In: Encyclopedia of Biostatistics. P. Armitage and T. Colton (eds). New York: Wiley, 4260-4264.

[4] Carter, R.L., Scheaffer, R.L. & Marks, R.G. (1987). The role of consulting units in statistics departments, *American Statistician* **40**, 260-264.

[5] DeMets, D.L., Anbar, D., Fairweather, W., Louis, T.A. & O'Neill, R.G. (1994). Training the next generation of biostatisticians, *American Statistician* **48**, 280-284.

[6] Derr, J.A. (1993). Biostatistics cores: improving the chances for funding, *American Statistician* **47**, 99-102.

[7] Derr, J.A. (1995). Statistics in nutrition, part 8: a review of good statistical practices, *Journal of Renal Nutrition* **5**, 208-209.

[8] Derr, J.A. & Rosenberger, J.L. (1992). A multi-objective course in statistical consulting, in *American Statistical Association 1992 Proceedings of the Section on Statistical Education*. American Statistical Association, Alexandria, pp. 269-272.

[9] Derr, J.A. & Stinnett, S.S. (1994). The interesting life of a practicing statistician, *Stats. The Magazine for Students of Statistics* **11**, 7-11.

[10] Ederer, F. (1979). The statistician's role in developing a protocol for a clinical trial, *American Statistican* **33**, 116-119.

[11] Feigl, P. (1980). The training of statisticians for clinical trials, *Biometrics* **36**, 677-678.

[12] Gardner, M.G. & Bond, J. (1993). An exploratory study of statistical assessment of papers published in the *British Medical Journal, Journal of the American Medical Association* **263**, 1355-1357.

[13] Gehan, E.A. (1980). The training of statisticians for cooperative clinical trials: a working statistician's viewpoint, *Biometrics* **36**, 699-706.

[14] Gibbons, J.D. & Freund, R.J. (1980). Organizations for statistical consulting at colleges and universities, *American Statistician* **34**, 140-145.

[15] Hammond, D. (1980). The training of clinical trials statisticians: a clinician's view, *Biometrics* **36**, 679-685.

[16] Hunter, W.G. (1981). The practice of statistics: the real world is an idea whose time has come, *American Statistician* **35**, 72-76.

[17] International Committee of Medical Journal Editors (1993). Uniform requirements for manuscripts submitted to biomedical journals, *Journal of the American Medical Association* **269**, 2282-2286.

[18] Kirk, R.E. (1991). Statistical consulting in a university: dealing with people and other challenges, *American Statistician* **45**, 28-34.

[19] McCulloch, C.E., Boroto, D.R., Meeter, D., Polland, R. & Zahn, D.A. (1985). An expanded approach to educating statistical consultants, *American Statistician* **39**, 159-167.

[20] Niland, J.C., Odom-Maryon, T.L., Lee, J. & Tilley, B.C. (1995). A survey of biostatistical consulting units throughout North America, *American Statistician* **49**, 183-189.

[21] Peterson, A.V. & Fisher, L.D. (1980). Teaching the principles of clinical trials design and management, *Biometrics* **36**, 687-697.

[22] Stinnett, S.S. (1990). Training statistical consultants using modules and mirrors, in *American Statistical Association 1990 Proceedings of the Section on Statistical Education*. American Statistical Association, Alexandria, pp. 194-199.

[23] Stinnett, S.S. (1991). Quality improvement procedures in statistical consulting education, in *American Statistical Association 1991 Proceedings of the Section on Statistical Education*. American Statistical Association, Alexandria, pp. 147-152.

[24] Stinnett, S.S. (1993). Are videotaping and psychology worth the effort for statistical consultants? in *American Statistical Association 1993 Proceedings of the Section on Statistical Education*. American Statistical Association, Alexandria, pp. 148-151.

[25] Williford, W.O., Kroll, W.F., Bingham, S.F., Collins, J.F. & Weiss, D.G. (1995). The multicenter clinical trials coordinating center statistician: "More than a consultant", *American Statistician* **49**, 221-225.

