Introduction to Correlation Analysis

Learning Objectives

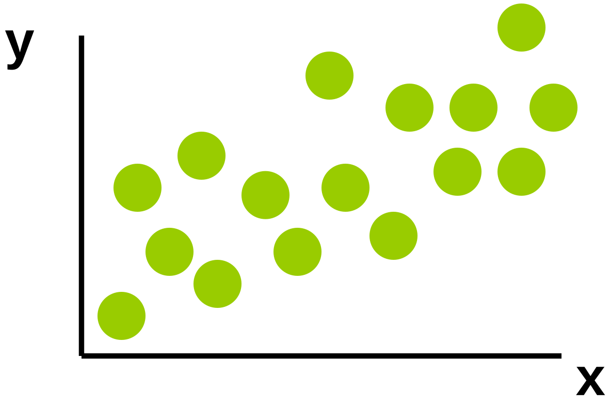
- Scatter Plots
- Correlation
- Correlation Coefficient
- Calculating the Correlation Coefficient
- Hypotheses
- Test statistic
- level of significance

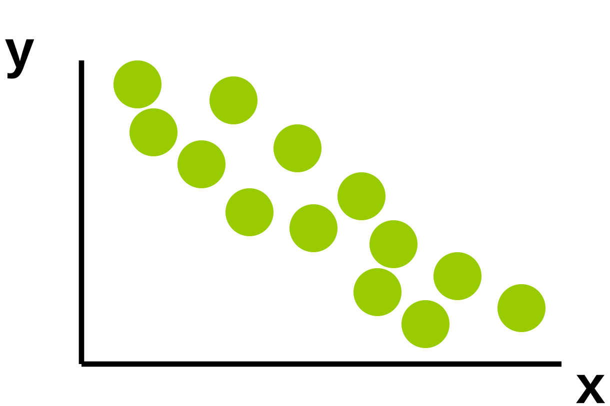
Scatter Plots and Correlation

- A **scatter plot** (or scatter diagram) is used to show the relationship between two variables
- **Correlation** analysis is used to measure strength of the association (linear relationship) between two variables
  - Only concerned with strength of the relationship
  - No causal effect is implied

