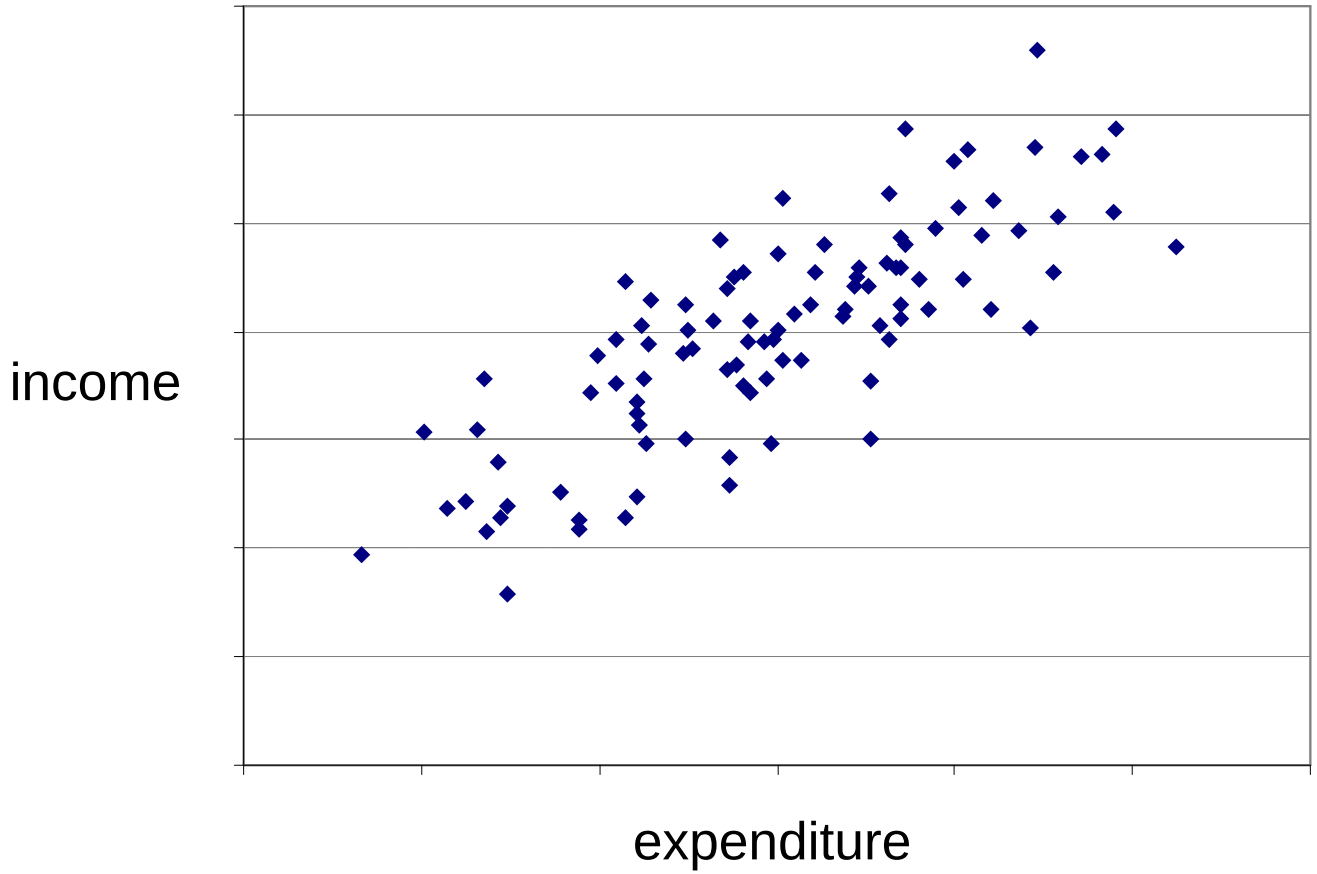
Correlation

Lecture 5

Correlation

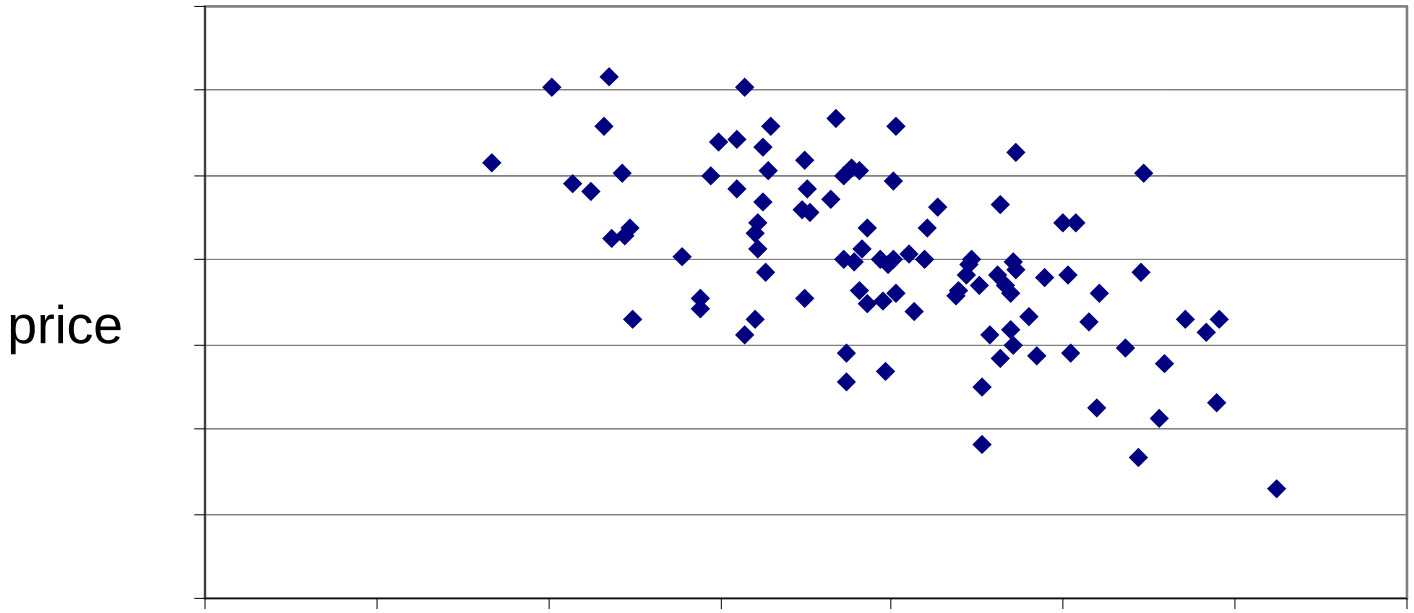
- Correlation examines the relationships between pairs of variables, for example
  - between the price of doughnuts and the demand for them
  - between economic growth and life expectancy
  - between hair colour and hourly wage
  - between rankings
- Such analyses can be useful for formulating policies

