

1. Overview

Up to now we have only looked at what are called "univariate" statistics. This means we are only studying a single variable in a given population or a given sample. For example, we might talk about mean height, mean weight, the standard deviation of an LSAT score. But now, we turn to relationships between two variables.

2. Basic Definitions & Graphical Summary

A. Scatterplot or Scatter Diagram -- a graphical representation 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)

B. 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.

C. Warning! Scatter diagrams only show association, but association does not mean causation (firefighters, fire damage)

In many disciplines, X and Y are usually called the INDEPENDENT and DEPENDENT variables respectively. They are given these names because the independent variable is thought to influence the dependent 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.

3. 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

Your text gives you a very long formula for calculating the correlation coefficient (pp 132-134) and I am not certain how useful it is. Instead, read the technical note on p. 134, the formula is reproduced here:

$$r = \frac{(\text{average of the products } xy) - ((\text{average } x) * (\text{average } y))}{(\text{Standard Deviation } x) * (\text{Standard Deviation } y)}$$

4. The SD Line

On the original front cover of your book is a scatter plot of father's height (x) vs. son's height (y). The dashed line is a SD line. The SD line, on a scatterplot, is a line that goes through the point of averages

(the point with values = mean of X and mean of Y). It's slope (Chapter 10) equal to the SD of Y divided by the SD of X. This will become more useful when we get to Chapter 10.

5. Properties of the correlation coefficient

- $-1 \leq r \leq 1$
- 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
- Pictures! (see pages 127 and 129 of your text)
- 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 .
- Invariant under addition. If some constant "a" is added to every one of the X or the Y values, the correlation is unchanged.
- Invariant under multiplication: if all of the x or the y values are multiplied by some positive constant "b", the correlation is unchanged. The correlation can change very dramatically if only ONE of the data points is changed.

6. Examples

Given the five points $\{(2,7), (3,3), (5,1), (8,4), (13,2)\}$, find r .

Answer: $r = -0.47$.

	x	y	product of x & y
	----	----	-----
	2	7	14
	3	3	9
	5	1	5
	8	4	32
	13	2	26
Average:	6.2	3.4	17.2
Stdev :	3.9699	2.0591	

$$r = \frac{(17.2) - (6.2 \times 3.4)}{(3.9699 \times 2.0591)} = -0.47$$

The dataset on the left has a correlation of $r = 0.415$. Find the correlation for the dataset on the right.

x	y	x	y
----	----	----	----
1	2	4	-12
1	3	6	-12
2	6	12	-11
3	5	10	-10
5	9	18	-8
7	8	16	-6
11	8	16	-2
13	4	8	0
13	7	14	0

Since the new list is just a transformation of the old list (i.e., the "new" $x = 2y$, and the "new" $y = x - 13$), the correlation is the same as in the previous list: $r = 0.415$. Note: If you modify only one of the lists (either the x or y) by adding or multiplying by a constant, it will not change the correlation either. But changing one value will change the entire list. If I change point (13,4) to (13,10) the new r would be .752