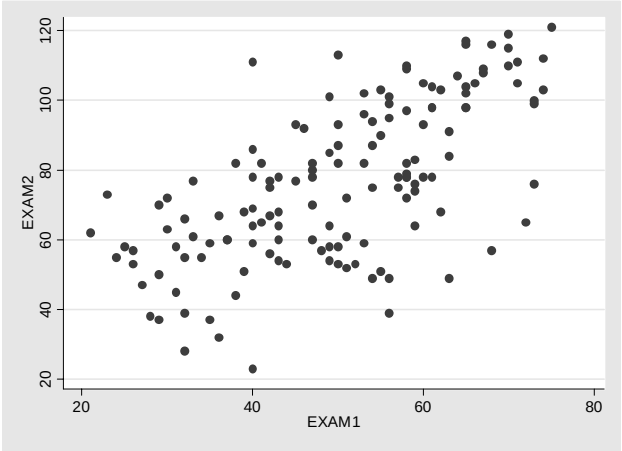
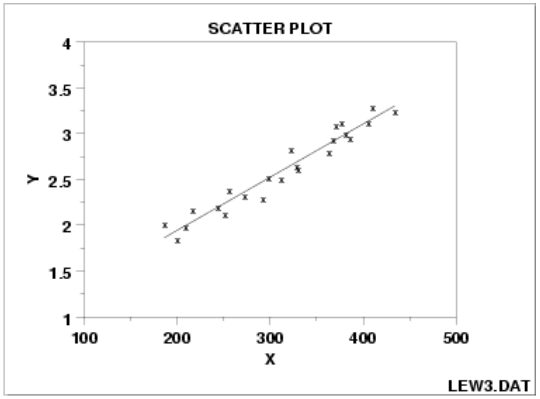
Correlation (Chapter 7)

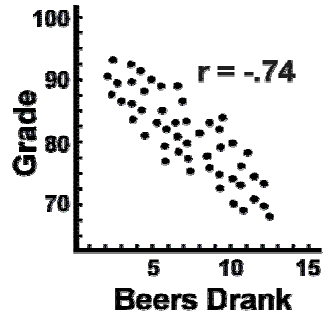
- **Scatterplot or Scatter Diagram** -- a graphical summary of a two-variable analysis.
  - A scatterplot or scatter diagram is a two dimensional plot of data. The horizontal dimension is called x, and the vertical dimension is called y.
  - Each point on a scatterplot or scatter diagram shows two values, an x value and a y value. Each point represents a single case. A single case could be a single person or object, but a single case could be a matched pair (e.g. father-son, twins, husband-wife)



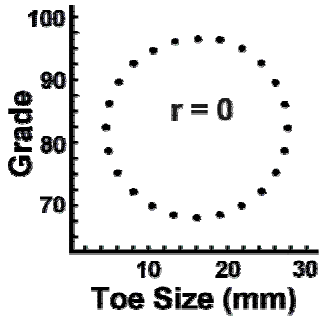
- **Positive and negative relationships**
  - There is a POSITIVE relationship if above-average values of x are associated with above-average values of y.
  - Conversely, there is a NEGATIVE relationship if above-average values of x are associated with below average values of y.
- **Form**
  - Ask – does the plot appear to be more or less like a line, or like a wedge, a circle, a U etc. The shape says something about the strength of the relationship. Also look for outliers.

A strongly positive relationship





Negative Relationship



No Relationship

Names for X and Y

- In many disciplines, X and Y are usually called the INDEPENDENT (AKA explanatory, predictor, treatment) and DEPENDENT (AKA response, predicted, outcome) variables respectively. They are given these names because the independent (X) variable is thought to influence the dependent (Y) variable. There is nothing to stop us from reversing the relationship. Designation of independent and dependent rely strongly on how the question is being asked.

Numerical Summary: The correlation coefficient r

- **The CORRELATION COEFFICIENT, denoted r, measures how close the data are to a straight line or in other words it measures the strength of association. It is a numerical summary of the scatter diagram graphic.**
- **The correlation coefficient can take values from -1 to +1. Values near zero mean that the data is not close to a straight line. Values near the ones (both positive and negative) mean that the data is very close to a straight line.**

Formula

- **Formula (YOU DO NOT NEED TO KNOW HOW TO CALCULATE ONE FOR THE FINAL)**
- Your text gives you a very long formula for calculating the correlation coefficient (footnote page 120) and this shorter one:

$$r = \frac{\sum Z_x Z_y}{n-1}$$

Properties of the correlation coefficient

- -1 <= r <= 1
- If r is close to 1 or -1, the data are close to a line
- If r is close to 0, the data are not close to a line
- The correlation r measures how close the data are to a line
- r does NOT tell what percentage of the data fall on the line
- The correlation between x and y is the same as the correlation between y and x.

