MDM4U **Scatter Plot & Linear Correlation** Date: _____

Scatter Plot

Scatter plots show the relationship between two variables displaying data points on a two-dimensional graph. The variable that might be considered as the **independent** (explanatory) variable is plotted on the horizontal axis, and the **dependent** (response) variable is plotted on the vertical axis.

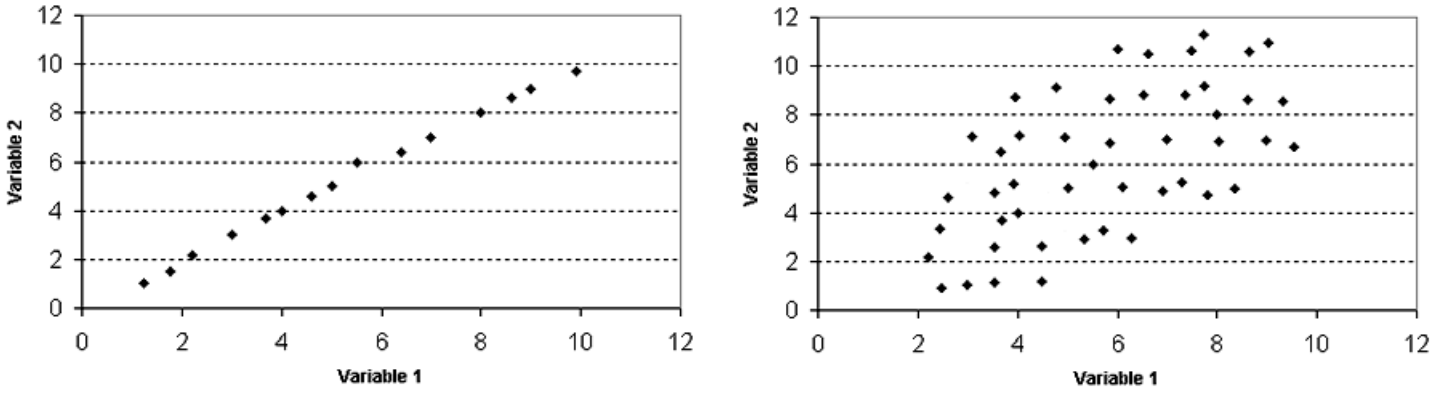
Scatter plots are especially useful when there is a large number of data. They provide the following information about the relationship between two variables:

- **Strength:** perfect, strong, moderate, or weak
- **Form:** linear, or non-linear such as quadratic, exponential, and trigonometric
- **Direction:** positive, or negative
- **Presence of outliers**
- **Spread of data:** concentrated, widely spread

Linear Correlation

Variables have a linear correlation if changes in one variable tend to be proportional to changes in the other. When the data points form a straight line on the graph, the linear relationship between the variables is stronger and the correlation is higher (Figure 1). The line of best fit is the straight line that passes as close as possible to all of the points on a scatter plot. The stronger the correlation, the more closely the data points cluster around the line of best fit.

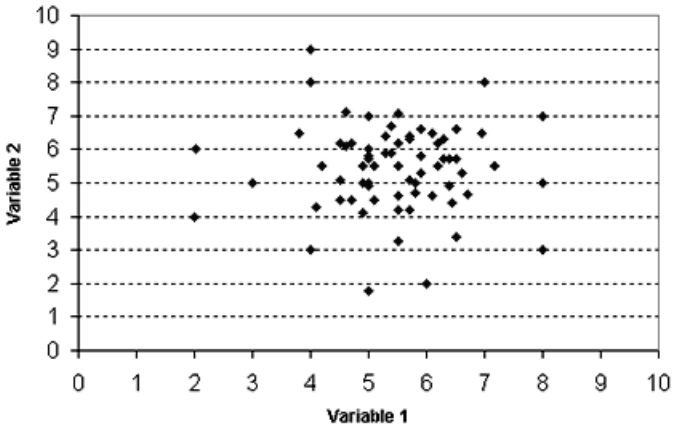
Figure 1. Strong vs. Weak Linear Correlation



Scattered data points

If the data points are randomly scattered, then there is no relationship between the two variables; this means there is a low or zero correlation between the variables (Figure 2).

