Journal of Educational & Behavioral Statistics
Summer 2000, Vol. 25, No. 2, pp. 185–201

Bayesian Procedures for Prediction Analysis of Implication Hypotheses in 2 × 2 Contingency Tables

Bruno Lecoutre
C.N.R.S. and Université de Rouen, France
Camilo Charron
IUFM de Bretagne, Saint-Brieuc, France

Keywords: 2 × 2 *contingency tables, implication hypotheses, predictive efficiency, multinomial model, conditional confidence intervals, Bayesian inference*

Procedures for prediction analysis in 2 × 2 *contingency tables are illustrated by the analysis of successes to six types of problems associated with the acquisition of fractions. According to Hildebrand, Laing, and Rosenthal (1977), hypotheses such as "success to problem type A implies in most cases success to problem type B" can be evaluated from a numerical index. This index has been considered in various other frameworks and can be interpreted in terms of a measure of predictive efficiency of implication hypotheses. Confidence interval procedures previously proposed for this index are reviewed and extended. Then, under a multinomial model with a conjugate Dirichlet prior distribution, the Bayesian posterior distribution of this index is characterized, leading to straightforward numerical methods.[1] The choices of "noninformative" priors for discrete data are shown to be no more arbitrary or subjective than the choices involved in the frequentist approach. Moreover, a simulation study of frequentist coverage probabilities favorably compares Bayesian credibility intervals with conditional confidence intervals.*

Measuring Predictive Efficiency of Implication Hypotheses Between Binary Attributes

Example: The Acquisition of Fractions Problems

While children learn rational numbers, fractions expressing a non-inclusive relationship (Part/Part) are more difficult to master than fractions expressing an inductive relationship (Part/Whole) (Vergnaud, 1983). This raises two questions: 1) To what extent the acquisition of Part/Whole fractions is necessary to master Part/Part fractions? 2) What precisely are the obstacles faced during the conceptualization of Part/Part fractions?

Eighteen problems concerning the use of fractions (Charron, 1995) were given to 165 pupils, 55 fifth-grade pupils (average age 10 years and 11 months),

The authors should like to thank the anonymous reviewers for very detailed and helpful suggestions on earlier versions.

55 seventh-grade pupils (average age 12 years and 10 months), 55 ninth-grade pupils (average age 15 years and 3 months). In each problem, a fraction (given or to be found) related a compared quantity to a reference quantity. Six problem types were considered by crossing two factors: 1) the *task*: computation of the Fraction (FR), computation of the Compared Quantity (CQ), or computation of the Reference Quantity (RQ), and 2) the *relationship*: Part/Whole relationship (PW) or Part/Part relationship (PP). For each problem type three situations were presented: slices of cake, clients in a restaurant, and trees in a forest. Examples of problems are given in Table 1.

A preliminary analysis did not show any effect of surface traits (cakes, clients, trees). Consequently, each type of problem defined by a task and a relationship (FRPW, CQPW, RQPW, FRPP, CQPP, RQPP) was coded as success when at least two out of three situations were correctly solved, and as failure otherwise. In what follows, we focus on the study of the oriented dependencies between the six binary attributes so defined for each subject. More generally, for each pair of problem types A and B, the four following questions can be investigated: 1) Does success to A imply success to B? 2) Does success to A imply failure to B, i.e. exclude success to B (*negative exclusion*)? 3) Does failure to A imply success to B, i.e., exclude failure to B (*exclusion*)? 4) Does success to B imply success to A?

A Measure of Predictive Efficiency

Let us consider a group of n subjects, with two sets of binary attributes, respectively $A = \{a_1, a_0\}$ and $B = \{b_1, b_0\}$. The number of occurrences of the event (a_i, b_j) ($i = 1{,}0$; $j = 1{,}0$) is denoted by n_{ij} ($\Sigma n_{ij} = n$). An *absolute* (or logical) implication from a_1 to b_1 (for instance) exists if all the subjects having the modality a_1 also have the modality b_1, whereas the converse is not necessarily true. But the hypothesis of an absolute implication, for instance "success to problem type A always implies success to problem type B," is of little practical interest, since a single observation of the event (a_1, b_0) is sufficient to falsify it. Consequently, we have to consider the weaker hypothesis "success to problem type A implies *in most cases* success to problem type B," denoted by $a_1 \approx\!> b_1$.

According to Hildebrand, Laing, and Rosenthal (1977, p. 68), the "prediction success" of this hypothesis can be quantified by the index:

$$H_{11} = 1 - \frac{n \times n_{10}}{(n_{11} + n_{10})(n_{10} + n_{00})} = 1 - \frac{f_{10}}{f_{1.}f_{.0}} \; (H_{11} \leq 1)$$

$$\text{where } f_{ij} = \frac{n_{ij}}{n} \; (\textstyle\sum f_{ij} = 1), f_{1.} = f_{11} + f_{10}, f_{.0} = f_{10} + f_{00}.$$

The prediction is perfect (there is an absolute implication) when $H_{11} = +1$. The closer to one H_{11} is, the more efficient the prediction, while $H_{11} \leq 0$ means that the hypothesis $a_1 \approx\!> b_1$ is a prediction failure.

