Population statistics cannot be used for reliable individual prediction

Richard Kennaway
School of Information Systems
University of East Anglia, Norwich NR4 7TJ, U.K.

17 April 1998

Abstract

It is known that predictions about individuals from statistical data about the population are in general unreliable. However, the size of the problem is not always realised. For a number of ways of predicting information about one variable from another with which it is correlated, we compute the reliability of such predictions.

For the bivariate normal distribution, we demonstrate that unless the correlation is at least 0.99, not even the sign of a variable can be predicted with 95% reliability in an individual case. The other prediction methods we consider do no better. We do not expect our results to be substantially different for other distributions or statistical analyses.

Correlations as high as 0.99 are almost unheard of in areas where correlations are routinely calculated. Where reliable prediction of one variable from another is required, measurement of correlations is irrelevant, except to show when it cannot be done.

Corresponding author and address: Richard Kennaway, School of Information Systems, University of East Anglia, Norwich NR4 7TJ, U.K.

Email: jrk@cmp.uea.ac.uk.

The support of the School of Information Systems, University of East Anglia, is acknowledged.

An empirical study of correlations reported in the sociological literature (McPhail 1971) found that only 1% of the correlations reported in the papers studied were over 0.4, and only two out of 281 correlations exceeded 0.6. In that area, a correlation of 0.8 is generally considered high, and a correlation of 0.2 is publishable as demonstrating a connection between two variables.

We consider the question of what such correlations imply for the task of reliably and/or accurately predicting the value of one variable from the other. We demonstrate that it is impossible to reliably estimate even the sign of the variable relative to its mean unless the correlation is at least 0.99. For lesser correlations, such a prediction will do better than chance on average, and for some purposes, this is all that is required. However, such correlations are useless for making reliable predictions in individual cases. Correlations of this level are virtually unheard of in almost every discipline where statistical methods are commonly employed.

Despite this, a frequent use of statistical trends is to make predictions about individuals. Aptitude tests and credit rating are two major applications, especially in the latter case if ratings are derived from rules generated from statistical analysis by data mining applications. An individual to whom such tests are applied is, in effect, participating in a lottery. If the test is valid, the lottery is biased to a greater or lesser extent in his favour, but it is a lottery nonetheless. Such tests say little about any individual being tested. Just how little, it is the purpose of this paper to demonstrate.

We are not suggesting an alternative to such assessment methods, or suggesting that they not be used. We are demonstrating what they can and cannot do.

We use the bivariate normal distribution as a case study, consider several different methods of gaining information about one variable from the other, and in each case calculate the reliability of the predictions given the correlation coefficient. We briefly consider the problems of multivariate analysis and estimation of unknown correlations.

We do not expect our conclusions to differ substantially for other distributions or statistical analyses. Only standard mathematical statistical theory is required. See (Kendall, Stuart, and Ord 1987) as a general reference.

The bivariate normal distribution

We recall for reference some basic properties of the bivariate normal distribution. This distribution over two random but correlated real variables X and Y whose means are both 0 has the probability density function:

$$P(x, y, a, b, c) \quad = \quad \frac{1}{2abc'\pi} \exp(-\frac{1}{2}(\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{2cxy}{ab}))$$

where $-1 < c < 1$ and $c' = 1/\sqrt{1-c^2}$. a is the standard deviation of X, given that $Y = y$. (This is independent of y.) We shall call this the *conditional standard deviation* of X. b is similarly the standard deviation of Y, given a fixed value $X = x$. c is the product-moment correlation coefficient of X and Y.

The conditional distribution of X, given that the value of Y is y, is a normal distribution with mean acy/b and standard deviation a.

If Y is unknown, the distribution of X is a normal distribution with mean 0 and standard deviation ac'. We term ac' the *unconditional standard deviation* of X. Similarly, bc' is the unconditional standard deviation of Y.

To visualise $P(x, y, a, b, c)$, it is helpful to look at its contour lines. These are the family of ellipses $\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{2cxy}{ab} =$ constant. Since a and b are just scaling factors, we take them to be 1. The axes of the ellipses are the lines $x = \pm y$, and they intersect the coordinate axes at $\pm k$ for any k. When $c = 0$ they are circles, and for positive c they are elongated along the line $x = y$. The ratio of the axes is $\sqrt{\frac{1+c}{1-c}}$. Figure 1 illustrates this for selected values of c. This gives an immediate feel for how useful a correlation is for estimating Y, given the value of X. A scatterplot of data drawn from these distributions will have roughly the shape of the ellipses. Figure 2 shows such plots for 100 random points chosen from the same distributions.

A correlation of 0.99 is about the point at which the data begin to resemble a straight line rather than a cloud. The slanting line in each figure is the regression line of Y against X. It passes through the origin and the points where the ellipse has a vertical tangent. Its gradient is c. This line represents the maximum likelihood estimate of Y, given X. Notice that it differs from the maximum likelihood estimate of X, given Y, which is a line through the origin of gradient $1/c$, meeting the ellipse at its horizontal tangents.

We will also use the probability distribution function of the (univariate) normal distribution, $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} dx$. This is the distribution function for mean 0 and standard deviation 1; for the general case of mean μ and standard deviation σ we define $\Phi_{\mu,\sigma}(x) = \Phi((x - \mu)/\sigma)$.

As the values we require of $\Phi(x)$ sometimes go far beyond the range of standard statistical tables, we have calculated them ourselves from formulae 7.1.26 (for small $|x|$) and 7.1.13 (for large $|x|$) from (Abramowitz and Stegun 1964), for the related error function $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2} dx$.

Analysis of variance

The function f for which the line $y = f(x)$ best fits the probability distribution, in the sense of minimising the expected squared error, is the straight line $y = cx$ which was plotted in Figure 1. The proportion of the variation of Y which is "explained" by this estimate (in the technical sense which the theory of analysis of variance assigns to this word) is c^2. The remaining, "unexplained" proportion is $1 - c^2$. The latter quantity is called the *coefficient of alienation* or *dissociation*, or more colloquially the *coefficient of uselessness*. The second and third columns of Table 1 tabulate these quantities for various values of c.

