

Web-based Resources and Activities for Teachers of AP Statistics

from the
NCSSM Statistics Leadership Institute 2000

This packet contains:

- Introduction to the use of the internet in AP Statistics
- Annotated list of websites, including URLs
- Ready-to-use web-based activities

Joe Joyner
Norview High School
Norfolk, Virginia
jjoyner@pinn.net

Rachel Levy
Carolina Friends School
Durham, North Carolina
raylevy@netpath.net

Mary Ellen Noyes
Mary Institute and Saint Louis Country Day School
Saint Louis, Missouri
mnoyes@micds.org

Jeane Swaynos
Seminole High School
Sandord, Florida
swaynos@aol.com

Table of Contents

Ways to Use the Internet in AP Statistics.....	page 3
Websites – Sources of Data.....	page 4
Websites -- Simulations, Demonstrations, and Analysis.....	page 5
Websites – Archives and Resources.....	page 6
Welcome to the Wild, Wired, Wonderful World of Statistics.....	page 7
Web-based Activity: Bookmark, internet search	
Central Limit Theorem with Simulation.....	page 10
Web-based Activity: JAVA applet, animation	
Polling for the President.....	page 12
Web-based Activity: Create a polling company	
Using an Applet on Power.....	page 13
A script to assist teachers using this applet in class	
Does Seeding Clouds Produce More Rain?.....	page 17
Web-based Activity: DASL, WebStat, transforming data, 2-sample t-test	
Canadian Crime.....	page 24
Web-based Activity: Statistics Canada, distributions in two-way tables	
A Nation of Movers? Exploring Categorical Data with the Calculator.....	page 27
Web-based Activity: General Social Survey data, chi-squared analysis	
Let’s Buy a Diamond Ring! Exploring Regression on a Computer.....	page 35
Web-based Activity: JSE data archive, statistical software	
Fit For Life - Fast Food Nutrition Comparison	page 44

Ways to Use the Internet in AP Statistics

Data

Teachers can find current, interesting data for projects, assignments, and tests by visiting favorite sites. Our list points teachers to some well-known sources.

Downloading data from websites into statistical software or word processors is not as easy as we might like. In some cases one can simply copy and paste. In other cases one must copy from the website, paste into a word processor, highlight the data, and use something like “Convert text to table.” Once the data are in a table, it is generally easy to copy and paste into the statistical software. But commas within numbers can cause problems. Sometimes one must delete all the commas before pasting the data into the statistical software.

The main problem with moving from a website to a statistical package is that the data appear to be in columns but actually one copies them as lines. In general, ASCII data are easier to copy and paste between different software packages.

Class Demonstrations

Some websites include Java applets that illustrate particular topics. These are often easier and quicker to use than a graphing calculator. A teacher with a projection system can run the demonstration or simulation and discuss the topic with the whole class. These are good to use particularly after students have had some experience with the topic, by either reading or practicing with their calculators.

Independent Assignments

Teachers with a computer lab available during class time can use full-period activities that include finding data, analyzing it using a statistical package, and creating a document or other product that illustrates the work. We include some sample activities on different topics.

Additionally, some websites have interactive applets that students can work through on their own, either during class or as a homework assignment.

Websites – Sources of Data

The **Data and Story Library (DASL)** has an extensive collection of datasets on a wide variety of topics. The interface is extremely friendly. Each dataset has a clear description, and the stories include brief analysis. Some of the topics are beyond the AP Statistics curriculum. The raw data are easily imported into statistical applications. The datasets tend to be a few years old. <http://lib.stat.cmu.edu/DASL/>

The **Journal of Statistics Education** has a data archive that includes datasets and the reference for the article that used them. The datasets are on a variety of subjects; the famous draft lottery dataset is included. The site also has links to datasets from some texts. <http://www.amstat.org/publications/jse/archive.htm>

The **U. S. Census Bureau** site has a variety of information. Some of the reports have tables of data useful for statistics classes. Teachers can use the “State and County Quick Facts” facility to obtain the latest census information for their state or county. The site includes housing and economic information as well as counts of residents. <http://www.census.gov/>

Statistics Canada is a good source of a variety of data, mainly on Canada and its citizens. Many of the datasets are free and displayed in tables. This is a user-friendly site. <http://www.statcan.ca/start.html>

The **Centers for Disease Control and Prevention** has a website that includes datasets and reports. The CDC Wonder database has such sets as mortality and AIDS cases. You can access it as an “Anonymous User.” The datasets generally appear as tables, but you may also find an option to have them displayed in ASCII format.

<http://www.cdc.gov/scientific.htm>

The **General Social Survey** website gives access to questions and results from the survey that started in 1972. The survey covers an extremely wide variety of topics; use the subject index to find specific questions. Many of the questions are asked on more than one survey, so one can compare the responses in different years. This is a good source for categorical data and for work with proportions. Many textbooks mention this survey. <http://www.icpsr.umich.edu/GSS99/>

The **Bureau of Labor Statistics** website has recent data on employment information, prices, and productivity. Some of the datasets are good examples of time series. Under the Data heading, the “Most Requested Series” is a good place to begin. <http://stats.bls.gov/blshome.htm>

Websites – Simulations, Demonstrations, and Analysis

WebStat is a basic statistical software package. For classes without software, or for students working outside of school, it is an excellent option. The sample data sets are good beginnings, or students can enter their own data. In addition, students can copy data from other websites and paste them into WebStat, following the directions from the Help page. WebStat does all the basic operations in exploratory data analysis and inference. Printing the graphs and the numerical output is possible, but some users might have trouble because of their own set-up. WebStat is an excellent way to introduce students to a statistical software package if your classroom has only calculators.
<http://www.stat.sc.edu/webstat/>

The **Rice University Virtual Statistics Lab** has extensive resources for a statistics course. It includes an online statistics text (including homework and assessment resources), a number of illustrative simulations and demonstrations, case studies with analysis, and software to analyze data. The software has the data from the case studies already loaded. The simulations are Java applets. Some of the material goes beyond the AP Statistics syllabus. The applets are short activities suitable for classroom demonstration or individual work; students could gain from them in 10 minutes or less. They are indexed by statistical topic.
<http://www.ruf.rice.edu/~lane/rvls.html>

The **VESTAC** site has simulations that demonstrate statistical concepts graphically. These simulations are appropriate for classroom demonstrations, but probably not as assignments for independent work for high school students. Instructors would need to explain the topic and use the simulation as a visual tool. <http://www.kuleuven.ac.be/ucs/java/index.htm>

On a page from the **University of Illinois** site, under a statistics course, is a game to guess the correlation of different scatterplots. This is suitable for either a class activity or an independent assignment. Students will quickly get a feel for how the value of the correlation coefficient varies with the shape of a scatterplot. Teachers should be aware that sometimes correlation coefficients given as choices differ by very small amounts and are very difficult to correctly discern. <http://www.stat.uiuc.edu/~stat100/java/GCApplet/GCAppletFrame.html>

The **University of South Carolina Department of Statistics** has an applet that demonstrates the effect of adding one point on a regression line. The applet is suitable for either a class demonstration or a directed activity. If one moves the added point along a line, for instance, one can watch how the slope of the regression line varies. Teachers might ask students questions such as, “Find the point that makes the correlation a minimum.”
<http://www.stat.sc.edu/~west/javahtml/Regression.html>

Websites – Archives and Resources

The American Statistical Association website has information on the professional organization's mission and activities. The ASA Center for Statistics Education has a section on statistical education and special resources for K-12 teachers. <http://www.amstat.org/> for educators:

Texas Instruments is a starting point for activities and articles about a variety of mathematical topics that have applications on the graphing calculator. Click on the *Educators* link under Starting Points. Click on the *High School* link under Mathematics. Scroll down to the *Classroom Activities* section. A variety of math topics are available--each check mark is a link. Click on the desired link depending on the topic and TI calculator type that you wish to use. Under *Statistics/Prob* alone, there are approximately 40 links to activities that have been submitted by high school teachers. <http://www.ti.com/calc/docs>

The Math Archives is an anthology of web site links to things mathematical. Under *Teaching Materials*, click on K-12 Teaching Materials. Next, click on *Lesson Plans*, then *Topics in Mathematics*. From this point, you may choose links to *Combinatorics*, *Statistics* or other topics. Once in Statistics, you can also link to *Probability Theory*, or other topics as desired. <http://archives.math.utk.edu/>

A list of statistics resources on the web can be found at **The University of San Diego Statistics Education Project** page. <http://www.acusd.edu/statpage/links.htm>

Welcome to the Wild, Wired, Wonderful World of Statistics!