Positive Correlation



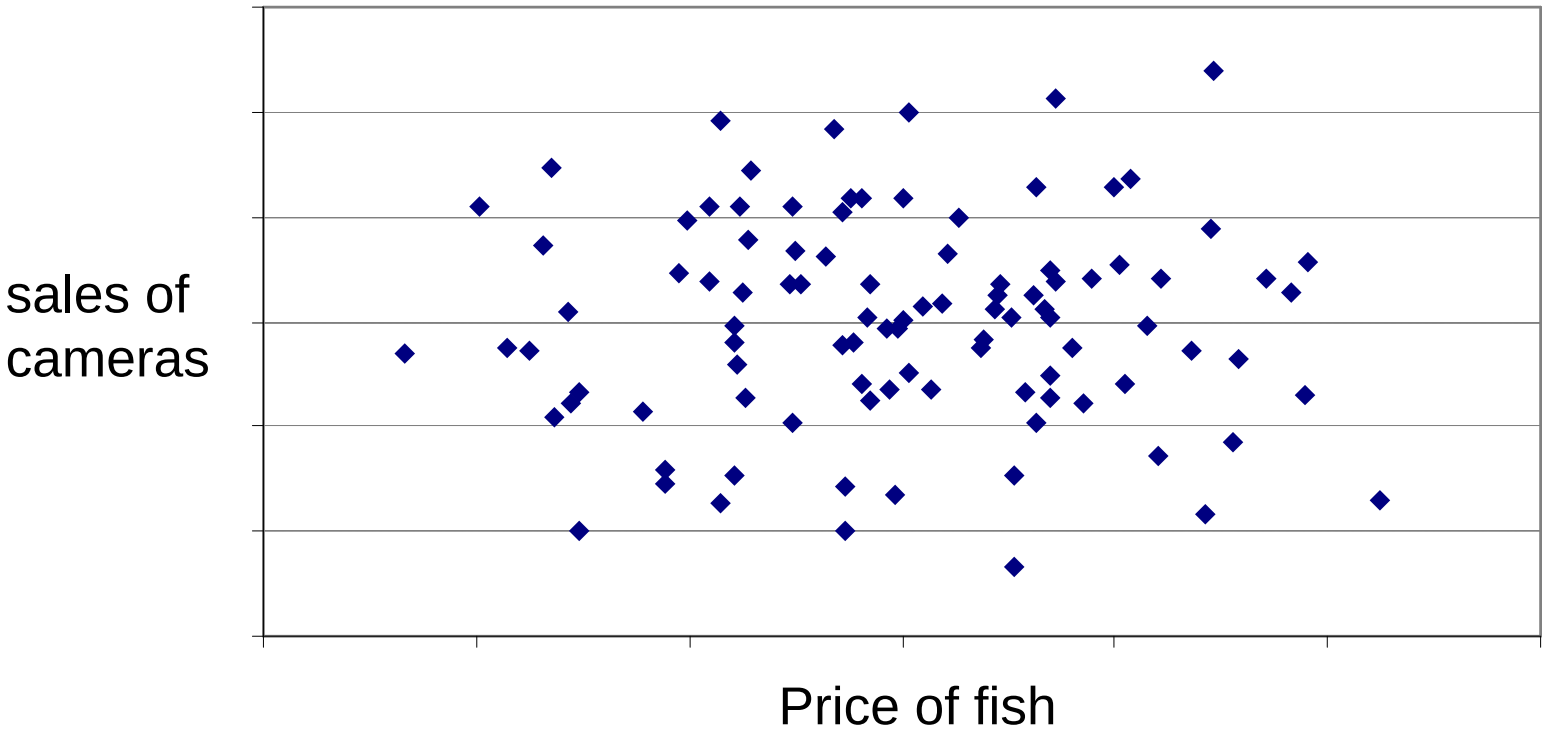
E.g. income and food expenditure

Negative Correlation



E.g. demand and price

Zero (absence of) Correlation



E.g. sales of cameras and the price of fish

The Correlation Coefficient, r

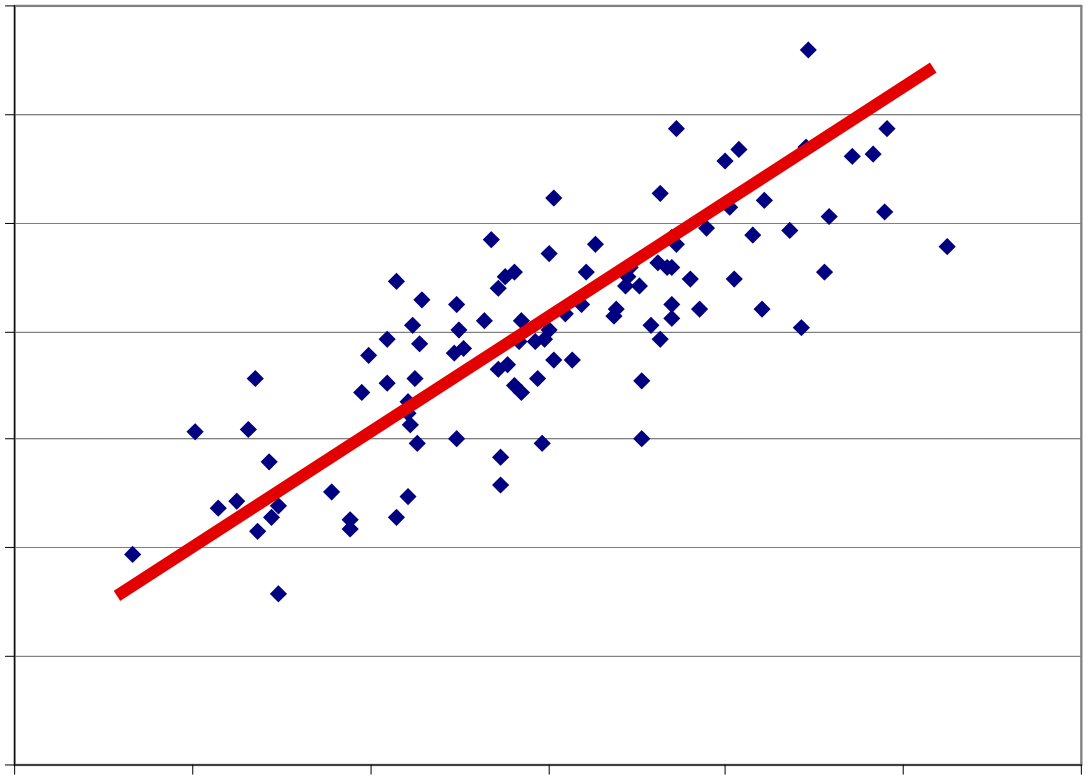
- Measures the strength of association between two variables, *X* and *Y*

$$-1 \leq r \leq +1$$

- Positive correlation: $r > 0$
- Negative correlation: $r < 0$
- Zero correlation: $r \approx 0$
- The closer *r* is to +1 (or -1), the closer the points lie to a straight line with positive (negative) slope

