	50	60	62	75	90	90
8	15	20	38	40	54	62	77	88	93
7	17	29	39	44	53	61	77	80	90
7	19	22	33	49	53	67	76	86	97

Compare the different sampling methods discussed by the farmer and his daughter, and record your findings in the table.

Method	Mean yield per plot	Estimate of total yield
convenience sample		
simple random sample		
vertical strata		
horizontal strata		

6. You have looked at four different methods of choosing plots. Is there a reason, other than convenience, to choose one method over the other?

7. How did your estimates vary according the different sampling methods you used?

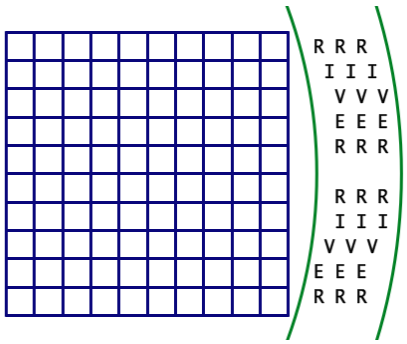
8. Compare your results to the rest of the class and discuss your findings.

9. Pool all the results from the class and make a boxplot of the data for the simple random sample, which we will call SRS. Do the same for the vertical strata and the horizontal strata. Discuss the results?

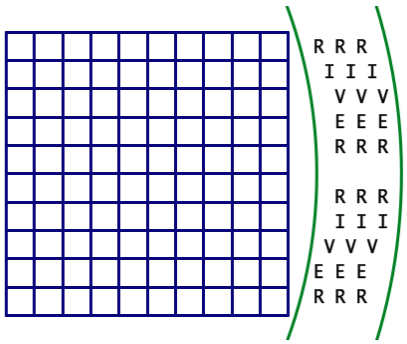
10. Which sampling method is best for the farmer?

11. What was the actual yield of the farmer's field, and how did the boxplots relate to this value?

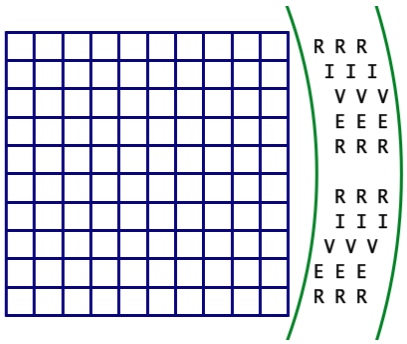
12. A year has gone by, and the farmer has installed an irrigation system to try and even out the yields in his field. He and his daughter decide to sample his plots using a SRS, and vertically and horizontally stratified samples. Repeat the process of generating a list of plots to sample for each method, and in each case, mark your chosen plots on the grids below:



SRS



Stratified Sample (Vertical)



Stratified Sample (Horizontal)

13. The actual harvest yields per plot post irrigation are shown in the table below:

79	81	95	69	65	59	88	65	66	91
80	75	88	80	82	66	76	99	62	61
97	50	92	92	91	84	75	85	63	89
99	71	55	75	65	66	66	86	96	50
57	95	51	79	98	71	70	86	89	76
57	53	90	71	50	76	56	91	85	64
69	95	98	90	93	97	79	95	73	90
58	99	75	51	67	81	55	63	89	74
98	62	73	54	50	76	91	50	90	55
91	59	69	59	71	72	85	85	86	97

Again, compare the different sampling methods and record your findings in the table.

Method	Mean yield per plot	Estimate of total yield
simple random sample		
vertical strata		
horizontal strata		

14. Compare the class box plots of the sample means obtained from the three sampling methods. Discuss with the class.

15. Based on the results of both the initial sampling and the post irrigation sampling under what conditions is it more useful to use stratified sampling? Random sampling?

Anscombe's Quartet

Questions:

Statistics is a relatively new topic of research and application. Statistics is a very active and dynamic area of research and application.

In this Activity:

Students will read a mathematical article from 1973, and consider an important data set called Anscombe's Quartet.

Materials:

You will need the following table of data:

Anscombe's Quartet

X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Procedure:

1. Read the Introduction to the article.

2. Use the data table to create your own scatter plots of the data and discuss with your classmates.

3. Read the rest of the article for homework and see how Anscombe's ideas compare with your and your classmates' ideas.

Graphs in Statistical Analysis

F. J. Anscombe

The American Statistician, Vol. 27, No. 1. (Feb., 1973), pp. 17-21.

Stable URL:
http://links.jstor.org/sici?sici=0003-1305%28197302%2927%3A1%3C17%3AGISA%3E2.0.CO%3B2-J

The American Statistician is currently published by American Statistical Association.

---

tion required (e.g. moments, estimation, application) can be cross-classified, as they are common to all distributions.

REFERENCES

1. Box, G. E. P. (1949). *Biometrika*, **36,** 317–346.
2. Box, G. E. P. and Andersen, S. L. (1955). *J. Roy Statist. Soc. Ser. B*, **17,** 1–26.
3. Bowman, K. O. and Shenton, L. R. (1965, 1966). Reports K-1633, K-1643, ORNL-4005, Union Carbide Corporation.
4. Haight, F. A. (1967). *Handbook of the Poisson Distribution*, New York: John Wiley and Sons.
5. James, A. T. (1954, 1960, 1964). *Ann. Math Statist.*, **25,** 40–75; **31,** 151–8; **35,** 475–97.
6. Johnson, N. L. and Kotz, S. (1969, 1970, 1972). *Distributions in Statistics*, Vol. I (Discrete). Vols. II and III (Continuous Univariate), Vol. IV (Continuous Multivariate), New York: John Wiley and Sons.
7. Kotz, S. and Johnson, N. L. (1969). *Distribution Theory in Statistical Literature*, Proc. 37th Session of the ISI, 303–305.
8. Lancaster, H. O. (1969). *The Chi-squared Distribution*, New York: John Wiley and Sons.
9. Milton, R. C. (1969). Computer Implementation of Distribution Function Algorithms. *Proc. Conference on Statistics and Computers*, University of Wisconsin, Madison, pp. 181–198.
10. Sichel, H. S. (1947). *J. Roy. Statist. Soc. Ser. A*, **110,** 337–47; (1949) *Biometrika*, **36,** 404–25.
11. Weiss, L. and Wolfowitz, J. (1966, 1968). *Teoriya Veroyatnostei i ee Primeneniya*, **11,** 68–93; **13,** 657–662. (English version of the journal: **11:** 58–81; **13,** 622–627.)

Graphs in Statistical Analysis*

F. J. ANSCOMBE**

Graphs are essential to good statistical analysis. Ordinary scatterplots and "triple" scatterplots are discussed in relation to regression analysis.

1. *Usefulness of graphs*

Most textbooks on statistical methods, and most statistical computer programs, pay too little attention to graphs. Few of us escape being indoctrinated with these notions:

(1) numerical calculations are exact, but graphs are rough;

(2) for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;

(3) performing intricate calculations is virtuous, whereas actually looking at the data is cheating.

A computer should make *both* calculations *and* graphs. Both sorts of output should be studied; each will contribute to understanding.

Graphs can have various purposes, such as: (i) to help us perceive and appreciate some broad features of the data, (ii) to let us look behind those broad features and see what else is there. Most kinds of statistical calculation rest on assumptions about the behavior of the data. Those assumptions may be false, and then the calculations may be misleading. We ought always to try to check whether the assumptions are reasonably correct; and if they are wrong we ought to be able to perceive in what ways they are wrong. Graphs are very valuable for these purposes.

Good statistical analysis is not a purely routine matter, and generally calls for more than one pass through the computer. The analysis should be sensitive both to peculiar features in the given numbers and also to whatever background information is available about the variables. The latter is particularly helpful in suggesting alternative ways of setting up the analysis.

