MINITAB PROJECTS IN INTRODUCTORY STATISTICS

Marsha Davis and Pete Johnson
Eastern Connecticut State University
Mathematics Department
83 Windham Street, Willimantic, CT 06226
davisma@easternct.edu and johnsonp@easternct.edu

Abstract

This paper discusses material presented during a Minitab minicourse in which participants had the opportunity to test out laboratory projects designed for introductory statistics courses.

Introduction

Enrollments in introductory statistics have increased dramatically in recent years (Blair, Kirkman, & Maxwell, 2013). Students enrolled in this course have diverse backgrounds, interests, and reasons for taking the course. Some students take the course because they recognize that they need to know how to deal with data. Other students take the course to meet a variety of requirements. Unfortunately, not all students find the course engaging and many find it difficult. Given the importance of statistics and statistical reasoning in an increasingly complex and information-rich world, ways must be found to engage students in activities that support learning the basic elements of statistical thinking and the important concepts that underlie statistical reasoning. More than 20 years ago, George Cobb (1991) provided a direction for improving the teaching of introductory statistics:

> Almost any course in statistics can be improved by more emphasis on data and concepts, at the expense of less theory and fewer recipes. To the maximum extent feasible, calculations and graphics should be automated. Any introductory course should take as its main goal helping students to learn the basics of statistical thinking.

More recently, the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report* (2010), put forth the following recommendations:

- Emphasize statistical literacy and develop statistical thinking
- Use real data
- Stress conceptual understanding, rather than mere knowledge of procedures
- Foster active learning in the classroom
- Use technology for developing conceptual understanding and analyzing data
- Use assessments to improve and evaluate student learning

The *GAISE* report recognizes that some progress has been made in introductory statistics courses since the release of the Cobb report. Data analysis is more prominent in the curriculum, there is an increased use of technology for calculations, and some instructors

include projects or labs as part of their assessment and as a means to actively involve students in doing statistics.

In the spirit of Cobb's statement and the *GAISE* recommendations, this minicourse included laboratory projects based on data that included birth weights, heights of students and their parents, marketing research data, demography, and body temperature as well as data from the study *Monitoring the Future: A Continuing Study of American Youth*. In addition, the project on the sampling distribution of the sample mean compared the means of simulated data from several different distributions.

We use the statistical package Minitab for our laboratory projects. Minitab has several advantages that make it ideal for use in introductory statistics. Minitab has an intuitive design that is simple to learn and yet is very powerful. Students can quickly analyze large, complex, and therefore very realistic, data sets. In addition, it is affordable for students and educators.

This paper shares two of the lab projects presented in the minicourse. The two labs focus on descriptive statistics, with an extension in the second lab into inferential statistics. Both labs deal with multivariate data sets and thus are more complex than standard textbook exercises. Students are encouraged to work in groups on these labs. During labs, they discuss statistical ideas and argue over different interpretations of what they see in the data. Because instructions on the use of Minitab are embedded in each lab, instructors need only provide a brief introduction to Minitab at the start of the semester. There is no further need for the instructor to take class time to show students how to use Minitab.

Minitab Lab: Jeans

Background: Minitab includes folders with data sets from varied contexts. This lab is based on one of these data sets, JEANS.mtw. JEANS contains marketing research data on jeans -- the price of jeans, who buys them, where they are sold, etc. In this lab, students learn about missing value codes, how to use histograms to find anomalies in data, how to make stem-and-leaf plots (and adjust the increment) and boxplots (including comparative plots), how to impose a logical order on categorical values, how to make a frequency table, and how to save a project. This lab can be introduced early in the course.

The Lab: The lab begins with instructions to students followed by items for them to work through and questions for them to answer.

Jeans Lab

As you work through this Minitab lab, record your answers in a Word document. Put the name of the group members as a header or footer in this document and make sure pages are numbered. Turn in **one** copy of the completed lab write-up per group. As you work through this lab, I will spend most of my time circulating around the room; *don't be afraid to ask me questions*!

