ASSOCIATION BETWEEN VARIABLES: SCATTERGRAMS
(Like Father, Like Son)

Though it is not especially relevant to political science, suppose we want to research the following bivariate hypothesis that involves two *interval and continuous variables*.

FATHER'S HEIGHT	+	ADULT SON'S HEIGHT	[*father-son*
(actual height)	=========>	(actual height)	*pairs*]

We select a random sample of $n = 1078$ pairs of fathers and their adult sons and collect the relevant data, the first five cases of which appears below. Note that the cases are *father-son pairs* and observed values have been very precisely measured and recorded, so probably each case has a unique value on each variable.

Pair ID	*Father's Height (inches)*	*Son's Height (inches)*
1	66.67	68.42
2	69.83	70.32
3	65.19	69.76
4	65.15	73.85
5	64.66	70.17

Note that we cannot as a practical matter crosstabulate these variables, because the variables are continuous and each case would need its own row and column. So what should we do? One possibility is to create *class intervals* for both variables (as discussed in Handout #5 on histograms) — in effect, turning them into discrete variables and proceed as before. We might create these class intervals for both variables:

Short	less than 65 inches
Medium	65-70 inches
Tall	70 or greater inches

We then can set up this table worksheet and begin to tally the cases as shown below.

	FATHER'S HEIGHT		
SON'S HEIGHT	Short	Medium	Tall
Short			
Medium		#1 #3	
Tall	#5	#2 #4	

We can quickly see that this is not very satisfactory. One would be unlikely to notice the height difference of about ½ inch between father and son in pair #2, while the height difference of nearly 9 inches between father and son in pair #4 would be immediately noticeable and indeed striking. Yet given the (perfectly reasonable) class intervals we created, these very different pairs fall into the same cell of the table (tall son of medium father). On the other hand, the fathers and

sons in pairs #3 and #5 have just about the same height difference (about 5 inches), yet both members of pair #3 fall in the same (medium son of medium father) class interval while the members of pair #5 fall in the extreme opposite class intervals (tall son of short father).

These considerations illustrate the fact that creating class intervals when you have gone to the trouble of measuring continuous variable quite precisely entails *throwing away valuable information that bears on the hypothesis of interest*. This problem could be mitigated by creating more refined class intervals, but what we really should do is use the very precise information we have collected. A very nice analytical device called a *scattergram* (or *scatterdiagram* or *scatterplot*) is similar in its basic logic to a crosstabulation and allows us to do just this.

First we need to set up a scattergram template or worksheet, which is similar in logic to that for a crosstabulation but reflects the continuous character of both variables. Figure 1 shows the general template for such a scattergram. We draw a *horizontal interval scale* representing values of the *independent variable*. This scale should be appropriately labeled and calibrated to encompass the full range of observed values found in the data (but it needn't, and probably shouldn't, be much wider than this). We then erect a *vertical interval scale* that similarly represents values of the *dependent variable*. It is standard to draw the scales on the left and bottom margins of the scattergram and scale them so that the scattergram is either approximately square or is somewhat wider than tall (SPSS does the latter).