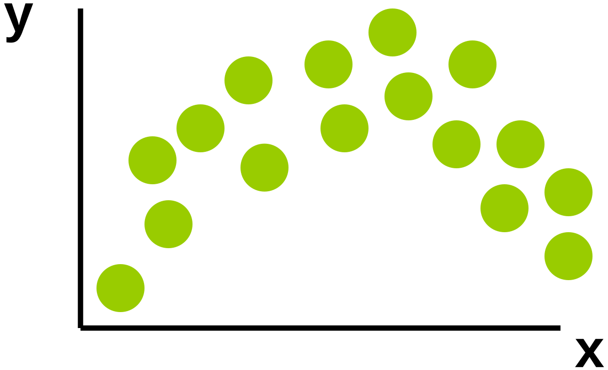
Scatter Plot Examples

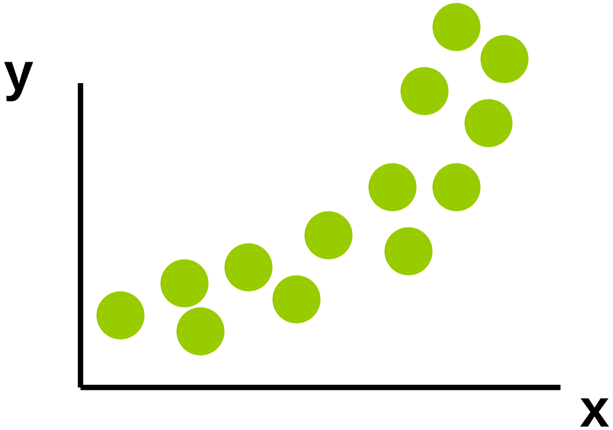
Linear relationships





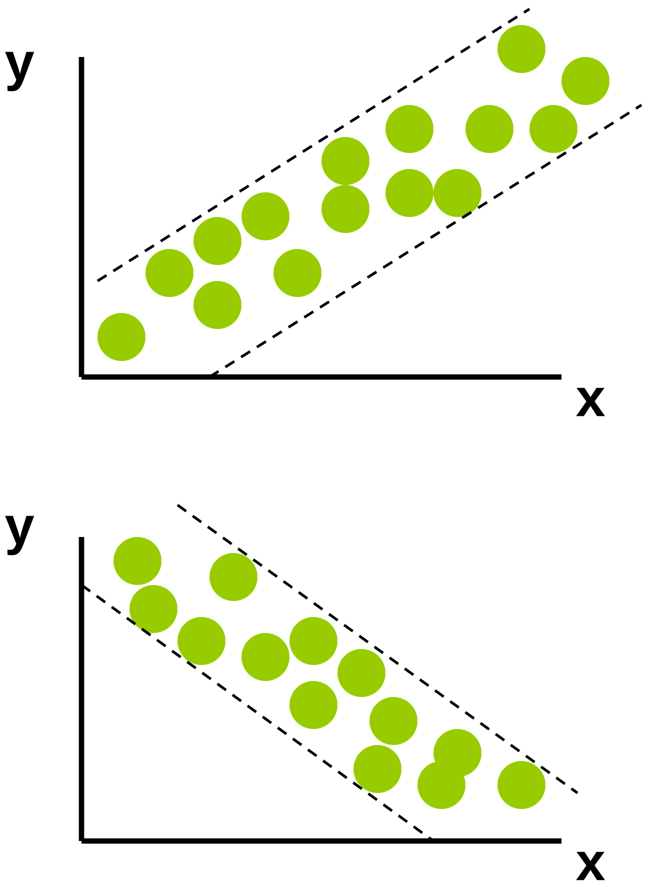
Curvilinear relationships



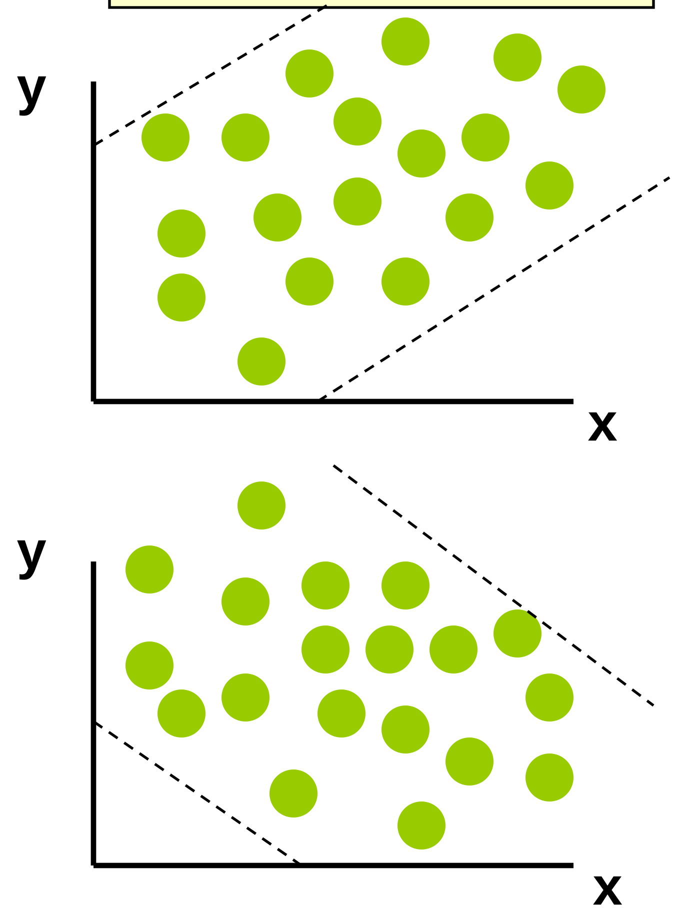


Scatter Plot Examples

Strong relationships

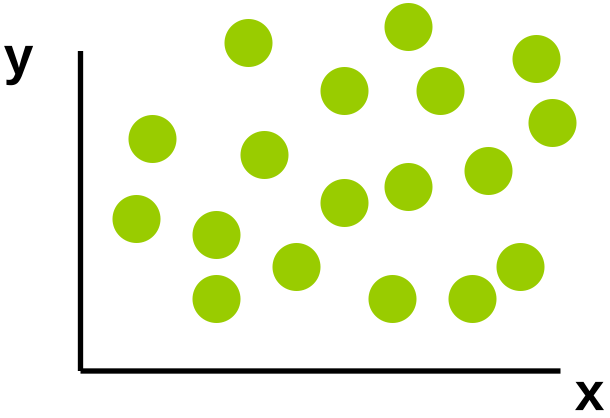


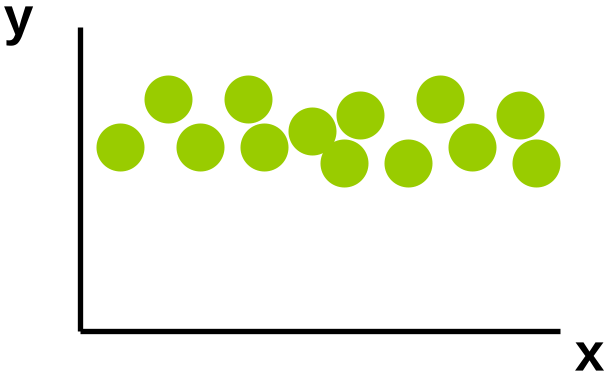
Weak relationships



