STAT 100 Chance and Data Analysis October 9, 2002

Friday: Mid-term Test I - See specs in Notes Oct 7 (at beginning), Sept. 30 (at end) and the sample midterm Oct 2 (at end).

Today: Correlation - and astronomy article pp 268-274

Lay meaning of "correlation": degree to which two events happen together

e.g.-1. 1. New York Stock Market Index declines
2. Toronto Stock Market Index declines

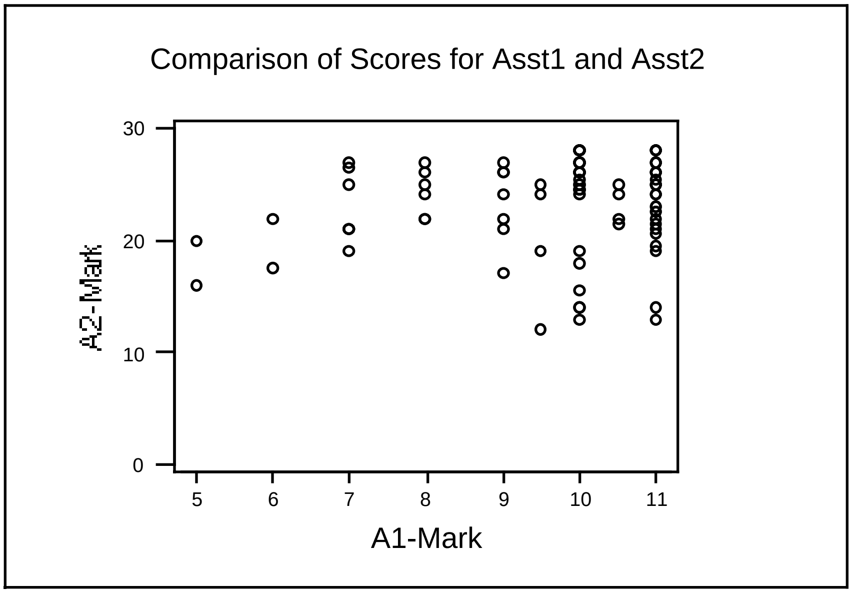
e.g.-2. 1. Ice on roads
2. More fender-benders than usual

e.g.-3 1. Vancouver Canucks win hockey game
2. Bigger sales at the bars

In statistics, "correlation" has a more exact meaning, but it only applies to two quantitative measurements and the degree to which two measurements are large together and small together. If one measurement is large when the other is large, and small when the other is small, the two measurements are said to be positively correlated. If one measurement is large when the other is small, and small when the other is large, the two measurements are said to be negatively correlated. The correlation coefficient varies from -1 (negatively correlated) through 0 (un-correlated) to +1 (positively correlated).

So how is this index computed?

Consider your marks on assignment one (out of 28) and assignment 2 (out of 11).



To what extent do students have the same performance on both assignments? Does a relatively high mark on assignment 1 tend to go with a relatively high mark on assignment 2? In this case the correlation between the two sets of scores is +0.2, and this is closer to 0 than to 1 so we would say the positive correlation is weak. What is obvious from the graph is that the two assignment marks are not "correlated" in the sense that they are large and small together.

How is the correlation coefficient computed?

1. First compute the mean and SD of each variable:

In this case A1 has mean 23.4 and SD 4.5
and A2 has mean 9.7 and SD 1.5

2. Calculate what is called the standardized values of each variable:
(measurement - mean)/SD

which in this case is S-A1=(A1 mark - 23.4)/4.5 and
S-A2=(A2 mark - 9.7)/1.5)

(By using all the marks, this generates two new columns of data - so instead of columns for A1 and A2, one has columns for S-A1 and S-A2)

3. Then multiply the column components of S-A1 and S-A2 together to result one column of products.

4. Then average this column of products to get the correlation coefficient.

A portion of the table of marks is

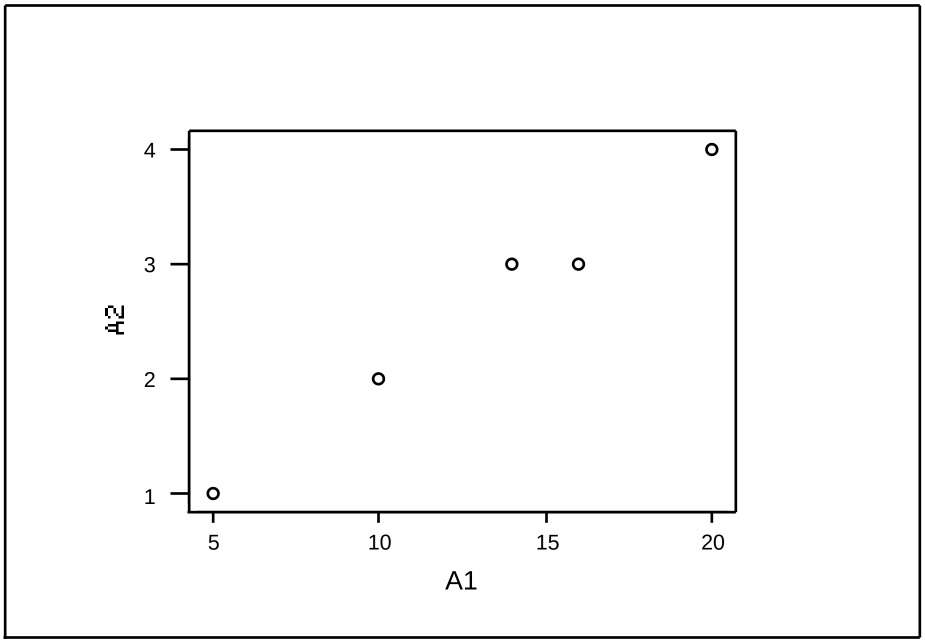
A1	A2	S-A1	S-A2	S-A1 x S-A2
28.0	11.0	1.025	0.859	0.88
14.0	10.0	-2.114	0.185	-0.39
28.0	10.0	1.024	0.185	0.19