Note that correlation does not imply causality. It does not follow merely from an analysis of variance of this sort that either variable has any causal influence on the other.

Improvement ratios

A further feeling for what the correlation coefficient means can be gained by considering the ratio of the standard deviation of Y to the standard deviation of Y given X, when the correlation is c. This ratio is c' (independent of X, a, and b). Call this the *improvement* of our knowledge of Y. This is a real number greater than or equal to 1. 1 means no improvement, 10 means that knowing X reduces the standard deviation of Y by a factor of 10, etc. The fourth column of Table 1 shows the improvement ratio for various correlations.

Mutual information

For two jointly distributed random variables X and Y, the mutual information between X and Y is the maximum amount of information about either variable which can be obtained from knowing the exact value of the other. It is denoted by $I_{X|Y}$ and is measured in bits. For a joint normal distribution with correlation c, $I_{X|Y} = \lg c'$. (lg is the binary logarithm.) This is tabulated in the fifth column of Table 1.

It is illuminating to consider some examples. Suppose we require that knowing the value of Y should give 1 bit of information about X. Then $\lg c'$ must be 1, which implies $c = 0.866$. Suppose we require to predict Y to within 1 part in N. Then $I_{X|Y} = \lg N$, and $c = \sqrt{1 - 1/N^2}$. If N is at least 5, then this is well estimated by $1 - 1/2N^2$.

Taking $N = 10$ — one decimal digit of accuracy — gives $c = 0.995$. To obtain two decimal digits accuracy requires $c = 0.99995$. Clearly, when obtaining results of such accuracy, one does not trouble to measure correlation coefficients at all. The variables being measured are obviously correlated, and one is more likely to concern oneself with the precise form of their relationship and the mechanism which produces it.

Now consider correlations such as are commonly reported in statistical studies. For example, it is claimed (see, for example, (Herrnstein and Murray 1994) and the references therein) that between 40% and 80% of variation in intelligence is explained by variation in genes. We are not here concerned with the validity of the studies which have produced this figure, but only with what it implies for the problem of predicting one variable from the other in an individual case. If $c = \sqrt{0.8} = 0.89$, then $c' = 2.236$, and the mutual information is $\lg 2.236 = 1.16$. For $c = \sqrt{0.4}$, the mutual information is 0.35. In other words, knowing everything about someone's genes gives somewhere between one-third of a bit and just over one bit of information about their IQ.

(Herrnstein and Murray 1994) also makes this general point, that the many statistical correlations they discuss (typically no higher than the $c = 0.4$ level) do not allow drawing conclusions about individuals. However, every time a statistically validated test is used to make a decision about an individual (which is the main purpose of most such tests), either one is drawing such a conclusion, or one is indifferent to the accuracy of the decision with regard to the specific individual, just as a casino is indifferent to the outcome of any particular bet. This issue arises not only for the correlation between the thing being tested for and the thing being predicted, but also for the correlation between the test score and the entity the test is intended to measure. We have no alternative to suggest; we are simply pointing out the scale of the problem.

Far lower correlations than 0.89 are often published. The previously cited survey (McPhail 1971) found that only 1% of the correlations reported were over 0.4, corresponding to a mutual information of 0.125. For practical purposes, this is no information at all in the individual case. Yet because a sufficient quantity of data has been amassed to be statistically sure at a high confidence level that the correlation differs from 0, such a result can be published and claimed to demonstrate a connection between two variables, when in practice it demonstrates a lack of connection.

Practical predictions

The mutual information is the maximum amount of information that could possibly be obtained about Y by knowing X. It does not give any method for actually extracting that much information, and in general it may not be possible. We consider in this section three practical methods of attempting to estimate Y from X. For simplicity, we take the means of X and Y to be zero, and the conditional standard deviations a and b to be 1. We also assume that c is non-negative.

Sign estimation

Suppose we attempt to predict just one bit of information about Y, given X, one yes-no decision: whether Y is positive. If the correlation is positive then the best we can do is to estimate the sign of Y by that of X. What proportion of correct predictions will we make overall? The formula for this is $\frac{1}{\pi}\cos^{-1}(-c)$, tabulated in the second column of Table 2.

At a correlation of 0.5, classification is correct 2/3 of the time, compared with the 50% that would be obtained by blind guessing. There is no way to tell which 1/3 of the decisions are the incorrect ones. 0.8 correlation gives four out of five correct, not good odds when the decision is of any importance to the individual. If the individual demands that the same standard of confidence apply to his case as are commonly used in the detection of statistical trends, i.e. 95% or 99% confidence of receiving a correct decision, then the required correlation is 0.988 or 0.9995 respectively.

The task of sign estimation assumes that one knows the means of X and Y. We can devise a similar task which does not depend on this information. Given two random individuals, we can estimate which has the higher value of y by seeing which has the higher value of x. The proportion of correct decisions turns out to be the same as the probability that for a single individual, y has the same sign as x, which we have just computed.

Screening tests

Although, as we mentioned above, we cannot tell which of the predictions of Y from X are the correct ones, we can be more confident of those where the magnitude of X is large. We therefore pose the questions: How large must the magnitude of X be, for the estimate of the sign of Y to be correct at least $1-\epsilon$ of the time? And for what proportion of the population does X have such a magnitude? To the first question, the answer is that $|x|$ must be at least $\delta = \frac{1}{c}\Phi^{-1}(1-\epsilon)$. To the second, the answer is $S(c,\epsilon) = 2(1-\Phi(\delta/c'))$. S is tabulated for various values of c and for $\epsilon = 5\%$ and 1% in the third and fourth columns of Table 2.