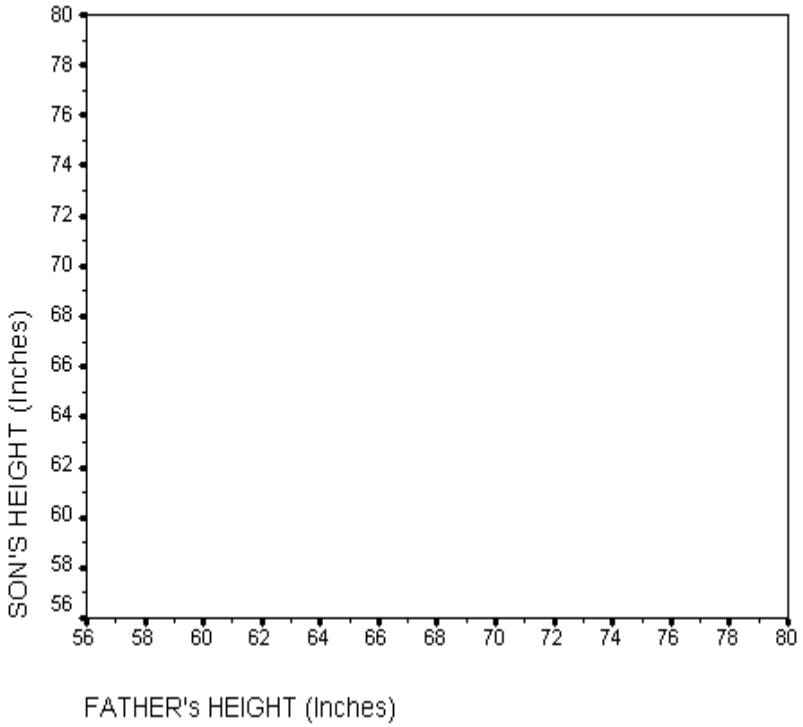

Figure 1

Just as cases are placed in cells of crosstabulation defined by the intersection of the row and column corresponding to the particular combination of (discrete) variable values that characterizes the case, each case in a scattergram is plotted at the point defined by the intersecting of horizontal and vertical lines corresponding to the particular combination of (continuous) variable values that characterizes the case. In a sense, each case falls into its own (almost always) unique and tiny cell.

Figure 2 shows a scattergram worksheet and the plotted points for each of the five father-son pairs listed above. (Clearly graph paper is very useful for making hand-drawn scattergrams, as you will do in Problem Set #11.)

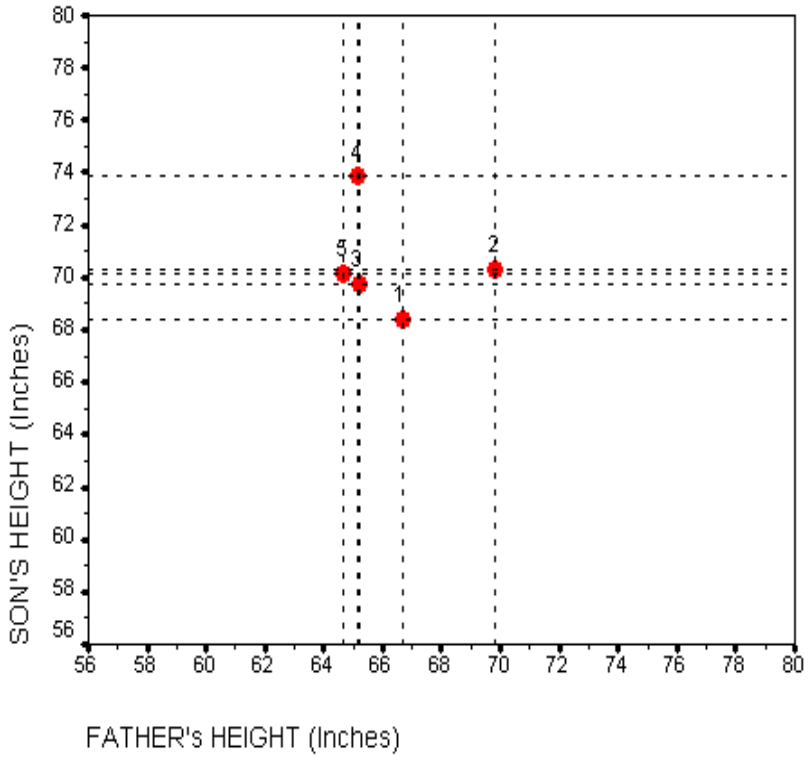

Figure 2

Such a study of father-son pairs was actually conducted by a statistician named Karl Pearson over 100 years ago in England. Having collected height data on 1078 father-son pairs, he realized that a list of 1078 pairs of numbers would be impossible to grasp as raw data and that a crosstabulation using class intervals would have the problems discussed above. Pearson therefore developed the *scattergram* and an alternate analytic device appropriate for analyzing association in such data. Figure 3 shows the scattergram of Pearson's data. (This scattergram is taken from David Freedman *et al.*, *Statistics*, p. 110. For the moment, ignore the diagonal line and the two vertical dotted lines; we will discuss them in class.) You should be able to see that the scattergram displays a substantial but far from perfect association between the two variables.

