Community Ophthalmology

Basic Biostatistics for Non-statisticians - III

A. Padmavathi MSc., & A. Karthika MSc., Biostatistics Department, LAICO

Correlation and regression

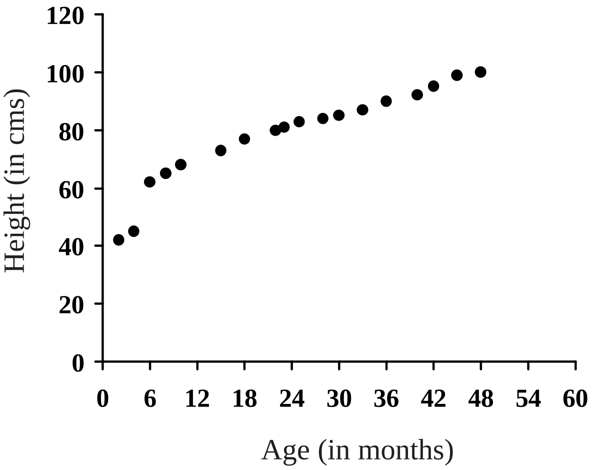
So far in the last two issues we saw how to represent all the individual values of a single variable in tabular, graphical and summarized forms. When we have more than one variable in hand, the natural tendency is to find out whether there exists any relationship between those variable.

Why do we want to find out such a relationship between those variables? The answer is that if there is some relationship between the variables then, one can predict the pattern of change between those variables. Hence the prediction of one variable with respect to another variable is possible. For example, if we happen to find out that there exists some relationship between age of persons and their corresponding blood pressure, we can identify the pattern of increase or decrease in blood pressure according to increase in age, so that one can estimate the blood pressure value for any given age.

Correlation is a concept which tells us whether such relationship exists between variables or not and if exists, is it a positive or negative relationship. Correlation is denoted by the symbol “r”, otherwise known as Correlation coefficient and its value always lies between -1 and +1. Correlation is said to be positive if one unit increase/decrease in the values of one variable causes a corresponding increase/decrease in the other variable, then the relationship between the variables is said to be positive or both the variables are **Positively correlated** and $0<r\leq1$. When r=1, the variables are said to be in a Perfect Positive correlation. On the other hand correlation is said to be negative if one unit increase in the values of one variable causes a corresponding decrease in the other variable and vice versa, then the relationship between the variables is said to be negative or both the variables are said to be **Negatively correlated** and $-1\leq r<0$. When r = -1, the relationship is said to be Perfectly Negatively correlated. Similarly if r=0, then it is known as **Zero correlation or null correlation.**

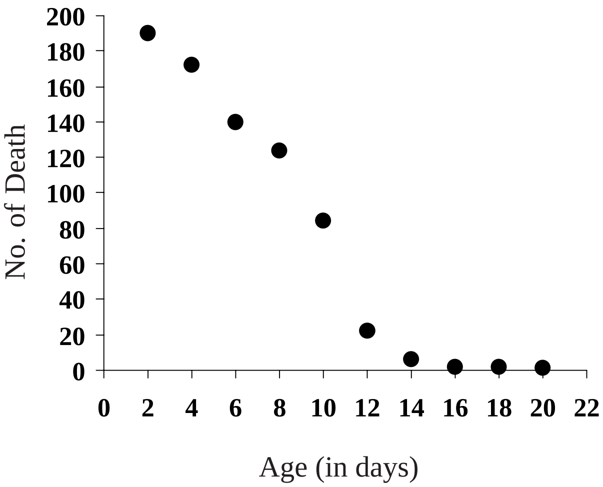
The initial step in finding out the relationship between two variables is the **Scatter plot**. When the units of variables are marked in X and Y-axis and their corresponding values are plotted then the graph that we get is a scatter plot. From the scatter plot one can say whether the two variables are correlated or not. Consider the following example.

Suppose if we plot the age of children (between 6th month - 5 yrs.) and their corresponding average heights, we get the below scatter plot,



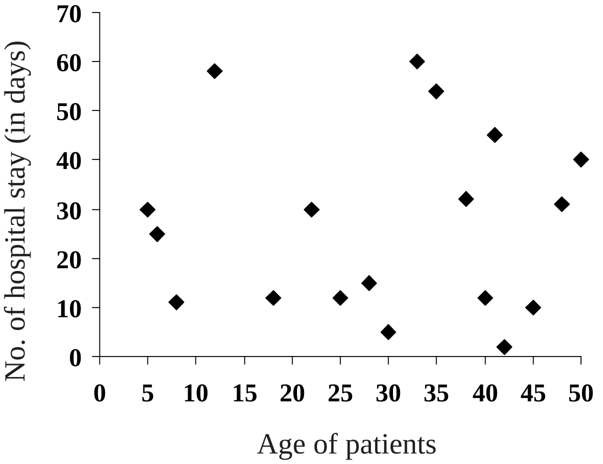
Graph - 1

From the plot, we can say that there exists a positive correlation between age and height. i.e., as the age increases, their corresponding height also increases. Now, consider the variables ‘age of children’ (between 1 day - 20 days) and the ‘number of deaths’ in a particular village during a particular period.



