Correlation Machines

Andries de Man

Dit artikel is eerder gepubliceerd in de proceedings van de International Meeting in Trento, september 2016.

Correlation

Correlation is a statistical concept that is used in many scientific areas. Two or more quantities are correlated if there is a "connection" between the pairwise values they can assume. Correlation is usually quantified with the "Pearson coefficient of correlation" r, spanning -1 to 1.

In this presentation we will use two quantities: X and Y. Figure 1 shows a number of datasets with these quantities, and the corresponding Pearson coefficient of correlation r

The interpretation of correlation coefficients is tricky. A high (absolute) coefficient of correlation does not necessarily indicate a *linear* relation between X and Y. The four datasets in the lower row of Figure 1 all have the same r, but look quite different. Data with a low (absolute) coefficient of correlation can still show a clear pattern. The datasets in the third row of Figure 1 have a zero coefficient of correlation, but still seem to have some connection between X and Y.

The Pearson coefficient of correlation was introduced in 1895 by Karl Pearson and started to be used at a large scale in the beginning of the 20th century by, amongst others, economists, psychologists, agricultural experts and brewers.

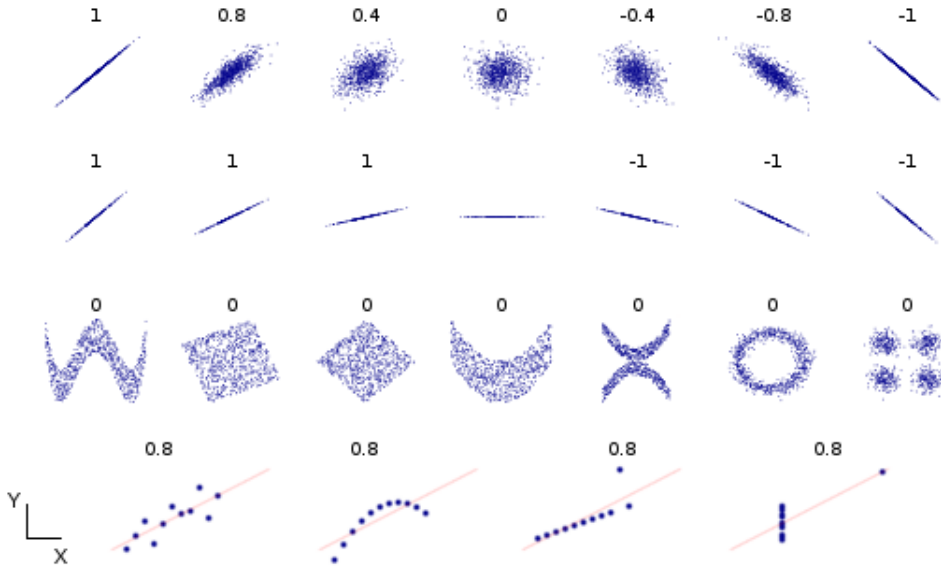

Figure 1: Pearson r value for various datasets (Source: Wikipedia)

How is *r* calculated?

If there are n "subjects" in the dataset, and the subject i has a value x_i for X and y_i for Y, than r is given by:

$$r = \frac{n\sum x_i\, y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

TABLE IV.—SUPPLEMENTARY CORRELATION SHEET
x-variable 1, y-variable 2, Data sheet (Table I)

$(x+y)$	Tabulation	f	$(x+y)^2$
29	1	1	841
28			784
27	11	2	729
26	1	1	676
25	1	1	625
24			576
23			529
22			484
21			441
20	11	2	400
19			361
18			324
17			289
16			256
15	1	1	225
14	11	2	196
13	1	1	169
12	1	1	144
11	11	2	121
10	11	2	100
9	..		81
8	1	1	64
7	1	1	49
6	11	2	36
5	.		25
4	1	1	16
3	..		9
2	1	1	4
1	11	2	1
0	.		0

$E(x+y) = 319$
$E(x+y)^2 = 5979$

$(x-y)$	Tabulation	f	$(x-y)^2$
15	.	.	225
14			196
13			169
12			144
11			121
10		.	100
9			81
8			64
7			49
6			36
5	..		25
4	IIII	5	16
3	111	3	9
2	IIII	5	4
1	IIII 111	8	1
0	111	3	0