Positive Correlation

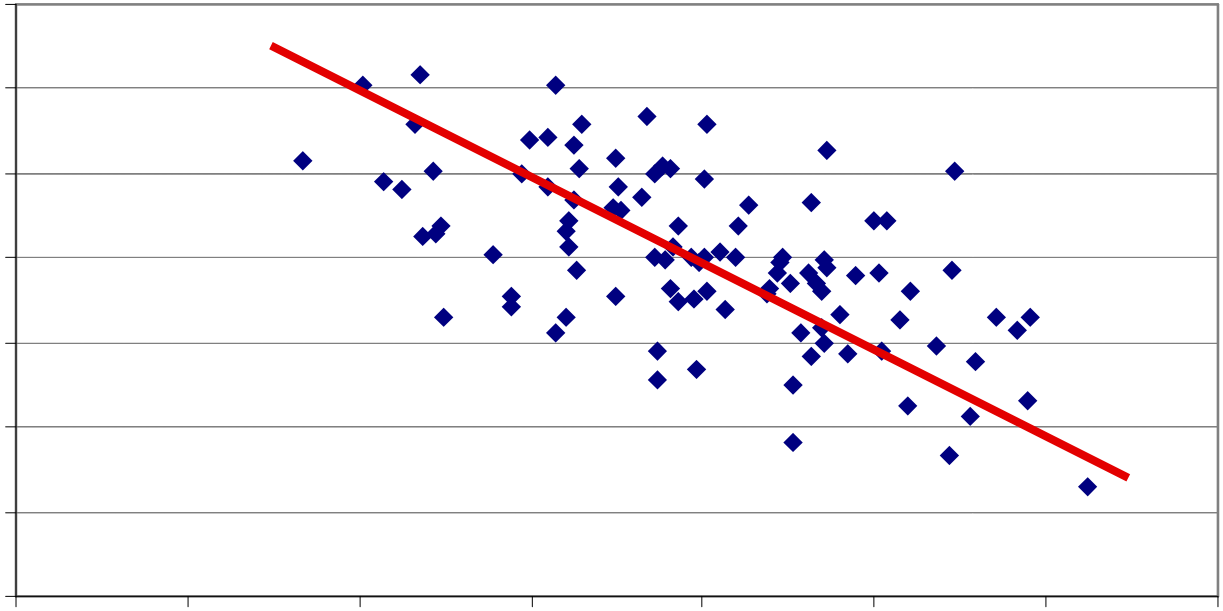
$r = 0.8$



E.g. income and food expenditure

Negative Correlation

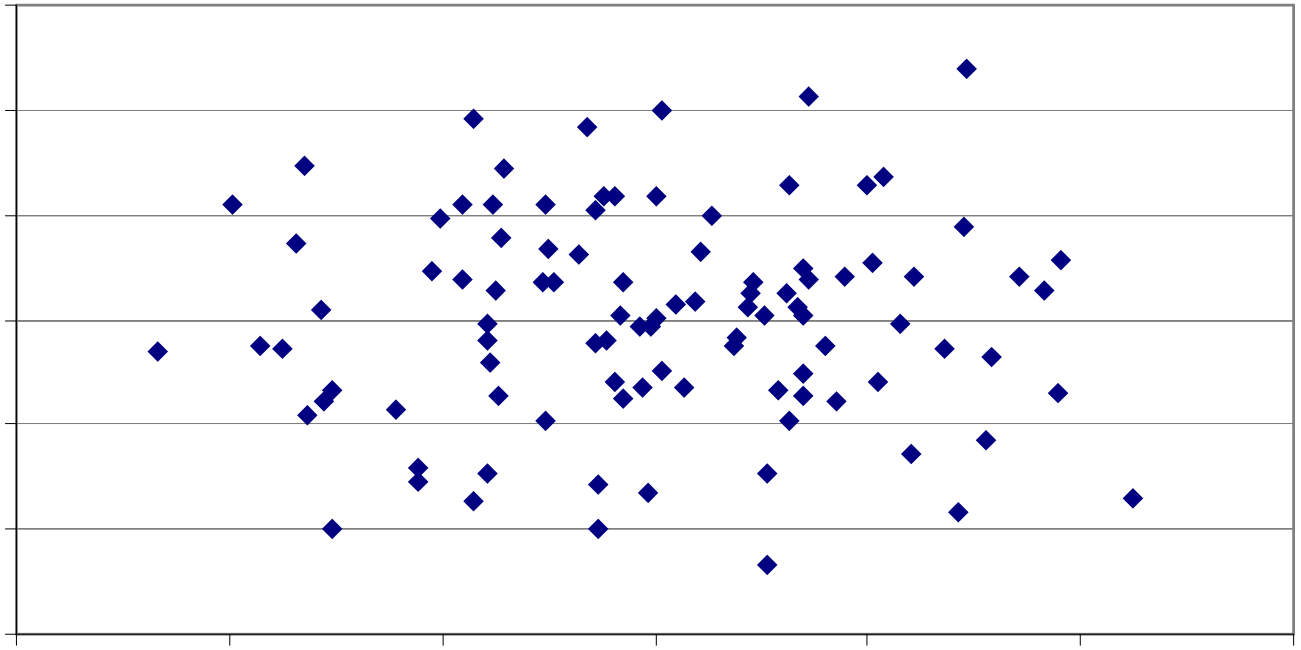
r = -0.7



E.g. demand and price

Zero (absence of) Correlation

$r = 0$



