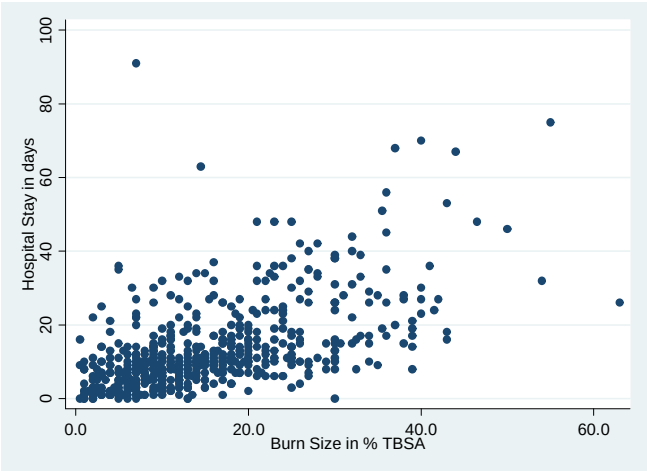
Correlation

We learned to examine the association between two categorical variables by chi-squared test. But what a bout two quantitative variables? Two quantitative variables may be related to each other i.e. one variable will vary according to the value of the other and this could be because one of them affects the other one. the values of one variable may become higher when the values of the other variable are higher or the values of one variable may decrease when the values of the other variable are higher. To examine this relationship between two quantitative variables we need to understand correlation. Correlation allows us to assess the degree to which two quantitative variables are associated with each other

Display of the relationship between two quantitative variables

A graph called a scatter plot (see figure below) is usually used to display the relationship between two quantitative variables. This is called a scatter plot because it is formed from a collection of points scattered around in the graph area. In a scatter plot, one variable is drawn on the x axis and the other on Y axis. The X axis variable is called the explanatory variable or the independent variable and the Y axis variable is called the response variable or the dependent variable. The value the response (dependent) variable is thought to depend or vary in response to the value of the explanatory (independent) variable. Each observation in a scatter plot represents on observation of the dataset.

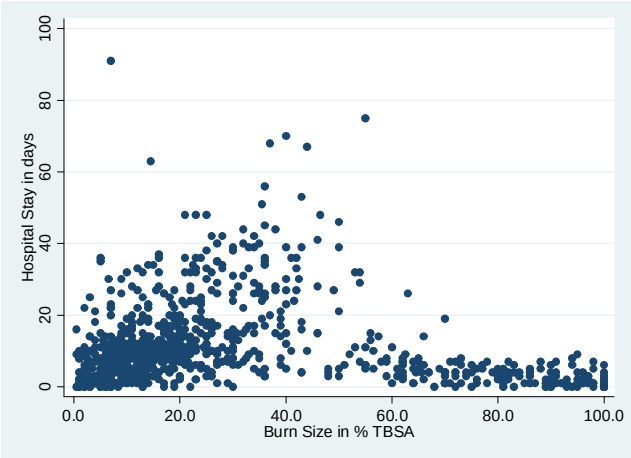


The scatter plot opposite display relationship between hospital stay in days (dependent variable) and burn size in %Total body surface area (independent variable) for patients admitted in 2008 to Sulaymaniyah Burns Centre and survived. Each dot on the graph is one patient. What does this scatterplot tell us?

The scatterplot displays hospital stay against burn size for patients admitted and survived in Sulaimaniyah 2008. It shows how the length of hospital stay varies according to the percentage of burn size. A quick look tells us that majority of the patients are in the inner corner of the graph area where Y and X coordinates start from zero. This indicates that more patients were in the group with lower burn size i.e.below 20% TBSA and with less hospital stay i.e. 20 days. There is one solitary patient with more than 80 days in hospital but with a burn size of about 10%. Such extreme values which are outside the main bulk of observations is called an outlier because it lies outside the main group.

Correlation

What else does the scatter plot tell us? If you look at the plot you see it is not scattered in the whole graph area, but it has a trend, a direction which rises slowly from the inner corner towards the outer border of the graph area. The spread of the dots (observations) in the graph area tells us the relationship between the tw variables. The shape of this distributions tells us that generally when the burn size increases the length of hospital stay also increases. This association is called correlation.