Thought and ingenuity devoted to devising good graphs are likely to pay off. Many ideas can be gleaned from the literature, of which a sampling is listed at the end of this paper. In particular, Tukey [7, 8] has much to say on the topics presented here.

A few simple types of statistical analysis are now considered.

2. *Regression analysis—the simplest case*

Suppose we have values for one "dependent" variable y and one "independent" (exogenous, predictor) variable x. Before anything else is done, we should scatterplot the y values against the x values and *see* what sort of relation there is—if any. Many different kinds of things can happen:—

(1) the (x, y) points lie nearly on a straight line;

(2) the (x, y) points lie nearly on a smooth curve, not a straight line;

(3) the y-values are scattered, without relation to the x-values;

(4) something intermediate between (1) or (2) and (3);

(5) most of the (x, y) points lie close to a line or smooth curve, but a few are scattered a long way away.

Case (5) is particularly interesting, because there is an effect to be noticed, but the ordinary calculations for linear regression may miss it. Whenever we see "outliers", it is usually wise first to check that the

* Prepared in connection with research supported by the Army, Navy, Air Force and NASA under a contract administered by the Office of Naval Research.

** Dept. of Statistics, Yale Univ., Box 2179, Yale Station, New Haven, Conn. 06520.

values used really are correct, that is, not copied wrongly nor obviously faulty in some way. Then, if we are satisfied that these readings are authentic, we may perhaps set them aside for special study, and fit a regression relation to the remainder of the data. Special study of the outliers may prove very rewarding.

Case (1) would usually be considered ideal. Case (2) can sometimes be brought back to case (1) by transforming the x-scale or the y-scale or both.

The ordinary least-squares regression calculation is based on the following theoretical description or "model": the given number pairs (x_i, y_i) are related by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad (i = 1, 2, \ldots, n), \quad \text{(A)}$$

where β_0 and β_1 are constants and the "errors" $\{\epsilon_i\}$ are drawn independently from a "normal" (Gauss-Laplace) probability distribution having zero mean and constant variance. The regression calculation leads to estimates b_0 and b_1 for β_0 and β_1, to the "fitted values"

$$\hat{y}_i = b_0 + b_1 x_i,$$

and to the "residuals"

$$e_i = y_i - \hat{y}_i.$$

The sum of squares of the latter, generally called the "residual sum of squares" or "error sum of squares", leads to an estimate of the variance of the distribution of errors. If the theoretical description were exactly correct (and all calculation were exact, without round-off error), these calculations would be entirely satisfactory, in the sense that b_0, b_1 and the residual sum of squares, together with the number of readings n and the first two moments of the x-values, would constitute sufficient statistics for the unknowns and could substitute for the original data for all purposes with no loss of information. In practice, we do not know that the theoretical description is correct, we should generally suspect that it is not, and we cannot therefore heave a sigh of relief when the regression calculation has been made, knowing that statistical justice has been done.

After the regression calculation, the residuals $\{e_i\}$ should be plotted against the $\{x_i\}$. One might think this would show nothing that could not be seen in the original plot of $\{y_i\}$ against $\{x_i\}$. However, the residual plot will probably have a larger scale for the ordinates, and with the linear regression removed the residual behavior is easier to see. Usually it is a good idea to specify that the residual plot should be of the residuals $\{e_i\}$ against the fitted values $\{\hat{y}_i\}$, rather than $\{x_i\}$, with the *same scale* for ordinates and abscissas. This plot, besides showing how the residuals behave in relation to the x-values, also from its overall shape shows at a glance the relative dispersion of fitted values and residuals. In the decomposition

$$y_i = \hat{y}_i + e_i$$
$$(\text{observation} = \text{fitted value} + \text{residual}),$$

hopefully the fitted values follow the observations closely and have a greater variability than the residuals. One should be aware of their relative contributions.

If the theoretical description of the observations were exactly true, the residuals would appear to be normally distributed with zero mean and common variance, the same for all x-values. [That statement is not quite correct, but near enough for most practical purposes. The residuals would usually not have exactly equal variances, and they would be variously correlated.] Things to look for in a plot of $\{e_i\}$ against $\{\hat{y}_i\}$ or $\{x_i\}$ are:—

(1) a few of the residuals much larger in magnitude than all the others—outliers;
(2) a curved regression of residuals on fitted values;
(3) progressive change in the variability of the residuals as the fitted values increase;
(4) a skew (or other nonnormal) distribution of the residuals.

Sometimes, if we are lucky, effects (2), (3) and (4) can be removed simultaneously by a transformation of the scale in which y is expressed, as by taking logarithms. Alternatively, effect (2) may be allowed for by transforming the x-scale, or by adding another term on the right side of the theoretical description (A), making it perhaps

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i.$$

Instead of looking at a scatterplot of $\{e_i\}$ against $\{\hat{y}_i\}$, we could detect effects such as those just listed by calculating suitable test statistics, and we could assess their significance. But the plot shows a variety of features quickly and vividly, and formal tests often seem unnecessary.

There is indeed another reason for examining a scatterplot of residuals against fitted values, that may be important even when there is no indication of inadequacy in the theoretical description (A). Possibly one (or a few) observations have x-values widely separated from the others, leading to (one or more) outliers among the fitted values. Even though (A) should seem to fit all observations satisfactorily, with no outliers among the residuals, we may feel less comfortable about postulating (A) and basing conclusions on it, than if there had been no greatly outlying fitted value. That is because an outlying x-value contributes much more to the determination of the regression coefficient than other x-values. If an observation with an outlying x-value were affected by some special circumstance, not common to other observations, our fitted regression relation might be misleading. Often the y-value corresponding to an outlying x-value could be altered considerably without much effect on the goodness of fit of the regression relation but with marked effect on the estimated relation itself. We are usually happier about asserting a regression relation if the relation is still apparent after a few observations (any ones) have been deleted—that is, we are happier if the regression relation seems to permeate all the observations and does not derive largely from one or two.

All these various features that can so greatly change the significance we attach to a calculated regression are

invisible if we see only the usual quadratic summaries—the regression line, the analysis of variance, the multiple correlation coefficient R^2.

3. *An example*

Some of these points are illustrated by four fictitious data sets, each consisting of eleven (x, y) pairs, shown in the table. For the first three data sets the x-values are the same, and they are listed only once.

Data set		1-3	1	2	3		4	4
Variable		x	y	y	y		x	y
Obs. no.	1 :	10.0	8.04	9.14	7.46	:	8.0	6.58
	2 :	8.0	6.95	8.14	6.77	:	8.0	5.76
	3 :	13.0	7.58	8.74	12.74	:	8.0	7.71
	4 :	9.0	8.81	8.77	7.11	:	8.0	8.84
	5 :	11.0	8.33	9.26	7.81	:	8.0	8.47
	6 :	14.0	9.96	8.10	8.84	:	8.0	7.04
	7 :	6.0	7.24	6.13	6.08	:	8.0	5.25
	8 :	4.0	4.26	3.10	5.39	:	19.0	12.50
	9 :	12.0	10.84	9.13	8.15	:	8.0	5.56
	10 :	7.0	4.82	7.26	6.42	:	8.0	7.91
	11 :	5.0	5.68	4.74	5.73	:	8.0	6.89

TABLE. Four data sets, each comprising 11 (x, y) pairs.

Each of the four data sets yields the same standard output from a typical regression program, namely

Number of observations $(n) = 11$
Mean of the x's $(\bar{x}) = 9.0$
Mean of the y's $(\bar{y}) = 7.5$
Regression coefficient (b_1) of y on $x = 0.5$
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x} = 110.0$
Regression sum of squares = 27.50 (1 d.f.)
Residual sum of squares of $y = 13.75$ (9 d.f.)
Estimated standard error of $b_1 = 0.118$
Multiple $R^2 = 0.667$

These calculations express in various (redundant) ways the sufficient statistics for the theoretical description (A), when that is assumed to be correct. Some typical computer programs also yield a print-out of the residuals, in the order in which the data were entered. Since in the present case the data have been listed in a random order, probably little would be seen if the eye were run down such a print-out (especially if it were in abominable floating-point notation).