Objectives:

1. Use a browser to enter a URL and find a given internet site.
2. Use a web based tutorial to review internet and search engine basics.
3. Create and save a file of bookmarks using a word processor; cut and paste.
4. Save the bookmark file to appropriate place.
5. Conduct an internet search using a search engine.
6. Quote a web page.
7. Email a document.

(In the blanks below, fill in the information for the computer you are using.)

Using a Browser to enter a URL

Turn on your computer and open the Browser.

The browser our school uses is called _____

To find the browser, _____

At the top of the browser is a place to type in internet addresses, called URL's.

Find a tutorial called "Finding Information on the Internet" by entering this URL exactly:

<http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/FindInfo.html>

Using a Web-based tutorial

Beginners: Look through Tutorial Part I. Learn and practice 5 new things. Write them down. **All students:** Look through Tutorial Part II. Learn and practice 5 new things. Write them down.

Creating a file of bookmarks

Our word processor is called _____

To run the word processor and create a document _____

To name and save the document (do this now!) _____

Go to the browser window and highlight the URL. Copy the URL using the copy command (CTRL C on a PC; apple C on a Macintosh).

Go to the word processing window and paste the URL using the paste command (CTRL V on a PC; apple V on a Macintosh). Give the URL a name and even a description so that you will remember where it leads.

Now, go back to the tutorial and make bookmarks for each of the 5 recommended search engines.

Using a Search Engine

You tell a search engine what you want to know about by using search terms. Search terms are words or phrases that the search engine looks for in web pages. It is important to pick search terms that are specific enough to get the right information and general enough that you get results.

Wide search → general words → lots of results → hard to pick the right site to view

Narrow search → specific words → few results → few sites to choose from

Search Engine Practice

Suppose you wanted to find out about places to hike in Minnesota. First, ask yourself if “places to hike in Minnesota” is the best thing to type into a search engine. Is it too general? Too specific? Wrong terms? Is there a better way to get to where you want? Maybe you want to find a general site about tourism in Minnesota and then look for trails. Maybe you want to find a map and find an area that looks interesting. Should you think about where hiking trails usually exist? What types of sites would have information on hiking?

Pick a search engine and try it out. Then use a different search engine and compare your results.

More practice: Find out whether any Connecticut college teams won a league championship in ice hockey last year. Hint: You probably need to find final league standings and then answer the question.

Quotes From Web Pages

If you ever find something on a web page that you want to quote in your document, just highlight it, copy (CTRL or apple-C), move to the word processor and paste (CTRL or apple-V). You must enclose the item in quotes and cite the reference.

Emailing documents

If you want to email a document to someone (such as your teacher), you have two options. In a new message, you can use INSERT FILE and the file will be an attachment. Otherwise cut and paste from your document to the mail message. You might lose some formatting with this method.

Homework assignment:

Suppose you are preparing a report on the death penalty in the United States. You want to be able to discuss which states have a death penalty and how often they use it. You also want to address issues of gender and race. So you need the data. Where are they?

Use the search engines to conduct the appropriate searches. Turn in a document that includes which search engines you used, how you searched and where the best results were found (URL). Copy the results (or just some of them if they are extensive). Your document should be no more than one page. Make sure you identify which material is yours and which you copied. E-mail the document to me by _____. Please put your name and the date on the document.

Central Limit Theorem with Simulation

Objective

The objective of this activity is to have students investigate the distribution of sample means in preparation for the study of the Central Limit Theorem. The student will explore sample means from various populations (normal, skewed, uniform) and discover how the size of the sample affects the distribution of the sample means.

GO TO: <http://www.ruf.rice.edu/~lane/rvls.html>

**Choose: Simulations/Demonstrations Sampling Distribution.
Read the information concerning Sampling Distribution Simulation.
Read the instructions for Sampling Distribution and press “BEGIN.”**

The display will show four graphs. The first graph is the parent function. The choice will display NORMAL. Move the arrow key and find the other choices of graphs:

The second graph is “Sample Data”. If you want your sample to be displayed on the graph you must choose Animate Sample.

1. Click Animated Sample once. Describe what happens to graphs two and three. Repeat this for a total of 5 times. Describe what is happening to the graphs. Repeat 25 times and describe what is happening. On graph 2 choose 1000 samples and describe what is happening.
2. On graphs three and four there are choices for variables and the size of your sample could be specified. List the graph choices and the sample size choices:
3. Set graph 3 to MEAN with $N = 5$ and set graph 4 to “NONE”. Click Animated Sample. Explain what the program is displaying on graphs 2 and 3.
4. How do you clear the graph to start again? Clear the lower three graphs.

- Using a **NORMAL** distribution for the parent function complete the following activity:

On graph 3 choose Mean and $N = 5$. Choose 10,000 samples. Describe the graph of the sample means (graph 3).

- Change graph 3 to $N = 25$ and complete problem 5 again. Describe the graph of sample means (graph 3) in statistical terms.

- Compare the distribution of sample means for $N = 5$ and $N = 25$.

- Complete problems 5, 6, and 7 using a **UNIFORM** distribution. Give specific characteristics using statistical language.

- Complete problems 5, 6, and 7 using a **SKEWED** distribution. Give specific characteristics using statistical language.

- Given a normal, skewed, or uniform distribution with repeated sampling, what happens to the shape, center, and variation of the sampling distribution of sample means when you change the sample size?

Optional: Choose "Custom" on Graph 1, draw your own graph, and repeat the exercise.

For further exploration go back to the window containing the original instructions and click on "exercises".

Polling for the President

Your teacher is running for the office of President of the United States. Presidential candidates often use polls to gauge public opinion. There are many polling organizations that supply this sort of information. The candidates must decide which organization to use for their campaigns. The class will be divided into groups. Each group will create a polling company. Your group must convince the presidential candidate to hire your company.

Web Resources

www.census.gov
www.people-press.org
www.gallup.com
www.pollingreport.com
www.princeton.edu/~abelson/xsurvey.html

Criteria to consider

Population: whom you will be polling (such as age, location)
Possible strata for sampling (such as geographical region)
Types of data collected (what does the candidate want to know?)
Sampling: how data are collected
How data are analyzed
How data and analysis are presented
Possible biases in the data
Margin of error
Specific wording of survey questions

Presentation of Results

Your group will be allowed a 5-minute presentation (strict time limit) to advertise your company. You will also be allowed to leave a one-page pamphlet (8.5x11 piece of paper) with the president.

Grade Criteria

- | | |
|-----------|---|
| 20 points | Individual: Each group member must create a one-page report on one of the criteria for your company (only one person per criterion). |
| 20 points | Individual: Each group member should create a list of URLs for sites you used to learn about your criterion. Include a title for the site and a paragraph about the kinds of information that site provided. You are not restricted to the sites suggested above. |
| 20 points | Group: Five-minute presentation should involve all group members, not just one spokesperson. |
| 20 points | Group: One-page pamphlet should incorporate ideas from each of the individual tasks. Your pamphlet should look professional. It should be clear what company you represent. |
| 20 points | Group cooperation, communication and effective use of class time. |
| Bonus | Write a second one-page report on another criteria (10 points) |

Script for a class demonstration:

Using an Applet on Power in AP Statistics

Power is sometimes a difficult topic for statistics students to understand. Some students immediately understand the explanation and drawing in the text, but many need multiple experiences to fully grasp the concept. Teachers may find an applet a useful alternative to their own drawing on the board. This document describes how a teacher might use the applet in class. We assume the students have either read about power or heard the term prior to this demonstration.

One applet on power is located at

<http://www.projects.cgu.edu/wise/appletsf.shtml>

Click on the Introduction to Statistical Power applet *Introduction to Statistical Power*.