A correlation of 0.2 is clearly useless. To give concrete meaning to just how useless it is, suppose that such a test were being applied to the entire human population of the world (currently about 6×10^9 people). There is only about one chance in 200,000 that *anyone* would be reliably classified at the 5% level. At the 1% level, the probability of a given individual being reliably classified is 1 in 2.3×10^{29}. The latter number is the number of atoms in about 2.5 tons of water. (These extreme figures must be taken with some common sense. In practice, no real distribution can even be observed at such a large distance from the mean, let alone measured. Nevertheless, for any distribution with rapidly decreasing tails, the qualitative conclusion holds.)

At a correlation of 0.5, fewer than 5 in 1000 of the population are reliably classified. This is still useless. A correlation of 0.9 reliably classifies less than half the population at the 5% confidence level. Only at a correlation of 0.99 does it begin to be useful — four fifths of the population are reliably classified at the 5% confidence level, and nearly three quarters at the 1% confidence level.

If an individual requires a 95% chance of receiving a prediction that has a 95% chance of being correct, the correlation must be over 0.99995, and for a 99% chance of receiving a prediction that is 99% likely to be accurate, the required correlation will be in practice unmeasurably close to 1.

If Y is impossible to measure at the time when a prediction of its value is required, then with a correlation of less than 0.99 one must choose between making many unreliable predictions, refusing to make predictions in many cases, or finding something more useful to measure. If Y can be measured, then we can predict it from X when $|X| > \delta$, and measure Y in the remaining cases. This is useful when Y is more difficult or expensive to measure than X. That is, the measurement of X is used as a screening test. Whether this is worth doing depends on the relative cost of the two measurements and the proportion of the population which is reliably classified by the measurement of X. If a direct measurement of Y costs K times as much as a measurement of X, then using the screening test will give a relative saving of $S(c, \epsilon) - 1/K$ over the cost of measuring Y in every case. For this to be an improvement requires $K > 1/S(c, \epsilon)$. The maximum possible cost saving, in the limit where testing X is free, is S. At a correlation of 0.9, the cost ratio is 2.3, and the cost saving can be at most 42.6%.

At a correlation of 0.99, four fifths of the population can be given just the simple test at the 5% confidence level, and nearly three quarters at the 1% confidence level. The break-even point for costs is a ratio of 1.2 or 1.33 respectively. Nevertheless, it must not be forgotten that if a screening test is only accurate at the 1% confidence level, then about 1% of those it is given to *will* be misclassified. What proportion of misclassifications is acceptable depends on the purpose of the test and on the context in which it is being used.

As for the case of sign estimation, we can devise a version of the screening test which does not require knowing the mean of X. Given a correlation c, and two random samples (x, y) and (x', y'), suppose we predict that $y' > y$ if $x' - x$ is at least a certain amount δ. How large should δ be to ensure that when we make this prediction, we can have at least a certain confidence in it, and given the confidence we require, what proportion of pairs of random samples allow such a prediction to be made? As for simple sign estimation, this problem reduces by a change of variables to the screening problem we have just studied, and the same figures apply.

Decile estimation

Just as a correlation sufficient to give one bit of mutual information is not enough to perform reliable sign estimation, so a correlation giving one decimal digit's worth of information is not enough to reliably estimate the magnitude of the dependent variable with that accuracy.

If we use bcx/a to estimate y, we can ask the question, how likely is it for a random data point (x, y), that bcx/a and y differ by no more than half a decile either way of the unconditional distribution of Y? The probability is $\int_{x=-\infty}^{x=\infty} \Phi_{cx,1}(y_1) - \Phi_{cx,1}(y_0)$ where $y_0 = \Phi_{0,c'}^{-1}(\Phi_{0,c'}(cx) - \alpha/2)$, $y_1 = \Phi_{0,c'}^{-1}(\Phi_{0,c'}(cx) + \alpha/2)$, and $\alpha = 0.1$. This is tabulated in Table 3. For an estimate to this accuracy to be wrong in no more than 5% of cases requires a correlation of at least 0.997.

Accuracy of the sample correlation coefficient

So far, we have considered the correlation coefficient of the population to be given. In practice, it is estimated from a sample, and there will be some uncertainty in the estimate. The smaller the sample, the larger the uncertainty.

To make this estimate, it is not sufficient to merely establish that the sample correlation and the sample size imply that the population correlation is very likely to be positive. As we have seen, the mere knowledge that a correlation is positive does not imply any useful relationship between the variables. We must obtain a numerical estimate of the minimum likely value of the population correlation. (Usually, one will not be concerned to place an upper bound on the correlation — the higher the correlation, the better.)

To do this, we need to know the distribution of the sample correlation, given the population correlation and the sample size. For the bivariate normal distribution, this may be accurately estimated via Fisher's z-transformation ((Kendall, Stuart, and Ord 1987, §16.33), (Wilks 1962, p. 594)). For a sufficiently large sample size n, a sample correlation c, and a population correlation ρ, let $z = \frac{1}{2} \log \frac{1+c}{1-c}$ and $\zeta = \frac{1}{2} \log \frac{1+\rho}{1-\rho}$, where log is the natural logarithm. Then $(z - \zeta)\sqrt{n-1}$ is very close to being normally distributed with mean 0 and standard deviation 1.

From this result we can calculate a confidence bound for an estimate of population correlation. If we observe a positive sample correlation c in a sample of size n, for what value ρ can we conclude with confidence that the population correlation is at least ρ? Table 4 tabulates this for sample sizes of 100 and 20 and confidence levels of 95% and 99%.

To be sure at the 5% confidence level that the population correlation is at least, say, 0.866 (giving exactly one bit of mutual information between the variables), the correlation observed for a sample of 100 must be at least 0.9. If the sample size is only 20, one needs to observe a sample correlation of at least 0.935. At the 1% confidence level the sample correlation needs to be even higher: 0.914 and 0.952 respectively.

At lower correlations the difference between sample correlation and minimum likely population correlation is even larger. If one is interested in knowing with 95% confidence that the population correlation is at least 0.2, a sample of 20 must display a correlation of at least 0.52 to allow this inference to be drawn.