E.g. sales of cameras and the price of fish

Formula for the Correlation Coefficient

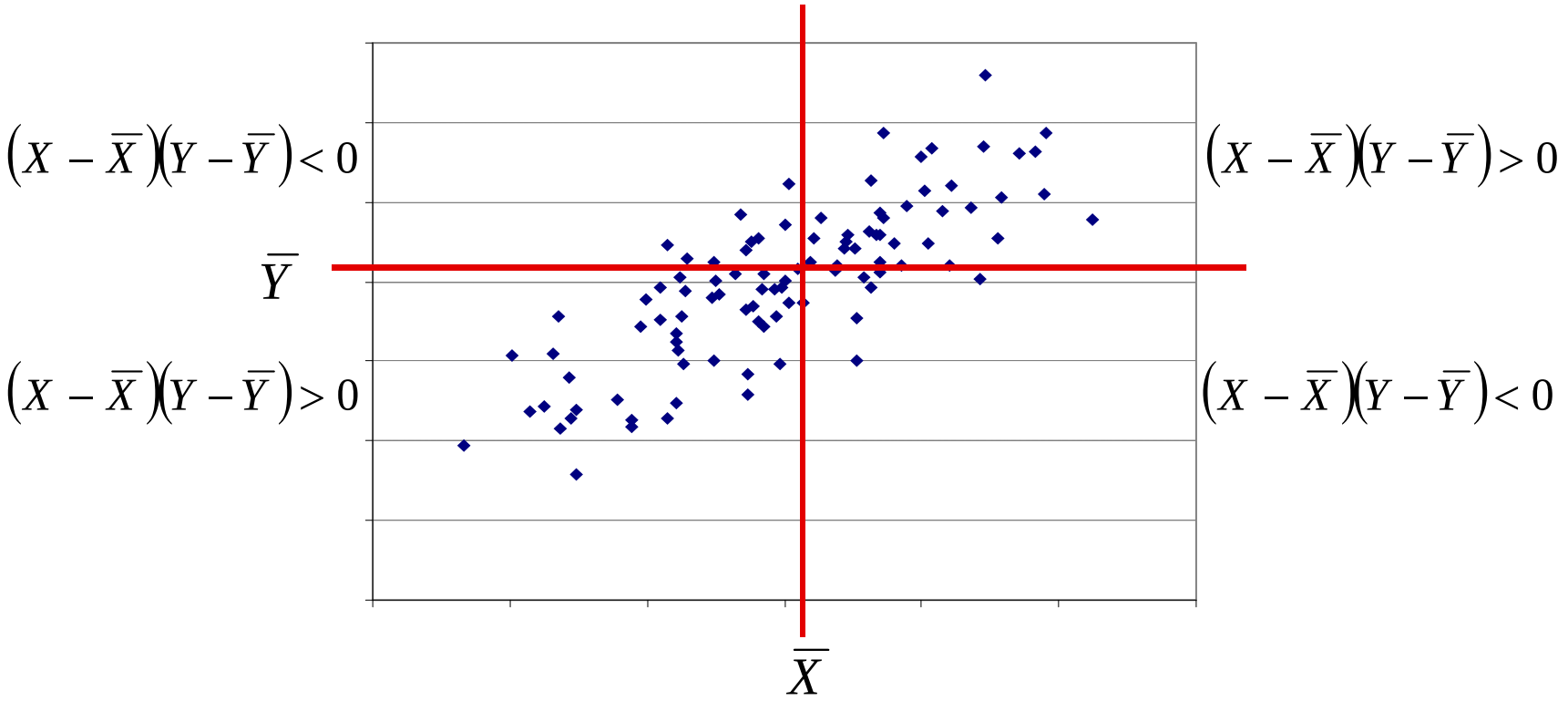
- Use either

$$r = \frac{\sum \left(X - \bar{X}\right)\left(Y - \bar{Y}\right)}{\sqrt{\sum \left(X - \bar{X}\right)^2 \times \sum \left(Y - \bar{Y}\right)^2}}$$

