Your Name: ____________________

Section: ____________________

36-201 INTRODUCTION TO STATISTICAL REASONING
Computer Lab #4
Scatterplots and Regression

Objectives:

1. To learn how to interpret scatterplots. Specifically you will investigate, using a graphical display, the direction and the strength of the relationship between two quantitative variables.
2. To develop skills in describing the relationship between two quantitative variables.
3. To use case identification to help interpret scatterplots.

Getting Started

COPYING FILES

1. Open the **Student** folder. To do this, double click on the "Cluster HD" icon. Then double click the "documents" folder. You will eventually move the data into this folder.
2. In the following order: From the menu in the upper left hand corner of the titlebar, choose **Chooser**. In the dialog box that appears *click* **AppleShare**. In the bottom left box called "AppleTalk Zones" *select* **BH**. In the top right box where it says "Select a file server" select **HSSHELIOS**. *Click* **Ok**.
3. In the dialog box that appears login using "36201" as your Name and "36201" as your password. Click **Ok**. In the dialog box that appears highlight "Class." Click **Ok**.
4. A "Class" icon will appear to the right. Double click on the icon. Then double click on 36201. You will see many files in a folder.
5. Copy all of the files that you will need to the **Student** folder. Do this by dragging the files into the Student folder. For today, the data files you need are *memory.mtw* and *lab4.mtw*.
6. Close all windows.
7. **IMPORTANT:** Drag the *class* icon into the trash. This is necessary to let others in the lab get access to the data.

To start Minitab:

1. *Double-click* the "Cluster HD" icon.
2. *Double-click* the Applications icon.
3. *Double-click* the Minitab 10.5 folder.
4. *Double-click* Minitab 10.5.

Part I. Introduction to Scatterplots

In this section you will learn about **scatterplots**, a graphical display for examining the relationship between two quantitative variables.

Statistical Background

When we are interested in the relationship between two variables, an observation consists of a pair of measurements. For example, we may be interested in the relationship between SAT score and Freshman QPA. An observation for each subject would consist of that subject's SAT score and that subject's Freshman QPA. A **scatterplot** is a graphical display that shows the *direction* and the *strength* of the relationship between two quantitative variables.

Example: Memory Experiment

To discover how rapidly people forget, a psychologist slowly read lists of 20 words to five people and asked them later to recognize the words when mixed with 20 other words. A score was assigned to indicate the average percent of words correctly recognized. The time interval before the subject was asked to recognize a list of words was varied from 1 minute to 1 week.

Question #1 As time increased what do you think happened to the average score?

You will now read the *memory.mtw* data file into Minitab. This file has been saved in a special Minitab format.

- From the **File** menu, choose **Open Worksheet**.
- *Click* the **Desktop** button on the right side of the dialog box. Select "Cluster HD." Click **Open.** Select the "Student" folder. Click **Open.**
- Select the file called "memory.mtw" and then *click* **Open**.

In the Worksheet window, the first variable, SCORE, is the average percent of words correctly recognized, and TIME, the second variable, indicates the amount of time elapsed since the subjects were introduced to the list of words. An observation consists of the pair of variables, the SCORE obtained at a given TIME. In this example SCORE is called the **response** or **Y** variable and TIME is the **explanatory** or **X** variable because we wish to see the effect of time on score.

To create a *scatterplot* of SCORE versus TIME, go to the **Graph** menu, choose **Plot**. Make sure the cell in the first row under the column labeled Y is highlighted. Type SCORE and press the **Right** arrow key on the keyboard (or just double-click on SCORE in the list of variable names on the left). Make sure the cell in the first row under the column labeled X is now highlighted. Type or double click on TIME. *Click* **OK**. A scatter plot will appear. It will help to make the scatterplot window larger so that you can see the plot better.

Question #2 Find the point on the scatterplot that has the value TIME approximately equal to 6000. What is the value of SCORE?

♣ **Question #3 i) Based on the scatterplot how would you describe the relationship between** SCORE

and TIME. Specifically, how does SCORE change as TIME increases? Would you say the relationship between average SCORE and TIME is an increasing relationship, decreasing relationship, or that there is no relationship? ii) Would you say the relationship between average SCORE and TIME was linear (i.e. the data fall approximately on a line) or nonlinear (i.e. the data fall approximately on a curve)?