$\Sigma(x-y)^2 = 135$

$\Sigma x = 167.$
$\Sigma y = 152.$
$\Sigma x^2 = 1599$
$\Sigma y^2 = 1458$

$m_x = 6.958$
$m_y = 6\ 333$
$s_x^2 = 18\ 211$
$s_x = 4\ 27$
$s_y^2 = 20\ 643.$
$s_y = 4\ 54$
$N = 24$

Preliminary checks		
Operation	Value	Check
$\Sigma x + \Sigma y$	319	$\Sigma(x+y)$
$2\Sigma x^2 + 2\Sigma y^2$	6114	$\Sigma(x+y)^2 + \Sigma(x-y)^2$
Computations		
$\frac{\Sigma(x+y)^2}{N}$	249 125	$\aleph_{(x+y)}$
$\frac{\Sigma(x-y)^2}{N}$	5 625	$\aleph_{(x-y)}$
$m_x + m_y$	13 291	$(m_x + m_y)$
$(m_x + m_y) \times (m_x + m_y)$	176 651	$(m_x + m_y)^2$
$m_x - m_y$	625	$(m_x - m_y)$
$(m_x - m_y) \times (m_x - m_y)$	391	$(m_x - m_y)^2$
$2s_x^2 + 2s_y^2 + (m_x + m_y)^2 + (m_x - m_y)^2$	254.570	$\aleph_{(x+y)} + \aleph_{(x-y)}$ (check)
$\sqrt{4s_x^2 s_y^2}$	38 77	$2s_x s_y$ (check)
$\aleph_{(x+y)} - (m_x + m_y)^2 - s_x^2 - s_y^2$	33 620	n
$s_x^2 + s_y^2 + (m_x - m_y)^2 - \aleph_{(x-y)}$	33.620	n (check)
$\frac{n}{2s_x s_y}$	87	r

(Do the last computation twice)

Figure 2: Correlation form, for a method using ($x_i + y_i$) [Cureton1929]

So one needs the following sums: $\sum x_i$, $\sum y_i$, $\sum x_i^2$, $\sum y_i^2$ and $\sum x_i y_i$. The first four sums are also needed for the calculation of averages and standard deviations. If there are more quantities (X,Y,Z,...) one will need all primary sums, square sums and (at least) pairwise cross sums. And one would like to do all these calculations by entering the data only once.

By 1900 Hollerith tabulators and punch cards could be used for this purpose, but this equipment was large and expensive. Another method consisted of employing human "computers" who, without any knowledge of statistics, performed calculations using pre-printed forms. A large variety of these forms were offered commercially, with small variations of the formula for r and built-in error checking (Figure 2).

The first step in the calculation of a correlation coefficient usually consisted in "transmutation" of the data: the range of the values was normalized to (generally) integer numbers between 1 and 20. After this step one could use a printed multiplication table, and mental addition to calculate the correlation. This calculation could also be performed with a "normal" mechanical calculating machine, preferably one with a very wide result register. The appendix describes how these calculations were done.

