SECTION 4.1 – Scatter Diagrams and Correlation

Response and Explanatory Variables

The **response variable** is the variable whose value can be explained by the value of the **explanatory** or **predictor variable**.

Scatter Diagram (or Scatterplot)

A **scatter diagram** (**or scatterplot**) is a graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram. The explanatory variable is plotted on the horizontal axis, and the response variable is plotted on the vertical axis.

Positively Associated

Two variables that are linearly related are positively associated when above-average values of one variable are associated with above-average values of the other variable and below-average values of one variable are associated with below-average values of the other variable. That is, two variables are positively associated if, whenever the value of one variable increases, the value of the other variable also increases.

Negatively Associated

Two variables that are linearly related are negatively associated when above-average values of one variable are associated with below-average values of the other variable. That is, two variables are negatively associated if, whenever the value of one variable increases, the value of the other variable decreases.

Linear Correlation Coefficient, *r*

The **linear correlation coefficient** or Pearson product moment correlation coefficient is a measure of the strength and direction of the linear relation between two quantitative variables. The Greek letter ρ (rho) represents the population correlation coefficient, and r represents the sample correlation coefficient. We present only the formulas for the sample correlation coefficient.

Conceptual Formula

$$r = \frac{\Sigma\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$$

Computational Formula

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

where x_i is the ith observation of the explanatory variable

$\bar{x}$ is the sample mean of the explanatory variable

s_x is the sample standard deviation of the explanatory variable

y_i is the ith observation of the response variable

$\bar{y}$ is the sample mean of the response variable

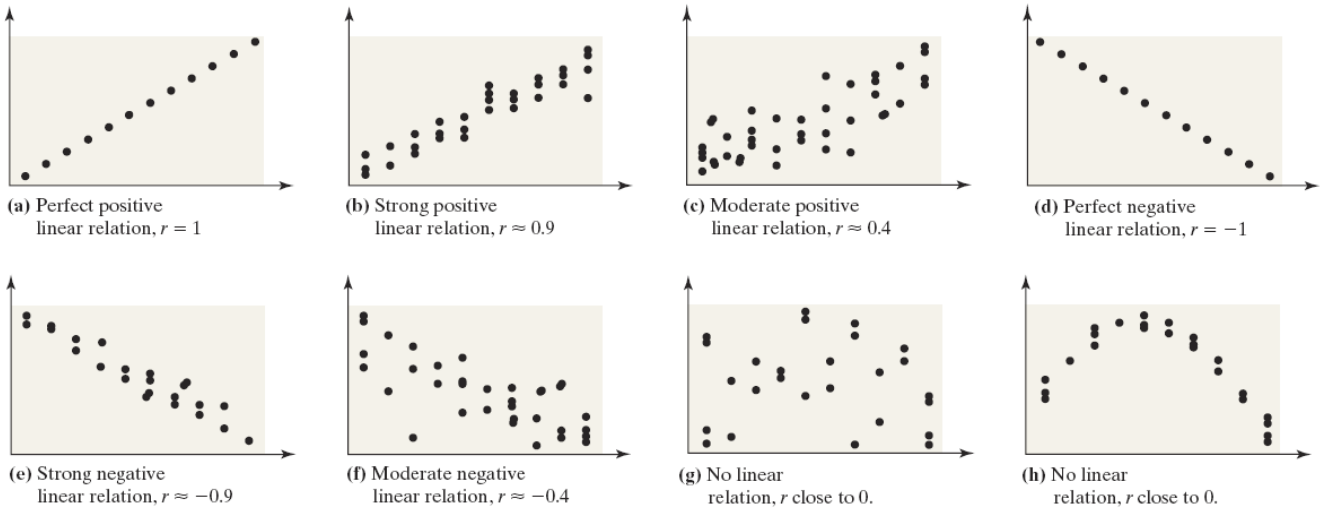
s_y is the sample standard deviation of the response variable

n is the number of individuals in the sample

SECTION 4.1 – Scatter Diagrams and Correlation

Properties of the Linear Correlation Coefficient

1. The linear correlation coefficient is always between –1 and 1, inclusive. That is, $-1 \le r \le 1$.
2. If $r = +1$, then a perfect positive linear relation exists between the two variables. See Figure (a).
3. If $r = -1$, then a perfect negative linear relation exists between the two variables. See Figure (d).
4. The closer r is to +1, the stronger is the evidence of positive association between the two variables. See Figures (b) and (c).
5. The closer r is to –1, the stronger is the evidence of negative association between the two variables. See Figures (e) and (f).
6. If r is close to 0, then little or no evidence exists of a linear relation between the two variables. So, r close to 0 does not imply no relation, just no linear relation. See Figures (g) and (h).
7. The linear correlation coefficient is a unitless measure of association. So, the unit of measure for x and y plays no role in the interpretation of r.
8. The correlation coefficient is not resistant. Therefore, an observation that does not follow the overall pattern of the data could affect the value of the linear correlation coefficient.



