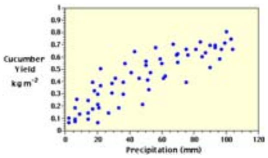
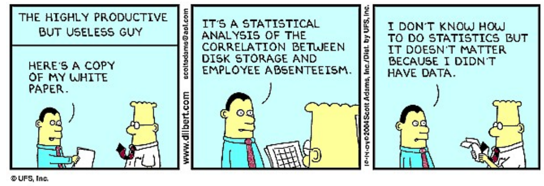
Correlation

Initially developed by Sir Francis Galton (1888) and Karl Pearson (1896)



Sir Francis Galton 1822-1911

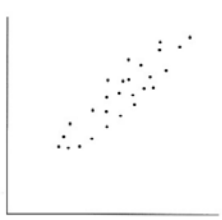


- correlation is a much abused word/term
- correlation is a term which implies that there is an association between the paired values of 2 variables, where association means that the fluctuations in the values for each variable is sufficiently regular to make it unlikely that the association has arisen by chance
- assumes: independent random samples are taken from a distribution in which the 2 variables are together normally distributed

- example 1:
- variable A (income of family) (1000s of Swiss francs)
- variable B (# of autos owned)
- Here there is a perfect and positive correlation as one variable increases in precisely the same proportion as the other variate increases

	paired values				
A	3	6	9	12	15
B	1	2	3	4	5

Case 1:
Positive

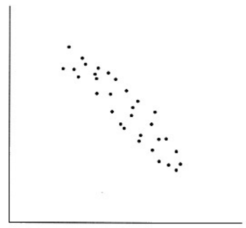


example 2

- variable A (income of family) (1000s of Zambian pounds)
- variable B (# of children)
- here is a perfect and negative correlation as one variate decreases in precisely the same proportion as the other variate increases

	paired values				
A	3	6	9	12	15
B	5	4	3	2	1

Case 2:
Negative

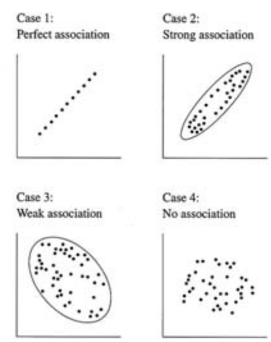


example 3

	paired values				
A	3	6	9	12	15
B	4	1	3	5	2

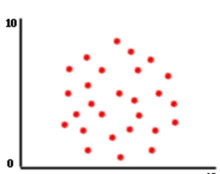
- variable A (income of family)
- variable B (last number of postal code)
- here there is almost no correlation because one variate does not systematically change with the other. Any association is caused by A and B being randomly distributed

Case 3:
Neutral



Generalized Scattergrams Showing Strength of Association

Correlation= 0



Examples

Palm Reading

Some people believe that the length of their palm's lifeline can be used to predict longevity. In a letter published in the *Journal of the American Medical Association*, authors M. E. Wilson and L. E. Mather refuted that belief with a study of cadavers. Ages at death were recorded, along with the lengths of palm lifelines. The authors concluded that there is no significant correlation between age at death and length of lifeline. Palmistry lost, hands down.

Student Ratings of Teachers

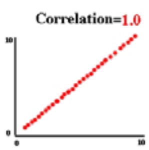
Many colleges equate high student ratings with good teaching—an equation often fostered by the fact that student evaluations are easy to administer and measure.

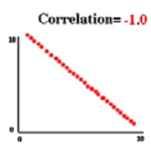
However, one study that compared student evaluations of teachers with the amount of material learned found a strong *negative* correlation between the two factors. Teachers rated highly by students seemed to induce less learning.

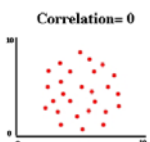
In a related study, an audience gave a high rating to a lecturer who conveyed very little information but was interesting and entertaining.

- correlation is a method whereby a coefficient is calculated to describe the degree of association between sets of paired values, and then tested to determine the probability that the association might be due to chance variation
- i.e. Can show there is only a 5% chance or less of the association being caused by a random influence
  - but this does not mean that one variables is causing fluctuations in the other
- no causal link can be deduced from a correlation alone- it requires other evidence and good judgment

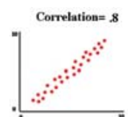
- in the following examples
- example 1 - correlation coefficient =1
- example 2 - correlation coefficient =-1
- example 3 - correlation coefficient =0
- the correlation coefficient for the parametric case is called the Pearson product moment correlation coefficient (r)