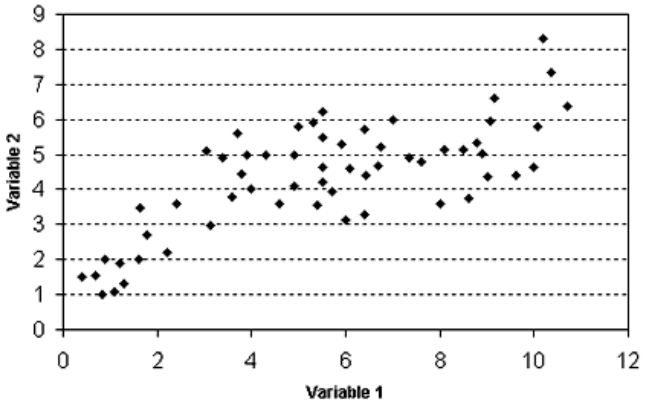
Figure 2. Scattered data points



Non-linear patterns

Very low or zero correlation may result from a non-linear relationship between two variables. If the relationship is, in fact, non-linear (i.e., points clustering around a curve, not a straight line), the correlation coefficient will not be a good measure of the strength of the relationship (Figure 3).

Figure 3. Very low or zero correlation



Positive or Direct Correlation vs. Negative or Inverse Correlation

If the trend of the data points rises to the right, then the relationship between the two variables is positive or direct (Figure 4). As X increases, Y also increases (same direction of change).

If the trend of the data points falls down to the right, then the relationship between the two variables is negative or inverse (Figure 5). As X increases, Y decreases (opposite direction of change).

Figure 4. Positive or Direct Relationship (Correlation)

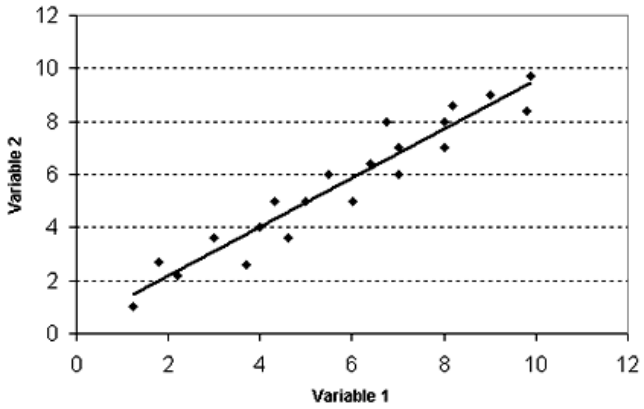
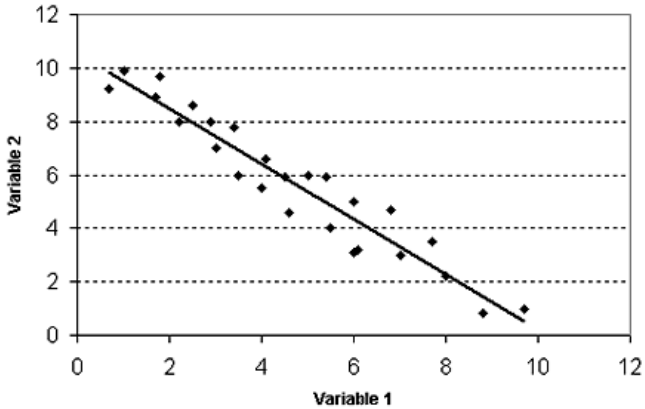


Figure 5. Negative or Inverse Relationship (Correlation)



Spread of data

A scatterplot will also illustrate if the data are widely spread or if they are concentrated within a smaller area (Figures 6 and 7).