The relation between population, sample correlation, and confidence level can be looked at in another way. If one desires to detect at confidence level ϵ that the population correlation is at least c, then if the population correlation is actually only c, the probability that one's test will detect it is only ϵ.

Relationships for populations and individuals

Suppose that the bivariate data arise from taking some number of individuals, and obtaining from each individual some number of pairs (x, y). The set of all the data will have a certain correlation c between x and y. The set of data from the ith individual will have a correlation c_i. What relation may hold between c and the individual c_i? What relation may hold between the regression line for the whole data and the regression lines for individuals?

No relation need hold at all. To visualise why this is, imagine the scatterplot of the whole set of data. If c is positive, this will have the general shape of an oval in the xy plane whose long axis has positive gradient, as in Figure 1. Each individual's data will consist of some subset of those points. Clearly, it is possible to cover the oval with smaller ovals whose eccentricities and long axis directions bear no relationship to each other nor to those properties of the whole oval.

This has nothing to do with the particular distribution, or with c as the statistic; clearly, the same situation obtains for any statistic calculated from any population distribution other than highly degenerate ones, such as a population of identical data points.

The moral of this is that no argument can be made from a relationship between variables shown by a population, and the relationship between them for any individual. The population relationship is a property only of the population, and not of any individual in it; excluding degenerate cases, *any* relationship between the population variables is consistent with *any* individual relationship, or *any* combination of individual relationships. See (Powers 1990) for a simulated example in which the correlation between an independent variable and a dependent variable had, for every individual in a population, a sign opposite to that of the correlation over the whole population.

If enough data points are taken for each individual to estimate the relationship between X and Y for an individual, then the group statistics are irrelevant to that task. If not enough data points have been taken, the task is impossible, and the group statistics are still irrelevant. In many situations, only one data point (x, y) is measured per individual. This is the case, for example, in most data collected from surveys. It is also inherent in the nature of some experiments, especially in the area of learning. It is not possible to predict from such a data set what y would have been for a given individual if x had been different. However, it is easier to overlook the error, since an actual distribution of data from that individual is not available as a standard to compare with the prediction.

Predictions for non-random subpopulations

The preceding section assumes that the quantity of data from a single individual is a small proportion of the whole data set. For larger non-random subpopulations, it is possible to say something about the relationship between population and subpopulation statistics, but not very much. For bivariate normal correlations we shall ask: given a population correlation of c, and a subpopulation containing a proportion κ of the whole, what lower bound can be placed on the subpopulation correlation c_1? We give a simple analysis based on the pictures of Figure 1, which will place an upper bound on this lower bound. That is, we shall exhibit a particular subpopulation designed to have a low c_1. In general it will not be the lowest possible.

If we inscribe the largest possible circle inside an elliptical contour, this represents a contour for a subpopulation with a bivariate normal distribution with correlation zero. The relative size of that subpopulation is the ratio of the areas of the circle and the ellipse, which is the ratio of the minor and major axes of the ellipses. This is $\sqrt{\frac{1-c}{1+c}}$. More generally, consider a population with a correlation of c, and a subpopulation with the same means, and a correlation of c_1. Given c and $c_1 < c$, how large can the subpopulation be, assuming both are bivariately normally distributed? This is the same as asking for the largest possible relative area of a contour ellipse for correlation c_1 contained in one having correlation c, and is $\kappa = \sqrt{\frac{(1-c)(1+c_1)}{(1+c)(1-c_1)}}$. If c and κ are given, this formula gives an upper bound on the minimum possible correlation c_1 of a subpopulation of relative size κ in a population of size c. This is tabulated in Table 5 for several values of c and a range of subpopulation sizes.

For a subpopulation of half the total population, if the population correlation is 0.95, the subpopulation correlation could be as low as 0.81. For c as low as 0.5, the subpopulation correlation could be negative. If the subpopulation is not constrained to be bivariate normal, lower correlations are possible. Finiteness of the population and subpopulation also widen the bounds of possible subpopulation correlations — barring degenerate cases, a subpopulation of two will have a correlation of ± 1.

Note that a non-random subpopulation can have a distribution bearing no resemblance to the parent distribution. Depending on the source of the data, it may be possible to justify an assumption that a certain subpopulation has a distribution resembling that of the total population, but its distribution cannot be inferred from the total distribution.

If the population has an arbitrary distribution, then one can construct examples where the population has correlation c and a subpopulation which omits just one point has a correlation c_1, for any values of c and c_1 with $-1 < c < 1$ and $-1 \leq c_1 \leq 1$.

Multivariate analysis

We shall only touch on the complexities of multivariate analysis. If one's goal is to predict a variable X_n from the values of $n-1$ variables $X_1, \ldots, X_{n-1}$, then the difficulties we have discussed in the bivariate case are magnified. If the correlations between each of $X_{1\ldots n-1}$ and X_n are small, then correlations among $X_{1\ldots n-1}$ can easily defeat an attempt to explain all of the variation in X_n.

It is possible to synthesize artificial problems in which, for example, four variables X_1, X_2, X_3, and X_4, each correlating 0.5 with a variable X_5, together account perfectly for all the variation in X_5. (This will be so, for example, when each of $X_{1...4}$ is identically and independently normally distributed, and $X_5 = X_1 + X_2 + X_3 + X_4$.) However, it is questionable whether in actual experimental situations one ever manages to account for a large proportion (i.e. over 99%) of the variance of a variable one wishes to explain, by amassing more and more small correlations of that variable with other variables (a procedure which might be dubbed "career MANOVA"). The author would be interested to hear of any real study in which this has been achieved.

Even that variation in X_n which can be associated with the variables $X_1, \ldots, X_{n-1}$ cannot necessarily be associated with any of them in particular. Consider the trivariate case. The multivariate normal distribution over a vector of random variables $\mathbf{X} = (X_1, \ldots, X_n)$ has the probability density function $P(\mathbf{X}) = K \exp(-\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{X})$, where K is a constant, and $\mathbf{A}$ is a matrix subject to the constraint that the contours of P be bounded surfaces.