Initially the picture looks like the drawing of power shown in many textbooks. The hypotheses are

$$H_0 : \mu = 100$$

$$H_a : \mu > 100$$

The applet uses a z-test, so the standard deviation is known. On the right of the screen, you can see that $\sigma = 15$. The alternative value of μ (labeled μ_1) is 115 and α is 0.05. (You might have to either scroll down or enlarge your window to see the whole display.) The power is represented by the pink and dark blue region, and its area is 0.639. The alternate distribution is in red and the distribution under the null hypothesis is in blue. The value of β is 0.361, which is $(1 - \text{power})$. That is the area of the red region.

If you click on the label for each of these quantities, a text box appears that describes the quantity. For instance, if you click on σ , a box appears with the text, “sigma is the population standard deviation for both population distributions.” Click again to make it disappear. Note: the value of d is $\frac{\mu_1 - \mu_0}{\sigma}$, which is the standardized distance between the true mean and the hypothesized mean.

Starter questions to ask the students:

1. What is the blue curve?

The blue curve is the density curve for a normal distribution with mean 100 and standard deviation 15. This curve represents the population with mean specified by the null hypothesis.

2. What is the red (or green) curve?

The density curve for a normal distribution with mean 115 and standard deviation 15. This curve represents the population with the true mean.

3. What does the red dotted line represent?

The critical value, which is the value of the mean beyond which we would reject the null hypothesis for our value of alpha.

4. What would happen if we took a sample and got a value less than the dotted red line?

We would not reject the null hypothesis, and that would be a mistake. That's called Type II error.

5. What if we took a sample and got a value greater than the dotted red line?

We would reject the null hypothesis, which would be the right decision, since the true mean is 115.

We want to find out how likely it is that we reject the null hypothesis. The probability of doing that is called the **power** of the test.

6. Which part of the graph shows us the power?

The region under the red (green) curve to the right of the dotted line. It is shaded in dark blue and pink.

Notes to teacher:

The applet simulates taking a sample from the alternative distribution and computing its mean, and it gives results in the white box below the parameter values. Click on Sample and see what happens.

On the graph at the top of the window the four data values are shown, with a red arrow indicating their mean. The red arrow also appears in the main graph. In the box the value of the mean is given, along with the z score and the probability z is greater than that z score (the

area under the blue curve to the right of the red arrow). The interesting issue here is whether this sample leads to rejection of the null hypothesis. Since the alternative value is the true value, we want to reject the null hypothesis. If you keep pressing Sample, the applet repeatedly samples and shows the result. The main graph shows each of the sample means. You could take 10 samples and see how many result in rejection. Since the power is 0.639, you would expect to see about 6 of those 10 samples result in rejection.

Example dialog with the class:

Let's take a sample of 4 items from the alternative distribution and see if we reject the null hypothesis. [*Press Sample.*] Did we? *Answer depends on your sample.* What are the boxes and the arrows on the graph? *The arrow is the sample mean, the boxes show the individual values in the sample.*

If we do this 10 times, how many times do you think we will reject H_0 ? *About 6, on average. This is the power times 10.* Let's try it and see. [*Press Sample 9 more times.*] What are those boxes on the lower graph? *They are a histogram of the 10 sample means.* How many times did we correctly reject H_0 ? *Answer depends on your sample.*

After the initial discussion of the given display, you can start to change values and examine the results. The applet allows you to change some values simply by clicking the mouse, or else you can type in new values into the boxes and hit enter to update the display.

Here is an example of how you can investigate the effects of changing the various parameters on the power:

Let's suppose the true mean is actually 118, not 115. What do you think will happen to the power? *It will increase.* Let's see. [*Change the alternative value to 118 in the box next to the red μ_1 . Be sure to hit enter to update the display.*] How did the power change? *It increased to 0.775.* Why? *Since the true value is farther from μ_0 , we are more likely to get a sample that leads to rejection.* Let's try a few more values and see what happens. Make a general statement relating the power to the true value of the mean. *As the true value gets farther from μ_0 , the power increases. Note that if you make your true value below μ_0 , the power goes to the value of α . You can also vary the value of μ_1 by clicking on the distribution curve, holding down the mouse, and dragging.*

What do you think will happen to the power if we change σ ? Explain.

If σ is larger, the power will be smaller. If σ is smaller, the power will be larger. Students might not have a good feel for this initially, so you might go ahead and change σ and let them see what happens. Let's change σ to 10 and see what happens. When the alternative value is 118, the power increases to 0.975. The graph should show the curves taller and narrower. Unfortunately, in the window the curves don't change in shape, but the scales

on the axes change. What do you expect to see if σ is 20? Let's try it. The power drops to 0.562, when the alternative value is 118. The curves are wider and shorter, but you can only tell this from the axes. Make a general statement relating power and σ . Power and σ are inversely related. As one increases the other decreases.

So far we have kept our sample size at 4. What if it were 40? How is sample size related to variability of the mean? *Means from larger samples have smaller variability. Students might recall that the sampling distribution of \bar{x} has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, so as n increases this standard deviation decreases.* So what should happen to the shape of the normal curve representing the sampling distribution of means if we increase the sample size? *It should get taller and narrower.* Let's change the sample size to 40. What happened? *The curves do get taller and narrower. The power increases to 1.000 to three decimal places.* Let's try a few other values of n and see what happens. Make a general statement relating power and sample size. *With a few examples, it is clear that as sample size increases, so does the power.*

At this point, students should see the relationship between power and n , σ , and μ_1 . For further exploration, you can vary α and see that larger α 's lead to higher power. You might also check to see if the students have noticed that β is $1 - \text{power}$, and it is represented by the dark red shading in the main graph.

Teachers might also use this applet as an investigation for students to work through independently.

Does Seeding Clouds Produce More Rain?

An activity exploring data transformations and inference using WebStat

Objectives

Import data from DASL into WebStat

Perform exploratory data analysis on raw and transformed data

Perform two-sample t inference procedures on transformed data

Introduction for Instructors

This activity uses a DASL dataset on rainfall from clouds. The user will import the data into WebStat and then use WebStat for analysis. The dataset is particularly interesting as an example of when transformations are necessary prior to inference. Since most classes study inference months after transformations, this activity will provide an opportunity for review. In addition, teachers will want to discuss robustness and outliers when examining the data.

Below is a student activity. Comments for the teacher are included in italics. We have appended a student copy after the annotated one.

Procedure

Years ago, researchers wanted to determine whether seeding clouds with silver nitrate increased rainfall. The DASL website contains a dataset from two sets of randomly selected clouds, one that was seeded and the other that was not.

Go to the DASL site,

<http://lib.stat.cmu.edu/DASL/>

and search on data subjects. Look under Environment. Choose the Clouds Datafile.

Variables:

Unseeded_clouds: Amount of rainfall from unseeded clouds (in acre-feet)

Seeded_clouds: Amount of rainfall from clouds seeded with silver nitrate (in acre-feet)

Highlight everything under the line that says **The Data:**, and copy.

Now go to the WebStat site,

<http://www.stat.sc.edu/webstat/>

and when the orange button appears, click on it. Soon you will have a new window open; the top part resembles a spreadsheet.

Pasting data into WebStat can be a bit tricky, because the browser you use makes a big difference. The window you need to paste into sometimes comes up behind another window. Go to the WebStat Help menu (see Help on the WebStat menu bar) and load the help page. Scroll down to where you see a link to the instructions (under the data column) for loading data by pasting. If you follow the directions closely, you can paste the data into WebStat, and the headings should be included. Start by going to the **Data** menu and choosing **Paste data**.

*If you are using Netscape, you should be able to just paste the data in using the WebStat “Paste data” command. If you are using Internet Explorer, after you use the “Paste data” command, do NOT click anything on the foremost window. Go to the **back** window (which may be partly hidden but should extend as a white window horizontally) and paste. Click on Submit. Another window appears that says your data were successfully loaded. Close that window, and then on the remaining dialog box click OK.*

The first step in analysis, of course, is to look at the data. Go to the Graphics menu and choose one of the univariate data pictures, Boxplot. You can select both variables by clicking on each, and WebStat produces side-by-side boxplots. Describe the shapes of the two distributions.