Pearson's Scattergram

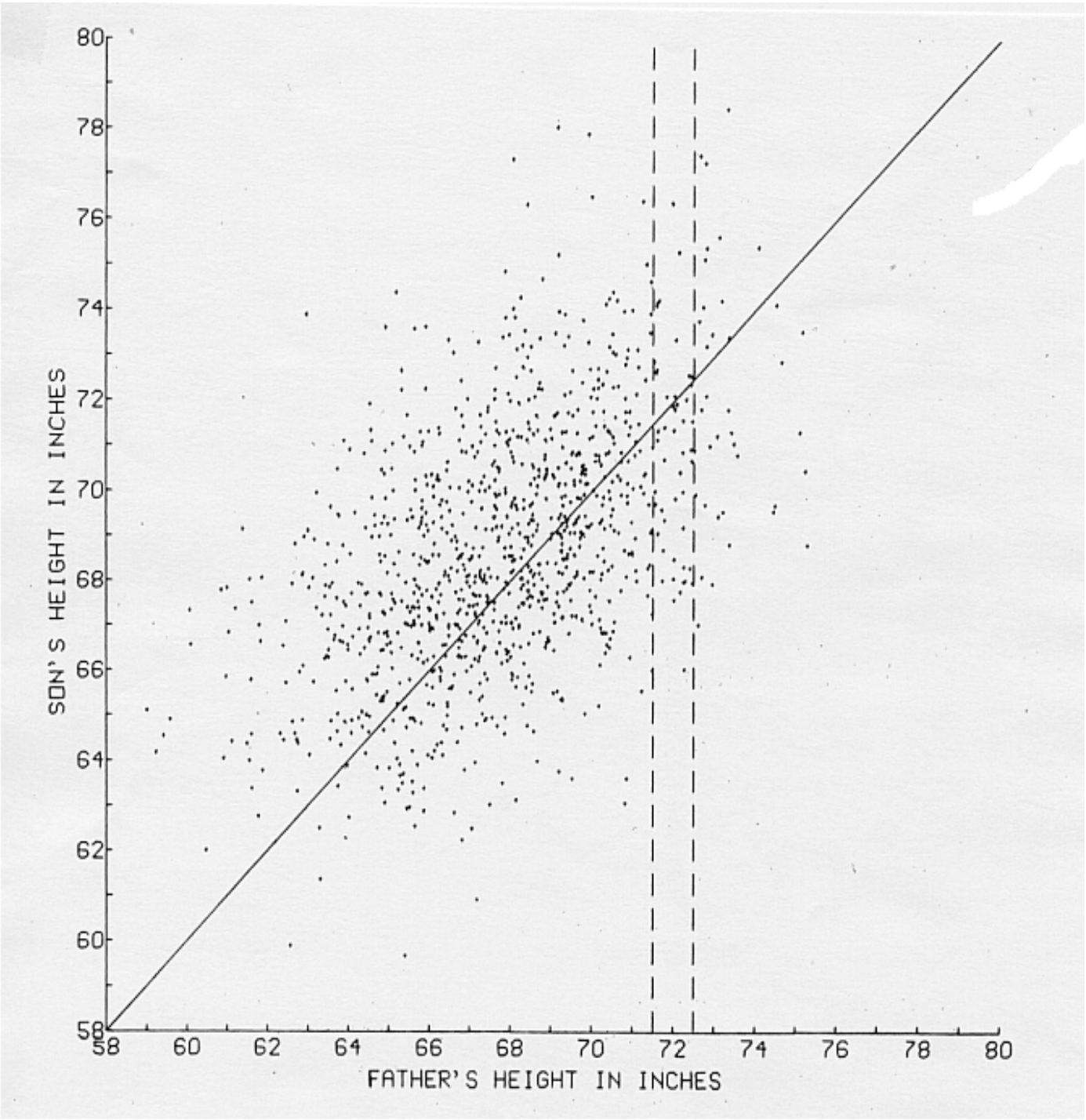

Figure 3

Figures 4-8 show other scattergrams for small hypothetical data sets. These were produced by the **Statistical "Applet"**on Correlation and Regression Demo available at the course web site

(which you are invited to play with). These five scattergrams correspond directly to Tables 1B-1F near the beginning of the previous handout on crosstabulations, in that they show differing degreees of association running from high positive through zero to high negative. The (Pearson) *correlation coefficient* is the standard measure of association between two interval variables.