(a) Perfect positive linear relation, $r = 1$

(b) Strong positive linear relation, $r \approx 0.9$

(c) Moderate positive linear relation, $r \approx 0.4$

(d) Perfect negative linear relation, $r = -1$

(e) Strong negative linear relation, $r \approx -0.9$

(f) Moderate negative linear relation, $r \approx -0.4$

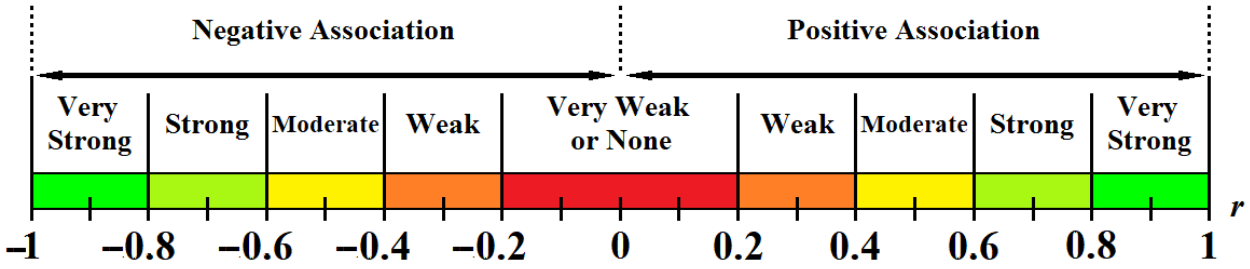
(g) No linear relation, r close to 0.

(h) No linear relation, r close to 0.

Note: As mentioned on the page previously, the correlation coefficient r is also called Pearson's product moment correlation, as well as Pearson's r. This was named after Karl Pearson, an English mathematician and biometrician, who is considered to be the founder of modern statistics.

SECTION 4.1 – Scatter Diagrams and Correlation

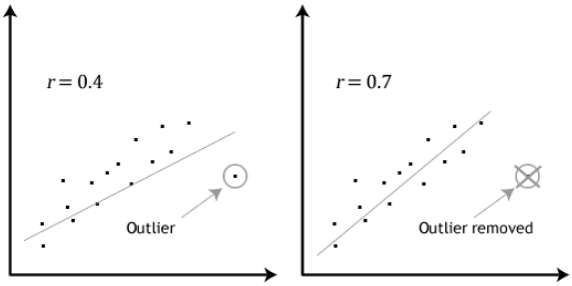
Here is an alternate scale or guide on the strength of the correlation coefficient, *r*. (These are only crude estimates for interpreting strengths of correlations and these estimates are not discussed in the text).



Algebraically, these are the guidelines:

$0 \le \lvert r \rvert < 0.2$	Very weak or no correlation	$0.6 \le \lvert r \rvert < 0.8$	Strong (or moderately strong) correlation
$0.2 \le \lvert r \rvert < 0.4$	Weak correlation	$0.8 \le \lvert r \rvert < 1$	Very Strong correlation
$0.4 \le \lvert r \rvert < 0.6$	Moderate correlation	$\lvert r \rvert = 1$	Perfect correlation

Pearson's *r* is sensitive to outliers, which can have a very large effect on the Pearson correlation coefficient as well as the line of best fit (which will be discussed in the next section), leading to very difficult conclusions regarding your data. Therefore, it is best if there are no outliers or they are kept to a minimum. Here is an example of how the removal (or inclusion) of an outlier changes the value of the correlation coefficient.



Don't ever assume the relationship is linear just because the correlation coefficient is high. A relationship can be strong and yet not significant. Conversely, a relationship can be weak but significant. The key factor is the size of the sample. For small samples, it is easy to produce a strong correlation by chance and one must pay attention to significance to keep from jumping to conclusions. So, to determine whether a relationship is linear or not linear, we must conduct the following test.

Testing for a Linear Relation

Step 1 Determine the absolute value of the correlation coefficient.
Step 2 Find the critical value in Table II below or from Appendix A or website for the given sample size.
Step 3 If the absolute value of the correlation coefficient is greater than the critical value, we say a linear relation exists between the two variables. Otherwise, no linear relation exists.