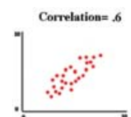


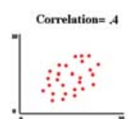


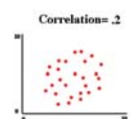


Intermediate positive values

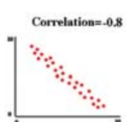


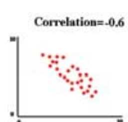


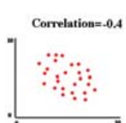


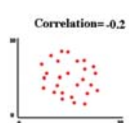


Intermediate negative values









- it is powerful but data has to satisfy ‘normal’ conditions
- calculation
  - x, y are values of the 2 variables
  - S_x, S_y are the sample standard deviation

equations

$$r = \frac{\frac{\sum xy}{n} - \bar{x}\bar{y}}{s_x s_y}$$

$$s_x = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$$

$$s_y = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2}$$

	total proteins consumed	log of income/capital			
	X	Y	X^2	Y^2	XY
Argentina	98	2.715	9604	7.37	266.1
Brazil	61	2.401	3721	5.77	146.5
Denmark	92	3.289	8464	10.82	302.6
Spain	71	2.849	5041	8.12	202.3
Turkey	73	2.476	5329	6.13	180.7
UK	86	3.193	7396	10.20	274.6
US	92	3.519	8464	12.38	323.7
$\sum$	573	20.45	48019	60.79	1696.5
	n=7	n=7			
	$\bar{x}$=81.9	$\bar{y}$=2.92	$\bar{x}^2$=6707.6	$\bar{y}^2$= 8.52	$\bar{x}\bar{y}$=239.15

$$s_x = \sqrt{\frac{48019}{7} - 6707.6} = \sqrt{6859.9 - 6707.6} = 12.34$$

$$s_y = \sqrt{\frac{60.79}{7} - 8.52} = \sqrt{8.68 - 8.52} = .4$$

$$r = \frac{\frac{1695.5}{7} - 239.15}{12.34(.4)} = \frac{242.4 - 239.15}{4.94} = 0.66$$

$$r^2 = 0.43$$

- testing the significance of r
- H_0: r is not significantly different than 0
- H_1: r is significantly different than 0

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$t = \frac{0.66\sqrt{7-2}}{\sqrt{1-0.66^2}} = \frac{0.66(2.24)}{0.73} = 2.03$$

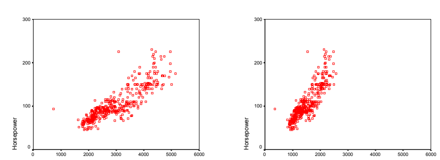
df=N-2
$t_{critical}(\alpha=0.05)=2.571$
we must accept the null hypothesis

Correlation Coefficient Rule of Thumb

Size of Coefficient	General Interpretation
0.8 to 1.0	Very Strong Relationship
0.6 to 0.8	Strong relationship
0.4 to 0.6	Moderate relationship
0.2 to 0.4	Weak relationship
0.0 to 0.2	Very Weak or No relationship

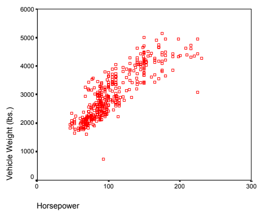
Correlation

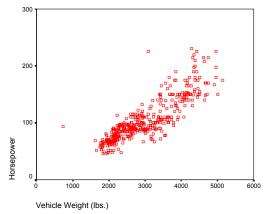
- Insensitive to scale; r = .86 in both cases (why?)



Correlation

- Symmetric with respect to XY orientation





Spurious correlations

- A correlation although strong doesn't make logical sense
- Spurious correlation is normally due to other extraneous variables (a lurking variable?) that are associated with the independent and dependent variables focused on at the time
  - The more bars a city has the more churches it has as well → religion causes drinking?
  - Students with tutors have lower test scores → tutoring lowers test scores?

A view of correlation

- A zero correlation represents complete independence and -1.00 or 1.00 indicates complete dependence. Independence viewed in this way is called *statistical independence.*
- Two variables are then statistically independent if their correlation is zero.
  - This a necessary but not sufficient condition

- As a matter of routine it is the squared correlations that should be interpreted. This is because the correlation coefficient is misleading in suggesting the existence of more covariation than exists, and this problem gets worse as the correlation approaches zero. Consider the following correlations and their squares.

- Note that as the correlation r decrease by tenths, the r^2 decreases by much more.
  - A correlation of .50 only shows that 25 percent variance is in common; a correlation of .20 shows 4 percent in common; and a correlation of .10 shows 1 percent in common (or 99 percent not in common).
- Thus, squaring should be a healthy corrective to the tendency to consider low correlations, such as .20 and .30, as indicating a meaningful or practical covariation.

r	r^2
1.00	1.00
.90	.81
.80	.64
.70	.49
.60	.36
.50	.25
.40	.16
.30	.09
.20	.04
.10	.01
.0	.0

Last word

- A key thing to remember when working with correlations is never to assume a correlation means that a change in one variable *causes* a change in another.