Example 1

- This is a correlation table, it's common to a lot of software, learn to recognize and interpret it, there are 7 variables here and we're looking at all possible parings of the variables:

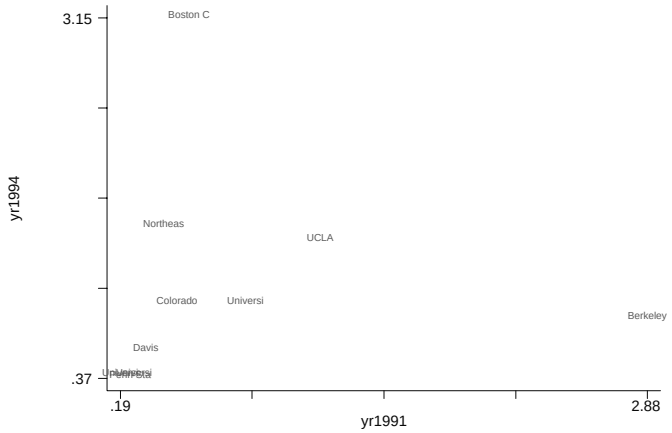
	sei	spsei	tvhours	income98	age	sibs	educ
sei	1.0000						
spsei	0.2281	1.0000					
tvhours	-0.2433	-0.1130	1.0000				
income98	0.3698	0.3949	-0.2209	1.0000			
age	-0.0423	0.0734	0.0797	-0.0116	1.0000		
sibs	-0.1432	-0.1078	0.0797	-0.1120	0.0821	1.0000	
educ	0.5835	0.1740	-0.2791	0.3268	-0.2029	-0.2599	1.0000

Another Example

- The correlation coefficient here is .0584, very near to zero and slightly positive. Implies virtually no relationship

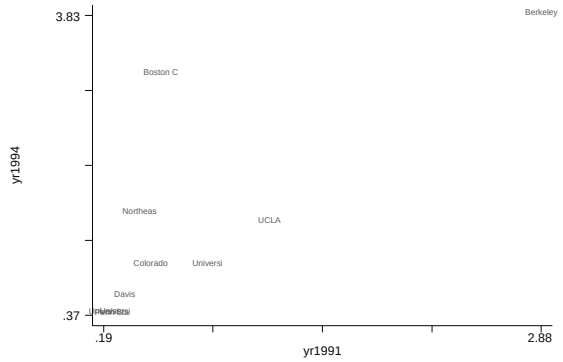
SCHOOL	1991	1994
Berkeley	2.88	0.83
UCLA	1.21	1.43
Davis	0.32	0.58
Colorado State	0.48	0.94
University of Florida	0.83	0.94
University of Minnesota	0.26	0.39
University of Iowa	0.19	0.39
Penn State	0.24	0.37
Northeastern	0.41	1.54
Boston College	0.54	3.15
Average	0.736	1.056
SD	0.7736	.8013

A Scatter Diagram with correlation r = .0584



Suppose the Berkeley pair is wrong and it should have been 2.88 and 3.83 instead of 2.88 and 0.83. Can a single error change the correlation coefficient?

YES r = .7655 here

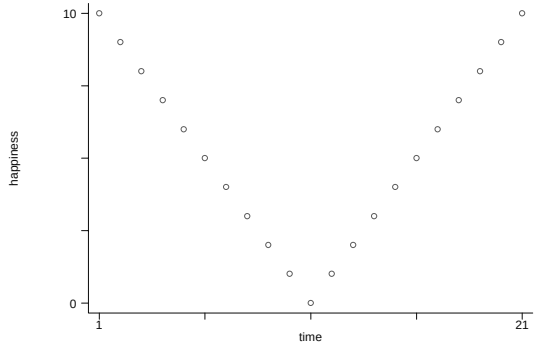


Other Problems - Outliers

- The correlation is useful when your data points are football shaped. But sometimes r will mislead. A single outlier can counterbalance a strong linear relationship. Consider Boston College in the previous graphics. If I were to remove that observation from the dataset, the correlation would jump from .7655 to about .952 so from something football shaped to virtually a straight line.

Another Problem – Non-Linearity

- Two variables are related, but their relationship is not well described by a straight line.



Two Remaining Issues

- **Ecological Correlations**
  - Correlations for aggregated information (e.g. information by state) are almost always stronger than correlations for individuals. Often people calculate correlations based on rates or averages that are based on many individuals combined. This tends to make a relationship look stronger than it is. Beware when people do this.
- **Association is NOT causation**
  - Remember: a correlation tells you how strongly two variables are related in a linear fashion. It doesn't mean one variable causes another to happen. The problem here is one of CONFOUNDING, that is, a third variable may be present that has not be taken into account.