As an example, take $\mathbf{A} = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$. This describes a distribution in which, when any of the variables X_1, X_2 and X_3 is fixed, the correlation between the other two is 0.5. If we estimate X_3 by X_1 (which amounts to choosing the major axis of the $c = 0.5$ ellipse in Figure 1 instead of the least-squares line) then the errors $X_3 - X_1$ will have a correlation of zero with X_2. If, on the other hand, we estimate X_3 by X_2, then the errors $X_3 - X_2$ have a correlation of zero with X_1. There is no way, given only the probability distribution or a sampling from it, of separating the relationships of X_3 with X_1 and X_2.

Other correlational problems

We have considered only bivariate and multivariate correlations for normal distributions. We have not considered analyses where the values of some variables are set by the experimenter and the values of others observed, nor have we considered non-normal distributions. However, although the precise figures will be different, we do not expect our conclusions to be substantially different in these or related situations. See (Brown 1975; Runkel 1990) for further discussion of the role of correlational studies in experimental investigations.

References

Abramowitz, M. and I. A. Stegun (Eds.) (1964). *Handbook of Mathematical Functions.* National Bureau of Standards.

Brown, D. J. (1975). Mirror, mirror... down with the linear model. *American Education and Research Journal 12*(4), 491–505.

Herrnstein, R. J. and C. Murray (1994). *The Bell Curve: Intelligence and Class Structure in American Life.* Free Press.

Kendall, M., A. Stuart, and J. K. Ord (1987). *Kendall's Advanced Theory of Statistics, 5th. ed., 3 vols.* London: Edward Arnold.

McPhail, C. (1971). Civil disorder participation. *American Sociological Review 36*, 1058–1072.

Powers, W. (1990, September/October). Control theory and statistical generalizations. *American Behavioral Scientist 34*(1), 24–31.

Runkel, P. (1990). *Casting Nets and Testing Specimens: Two Grand Methods of Psychology*. New York: Praeger Publishers.

Wilks, S. S. (1962). *Mathematical Statistics*. John Wiley and Sons.

Figure legends

Figure 1: Contour and regression lines for various correlation coefficients.

Figure 2: Scatter plots of 100 points for various correlation coefficients.

Figure 1: Contour and regression lines for various correlation coefficients.

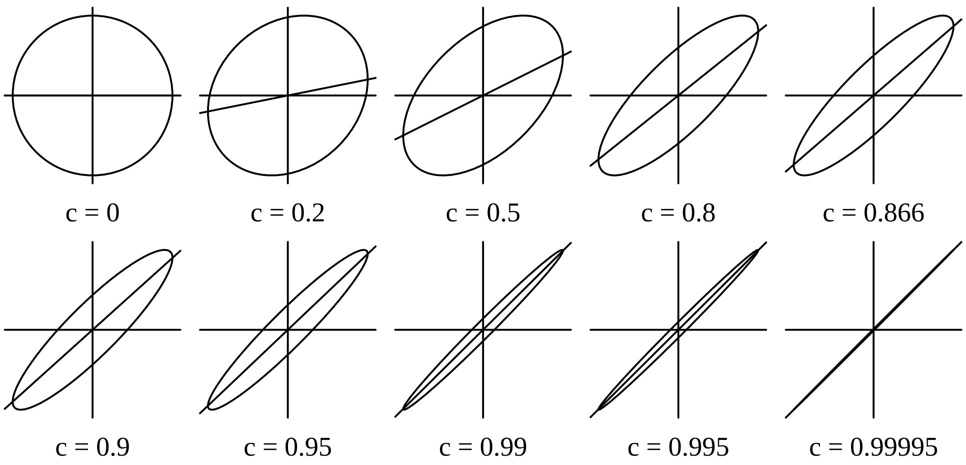

Figure 2: Scatter plots of 100 points for various correlation coefficients.

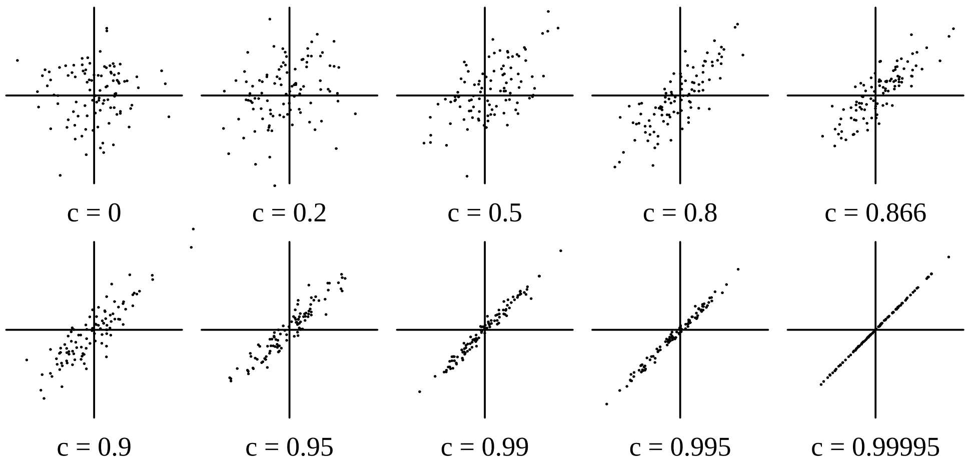

Table 1: Correlation, variance, and mutual information.

	Variance in Y		Improvement	Mutual
Correlation	attributed to X	unaccounted for	ratio	information (bits)
0	0%	100%	1	0
0.2	4%	96%	1.02	0.028
0.5	25%	75%	1.15	0.20
0.8	64%	36%	1.67	0.74
0.866	75%	25%	2	1
0.9	81%	19%	2.29	1.20
0.95	90.25%	9.75%	3.20	1.68
0.99	98%	2%	7.09	2.83
0.995	99%	1%	10	3.32
0.99995	99.99%	0.01%	100	6.64

Table 2: Binary classification and screening.