CORRELATION COEFFICIENT

A *correlation coefficient* is a numerical summary indicating the direction and the strength of the linear relationship between two quantitative variables. The correlation coefficient, denoted by r, takes values between –1 and 1.

- A value of r greater than 0 indicates a *positive association* between the two variables;
- A value of r less than 0 indicates a *negative association* between the two variables;
- A value of r close to 0 indicates *none or very little linear association* between the two variables.
- A value of r close to 1 or –1 indicates a *strong linear association.*

To find the correlation coefficient to summarize numerically the relationship between SCORE and TIME, go to the **Stat** menu, and from the **Basic Statistics** sub-menu, choose **Correlation**. *Click* the box under "Variables" and type SCORE TIME (or double click on the names SCORE and TIME in the list of variable names). *Click* **OK**. The correlation coefficient for the two variables appears in the Session window.

Question #4 What is the value of the correlation coefficient between SCORE and TIME?

Question #5 Does the sign of the correlation coefficient (i.e., negative or positive) agree with your answer to question #3? Is the value of the correlation coefficient closer to –1 or to 0? Does this correlation coefficient indicate a strong, mild, or weak relationship between SCORE and TIME? Is the relationship *really* linear?

You have now finished this section. From the **File** menu, choose **Restart Minitab**. It will ask you if you want to save changes to the Session window. Answer **No**. You should see blank Session and Worksheet windows.

Part II: Scatterplots and Regression

To illustrate a more detailed analysis of the relationship between two variables using scatterplots, we will look at height and weight data for 36 male and female students enrolled in a statistics class.

- From the **File** menu, choose **Open Worksheet**.
- *Click* the **Desktop** button on the right side of the dialog box. Select "Cluster HD." Click **Open.** Select the "Student" folder. Click **Open.**
- Select the file called "lab4.mtw" and then *click* **Open**.

The first variable, GENDER takes the values $0 =$ male, and $1 =$ female. The second variable is HEIGHT in inches and the third variable is WEIGHT in pounds.

We are interested in studying the relationship between weight and height for these students and, specifically, would like to know whether this relationship is different for male students than for female students.

Question #6 State what type of variable (quantitative or qualitative) each of the variables GENDER, HEIGHT, and WEIGHT are. Find the first male in the data set and write down his height and weight.

Make a boxplot for WEIGHT. Make a separate boxplot for HEIGHT. Does the boxplot for height tell you anything about the distribution of weight? The answer is no. There is nothing in the boxplot of height that directly relates the height of a person to that person's weight or vice versa.

Question #7 i) Looking at the boxplot for weight, do you think that the mean of this distribution is greater than, less than, or equal to the median? Why? ii) Looking at the boxplot for height, do you think that the mean of this distribution is greater than, less than, or equal to the median? Why?

Close the boxplot windows. We now want to study the relationship between HEIGHT and WEIGHT using the scatterplot, a graphical display that does tell you something about the relationship between two variables. When we say that two variables are related to each other, we mean that if we know something about the value of one of the variables that will tell us something about the value of the other variable. For example, in the first part of this lab, we looked at data from the memory study. If you knew that the value of *time* (the explanatory variable) was large then you knew that the value of *score* (the response variable) was small. This fact was reflected in the scatterplot by a very clear decreasing relationship between *score* and *time*. Also it was reflected in a correlation coefficient that was close to -1. Recall that the response variable is usually denoted by Y and the explanatory variable is denoted by X.

Exploratory scatter plot.

Question #8 For these height and weight data, suppose we wish to use height to predict weight. Which is the response? Which is the explanatory variable?

Recall, to make a scatterplot of WEIGHT (Y-variable) versus HEIGHT (X-variable), from the **Graph** menu, choose **Plot**. Type WEIGHT in the cell in the first under the column labeled Y and press the **Right** arrow key, or just double click on WEIGHT. Make sure the cell in the first row under the column labeled X is highlighted. Type (or double click on) HEIGHT and *click* **OK**.

