1) Two-way scatter plot

- Thus far we have focused on only a single variable, say $X$. Suppose however, that we are interested in a pair of continuous variables, say $X$ and $Y$, which measure two different attributes of elements in a population. One question of natural interest is whether there is any relationship, or association, between these variables. Correlation is defined as the quantification of the degree of association.

- As an example, consider a study of the effectiveness of childhood DPT vaccination. A random sample of 20 countries is taken, with $X$ representing the % of children immunized by age one, and $Y$ the under-five mortality rate. Each country has a pair of $(x_i, y_i)$ outcomes, as listed in Table 17.1. These data may also be plotted, to give a scatter plot, as on Figure 17.1. It seems apparent that there is a definite association between $X$ and $Y$, specifically that $Y$ tends to decrease as $X$ increases, as one might expect. A correlation analysis will quantify the nature and degree of association.

2) Pearson correlation coefficient

- The common measure of linear association in a population is denoted $\rho$. Conceptually it is the average of the product of the standard normal deviates of $X$ and $Y$, i.e.

$$\rho = \text{average} \left[ \frac{(X - \mu_x)(Y - \mu_y)}{\sigma_x \sigma_y} \right].$$

How could one interpret this?

- The estimate of $\rho$ is called the (Pearson) correlation coefficient $r$. It is defined as

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

The computational formula is

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (1)$$

- The $r$ statistic has several properties:

1) $-1 \leq r \leq 1$
2) $r = 1$ and $r = -1$ indicate that all points lie on a straight line, with perfect positive or negative correlation, respectively.
3) $r = 0$ indicate the absence of linear correlation.

Sample scatter plots are given in Figure 17.2.
Let’s illustrate the calculation of $r$ using (1), as follows:

<table>
<thead>
<tr>
<th>Obs</th>
<th>$x$</th>
<th>$y$</th>
<th>$x - \bar{x}$</th>
<th>$y - \bar{y}$</th>
<th>$(x - \bar{x})^2$</th>
<th>$(y - \bar{y})^2$</th>
<th>$(x_i - \bar{x})(y_i - \bar{y})$</th>
<th>$(y_i - \bar{y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>77</td>
<td>118</td>
<td>-.4</td>
<td>.16</td>
<td>-23.6</td>
<td>3481</td>
<td>-23.6</td>
<td>3481</td>
</tr>
<tr>
<td>2</td>
<td>69</td>
<td>65</td>
<td>-8.4</td>
<td>6</td>
<td>70.56</td>
<td>36</td>
<td>-50.4</td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>184</td>
<td>-45.4</td>
<td>125</td>
<td>2061.16</td>
<td>-5675</td>
<td>15625</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>90</td>
<td>9</td>
<td>12.6</td>
<td>-50</td>
<td>158.76</td>
<td>-630</td>
<td>2500</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1548</td>
<td>1180</td>
<td>0</td>
<td>0</td>
<td>10630.8</td>
<td>-22706</td>
<td>77498</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>77.4</td>
<td>59.0</td>
<td>0</td>
<td>0</td>
<td>10630.8</td>
<td>-22706</td>
<td>77498</td>
<td></td>
</tr>
</tbody>
</table>

Note $r = \frac{-22706}{\sqrt{(10630.8)(77498)}} = -0.79$

Clearly, $r$ indicates a negative association in the sample data. One chief objective is to test whether we have statistical evidence of (negative) association in the population, i.e. to test

$H_0 : \rho = 0$.

The estimated standard error of $r$ is

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

If $X$ and $Y$ have normal distributions, one can find a $t$ statistic with $n - 2$ df as

$t = \frac{r}{s_r}$

$$= \frac{r}{\sqrt{\frac{n-2}{1-r^2}}} \quad (2)$$

For example, with our data the test would be

1) $H_0 : \rho = 0$

2) $H_0 : \rho < 0$

3) $TS : \quad t = r / s_r$

4) $RR : \quad$ for $\alpha = .05$ with 18 df

$t < -1.734$

5) Calculations: $t = -0.79 \sqrt{\frac{18}{1-.79^2}} = -5.47$

hence reject $H_0$.

In summary, the data provides strong evidence ($p < .0005$) of negative correlation between the childhood immunization rate and infant mortality of a country.
• It is more difficult to provide confidence intervals for \( \rho \). Specialized tables are available for this in many books, including *Biometrika Tables for Statisticians* by E. Pearson and H. O. Hartley.

• Clearly, this tests only linear association. The test is also sensitive to outliers, which leads to the following alternative analysis.

3) **Spearman rank correlation coefficient.**

• If the data are highly skewed, with outliers, one may use the Spearman rank correlation coefficient, \( r_s \). The logic is simple, one would replace the raw \( X \) and \( Y \) data with their respective ranks, and then calculate the correlation coefficient from (1). Data consisting of ranks are the basis of nonparametric methods in Chapter 13. Rank data clearly do not have outliers.

• Though one could substitute the rank data into (1), an equivalent, easier formula is

\[
r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}
\]

where \( d_i \) is the difference between the rank of \( x_i \) and the rank of \( y_i \). Because \( r_s \) is a special case of \( r \), it has the same properties.

• To illustrate, suppose the previous observations are reordered based on their \( X \) rank. This is not necessary, but sometimes clarifies the presentation, as the book illustrates. If one has ties, the average rank is used, as illustrated, for example, with India and Egypt which both have \( X = 89 \). They constitute the 11\textsuperscript{th} and 12\textsuperscript{th} values, hence each is given rank 11.5. The table of \( X \) and \( Y \) ranks, with the calculations are:

<table>
<thead>
<tr>
<th>Obs</th>
<th>( X ) rank</th>
<th>( Y ) rank</th>
<th>( d_i )</th>
<th>( d_i^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>20</td>
<td>-19</td>
<td>361</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>19</td>
<td>-17</td>
<td>289</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>18</td>
<td>-15</td>
<td>225</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>11</td>
<td>11.5</td>
<td>17</td>
<td>-5.5</td>
<td>30.25</td>
</tr>
<tr>
<td>12</td>
<td>11.5</td>
<td>13</td>
<td>-1.5</td>
<td>2.25</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>8</td>
<td>12</td>
<td>144</td>
</tr>
</tbody>
</table>

\[
\sum d_i^2 = 2045.5
\]

Therefore
\[ r_s = 1 - \frac{6(2045.5)}{20(399)} = -0.54 \]

- If \( n \geq 10 \) (and the data are a random sample), we can use the test statistic in (2) to test \( H_0 : \rho = 0 \).

In this case,

\[ t = -0.54 \sqrt{\frac{18}{1 - 0.54^2}} = -2.72 \]

The rejection region is again \( t < -1.734 \) hence we would reject \( H_0 \).

- If \( n < 10 \), there are specialized tables which give the distribution of \( r_s \) assuming \( H_0 \) is true. This may be used to test \( H_0 \).

- In the example, we conclude that the ranks are linearly correlated. In terms of the original \( X, Y \) data, this implies that there is a monotonically increasing (or decreasing) association between the variables.