TABLE 1
Statement examples for each of the six tests

FRPW—Computation of the FRaction of the Part/Whole relationship
"A large cake is made up of 90 slices. 36 slices have been eaten. What fraction has been eaten? Reduce the fraction to the simplest form possible (a fraction that cannot be reduced any further)."
CQPW—Computation of the Compared Quantity of the Part/Whole relationship
"In a forest, there are 80 trees and 4/5 of the trees are chestnut trees. Find the number of chestnut trees in the forest."
RQPW—Computation of the Reference Quantity of the Part/Whole relationship
"At a restaurant, 30 clients have finished eating, that is 3/5 of all the clients in the restaurant. Find the number of clients in the restaurant"
FRPP—Computation of the FRaction of the Part/Part relationship
"A cake has 49 decorated slices. 14 slices are not decorated. What fraction represents the number of non-decorated slices, compared to the number of decorated slices? Reduce the fraction to the simplest form possible (a fraction that cannot be reduced any further)."
CQPP—Computation of the Compared Quantity of the Part/Part relationship
"At a restaurant, 70 people order meat and the others order fish. The number of people who order fish represents 2/5 of the people who order meat. Find the number of people who order fish."
RQPP—Computation of the Reference Quantity of the Part/Part relationship
"In a forest of cedar and pine trees, there are 60 cedars, 3/5 the number of the pine trees. Find the number of pine trees in the forest."

This index has been considered in widely varying frameworks with different approaches. It was actually proposed by Quetelet (quoted by Yule, 1900, p. 280) as early as 1832 (for historical context, see especially Goodman & Kruskal, 1959, and Hildebrand, Laing, & Rosenthal, 1977, p. 73). Some presentations which can make the interpretation easier are briefly summarized below. It is a common viewpoint that H_{11} can be expressed in terms of a measure of predictive efficiency, when predicting the outcome of B given a_1. Since the statement $a_1 \not\Rightarrow b_1$ makes no particular prediction when $A = a_0$, only the n_{10} occurrences of the event a_1b_0 constitute prediction *errors*, while the n_{11} occurrences of the event a_1b_1 are *correct* predictions.

Thus, for Hildebrand, Laing, and Rosenthal (1977), H_{11} is a proportionate reduction in prediction error when applying the absolute implication $a_1 \Rightarrow b_1$ given knowledge of the attribute A over that expected when applying the prediction to the same number (i.e. $n_{11} + n_{10}$) of randomly selected a_1:

$$H_{11} = \frac{\text{Errors for randomly selected } a_1 - \text{Errors given knowledge of } A}{\text{Errors for randomly selected } a_1} = \frac{f_{1.}f_{.0} - f_{10}}{f_{1.}f_{.0}}.$$

According to Loeber and Dishion (1983), H_{11} is intended for comparing correct prediction against prediction based on chance alone, hence the name of Relative Improvement Over Chance (RIOC) index. It can be defined by the following formula, which corrects for chance and for the maximum ceiling:[2]

$$H_{11} = \frac{\text{Correct} - \text{Chance Correct}}{\text{Maximum Correct} - \text{Chance Correct}} = \frac{f_{10} - f_{1.}f_{.1}}{f_{1.} - f_{1.}f_{.1}}.$$

In the context of epidemiology, H_{11} is equal to the Levin's measure of Attributable Risk (Levin, 1953), that can be thought of as the proportion by which $P(b_0)$ (where b_0 is typically a disease) can be reduced among people exposed to the risk factor a_1 if this risk factor could be eliminated (Fleiss, 1981).[3] This leads to the comparison of the conditional proportion of errors b_0 given a_1 to its marginal proportion:

$$H_{11} = \frac{P(b_0) - P(b_0|a_1)}{P(b_0)} = \frac{f_{.0} - \frac{f_{10}}{f_{1.}}}{f_{.0}},$$

which is again equivalent to Fleiss's formula:

$$H_{11} = \frac{P(b_0|a_1)P(a_1) - P(b_0|a_0)P(a_1)}{P(b_0)} = \frac{\frac{f_{10}}{f_{1.}}f_{1.} - \frac{f_{00}}{f_{0.}}f_{1.}}{f_{.0}}.$$

Lastly, it can be noted that H_{11} is equal to

$$H_{11} = 1 - \frac{n_{10}}{E(n_{10})}$$

where $E(n_{10})$ is the expectation of n_{10} for two sets of independent binary attributes. Hence it can again be considered as a residual (Expected-Observed)/Expected under an independence model. But this can lead to misunderstanding since, as stressed by Hildebrand, Laing, and Rosenthal (1977, p. 71), H_{11} measures error reduction and not just deviation from statistical independence. Actually, in the general case of $R \times C$ contingency tables, $H_{11} = 0$ does not imply statistical independence as soon as $R > 2$ or $C > 2$ (see Bernard & Charron, 1996). Moreover, the interpretation of the index as a departure from independence requires some caution. For example, a simple Rasch model with sufficient heterogeneity on the subject level parameter would lead to a large value (possibly one), even though conditionally of this subject specific effect the two questions were independent.

Inferential Procedures: Confidence Intervals and Bayesian Inference

Assume now that the group of n subjects constitutes a random sample from some population. More precisely the joint sampling distribution of the random-variables, n_{11}, n_{10}, n_{01}, n_{00} is assumed to be a *multinomial* distribution, with parameters n, φ_{11}, φ_{10}, φ_{01}, φ_{00} ($0 \leq \varphi_{ij} \leq 1$, $\Sigma\varphi_{ij} = 1$). Let η_{11} denote the numerical parameter $1 - \varphi_{10}/(\varphi_{1.}\varphi_{.0})$ (where $\varphi_{1.} = \varphi_{11} + \varphi_{10}$ and $\varphi_{.0} = \varphi_{10} + \varphi_{00}$), which measures the prediction success of the hypothesis $a_1 \Rrightarrow b_1$ in the parent population.