- or, equivalently

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

Why the Formula Works



More +ve than -ve points, hence $r > 0$

Calculation of *r* Between Growth and Birth Rates

Country	Birth rate	GNP growth			
	Y	X	Y^2	X^2	XY
Brazil	30	5.1	900	26.01	153.0
Colombia	29	3.2	841	10.24	92.8
Costa Rica	30	3.0	900	9.00	90.0
India	35	1.4	1,225	1.96	49.0
Mexico	36	3.8	1,296	14.44	136.8
Peru	36	1.0	1,296	1.00	36.0
Philippines	34	2.8	1,156	7.84	95.2
Senegal	48	–0.3	2,304	0.09	–14.4
South Korea	24	6.9	576	47.61	165.6
Sri Lanka	27	2.5	729	6.25	67.5
Taiwan	21	6.2	441	38.44	130.2
Thailand	30	4.6	900	21.16	138.0
Total	380	40.2	12,564	184.04	1,139.7

Calculation of *r* Between Growth and Birth Rates (cont.)

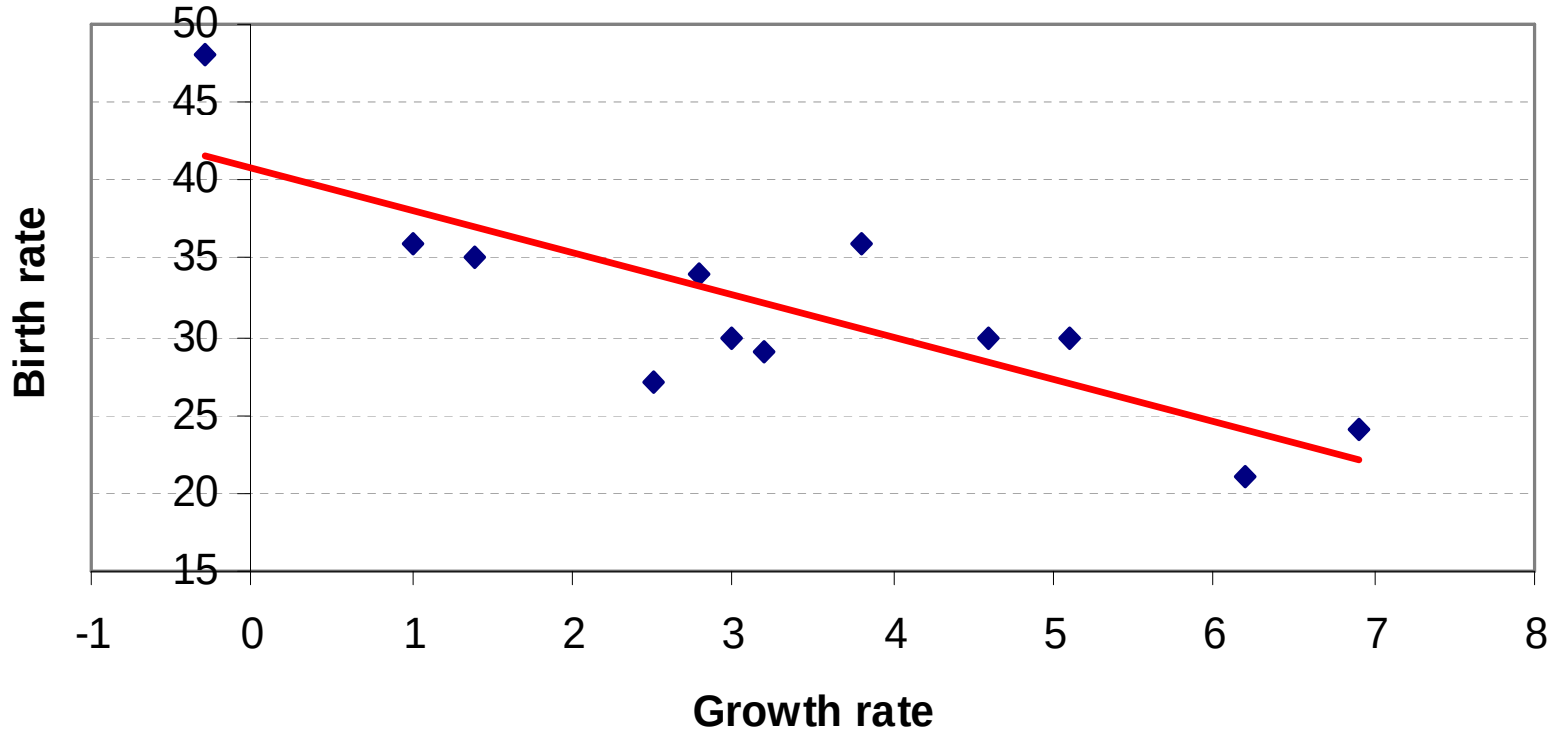
- Using the second formula,

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

- we obtain

$$r = \frac{12 \times 1{,}139.7 - 40.2 \times 380}{\sqrt{(12 \times 184.04 - 40.2^2)(12 \times 12{,}564 - 380^2)}} = -0.824$$

Chart of Birth Rate Against Growth Rate



Notes About *r*

- The correlation between *Y* and *X* is the same as between *X* and *Y*
  - it does not matter which variable is labelled *X* and which *Y*
- *r* is independent of units of measurement
  - If the birth rate were measured as births per 100 population (3.0, 2.9,...) *r* would still be -0.824
- Correlation does not imply causality

Is the Result Statistically Significant?

- The correlation coefficient *r* is just like any other summary statistic
- H_0: $\rho = 0$ versus H_1: $\rho \neq 0$ where ρ is the population correlation coefficient
  - The null asserts no genuine association between *X* and *Y;* the sample correlation observed is just due to (bad) luck
- The test statistic is

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