♣ Question #9 i) Describe the relationship between weight and height. Is the relationship increasing, decreasing or is there no relationship? Is it linear or nonlinear? This relationship is called the *trend*.

ii) Are there any points that are outliers, i.e., points that do not seem to follow the trend?

iii) As height increases comment on what happens to the spread of the weight variable?

Exploratory regression.

To get a better understanding of the trend, we can use "exploratory regression" or "smoothing." To do this, imagine dividing the X-axis into intervals. Then form side-by-side boxplots and join the medians. This is called a "median trace plot." Minitab can make a plot similar to a median trace plot, called a "Lowess plot;" actually, it will display the final trace plot without drawing the boxplots.

From the **Graph** menu, choose **Plot.** You should still have WEIGHT as the Y-variable and HEIGHT as the X-variable. Under **Data Display**, in the column marked "Display", highlight the second cell (the blank cell below "Symbol"). From the pop-up (arrow) menu next to **Display**, choose **Lowess.** Now highlight the second cell in the column marked "For each". From the pop-up (arrow) menu next to **For Each**, choose **Graph**. Click **Ok**.

♣ Question #10 Describe the lowess plot. Is there an increasing or decreasing relationship? Is the relationship linear or nonlinear?

Linear regression.

The "regression line" is the line that best fits the data. We can also get Minitab to plot the regression line. Select the **Stat** menu, then select **Regression** followed by **Fitted Line Plot**. In the box next to **Response (Y)** type (or double click on) WEIGHT. In the box next to **Predictor (X)** type (or double click on) HEIGHT. Click **Ok.**

Question #11 Does the regression line fit the data well? How does it compare to the lowess plot?

We can also get Minitab to give us the equation for the regression line. Select the **Stat** menu, then select **Regression** followed by **Regression**. In the box next to **Response (Y)** type (or double click on) WEIGHT. In the box next to **Predictor (X)** type (or double click on) HEIGHT. Click **Ok.** Lots of information will appear in the Session window. Somewhere it will say "The regression equation is ..."

♣ **Question #12. Write down the regression equation. Suppose you have a friend who is 65 inches tall. Use the regression equation to predict her weight.**

Question #13 What is the value of the correlation coefficient between WEIGHT and HEIGHT? (Recall you can find the correlation coefficient by choosing from the **Stat** menu, **Basic Statistics**, and click-on **Correlation** in the sub-menu. Drag over the height and weight variables in the list box and *click-on* **Select**. Then *click* **OK**.

Question #14 Does the sign of the correlation coefficient (i.e., negative or positive) agree with your answer to question #9? Is the value of the correlation coefficient closer to 1 or to 0? Does this correlation coefficient indicate a strong, mild, or weak relationship between WEIGHT and HEIGHT?

<u>Case Identification</u>

We now want to indicate in the scatterplot which observations are males and which are females, i.e., we want to include a third variable in the scatterplot in order to see how the relationship between weight and height might be different by sex. In other words we want to identify cases that are males and cases that are females. From the **Graph** menu, choose **Plot**. This should bring up the Dialog Box for the scatterplot that you previously created.

- Click on the first cell in the column labelled "For each", then from the pop-up (arrow) menu next to "For Each", choose **Group**. Do the same thing for the second cell in the "For each" column.
- Click on the first cell in the "Group Variables" column. Type or double click on GENDER. Do the same thing for the second cell in the "Group variables" column.

Click **OK**. You should now see two types of symbols on the scatterplot, one for males and one for females.

Question #15 Describe the general location in the scatterplot where the females fall and where the males fall with respect to height and weight. Describe why you think this is so.

♣ **Question #16 Describe the relationship between** HEIGHT **and** WEIGHT **in the scatterplot, taking gender into account. Specifically, does it look like the weight of females increases with an increase in height as quickly as the weight of males increases with a corresponding increasing change in height?**

To quit Minitab, from the **File** *menu, choose* **Quit***. Do not save any files. Remember to* **delete** *any files and folders that you might have created.*