The two plots show strong right skewness. [WebStat does not produce modified boxplots, so the outliers are not clearly identified.]

Now consider a second picture of the data, a stemplot (“stem and leaf plot”). What do the stemplots show us that the boxplots did not?

The stemplots identify outliers. The plots appear in the lower window. Note that the stemplots round to the tens place and truncate the rest, so the stems are the hundreds place and the leaves are the tens place. WebStat does not make back-to-back stemplots.

The plots made above indicate that these data are not normally distributed. Have WebStat make a normal probability plot to verify that these data are not normally distributed.

Students should find the QQ Plot on the Graphics menu. That is the same as a normal quantile plot or normal probability plot. Your text may use any of those terms, and you may need to guide the students to the plot.

The plots are dramatically NOT linear, so the data are not normally distributed.

The researchers wanted to find out if the seeding produced a significant increase in rainfall. Write the hypotheses for the appropriate significance test.

Here are the hypotheses:

$H_0: \mu_u = \mu_s$, where μ_u is the mean amount of rainfall in the unseeded clouds,
and μ_s is the mean amount of rainfall in the seeded clouds

$H_0: \mu_u < \mu_s$

What conditions must be satisfied before we apply our inference procedures? Are these conditions satisfied? Explain.

In order to perform the two-sample t test, we must have a random assignment of subject (cloud) to treatment group (seeded or unseeded). We appear to have that, according to the DASL site. [Your students might get caught up in wondering how they randomly selected clouds. That information is not on the DASL page. You might find the discussion useful, or you might want to refer the students to the reference and continue on.] We also need to have normally distributed data, or at least no strong skewness or outliers. The data do not satisfy that condition, so we should not proceed with the two-sample t -test.

One method to resolve this situation is to transform the data before performing inference. WebStat makes transformation easy. Go to the Data menu and choose Transform data. You have four options. Considering the skewness you noted earlier, which transformation would you expect to mitigate that? Explain your choice.

The first three options are reasonable choices. Both of the logarithmic transformations should bring the high outliers closer to the rest of the data. Taking the square root might have a similar effect, but perhaps not as strong. Squaring the data would make the high outliers even more pronounced.

Transform the data using a logarithmic transformation (you may specify which one you want your students to use) and the square root transformation. WebStat stores the transformed data in new columns. You can change the labels by going to the heading lines, backspacing over the “var” names, and typing a new label (such as “log_Un”). Using the same procedures as before, analyze the data sets. Does either of these transformations lead to data satisfying the conditions of our inference procedures? Explain.

The logarithmic transformation leads to a quite symmetric data set, but produces a few low outliers. The normal quantile plots are much more linear than before. Overall, the logarithmic transformation of the two data sets is our best option. The square root transformation does not eliminate the high outliers and does not produce symmetric distributions.

Now it’s time to find out the answer to our question—does cloud seeding really work? WebStat performs the two-sample t procedure very quickly. Go to Stat > T statistics > Two Sample. Choose the appropriate two variables. Should you pool the variances? Discuss this with your teacher if you are unsure. *In general, there is no reason to pool variances unless you have some information that compels you to believe the two variances are equal. In this case it will make little to no difference.* Then press the Next button. You have the option to perform a significance test or create a confidence interval. Choose the “hypothesis test” (significance test), and specify the correct alternative hypothesis.

In the lower window you should see the results of the test. Note especially the p -value. What do you conclude?

The p -value is small, 0.007. Since this is less than 0.05, the data provide strong evidence that the means of the logarithm of rainfall amounts were indeed different, because it is

extremely unlikely that this result occurred by chance. Cloud seeding appears to increase rainfall amounts.

If your students get a p-value of 0.0269, they performed the test on the original data instead of the transformed data.

WebStat allows you to save the results from that lower window. If you go to Stat > Save/Print results it will open a new window with the contents of the lower window. From that you can either print or copy and paste into a word processing document.

When you are finished with WebStat, remember to go to the WebStat menu and choose Exit.

Extensions for the teacher

If you have time, ask the students to repeat the analysis after eliminating the low outliers in the transformed data. Did the conclusion change?

Eliminating the three low outliers leads to a p-value of 0.0012, so the conclusion remains the same.

You might want your students to investigate the two different logarithmic transformations. Both produce the same p-value.

You can also discuss the degrees of freedom, t statistic, and standard error given in the output. Students might also try entering the data into their calculators to see if the results mirror those produced by WebStat.

Does Seeding Clouds Produce More Rain?

An activity exploring data transformations and inference using WebStat

Procedure

Years ago, researchers wanted to determine whether seeding clouds with silver nitrate increased rainfall. The DASL website contains a dataset from two sets of randomly selected clouds, one that was seeded and the other that was not.

Go to the DASL site,

<http://lib.stat.cmu.edu/DASL/>

and search on data subjects. Look under Environment. Choose the Clouds Datafile.

Variables:

Unseeded_clouds: Amount of rainfall from unseeded clouds (in acre-feet)

Seeded_clouds: Amount of rainfall from clouds seeded with silver nitrate (in acre-feet)

Highlight everything under the line that says **The Data:**, and copy.

Now go to the WebStat site,

<http://www.stat.sc.edu/webstat/>

and when the orange button appears, click on it. Soon you will have a new window open; the top part resembles a spreadsheet.

Pasting data into WebStat can be a bit tricky, because the browser you use makes a big difference. The window you need to paste into sometimes comes up behind another window. Go to the WebStat Help menu (see Help on the WebStat menu bar) and load the help page. Scroll down to where you see a link to the instructions (under the data column) for loading data by pasting. If you follow the directions closely, you can paste the data into WebStat, and the headings should be included. Start by going to the **Data** menu and choosing **Paste data**.

The first step in analysis, of course, is to look at the data. Go to the Graphics menu and choose one of the univariate data pictures, Boxplot. You can select both variables, and WebStat produces side-by-side boxplots. Describe the shapes of the two distributions.

Now consider a second picture of the data, a stemplot (“Stem and leaf plot”). What do the stemplots show us that the boxplots did not?

The plots made above indicate that these data are not normally distributed. Have WebStat make a normal probability plot to verify that these data are not normally distributed.

The researchers wanted to find out if the seeding produced a significant increase in rainfall. Write the hypotheses for the appropriate significance test.

What conditions must be satisfied before we apply our inference procedures? Are these conditions satisfied? Explain.

One method to resolve this situation is to transform the data before performing inference. WebStat makes transformation easy. Go to the Data menu and choose Transform data. You have four options. Considering the skewness you noted earlier, which transformation would you expect to mitigate that? Explain your choice.

Transform the data using a logarithmic transformation and the square root transformation. WebStat stores the transformed data in new columns. You can change the labels by going to the heading lines, backspacing over the “var” names, and typing a new label (such as

“log_Un”). Using the same procedures as before, analyze the data sets. Does either of these transformations lead to data satisfying the conditions of our inference procedures? Explain.

Now it's time to find out the answer to our question—does cloud seeding really work? WebStat performs the two-sample t procedure very quickly. Go to Stat > T statistics > Two Sample. Choose the appropriate two variables. Should you pool the variances? Discuss this with your teacher if you are unsure. Then press the Next button. You have the option to perform a significance test or create a confidence interval. Choose the “hypothesis test” (significance test), and specify the correct alternative hypothesis.

In the lower window you should see the results of the test. Note especially the p-value. What do you conclude?

WebStat allows you to save the results from that lower window. If you go to Stat > Save/Print results it will open a new window with the contents of the lower window. From that you can either print or copy and paste into a word processing document.

When you are finished with WebStat, remember to go to the WebStat menu and choose Exit.

Canadian Crime

Distributions in Two Way Tables

Go to:

<http://www.statcan.ca/english/Pgdb/State/Justice/legal04a.htm>

(This is the lower case “l” not the number one.)

All of the following questions refer to 1999 Canadian crime data. Round all answers to three decimal places.