The data sets are graphed in the figures, together with the fitted line. Figure 1, corresponding to data set 1, is the kind of thing most people would see in their mind's eye, if they were presented with the above calculated summary. The theoretical description (A) seems to be perfectly appropriate here, and the calculated summary seems fair and adequate. Figure 2 suggests forcefully that data set 2 does not conform with the theoretical description (A), but rather y has a smooth curved relation with x, possibly quadratic, and there is little residual variability. Figure 3 similarly suggests that (A) is not a good description for data set 3: all but one of the observations lie close to a straight line (not the one yielded by the standard regression calculation), namely

$$y = 4 + 0.346\,x;$$

and one observation is far from this line. Those are the essential facts that need to be understood and reported.

Figure 4, like Figure 1, shows data apparently conforming well with the theoretical description (A). If all observations are considered genuine and reliable, data set 4 is just as informative about the regression relation as data set 1; there is no reason to prefer either to the other. Yet in most circumstances we should feel that there was something unsatisfactory about data set 4. All the information about the slope of the regression line resides in one observation—if that observation were deleted the slope could not be estimated. In most circumstances we are not quite sure that every observation is reliable. If any one observation were discredited and therefore deleted from data set 1, the remainder would tell much the same story. That is not so for data set 4. Thus the standard regression calculation ought to be accompanied by a warning that one observation has played a critical role.

Each of data sets 2, 3, 4 illustrates a peculiar effect in an extreme form. In less extreme forms such effects are often encountered in statistical analysis. For an example of the last effect: in a study (to be published

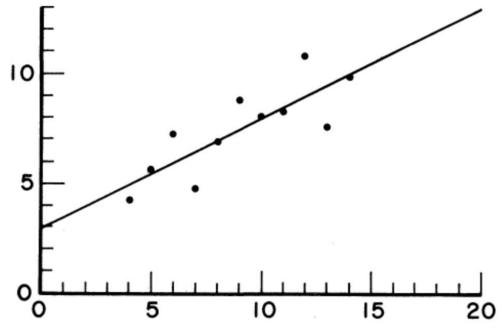


Figure 1

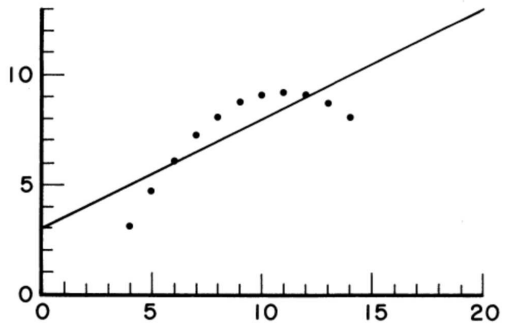


Figure 2

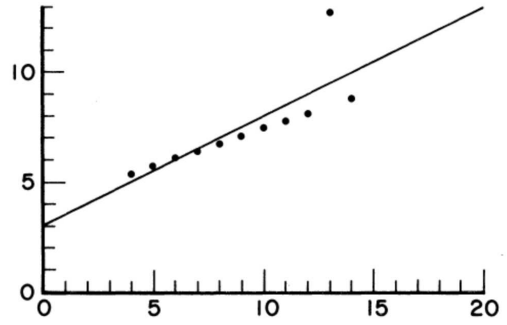


Figure 3

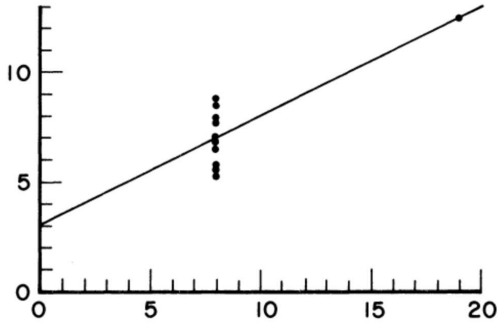


Figure 4

elsewhere) of per capita expenditures on public school education in each of the fifty states of the Union, together with the District of Columbia, it was found that the expenditures had a satisfactory linear regression on three likely predictor variables, with multiple R^2 about 0.7 and well behaved residuals. However, one of the states, namely Alaska, was seen to have values for the predictor variables rather far removed from those of the other states, and therefore Alaska contributed rather heavily to determining the regression relation. Of course Alaska is an abnormal state, and the thought immediately occurs that perhaps Alaska should be excluded from the study. But there are other extraordinary states, Hawaii, the District of Columbia (counted here as a state), California, Florida, New York, North Dakota, . . . Where does one stop? Rather than merely exclude Alaska, a preferable course seems to be to report the regression relation when all states are included, but add that Alaska has contributed heavily and say what happens if Alaska is omitted—the regression relation is not greatly changed, but the standard errors are increased somewhat and multiple R^2 is reduced below 0.6. We need to understand *both* the regression relation visible in all the data *and also* Alaska's special contribution to that relation.

4. *More general regression analysis*

Much of what has been said about regression of one dependent on one independent variable applies to more general regression analyses. Suppose there is one dependent variable y but two "independent" variables $x^{(1)}$ and $x^{(2)}$, so that the theoretical description reads

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \epsilon_i, \qquad \text{(B)}$$

where the β's are constants and the ϵ's are distributed as before.

We cannot simply make on a two-dimensional surface a three-dimensional plot of y against $x^{(1)}$ and $x^{(2)}$ simultaneously. There are indeed expensive visual devices for suggesting such a thing. If we confine ourselves to what can be done with a line printer or typewriter terminal, there are two approaches to visualizing relations between the three variables y, $x^{(1)}$ and $x^{(2)}$, before any regression calculation.

(a) Make ordinary scatterplots of the three pairs of variables, y against $x^{(1)}$, y against $x^{(2)}$, $x^{(1)}$ against $x^{(2)}$. The third of these shows whether it will be possible to distinguish the effects of $x^{(1)}$ and $x^{(2)}$ on y. For if $x^{(1)}$ and $x^{(2)}$ are closely related to each other, one of them having a regression (not necessarily linear) on the other with little residual variation, then any apparent relation of y with $x^{(1)}$ and $x^{(2)}$ may perhaps be expressible equally well as a relation of y with either $x^{(1)}$ or $x^{(2)}$ alone.

(b) Make a scatterplot of two of the variables, say $x^{(1)}$ and $x^{(2)}$, marking each point by a symbol that roughly indicates the value of the third variable y. The values of the third variable can be coded numerically, say by dividing the range into ten intervals and representing values, according to the interval they fall in, by single digits, 0, 1, 2, . . . , 9—or possibly by dividing the range into not more than twenty-six intervals and using letters of the alphabet. Alternatively, values of the third variable may be coded by symbols whose physical appearance (size and blackness) indicates magnitude—for example, with an APL typeball, symbols of increasing weight such as

. ∘ ○ ⊖ □ ⊠

or these representing steps from large-negative to large-positive (M standing for minus and P for plus)

M M - ∘ + P P

Such a plot is equivalent to an ordinary scatterplot of the first two variables and also indicates, well enough for many purposes, how the third variable is related to the other two. This kind of plot will be called a *triple scatterplot* (TSCP).

After the ordinary regression calculations have been made, yielding the regression coefficients b_0, b_1, b_2, the fitted values and the residuals, the single most useful plot is an ordinary scatterplot of residuals against fitted values, preferably on the same scale. Interpretation is as indicated before.

Another possibility is to make a TSCP, taking as the first two variables the contributions of $x^{(1)}$ and $x^{(2)}$ to the fitted values, that is, plot $\{b_1 x_i^{(1)}\}$ against $\{b_2 x_i^{(2)}\}$ on the same scale, with the residuals $\{e_i\}$ coded as the third variable. This plot can show association of residual behavior with $x^{(1)}$ and $x^{(2)}$ individually.

