Part II: Two - Variable Statistics - Correlation

Statistical studies often involve more than one variable. We are interested in knowing if there is a relationship between the two .

Example: A person's age and the time spent using a mobile phone.

When the data is quantitative (numbers), the variables can be written as an ordered pair (x, y) and graphed on a Cartesian plane (called a scatterplot) .

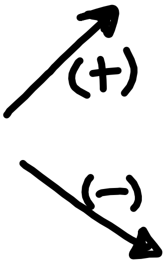
Correlation is the study and description of the relationship (if any) that exists between the variables.

A) Qualitative Interpretation of Correlation

Data can be organised and displayed in a scatterplot (Cartesian plane) or a contingency table.

By looking, we can describe the correlation - the direction, and the intensity (or strength) of the relation between the variables.

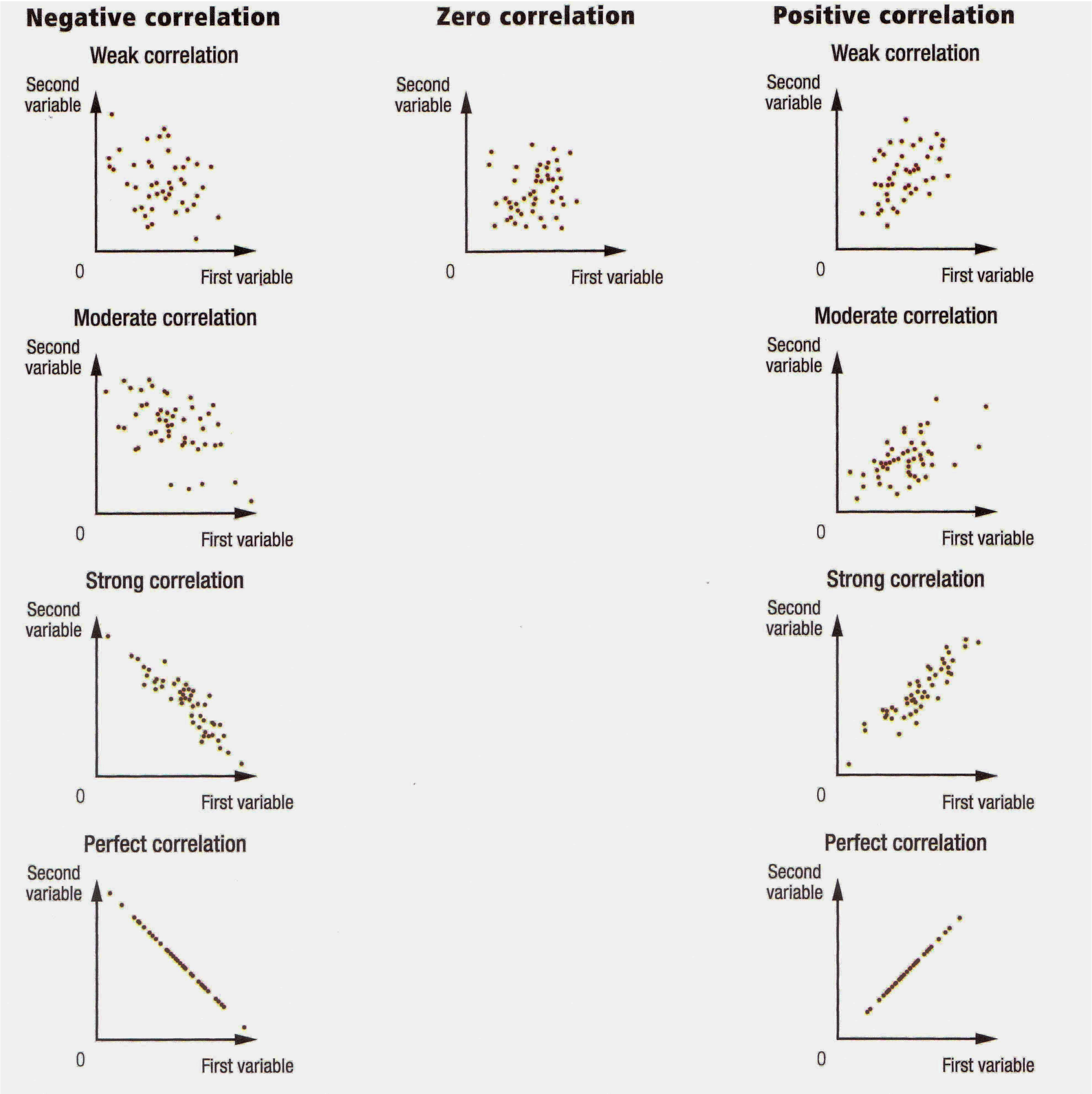
Direction: If both variables move in the same direction (increase together or decrease together), then the direction is positive.



If both variables move in opposite directions, then the direction is negative.

Intensity: Strength may be categorised as...
Zero, weak, moderate, strong or perfect.

Since we are doing linear correlation, the relationship is stronger the more the graph resembles a straight line.


Negative correlation
Zero correlation
Positive correlation
Weak correlation
Second variable
0
First variable
Moderate correlation
Strong correlation
Perfect correlation

B) Quantitative Interpretation of Correlation

The correlation will be represented by a number, called the correlation coefficient.

This coefficient will range from -1 to $+1$.

Its symbol is r.