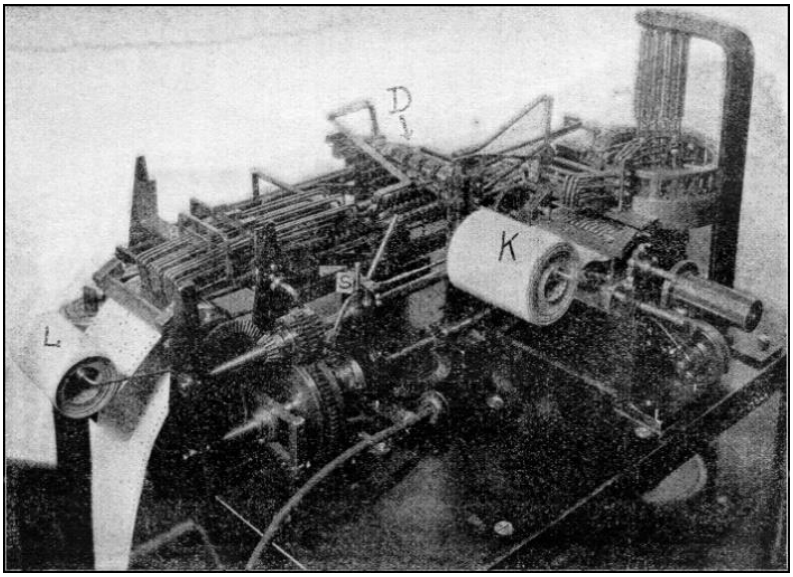

Figure 3: Hull's machine

It is clear that this was a time-consuming error-prone job. That's why a need arose for dedicated, inexpensive correlation calculators.

Dedicated machines

Hull

By 1921 Clark Hull, a psychologist from Wisconsin, constructed a purely mechanical machine (Figure 3) for calculating the sums mentioned above [Hull1925]. The data, having integer values between 0 and 999, was punched into paper tape that had to be fed multiple times through the machine. Hull could the calculate correlations between a large number of quantities. The resulting averages, standard deviations and correlation values could be registered in thin metal strips for use in correlation-based predictions. The machine was used to give occupational advice for 40 professions based on 60 psychological parameters [Mechanix 1929].

Two versions of this machine were built, one for the Wisconsin Psychological Laboratory and one for the National Research Council. Because some researches started sending Hull data for processing, he proposed to establish a Central Correlation Bureau that would compute correlations on demand.

At the time, it was thought that two machines would be enough for all the correlation-needs in the United States.

Dodd

Around 1925 Stuart C. Dodd, a psychologist at Princeton University, made a simpler correlation calculator [Dodd1926]. This machine (Figure 4) contained drums on which square numbers where represented by pins of different lengths (separate pins for units and tens). These drums are the square number equivalent of the multiplication bodies as used in the Millionaire calculator. Dodd designed different

versions of this device. Later development was continued by the Cambridge Instrument Co. Inc., New York, who sold these correlation machines to the universities of Harvard, Berkeley, and Chicago.

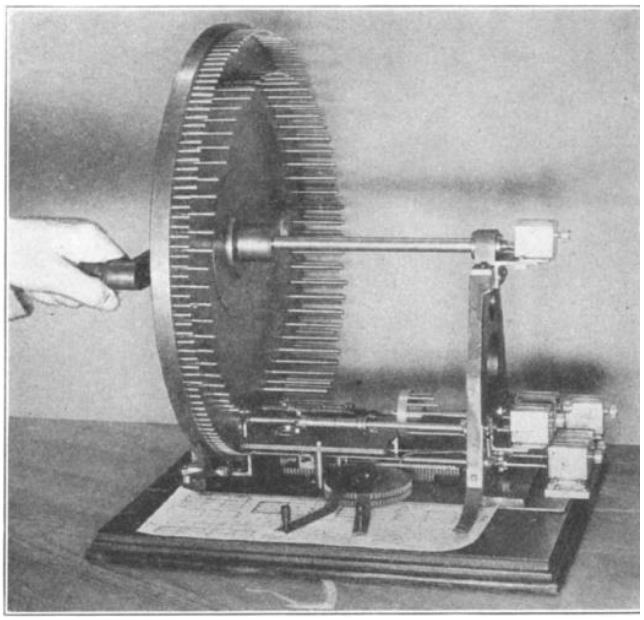

Figure 4: Prototype of Stuart Dodd's machine

Seashore

A third correlation machine (Figure 5) carries the name of Carl Emil Seashore (Sjöstrand), again a psychologist. This machine was sold around 1930 by the C.H. Stoelting Co. from Chicago, for $550. The machine calculated $\sum x_i$, $\sum y_i$, $\sum x_i^2$, $\sum y_i^2$ and $\sum (x_i - y_i)^2$, using a slightly different formula for r.

Analog methods

If X and Y are regarded as coordinates of point masses in a two-dimensional space, then $\frac{1}{n}\sum x_i$ and $\frac{1}{n}\sum y_i$ give the centre of mass, and $\sum x_i^2$ and $\sum y_i^2$ give the moments of inertia with respect to the X and Y axes.