To study the dependence of y on one of the independent variables, say $x^{(1)}$, with the effect of the other eliminated, one may scatterplot $\{y_i - b_2 x_i^{(2)}\}$ against $\{x_i^{(1)}\}$. This would be useful in planning a transformation of the $x^{(1)}$-scale.

When we pass from a regression problem with only two "independent" variables to one with many, we find it harder to see all that is going on by looking at graphs. But that is as it should be—the possibilities are now so much greater. The likelihood that we fool ourselves by *only* carrying out some ordinary regression calculations is much greater too. Usually when there are many "independent" variables they are mutually related and we are interested in performing regression on subsets of them, possibly by a "stepwise" procedure; so even the standard calculation is not so simple.

In any case, whenever a regression calculation has been carried out, whether on all the "independent" variables or on a subset of them, it will be useful to see a simple scatterplot of residuals against fitted values (on the same scale).

If the independent variables are separated into two sets, we may be interested to see a TSCP, in which the two coordinates represent the contributions of each of the two sets to the fitted values (on the same scale) and the plotting code represents the residuals.

5. *Two-way tables*

The analysis of a two-way table by calculating row means, column means, residuals and what R. A. Fisher called the analysis of variance, may be regarded as a special instance of regression analysis. The structure is now sufficiently rich that graphical presentation in advance of numerical calculation is probably not too useful. But after the calculations the same sorts of graphical treatment as for ordinary regression have the same effectiveness. Residuals may be scatterplotted against fitted values on the same scale. Row effects can be plotted against column effects, on the same scale, in a TSCP with coded residuals. (It was Tukey's elegant use of a kind of TSCP for two-way tables that introduced me to the idea; see Chapter 16 in [7].) If the rows or columns have a meaningful natural order, the residuals should also be presented in that order.

Rectangular tables (crossclassifications) in two or more dimensions, with some modes of classification perhaps "nested" rather than "crossed", are of common occurrence. Whenever any set of main effects and interactions has been calculated, the residuals should be scatterplotted against the fitted values, and various sorts of TSCP may be interesting.

This article is emphatically not a catalog of useful graphical procedures in statistics. Its purpose is merely to suggest that graphical procedures are useful. Only two types of graph have been mentioned, the ordinary scatterplot and the triple scatterplot, and these have been considered in only one sort of context (regression). There are other types of graphs and display devices that can make quantitative relations visible and comprehensible, and other sorts of statistical tasks than regression.

6. *Implementation*

Graphical output such as described above is readily available to anyone who does his own programming. I myself habitually generate such plots at an APL terminal, and have come to appreciate their importance. A skilled Fortran or PL/1 programmer, with an organized library of subroutines, can do the same (on a larger scale).

Unfortunately, most persons who have recourse to a computer for statistical analysis of data are not much interested either in computer programming or in statistical method, being primarily concerned with their own proper business. Hence the common use of library programs and various statistical packages. Most of these originated in the pre-visual era. The user is not showered with graphical displays. He can get them only with trouble, cunning and a fighting spirit. It's time that was changed.

REFERENCES

[1] Anderson, Edgar, "A semigraphical method for the analysis of complex problems," *Proceedings of the National Academy of Sciences*, 13 (1957), 923–927. Reprinted in *Technometrics*, 2 (1960), 387–391.

[2] Andrews, D. F., "Plots of high-dimensional data," *Biometrics*, 28 (1972), 125–136.

[3] Bachi, Roberto, *Graphical Rational Patterns*, Jerusalem: Israel Universities Press, 1968.

[4] Bertin, Jacques, *Sémiologie Graphique*, Paris: Mouton and Gauthier-Villars, 1967.

[5] Daniel, Cuthbert, "Use of half-normal plots in interpreting factorial two-level experiments," *Technometrics*, 1 (1959), 311–341.

[6] Draper, N. R., and Smith, H., *Applied Regression Analysis*, New York: Wiley, 1966.

[7] Tukey, John W., *Exploratory Data Analysis*, limited preliminary edition, three volumes, Reading: Addison-Wesley, 1970–71.

[8] Tukey, John W., "Some graphic and semigraphic displays," *Statistical Papers in Honor of George W. Snedecor* (ed. T. A. Bancroft), Ames: Iowa State University Press, 1972, pp. 293–316.

[9] Tukey, John W., and Wilk, M. B., "Data analysis and statistics: techniques and approaches," *The Quantitative Analysis of Social Problems* (ed. E. R. Tufte), Reading: Addison-Wesley, 1970, pp. 370–390.

[10] Wilk, M. B., and Gnanadesikan, R., "Probability plotting methods for the analysis of data," *Biometrika*, 55 (1968), 1–17.

[11] "Statistical analysis, special problems of, I. Outliers, II. Transformations of data," *International Encyclopedia of the Social Sciences*, Macmillan and Free Press, 1968, vol. 15, pp. 178–193.

Cereal Box Problem

Before we begin:

This lab is based on the following article:

The Cereal Box Problem Revisted Jesse L. M. Wilkins, *Virginia Polytechnic Institute and State University School of Science and Mathematics* (Volume 99(3), March 1999),

http://eric.ed.gov/?id=EJ590348

Questions:

Fastbreak Cereal Company wanted to promote their Star Oats cereal, so they made Harry Potter figurines, Harry, Ron, Hermione, Hagrid, Dumbledore and Draco, and they placed a character in each box of Star Oats. You can image what a stir this caused. Assuming equal chances of getting any of the six characters, (i.e. Fastbreak did not withhold all of the boxes with Dumbledore from shipping to increase demand for its cereal beyond the increase certainly guaranteed by the insertion of prizes in its packages.) how many boxes of Star Oats should any kid expect to have to buy, open, consume in order to get a full set of six characters?

In this Activity:

Students will be introduced to Monte Carlo simulation, the process of simulating an event that is difficult, expensive or otherwise impossible to affect. They will explore the answer to the above question and discover an empirical answer. They will also explore ways to understand and generate a theoretical answer to the question of how many boxes of cereal one must open to collect all six prizes.

Materials:

You will need one die for each two students and pencils.

Procedure:

1. Break into teams of two, pick a die and assign one of the six numbers to each of the Harry Potter characters. This die will allow us to model buying, opening and consuming boxes of cereal without actually buying, opening and consuming boxes of Star Oats.

Characters' numbers

Character	Harry	Ron	Hermione	Hagrid	Dumbledore	Draco
Number						

2. Have one person roll the die (buy, open and consume boxes of Star Oats) until they obtain all 6 numbers (characters). Using the table below, the other person can record the numbers as they are obtained and after they have all been obtained, count the number of rolls. Divide the work of generating 100 simulated trials amongst the pairs in your class.

Trial Outcomes

Trail	**1**	**2**	**3**	**4**	**5**	**6**	**N**
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							

3. When your class is done collecting the data, find the average number of purchases required to obtain all six characters. This average is the expected number of purchases or the expected outcome. Discuss this number with your classmates.

4. If you have a TI89 or TI84, ask your instructor for the procedure which will simulate as many rolls of the die as you would like and find the average value of all of the simulations. This is a Monte Carlo method. Give it a try, and then discuss your results with your classmates.

5. The theoretical answer to the question of how many boxes of cereal a person should expect to open in order to find all 6 prizes is 14.7. How does this answer compare with your findings?

6. If you want, you can modify the procedure given to help you understand the smaller components that go into answering the complicated question posed in this lab. Once you have opened one box and found a prize, how many boxes can you expect to open before you find a second different prize?