Scatter Plot Examples

No relationship





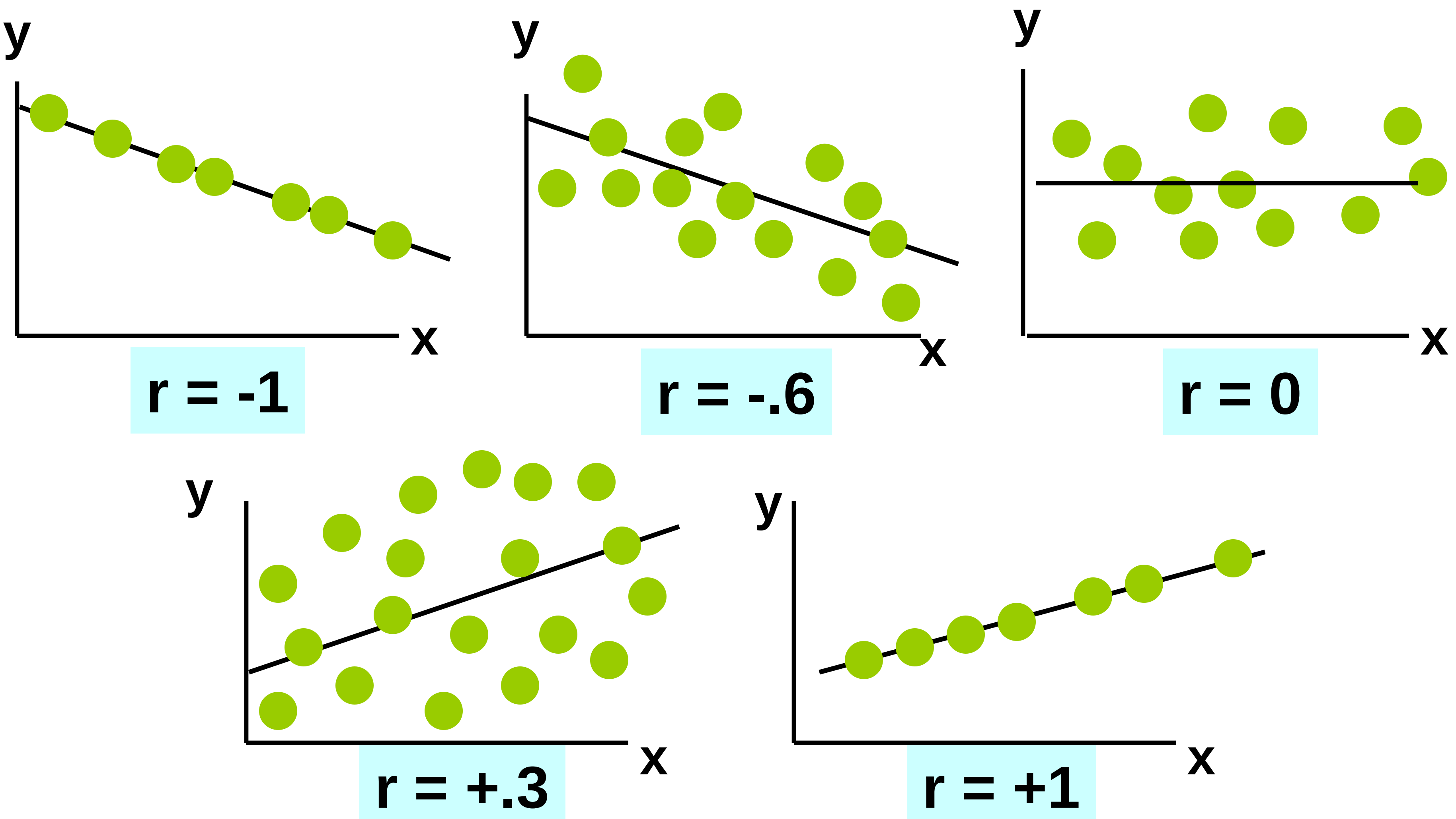
Correlation Coefficient

- The **population correlation coefficient ρ** (rho) measures the strength of the association between the variables
- The **sample correlation coefficient r** is an estimate of ρ and is used to measure the strength of the linear relationship in the sample observations