Price

In 1935 the psychologist Bronson Price [Price1935] proposed to use these properties for the calculation of: construct a bed of nails on which the data is attached using thin rings, and determine the centre of mass and the moments of inertia of the whole shebang. The practical problem was that the bed and the nails had to be very light, or the rings had to be very heavy. Price never tried it himself.

Figure 5: Seashore's machine (Stoelting catalog, 1930)

Harsh and Stevens

C.M. Harsh and Stanley Smith Stevens, psychologists at Harvard, created an analog correlation machine working with small balls (Figure 6).

Platt

Later John R. Platt revived the idea [Platt1943]. Platt was a biophysicist from Michigan who ended up in sociology. He used a flat metal sieve in which the data was set with lead pins (Figure 7).

The centre of mass was determined by hanging the thing twice, from different vertices, getting the two plumb lines from these vertices, and then obtaining their intersection. For the determination of the moments of inertia the thing was regarded as a torsion pendulum: the whole thing was given a rotation around a vertically placed X-, Y- or diagonal-axis and then released, and the time needed for at least 20 to-and-fro rotations was measured. Platt claimed that, this way, he could determine r with an accuracy of 0.01.

The hydraulic device of Schumann

A somewhat more complicate, and more dangerous, device was made by the South-African meteorologist Theodor Eberhardt Werner Schumann [Schumann1940]. This device contained glass tubes filled with mercury in which iron rods were floating. His machine was also used to solve sets of linear equations. A trained human computer would need $5m^2 + m^3/4$ minutes to solve a set of m equations with m variables, while Schumann's machine could do that in $6m$ minutes.

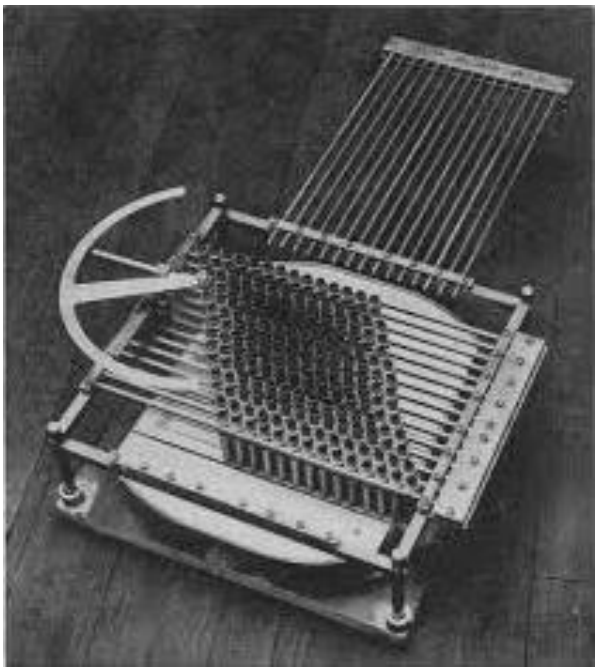

Figure 4: Harsh and Stevens' analog "instrument" [Harsh1938]

The electrical device of Ford

Adelbert Ford, a psychologist of Michigan University, built an electrical correlation machine in 1931 [Ford1931]. The data was entered on a panel with 100 potentiometers (Figure 8). The position of a potentiometer corresponds with the (transmuted) X and Y value of a data point. For each data point the corresponding potentiometer is turned 1 unit. The actual rotation angle for one unit changes with the position of the potentiometer, and this way squares and cross products are "calculated". Each potentiometer is connected to its own coil, and all these coils together form the secondary side of a transformer. To the primary side a voltmeter is attached which is used to read $\sum x_i y_i$. Since the transmuted data is required to average to zero, for the calculation of r Ford only needs $\sum x_i y_i$ and $\sqrt{\sum x_i^2 \sum y_i^2}$, with the latter being set with a complicated calibration procedure.