Graph - 2

Which shows a negative relationship. ie., as age of children increases, the number of deaths decreases. Now let us consider the following hypothetical situation, which is an example for zero correlation or null correlation. The 'age of patients' is taken in X-axis and the 'number of days stayed in hospital' is taken in Y-axis. Hence the scatter plot is,



Graph - 3

There seems no particular pattern in the above scatter diagram. The points are scattered everywhere. From the diagram we cannot say that the variables 'Age of patients' and 'Number of hospital stay are related. This is known as **'Zero Correlation'** or **'Null Correlation'.**

Scatter plot is the graphical method to assess the relationship, whereas to measure the degree of association between the two variables mathematically Karl Pearson coefficient of correlation as well as Spearman's rank correlation coefficient will help. These methods result in a value that lies between -1 and +1.

Another important assumption in correlation is that the relationship, if exists between two variables, is linear. Ie. If we plot the points of the variables in a graph, it should scatter around a straight line. Correlation coefficients calculated using mathematical formulae are valid only when the relationship is linear.

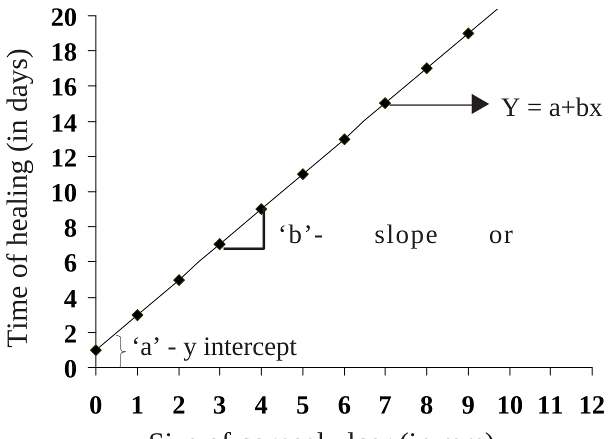
When the sample is very small, the correlation may sometimes be purely of chance. For example, the factors income and weight may have a very high correlation of r=0.93. And, as it is very obvious that increase in income will not have any direct impact on weight, this is a spurious correlation and this type of situation may occur due to chance. Prof. Yule's called it as "Non-sense Correlation". This is the major limitation of correlation.

Regression

In correlation, we try to measure the degree with which the variables vary together and we neither worry about the dependency or independency of the variables nor about the cause-effect relationship. Hence, the next step after confirming a relationship between the two variables is to identify the dependency of the variables and the cause-effect relationship. This identification of cause and effect variables helps us to predict one variable in terms of rate of change in another variable. Hence, regression coefficient helps us to predict the dependent variable in terms of independent variable.

For example, growth of crops and rainfall, rainfall and cholera, maternal health and childbirth weight etc are examples of regression.

Regression is calculated using a mathematical equation which is given as ***y=a+bx***. This is most commonly named as straight-line equation. Here ***'y'*** is the dependent variable and ***'x'*** is the independent variable. Based on the unit change in x, the value of y is calculated using this formula. Graphically, regression can be explained with the following example: In a study to find the whether the 'Duration of ulcer healing' depends upon the 'Size of ulcer'; we plot the values of 'Size of ulcer' (Independent variable) on the X-axis and the 'duration of healing' (Dependent variable) on the Y-axis.



Graph - 4

Here ***'a'*** is known as 'y intercept' (ie) the point at which it cuts the 'y-axis' when x=0 and ***'b'*** is the regression coefficient or in simple terms, it is known as 'slope of straight line'. (Standard textbooks can be referred for the formula of correlation and regression coefficients). Applying and extending this concept to find the relationship between three or more variables is known as 'Multiple Regression'. Here we assume the linear relationship between one dependent variable '*y*' and more than one independent variables, $x_1,x_2,x_3,......x_{n.}$

The major limitation of regression is that at times it is not very clear which variable acts as cause and which corresponds to effect even though they exists correlation between them. For example, though there exists correlation between 'Price and Demand', we cannot say, at some times, whether price is the cause for demand or demand is the cause for price. But anyhow regression gives us useful information about the average change of one variable to another and enables us for prediction of that dependent with respect to the average change in the independent variable.

Suggested readings

1. *S*undar Rao PSS, Richard J. *An Introduction to Biostatistics - A Manual for Students in Health Sciences* .Prentice-Hall of India, New Delhi, 1997.

2. Bernard Rosner. *Fundamentals of Biostatistics* .Duxbury Thomson Learning, USA, 2000.7

3. T D V Swinscow. *Statistics at Square one.* BMJ Publishing Group, London, 1997