For this lab you will use the Minitab worksheet Jeans.mtw, which contains the kind of data used in marketing research. To open it,

- Go to **File>Open Worksheet**
- Click on the "Look in Minitab Sample Data Folder" button at the bottom of that window.
- Select the folder "Student12," then select "Jeans".

If you look at the data, much of it probably won't make sense to you without some explanation of what the numbers and codes stand for. To find a description of this worksheet,

- Click **Help>Help.**
- Enter Data Sets.
- Select Student Data Set Descriptions and click Display.
- Locate Jeans and click to open.

Before you start to answer the questions in this lab, look at the data description; you should probably leave the Help window open while you are doing this lab.

1. Make a histogram of the variable *PayJean*. (Recall: **Graph>Histogram,** click on the "Simple" graph, then select the variable *PayJean*.)

   QUESTION: If you either look at the graph or page down in the Minitab spreadsheet, you'll notice quite a few entries under *PayJean* are "999". These values look suspicious to me. What do you think "999" stands for?

2. a) It will be easier to analyze the cost of jeans if we delete the "999" values by recoding them as missing values. Here's how:
   - Click at the top of the *PayJean* column to highlight the entire column.
   - Go to **Editor>Replace**. Enter 999" for "Find What" and leave the "Replace With" line blank.
   - Click Replace All.

   Scroll down the column. What symbol did Minitab replace the 999s with? This is Minitab's missing value symbol for *numeric data*.
   - Click anywhere on the worksheet. Repeat Editor>Replace to recode all 999s for the other variables to missing values. (Notice that the 999s for categorical variables are replaced by a blank cell.)

   Once you've replaced the 999s, make another histogram. Make any adjustments to the scaling that will make the histogram easier to interpret. (For example, would it be better to use cutpoints or midpoints? Do you want to adjust the scaling?)

   b) Make a stem-and-leaf plot (**Graph>Stem-and-Leaf** and select the variable *PayJean*). While your histogram in (a) is fairly easy to read, your stem-and-leaf plot is not.

- Try making a second stem-and-leaf plot, only this time specify "2" for "Increment" (box near the bottom of the stem-and-leaf window). This should give a graph that is easier to read.
- Next remake the stemplot but this time click the box for "Trim Outliers."

(***NOTE:*** *In Minitab, the numbers in the far left column of the stemplot do not stand for data points. Instead, they show the number of data points, counting either down from the top or up from the bottom; the row containing the middle number is noted by writing parentheses around the number of values in that row. Also note that by clicking the box that says "Trim Outliers," Minitab gives a list of outliers at the bottom of the stem-and-leaf plot that are not actually drawn in the graph itself.*)

c) Make a boxplot (**Graph>Boxplot** and select the variable *PayJean*). Point the cursor at one of the outliers – you should be able to read off its value. To see the plot displayed horizontally rather than vertically. Double click on the horizontal axis. Toward the bottom of the dialog box there is a small open box next to the phrase "Transpose values and category scales." Click on that box and then on OK.

QUESTIONS: Look at the final histogram, the stem-and-leaf plot, and the boxplot. Would you call the graph symmetric or skewed? Would you classify the distribution as unimodal, bimodal or multimodal? (In other words, does it have one or more than one peak?) Do the numbers look fairly close together, or are they more spread out? *Looking only at your histogram*, do you see any outliers in the graph? Does it seem like the outliers that Minitab gives you with the stem-and-leaf plot or boxplot are "really" outliers? Now imagine I am giving an exam where *I allow you to look **only** at the stem-and-leaf plot and not at the actual worksheet with the original data*. Would you be able to compute the mean amount of money paid for a pair of jeans? The median? The mode? *If I allowed you to look **only** at your histogram*, would your answer about the mean, median, and mode change? Explain.

3. In your spreadsheet, you will notice that only three of the variables have numbers for data (*Earners*, *JeanShop*, and *PayJean*), and the rest are categorical.

   QUESTIONS: If you calculate summary statistics **Stat> Basic Statistics>Display Descriptive Statistics** (mean, median, etc.) for each of these three numerical variables, are those summary statistics meaningful? (Note: You should look at **Jeans Description** to see what each of these variables represents before answering this question.) If there are any 999 values in these columns, you need to replace them with * before calculating summary statistics (as you did for PayJean).