Correlation	Prob. of correct sign estimation	Rate of reliable classification 5% confidence	1% confidence
0	50%	0%	0%
0.2	56%	7.5×10^{-14}%	4.3×10^{-28}%
0.5	67%	0.4%	0.006%
0.8	80%	21.7%	8.1%
0.866	83.3%	34.2%	17.9%
0.9	85.6%	42.6%	26.0%
0.95	89.9%	58.9%	44.4%
0.99	95.5%	81.5%	74.0%
0.995	97.8%	86.9%	81.5%
0.99995	99.68%	98.7%	98.1%

Table 3: Decile estimation.

Correlation	Prob. of estimation within $\pm$ half a decile
0	10%
0.8	25%
0.9	36%
0.95	47%
0.99	78%
0.995	89%
0.997	95%
0.9985	99%

Table 4: Estimation of population correlation.

Desired lower bound on ρ	Minimum c that must be observed $\epsilon = 0.05$ Sample size 100	Sample size 20	$\epsilon = 0.01$ Sample size 100	Sample size 20
0	0.16	0.36	0.23	0.49
0.2	0.35	0.52	0.41	0.63
0.5	0.61	0.73	0.65	0.79
0.8	0.85	0.90	0.87	0.926
0.866	0.90	0.935	0.914	0.952
0.9	0.927	0.952	0.936	0.964
0.95	0.964	0.976	0.968	0.983
0.99	0.9928	0.9953	0.9937	0.9966
0.995	0.9964	0.9976	0.9969	0.9983
0.99995	0.999964	0.999976	0.999969	0.999983

Table 5: Correlation in nonrandom subpopulations.

Subpopulation relative size κ	Min. subpop. correl. c_1 for pop. correl. c			
	$c = 0.5$	$c = 0.8$	$c = 0.9$	$c = 0.95$
10%	−0.94	−0.83	−0.68	−0.44
20%	−0.79	−0.47	−0.14	0.22
40%	−0.35	0.18	0.50	0.72
50%	−0.14	0.38	0.65	0.81
60%	0.04	0.53	0.74	0.87
80%	0.32	0.70	0.85	0.92
100%	0.50	0.80	0.90	0.95

A Appendix to "Population statistics cannot be used for reliable individual prediction"

This appendix provides proofs of theorems and derivations of formulas stated in the body of the paper. It is primarily for the benefit of the referees and of interested readers of the main paper and is not necessarily intended to appear in print.

Most of the mathematics is standard statistical theory, for which see (Kendall, Stuart, and Ord 1987) as a general reference. Some of the calculations here have not to my knowledge been previously published.

A.1 The bivariate normal distribution

When integrating with respect to one variable, the following form of P is useful:

$$P(x, y, a, b, c) = \frac{1}{2abc'\pi} \exp(-x'^2/c'^2) \exp(-(y' - cx')^2/2)$$

where $x' = x/a$, $y' = y/b$, and $c' = 1/\sqrt{1-c^2}$. In addition, we recall the following facts about normal distributions:

$$\int_x e^{-x^2/2k^2} = k\sqrt{2\pi}$$

$$\int_x x^2 e^{-x^2/2k^2} = k^3\sqrt{2\pi}$$

That is, $\frac{1}{k\sqrt{2\pi}} e^{-x^2/2k^2}$ is the probability density function for a normal distribution with mean 0 and standard deviation k.

THEOREM A.1

1. *P is a probability distribution, i.e. it is everywhere non-negative, and its integral over all x and y is 1.*
2. *If Y is fixed at y, the distribution of X is a normal distribution with mean acy/b and standard deviation a.*
3. *If Y is unknown, the distribution of X is a normal distribution with mean 0 and standard deviation ac'.*

PROOF.

1. P is obviously non-negative.

$$\begin{aligned}
\int_x \int_y P(x, y, a, b, c) &= \frac{1}{2abc'\pi} \int_x \int_y \exp(-\frac{1}{2}(\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{2cxy}{ab})) \\
&= \frac{1}{2c'\pi} \int_x \int_y \exp(-\frac{1}{2}(x^2 + y^2 - 2cxy)) \\
&= \frac{1}{2c'\pi} \int_x \exp(-x^2/2c'^2) \int_y \exp(-(y - cx)^2/2) \\
&= \frac{1}{2c'\pi} \int_x \exp(-x^2/2c'^2) \int_y \exp(-y^2/2) \\
&= \frac{1}{c'\sqrt{2\pi}} \int_x \exp(-x^2/2c'^2) \\
&= 1
\end{aligned}$$

2.

$$\begin{aligned}
P(x,y,a,b,c) &= \frac{1}{2abc'\pi}\exp(-\frac{1}{2}(\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{2cxy}{ab})) \\
&= \frac{1}{2abc'\pi}\exp(-\frac{1}{2}(\frac{y^2}{b^2c'^2}+(\frac{x}{a}-\frac{cy}{b})^2) \\
&= K\exp(-\frac{1}{2}(\frac{x}{a}-\frac{cy}{b})^2)
\end{aligned}$$

where K is independent of x. This is a normal distribution of X as described.

3. Most of the calculation is included in the proof of part 1. Just leave out the integration over x and the change of variables $x/a \to x$.

□

THEOREM A.2 *If X and Y are distributed according to P, then the correlation of X and Y is c.*

PROOF. The product-moment correlation coefficient of any bivariate probability distribution $P(X,Y)$ is defined to be

$$C = \frac{\int_x\int_y(x-\overline{x})(y-\overline{y})P(x,y)}{\sqrt{\int_x\int_y(x-\overline{x})^2P(x,y)\int_x\int_y(y-\overline{y})^2P(x,y)}}$$

where $\overline{x} = \int_x\int_y xP(x,y)$ and $\overline{y} = \int_x\int_y yP(x,y)$.