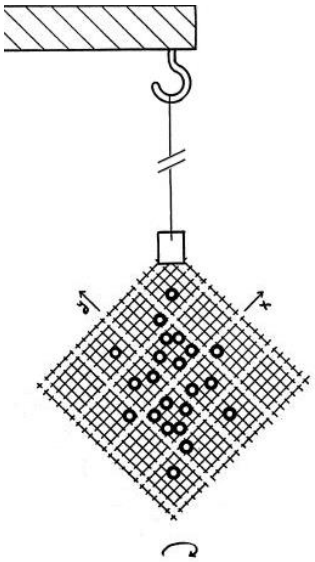

Figure 5: Platt's analog "instrument" [Platt1943]

Post processing

The mechanical correlation machines only calculated sums. To calculate r, square roots and quotients had to be computed using a slide rule or log table (or a mechanical calculating machine, if you time to spare). Fortunately, the required accuracy for r was usually only one decimal (in a −1 to 1 range).

Epilogue

The correlation machines present a unique contribution of psychologists, sociologists and the like to the development of computing devices. This was clearly brought about by practical needs and a contemporary inclination towards technology among psychologists [Draaisma1992].

I have been unable to find correlation machines of European origin. On the contrary: in 1929 the German mathematician Wilhelm Cauer attempted to buy a Hull machine for Göttingen University [Petzold2000].

Of the mentioned machines no surviving specimens are known to me, except for the Hull machine that was rediscovered in 1997 by Hartmut Petzold in a depot of the National Museum of American History [Petzold2000]. Psychological institutes seem to be as careless regarding their material heritage as are institutes of science and technology.

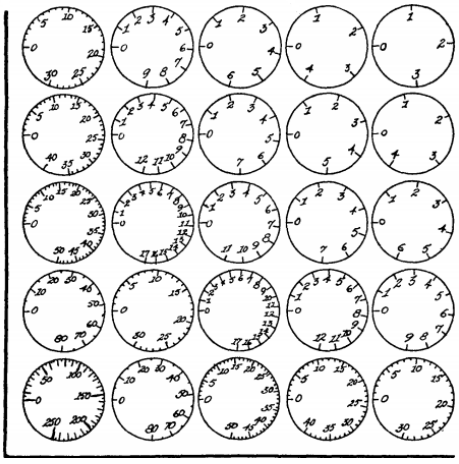

Figure 6: Ford's Correlator: a quarter section of the panel with potentiometers for entering data

References

- [Pearson1895] Karl Pearson, "Notes on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London,* 1895; 58: 240–242. To be fair, this correlation measure was introduced in 1880 by Francis Galton, and used to be called "Galton's function".
- [Cureton1929] E.E. Cureton, "Computation of correlation coefficients", *Journal of Educational Psychology*, 1929;208:588-601
- [Hull1925] C.L. Hull, "An automatic correlation calculating machine", *Journal of the American Statistical Association,* 1925; 201(52):522-531
- [Dodd1926] Stuart C. Dodd, "The applications and mechanical calculation of correlation coefficients", *Journal of the Franklin Institute*, 1926;201(3):337-349
- [Mechanix 1929] "Intricate machine fits men to jobs", *Modern Mechanix,* Mei 1929
- [Price1935]Bronson Price, "A proposed method for the direct measurement of correlation", *Science,* 1935; 822(134):497-498
- [Harsh1938] C.M. Harsh, S.S. Stevens, "A Mechanical Correlator", *Am. J. Psych.*, 1938; 51:772-730
- [Platt1943] John R. Platt, "A Mechanical Determination of Correlation Coefficients and Standard Deviations", *Journal of the American Statistical Association,* 1943; 38(223): 311-318.
- [Schumann1940] T.E.W. Schumann, "The principles of a mechanical method for calculating regression equations and multiple correlation coefficients and for the solution of simultaneous linear equations", *Phil. Mag. Series 7,* 1940; 29:258-273
- [Ford1931] A. Ford, "The Correlator", *Journal of Experimental Psychology,* 1931; 142:155-163
- [Draaisma1992] Douwe Draaisma, "Een laboratorium voor de ziel", Universiteitmuseum Groningen, 1992
- [Petzold2000] Hartmut Petzold, "Wilhelm Cauer and his Mathematical Device", in Bernard Finn (ed), "Exposing Electronics", CRC Press, 2000

