Correlation

Ph. Grosjean <phgrosjean@sciviews.org>

March 24, 2012

Part I
Introduction

Correlation is a form of **association** between two random variables or samples of these in statistics. Dependence is a synonym of correlation (???) Aspects to develop:

- ▷ Three degrees of association: correlation, relationship and causality.
- ▷ Correlation does not mean causality.
- ▷ There are several kinds of correlation coefficients, with basic hypotheses one should know.
- ▷ It is relatively easy to measure independence, but not dependence, because correlation coefficient measure only one given type of dependence (linear for Pearson's product-moment). Should derive something from the picture in Wikipedia that illustrates several cases).
- ▷ Correlation coefficient is invariant for the slope of relationship,
- ▷ Pearson's is not always defined (when there is no variation in X or in Y), same for Spearmann's and Kendall's (???) => but not useful in this case (!?).
- ▷ In `cor.test()`, the confidence interval on the correlation coefficient is only calculated for Pearson's and when there are at least 4 complete cases. Also look at resampling estimation and test???
- ▷ More robust correlation coefficients and less sensitive to nonlinear relationship.
- ▷ Pearson's is not restricted to bivariate Normal distribution (!), but correlation hypothesis test is (=> use rank correlation to create test and confidence interval in case you suspect a different distribution)!
- ▷ Show also problem with multimodal (multipopulation) data + solution = coef calculated per group.

- ▷ A diagram indicating independence versus dependence + various types of dependencies (any kind, monotonous, linear) and the corresponding correlation coefficient that applies.
- ▷ Useful simple examples to compare Pearson's and rank correlation in Wikipedia.
- ▷ Correlation coefficients should be independent from translations and scaling, i.e., $X' = aX + b$, and $Y' = cY + d$. Rank coefficients are insensitive to monotone transformation too.
- ▷ Unbiased? Asymptotically consistent?
- ▷ Problem of restricted range for one or both variables: example in Wikipedia + the case of a sinusoidal signal!
- ▷ Correlation matrices + properties + they are var/covar matrices of standardized (scaled) variables => demonstrate!
- ▷ Manual calculation + tricks to speed up calc.
- ▷ Nice example of correlation versus linearity in Anscombe's quartet (see Wikipedia).
- ▷ Introduce and illustrate partial correlation.
- ▷ See also **corpcor**, **mvoutlier**, **corrperm** (for permutation tests of correlation with repeated measurements using `cp.test()`) packages. Also, **pcalg** for robust estimation and causal inference.

1 Correlation in R

Correlation in R is `cor()` and `cor.test()` + `cov.wt(cor = TRUE)$cor` for weighted correlation matrix and `cov2cor()` to convert efficiently a covariance matrix into a correlation matrix. `cor.test()` is a generic function that provides both a default and a formula interface in the **stats** package.