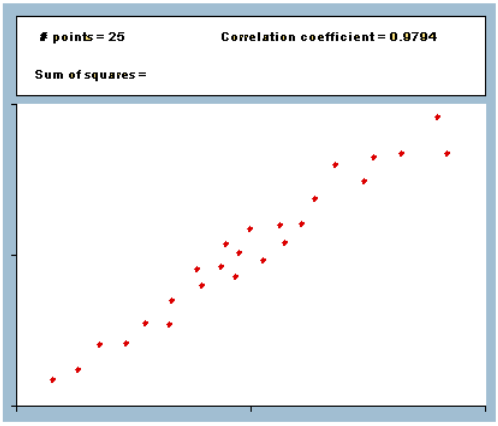

Figure 4

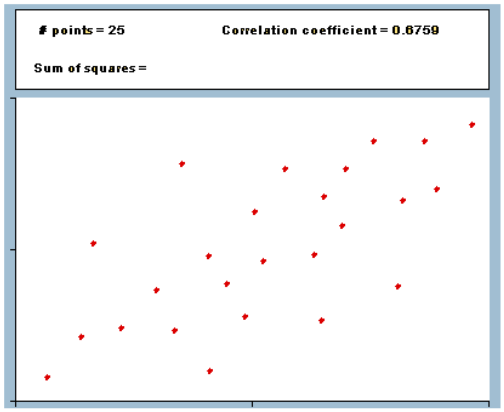

Figure 5

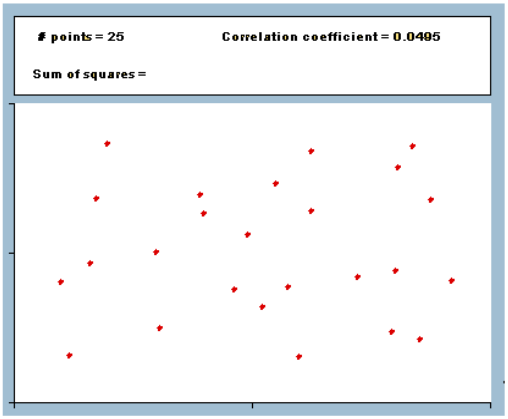

Figure 6

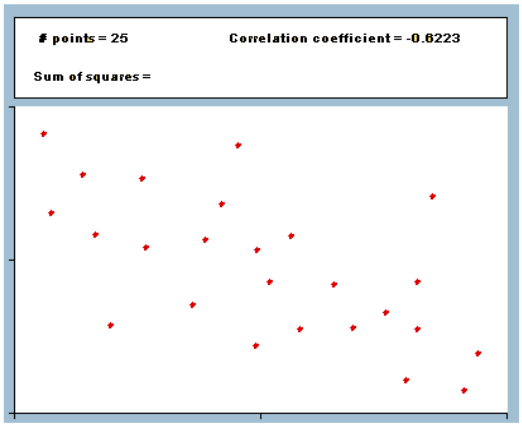

Figure 7

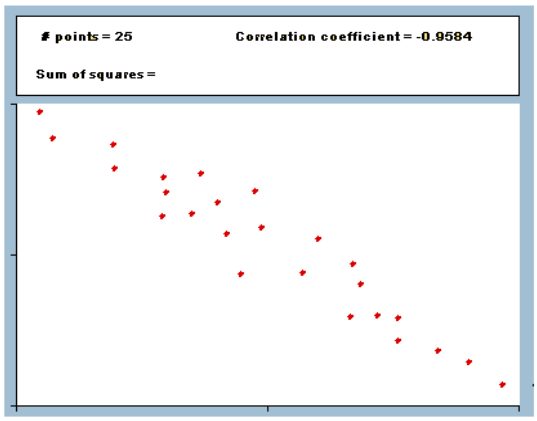

Figure 8

As the figures should suggest, a positive association exists between the two variables if the plotted points tend to form a cloud of points typically in the shape of an ellipse (or football) that is oriented from the "southwest" to the "northeast." The longer and thinner it is (the greater the *eccentricity* of the ellipse), the stronger the positive association. A negative association produces a similar cloud but with the opposite orientation from "northwest" to "southeast." A zero or very weak association produces a cloud of points of no distinct shape or orientation.

It may be puzzling at first why in a standard crosstabulation a concentration of cases running from "northwest" to "southeast" reflects a positive association while in a scattergram