1. Find the proportion of crimes in Canada that involve violence. _____
2. Find the proportion of crimes in Canada that are property crimes. _____
3. What proportion of Canada’s violent crimes happened in New Brunswick? _____
4. What percentage of Canada’s violent crimes happened in Alberta? (You need to return to the box at the top of the screen and choose Alberta in the menu.) _____
5. What percentage of the crimes in Alberta involved violence? _____
6. Find the percentage of crimes in Canada that were breaking and entering or theft (not including motor vehicle theft). _____
7. Find the percentage of crimes in Yukon that involved breaking and entering or theft (not including motor vehicle theft). _____

8. Given a crime was committed in Yukon, what is the probability that the crime was fraud?

9. Given a person was a victim of a property crime in Yukon, what is the probability that the crime committed against them was fraud? _____

10. Of all the property crimes in Yukon, what proportion were motor vehicle theft? _____

11. What proportion of Canada's murders were committed in Quebec? _____

12. Suppose you have been offered jobs in Quebec, British Columbia, New Brunswick, and Alberta. If you were concerned with crime, which province would appear safest to you? Explain your answer using complete sentences and mathematical reasoning. (Further investigation of this website may be needed to support your answer.)

Canadian Crime

Distributions in Two Way Tables

Teacher Notes:

Objective: The objective of this activity is for students to read two-way tables and calculate marginal and conditional distributions. You may want to include a map of Canada for students. Answers are based on 1999 figures, but you may update this activity when more recent data become available.

Answers:

1. 0.118
2. 0.525
3. 0.025
4. 10.8%
5. 11.3%
6. 41.2%
7. 14.0%
8. 0.035
9. 0.166
10. 0.076
11. 0.259

12. Answers may vary:

If students look at the proportion of Canadian crimes committed in each province, they will find that Quebec has 18.4%, British Columbia has 19.9%, New Brunswick has 2.2% and Alberta has 11.2%. But the crime rate (number of crimes/population) is a more meaningful measure. Populations are given at

<http://www.statcan.ca/english/Pgdb/People/Population/demo02.htm>

For 1999, the rates are

$$\text{Quebec: } \frac{456427}{7345400} = 0.062$$

$$\text{British Columbia: } \frac{493649}{4023100} = 0.123$$

$$\text{New Brunswick: } \frac{55443}{755000} = 0.073$$

$$\text{Alberta: } \frac{277995}{2964700} = 0.094$$

From these we would conclude that Quebec is safest province of these choices in regard to overall crime.

A Nation of Movers?

Exploring Categorical Data with the Calculator

Introduction

The General Social Survey is conducted regularly in the U. S., and the interviewers ask questions on a wide variety of topics. They repeat some questions in different years, to see if people’s attitudes or status change. The surveys have labels that are years; you can think of the years as categorical data. Today we will investigate whether Americans tend to stay in their hometowns.

The web address for the gateway to the data is

<http://www.icpsr.umich.edu/GSS99/>

Once there, scroll down until you see the link labeled **Main Codebook Pages**. Choose that. Then choose **Index by Subject**. Under the letter G choose **Geographic Mobility**. One choice appears, **MOBILE16**. Choose that dataset.

Both the question and the data table appear. The “Punch” column is simply the code the interviewer used for each response.

Investigative Questions

1. What is the statistical term for the numbers in the rightmost column?
2. Notice that some of the responses are “No answer.” Why might some respondents have refused to answer this question? Are there other situations that may have led the interviewer to record “No answer” for the response?
3. Identify the column variable for the data table.
4. Identify the row variable for the data table.
5. Calculate the column totals and record them here.

1972-82	1982B	1983-87	1987B	1988-91	1993	1994	1996

6. Give the total number of respondents over all the surveys.
7. What percent of all the respondents live in their hometown?
8. In 1972-82, what percent of the respondents lived in their hometown?

9. In 1996, what percent of the respondents lived in their hometown?
10. Does it appear that the mobility of Americans has changed over these years? Explain your answer.

More Advanced Analysis

11. As statisticians, we would like to determine whether any differences you have noted are significant. What tools can we use to answer that question?
12. Give the hypotheses for the analysis you wish to do.
13. Look back at the data table. Identify any of the responses that might not be of interest in investigating the mobility of Americans.

We will now proceed with the inference. Disregarding the row(s) you identified above, enter your data into your calculator and run the chi-squared test.

14. Give the value of the test statistic, the degrees of freedom, and the p-value.
15. Before we draw a conclusion, we must examine the expected counts. Are any of them less than 5? Are any of them less than one?
16. What do you conclude?

Digging Deeper

So far we have examined the data over eight different surveys. Now we will focus on a subset of the surveys and look for relationships.

Earlier you compared one part of the 1972-82 survey with the 1996 one and found a difference in one response. Let's look at the distributions for those two surveys.

17. Follow the same analysis you performed above. What do you conclude?

Perhaps the mobility in our country has changed significantly since the seventies, but has it been changing significantly more recently? Now we will examine the data from the 1993, 1994, and 1996 surveys.

18. Using the data from those three surveys, follow the same analysis you performed above. What do you conclude?

19. Now focus on just 1993 and 1994. Follow the same analysis you performed above. What do you conclude?

20. Now focus on just 1994 and 1996. Follow the same analysis you performed above. What do you conclude?

21. Look back at the previous three results. Explain what the results appear to show about those three years.

22. Create the conditional distributions of responses for each of the three years, giving your answers in percents.

	1993	1994	1996
Same State, same City			
Same State, different City			
Different State			

What do these results tell you?

Solutions: A Nation of Movers?

Exploring Categorical Data with the Calculator

Objectives

Obtain data from the General Social Survey.

Perform chi-squared analysis to explore changes in the mobility of U. S. citizens.

Introduction for the Instructor

This activity is meant for students to work through, with some guidance from the instructor. Students must have statistical calculators that can perform a chi-squared analysis. Some instructors might adapt this to a statistical software package for more extensive analysis.

We have included answers to the questions, printed in italics. A student copy without answers is included after this annotated version.

Introduction

The General Social Survey is conducted regularly in the U. S., and the interviewers ask questions on a wide variety of topics. They repeat some questions in different years, to see if people's attitudes or status change. The surveys have labels that are years; you can think of the years as categorical data. Today we will investigate whether Americans tend to stay in their hometowns.

The web address for the gateway to the data is

<http://www.icpsr.umich.edu/GSS99/>

Once there, scroll down until you see the link labeled **Main Codebook Pages**. Choose that. Then choose **Index by Subject**. Under the letter G choose **Geographic Mobility**. One choice appears, **MOBILE16**. Choose that dataset.

Both the question and the data table appear. The "Punch" column is simply the code the interviewer used for each response.

Investigative Questions

1. What is the statistical term for the numbers in the rightmost column?

They are the row totals, and they constitute the marginal distribution of responses.

2. Notice that some of the responses are "No answer." Why might some respondents have refused to answer this question? Are there other situations that may have led the interviewer to record "No answer" for the response?

Many answers are possible. Some respondents may have been suspicious of the interviewer. Some might have been fugitives, illegal aliens, or runaways. Some respondents might not have spoken English sufficiently well to understand the question. Some might have been mentally incapacitated.

3. Identify the column variable for the data table. *survey year*
4. Identify the row variable for the data table. *response*
5. Calculate the column totals and record them here.

1972-82	1982B	1983-87	1987B	1988-91	1993	1994	1996
13626	354	7542	353	5907	1606	2992	2904

6. Give the total number of respondents over all the surveys. *35,284*
7. What percent of all the respondents live in their hometown? $\frac{14571}{35284} = 41.3\%$
8. In 1972-82, what percent of the respondents lived in their hometown? $\frac{5759}{13626} = 42.3\%$
9. In 1996, what percent of the respondents lived in their hometown? $\frac{1092}{2904} = 37.6\%$
10. Does it appear that the mobility of Americans has changed over these years? Explain your answer.

With just the numbers calculated above, it appears that more Americans lived in their hometown in the seventies than in the mid-nineties. Students might calculate other conditional distributions (such as the percentage of Americans who lived in a different state according to the two surveys) to support their answers.

