POINTS OF SIGNIFICANCE

Association, correlation and causation

Correlation implies association, but not causation. Conversely, causation implies association, but not correlation.

Most studies include multiple response variables, and the dependencies among them are often of great interest. For example, we may wish to know whether the levels of mRNA and the matching protein vary together in a tissue, or whether increasing levels of one metabolite are associated with changed levels of another. This month we begin a series of columns about relationships between variables (or features of a system), beginning with how pairwise dependencies can be characterized using correlation.

Two variables are independent when the value of one gives no information about the value of the other. For variables X and Y, we can express independence by saying that the chance of measuring any one of the possible values of X is unaffected by the value of Y, and vice versa, or by using conditional probability, $P(X|Y) = P(X)$. For example, successive tosses of a coin are independent—for a fair coin, $P(H) = 0.5$ regardless of the outcome of the previous toss, because a toss does not alter the properties of the coin. In contrast, if a system is changed by observation, measurements may become associated or, equivalently, dependent. Cards drawn without replacement are not independent; when a red card is drawn, the probability of drawing a black card increases, because now there are fewer red cards.

Association should not be confused with causality; if X causes Y, then the two are associated (dependent). However, associations can arise between variables in the presence (i.e., X causes Y) and absence (i.e., they have a common cause) of a causal relationship, as we've seen in the context of Bayesian networks[1]. As an example, suppose we observe that people who daily drink more than 4 cups of coffee have a decreased chance of developing skin cancer. This does not necessarily mean that coffee confers resistance to cancer; one alternative explanation would be that people who drink a lot of coffee work indoors for long hours and thus have little exposure to the sun, a known risk. If this is the case, then the number of hours spent outdoors is a confounding variable—a cause common to both observations. In such a situation, a direct causal link cannot be inferred; the association merely suggests a hypothesis, such as a common cause, but does not offer proof. In addition, when many variables in complex systems are studied, spurious associations can arise. Thus, association does not imply causation.

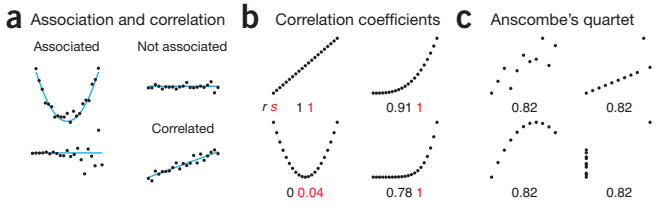

Figure 1 | Correlation is a type of association and measures increasing or decreasing trends quantified using correlation coefficients. (**a**) Scatter plots of associated (but not correlated), non-associated and correlated variables. In the lower association example, variance in y is increasing with x. (**b**) The Pearson correlation coefficient (r, black) measures linear trends, and the Spearman correlation coefficient (s, red) measures increasing or decreasing trends. (**c**) Very different data sets may have similar r values. Descriptors such as curvature or the presence of outliers can be more specific.

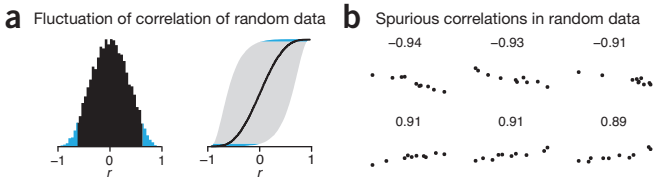

Figure 2 | Correlation coefficients fluctuate in random data, and spurious correlations can arise. (**a**) Distribution (left) and 95% confidence intervals (right) of correlation coefficients of 10,000 $n = 10$ samples of two independent normally distributed variables. Statistically significant coefficients ($\alpha = 0.05$) and corresponding intervals that do not include $r = 0$ are highlighted in blue. (**b**) Samples with the three largest and smallest correlation coefficients (statistically significant) from **a**.

In everyday language, dependence, association and correlation are used interchangeably. Technically, however, association is synonymous with dependence and is different from correlation (**Fig. 1a**). Association is a very general relationship: one variable provides information about another. Correlation is more specific: two variables are correlated when they display an increasing or decreasing trend. For example, in an increasing trend, observing that $X > \mu_X$ implies that it is more likely that $Y > \mu_Y$. Because not all associations are correlations, and because causality, as discussed above, can be connected only to association, we cannot equate correlation with causality in either direction.