4. Make a frequency table for the variable *Income*. (**Stat>Tables>Tally Individual Variables**, and check the boxes that say "Counts" and "Percents". Make sure you have replaced values of 999 with Minitab's missing value designation.) You'll notice that the incomes do not come out in a logical order, because Minitab is trying to sort

them alphabetically (*Income* is an ordinal categorical variable). You can change this by setting up a new "sort order." To do this in Minitab:

- Select the *Income* column (you can click at the very top of the column, where it says "C7-T").
- Once this is highlighted, go to **Editor>Column>Value Order**, then click on "User-Specified Order."
- In the window that appears, scroll to the top. Cut and paste to rearrange the values of Income in a logical order: Under15, 15to24.9, 25to34.9, 35to49.9, 50to74.9, 75to99.9, and 100Plus.
- Click Add Order.

Once this is done, go back to **Stat>Tables> Tally Individual Variables** and make another frequency table.

QUESTIONS: Based on what you know, does this set of data seem to have lots of people that are either "high" or "low" in income, or are they fairly "typical" earners? Do you think the range of incomes would be similar to the people who live in your state?

5. Minitab will compute mean, median, and mode, along with a number of other statistics, by choosing **Stat>Basic Statistics>Display Descriptive Statistics**. Click on the **Statistics** button and select mean, median and mode. (Remove the check marks from the other boxes.) Compute these descriptive statistics for the variable *PayJean*.

   QUESTIONS: You will notice that the mean of *PayJean* is larger than the median. Could you have predicted this from your graphs from question (2)? Why or why not?

6. You can also compute descriptive statistics for *PayJean* sorted by other variables, such as marital status. To do this, repeat question (5) but also click on the "By Variable" box, and select *Marital*.

   QUESTION: Does it appear that marital status has an influence on how much is spent on jeans? Explain.

7. Make comparative boxplots for *PayJean* by *Marital* status. Notice that you will see a line for a boxplot corresponding to the missing value *Marital* status. You need to remove this from your graphic display. Here's how:
   - Graph>Boxplot
   - Click With Groups
   - Click the Data Options button and then the Group Options tab. Click both boxes (to uncheck them) to remove the missing values and missing cells.
   - Click OK and then OK.
   - If you don't like the orientation of the plot, double click on the horizontal axis. Check the Transpose value and category box at the bottom of the dialog box. Click OK.

What, if anything, can you learn from this display? What does it tell you about the spending habits (on jeans) of people of different marital status?

8. Save your work at the end of class. Go to **File>Save Project As** and select where you want the files saved. Note that when you save your work *as a project* and open it later, Minitab will save your open graphs, calculations, and the worksheet.

Minitab Lab: Student and Parent Height Relationships

Background: Although it is a valuable experience for students to work with data from important studies, students are often more interested in data directly related to them or, better yet, data that they collect themselves. The data for this lab consist of female student heights and their mothers' heights, male student heights and their fathers' heights, which were collected from three sections of an introductory statistics course (Spring 2013). Consider collecting your own data. However, if you can't spare the time, use the data in Table 1, which follows the lab.

In this lab, students make scatterplots (including one that has male and female data on the same plot), fitted line plots, run regression to find the equation of the least-squares line, make residual plots, compute correlation, and make predictions. In an extension designed to be used later in the semester, students use multiple regression, decide if each of their two predictor variables adds significantly to the model, make a prediction (including a prediction interval), and compare their prediction to one made using a simple linear regression model. The extension shows how a lab given while covering descriptive statistics can be revisited later in the semester when the topic turns to inference.

The Lab: The lab begins with a brief introduction, which includes the source of the data and some overarching questions that students should be thinking about during the lab.

Student and Parent Height Lab

The data for this lab were collected from three sections of an introductory statistics course. In this lab, you will investigate the relationships between female students' heights and their mothers' heights, male students' heights and their fathers' heights, and students' heights and their parents' heights. How different are these relationships? What role does gender play? You'll have to work through the lab to find out.

1. Begin by finding descriptive statistics for each variable.

QUESTIONS :

a) Compare the means: Is the mean height of female students close to the mean height of the mothers? What about the mean height of male students compared to the mean heights of their fathers? Is the mean height of male students greater than the mean height of female students?
b) Now that you've compared means, compare the variability.

2. a) Construct a scatterplot of *Female Student Height* versus *Mother Height*. (In other words, use *Mother Height* as the explanatory variable – so it goes on the horizontal axis.).

- Graph>Scatterplot
- Select Simple and click OK.
- Enter *Female Student Height* as Y and *Mother Height* as X.
- Click OK

b) Construct a scatterplot of *Male Student Height* versus *Father Height*.

QUESTIONS:

a) Compare the patterns of the two scatterplots. Would you describe either pattern or both as linear?

b) From a quick look at the scatterplots, which relationship appears stronger, the one for the males or the one for the females?

3. Make a scatterplot in which you can visualize three variables: student height, parent height, and student gender. Use *Parent Height* for the horizontal axis. Use different symbols for data from males and females. In order to do this, stack the height data. Here's how:

- Label columns C14, C15, and C16 *Student Height*, *Parent Height*, and *Gender* respectively.
- Data>Stack>Columns.
- For Stack the following columns: select *Female Student Height* and *Male Student Height* in that order.
- Click the circle for Column of current worksheet and select *Student Height*.
- Store subscripts in *Gender*.
- Click OK.
- Data>Stack>Columns; for Stack the following columns: select *Mother Height* and *Father Height*. Store these data in the *Parent Height* column. Do not store subscripts. Click OK.

Next, we'll make the scatterplot.

- Graph>Scatterplot.
- Select With Groups and click OK.
- Enter the variables *Student Height* and *Parent Height* in the appropriate order.
- Enter *Gender* for Categorical variables for grouping. Click OK.

QUESTION:

Write a brief description about what can be learned from the scatterplot in question 3.

4. Next, we want to find a relationship that can be used to predict (or describe) *Female Student Height* (Y) in terms of *Mother Height* (X). Determine the equation of the least-squares line (estimated regression line) and the corresponding residuals. In addition, overlay a graph of the least-squares line on a scatterplot of the data. Also find the correlation between *Female Student Height* and *Mother Height*. Instructions follow:

Regression, Residuals

- Select: Stat > Regression > Regression
- Dialogue box: Select *Female Student Height* for the response variable and *Mother Height* for the predictor variable (explanatory variable).
- Click the Storage button. Under Storage, click the box for Residuals. Click OK twice.

Scatterplot of data and graph of least-squares line

- Stat>Regression>Fitted Line Plot
- Dialogue box: Select *Female Student Height* for the response variable and *Mother Height* for the predictor variable.
- The "Type of Regression Model" should be Linear.
- Click OK.

Checking the adequacy of the model: Residual plot

Here are two methods for creating a residual plot:

- Select Graph>Scatterplot; make a scatterplot of the Residuals (RESI1) versus *Mother Height*.
- Select Stat>Regression>Regression; click Graphs; in the Residuals-versus-the-variables box, select *Mother Height*. Click OK twice.

Finding the correlation

- Stat>Basic Statistics>Correlation
- Select *Female Student Height* and *Mother Height* and click OK.

QUESTIONS:

a) What is the equation of the least-squares line? Does the line appear to summarize the pattern in the dots?
b) Notice that an additional column of data appears on your worksheet: RESI1. Focus on the residuals in RESI1. Do the residuals appear balanced between positive and negative values? (You might want to use a graphic display to help answer this question.)
c) Is the least-squares line adequate to describe the pattern in these data? Base your answer on your residual plot.
d) If a mother's height was 64 inches, how tall would you expect her daughter to be?
e) If two mothers' heights differed by 1 inch, by how much would you expect their daughters' heights to differ? Explain.
f) What is the correlation between Female Student Height and Mother Height? Based on this correlation, would you describe the strength of the linear relationship as strong, moderate, or weak? Explain.