Nijsse (1992) and Bernard and Charron (1996) reviewed confidence interval procedures previously proposed. Two main difficulties are encountered with these procedures, the discreteness of the sampling distributions and the presence of nuisance parameters. Simple asymptotic ("large-sample") procedures are easily available, but they become clearly inappropriate when the expected frequency of one or more cells is too small. In these cases, as well as analyzing small samples, the "exact" conditional approach for constructing confidence intervals is undoubtedly preferable, but Bayesian inference will be shown to be a valuable alternative.

Asymptotic Confidence Intervals

A first procedure is based on a *normal* approximation of the sampling distribution of H_{11}:

$$H_{11}|\varphi_{11}, \varphi_{10}, \varphi_{01}, \varphi_{00} \approx N\left(\eta_{11}, \frac{f_{01}(f_{10}(f_{10}f_{01} - f_{11}f_{00}) + f_{11}f_{00})}{n(f_{1.}f_{.0})^3}\right)$$

where the approximate variance is obtained from the formula for the asymptotic variance derived in particular by Walter (1976) and Copas and Loeber (1990).[4]

An alternative procedure, suggested by Walter (1975), is obtained from the asymptotic variance of $\log(1-H_{11})$ (Fleiss, 1981, p. 76):

$$\log(1 - H_{11})|\varphi_{11}, \varphi_{10}, \varphi_{01}, \varphi_{00} \approx N\left(\log(1 - \eta_{11}), \frac{f_{10} + H_{11}(f_{11} + f_{00})}{n \times f_{10}}\right).$$

Conditional Confidence Intervals

A well-known result is that the sampling distribution of n_{11} given fixed marginal values $n_{1.}$ and $n_{.1}$ is (see Cox, 1970, p. 4):

$$P(n_{11} = k|n_{1.},n_{.1},\varphi_{11},\varphi_{10},\varphi_{01},\varphi_{00}) = \frac{\binom{n_{1.}}{k}\binom{n-n_{1.}}{n_{.1}-k}\rho^k}{\sum_{j=\max(0,n_{1.}+n_{.1}-n)}^{\min(n_{1.},n_{.1})}\binom{n_{1.}}{j}\binom{n-n_{1.}}{n_{.1}-j}\rho^j}.$$

This distribution depends only of the cross-ratio:

$$\rho = \frac{\varphi_{11}\varphi_{00}}{\varphi_{10}\varphi_{01}}.$$

Thus the null hypothesis $\rho = \rho_0$ can be tested against the alternative $\rho > \rho_0$ by declaring the result significant at level α if

$$\overline{p}^{\rho_0}_{inc} = \sum_{k=n_{11}}^{\min(n_{1.},n_{.1})} P(n_{11} = k|n_{1.},n_{.1},\rho = \rho_0) \leq \alpha.$$

In the particular case $\rho_0 = 1$, this test is the Fisher's randomization test of the null hypothesis $\rho = 1$ (i.e., $\eta_{11} = 0$) against $\rho > 1$ ($\eta_{11} > 0$).

Note that in this solution the summation is over all k, consistent with the fixed margins, which are greater than *or equal to* the observed frequency n_{11} (hence the subscript *inc* for *inclusive*). Since Fisher's test can be highly conservative, an alternative (but liberal) solution would be to exclude of the summation the observed frequency value. Let $\overline{p}^{\rho_0}_{exc}$ be the corresponding probability. Moreover some solutions defined as intermediate between the inclusive and exclusive values can be thought to be preferable. So we will also consider here the *mid* probability: $\overline{p}^{\rho_0}_{mid} = \frac{1}{2}(\overline{p}^{\rho_0}_{inc} + \overline{p}^{\rho_0}_{exc})$.

Then, for each of these solutions, a lower confidence limit for ρ can be found by solving $\overline{p}^{\rho_0} = \alpha$. In the same way an upper confidence limit is given by $\underline{p}^{\rho_0} = \alpha$, where:

$$\underline{p}^{\rho_0}_{inc} = \sum_{k=\max(0,n_{1.}+n_{.1}-n)}^{k=n_{11}} P(n_{11} = k|n_{1.},n_{.1},\rho = \rho_0) = \alpha,$$

the exclusive $\underline{p}^{\rho_0}_{exc}$ and mid $\underline{p}^{\rho_0}_{mid}$ solutions being defined as before.

The equations above for the confidence limits associated with the exclusive probabilities were given by Copas and Loeber (1990, p. 304). However, these authors did not consider their exact solutions, but rather obtained confidence limits for log ρ using a *normal* approximation, with mean:

$$\widehat{\log\rho} = \log \frac{\left(n_{11} + \frac{1}{2}\right)\left(n_{00} + \frac{1}{2}\right)}{\left(n_{10} + \frac{1}{2}\right)\left(n_{01} + \frac{1}{2}\right)}$$

and estimated variance:

$$\frac{(n_{1.} + 1)(n_{1.} + 2)}{n_{1.}(n_{11} + 1)(n_{10} + 2)} + \frac{(n + 1 - n_{1.})(n + 2 - n_{1.})}{(n - n_{1.})(n_{01} + 1)(n_{00} + 1)}.$$

Insofar as this approximation, which involves a continuous distribution, is accurate, it can also be expected to give intermediate values between the exact inclusive and exclusive limits. Note also that adding $\frac{1}{2}$ to the cell frequencies in the estimate of log ρ avoids impossible calculations due to null frequencies (but in an arbitrary way).