For our particular probability distribution, this simplifies to

$$\begin{aligned}
C &= \frac{\int_x\int_y xyP(x,y,a,b,c)}{\sqrt{a^2c'^2b^2c'^2}} \\
&= \frac{1}{2a^2b^2c'^3\pi}\int_x\int_y xy\exp(-\frac{1}{2}(\frac{x^2}{a^2}+\frac{y^2}{b^2}-\frac{2cxy}{ab})) \\
&= \frac{1}{2c'^3\pi}\int_x\int_y xy\exp(-\frac{1}{2}(x^2+y^2-2cxy)) \\
&= \frac{1}{2c'^3\pi}\int_x\left(x\exp(-x^2/2c'^2)\int_y y\exp(-\frac{1}{2}((y-cx)^2))\right) \\
&= \frac{1}{2c'^3\pi}\int_x\left(x\exp(-x^2/2c'^2)\int_y(y+cx)\exp(-y^2/2)\right) \\
&= \frac{1}{2c'^3\pi}\left(\int_x cx^2\exp(-x^2/2c'^2)\right)\left(\int_y\exp(-y^2/2)\right) \\
&= \frac{c}{c'^3\sqrt{2\pi}}\int_x x^2\exp(-x^2/2c'^2) \\
&= c
\end{aligned}$$

□

The fact that the regression lines of Y against X and of X against Y are not identical often seems counterintuitive to beginning students of statistics. Figure 1 demonstrates why each line must stand in the relation it does to the ellipse (at least for a bivariate normal distribution), and hence why the two lines must be different, and why they approach each other as c tends to ± 1.

As the ellipses are all similar, we fix on the contour lines for $k = 1$ and consider how the shape of the ellipse changes with c. (The contour height is $P(1,0,1,1,c) =$

$1/2\pi\sqrt{e} \sim 0.097$.) This contour is defined by the equation $x^2 + y^2 - 2cxy = 1$. When $c = 0$, this is the unit circle. As c increases, it becomes an ellipse whose major axis is along the line $x = y$, and stretches more and more as c approaches 1. In the limit as c approaches 1, the ellipse stretches without bound, approximating to the pair of straight lines $x - y = \pm 1$. For negative c the symmetrical transformation happens along the line $x = -y$. Figure 1 shows this for selected values of c, scaled so that the variables have unit unconditional standard deviation. The ellipses tend to the finite line segment from $(-1, -1)$ to $(1, 1)$ as c approaches 1.

The values of $\Phi(x)$ we require sometimes go far beyond the range of standard statistical tables. We have therefore calculated them from formulae 7.1.26 (for small $|x|$) and 7.1.13 (for large $|x|$) from (Abramowitz and Stegun 1964), for the related error function $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-x^2} dx$. $\mathrm{erf}(x)$ and $\Phi(x)$ are related by $\Phi(x) = \frac{1}{2}(1 + \mathrm{erf}(x/\sqrt{2}))$.

A.2 Analysis of variance

Suppose we try to fit a function $Y = f(X)$ to a probability distribution $P(X, Y)$ in the plane. One measure of the goodness of fit is the average squared error:

$$\int_x \int_y P(x, y)(f(x) - y)^2$$

If f is to be chosen to minimise this quantity, then it is clear that for each x, $f(x)$ should be chosen to minimise $\int_y P(x, y)(f(x) - y)^2$. Let $P_x(y)$ be the distribution of Y given $X = x$. Thus $P_x(y) = K_x P(x, y)$, where $K_x = 1/\int_y P(x, y)$. So $f(x)$ should minimise $\int_y P_x(y)(f(x) - y)^2$. Since this is equal to $(f(x) - \overline{y}_x)^2 + \int_y (\overline{y}_x - y)^2$, where $\overline{y}_x = \int_y y P_x(y)$, it is minimised by taking $f(x) = \overline{y}$. In other words, the least squares estimate of a random variable is its mean.

When this is done, the total variance of Y can be divided into two parts: the variance in $f(X)$ due to variation in X, and the remainder, the variance of $f(X) - Y$. We can say that the first component is the part of the variance of Y which is explained by variation in X. This is a technical mathematical meaning of the word "explain" which bears little relation to its everyday use. This calculation can be made regardless of the meaning of X and Y. To draw a line through a set of points does not in itself constitute an explanation of anything.

The total variance of Y, assuming Y to have mean 0, is $\int_x \int_y y^2 P(x, y)$. The division into two components is constructed thus:

$$\begin{aligned}\int_x \int_y y^2 P(x, y) &= \int_x \left(\overline{y}_x{}^2 P(x) + \int_y (\overline{y}_x - y)^2 P(x, y) \right) \\ &= \mathrm{Var}(f(X)) + \int_x \int_y (f(x) - y)^2 P(x, y)\end{aligned}$$

For the bivariate normal distribution, $\overline{y}_x = bcx/a$. Therefore $f(X) = bcX/a$ is the function which best estimates Y from X, and since it happens to be linear, it is also the best straight line. The two components of the variance are $c^2 b^2 c'^2$ and b^2, their sum being $b^2 c'^2$. The proportion of the variance due to X is c^2, and the remaining proportion is $1 - c^2$.

A.3 Mutual information

For a joint normal distribution we shall compute the entropy H_X of X given no information about Y, and the entropy $H_{X|Y}$ of X given Y, where the entropy $H(P)$ of a probability distribution P is the integral of $-P \lg P$ over the whole space. This enables us to calculate $I_{X|Y} = H_X - H_{X|Y}$, the information given about X by knowing the value of Y. (Note that this quantity is actually symmetrical in X and Y: it is also the information given about Y by knowing the value of X. This is true for all distributions.)