7. There are many approaches to determining the theoretical answer to the question posed in this lab. Here is an outline of one way. First, the first box must be opened, and a prize obtained, call it P. Once that box has been opened, the chance that a new prize is in the next box is $\frac{5}{6}$, and the chance that you have to open 2 boxes obtain a new prize is $\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)$ because you have 1/6 of a chance of getting P again, and then 5/6 of a chance of getting a prize other than P. This goes on for a possibly infinite number of times, but the chance of needing to open yet another box decreases significantly after a certain point. If we multiply the number of boxes opened after the first one by the probability of finding our second new prize in that one, and add all of these numbers together, we obtain the expected number of boxes. It looks like this:

$$1\left(\frac{5}{6}\right) + 2\left(\frac{5}{6}\right)\left(\frac{1}{6}\right) + 3\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)^2 + \cdots + = \sum_{n=1}^{\infty} n\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)^{n-1}.$$

And if we perform the same calculation for the third new prize, the fourth new prize, the fifth new prize and the final prize, we find that the expected number of prizes is:

$$1 + \sum_{n=1}^{\infty} n\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)^{n-1} + \sum_{n=1}^{\infty} n\left(\frac{4}{6}\right)\left(\frac{2}{6}\right)^{n-1} + \sum_{n=1}^{\infty} n\left(\frac{3}{6}\right)\left(\frac{3}{6}\right)^{n-1} + \sum_{n=1}^{\infty} n\left(\frac{2}{6}\right)\left(\frac{4}{6}\right)^{n-1} + \sum_{n=1}^{\infty} n\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^{n-1}.$$

We will use what we know about convergent geometric series and differentiation to compute this value. You can see that the general form of each of the above sums is:

$$S = \sum_{n=1}^{\infty}(1-r)n(r)^{n-1} = A\sum_{n=0}^{\infty}(n+1)r^n,$$

where $0 < r < 1$ and $A = 1 - r$. And understanding that these series are convergent, we have:

$$S = A\sum_{n=0}^{\infty} r^n(n+1) = A\frac{d}{dr}\sum_{n=0}^{\infty} r^{n+1} = A\frac{d}{dr}\frac{r}{1-r} = A\frac{1}{(1-r)^2}.$$

Now remember that $A = 1 - r$ to see that $S = 1/(1-r)$. Applying this fact 5 times above, we obtain:

$$1 + 6/5 + 6/4 + 6/3 + 6/2 + 6/1 = 14.7.$$

Nine-Block

Before we begin: The words of the authors of this activity:

In sum, the 9-block form is designed to scaffold student insight into the systematicity and rigor of combinatorial analysis by serving as a template that helps students initially see, create, and organize the combinatorial space.

This lab was adapted from:

There Once Was a 9-Block ... - A Middle-School Design for Probability and Statistics. Abrahamson, D., Janusz, R. M., and Wilensky, U. *Journal of Statistics Education [Online]* (2006).

www.amstat.org/publications/jse/v14n1/abrahamson.html

Questions:

How many different ways are there of filling up a nine-block grid with blue and green blocks? How would you count the number of ways?

In this Activity:

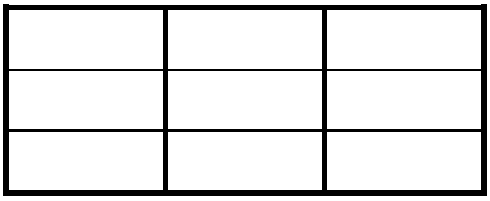
In this lab students create 9-blocks and discuss ways to count and record all possibilities. The goal is to create a tower of 9-blocks, and then have that as a visual from which to draw conclusions and test hypotheses.

Materials:

9-block grid sheets, colored pencils, crayons or dot markers.

Procedure:

1. Fill in each of the squares below with blue or green. Compare your square with your neighbor.



2. On the grid below draw 5 different possible blue and green 9-blocks. Compare your squares with your neighbor's squares. How many are the same? How many are different?

3. How many possible 9-blocks can be made? How many different 9-blocks has your class made so far?

4. Together devise a strategy for creating all 9-blocks.

5. Together devise a strategy for displaying all 9-blocks.

6. How many different 9-blocks have 4 blue squares?

7. How many different 9-blocks have 6 blue squares?

8. How many different 9-blocks have a blue square in the top left-hand corner?

9. How many different 9-blocks have a blue square in each of the corners?

10. HOMEWORK: Make and display all 4-block possibilities and write three probability questions you might use your display to answer.

Homework, Activities and Exercises

1. Confirm that the five points in the table all lie on a single line. Write an equation for the line. Use your calculator to make a *scatter plot*, and graph the line on the same system of axes.

x	y
-3	7
-2	5
-1	3
0	1
1	-1

2. Given the line whose equation is $y = 2x + 3$ and the points $A = (0, 0)$, $B = (1, 9)$, $C = (2, 8)$, $D = (3, 3)$, and $E = (4, 10)$, do the following:
(a) Plot the line and the points on the same axes.
(b) Let A' be the point on the line that has the same x-coordinate as A. Subtract the y-coordinate of A' from the y-coordinate of A. The result is called the *residual of* A.
(c) Calculate the other four residuals.
(d) What does a residual tell you about the relation between a point and the line?

3. The table at right shows data that Morgan collected during a 10-mile bike ride that took 50 minutes. The cumulative distance (measured in miles) is tabled at ten-minute intervals.
(a) Make a scatter plot of this data. Why might you expect the data points to line up? Why do they not line up?
(b) Morgan's next bike ride lasted for 90 minutes. Estimate its length (in miles), and explain your method. What if the bike ride had lasted t minutes; what would its length be, in miles?

time	*dist*
0.0	0.0
10.0	2.3
20.0	4.4
30.0	5.7
40.0	8.2
50.0	10.0

4. Let $P = (1.35, 4.26)$, $Q = (5.81, 5.76)$, $R = (19.63, 9.71)$, and $R' = (19.63, y)$, where R' is on the line through P and Q. Calculate the residual value $9.71 - y$.

5. (Continuation)
(a) Given that $Q' = (5.81, y)$ is on the line through P and R, find y. Calculate $5.76 - y$.
(b) Given that $P' = (1.35, y)$ is on the line through Q and R, find y. Calculate $4.26 - y$.
(c) Which of the three lines best fits the given data? Why do you think so?

6. Verify that $P = (-1.15, 0.97)$, $Q = (3.22, 2.75)$, and $R = (9.21, 10.68)$ are not collinear.
(a) Let $Q' = (3.22, y)$ be the point on the line through P and R that has the same x-coordinate as Q has. Find y, then calculate the *residual* value $2.75 - y$.
(b) Because the segment PR seems to provide the most accurate slope, one might regard PR as the line that best fits the given data. The point Q has as yet played no part in this decision, however. Find an equation for the line that is parallel to PR and that makes the sum of the three residuals zero. In this sense, this is the *line of best fit*.

7. Given $T = (1.20, 7.48)$, $U = (4.40, 6.12)$, and $V = (8.80, 2.54)$, find an equation for the line that is parallel to the line TV and that makes the sum of the three residuals zero. This line is called the *zero-residual line* determined by T, U, and V.

1. Consider the points $A = (-0.5, -8)$, $B = (0.5, -5)$, and $C = (3, 4.5)$. Calculate the residual for each of these points with respect to the line $4x - y = 7$.
2. Show that the zero-residual line of the points P, Q, and R goes through their centroid.

3. (Continuation) The zero-residual line makes the sum of the residuals zero. What about the sum of the *absolute values* of the residuals? Is it possible for this sum to be zero? If not, does the zero-residual line make this sum as small as it can be?

4. Let $P = (2, 6)$, $Q = (8, 10)$, and $R = (11, 2)$. Find an equation for the zero-residual line, as well as the line of slope 2 through the centroid G of triangle PQR. Find the sum of the residuals of P, Q, and R with respect to the second line. Repeat the investigation using the line of slope -1 through G. Use your results to formulate a conjecture.