For quantitative and ordinal data, there are two primary measures of correlation: Pearson's correlation (r), which measures linear trends, and Spearman's (rank) correlation (s), which measures increasing and decreasing trends that are not necessarily linear (**Fig. 1b**). Like other statistics, these have population values, usually referred to as ρ. There are other measures of association that are also referred to as correlation coefficients, but which might not measure trends.

When "correlated" is used unmodified, it generally refers to Pearson's correlation, given by $\rho(X, Y) = \mathrm{cov}(X, Y)/\sigma_X\sigma_Y$, where $\mathrm{cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y))$. The correlation computed from the sample is denoted by r. Both variables must be on an interval or ratio scale; r cannot be interpreted if either variable is ordinal. For a linear trend, $|r| = 1$ in the absence of noise and decreases with noise, but it is also possible that $|r| < 1$ for perfectly associated nonlinear trends (**Fig. 1b**). In addition, data sets with very different associations may have the same correlation (**Fig. 1c**). Thus, a scatter plot should be used to interpret r. If either variable is shifted or scaled, r does not change and $r(X, Y) = r(aX + b, Y)$. However, r is sensitive to nonlinear monotone (increasing or decreasing) transformation. For example, when applying log transformation, $r(X, Y) \neq r(X, \log(Y))$. It is also sensitive to the range of X or Y values and can decrease as values are sampled from a smaller range.

If an increasing or decreasing but nonlinear relationship is suspected, Spearman's correlation is more appropriate. It is a nonparametric method that converts the data to ranks and then applies the formula for the Pearson correlation. It can be used when X is ordinal and is more robust to outliers. It is also not sensitive to monotone

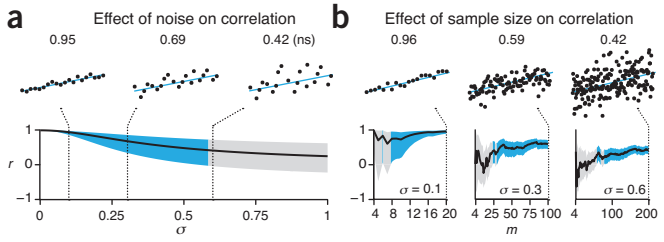

Figure 3 | Effect of noise and sample size on Pearson's correlation coefficient *r*. (**a**) *r* of an $n = 20$ sample of $(X, X + \varepsilon)$, where ε is the normally distributed noise scaled to standard deviation σ. The amount of scatter and value of *r* at three values of σ are shown. The shaded area is the 95% confidence interval. Intervals that do not include $r = 0$ are highlighted in blue ($\sigma < 0.58$), and those that do are highlighted in gray and correspond to nonsignificant *r* values (ns; e.g., $r = 0.42$ with $P = 0.063$). (**b**) As sample size increases, *r* becomes less variable, and the estimate of the population correlation improves. Shown are samples with increasing size and noise: $n = 20$ ($\sigma = 0.1$), $n = 100$ ($\sigma = 0.3$) and $n = 200$ ($\sigma = 0.6$). Traces at the bottom show *r* calculated from a subsample, created from the first *m* values of each sample.

increasing transformations because they preserve ranks—for example, $s(X, Y) = s(X, \log(Y))$. For both coefficients, a smaller magnitude corresponds to increasing scatter or a non-monotonic relationship.

It is possible to see large correlation coefficients even for random data (**Fig. 2a**). Thus, *r* should be reported together with a *P* value, which measures the degree to which the data are consistent with the null hypothesis that there is no trend in the population. For Pearson's *r*, to calculate the *P* value we use the test statistic $\sqrt{[\text{d.f.} \times r^2/(1 - r^2)]}$, which is *t*-distributed with d.f. $= n - 2$ when (X, Y) has a bivariate normal distribution (*P* for *s* does not require normality) and the population correlation is 0. Even more informative is a 95% confidence interval, often calculated using the bootstrap method[2]. In **Figure 2a** we see that values up to $|r| < 0.63$ are not statistically significant—their confidence intervals span zero. More important, there are very large correlations that are statistically significant (**Fig. 2a**) even though they are drawn from a population in which the true correlation is $\rho = 0$. These spurious cases (**Fig. 2b**) should be expected any time a large number of correlations is calculated—for example, a study with only 140 genes yields 9,730 correlations. Conversely, modest correlations between a few variables, known to be noisy, could be biologically interesting.