5. This time fit a least-squares line to the *Male Student Height* and *Father Height* data.

QUESTIONS:

a) Is the least-squares line adequate to describe the relationship between *Male Student Height* and *Father Height*? Justify your answer.

b) Is the strength of the relationship between *Male Student Height* and *Father Height* stronger, weaker, or similar in strength to the relationship between *Female Student Height* and *Mother Height*? Justify your answer.

6. Finally, fit a least-squares line to the *Student Height* and *Parent Height* data. Check the adequacy of your model.

QUESTIONS:

a) Compare the three models, the three least-squares equations.
b) Interpret their slopes in the context of these data.
c) Based on correlation, which of the three relationships is the strongest?

Extension – Multiple Regression and Inference for Regression

Build a model that predicts student height from parent height and gender. How does adding Gender change R-SQ and s?

1. Recode the data in Gender to 0 for Females and 1 for Males.

- Label a column Gender Num.
- Data>Code>Text to Numeric.
- Code data from columns: *Gender*; Store coded data in *Gender Num*.
- Original Values: Use quotes – "Female Student Height" code to 0; "Male Student Height" code to 1.
- Click OK.

2. Create a model for predicting student height using two predictors – *Parent Height* and *Gender Num.*

- Stat>Regression>Regression
- Enter Student Height for the response and Parent Height and Gender Num for the Predictors.
- Click Graphs and check Residuals versus Fits.
- Click OK.

QUESTIONS:

a) Discuss the adequacy of the model in describing the pattern in these data.
b) Do both predictor variables *Parent Height* and *Gender Num* contribute significantly to the model?
c) How does the addition of the variable *Gender Num* affect the values of both R-SQ and *s*?
d) Predict the height of a female student whose mother is 67 inches. Give both a point estimate and a 95% prediction interval estimate. Compare this prediction to predictions made using the model from question 4 above.
    - Stat>Regression>Regression
    - Enter *Student Height* for the response; enter *Parent Height* and *Gender Num* for the Predictors.

- Click Options and enter 67 0.
- Click OK.
- Scroll to the bottom of output and read off values for Fit and 95% PI.

Females				Males			
Student	Mother	Student	Mother	Student	Father	Student	Father
62	61	67	62	75	74	71	72
66	66	67	64	72	70	72	70
68	64	65	64	70	68	74	72
60	62	66	65	73	74	73	68
68	66	62	62	68	71	69	72
64	61	64	69	73	70	72	70
62	65	69	70	67	61	80	70
67	64	64	66	68	68	69	68
66	70	57	62	65	67	72	78
63	66	64	60	71	71	70	68
68	66	63	65	70	68	70	68
64	63	60	63	75	76	68	68
60	62	62	59	73	70	66	68
69	65			70	67	69	68
66	65			71	74	72	68
66	67			69	68	72	68

Table 1. Data on students' and parents' heights in inches.

Conclusion

We have found Minitab to be an ideal statistical package for use in introductory statistics. The commands for carrying out calculations and for producing visual displays are fairly intuitive and straightforward for students to follow. Because we embed the commands directly into our laboratory assignments (at least for the first time they are needed), students are able to carry out these commands quickly and easily, allowing them to focus most of their attention on statistical concepts. Often labs are used for guided discovery and serve as a vehicle to introduce concepts before extensive class discussion takes place.

During the minicourse, we shared additional Minitab labs that cover other statistical concepts such as correlation, the sampling distribution of a sample mean, confidence intervals, and statistical inference for population means. Copies of these Minitab labs are available from the authors upon request.

References

Blair, Richelle M., Ellen E. Kirkman, and James W. Maxwell (2013). *Statistical abstract of undergraduate programs in the mathematical sciences in the United States, Fall 2010 CBMS Survey* (Providence, RI: American Mathematical Society).

Cobb, George W. (1991). "Teaching statistics: more data, less lecturing," *Amstat News*, December 1991, pp. 1, 4.

Zieffler, Andrew, and Karl, Stacy, Eds. (2010), *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report*, American Statistical Association.