More Advanced Analysis

11. As statisticians, we would like to determine whether any differences you have noted are significant. What tools can we use to answer that question?

We would like to use the tools of inference. Specifically, we would like to use a chi-squared test to see whether year and response are independent variables.

12. Give the hypotheses for the analysis you wish to do.

H_0 : The survey year and response are independent.

H_a : The survey year and response are not independent.

or

H_0 : There is no association between survey year and response.

H_a : There is an association.

13. Look back at the data table. Identify any of the responses that might not be of interest in investigating the mobility of Americans.

“Don’t know” and “No answer” are probably not relevant. We are interested here in whether people move, not whether they can answer the question.

We will now proceed with the inference. Disregarding the row(s) you identified above, enter your data into your calculator and run the chi-squared test.

On a TI-83, students should enter their data into a 3 by 8 matrix and then use the chi-squared test function on the Stat > Test menu.

14. Give the value of the test statistic, the degrees of freedom, and the p-value.

$$\chi^2 = 157.9, df = 14, p\text{-value} = 1.86E-26$$

15. Before we draw a conclusion, we must examine the expected counts. Are any of them less than 5? Are any of them less than one?

All expected counts are at least 87, so we are comfortable proceeding. (The TI-83 has the expected counts stored in the matrix specified when running the test.)

16. What do you conclude?

The extremely low p-value indicates that the variation among the years is unlikely to have happened by chance. We have strong evidence to reject the null hypothesis and conclude that the distribution of responses is different for the different surveys. The mobility of Americans appears to have changed over time.

Digging Deeper

So far we have examined the data over eight different surveys. Now we will focus on a subset of the surveys and look for relationships.

In problems 8 and 9 you compared one part of the 1972-82 survey with the 1996 one and found a difference in one response. Let’s look at the distributions for just those two surveys.

17. Using those two surveys alone, follow the same procedures for the chi-squared analysis you performed above. What do you conclude?

$$\chi^2 = 48.9, df = 2, p\text{-value} = 2.36E-11 \text{ Once again, we have strong evidence of a difference in the distribution of responses.}$$

Perhaps the mobility in our country has changed significantly since the seventies, but has it been changing significantly more recently? Let’s examine the data from the 1993, 1994, and 1996 surveys.

18. Using the data from those three surveys, follow the same analysis you performed above. What do you conclude?

$\chi^2 = 15.06$, $df = 4$, $p\text{-value} = 0.005$ This $p\text{-value}$ is quite low and we have strong evidence that the distribution of responses is different for the three surveys. Mobility appears to have changed in recent years.

19. Now focus on just 1993 and 1994. Follow the same analysis you performed above. What do you conclude?

$\chi^2 = 4.42$, $df = 2$, $p\text{-value} = 0.11$ This $p\text{-value}$ is not less than 0.05. We do not have strong evidence that the distribution of responses from 1993 is significantly different from the distribution of responses from 1994.

20. Now focus on just 1994 and 1996. Follow the same analysis you performed above. What do you conclude?

$\chi^2 = 4.45$, $df = 2$, $p\text{-value} = 0.11$ This $p\text{-value}$ is not less than 0.05. We do not have strong evidence that the distribution of responses from 1994 is significantly different from the distribution of responses from 1996.

21. Look back at the previous three results. Explain what the results appear to show about those three years.

The distributions from the three surveys are significantly different, but when we pair 1993 with 1994 and 1994 with 1996 we do not find a significant difference. This leads us to believe that the 1993 survey was significantly different from the 1996 survey. Your students should recognize that the $p\text{-value}$ and χ^2 statistic do not follow an additive or transitive property.

22. Create the conditional distributions of responses for each of the three years, giving your answers in percents.

	1993	1994	1996
Same State, same City	41.3	39.5	37.8
Same State, different City	26.6	25.3	24.4
Different State	32.1	35.2	37.9

What do these results tell you?

We can now see that 1993 and 1996 differed more than the other pairings. In 1993 the most popular response was remaining in one's hometown. In 1996 living in a different state from one's hometown was the most common response. It appears that the mobility of Americans changed significantly in the mid-nineties.

Careful students will point out that we need more recent data to see if 1996 was the beginning of a trend or just an unusual year.

Extensions

Students might examine other subsets of the surveys in a similar fashion.

You could change this to a problem using inference procedures for proportions instead of chi-square, if you focus on one response and two surveys.

If you have software available, you can enter the data into it and perform the same analysis as above. If you would like to use WebStat (<http://www.stat.sc.edu/webstat/>), you can enter the data after choosing Stat > Contingency Table and entering the numbers of rows and columns.

Let's Buy a Diamond Ring!

Exploring Regression on a Computer

Introduction

When people shop for diamonds, they learn that stones vary in cut, clarity, color, and carat. Carat is a measure of weight, one carat = 0.2 g. Today we will only focus on carat. We want to discover how the price varies with the carat of a diamond.

Fortunately for us, someone has already found data on the price of diamond rings and their carats. The price happens to be in Singapore dollars.

To find the data, go to

<http://www.amstat.org/publications/jse/archive.htm>

Once there, scroll down until you see the information beginning with “diamond.dat” (the datasets are listed alphabetically).

1. How many observations are in the data set?

If you read the “diamond.txt” file, you will learn that the diamond ring data comes from a newspaper advertisement. All the rings had the same sort of band, so only the stones differed.

Loading the Data

We can load the data either by saving it first to our own computer or by opening it, copying the contents, and pasting.

JMPIN 4 Directions: Click on diamond.dat. Save the file to your own computer. Then open JMPIN. Click on Open Data Table. Find the appropriate directory for the data file. The file is saved as .dat, so in the “Files of type” box, select “Text Import Files (*.TXT, *.CSV, *.DAT).” Then select “Attempt to Discern Format” at the bottom of the window. Select your file in the “File Name” box. Click Open. The data file should appear in its own window. Change the labels on the columns to “Carat” and “Price.”

Minitab version 11 Directions: Right click on diamond.dat. Save the file to your own computer. Then open Minitab. Go to File > Open Worksheet. Find the directory where you saved the data file. Under “List Files of Type,” choose “Data (*.dat).” Highlight diamond.dat. Click on Options. A box opens. In the text in the left column, under “First Row of Data,” select “Use Row” and enter 1. Click OK to close this box. Then choose “Preview” in the box that appears. The cursor should appear in the box next to “Name.” Type in “Carat” in the first box and “Price” in the one next to it. Click OK, and the data should appear with the columns correctly labeled.

Examine the Data

2. Before we get too deep in the analysis, let's take a moment to think. What do you expect to be the relationship between the carat of the diamond ring and the price?
3. Which variable is the explanatory variable and which is response?
4. Create a scatterplot of the data and describe it here.

JMPIN 4 Directions: Choose Graph > Overlay Plot. Select the appropriate variable for Y and the other for X. Click on OK.

Minitab version 11 Directions: Choose Graph > Plot and put the appropriate variables for Y and X. Click on OK.

5. Is it reasonable for us to proceed with a linear model? Or does another model appear to be more appropriate? Explain.

Linear Regression (Least Squares Method)

Let's put our computers to work. Remember, statistical software will perform many calculations, but it's up to the user to identify what's meaningful and appropriate!

6. Find the least squares regression line for these data.

JMPIN 4 Directions: Choose Analyze > Fit Y by X. Enter the correct Response and Factor variables. Click OK. A window appears. Click on the red arrow next to the word, "Bivariate," and choose Fit Line. A line appears on the graph and information appears below.

Minitab version 11 Directions: Choose Stat > Regression > Regression. Enter the appropriate variables for Response and Predictors. Click OK. This will give you information in the Session window, but no scatterplot. Alternatively, choose Stat > Regression > Fitted Line Plot and enter the variables and OK.

7. What is R^2 for these data? What information does R^2 provide?

8. Examine a residual plot and describe it here. Does it support the use of a linear model for these data? Explain.

JMPIN 4 Directions: Click on the red arrow next to the word, “Linear fit” below the scatterplot. Choose Plot Residuals. At the bottom of the window a residual plot appears.