Illustrations

- Fig 1: Pearson r value for various datasets (Source: Wikipedia)
- Fig 2: Correlation form, for a method using $(x_i + y_i)$ [Cureton1929]
- Fig 3: Hull's machine
- Fig 4: Prototype of Stuart Dodd's machine
- Fig 5: Seashore's machine (Stoelting catalog, 1930)
- Fig 6: Harsh and Stevens' analog "instrument" [Harsh1938]
- Fig 7: Platt's analog "instrument" [Platt1943]
- Fig. 8: Ford's Correlator: a quarter section of the panel with potentiometers for entering data

Appendix

The calculation of the sums for *r* using a mechanical calculator:

Suppose we have a simple pinwheel calculator with 12 digits in the input- and revolution register, and 20 in the result register:

Revolution											0	0	0	0	0	0	0	0	0	0	0
Input											0	0	0	0	0	0	0	0	0	0	0
Result	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Take $x_i = 12$ and $y_i = 13$.

Put x_i and y_i in the input register (separated by zeroes!)

Revolution											0	0	0	0	0	0	0	0	0	0	0
Input											0	0	0	0	1	2	0	0	0	1	3
Result	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Shift the input register over half its width:

Revolution											0	0	0	0	0	0	0	0	0	0	0
Input					0	0	0	0	1	2	0	0	0	0	1	3					
Result	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Turn the crank 2 times (units of x_i)

Revolution											0	0	0	0	0	2	0	0	0	0	0
Input					0	0	0	0	1	2	0	0	0	0	1	3					
Result	0	0	0	0	0	0	2	4	0	0	0	0	2	6	0	0	0	0	0	0	0

Shift the input register over 1 position and turn the crank 1 time (tens of x_i)

Revolution											0	0	0	0	1	2	0	0	0	0	0
Input				0	0	0	0	1	2	0	0	0	0	1	3						
Result	0	0	0	0	0	1	4	4	0	0	0	1	5	6	0	0	0	0	0	0	0

We now have x_i in the revolution register, and x_i^2 and $x_i y_i$ in the result register.

Shift the input register back completely.

Revolution											0	0	0	0	1	2	0	0	0	0	0
Input											0	0	0	0	1	2	0	0	0	1	3
Result	0	0	0	0	0	1	4	4	0	0	0	1	5	6	0	0	0	0	0	0	0

Turn the crank 3 times (units of y_i)

Revolution											0	0	0	0	1	2	0	0	0	0	3
Input											0	0	0	0	1	2	0	0	0	1	3
Result	0	0	0	0	0	1	4	4	0	0	0	1	9	2	0	0	0	0	3	9	

Shift the input register and turn the 1 time (tens of y_i)

Revolution											0	0	0	0	1	2	0	0	0	1	3
Input							0	0	0	0	1	2	0	0	0	0	1	3			
Result	0	0	0	0	0	1	4	4	0	0	0	3	1	2	0	0	0	1	6	9	

We now have x_i and y_i in the revolution register, and x_i^2 and $2x_iy_i$ and y_i^2 in the result register. Shift the input register back completely, and repeat the procedure for $x_{i+1}$and y_{i+1} *without clearing* the revolution- and result registers. Finally the revolution register will contain $\sum x_i$ and $\sum y_i$, and the result register $\sum x_i^2$ and $2\sum x_iy_i$ and $\sum y_i^2$

We see that each value of x_i and y_i has to be entered *twice*: once in the input register and once when cranking. But, because the revolution register is not cleared, it is difficult, after the first pair of values, to check if the correct value has been "cranked in". Electrically driven machines that allow entering multiplicators via a separate keyboard or pin setting would be a great help in this case.

It is also clear that, depending on the number of data-points and the range of the data, the registers should be rather large: for 100 data points with a range of 0…100 (integer numbers!) $\sum x_i^2$ can grow to 10^6, needing 7 digits.
The result register will have to accommodate 3 sums of this size, so should have at least 21 digits.