Now, going back to our problem, confidence limits for η_{11} are found by substituting the corresponding limits for ρ in the following expression that gives η_{11} as a function of ρ (see Copas & Loeber, 1990, p. 305):

$$\eta_{11} = \frac{1+(\rho-1)(\varphi_{1.}+\varphi_{.1}-2\varphi_{1.}\varphi_{.1})-[1+(\varphi_{1.}+\varphi_{.1})(\rho-1)^2-4\varphi_{1.}\varphi_{.1}\rho(\rho-1)]^{1/2}}{2(\rho-1)\varphi_{.1}(1-\varphi_{1.})}.$$

Unfortunately these limits depend on the true margin values $\varphi_{.1}$ and $\varphi_{1.}$. Following a common procedure, these parameters are simply replaced by their estimates $f_{.1}$ and $f_{1.}$

Bayesian Inference

Let us assume for φ_{11}, φ_{10}, φ_{01}, φ_{00} a joint prior *Dirichlet* distribution, with parameters υ_{11}, υ_{10}, υ_{01}, υ_{00} (natural conjugate prior). In this case, the posterior distribution is also *Dirichlet* with parameters $n_{11}+\upsilon_{11}$, $n_{10}+\upsilon_{10}$, $n_{01}+\upsilon_{01}$, $n_{00}+\varphi_{00}$ (e.g., Novick & Jackson, 1974, p. 345). Then the marginal posterior distribution for η_{11} can be characterized as follows:

Let

$$Z = \varphi_{10}$$

$$Y = \frac{\varphi_{00}}{1-\varphi_{10}} = \frac{\varphi_{00}}{1-Z}$$

$$X = \frac{\varphi_{11}}{1-\varphi_{10}-\varphi_{00}} = \frac{\varphi_{11}}{(1-Y)(1-Z)}$$

$$\text{hence } \eta_{11} = 1 - \frac{Z}{(Z+X(1-Y)(1-Z))\,(Z+Y(1-Z))}.$$

From the basic properties of the *Dirichlet* distribution (see Bernardo & Smith, 1994, p. 135), the three numerical variables X, Y, Z have independent posterior *beta* distributions. More precisely,

$$X\,|\,data \sim beta(\eta_{11}+\upsilon_{11},\, n_{01}+\upsilon_{01})$$
$$Y\,|\,data \sim beta(n_{00}+\upsilon_{00},\, n_{11}+\upsilon_{11}+n_{01}+\upsilon_{01})$$
$$Z\,|\,data \sim beta(n_{10}+\upsilon_{10},\, n_{11}+\upsilon_{11}+n_{01}+\upsilon_{01}+n_{00}+\upsilon_{00}).$$

Consequently, straightforward procedures to compute, for any limit $\underline{\ell}$, the probability $Pr(\eta_{11} > \underline{\ell})$ can be derived. First, $Pr(\eta_{11} > \underline{\ell})$ can be converted into the probability that the three-dimensional variable (X,Y,Z) falls in a given region of the space $[0,1]^3$. Then dividing this space into small rectangular regions and using the independence of the marginal distributions, this last probability can be approximated from the usual *incomplete beta function*, with any desired degree of accuracy, as the sum of the probabilities of these regions. This procedure generalizes the method described by Novick and Jackson (1974, p. 338) for the difference of two independent *beta* distributions, and used in Lecoutre, Derzko, and Grouin (1995). Alternatively, the posterior distribution of η_{11} can be approximated by simulating a large sample from three independent *beta* distributions. These two methods need some refinement for extreme cases (with some of the n_{ij} close to zero), but simulations appeared to us to require greater caution. On the other hand simulation procedures are more generally applicable to higher dimensionality.

Choice of the Prior and Conditional Confidence Intervals

When analyzing experimental data, a non-informative prior is generally wanted for objective report in publication. For discrete data there is no consensus for the choice of a particular distribution. Recent approaches (Bernard, 1996 and Walley, 1996) suggest the replacement of the notion of "ignorance prior" by the notion of "ignorance zone." Then the whole set of inferences (or at least extreme inferences) provided by varying the prior within this zone can be considered in order to determine the sensitivity of the Bayesian analysis. For the multinomial model, such a zone consists of all the Dirichlet priors with $0 \leq \upsilon_{ij} \leq 1$ for each (i,j). This includes in particular the *zero* prior [0 0 0 0] and the *uniform* prior [1 1 1 1]. With regard to the hypothesis of a high degree of implication, the two priors [0 1 1 0] and [1 0 0 1] have the privileged status of extreme cases. Indeed, they are respectively (within the ignorance zone) the more *unfavorable* and the more *favorable* priors. Moreover, Altham (1969) demonstrated that the Bayesian posterior probability that the cross-ratio ρ is less than one (i.e. $\eta_{11} < 0$), assuming the prior [0 1 1 0], is just the observed level $\overline{p}^{1}_{inc}$ of Fisher's randomization test. The same equality can be shown between the posterior probability associated with the prior [1 0 0 1] and the observed level $\overline{p}^{1}_{exc}$ of the exclusive solution. Such equalities do not hold for null hypothesis values other than one. Nevertheless it can be thought that the Bayesian lower credibility limits associated with these two priors are respectively relatively closed to the *inclusive* and *exclusive* frequentist interval limits (with the reverse for upper limits). In the same way, a parallel can be established between the *mid* solution and an "intermediate" symmetrical prior such that the average of the two extreme priors $[\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}]$ or again the Perk's (1947) prior $[\frac{1}{4}\frac{1}{4}\frac{1}{4}\frac{1}{4}]$.

