Correlations

- When we wnat to consider if there is any relationship between two variables we can look for corelation or association
- A correlation is a menasure of the linear relationship between variables. They could be
  - positively related
  - negatively related
  - not related at all
- Often a useful exploratory tool. But care needed; Spurious correlation -see Darrell Huffs book
- To compare two series we can describe their distribution about their mean. We would expect that if they are related they will vary from the mean in the same way for a partcular observation. As we have seen the measure of the deviation from the mean of a sample is the variance:

$$Variance = s^2 = \frac{\sum (X_j - \bar{X})^2}{N - 1}$$
$$= \frac{\sum (X_j - \bar{X})(X_j - \bar{X})}{N - 1}$$

- Note that *s* is the standard deviation (the square root of the variance)
- So one way of seeing if two variables is related is to see how they *covary.* Rather than squaring the difference we can multiply one differrence by the other

$$Covariance = cov(X, Y) = \frac{\sum (X_j - \bar{X})(Y_j - \bar{Y})}{N - 1}$$

- A posiive covariance will indicates that as one variable deviates from the mean in one direction so does the other.
- But this measure will depend upon the scale of the variables used -its just a value
- Need to standardise it so we get some idea of how strong the relation is and can compare such a relation between different variable pairs
- To do this we use the correlation coefficient *r* which looks at the covariance relative to their standard deviations

$$r = \frac{Cov(X, Y)}{s_X s_Y}$$
$$= \frac{\sum (X_j - \bar{X})(Y_j - \bar{Y})}{(N - 1) s_X s_Y}$$

- This is the Pearson product-moment correlation coeficient
- By standardising the covariance we have a value that has to lie between -1 and 1:
  - A correlation of +1 implies the variables are perfectly positively correlated
  - A correlation of -1 implies the variables are perfectly negatively correlated
  - A correlation of 0 implies the variables are not correlated
- NB this does not imply causality -association
- In SPSS can do correlations easily, but always worth doing a scatterplot first
- Problem with this measure is that as it looks at deviations from the mean it needs parametric data -need interval or ratio data
- If data are non parametric -when they are ordinal rather than cardinal or qualitaitive rather than quantitative- need to look at other methods provided by SPSS -rank correlation coefficients
  - Spearman's rho:
  - Kendall's tau
- Spearman's rho works by
  - first ranking the data and
  - then applying the Pearson equation for *r*

$$r = \frac{Cov(X, Y)}{s_X s_Y}$$

- Kendall's tau should be used when there is a small data set with a number of tied ranks