5. What is the sine of the angle whose tangent is 2? First find an answer *without* using your calculator (draw a picture), then use your calculator to check.

6. Consider the line $y = 1.8x + 0.7$.
(a) Find a point whose residual with respect to this line is -1.
(b) Describe the configuration of points whose residuals are -1 with respect to this line.

7. The *median* of a set of numbers is the middle number, once the numbers have been arranged in order. If there are two middle numbers, then the median is half their sum. Find the median of **(a)** 5, 8, 3, 9, 5, 6, 8; **(b)** 4, 10, 8, 7.

8. A *median-median point* for a set of points is the point whose x-value is the median of *all* the given x-values and whose y-value is the median of *all* the given y-values. Find the median-median point for the following set of points: $(1, 2)$, $(2, 1)$, $(3, 5)$, $(6, 4)$, and $(10, 7)$.

9. The table shows the population of New Hampshire at the start of each of the last six decades.

year	*pop*
1960	606921
1970	746284
1980	920610
1990	1113915
2000	1238415
2010	1316472

(a) Write an equation for the line that contains the data points for 1960 and 2010.
(b) Write an equation for the line that contains the data points for 2000 and 2010.
(c) Make a scatter plot of the data. Graph both lines on it.
(d) Use each of these equations to predict the population of New Hampshire at the beginning of 2020. For each prediction, explain why you could expect it to provide an accurate forecast.

10. The zero-residual line determined by $(1, 2)$, $(4, k)$, and $(7, 8)$ is $y = x - \frac{2}{3}$. Sketch the line, plot the points, and find the value of k. Be prepared to explain your method.

1. Plot the following nine non-collinear points:
$(0.0, 1.0)$ $(1.0, 2.0)$ $(2.0, 2.7)$ $(3.0, 4.0)$ $(4.0, 3.0)$ $(5.0, 4.6)$ $(6.0, 6.2)$ $(7.0, 8.0)$ $(8.0, 8.5)$
(a) Use your ruler (clear plastic is best) to draw the line that seems to best fit this data.
(b) Record the slope and the y-intercept of your line.

2. (Continuation) Extend the zero-residual-line technique to this data set as follows: First, working left to right, separate the data into three groups of equal size (three points in each group for this example). Next, select the *summary point* for each group by finding its median-median point. Finally, calculate the zero-residual line defined by these three summary points. This line is called the *median-median line*. Sketch this line, and compare it with your estimated line of best fit.

3. (Continuation) If the number of data points is not divisible by three, the three groups cannot have the same number of points. In such cases, it is customary to arrange the group sizes in a symmetric fashion. For instance:
(a) Enlarge the data set to include a tenth point, $(9.0, 9.5)$, and then separate the ten points into groups, of sizes three, four, and three points, reading from left to right. Calculate the summary points for these three groups.
(b) Enlarge the data set again to include an eleventh point, $(10.0, 10.5)$. Separate the eleven points into three groups and calculate the three summary points.

4. An avid gardener, Gerry Anium just bought 80 feet of decorative fencing, to create a border around a new rectangular garden that is still being designed.
(a) If the width of the rectangle were 5 feet, what would the length be? How much area would the rectangle enclose? Write this data in the first row of the table.
(b) Record data for the next five examples in the table.
(c) Let x be the width of the garden. In terms of x, fill in the last row of the table.

width	*length*	*area*
5		
9		
16		
22		
24		
35		
x		

(d) Use your calculator to graph the rectangle's area versus x, for $0 \leq x \leq 40$. As a check, you can make a scatter plot using the table data. What is special5about the values $x = 0$ and $x = 40$?
(e) Comment on the symmetric appearance of the graph. Why was it predictable?
(f) Find the point on the graph that corresponds to the largest rectangular area that Gerry can enclose using the 80 feet of available fencing. This point is called the *vertex*.

5. Using only positive numbers, add the first two odd numbers, the first three odd numbers, and the first four odd numbers. Do your answers show a pattern? What is the sum of the first n odd numbers?

6. It is often convenient to use what is called *sigma notation* to describe a series. For example, the preceding sum could have been written $\sum_{k=0}^{n}(2k+1)$ Use sigma notation to describe the sum of the first n even integers.

1. Jan had the same summer job for the years 1993 through 1996, earning \$250 in 1993, \$325 in 1994, \$400 in 1995, and \$475 in 1996.
(a) Plot the four data points, using the horizontal axis for "year". You should be able to draw a line through the four points.
(b) What is the slope of this line? What does it represent?
(c) Which points on this line are meaningful in this context?
(d) Guess what Jan's earnings were for 1992 and 1998, assuming the same summer job.
(e) Write an inequality that states that Jan's earnings in 1998 were within 10% of the amount you guessed.

2. The height h (in feet) above the ground of a baseball depends upon the time t (in seconds) it has been in flight. Cameron takes a mighty swing and hits a bloop single whose height is described approximately by the equation $h = 80t - 16t^2$. Without resorting to graphing on your calculator, answer the following questions:
(a) How long is the ball in the air?
(b) The ball reaches its maximum height after how many seconds of flight?
(c) What is the maximum height?
(d) It takes approximately 0.92 seconds for the ball to reach a height of 60 feet. On its way back down, the ball is again 60 feet above the ground; what is the value of t when this happens?

3. Consider the triangular arrangements of hearts shown below:

(a) In your notebook, continue the pattern by drawing the next triangular array.
(b) Let x equal the number of hearts along one edge of a triangle, and let y equal the corresponding number of hearts in the whole triangle. Make a table of values that illustrates the relationship between x and y for $1 \leq x \leq 6$. What value of y should be associated with $x = 0$?
(c) Is the relationship between x and y linear? Explain. Is the relationship quadratic? Explain.
(d) Is y a function of x? Is x a function of y? Explain.
(e) The numbers 1, 3, 6, 10, ... are called *triangular numbers*. Why? Find an equation for the triangular number relationship. Check it by replacing x with 6. Do you get the same number as there are hearts in the 6^{th} triangle?

1. These next few problems show how to compute the slope and intercept of the line that best fits the data. Suppose we knew the equation of the line of best fit, $y = ax + b$, then we could compute the residuals of our 26 data with respect to this line:

$$S = \sum_{k=1}^{n} (y_k - ax_k - b)^2.$$

First show how to arrive at the expanded expression:

$$S = \sum_{k=1}^{n} y_k^2 + a^2 x_k^2 + b^2 - 2ax_k y_k - 2by_k + 2abx_k$$

Next, since we are working with a finite number of finite values, we can rewrite the big sum as a sum of sums:

$$S = \sum_{k=1}^{n} y_k^2 + a^2 \sum_{k=1}^{n} x_k^2 - 2a \sum_{k=1}^{n} x_k y_k - 2b \sum_{k=1}^{n} y_k + 2ab \sum_{k=1}^{n} x_k + n(b^2)$$

Remember that all of the x_k and y_k come from our data points and are given. The unknowns in this equation are a and b! Rewrite the equation using the shorthand notation $\sum_{k=1}^{n} x_k = S(x)$, $\sum_{k=1}^{n} x_k^2 = S(x^2)$, $\sum_{k=1}^{n} y_k = S(y)$, $\sum_{k=1}^{n} y_k = S(y^2)$ and $\sum_{k=1}^{n} x_k y_k = S(xy)$.

2. Now rearrange the terms so that your equation has the form $S = Aa^2 + Ba + C$.

$$S = S(x^2)a^2 + (2bS(x) - 2S(xy))\, a + \left(S(y^2) - 2bS(y) + nb^2\right).$$

3. And now, use what you know about quadratic equations to find the a-value that minimizes this quadratic function.

4. Notice that your equation for a includes a b. Rewrite S as a quadratic function in b instead of a.

$$S = nb^2 + (2aS(x) - 2S(y))\, b + \left(S(y^2) + a^2 S(x^2) - 2aS(xy)\right)$$

5. And find the value of b that minimizes S. (Is this the same minimum value that you obtained in a previous problem?) Now you have two equations in two unknowns, and you can solve for a and b.