the same pattern reflects a negative association (and *vice versa*). This reflects only a cosmetic difference in setting things up: in a crosstabulation, the values of the row (vertical) variables are usually placed in descending order while in a scattergram the values of the vertical (row) variable are invariably (and more reasonably) placed in ascending order.

Finally, we can drive home the point that scattergrams and crosstabulations are essentially similar devices by showing how directly they are connected (and also how the former is much more informative that the latter). Suppose we want to construct a crosstabulation of Pearson data using the *short-medium-tall* class intervals we previously worked with. This can be accomplished simply by superimposing the appropriate grid on the scattergram (as shown in Figure 9) and then counting the number of plotted points in each resulting cell.

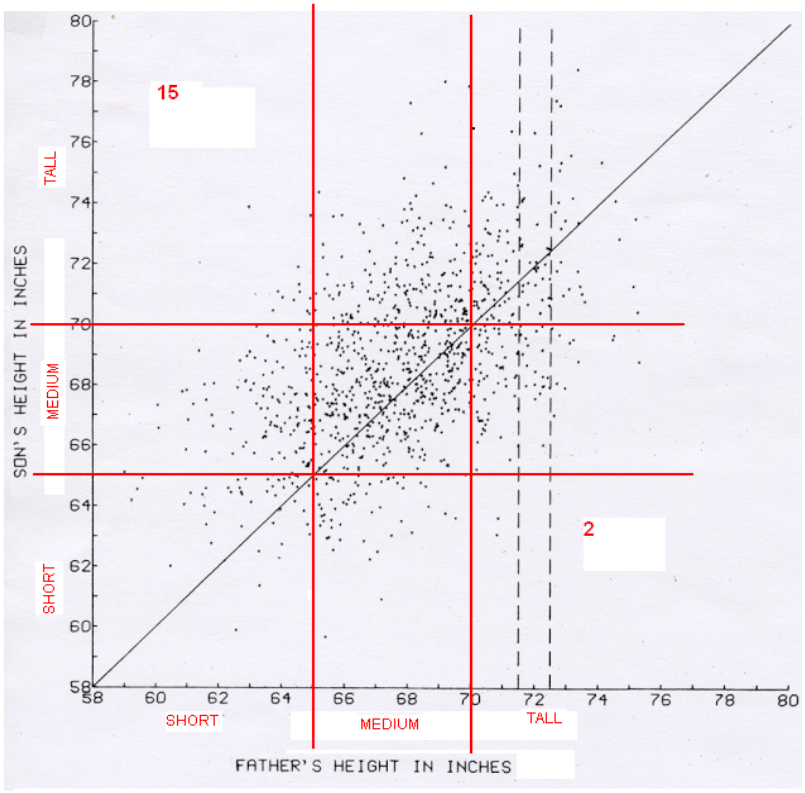

Figure 9

SPSS can readily create scattergrams. For example, let's:

(a) open the PRESELECT data file, giving state by state vote totals for each Presidential candidate in each election;

(b) find the variables dem2000, rep2000, dem2004, and rep2004;

(c) compute the DEMOCRATIC PERCENT OF THE TWO-PARTY VOTE in each election, by clicking on Transform => Compute and entering this expression in the Compute Variables dialog box: d2pc2004 = 100 * dem2000/(dem2000 + rep2004) and likewise for d2pc2004; and then

(d) produce the following scattergram by clicking on Graphs => Scatter... => Simple/Define and then in the Simple Scatterplot dialog box put d2pc2004 on the Y Axis and d2pc2004 on the X Axis.

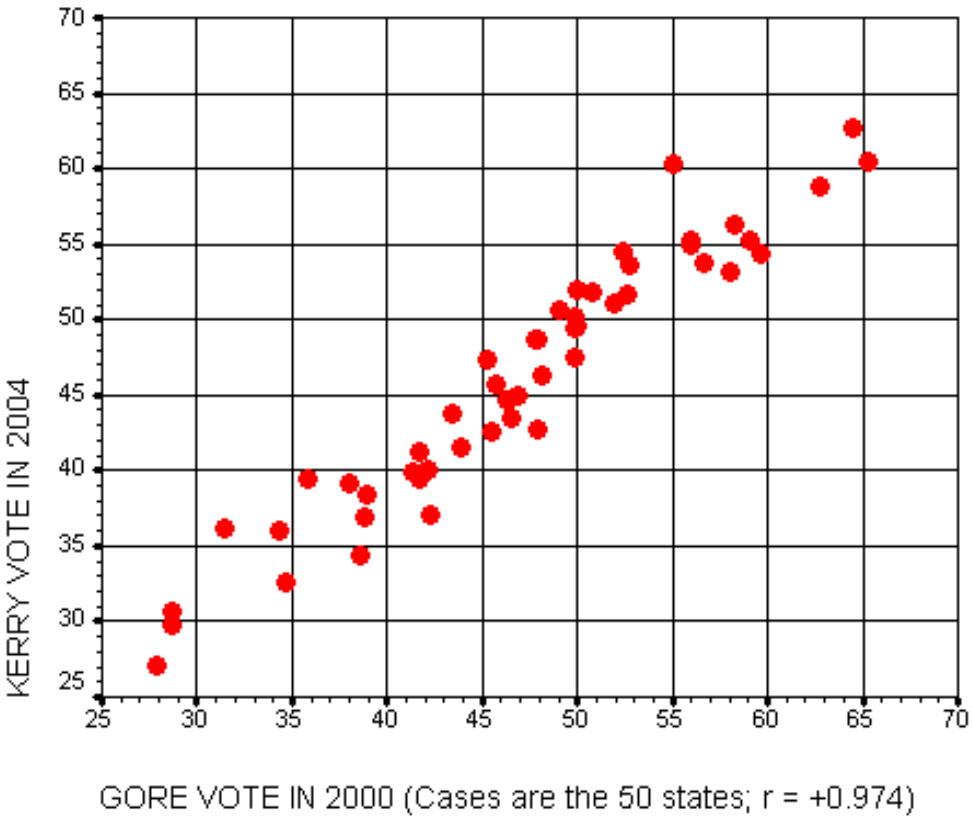

Figure 10

Figure 10 includes editing done within SPSS itself. Excel and similar programs also can produce scattergrams/scatterplots (sometimes called X-Y charts).