Hypothesis Testing

- Choose α = 5%. This implies t^*_{10} = 2.228
- Calculate the test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.824\sqrt{12-2}}{\sqrt{1-(-0.824)^2}} = -4.59$$

- Hence we reject H_0. There does seem to be genuine association between *growth* and *birth rates*

Spearman Rank Correlation Coefficient, r_s

- Used for examining relationships between ranks of variables
  - e.g. rankings of school performance and spending per pupil
- Can be useful if data contains extreme values that may distort the means and other summary statistics
- Based on looking at differences in ranks by each variable

Country	Birth rate	GNP growth				
	Y	*X*	Rank *Y*	Rank *X*	*d*	d^2
Brazil	30	5.1	7	3	4	16
Colombia	29	3.2	9	6	3	9
Costa Rica	30	3.0	7	7	0	0
India	35	1.4	4	10	-6	36
Mexico	36	3.8	2.5	5	-2.5	6.25
Peru	36	1.0	2.5	11	-8.5	72.25
Philippines	34	2.8	5	8	-3	9
Senegal	48	–0.3	1	12	–11	121
South Korea	24	6.9	11	1	10	100
Sri Lanka	27	2.5	10	9	1	1
Taiwan	21	6.2	12	2	10	100
Thailand	30	4.6	7	4	3	9
Total	380	40.2				479.5

$$r_s = 1 - \frac{6 \times \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 479.5}{12(144 - 1)} = -0.676$$

Hypothesis Testing With r_s

- r_s does not follow a standard distribution so we have to use special tables
- H_0: $\rho_s=0$ vs H_1: $\rho_s\neq0$
- so we can reject the null

n	10%	5%	2%	1%
5	0.9			
6	0.829	0.896	0.943	
..				
11	0.523	0.623	0.763	0.794
12	0.497	0.591	0.703	0.78

See Table A6 in Barrow for full version

Summary

- Correlation measures the association between two variables
  - correlation coefficient, r, if we have data values
  - Spearman rank correlation coefficient, if just have ranks, or have extreme values
- Note that association does not infer causation