Features of ρ and r

- Unit free
- Range between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

Examples of Approximate r Values



Calculating the Correlation Coefficient

Sample correlation coefficient:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

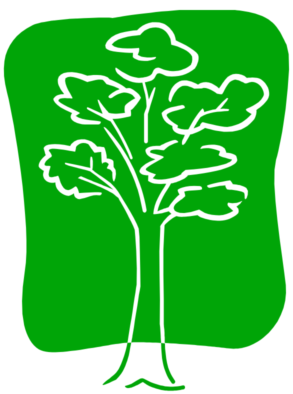
or the algebraic equivalent:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

where:

r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

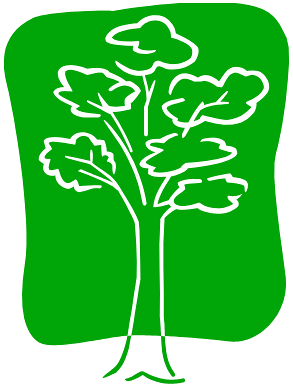
Calculation Example



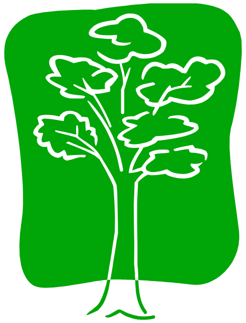
Tree Height	Trunk Diameter			
y	**x**	**xy**	**y^2**	**x^2**
35	8			
49	9			
27	7			
33	6			
60	13			
21	7			
45	11			
51	12			
$\sum =$	$\sum =$	$\sum =$	$\sum =$	$\sum =$

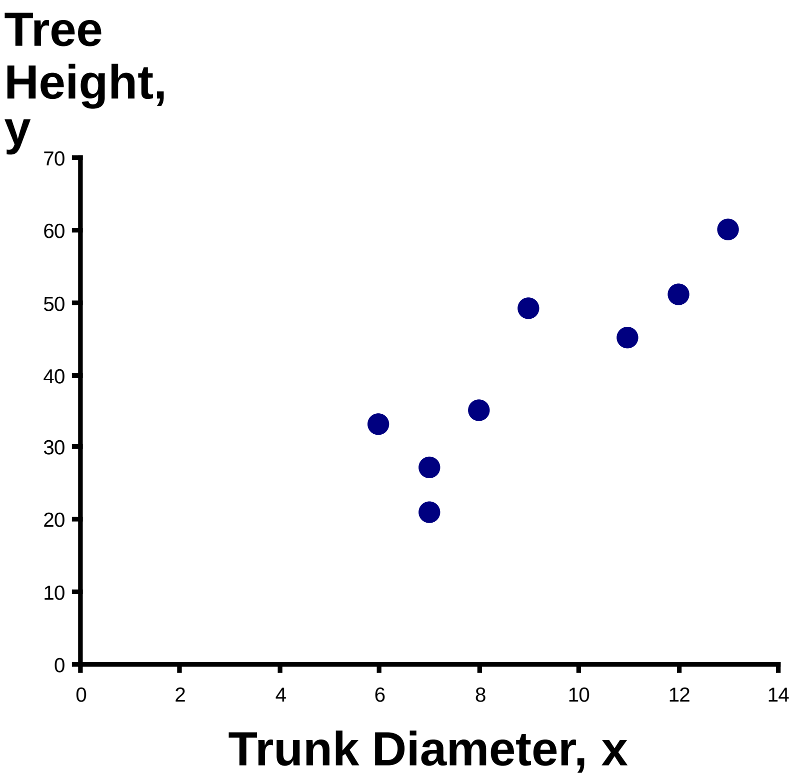
Calculation Example

Tree Height	Trunk Diameter			
y	**x**	**xy**	**y^2**	**x^2**
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
$\sum$ **=321**	$\sum$ **=73**	$\sum$**=3142**	$\sum$**=14111**	$\sum$ **=713**



