**Topic 27**

**Correlation Coefficient or “Measures of Association”**

The correlation coefficient, r, *measures the degree to which two variables are associated*.

Two more symbols to add, r, and rho (pronounced r),

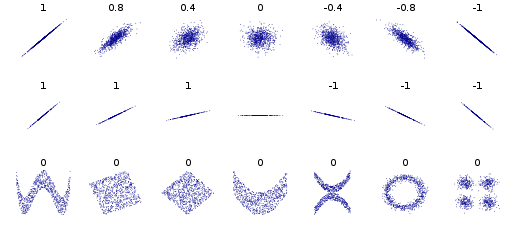
The **SAMPLE CORRELATION COEFICIENT** :

r = \frac{1}{n-1} \sum ^n _{i=1} \left( \frac{X_i - \bar{X}}{s_X} \right) \left( \frac{Y_i - \bar{Y}}{s_Y} \right)

**POPULATION CORRELATION COEFICIENT** :

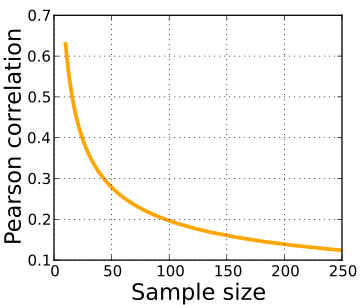
 \rho_{X,Y}={\mathrm{cov}(X,Y) \over \sigma_X \sigma_Y} ={E[(X-\mu_X)(Y-\mu_Y)] \over \sigma_X\sigma_Y} 

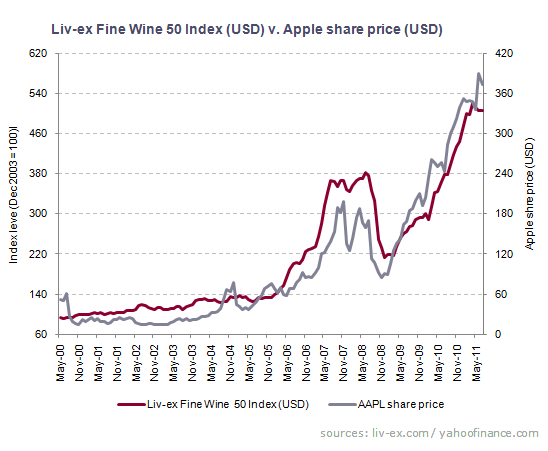
R is king. It is one of the most common and relentlessly pervasive statistics used.

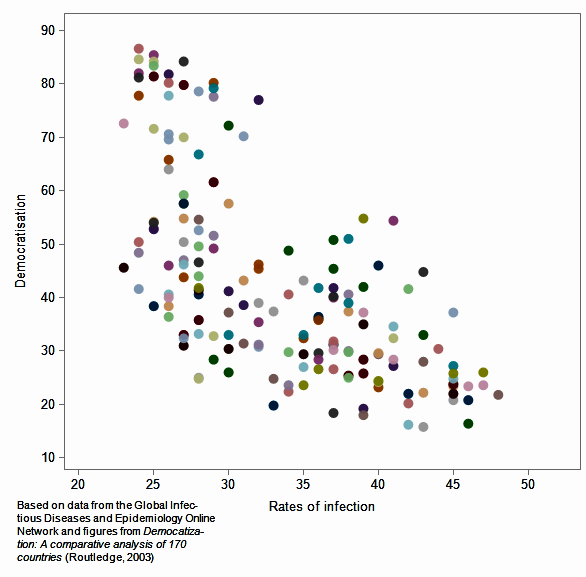


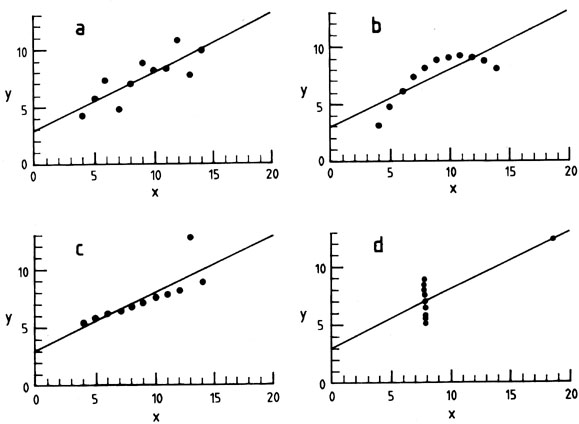
*However*

* R is weak
* R is only for linear relationships
* Big Data does not need a high r value to be significant
* Associations between unrelated variables are not uncommon
* Be aware of spurious associations (due to lurking variable)
* Conditions to use R are fairly restrictive
* R should be used secondary to visualizing data









Notes:

The restrictive conditions on correlation are the following:

* Normally distributed
* Interval

R2 , the [coefficient of determination](http://en.wikipedia.org/wiki/Coefficient_of_determination), estimates the fraction of the variance in *the response variable*  that is explained by *the explanatory variable*  in a linear regression

There are Non-parametric alternatives to the Spearman statistic, rho. Non-parametric means that the data can take a much broader class of shapes.

To confuse matters, one is called the Spearman rho (use for ordinal variables)

The other is the Kendall Tau

it is invariant in magnitude to separate changes in location and scale in the two variables