Simulation Study

In order to investigate the frequentist coverage probabilities of the different solutions, a simulation study was conducted. Ten sets of parameters (φ_{11}, φ_{10}, φ_{01}, φ_{00}) were randomly generated by a uniform distribution, such that they had approximate values of η_{11} ranking from 0 to 0.90 by step of 0.10. Four sets with extreme parameter values ($\varphi_{ij} < 0.01$ or margins less than 0.10) were considered separately. For each of these sets, 10^4 multinomial samples of size $n = 55$ were generated and the corresponding 95% lower $\underline{\ell}$ and upper $\overline{\ell}$ limits were computed. When these limits were not calculable (in some cases of null cell frequencies) ad hoc corrections were used. For the asymptotic procedures, the arbitrary corrections proposed by the authors were used. For the conditional limits, the Copas and Loeber's approximate solution was computed instead of the "exact" solution. For the Bayesian procedures, null values of $n_{ij} + \upsilon_{ij}$ were arbitrarily replaced by 0.10.

The proportion of samples for which respectively $\underline{\ell} > \eta_{11}$ and $\overline{\ell} < \eta_{11}$ are reported in Table 2. Concerning the Bayesian procedures, it can be seen that the two frequentist error rates associated with the *extreme* priors [0 1 1 0] and [1 0 0 1] always includes 5%. Hence, simultaneously considering the two corresponding credibility limits (as advocated by Walley, 1996) protects the user both from erroneous acceptation and rejection of hypotheses about η_{11} at level 0.05. Moreover, if a single limit is wanted for summarizing and reporting results, the symetrical *intermediate* priors $[\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}]$ and $[\frac{1}{4}\frac{1}{4}\frac{1}{4}\frac{1}{4}]$ have fairly good coverage properties, including the cases of moderate sample sizes and small parameter values. Of course, the difference between the different priors in the *ignorance zone* is less for small or medium values of η_{11} and vanishes as the sample size increases.

It also appears that the *inclusive, exclusive*, and *mid* conditional confidence intervals are less efficient than their Bayesian parallels. In particular the two error rates associated with the "extreme" solutions do not always include 5%. A possible avenue for further investigations would be to search for improvements. First, the presence of nuisance parameters could be overcome by considering some weighted average over the margin parameters $\eta_{1.}$ and $\eta_{1.}$ of the associated conditional sampling probabilities. Such a principle has been previously applied by Rice (1988) for comparing two binomial proportions. Second, concerning the discreteness of the sampling distributions, the relevant probability should also be a weighted average of the exclusive and inclusive probabilities. For instance, random weights from a uniform distribution could be used, according to the general solution for discontinuous variables proposed by Toecher (1950). At least, more efficient solutions should be found for the cases of missing information due to unobserved events.

In any case, Bayesian inference copes with the problem of nuisance parameters. Moreover, it explicitly handles the problems of discreteness and unobserved events by way of the prior distribution. At the very least, it should be recognized as being no more arbitrary and subjective than frequentist procedures. The present results confirm that the notion of *ignorance prior zone* provides realistic and efficient routine procedures for analyzing data with regard to the need for objective inference. Moreover, various "informative" prior distributions, including *skeptical* and *enthusiastic* priors (see Spiegelhalter, Freedman, and Parmar, 1994), can be investigated to assess the sensitivity of the conclusions vis-à-vis additional information.

TABLE 2

Simulation study based on 10^4 multinomial samples of size $n = 55$: Proportions of samples for which $\underline{\ell} > \eta_{11}$ and for which $\overline{\ell} < \eta_{11}$

95% lower limit $\underline{\ell}$					asymptotic		conditional				Bayesian			
η_{11}	π_{11}	π_{10}	π_{01}	π_{00}	*norm*	*log*	*incl*	*excl*	*mid*	*C&L*	0110	1001	$\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}$	$\frac{1}{4}\frac{1}{4}\frac{1}{4}\frac{1}{4}$
−0.0086	0.4088	0.1260	0.3576	0.1076	0.059	0.014	0.027	0.099	0.044	0.045	0.025	0.100	0.052	0.055
0.0955	0.3139	0.2961	0.1495	0.2406	0.042	0.023	0.027	0.091	0.046	0.045	0.027	0.087	0.051	0.054
0.1904	0.2427	0.3184	0.0564	0.3825	0.031	0.015	0.035	0.115	0.058	0.050	0.017	0.068	0.034	0.042
0.3037	0.3669	0.0910	0.3475	0.1945	0.079	0.005	0.024	0.097	0.042	0.036	0.021	0.092	0.050	0.057
0.3967	0.5288	0.1950	0.0247	0.2516	0.046	0.013	0.061	0.162	0.093	0.090	0.024	0.073	0.044	0.049
0.4939	0.1232	0.0658	0.1892	0.6218	0.100	0.000	0.020	0.118	0.040	0.040	0.018	0.108	0.045	0.061
0.6084	0.1975	0.0736	0.1090	0.6198	0.096	0.000	0.027	0.128	0.049	0.042	0.020	0.099	0.048	0.061
0.7064	0.3885	0.0764	0.0518	0.4833	0.085	0.000	0.032	0.145	0.066	0.062	0.014	0.073	0.057	0.066
0.8053	0.4244	0.0483	0.0503	0.4770	0.111	0.000	0.025	0.156	0.061	0.061	0.000	0.074	0.067	0.067
0.8914	0.3672	0.0160	0.2470	0.3698	0.405	0.000	0.000	0.000	0.000	0.000	0.000	0.403	0.000	0.002
0.3882	0.0190	0.0250	0.0513	0.9037	0.312	0.148	0.100	0.150	0.117	0.186	0.005	0.198	0.052	0.082
0.4074	0.0700	0.0279	0.4489	0.4532	0.227	0.009	0.006	0.034	0.019	0.026	0.002	0.212	0.044	0.109
0.7055	0.2921	0.0767	0.0019	0.6293	0.093	0.000	0.053	0.124	0.099	0.112	0.011	0.082	0.053	0.061
0.7803	0.0343	0.0090	0.0183	0.9384	0.639	0.160	0.107	0.107	0.107	0.179	0.000	0.307	0.000	0.009