[26] Wilson, W.J. (1992). Statistical consulting is scholarship, *American Statistician* **46**, 295-298.

[27] Zahn, D.A. & Isenberg, D.J. (1983). Nonstatistical aspects of statistical consulting, *American Statistician* **37**, 297-302.

SANDRA S. STINNETT, JANICE A. DERR & EDMUND A. GEHAN

Statistical Dependence and Independence

Statistical dependence is a type of relation between any two features of units under study. These units may, for instance, be individuals, objects, or various aspects of the environment. Deterministic dependence and statistical independence can be regarded as the two opposite extreme types of relation, but also as being qualitatively distinct from the possible other forms of relation. If deterministic dependence and independence are excluded, then the remaining intermediate types of statistical dependence involve both features as proper variables such that there are differences in the distributions of one variable for at least some of the levels of the other.

If proper variables are statistically independent, then the distribution of one of them is the same no matter at which fixed levels the other variable is considered and observations for such variables will lead correspondingly to nearly equal frequency distributions. If there is deterministic dependence, then the levels of one of the variables vary in an exactly determined way with changing levels of the other. In other words, under independence, knowledge about one feature remains unaffected by information provided about the other, while under deterministic dependence it follows with certainty which level of one variable occurs as soon as the level of the other variable is known.

The definition of these opposite extreme types of relation is symmetrical between the two features involved, but in its intermediate forms, statistical dependence may or may not be considered in a symmetric way, depending on the substance matter context. A symmetrical type of dependence will be appropriate if the variables involved are considered to be on an equal footing, such as symptoms of a disease, or as length, height and depth of produced objects, or as personality characteristics of individuals. By contrast, an asymmetrical form of dependence is of main interest if, instead, one of the variables is considered as a possible **response** to the other, such as weight to caloric intake, or as depression to anxiety. The terms symmetric **association** and directed association are often used to capture this distinction.

Given observations on independent units, statistical dependence shows in a number of different ways depending on several aspects. Important are, in particular, the types of variable involved, the conditions under which the relation is recorded, and the type of association measures used to summarize the data. These issues are addressed next, in turn.

Relations Depending on Types of Variable

One important distinction for variables is whether they are qualitative or quantitative. Quantitative variables have levels that are numerical values with a substantive meaning, such as kilograms, as ranks, or as sumscores of questionnaires. Qualitative variables have, instead, categories as possible levels. With a nominal scale the categories are just of a qualitatively similar kind such as **blood groups**; numbers possibly assigned to them play the role of codes; that is, of mere labels. In the case in which levels of a qualitative variable can be ranked, the scale becomes ordinal. This information may sometimes be exploited to improve formal analysis (*see* **Measurement Scale**).

First, data summaries appropriate to detect the form of **pairwise dependence** change with the types of variable involved. They are, typically, **contingency tables** for qualitative or discretized quantitative variables, scatter plots for quantitative variables and frequency distributions (or at least selected characteristics of the distributions) of the quantitative variable displayed within each category of the qualitative variable (*see* **Graphical Displays**).

Accordingly, a great variety of more formal techniques is available. In the case of symmetric associations examples are **loglinear models** for qualitative variables, covariance selection for quantitative variables (*see* **Variable Selection**), and mixed **interaction** models for both qualitative and quantitative variables. In the case of directed associations examples are logistic [2] and probit [6] regression for discrete responses (*see* **Quantal Response Models**), **linear regression** for quantitative responses, and combinations of these for mixed joint responses. In any case it is essential to check systematically [4] for more complex dependencies involving several variables or, possibly, nonlinear relations among quantitative variables.