The average of the last column for all the data turns out to be 0.2 in this case.

Consider the following hypothetical marks:

A1	A2
10	2
14	3
5	1
16	3
20	4

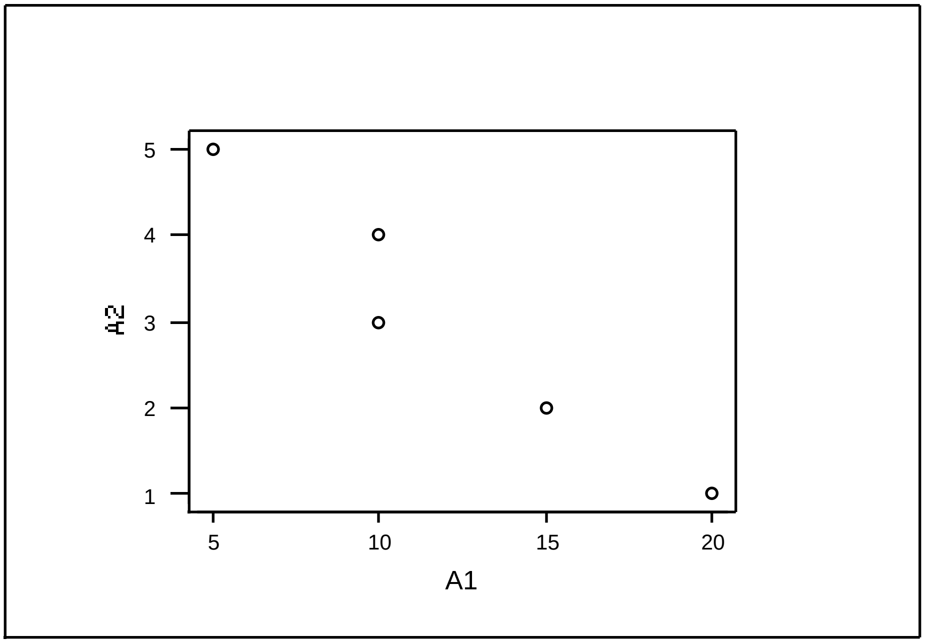
What would the scatterplot look like?



Almost a straight line. And the correlation in this case is .79 (your calculator may give .99 since it divides the total of the products my n-1 instead of n, in this case by 4 instead of 5. For realistic data sets with 20 or more data values, the difference is negligible.)

On the other hand if the data were

A1	A2
10	4
5	5
15	2
10	3
20	1



The correlation would be -.79.

A negative correlation between two variables means that the large value of one are associated with the small values of the other, and vice versa .

The application in the article is a non-standard use of the correlation coefficient, but one that uses its properties in a proper way.

Table 1 shows three "variables", representing the measurements at different points of time. The correlation can summarize similarity in the measurements two variables, which in this application is two times. So we can compare time 1 with time 2, and time 2 with time 3. (Of course we can compare time 1 with time 3 but this is not so useful in this example).

			Points		
Time	1	2	3	4	5
1	1.5	0.5	0	-0.5	-1.5
2	1.4	0.7	0	-0.7	-1.4
3	1.0	-1.4	0.4	-1.0	1.0

It is easy to see that time 1 and time 2 have similar readings, but both are quite different from time 3.   The correlation coefficient is close to 1 for time 1 and 2, and close to 0 for time 2 and 3. (The method for the correlation coefficient in the book would be the same as ours if the data were really standardized - it is almost, but not quite: means are 0 but SDs are a bit larger than 1).

Don’t worry too much about the physical interpretation described in the article - it is more complicated than is useful for this course. If you get the point that summary measures like the correlation coefficient can have non-trivial uses in astonomical research, then that is enough for now.

Guess the correlation:

1. Money in wallet and number of credit or bank cards
2. Outside temperature and number of students wearing coats
3. Outside temperature and rainfall.
4. Height and course grade