

### **Chapter 3 Association: Contingency, Correlation, and Regression**

#### **Simpson's Paradox**

The University of California at Berkeley was charged with having discrimination against women in the graduate admissions process for the fall quarter of 1973. The table below identifies the number of acceptances and denials for both male and female applicant in each of the six largest graduate programs at the institution at that time.

	Men accepted	Men denied	Women accepted	Women denied
Program A	511	314	89	19
Program B	352	208	17	8
Program C	120	205	202	391
Program D	137	270	132	243
Program E	53	138	95	298
Program F	22	351	24	317
Total				

1. Start by ignoring the program distinction, collapsing the data into a two-way table of gender by admission status. To do this, find the total number of men accepted and denied and the total number of women accepted and denied. Complete the table below.

	Accepted	Denied	Total
Men			
Women			
Total			

2. Consider for the moment just the *men* applicants. Of the men who applied to one of these programs, what proportion was accepted? Now consider the *women* applicants; what proportion of them were accepted? Do these proportions seem to support the claim that men were given preferential treatment in admissions decisions?

3. To try to find the program(s) responsible for the male preferential treatment, calculate the proportion of men and the proportion of women *within each program who* were accepted. Record your results in the table below.

	Proportion of men accepted	Proportion of women accepted
Program A		
Program B		
Program C		
Program D		
Program E		
Program F		

4. Does it seem as if any program is responsible for the large discrepancy between men and women in the overall proportions accepted?

5. Reason from the data given to explain how it happened that men had a much higher rate of admission overall even though women had higher rates in most programs and no program favored men very strongly.

### Constructing a Scatterplot and Calculating Correlation and Regression on a TI-83

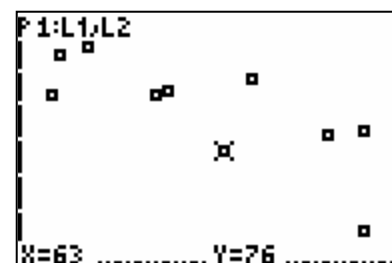
The following table gives the times (in minutes) that a random sample of ten students took to take a statistics test and their scores on the test.

Time	63	53	39	78	67	83	43	83	55	38
Score	76	88	97	79	92	58	99	80	89	88

- **Step 1: Enter the data.** Press [STAT] [1] and enter the times in L1 and the scores in L2.
- **Step 2: Setting up the scatterplot.** Press [2nd] [Y=] to go to the Stat Plot window. Under **Type:**, the first one is the scatterplot (press [ENTER] on that) and make the **Xlist** L1 and the **Ylist** L2.
- **Step 3: Setting up the viewing window.** To set up an appropriate viewing window simply press [ZOOM] [9]. By pressing [TRACE] you can see the points on the scatterplot.

L1	L2	L3	2
67	92		
83	58		
43	99		
83	80		
55	89		
38	88		
-----			
L2(11) =			

Plot1	Plot2	Plot3
Off	Off	Off
Type:		
Xlist: L1		
Ylist: L2		
Mark:		



- **Step 4: Calculating correlation and regression.** To find the regression equation that can be used to predict the score on the test given the time needed to take the test and to find the correlation between these two variables you need to go to one of the linear regression functions on the calculator.

To do this press [STAT] [2] [8]. You should now see **LinReg(a+bx)** on the home screen of your calculator.

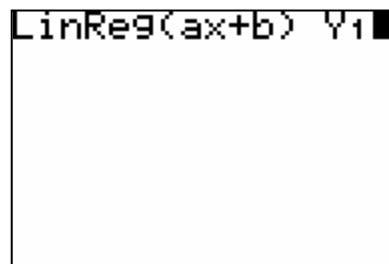
Press [ENTER] to get the  $y$ -intercept and slope of the regression equation along with  $r$  and  $r^2$ . [Hint: If the data are in columns other than L1 and L2 then you also have to select the columns after the **LinReg(a+bx)** appears.] If your calculator does not give you the correlation, a “switch” has been turned off and needs to be turned back on. To do this you need to go to the **CATALOG**, which is [2nd] [0], and toggle down to **DiagnosticOn** and press [ENTER] twice.

LinReg
y=a+bx
a=115.9006453
b=-.5199442734
r <sup>2</sup> =.5705547574
r=-.7553507512

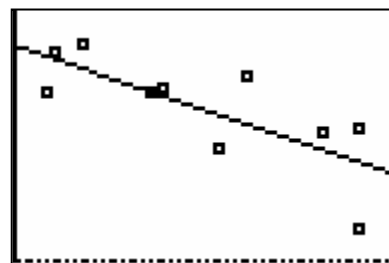
CATALOG
Degree
DelVar
DependAsk
DependAuto
det(
DiagnosticOff
DiagnosticOn

- **Step 5: Graphing the line along with the scatterplot.**

To graph the line along with the scatterplot you can copy down the equation and type it in after pressing  $\boxed{Y=}$ . There is a shortcut to this though. Before pressing  $\boxed{\text{ENTER}}$  in step 2, if you put **Y1** after **LinReg** your calculator will automatically transfer your equation in the appropriate place to be graphed. To get **Y1** after **LinReg** press  $\boxed{\text{VAR}} \boxed{\blacktriangleright} \boxed{1} \boxed{1} \boxed{\text{ENTER}}$ . Now by pressing  $\boxed{\text{ENTER}}$  again, your calculator will recalculate the regression equation and put it in place to be graphed. Making a scatterplot of the data you should now also see the regression line graphed as well.



LinReg(ax+b) Y1



### Correlation and Regression Questions

Hope College male students were asked how tall they were and their responses are the following reported heights. They were then measured and the results are the actual heights.

Reported	74.5	72	70.5	76.5	70	70.5	71	68	72	71	73	69	68	75	78.75
Actual	74.5	71.5	69.75	76.25	68.5	70	69	67.75	72	70.5	72.5	69	67.5	74.5	78

6. Make a scatterplot of these data with reported height as the explanatory variable and actual height as the response.
  - a) Is there a positive or negative relationship between the two variables?
  - b) What does that relationship mean?
7. On your scatterplot also include the line  $y = x$ .
  - a) What can you say about points that lie on that line?
  - b) What can you say about points that lie below that line?
  - c) Which point is the farthest below the line? What can you say about that person's reported height?

8. Find the correlation between reported and actual height.
9. Find the correlation between actual and reported height.
10. Suppose the last person on the list reported his height as 67 inches when he in fact was 78 inches. Change that data value, look at the new scatterplot, and find the new correlation.

Each member of a class at Hope measured the length of his or her feet and height (in centimeters). The results are shown below.

<b>Foot</b>	27	22	22	23	25	26	28	26	28	23
<b>Height</b>	183	155	163	163	163	188	178	175	185	157

11. Find the regression equation that can be used to predict someone's height given the length of their foot.
12. What does the slope of the regression equation mean in terms of foot length and height?
13. What does the  $y$ -intercept of the regression equation mean in terms of foot length and height?
14. Using the regression equation, how tall is someone with a foot length of 27 cm?
15. What is the residual for the first person on the list with a foot length of 27 cm and a height of 183 cm?
16. Make a scatterplot and include the regression equation along with the plot.
  - a) What can you say about points that are above the regression line?
  - b) What can you say about points that are below the regression line?
  - c) Which point has the largest residual?