Relations Depending on the Conditioning Set

Every statistical dependence among observed variables is a conditional relation, since there is always some conditioning, at least implicitly on time and location of the study. A more explicit form of conditioning may result by design or by statistical analysis involving several recorded variables. In that case the distinction between conditional and marginal dependence and conditional and marginal independence becomes relevant. Both may convey different information. A marginal dependence of a response on a potential explanatory variable may, for instance, be completely explainable in terms of a corresponding conditional independence statement given an intermediate variable, which itself is strongly related to both.

One example from the German labor market in 1986 is shown here with the following 2^3 contingency table, adapted from job placement statistics [1]. The response is successful job placement, A, the intermediate variable is field of study, B, and the potential explanatory variable is gender of the applicant, C. If the marginal dependence of job placement on gender is considered, i.e. the overall association of

Table 1 Overall dependence in spite of conditional independence

	B, field of qualification				Overall; that is, summed over *B*	
	Home economics		Mechanical engineering			
A, successful job placement	*C*, gender		*C*, gender		*C*, gender	
	Female	Male	Female	Male	Female	Male
Yes	15 (3.61%)	2 (3.64%)	4 (20.0%)	95 (21.1%)	19 (4.4%)	97 (19.2%)
No	400	53	16	355	416	408
Sum	415	55	20	450	435	505

pair (*A*, *C*), shown on the right-hand side of Table 1, it appears as if there were discrimination against women, since females have a much lower chance than men of obtaining a job.

This dependence can, however, be explained in the following way: home economics was a preferred field of qualification for women, while mechanical engineering was strongly preferred by men. At the same time there were many more successful job placements for mechanical engineers than for home economists, simply because many more job openings were available for the former. Within each of the two fields of qualification there was the same percentage of successful job placements for both, women and men. In other words, *A* is conditionally independent of *C* given *B* (*see* **Simpson's Paradox**).

This conditional independence, together with the strong marginal associations for pairs (A, B) and (B, C) both having variable *B* in common, imply the observed dependence for (A, C); that is, this dependence is generated by the intermediate variable *B*. The data are also an example of a simple **Markov chain** [8] and, more generally, of a graphical Markov model, a general framework (see [4], [5], and [7]) within which sequences of response, intermediate and **explanatory** variables, both types of variables, qualitative and quantitative, distinct levels of conditioning and interactive as well as nonlinear relations, may be modeled explicitly.

Judgment of Relations as Dependent on Measures of Association

In many contexts it is possible to summarize dependencies concisely with a few carefully chosen measures of association (*see* **Association, Measures of**). One example for a quantitative response and equally spaced levels of a quantitative explanatory variable is the set of coefficients of a **polynomial regression**. If, for instance, the dependence can be well captured by an orthogonal polynomial in three coefficients, then the dependence is additively decomposed into an overall mean, a linear, and a quadratic effect. A direct extension is, conceptually though not technically, the decomposition of a time dependence into a general level, a linear trend, and seasonal effects.

Some measures of association arise as parameters in multivariate distributions. In such distributions, it is typical that discrete random variables model qualitative features and continuous random variables model quantitative features. For symmetric associations one prominent example is the **exponential family** called the conditional Gaussian (CG) distribution, in which the continuous variables have a joint Gaussian distribution for each level combination of the discrete variables.

In the bivariate versions of the CG distribution, the canonical association parameters are log **odds ratios** for two discrete variables, multiples of the simple **correlation** coefficient for two continuous variables, and a weighted difference in means for the mixed case. In higher dimensions these association parameters are generalized in such a way that null values of all terms involving a particular pair of variables imply conditional independence of the pair given all remaining variables: the measures of association are then conditional log odds ratios, multiples of partial correlation coefficients, and weighted differences of means, corrected for effects of the remaining variables.

The obvious danger in using measures of association which are part of a well studied joint distribution is that the true distribution of the features under study may be quite different. For instance, if the judgment of dependencies among quantitative variables were based only on simple and partial correlation coefficients, then substantial misjudgments of the actual relations might result. If the simple correlation is zero, then strong nonlinear relations of a particular type may still be present, but at least, if the simple correlation is nonzero, the variable pair will always be marginally dependent. The situation is much worse with partial correlations.