95% upper limit $\bar{\ell}$					asymptotic		conditional				Bayesian			
η_{11}	π_{11}	π_{10}	π_{01}	π_{00}	*norm*	*log*	*incl*	*excl*	*mid*	*C&L*	0110	1001	$\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}$	$\frac{1}{4}\frac{1}{4}\frac{1}{4}\frac{1}{4}$
−0.0086	0.4088	0.1260	0.3576	0.1076	0.056	0.085	0.025	0.100	0.043	0.041	0.104	0.024	0.051	0.055
0.0955	0.3139	0.2961	0.1495	0.2406	0.055	0.065	0.032	0.100	0.055	0.057	0.098	0.030	0.054	0.054
0.1904	0.2427	0.3184	0.0564	0.3825	0.075	0.084	0.131	0.229	0.158	0.202	0.115	0.038	0.065	0.065
0.3037	0.3669	0.0910	0.3475	0.1945	0.051	0.088	0.030	0.103	0.050	0.061	0.103	0.029	0.056	0.053
0.3967	0.5288	0.1950	0.0247	0.2516	0.069	0.089	0.302	0.367	0.321	0.371	0.102	0.037	0.063	0.060
0.4939	0.1232	0.0658	0.1892	0.6218	0.058	0.093	0.040	0.138	0.061	0.078	0.104	0.026	0.052	0.050
0.6084	0.1975	0.0736	0.1090	0.6198	0.049	0.088	0.076	0.176	0.107	0.138	0.087	0.033	0.060	0.054
0.7064	0.3885	0.0764	0.0518	0.4833	0.044	0.083	0.171	0.256	0.191	0.254	0.099	0.036	0.058	0.054
0.8053	0.4244	0.0483	0.0503	0.4770	0.036	0.084	0.180	0.294	0.182	0.260	0.112	0.033	0.055	0.048
0.8914	0.3672	0.0160	0.2470	0.3698	0.012	0.073	0.036	0.164	0.062	0.140	0.127	0.024	0.059	0.048
0.3882	0.0190	0.0250	0.0513	0.9037	0.398	0.394	0.129	0.211	0.143	0.241	0.276	0.002	0.030	0.062
0.4074	0.0700	0.0279	0.4489	0.4532	0.059	0.117	0.023	0.134	0.043	0.072	0.138	0.020	0.053	0.053
0.7055	0.2921	0.0767	0.0019	0.6293	0.044	0.082	0.488	0.479	0.489	0.498	0.091	0.033	0.062	0.047
0.7803	0.0343	0.0090	0.0183	0.9384	0.241	0.245	0.263	0.398	0.285	0.390	0.288	0.019	0.090	0.082

Application to the Acquisition of Fractions Problems

Descriptive Results

Table 3 gives the success rates observed for the three groups of pupils of the experiment described in the first Section. For each pair of problem types, the four indexes H_{ij} have been computed. We adopt here the following conventional criteria. The predictive efficiency of the implication hypothesis $a_1 \gtrapprox b_1$ will be assessed *medium* if $0.20 \leq H_{ij} < 0.60$, *large* if $0.60 \leq H_{ij}$ 0.80, and *very large* if $0.80 \leq H_{ij} \leq 1$. The hypothesis will be considered as unsupported if H_{ij} is less than 0.20. But, bearing in mind the interpretations in terms of proportionate reduction in prediction error, it is clear that such criteria are strongly dependent on the context.

The observed results for each of the three school grades can be summarized by the *descriptive* graphs of "supported" implication hypotheses shown in Figure 1.[5]

Examples of Inferences

Table 4 gives four examples of observed contingency tables with different levels of predictive efficiency for the ninth-grade pupils. In each case the 95% lower and upper limits for all the solutions considered above are reported. The results agree with the simulation study. In particular the need for caution in using asymptotic confidence intervals is exemplified. The *normal* solution can give upper limits greater than one, while the *log normal* solution generally gives lower limits that are too weak. Except for extreme tables the "intermediate" conditional frequentist and Bayesian solutions are relatively close to each other. For Bayesian procedures, it can be verified that varying the prior in the ignorance zone constitutes a sensible and coherent methodology.

A further attractive feature of Bayesian methods is that the joint probability of posterior statements can be obtained in a straightforward way. A particular case of interest is to evaluate the predictive efficiency of *equivalence* hypotheses. For instance, with regard to the equivalence between a_1 and b_1, it must be demonstrated that both η_{11} (for $a_1 \gtrapprox b_1$) and η_{00} (for $a_0 \gtrapprox b_0$, i.e. $b_1 \gtrapprox a_1$) are large. Since η_{00} is also a function of the three independent Beta variables (X,Y,Z), a joint statement can be obtained by simulation. As a numerical illustration, for the "very large" value $H_{11} = 0.914$ in Table 1, the observed index $H_{00} = 0.679$ for the implication hypothesis $a_0 \gtrapprox b_0$ is also relatively large. For instance, for the prior $[\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}]$ the respective 95% lower credibility limits are found: 0.703 for η_{11} and 0.473 for η_{00}, with the joint statement $Pr(\eta_{11} > 0.703$ and $\eta_{00} > 0.473) = 0.908$. Alternatively, a common limit associated with a given probability can be searched for, hence, for instance, $Pr(\eta_{11} > 0.473$ and $\eta_{00} > 0.473) = 0.95$.