Figure 6. Data concentrated

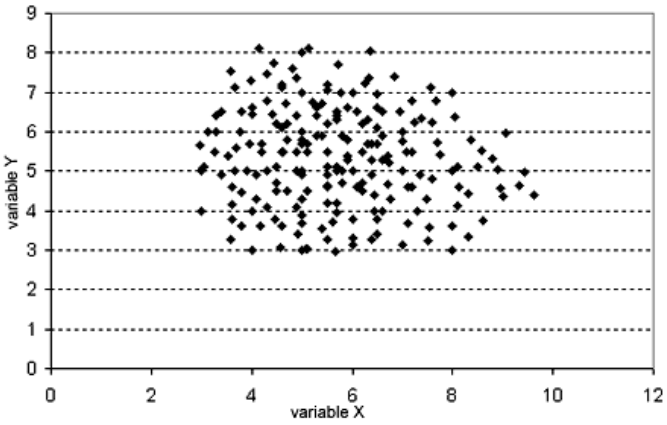
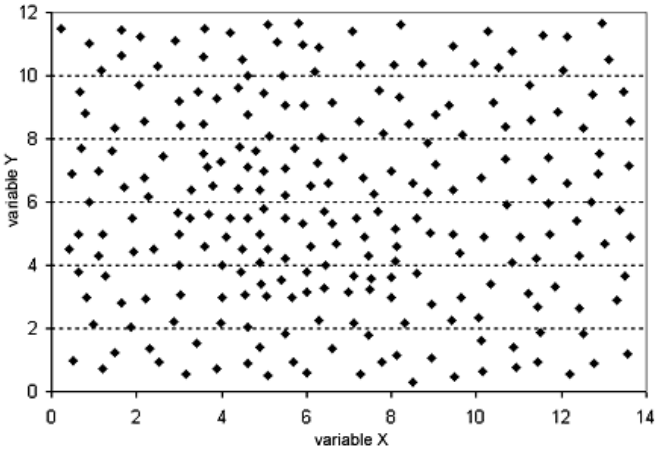


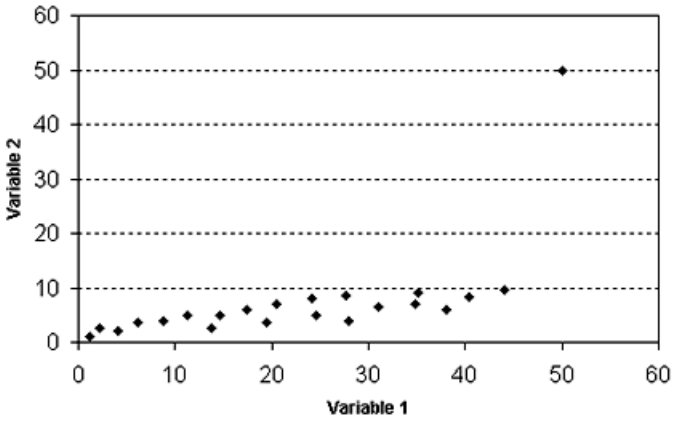
Figure 7. Data widely spread



Outliers

Besides portraying a non-linear relationship between the two variables, a scatterplot can also show whether or not there exist any outliers in the data (Figure 8).

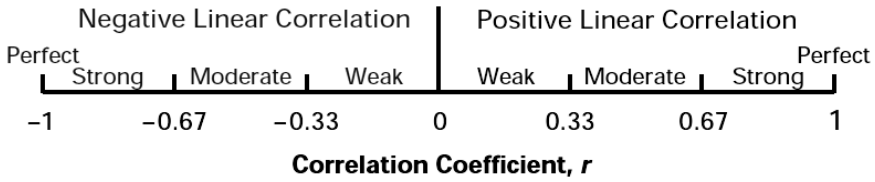
Figure 8. Outliers



Correlation Coefficient

The correlation coefficient, a concept from statistics, is a measure of how well trends in the predicted values follow trends in past actual values. It is a measure of how well the predicted values from a forecast model "fit" with the real-life data.

Correlation Coefficient (*r*) gives a quantitative measure of the strength of a linear correlation, indicating how closely the data points cluster around the line of best fit. The sign of *r* indicates the slope, while a number close to ±1 indicates a strong correlation and a number close to zero indicates a weak correlation. A perfect positive linear correlation between variables X and Y has a correlation coefficient of 1.



The Correlation Coefficient (*r*) is the covariance divided by the product of the standard deviations for X and Y.

$$r = \frac{s_{xy}}{s_x \times s_y}$$ where:

S_x is the standard deviation of X, S_y is the standard deviation of Y

Covariance of two variables in a sample is given by:

$$s_{xy} = \frac{1}{n-1}\sum (x-\bar{x})(y-\bar{y})$$ where:

n is the size of the sample

x represents individual values of the variable X, $\bar{x}$ is the mean of X

y represents individual values of the variable Y, $\bar{y}$ is the mean of Y

Using $\sum x = n\bar{x}$:

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