Every partial correlation coefficient is a simple correlation coefficient for **residuals** obtained after linear regression on some common set of further variables. As for the simple correlation, there may be strong nonlinear conditional associations even if a partial correlation coefficient is zero. However, the reverse may happen as well; that is, the partial correlation coefficient may be high in spite of conditional independence. This is best illustrated with an example.

Let Z, U, and V be mutually independent variables, each having a standardized Gaussian distribution; that is, in particular, each having mean zero and variance one. Define Y and X as follows:

$$Y = (Z^2 - 1) + U, \qquad X = (Z^2 - 1) + Z + V.$$

Then Y is conditionally independent of X given Z, written as $Y \perp\!\!\!\perp X|Z$, because given Z only U and V are variable, and they are independent by assumption. But the simple correlation between the residuals from linear regression is 2/3; that is, the partial correlation coefficient $\rho_{xy.z}$ is sizeable.

To see this, note that linear – instead of the appropriate nonlinear – regression of Y on Z and of X on Z would give as conditional means

$$\mathrm{E}_{\text{linear}}(Y|Z) = 0, \qquad \mathrm{E}_{\text{linear}}(X|Z) = Z,$$

and hence as residuals from these linear regressions

$$R_{Y.Z} = (Z^2 - 1) + U, \qquad R_{X.Z} = (Z^2 - 1) + V.$$

Since the square of a standardized Gaussian variable has a **chi-square distribution** on one **degree of freedom**, the variable Z^2 has mean 1 and variance 2 and the residuals both have zero means. Furthermore, both residuals have variance 3 and their covariance is

$$\mathrm{cov}(R_{X.Z}, R_{Y.Z}) = \mathrm{var}(Z^2 - 1) = 2,$$

so that $\rho_{xy.z} = \mathrm{cov}(R_{X.Z}, R_{Y.Z})\{\mathrm{var}(R_{X.Z})\,\mathrm{var}\,(R_{Y.Z})\}^{-1/2} = 2/3$ even though the corresponding conditional independence statement $Y \perp\!\!\!\perp X|Z$ holds. Of course, if for corresponding observations systematic checks for nonlinearities and interactions were used [3], then it would certainly be detected that nonlinear associations are present and hence it would be noticed that correcting for only linear relations of Y on Z and of X on Z is inadequate.

An alternative to assuming that a set of variables has a particular distribution is to define the joint distribution only implicitly via a sequence of recursive conditional distributions. This is typical for graphical Markov models corresponding to so-called chain graphs. In that case, conditional dependencies of potential explanatory variables are modeled separately for each response in accordance with available substance matter knowledge [4, 9]; nonlinear relations and interactions among continuous variables may be part of the model. In addition, for a given model it may often be deduced which independencies and associations are implied under other conditioning sets than those specified with the given model [10].

Another important additional advantage of such conditional modeling is that issues such as **censoring**, measurement error (*see* **Errors in Variables**), missing values, time dependencies, and effects of hidden random variables may in principle be directly integrated into the modeling process. To date, however, the actual implementation might for some combinations still require substantial further theoretic and technical developments.

References

[1] Bundesanstalt für Arbeit (1986). *Amtliche Nachrichten* **5**, 846–847.

[2] Cox, D.R. (1958). The regression analysis of binary sequences (with discussion), *Journal of the Royal Statistical Society, Series B* **20**, 215–242.

[3] Cox, D.R. & Wermuth, N. (1994). Tests of linearity, multivariate normality and the adequacy of linear scores, *Applied Statistics* **43**, 347–355.

[4] Cox, D.R. & Wermuth, N. (1996). *Multivariate Dependencies - Models, Analysis and Interpretation.* Chapman & Hall, London.

[5] Edwards, D. (1995). *Introduction to Graphical Modelling*. Springer-Verlag, New York.

[6] Finney, D.J. (1971). *Probit Analysis*, 3rd Ed. Cambridge University Press, Cambridge.