TABLE 3
Success rates observed for the three groups of pupils

	FRPW	CQPW	RQPW	FRPP	CQPP	RQPP
5th grade	0.36	0.73	0.24	0.24	0.67	0.01
7th grade	0.56	0.60	0.49	0.38	0.53	0.33
9th grade	0.69	0.82	0.58	0.45	0.78	0.31

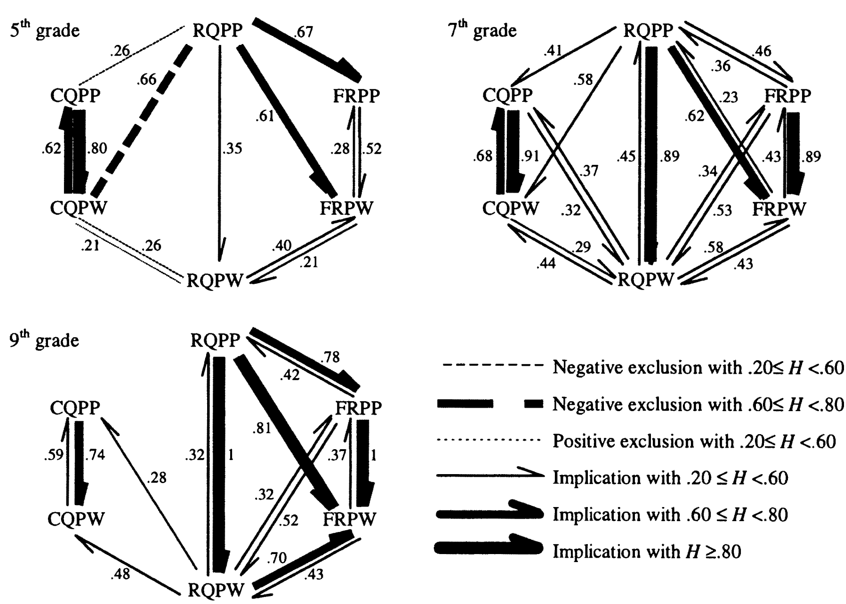

FIGURE 1. *Descriptive implication hypotheses graphs for the three school grades*

Bayesian Analysis

For each implication hypothesis, the lower limit $\underline{\ell}$ such that the posterior probability that η_{ij} is larger than $\underline{\ell}$ is equal to 0.90 has been computed. The corresponding results, obtained for the prior $[\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}]$, can be summarized by the *inductive* graphs of "supported" implication hypotheses shown in Figure 2. Note that no inductive conclusion of "unsupported" implication hypotheses ($\eta_{ij} < 0.20$) can be obtained from the observed indexes H_{ij} less than 0.20. However, when pooling the three age groups, most of the corresponding posterior probabilities $Pr(\eta_{ij} < 0.20)$ are found to be greater than 0.90.

TABLE 4

Four examples of observed contingency tables: 95% confidence and credibility limits for η_{11}

n_{11} n_{10} n_{01} n_{00}	14 31 3 7		20 5 12 18		21 1 5 28		17 0 15 23	
	$H_{11} = 0.003$		$H_{11} = 0.522$		$H_{11} = 0.914$		$H_{11} = 1$	
Procedure:	$\underline{\ell}$	$\overline{\ell}$	$\underline{\ell}$	$\overline{\ell}$	$\underline{\ell}$	$\overline{\ell}$	$\underline{\ell}$	$\overline{\ell}$
asymptotic normal	−0.067	0.072	0.254	0.789	0.779	1.049	/	/
asymptotic log normal	−0.069	0.070	0.163	0.727	0.588	0.982	0.073	0.999
conditional inclusive	−0.085	0.069	0.193	0.768	0.666	0.995	0.641	/
conditional exclusive	−0.053	0.047	0.294	0.692	0.783	0.961	/	0.992
conditional mid	−0.074	0.062	0.231	0.741	0.704	0.989	/	/
conditional Copas & Loeber	−0.071	0.053	0.226	0.714	0.694	0.962	0.705	0.989
Bayesian [0 1 1 0]	−0.099	0.053	0.187	0.709	0.644	0.967	0.633	0.993
Bayesian [1 0 0 1]	−0.059	0.085	0.284	0.784	0.768	0.995	/	/
Bayesian $[\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}]$	−0.079	0.069	0.235	0.747	0.703	0.984	0.757	1.000
Bayesian $[\frac{1}{4}\frac{1}{4}\frac{1}{4}\frac{1}{4}]$	−0.075	0.070	0.244	0.760	0.729	0.990	0.838	1.000
Bayesian [0 0 0 0]	−0.072	0.070	0.254	0.772	0.758	0.995	/	/
Bayesian [1 1 1 1]	−0.085	0.068	0.218	0.724	0.657	0.991	0.653	0.993

Comments

The Bayesian analysis of implication hypotheses between six types of problems associated with the acquisition of fractions has brought out three main results. First, it can be seen that for fifth-grade pupils a success in computing the compared quantity for a Part/Whole relationship (CQPW) often implies a failure in computing the reference quantity for a Part/Part relationship (RQPP). A subsequent analysis of the resolution process has shown that many pupils used the same strategy for both problems. This strategy is efficient for the first type of problems, but irrelevant for the second one (Charron, 2000). This result reveals a conceptual obstacle for younger pupils, that is overcome by older pupils. Second, in order to succeed in a Part/Part task, success in the corresponding Part/Whole task is generally needed. Thus the acquisition of the Part/Whole concept appears to be a prerequisite to the mastery of the Part/Part concept. Finally, the computation of fractions and the computation of reference quantity share many implication relations, particularly for seventh and ninth grades. These links suggest that these computations involve common resolution processes.