Theorem A.3 *The entropy of a one-dimensional normal distribution with standard deviation a is $\lg a + C$, where C is the constant $\lg(\frac{\sqrt{2\pi}}{e}) = -0.1169$.*

Proof. Let $P(x) = \frac{1}{a\sqrt{2\pi}} \exp(x^2/2a^2)$.

$$
\begin{aligned}
\int_x -P \lg P &= \frac{1}{\log 2} \int_x -P \log P \\
&= \frac{1}{a\sqrt{2\pi}\log 2} \int_x \exp(\frac{x^2}{2a^2})(-\frac{x^2}{2a^2} + \log(a\sqrt{2\pi})) \\
&= \lg(a\sqrt{2\pi}) - \frac{1}{a\sqrt{2\pi}\log 2} \int_x \frac{x^2}{2a^2} \exp(\frac{x^2}{2a^2}) \\
&= \lg(a\sqrt{2\pi}) - \frac{1}{\log 2} \\
&= \lg a + \lg \frac{\sqrt{2\pi}}{e}
\end{aligned}
$$

□

Theorem A.4 *$H_X = \lg(ac'\sqrt{2\pi}) - C$. $H_{X|Y} = \lg(a\sqrt{2\pi}) - C$.*

Proof. Immediate from Theorems A.1(parts 2 and 3) and A.3. □

Theorem A.5 *The* mutual information *$I_{X|Y}$ given about X by knowing the value of Y is $\lg c'$.*

Proof. By definition, $I_{X|Y} = H_X - H_{X|Y}$. The result follows from the preceding theorem. □

A.4 Sign estimation

Without loss of generality, we assume that X and Y are normalised to have standard deviation 1. The proportion of errors in estimating the sign of Y from the sign of X is the measure of the upper right and lower left quadrants of an elliptical contour line of the distribution, as a fraction of the area of the ellipse. The ratio of the lengths of the $x = y$ and $x = -y$ axes of the ellipse is $\sqrt{\frac{1+c}{1-c}}$. Squeezing the plane along the $x = y$ axis by this factor transforms the ellipse into a circle, and changes the angle between the x and y axes to 2θ, where $\tan\theta = \sqrt{\frac{1+c}{1-c}}$. The proportion sought for is now the ratio of the transformed sectors to the area of the circle, which is $2\theta/\pi$. The proportion of correct predictions for a correlation of c is therefore $\frac{2}{\pi}\tan^{-1}\sqrt{\frac{1+c}{1-c}} = \frac{1}{\pi}\cos^{-1}(-c)$.

Now consider the modified problem of estimating the sign of $Y' - Y$ by the sign of $X' - X$. (x, y, x', y') is distributed over a four-dimensional space according to $P(X, Y, X', Y') = P(X, Y)P(X', Y')$, where $P(X, Y)$ is the bivariate normal distribution. As before we take $P(X, Y)$ to be normalised so that both arguments have mean 0, standard deviation 1, and correlation c. The probability we require is the total measure of the four-dimensional subspace within which $x' - x$ and $y' - y$ have the same sign.

Observe that X and X' are independently distributed, and that $P(x, y, x', y')$ is circularly symmetric about the plane $y = 0$, $y' = 0$. The same is true of Y, Y', and the plane $x = 0$, $x' = 0$. Therefore $P(x, y, x', y')$ remains invariant under any four-dimensional rotation about these two planes. Consider a rotation of $\pi/4$ about each plane. That is, we change to coordinates (x_1, y_1, x'_1, y'_1) such that:

$$\begin{aligned} x_1 &= (x' + x)/\sqrt{2} \\ x'_1 &= (x' - x)/\sqrt{2} \\ y_1 &= (y' + y)/\sqrt{2} \\ y'_1 &= (y' - y)/\sqrt{2} \end{aligned}$$

Then the measure we require is of that part of the space where x'_1 and y'_1 have the same sign. Since $P(x, y, x', y') = P(x_1, y_1, x'_1, y'_1) = P(x_1, y_1)P(x'_1, y'_1)$, this is equal to the measure of that part of the distribution $P(X, Y)$ where X and Y have the same sign. This is exactly what we calculated earlier. Therefore a correlation of c implies that the proportion of correct predictions is $\frac{1}{\pi}\cos^{-1}(-c)$.

A.5 Screening tests

The distribution of Y, given $X = x$, is normal with standard deviation 1 and mean cx. When x is positive, the probability that Y is negative is $1 - \Phi^{-1}(cx)$. For this to be at most ϵ, x must be at least $\delta = \frac{1}{c}\Phi^{-1}(1 - \epsilon)$. A similar analysis applies for negative x. Since the unconditional standard deviation of x is c', the probability $S(c, \epsilon)$ that the magnitude of x is at least this great is $2(1 - \Phi(\delta/c'))$.

For the modified screening test problem (estimating the sign of $Y' - Y$ by the sign of $X' - X$ when $|X' - X| > \delta$), we apply the same four-dimensional rotation as for the modified sign estimation problem. This reduces it to the basic screening test problem.

The number of atoms in 2.5 tons of water is 2.5×10^6 grams $\times$ 3 atoms per molecule $\times 6 \times 10^2 3$ molecules per mole / 20 grams per mole of water $= 2.25 \times 10^{29}$.

A.6 Correlations and regression lines for non-random subpopulations

For the relation between population correlation c, subpopulation correlation c_1, and relative subpopulation size κ, we must find the largest ellipse associated with a correlation of c_1 which is contained in one of correlation c. This is an ellipse, one axis of which coincides with the minor axis of the larger ellipse. The ratio of areas of the ellipses is the ratio of their other axes. From the preceding formula it follows that this ratio is $\kappa = \sqrt{\frac{(1-c)(1+c_1)}{(1+c)(1-c_1)}}$. Given c and κ, define $d = \kappa^2\frac{1+c}{1-c}$. Then $c_1 = \frac{d-1}{d+1}$. This is the formula tabulated in Table 5.

We can similarly derive a lower bound for the maximum possible subpopulation correlation (not tabulated in the main paper): this is $\frac{1-d'}{1+d'}$ where $d' = \kappa^2\frac{1-c}{1+c}$.

An example such as is described at the end of the section on subpopulations consists of a subpopulation, bivariately normally distributed with means 0, unconditional standard deviations 1, and correlation c, and a population which adds a single point at (A, B), for some A and B. Choosing A large enough and $B = \pm A$ gives a population correlation arbitrarily close to ± 1. By the mean value theorem, there are other choices of A and B yielding any correlation between those limits.