

STATISTICS 101 - Homework 3  
Due Wednesday, June 8, 2005

- Homework is due on the due date at the end of lecture.
- You may talk with others about the homework problems but please write your solutions up independently.
- Please answer homework questions in complete sentences. Make sure to **staple** the pages of your assignment together. **Be sure to indicate your lab section on your paper.**
- You will have an opportunity to get help on homework during lab.

**Reading:** Chapters 7,8 and 9

**Problems:**

1. An educational foundation would like to give scholarships to high school seniors who will be successful in college. The foundation wishes to see if there is a relationship between the score on a verbal aptitude test as a predictor of success in college and thus help them decide who should get the scholarships. The verbal aptitude test is on a scale of 200 to 800 and GPA is on a scale from 0 to 4. The plot that is appended to the back of this assignment is a plot of GPA versus the verbal aptitude test score for 50 students randomly selected from all students at a large public university.
  - (a) From the plot, what is the lowest GPA? What verbal aptitude score is associated with the lowest GPA?
  - (b) From the plot, what is the highest GPA? What verbal aptitude score is associated with the highest GPA?
  - (c) Describe the general pattern of the relationship between verbal aptitude score and GPA.
  - (d) The value of the correlation coefficient for these 50 pairs of verbal aptitude score and GPA is 0.516. However, there appears to be an unusual pair or outlier. What are the verbal aptitude score and GPA for that apparent outlier? If this apparent outlier were removed, would the correlation coefficient calculated using the remaining 49 students be smaller than, about the same as, or larger than the 0.516? Explain briefly.
  - (e) Below are summary data for the 49 observations after eliminating the one outlier. Calculate the value of the correlation coefficient,  $r$ . Does this agree with your assessment in (d)?

$$n = 49 \quad \sum Y = 135.3 \quad \sum X = 29,674$$
$$\sum (X - \bar{X})(Y - \bar{Y}) = 1083.0 \quad \sum (Y - \bar{Y})^2 = 16.71 \quad \sum (X - \bar{X})^2 = 246,074.88$$

2. Can the length of a person's forearm be used to predict the length of a person's foot? We will collect data on the two variables during Lab 4. The data below were taken from a sample of 25 women during a previous semester.

Forearm (cm)	Foot (cm)	Forearm (cm)	Foot (cm)	Forearm (cm)	Foot (cm)
24	24	23	25.5	24.5	25.5
24	27.5	26	27	26	28
25	30	24.5	25	26	30
25.5	26.5	27.5	28	26	28
23.5	27	24.5	27.5	25	24
27	28.5	26	25.5	24	27
29	31	24	26	27	26.5
27	29	25	25	28	27
28	28				

- Start the computer program JMP. Select **File** → **New** → **Data Table** from the JMP Menu. Double-click on Column 1 in the Table and type in the name **arm**. Click on the red triangle next to **Columns (1/0)** and select **New Column**. In the window that appears, change the Column Name to **foot**. Then click OK. You should now have two columns in your data table, **arm** and **foot**. Enter the data from the table above into JMP. You should have 25 rows in your data table when you are finished.
  - We want to look at the relationship between the two variables, **arm** and **foot**. Specifically, we would like to predict a person's foot length from the length of their forearm. From the JMP menu, select **Analyze** → **Fit Y by X**. Select the column **foot** and click on the button **Y, Response**. Select the column **arm** and click on the button **X, Factor**. Then click on **OK**.
  - You should have a scatterplot of the two variables with the variable **arm** on the  $x$  axis and the variable **foot** on the  $y$  axis. To add the regression line to the scatterplot, click on the red triangle next to **Bivariate Fit of foot By arm** and select **Fit Line**. This should add a regression line to the scatterplot and statistics for the regression line to the window.
  - To get a complete picture of all regression lines, we must study the residual plot. Click on the **red triangle** next **Linear Fit** and select **Plot Residuals**. A residual plot should be added to the bottom of the window.
  - From the JMP menu, select **File** → **Print** to print your output. Turn this paper in with your assignment. You will need this output to answer the questions below.
- (a) What is the explanatory variable? What is the response variable? Briefly explain your choice.
  - (b) Describe the general relationship between the two variables.
  - (c) Give the value for the slope of the least squares regression line. Give an interpretation of this value within the context of the problem.
  - (d) Give the value for the intercept of the least squares regression line. Give an interpretation of this value within the context of the problem. Does this interpretation make sense? Explain your answer.