[7] Lauritzen, S.L. (1996). *Graphical Models*. Oxford University Press, Oxford.

[8] Markov, A.A. (1912). *Wahrscheinlichkeitsrechnung* (German translation of 2nd Russian Ed. 1908). Teubner, Leipzig.

[9] Wermuth, N. (1997). Graphical Markov models, in *Encyclopedia of Statistical Sciences*, S. Kotz, C. Read & D. Banks, eds. Wiley, New York, to appear.

[10] Wermuth, N. & Cox, D.R. (1998). On association models defined over independence graphs, *Bernoulli*, to appear.

(*See also* **Pairwise Independence**)

NANNY WERMUTH & D.R. COX

Statistical Forensics

When genetic evidence is used for individual identification, there are generally competing explanations for the observations. A typical forensic situation arises when biological material at the scene of a crime is typed, found to have some profile A, and the circumstances of the crime suggest that the material was left by the perpetrator P. A person S suspected of having committed the crime is also typed, and is found to have the same profile. The evidence E is that the two profiles are of type A.

The competing explanations are:

H_p: the crime sample is from S
H_d: the crime sample is not from S

and the relative merits of these two explanations are compared by means of a **likelihood ratio**. This compares the probability of the evidence under the two explanations:

$$L = \frac{\Pr(E|H_p)}{\Pr(E|H_d)}. \tag{1}$$

Values of L greater than 1 favor the explanation H_p over H_d. If there are prior odds $\Pr(H_p)/\Pr(H_d)$ on S being the contributor, then the posterior odds $\Pr(H_p|E)/\Pr(H_d|E)$ follow from **Bayes' Theorem** as

$$\text{posterior odds} = L \times \text{prior odds}.$$

One of the most common errors in interpreting genetic evidence is to confuse the posterior odds with the likelihood ratio. This transposition of the conditional is more commonly made by prosecutors, giving rise to the term "prosecutor's fallacy". It is generally the case that $\Pr(E|H_p) = 1$, and the value of $\Pr(E|H_d)$ might be 10^{-6}. The likelihood ratio is then one million, but the posterior odds depend on the prior odds. They are not a million to one on guilt. Although odds on guilt is very much the kind of information desired by courts, it cannot be found from genetic evidence alone.

Conditional Probabilities

Eq. (1) can be modified by the rules of **conditional probability**. If S_A and P_A mean that S and P, respectively, have genetic profile A, then

$$L = \frac{\Pr(S_A, P_A|H_p)}{\Pr(S_A, P_A|H_d)} = \frac{\Pr(P_A|S_A, H_p)\ \Pr(S_A|H_p)}{\Pr(P_A|S_A, H_d)\ \Pr(S_A|H_d)}.$$

It may generally be assumed that the profile type of S does not depend on either explanation of the matching profiles, so $\Pr(S_A|H_p) = \Pr(S_A|H_d)$, and that a match is certain under H_p, so $\Pr(P_A|S_A, H_p) = 1$, and then

$$L = \frac{1}{\Pr(P_A|S_A, H_d)}.$$

The focus on conditional probabilities greatly simplifies the interpretation of matching profiles. The question is clearly seen to be "What is the probability that the perpetrator of the crime is of type A given that S is of type A, when these two people are not the same?" The smaller this probability, the stronger the evidence against S. By emphasizing that L depends on the probability of an event, comparisons between L and the size of the population are avoided. There is no inconsistency between an L of one million and a population size of one thousand. One has nothing to do with the other.

In the special case that profile probabilities of different people S and P are independent, the likelihood ratio reduces to the reciprocal of the profile probability ("profile frequency")

$$L = \frac{1}{\Pr(P_A)}. \tag{2}$$

This equation will not hold if S and P are related, or if they both belong to the same subpopulation. In one case the two people are related by virtue of being in the same family, and in the second they are related in an evolutionary sense. Although the