Calculation Example





$$
\begin{aligned}
r &= \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} \\
&= \frac{8 \cdot 3142 - 73 \cdot 321}{\sqrt{[8 \cdot 713 - (73)^2][8 \cdot 14111 - (321)^2]}} \\
&= 0.886
\end{aligned}
$$

r = 0.886 → relatively strong positive linear association between x and y

Significance Test for Correlation

- Hypotheses

$$H_0: \rho = 0 \qquad \text{(no correlation)}$$

$$H_A: \rho \neq 0 \qquad \text{(correlation exists)}$$

- Test statistic
  - $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$ (with n – 2 degrees of freedom)

Example: Produce Stores

Is there evidence of a linear relationship between tree height and trunk diameter at the 0.05 level of significance?

H_0: ρ = 0 (No correlation)

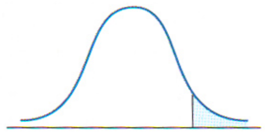
H_1: ρ ≠ 0 (correlation exists)

α =0.05 , df = 8 - 2 = 6

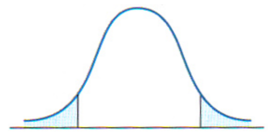
$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.886}{\sqrt{\frac{1-0.886^2}{8-2}}} = 4.68$$

TABLE B.2 THE *t* DISTRIBUTION

Table entries are values of *t* corresponding to proportions in one tail or in two tails combined.



One tail
(either right or left)



Two tails
combined

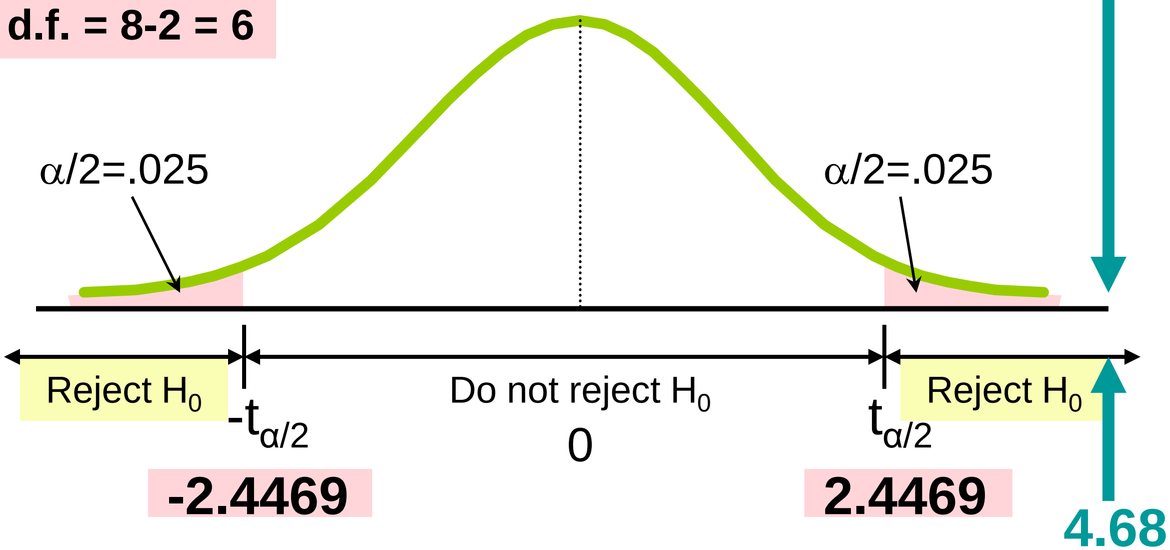
	PROPORTION IN ONE TAIL					
	0.25	0.10	0.05	0.025	0.01	0.005
	PROPORTION IN TWO TAILS COMBINED					
df	0.50	0.20	0.10	0.05	0.02	0.01
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
40	0.681	1.303	1.684	2.021	2.423	2.704
60	0.679	1.296	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
∞	0.674	1.282	1.645	1.960	2.326	2.576

ble III of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, 6th ed. London: Longman Group Ltd., 74 (previously published by Oliver and Boyd Ltd., Edinburgh). Adapted and reprinted with permission of the Addison Wesley Longman blishing Co.

Example: Test Solution

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.886}{\sqrt{\frac{1-0.886^2}{8-2}}} = 4.68$$

d.f. = 8-2 = 6



Decision:
Reject H_0

Conclusion:
There **is evidence** of a linear relationship at the 5% level of significance