Minitab version 11 Directions: Choose Stat > Regression > Regression. Enter the appropriate variables for Response and Predictors. Click on Graphs. In the box under “Residuals versus the variables,” select Carat.

9. Write a sentence interpreting the slope of the line in the context of these data.

10. Write a sentence interpreting the vertical intercept in the context of these data.

Solutions: Let's Buy a Diamond Ring!

Exploring Regression on a Computer

Objectives

Obtain data from the Journal of Statistics Education data archive.

Use technology to perform linear regression. Also perform regression inference, if desired.

Evaluate the validity of a model.

Introduction for the Instructor

This activity is meant for students to work through, with some guidance from the instructor. We obtain data from a website and analyze it using statistical software. We include basic instructions for different software packages, but individuals may have to adapt them to a specific version and machine. This activity can be used either when studying regression for the first time or when studying regression inference.

We have included answers to the questions, printed in italics. We have appended a student copy with the answers omitted.

Introduction

When people shop for diamonds, they learn that stones vary in cut, clarity, color, and carat. Carat is a measure of weight, one carat = 0.2 g. Today we will only focus on carat. We want to discover how the price varies with the carat of a diamond.

Fortunately for us, someone has already found data on the price of diamond rings and their carats. The price happens to be in Singapore dollars.

To find the data, go to

<http://www.amstat.org/publications/jse/archive.htm>

Once there, scroll down until you see the information beginning with “diamond.dat” (the datasets are listed alphabetically).

1. How many observations are in the data set? *48*

If you read the “diamond.txt” file, you will learn that the diamond ring data comes from a newspaper advertisement. All the rings had the same sort of band, so only the stones differed.

Loading the Data

We can load the data either by saving it first to our own computer or by opening it, copying the contents, and pasting.

Teachers should adapt these instructions to their own situations. When some users click on the diamond.dat link, they get a new window containing the data. Those users may have success simply by copying and pasting the data into the statistical software. The instructions below are for users who must save the file first.

JMPIN 4 Directions: Click on diamond.dat. Save the file to your own computer. Then open JMPIN. Click on Open Data Table. Find the appropriate directory for the data file. The file is saved as .dat, so in the “Files of type” box, select “Text Import Files (*.TXT, *.CSV, *.DAT).” Then select “Attempt to Discern Format” at the bottom of the window. Select your file in the “File Name” box. Click Open. The data file should appear in its own window. Change the labels on the columns to “Carat” and “Price.”

Minitab version 11 Directions: Right click on diamond.dat. Save the file to your own computer. Then open Minitab. Go to File > Open Worksheet. Find the directory where you saved the data file. Under “List Files of Type,” choose “Data (*.dat).” Highlight diamond.dat. Click on Options. A box opens. In the text in the left column, under “First Row of Data,” select “Use Row” and enter 1. Click OK to close this box. Then choose “Preview” in the box that appears. The cursor should appear in the box next to “Name.” Type in “Carat” in the first box and “Price” in the one next to it. Click OK, and the data should appear with the columns correctly labeled.

Newer versions of Minitab can load the data more easily, perhaps even just by opening the diamond.dat file.

Examine the Data

2. Before we get too deep in the analysis, let’s take a moment to think. What do you expect to be the relationship between the carat of the diamond ring and the price?

We expect that as carat increases, price increases.

3. Which variable is the explanatory variable and which is response?

Carat is explanatory and price is response.

4. Create a scatterplot of the data and describe it here.

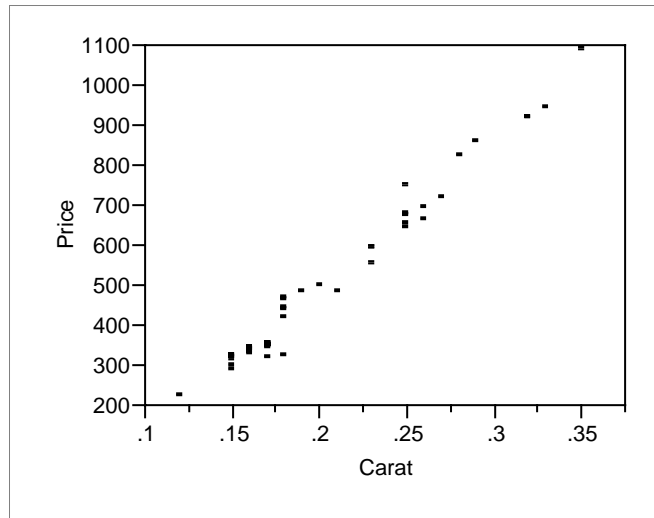
JMPIN 4 Directions: Choose Graph > Overlay Plot. Select the appropriate variable for Y and the other for X. Click on OK.

Minitab version 11 Directions: Choose Graph > Plot and put the appropriate variables for Y and X. Click on OK.

The scatterplot shows a linear pattern with positive slope. No obvious outliers are present.

Below is the graph from JMPIN.

Overlay Plot



5. Is it reasonable for us to proceed with a linear model? Or does another model appear to be more appropriate? Explain.

A linear model seems like an appropriate place to start, given this plot.

Linear Regression (Least Squares Method)

Let's put our computers to work. Remember, statistical software will perform many calculations, but it's up to the user to identify what's meaningful and appropriate!

6. Find the least squares regression line for these data.

JMPIN 4 Directions: Choose Analyze > Fit Y by X. Enter the correct Response and Factor variables. Click OK. A window appears. Click on the red arrow next to the word, "Bivariate," and choose Fit Line. A line appears on the graph and information appears below.

Minitab version 11 Directions: Choose Stat > Regression > Regression. Enter the appropriate variables for Response and Predictors. Click OK. This will give you information in the Session window, but no scatterplot. Alternatively, choose Stat > Regression > Fitted Line Plot and enter the variables and OK.

The equation is $Price = -260 + 3721 \text{ Carat}$, to the nearest integers.

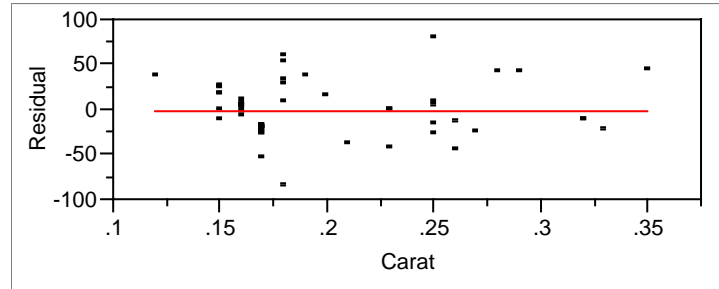
7. What is R^2 for these data? What information does R^2 provide?

$R^2 = 0.978$. This tells us that approximately 98% of the variation in price can be accounted for by this model.

8. Examine a residual plot and describe it here. Does it support the use of a linear model for these data? Explain.

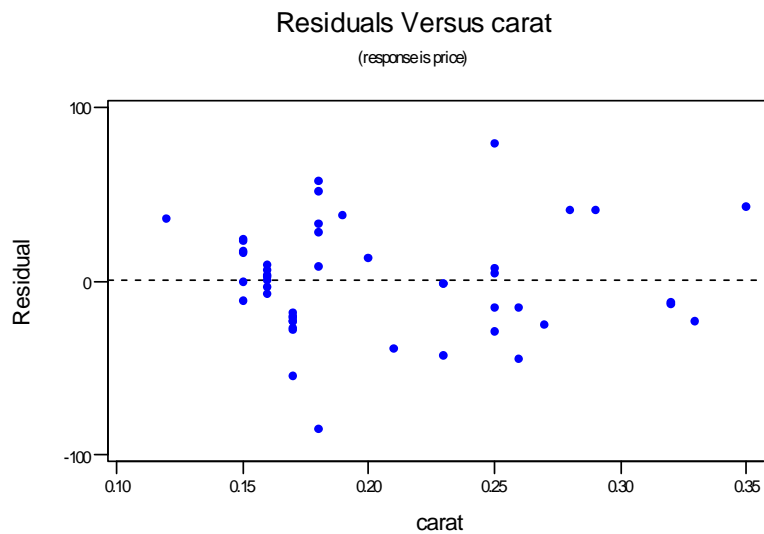
JMPIN 4 Directions: Click on the red arrow next to the word, “Linear fit” below the scatterplot. Choose Plot Residuals. At the bottom of the window a residual plot appears.