1. Use your equations, the ones we got are:

$$a = \frac{nS(xy) - S(x)S(y)}{nS(x^2) - (S(x))^2} \quad \text{and} \quad b = \frac{S(x) - aS(x)}{n},$$

to calculate a and b for our data set. Remember that you already have a column for x, x^2 and y in your calculator. You may want to add a column with xy. Compare with what the calculator gave you for the values of a and b.

2. Use your calculator to find the equation of the least-squares line (LinReg) for the five data points (2.0,3.2), (3.0,3.5), (5.0,5.0), (7.0,5.8), and (8.0,6.0). Let G be the *centroid* of these points — its x-coordinate is the average of the five given x-coordinates, and its y-coordinate is the average of the five given y-coordinates. Verify that G is on the least-squares line.

3. The table at right shows how many seconds are needed for a stone to fall to Earth from various heights (measured in meters). Make a scatter plot of this data. Explain how the data suggests that the underlying relationship is not linear.
(a) Calculate the squares of the times and enter them in a third column. A scatter plot of the relation between the first and third columns does suggest a linear relationship. Use LinReg to find it, letting x stand for height and y stand for the square of the time.
(b) It is now easy to write a nonlinear relation between h and t by expressing t^2 in terms of h. Use this equation to predict how long it will take for a stone to fall from a height of 300 meters.

height	*time*
10	1.42
20	2.01
30	2.46
40	2.85
50	3.18
60	3.49
70	3.76
80	4.02
90	4.27
100	4.50

4. (Continuation) This time, calculate the *square roots* of the heights and enter them in a new column. A scatter plot of the relation between the second column and the new column should reveal a linear relationship. Find it, then use it to *extrapolate* how much time is needed for a stone to fall 300 meters.

5. The following are data about how charged Sasha's laptop is:

Percent Charge

Time	9:11 am	9:27 am	9:36 am	9:48 am	9:55 am	10:08 am	10:17 am
Charge	41 %	56 %	64 %	74 %	79 %	86 %	91 %

a. What are the variables in this story?

b. Make a scatterplot of these data.

c. Draw a line that best fits the data and find its equation.

d. Interpret the slope and y-intercept in this context.

e. Based on your findings, when can Sasha expect to have a fully charged battery?

Standard Normal Probabilities

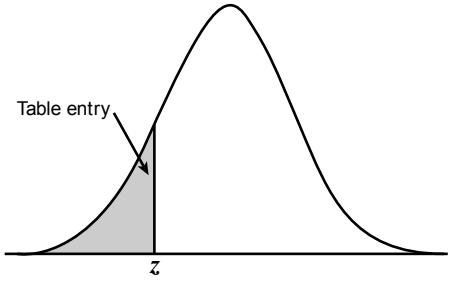


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
–1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
–1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
–1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
–1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
–1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
–1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
–0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
–0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
–0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
–0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
–0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
–0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
–0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
–0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
–0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
–0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

TABLE B

Random digits

Line								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335
106	68417	35013	15529	72765	85089	57067	50211	47487
107	82739	57890	20807	47511	81676	55300	94383	14893
108	60940	72024	17868	24943	61790	90656	87964	18883
109	36009	19365	15412	39638	85453	46816	83485	41979
110	38448	48789	18338	24697	39364	42006	76688	08708
111	81486	69487	60513	09297	00412	71238	27649	39950
112	59636	88804	04634	71197	19352	73089	84898	45785
113	62568	70206	40325	03699	71080	22553	11486	11776
114	45149	32992	75730	66280	03819	56202	02938	70915
115	61041	77684	94322	24709	73698	14526	31893	32592
116	14459	26056	31424	80371	65103	62253	50490	61181
117	38167	98532	62183	70632	23417	26185	41448	75532
118	73190	32533	04470	29669	84407	90785	65956	86382
119	95857	07118	87664	92099	58806	66979	98624	84826
120	35476	55972	39421	65850	04266	35435	43742	11937
121	71487	09984	29077	14863	61683	47052	62224	51025
122	13873	81598	95052	90908	73592	75186	87136	95761
123	54580	81507	27102	56027	55892	33063	41842	81868
124	71035	09001	43367	49497	72719	96758	27611	91596
125	96746	12149	37823	71868	18442	35119	62103	39244
126	96927	19931	36089	74192	77567	88741	48409	41903
127	43909	99477	25330	64359	40085	16925	85117	36071
128	15689	14227	06565	14374	13352	49367	81982	87209
129	36759	58984	68288	22913	18638	54303	00795	08727
130	69051	64817	87174	09517	84534	06489	87201	97245
131	05007	16632	81194	14873	04197	85576	45195	96565
132	68732	55259	84292	08796	43165	93739	31685	97150
133	45740	41807	65561	33302	07051	93623	18132	09547
134	27816	78416	18329	21337	35213	37741	04312	68508
135	66925	55658	39100	78458	11206	19876	87151	31260
136	08421	44753	77377	28744	75592	08563	79140	92454
137	53645	66812	61421	47836	12609	15373	98481	14592
138	66831	68908	40772	21558	47781	33586	79177	06928
139	55588	99404	70708	41098	43563	56934	48394	51719
140	12975	13258	13048	45144	72321	81940	00360	02428
141	96767	35964	23822	96012	94591	65194	50842	53372
142	72829	50232	97892	63408	77919	44575	24870	04178
143	88565	42628	17797	49376	61762	16953	88604	12724
144	62964	88145	83083	69453	46109	59505	69680	00900
145	19687	12633	57857	95806	09931	02150	43163	58636
146	37609	59057	66967	83401	60705	02384	90597	93600
147	54973	86278	88737	74351	47500	84552	19909	67181
148	00694	05977	19664	65441	20903	62371	22725	53340
149	71546	05233	53946	68743	72460	27601	45403	88692
150	07511	88915	41267	16853	84569	79367	32337	03316

Glossary

Association Association describes the strength of the given relationship between two variables.

Back-To-Back Stem Plot Used to compare two sets of data. The leaves for one set of data are on one side of the stem, and the leaves for the other set of data are on the other side. See **stem and leaf plot**.

Back-To-Back Stem (and Leaf) Plot A graphic device to compare two data sets. The two sets share common stems. The leaves of one set go to the right, the leaves of the other go to the left. See **Stem (and Leaf) Plot**.

Bar Graph A device to display categorical data using (usually) vertical bars, the heights of which represent the frequency of each category. The bars are separated, unlike a **histogram** (q. v.).

Bias Bias is the difference between an estimated expected value and the true value being estimated.

Bivariate data Ordered pairs of linked numerical observations.

Boxplot or Box-and-Whisker Plot A graphic device to display 1-variable data using the **minimum**, **maximum**, **median** and the upper and lower **quartiles** of the data.

Center A one-digit summary of a data set. Usually the mean or median, sometimes the mode.

Conditional Probability The probability that A will occur given the knowledge that B has occurred. The probability of A given B, denoted $Pr(A|B)$, is the probability of (A and B)/(probability of B).

Dot Plot A device to display one-variable data. Each data point is represented (usually) by a single dot. The number of dots corresponding to a value is the **frequency** of that value.

Empirical Probability An estimate of the probability of an event occurring based on a large number of trials.

Event A set of possible outcomes from a random situation.

Expected Value The weighted average of all possible outcomes of an event, the weights being the probabilities of each outcome, denoted $E(X) = \sum x_i p(x_i)$. .

Experiment A procedure to verify or refute a hypothesis. A random, controlled experiment allows valid inferences to be drawn. Compare Observational Study.

Five-Number Summary The minimum, maximum, median and quartiles of a set of observations; the skeleton of a boxplot.