Because *P* depends on both *r* and the sample size, it should never be used as a measure of the strength of the association. It is possible for a smaller *r*, whose magnitude can be interpreted as the estimated effect size, to be associated with a smaller *P* merely because of a large sample size[3]. Statistical significance of a correlation coefficient does not imply substantive and biologically relevant significance.

The value of both coefficients will fluctuate with different samples, as seen in **Figure 2**, as well as with the amount of noise and/or the sample size. With enough noise, the correlation coefficient can cease to be informative about any underlying trend. **Figure 3a** shows a perfectly correlated relationship (X, X) where *X* is a set of $n = 20$ points uniformly distributed in the range [0, 1] in the presence of different amounts of normally distributed noise with a standard deviation σ. As σ increases from 0.1 to 0.3 to 0.6, $r(X, X + \sigma)$ decreases from 0.95 to 0.69 to 0.42. At $\sigma = 0.6$ the noise is high enough that $r = 0.42$ ($P = 0.063$) is not statistically significant—its confidence interval includes $\rho = 0$.

When the linear trend is masked by noise, larger samples are needed to confidently measure the correlation. **Figure 3b** shows how the correlation coefficient varies for subsamples of size *m* drawn from samples at different noise levels: $m = 4$–20 ($\sigma = 0.1$), $m = 4$–100 ($\sigma = 0.3$) and $m = 4$–200 ($\sigma = 0.6$). When $\sigma = 0.1$, the correlation coefficient converges to 0.96 once $m > 12$. However, when noise is high, not only is the value of *r* lower for the full sample (e.g., $r = 0.59$ for $\sigma = 0.3$), but larger subsamples are needed to robustly estimate ρ.

The Pearson correlation coefficient can also be used to quantify how much fluctuation in one variable can be explained by its correlation with another variable. A previous discussion about analysis of variance[4] showed that the effect of a factor on the response variable can be described as explaining the variation in the response; the response varied, and once the factor was accounted for, the variation decreased. The squared Pearson correlation coefficient r^2 has a similar role: it is the proportion of variation in *Y* explained by *X* (and vice versa). For example, $r = 0.05$ means that only 0.25% of the variance of *Y* is explained by *X* (and vice versa), and $r = 0.9$ means that 81% of the variance of *Y* is explained by *X*. This interpretation is helpful in assessments of the biological importance of the magnitude of *r* when it is statistically significant.

Besides the correlation among features, we may also talk about the correlation among the items we are measuring. This is also expressed as the proportion of variance explained. In particular, if the units are clustered, then the intraclass correlation (which should be thought of as a squared correlation) is the percent variance explained by the clusters and given by $\sigma_b^2/(\sigma_b^2 + \sigma_w^2)$, where σ_b^2 is the between-cluster variation and $\sigma_b^2 + \sigma_w^2$ is the total between- and within-cluster variation. This formula was discussed previously in an examination of the percentage of total variance explained by biological variation[5] where the clusters are the technical replicates for the same biological replicate. As with the correlation between features, the higher the intraclass correlation, the less scatter in the data—this time measured not from the trend curve but from the cluster centers.

Association is the same as dependence and may be due to direct or indirect causation. Correlation implies specific types of association such as monotone trends or clustering, but not causation. For example, when the number of features is large compared with the sample size, large but spurious correlations frequently occur. Conversely, when there are a large number of observations, small and substantively unimportant correlations may be statistically significant.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Naomi Altman & Martin Krzywinski

1. Puga, J.L., Krzywinski, M. & Altman, N. *Nat. Methods* **12**, 799–800 (2015).
2. Kulesa, A., Krzywinski, M., Blainey, P. & Altman, N. *Nat. Methods* **12**, 477–478 (2015).
3. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 1139–1140 (2013).
4. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 699–700 (2014).
5. Altman, N. & Krzywinski, M. *Nat. Methods* **12**, 5–6 (2015).

Naomi Altman is a Professor of Statistics at The Pennsylvania State University. Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre.