**Pearson's Correlation Coefficient**

Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure. Pearson's correlation coefficient (r) is a **measure of the strength of the association** between the two variables.

In **Business**: - Correlation is another method of **sales forecasting**. Correlation looks at the **strength of a relationship**between two variables.

For marketing, it might be useful to know that there is a predictable relationship between sales and factors such as advertising, weather, consumer income etc.

Correlation is usually measured by using a scatter diagram, on which data points are plotted. For example, a data point might measure the number of customer enquiries that are generated per week (x-axis) against the amount spent on advertising (y-axis). This is illustrated below:

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

**Values of Pearson's correlation coefficient**

Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:

|  |  |  |
| --- | --- | --- |
| r = -1 | data lie on a perfect straight line with a negative slope | data lie on a perfect straight line with a negative slope |
| r = 0 | no linear relationship between the variables | no linear relationship between the variables |
| r = +1 | data lie on a perfect straight line with a positive slope | data lie on a perfect straight line with a positive slope |

Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa.

**Example Scatterplots**

Identify the approximate value of Pearson's correlation coefficient. There are 8 charts, and on choosing the correct answer, you will automatically move onto the next chart.

*(FLASH)*

Tip: that the square of the correlation coefficient indicates the proportion of variation of one variable 'explained' by the other (see Campbell & Machin, 1999 for more details).

Statistical significance of r

**Significance**

The t-test is used to establish if the correlation coefficient is significantly different from zero, and, hence that there is evidence of an association between the two variables. There is then the underlying assumption that the data is from a normal distribution sampled randomly. If this is not true, the conclusions may well be invalidated. If this is the case, then it is better to use Spearman's coefficient of rank correlation (for non-parametric variables). See Campbell & Machin (1999) appendix A12 for calculations and more discussion of this.

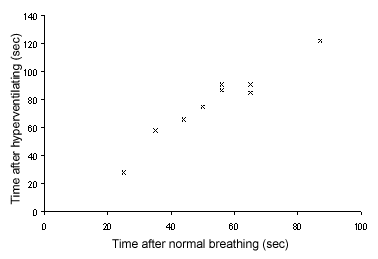
It is interesting to note that with larger samples, a low strength of correlation, for example r = 0.3, can be highly statistically significant (ie p < 0.01). However, is this an indication of a meaningful strength of association?

**NB Just because two variables are related, it does not necessarily mean that one directly causes the other!**

**Worked example:**

Nine students held their breath, once after breathing normally and relaxing for one minute, and once after hyperventilating for one minute. The table indicates how long (in sec) they were able to hold their breath. Is there an association between the two variables?

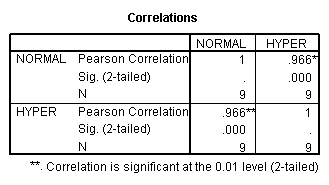
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Subject** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** |
| **Normal** | 56 | 56 | 65 | 65 | 50 | 25 | 87 | 44 | 35 |
| **Hypervent** | 87 | 91 | 85 | 91 | 75 | 28 | 122 | 66 | 58 |



The chart shows the scatter plot (drawn in MS Excel) of the data, indicating the reasonableness of assuming a linear association between the variables.

Hyperventilating times are considered to be the dependent variable, so are plotted on the vertical axis.

**Output from SPSS and Minitab are shown below:**

**SPSS  
Select Analysis>Correlation>Bi-variate**  


**Minitab  
Correlations: Normal, Hypervent**

Pearson correlation of Normal and Hypervent = 0.966  
P-Value = 0.000

**Conclusion**

In conclusion, the printouts indicate that the strength of association between the variables is very high (r = 0.966), and that the correlation coefficient is very highly significantly different from zero (P < 0.001). Also, we can say that 93% (0.9662) of the variation in hyperventilating times is explained by normal breathing times.