Visually by looking to scatter plots we can say whether the two variable are associated with each other or not (is there a correlation between them or not).
The second scatterplot is from the same dataset but also includes patients who have died in hospital. We can see there is a clear difference with the first scatter plot. The

distribution of the observations in this one is not uniform as the first scatter plot; the correlation is not as before. Here we can't say that hospital stay is more with greater burn size. We see a lot of dots with burn size over 60% but their hospital stay is short, near the zero marker. We can say from visual inspection of this scatter plot that the two variables are not correlated i.e. hospital stay is not associated with burn size.

To confirm our visual understanding of the correlation, we have to calculate and report a mathematical measure. This mathematical presentation of correlation is called Pearson's correlation coefficient.

Pearson's correlation coefficient

In order to quantify the association between two numerical variables, we calculate Pearson's correlation coefficient (r). The correlation coefficient which is calculated using the formula opposite, measures how the observations are spread (scattered) around an underlying linear pathway. You don't have to learn the formula of calculate r manually, it is calculated by statistical packages. The correlation coefficient can take any value from -1 to +1

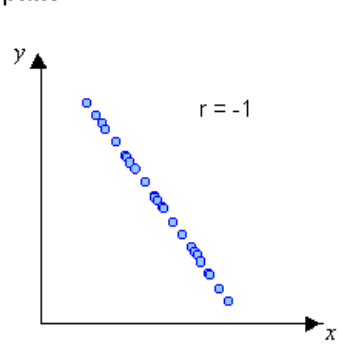
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

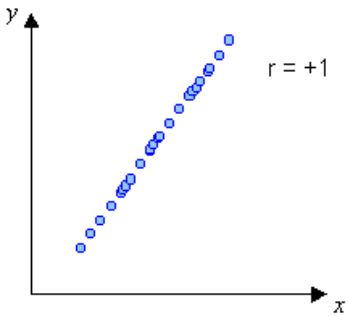
where,

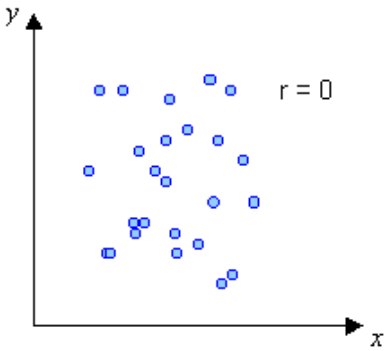
x_i, y_i are the individual observations

$\bar{x}, \bar{y}$ are the means for all observations of x and y .

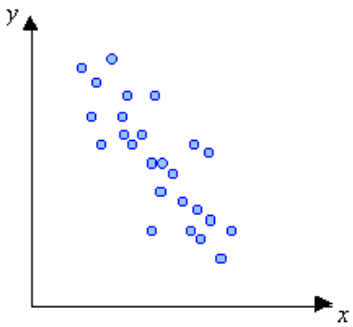
The correlation is positive if higher values of one variable are associated with higher values of the other variable. The correlation is negative if higher values of one variable is associated wth lower values of the other variable. The graphs below shoes correlation of 2 variables when r=1(perfect positive correlation) , r=-1 (perfect negative correlation), r=0 (no correlation). When r=1 or r=0 the observations will be on a straight line and when r=0 they will be scattered without a clear pattern.

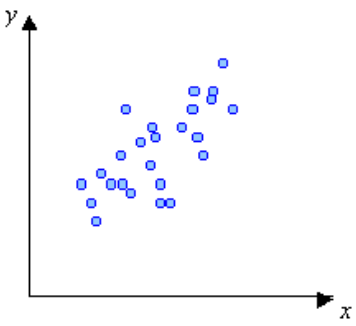






When r is more than (0.5) or less than (-0.5) the observations will not be on a straight line but still they will have a clear trend on pathway. For example in the figure below, r in the right plot is -0.8 (strong negative correlation) and in the left plot r is 0.8 (strong positive correlation).





Assumptions for use of correlation

Proper use of correlation coefficient requires tow conditions:

1- A Pearson's correlation coefficient may be calculated for any two numerical variables but it is more meaningful when the two variables have an approximately Normal distribution.

2- Al observations should be independent i.e. each observation in the data should be from a separate individual. In other words there should not be repeated observations from the same person.