Table II – Critical Values (CV) for Correlation Coefficient

n	CV		*n*	CV		*n*	CV		*n*	CV
3	0.997		10	0.632		17	0.482		24	0.404
4	0.950		11	0.602		18	0.468		25	0.396
5	0.878		12	0.576		19	0.456		26	0.388
6	0.811		13	0.553		20	0.444		27	0.381
7	0.754		14	0.532		21	0.433		28	0.374
8	0.707		15	0.514		22	0.423		29	0.367
9	0.666		16	0.497		23	0.413		30	0.361

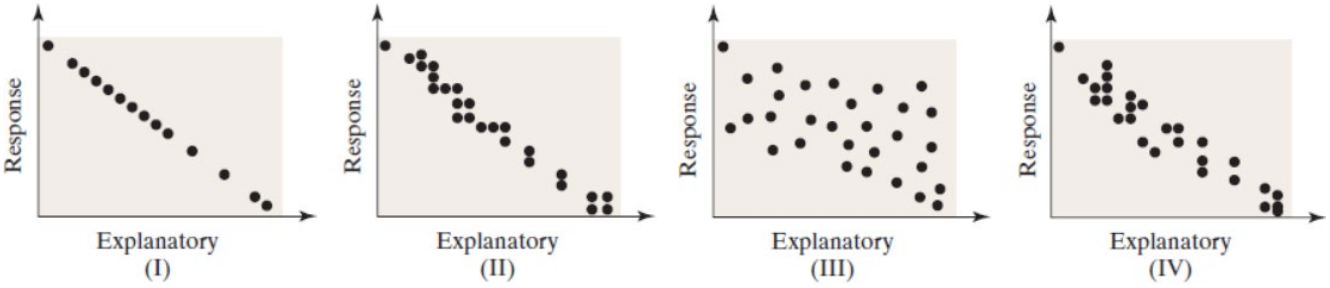
Note: The critical values above were obtained using a significance level of 5%. This terminology will be discussed in greater detail in Chapter 10.

SECTION 4.1 – Scatter Diagrams and Correlation

☺ **Exercises**:

1) Match the linear correlation coefficient to the scatter diagram. Assume the scales on the x- and y-axes are the same for each diagram.

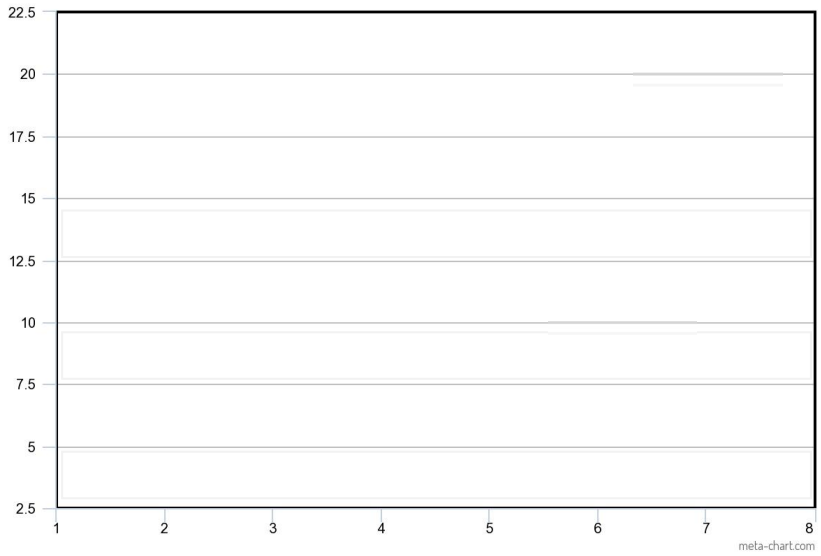
a) $r = -0.969$ **b)** $r = -0.249$ **c)** $r = -1$ **d)** $r = -0.992$



2) Consider the following data set:

x	2	4	6	6	7
y	4	8	10	13	20

a) Draw a scatter diagram of the data.



SECTION 4.1 – Scatter Diagrams and Correlation

☺ **Exercises**:

b) Compute the correlation coefficient.

x	y	xy	x^2	y^2

c) Based off the correlation coefficient, describe the association between x and y.

d) Determine whether there is a linear relation between x and y.

Testing for a Linear Relation

Step 1 Determine the absolute value of the correlation coefficient.

Step 2 Find the critical value in Table II or from Appendix A or website for the given sample size.

Step 3 If the absolute value of the correlation coefficient is greater than the critical value, we say a linear relation exists between the two variables. Otherwise, no linear relation exists.