- (e) Give the equation of the least squares regression line for this problem. Use this equation to predict the length of a person's foot given the length of their forearm is 29 cm.
- (f) One woman had a forearm length of 29 cm. What is the residual for this woman?
- (g) Would you use the least squares regression line to predict the length of a person's foot if their forearm length was 32 cm?
- (h) Give the value of  $R^2$  for this regression. Give an interpretation of this value in the context of the problem.
3. We often hear reports about the relationship between diet and health. The data below give fat intake (grams) *per capita* per day (X) and the death rate (deaths per 100,000 people) from colon cancer (Y) for thirty nations in the year 1975.

Nation	Fat Intake per day (X)	Death rate (Y)	Nation	Fat Intake per day (X)	Death rate (Y)
Phillipines	28	4.5	Czechoslovakia	95	15.0
Japan	39	3.0	Hungary	100	14.0
Taiwan	46	3.5	Spain	101	8.5
Colombia	47	5.1	Finland	115	14.0
Chile	52	9.0	Austria	118	17.2
Panama	55	7.5	Australia	128	19.0
Mexico	57	3.9	Norway	129	17.0
Bulgaria	68	9.0	Sweden	129	18.5
Portugal	70	13.0	Ireland	134	21.5
Yugoslavia	70	7.0	Switzerland	138	22.0
Hong Kong	71	10.0	Belgium	140	21.0
Puerto Rico	77	5.5	United Kingdom	142	24.5
Italy	87	16.0	Canada	143	23.0
Poland	90	10.5	United States	148	21.0
Greece	95	7.5	New Zealand	152	23.0

- Start the computer program JMP. Select **File** → **New** → **Data Table** from the JMP Menu. Double-click on Column 1 in the Table and type in the name **Fat** . Click on the red triangle next to **Columns (1/0)** and select **New Column**. In the window that appears, change the Column Name to **Death**. Then click OK. You should now have two columns in your data table, Fat and Death. Enter the data from the table above into JMP. You should have 30 rows in your data table when you are finished.
- First we want to look at the distributions of fat intake and death rate separately. Select **Analyze** → **Distribution** from the JMP menu. Select the column **fat** and click the button **Y, Columns**. Also, select the column **death** and click the button **Y, Columns**. Then click on **OK**.
- You should now have a histogram, boxplot, and statistics for both variables. We want to make a few changes to the information JMP has calculated. First, click on the **red triangle** next to **fat** and select **Stem and Leaf**. Now, click on the **red triangle** next to **fat** and select **Histogram Options** → **Count Axis**. Finally, click on the **red triangle** next to **fat** and select **Display Options** → **Horizontal Layout**. Repeat this process for the variable **death**.

- From the JMP menu, select **File** → **Print** to print your output. Turn this paper in with your assignment. You will use this output to answer the questions at the end of the assignment.
- Now, we want to look at the relationship between the two variables, **fat** and **death**. Specifically, we would like to predict the death rate due to colon cancer from the fat intake per capita. From the JMP menu, select **Analyze** → **Fit Y by X**. Select the column **death** and click on the button **Y, Response**. Select the column **fat** and click on the button **X, Factor**. Then click on **OK**.
- You should have a scatterplot of the two variables with the variable **fat** on the  $x$  axis and the variable **death** on the  $y$  axis. To add the regression line to the scatterplot, click on the red triangle next to **Bivariate Fit of death By fat** and select **Fit Line**. This should add a regression line to the scatterplot and statistics for the regression line to the window.
- To get a complete picture of all regression lines, we must study the residual plot. Click on the **red triangle** next **Linear Fit** and select **Plot Residuals**. A residual plot should be added to the bottom of the window.
- From the JMP menu, select **File** → **Print** to print your output. Turn this paper in with your assignment. You will need this output to answer the questions at the end of the assignment.

**Questions:** Use your output to answer the following questions.

- (a) Describe the distributions of fat intake and death rate. Make sure to include in your description the five number summary, the mean and standard deviation, and the shape of the histogram. Are there any outliers?
- (b) Describe the scatterplot of fat intake vs. death rate. Give the regression equation for predicting death rate from fat intake, give an interpretation of the slope of the regression equation, and give an interpretation of the  $R^2$  value for the regression. Finally describe the residual plot, and make note of any potential problems with the regression.

College GPA vs. Verbal Aptitude Score