Form The given data can look linear or non-linear, and its form is the basic function that describes the look of the data.

Frequency The number of occurrences of a particular value. Denoted by the height of the bars in a bar graph or histogram.

Histogram a device to display data by bars, the height of which represents the frequency of that observation or group of observations. The bars are drawn without gaps, the width of each bar being the same and representing individual values of a group of values of the same size.

Independent Events A and B are independent events if the probability of A and B occurring is the product of the probabilities of A and the probability of B. Knowing that one of the events has occurred has no effect on the probability of the other event occurring. For example, drawing two cards from a shuffled deck, noting the color of one and replacing it in a deck before drawing the second are independent events. If the first card is not replaced, the two events are not independent.

Inference The drawing of conclusions, usually based on a **sample**. The process of extrapolating information from a sample to the parent **population**.

Influenctial Point A data point for which a small change in position will have a disproportionate effect on the least squares residual line.

Interquartile Range $Q_3 - Q_1$ A measure of the variability of a set of data.

Least Squares Regression Line The Least Squares Regression Line is the line that makes the sum of squares of the **residuals** with respect to that line as small as possible.

Maximum The greatest value of a set.

Mean The arithmetic of the values of all the observations. Geometrically the mean is the balance point of a data set.

Mean Absolute Deviation A measure of the variability of a set computed by taking the mean of the absolute values of the difference between each observation and the median or the mean of the set. For the set $X = \{x_1, x_2, \ldots x_n\}$ the MAD is $\frac{1}{n}\sum_{i=1}^{n} |x_i - m(X)|$, where $m(X)$ is the mean or median of the set X.

Median In an ordered list of observations the middle number if there is an odd number of observations and the mean of the two middle numbers if there is an even number of observations. Half of the observations lie above the median, half below. Geometrically the median divides the data into two equal areas.

Minimum The least value of a set.

Mode The value of the most commonly occurring observation. A useful measure of the center of the distribution of categorical data. The **modal** value.

Normal Distribution The normal distribution with mean equal to μ and standard deviation equal to σ is given by $y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$.

Observational Study A method of collecting data from which it not appropriate to draw inferences.

Outcome The result of a random situation.

Outlier An observation that lies outside the overall pattern of the data. In a boxplot an outlier is arbitrarily defined to be a value that lies either 1.5 interquartile ranges below the lower quartile or 1.5 interquartile ranges above the upper quartile. A value which invites further study.

Parameter A parameter is a numerical value that states something about an entire population and is often hard to obtain.

Population The entire set of people or objects about which information is sought.

Quartile For a data set with an even number of observations, the **First** or **Lower** quartile, denoted Q_1 , is the **median** of the observations in the lower half of the set, the 25^{th} percentile. The **Upper** or **Third** quartile, the 75^{th} percentile, denoted Q_3, is the median of the observations in the upper half of the set.

Residual The residual of a point with respect to a line is the vertical distance between the y-value of the point and the y-value of the line both given at the x-value of the point of interest.

Sample A subset of a population. The goal of inference is to extrapolate information gleaned from a sample (about which everything is known) to the parent population.

Sample Space The set of all possible outcomes of a chance process. The sum of the probabilities of all the outcomes is 1 or 100 %.

Shape Some of the words used to describe the shape of a one-variable distribution are symmetrical, mound-shaped, skewed left or right, uniform, bimodal, fan-shaped.

Spread The variability of the observations in a data set. Some measures of spread are range, interquartile range, standard deviation, mean absolute deviation.

Standard Deviation is a measure of the variability in a data set. It is used almost exclusively in conjunction with the mean to summarize approximately normally distributed data. Every calculating or computing device will compute the standard deviation of a data set. It is computed as the square root of the mean of the sum of the squares of the deviations (signed differences) of each data point from the mean. For the set $X = \{x_1, x_2, \ldots x_n\}$ the standard deviation is $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m(X))^2}$, where $m(X)$ is the mean of the set X.
Compare Mean Absolute Deviation.

Statistic A statistic is a numerical value that states something about a sample of a population and can be different for different samples.

Statistics Statistics are more than one statistic.

Stem Plot, Stem-And-Leaf Plot A graphing technique that displays all the individual values in a set. The principal values, which vary for each set, form the stems which are arranged vertically. A stemplot is quickly drawn. The leaves extend horizontally from the stems. When rotated 90 degrees counterclockwise a stem plot looks just like a histogram.

Strength The strength of the pattern is an assessment of how tightly clustered the data points are around the underlying form.

Subjective Probability An estimate of the probability of an event occurring determined from an (educated) guess.

Theoretical Probability The mathematically determined probability of an event occurring.

Trend The pattern displayed by the data.

Reference Materials

Morris Code, Scrabble and the Alphabet. Mary Richardson, John Gabrosek, Diann Reischman, Phyllis Curtiss, *Grand Valley State UniJournal of Statistics Education Volume 12, Number 3* (2004).
www.amstat.org/publications/jse/v12n3/richardson.html

The Journal of Statistics Education An International Journal on the Teaching and Learning of Statistics, Editor of JSE Michelle Everson,
www.amstat.org/publications/jse/jse_users.html

An Exercise in Sampling: Rolling Down the River Doetsch, Flanangan-Hyde, Harrison, Tabor, Tiberio, *NCSSM Statistics Leadership Institute* (July 2000).
courses.ncssm.edu/math/Stat_inst01/PDFS/river.pdf.

There Once Was a 9-Block ... - A Middle-School Design for Probability and Statistics. Abrahamson, D., Janusz, R. M., and Wilensky, U. *Journal of Statistics Education [Online]* (2006).
www.amstat.org/publications/jse/v14n1/abrahamson.html

What is the Probability of a Kiss? (It's Not What You Think) Mary Richardson, Susan Haller, *Journal of Statistic Education Volume 10, Number 3* (2002).
www.amstat.org/publications/jse/v10n3/haller.html

The Cereal Box Problem Revisted Jesse L. M. Wilkins, *Virginia Polytechnic Institute and State University School of Science and Mathematics (Volume 99(3))* (March 1999)
http://eric.ed.gov/?id=EJ590348

Engaging Students in a Large Lecture: An Experiment using Sudoku Puzzles Brophy and Hahn (2014) *Journal of Statistics Education* Volume 22, Number 1, www.amstat.org/publications/jse/v22n1/brophy.pdf.

Graphs in Statistical Analysis F. J. Anscombe *The American Statistician, Vol. 27, No. 1* (Feb. 1973), pp.17-21.

illustrativemathematics.org

This site lists all the CCSSM standards in their complete wording. It also gives many examples of exercises which illustrate the standards. A very valuable resource.

Middle Grades Mathematics Project Probability Addison Wesley, 1986, ISBN 0-201-21478-4

Navigating Through Data Analysis in Grades 6-8 NCTM, Reston VA, 2003, ISBN 987-0-87353-547-2.

This is an excellent resource, written long before the CCSSM appeared. It contains lots of good examples and a CD.

Navigating Through Data Analysis in Grades 9-12, NCTM.

Workshop Statistics, Discovery with Data and Fathom, A Rossman, B Chance, R Lock, *Key Curriculum Press*, 2001. ISBN 1-930190-07-7.

This can be used as a stand-alone statistics text. It comes in versions for Fathom, the graphing calculator, Excel, Minitab.

Very highly recommended.

Statistics From Data To Decision A Watkins, R Scheaffer, G Cobb. *Wiley* 2011, ISBN 978-0470-45851-8.

This is only one of a collection of high school texts designed to prepare students for the AP Test. It contains masses of data, excellent examples and a wealth of supplementary material.

Math 1 Phillips Exeter Academy teachers, 1998

Math 2 Phillips Exeter Academy teachers, 1998

Math 3 Phillips Exeter Academy teachers, 1998

Math 4 Phillips Exeter Academy teachers, 1998