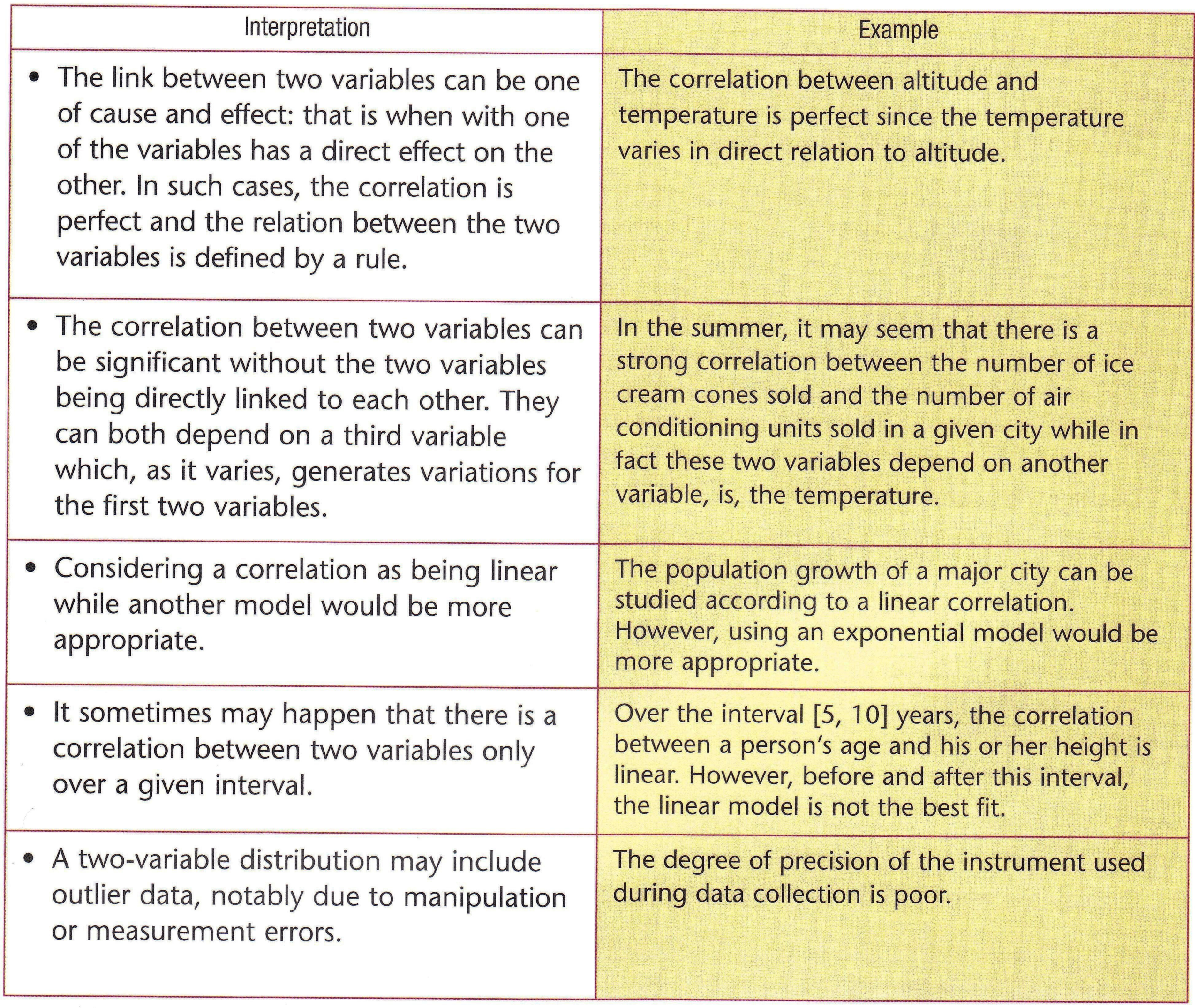
r	Meaning
Near 0	Zero correlation
Near ± 0.5	Weak correlation
Near ± 0.75	Moderate correlation
Near ± 0.87	Strong correlation
Near ± 1	Perfect correlation

Interpreting a Correlation

A strong correlation indicates that there is a statistical relationship between two variables.

It does not, however, explain the reason for the relationship or its nature.

There are other things to consider...



Interpretation	Example
• The link between two variables can be one of cause and effect: that is when with one of the variables has a direct effect on the other. In such cases, the correlation is perfect and the relation between the two variables is defined by a rule.	The correlation between altitude and temperature is perfect since the temperature varies in direct relation to altitude.
• The correlation between two variables can be significant without the two variables being directly linked to each other. They can both depend on a third variable which, as it varies, generates variations for the first two variables.	In the summer, it may seem that there is a strong correlation between the number of ice cream cones sold and the number of air conditioning units sold in a given city while in fact these two variables depend on another variable, is, the temperature.
• Considering a correlation as being linear while another model would be more appropriate.	The population growth of a major city can be studied according to a linear correlation. However, using an exponential model would be more appropriate.
• It sometimes may happen that there is a correlation between two variables only over a given interval.	Over the interval [5, 10] years, the correlation between a person's age and his or her height is linear. However, before and after this interval, the linear model is not the best fit.
• A two-variable distribution may include outlier data, notably due to manipulation or measurement errors.	The degree of precision of the instrument used during data collection is poor.