Here is the plot from JMPIN:



Minitab version 11 Directions: Choose Stat > Regression > Regression. Enter the appropriate variables for Response and Predictors. Click on Graphs. In the box under “Residuals versus the variables,” select Carat.

Here is the plot from Minitab:



We do not see any obvious patterns to make us doubt the worthiness of a linear model.

9. Write a sentence interpreting the slope of the line in the context of these data.

For every increase of one carat in weight of the diamond, the price would increase \$3721, on average.

10. Write a sentence interpreting the vertical intercept in the context of these data.

The cost of producing a ring with no diamond is -\$260. This obviously makes no sense and alerts us to the dangers of using this model for very small carat sizes.

Extensions—Notes to the Instructor

If your students are studying regression inference, the software packages have given the results of t tests on slope and intercept.

JMPIN Output:

Parameter Estimates

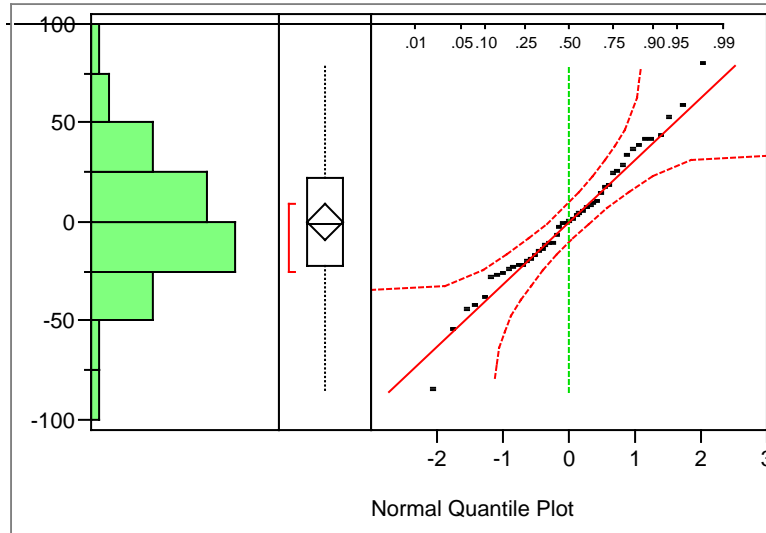
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-259.6259	17.31886	-14.99	<.0001
Carat	3721.0249	81.78588	45.50	<.0001

Minitab output:

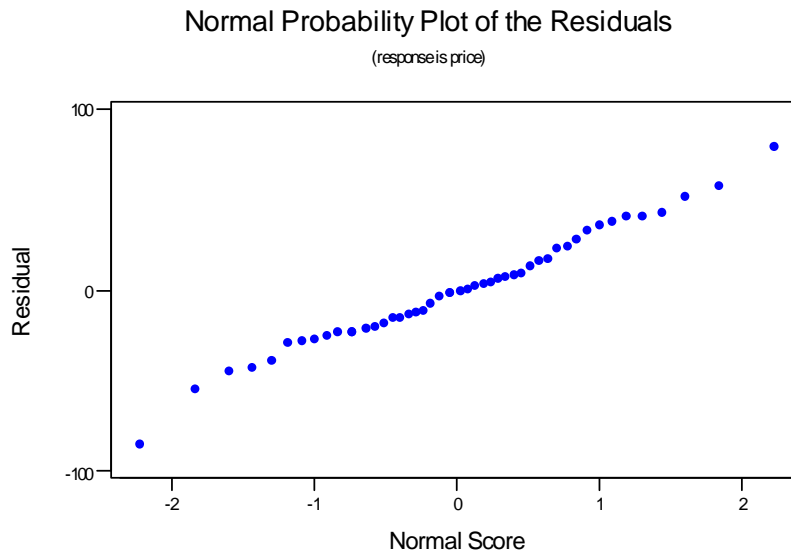
Predictor	Coef	StDev	T	P
Constant	-259.63	17.32	-14.99	0.000
Carat	3721.02	81.79	45.50	0.000

Students should notice that both the intercept and the slope are significantly different from zero, which should not be surprising! But they should also check to see if the residuals appear to arise from a normal distribution before expressing full confidence in the results. Here are directions for creating a normal quantile plot of the residuals:

JMPIN 4 Directions: Click on the red arrow next to “Linear fit” and choose Save Residuals. Go to the data window and the residuals have appeared. Go to Analyze > Distribution and put Residuals Price in the Y box. Click OK. Click on the red arrow next to Residuals Price. Choose Normal Quantile Plot and it will appear beside the histogram. Here are the graphs that should appear:



Minitab version 11 Directions: Choose Stat > Regression > Regression. Enter the appropriate variables for Response and Predictors. Click on Graphs and check “Normal plot of residuals.” The plot shown below should appear in its own window.



The plots show a linear trend, so students should express confidence in assuming the residuals arise from a normal distribution, and thus have confidence in the results of the significance test.

Concluding Notes

The article linked to the data file contains a discussion of how the author used the data in a course that is more advanced than AP Statistics. Teachers may gain insights through reading the discussion there. Students should not read that discussion until after performing their own investigation, as the final model given in the article is quite complicated.

Food Fit For Life - Fast Food Nutrition Comparison

Introduction:

More and more people are eating food prepared outside the home than ever before, and fast food constitutes a large proportion of that food. On one hand, people have less control over how the food is prepared. On the other hand, people have much more nutritional information that is made available to them by fast food restaurants, all of which want to convince consumers that they have “healthy” food.

Goal:

You need to compare nutritional information at fast food restaurants in our area. The Chamber of Commerce has hired you as a consultant in order to advise it on which fast food restaurant in our area is the “healthiest” one at which to eat. The Chamber plans to award its annual “Food Fit for Life” award to that fast food restaurant.

Procedure:

1. Determine which fast food restaurants there are in our area and which will be included in your study. You must have at least four restaurants.
2. Determine which food items from those fast food restaurants will be used in your study. Use the internet links provided below to help determine which food items you will include in your study.
3. Collect data about the selected food items served at the fast food restaurants in your study.
4. Organize your data first into a table using a spreadsheet or statistics software program. Then organize the data into a more appropriate graphical format for describing the distribution of data, making comparisons and drawing conclusions.
5. Calculate the necessary statistics to support your description of the distribution of the collected data. Show your setup of the calculation that produced each statistic.
6. Write a formal report for the Chamber of Commerce using your data, comparative descriptions, statistics and graphical displays. Your report must include the data table, appropriate graphs and statistics, and your written descriptions, comparisons and conclusions. *You will make a presentation to the Chamber of Commerce* (our class). The preferred medium for your report is a PowerPoint presentation file; another presentation software package may be used, but you must be able to make the presentation in class. If you do not have access to presentation software, you may type your report using a word processor.
7. The Chamber of Commerce will use your report on the fast food restaurants to award its “Food Fit for Life” award.

Internet Links

The links that follow are resources that will enable you to collect your data. If you find other resources, you may use them as long as you cite them in your report.

McDonalds: http://www.mcdonalds.com/countries/usa/food/nutrition_facts/index.html

Burger King: <http://www.burgerking.com/nutrition/ntables.htm>

Wendy's: http://www.wendys.com/the_menu/nut_frame.html

Subway: http://www.subway.com/our_food/nguide/usa/index.html

Food Finder: <http://www.olen.com/food/>

Considerations/Things to Address:

1. What fast food items were chosen for data collection? Explain why these food items were chosen.
2. What are the variables on which you chose to collect data? Explain why these variables were chosen.
3. What statistics did you calculate and how did you use them in your analysis and in reaching your conclusions?

Evaluation:

Your final project grade will be based on the following:

1. Your justification of why the data you chose to collect was the best data to achieve your goal. (20%)
2. Your spreadsheet/software tabular data, graphs and setups for calculated statistics. (60%)
3. Your class presentation of your report to the Chamber of Commerce. (20%)
4. Other materials that support your report. (10%)