References

Altham, P.M.E. (1969). Exact Bayesian analysis of a 2 × 2 contingency table and Fisher's "exact" significance test. *Journal of the Royal Statistical Society B, 31*, 261–269.

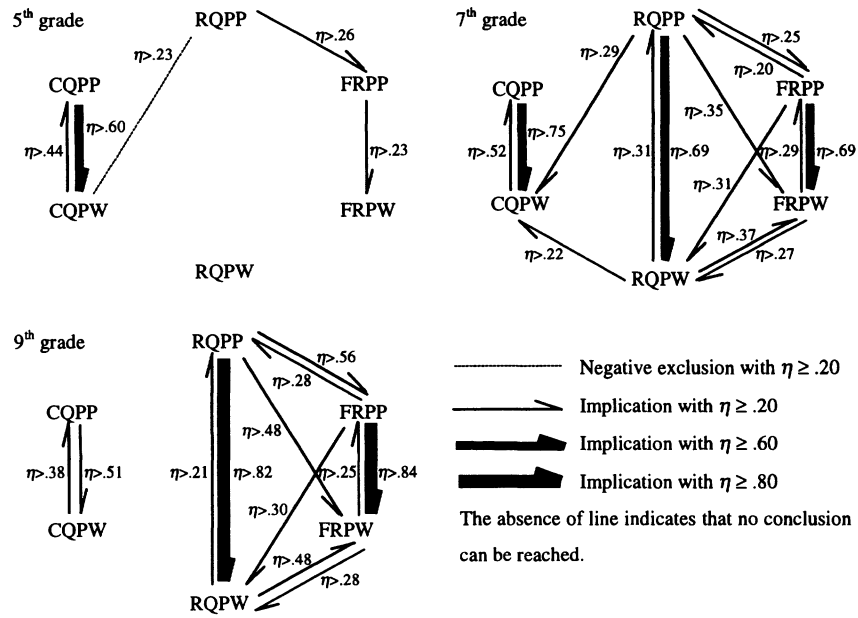

FIGURE 2. *Inductive implication hypotheses graphs for the three school grades*

Bernard, J.-M. (1996). Bayesian interpretation of frequentist procedures for a Bernouilli process. *The American Statistician, 50,* 7–13.

Bernard, J.-M., & Charron, C. (1996). L'analyse implicative bayésienne: Une méthode pour l'étude des differcences orientées. *Mathématiques, Informatique et Sciences Humaines, 134,* 5–38; *135,* 5–18.

Bernardo, J., & Smith, A.F.M. (1994). *Bayesian Theory.* New York: Wiley.

Charron, C. (1995). Individual variations in the construction of categories of problems involving fractions. *Proceedings of the Psychological Mathematical Education Meeting,* Onasbruck (Germany), 7–10.

Charron, C. (1999). Categorization of problems and conceptualization of fractions in adolescents. *European Journal of Psychology of Education,* in press.

Copas, J. B., & Loeber, R. (1990). Relative improvement over chance (RIOC) for 2 x 2 tables. *British Journal of Mathematical and Statistical Psychology, 43,* 293–307.

Cox, D. R. (1970). *The Analysis of Binary Data.* London: Methuen.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions* (2nd ed.). New York: Wiley.

Gefeller, O. (1992). An annotated bibliography on the attributable risk. *Biometrical Journal, 34,* 1007–1012.

Goodman, L. A., & Kruskal, W. H. (1959). Measures of association for cross classifications, II: Further discussions and references. *Journal of the American Statistical Association, 54,* 123–163.

Gras, R., & Larher, A. (1993). L'implication statistique, une nouvelle méthode d'analyse de données. *Mathématiques, Informatique et Sciences Humaines*, *120*, 5–31.

Hildebrand, D. K., Laing, J. D., & Rosenthal, H. (1977). *Prediction Analysis of Cross Classifications*. New York: John Wiley & Sons.

Lecoutre, B., Derzko, G., & Grouin, J.-M. (1995). Bayesian predictive approach for inference about proportions. *Statistics in Medicine*, 14, 1057–1063.

Levin, L. L. (1953). The occurrence of lung cancer in man. *Acta Unio Internationalis contra Cancrum*, *9*, 531–541.

Loeber, R., & Dishion, T. (1983). Early predictors of mail delinquency: A review. *Psychological Bulletin*, *94*, 68–99.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographics*, *61*, 4.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of scale analysis and factor analyses. *Psychological Bulletin*, *45*, 507–529.

Nijsse, M. (1992). Attributable risk and relative improvement over chance: A note on Copas & Loeber (1990). *British Journal of Mathematical and Statistical Psychology*, *45*, 329–332.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.

Perks, F.J.A. (1947). Some observations on inverse probability including a new indifference rule (with discussion). *Journal of the Institute of Actuaries*, *73*, 285–334.

Rice, W. R. (1988). A new probability model for determining exact P value for 2 × 2 contingency tables. *Biometrics*, *44*, 1–22.

Spiegelhalter, D. J., Freedman, L. S., & Parmar, M.K.B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society A*, *157*, 357–416.

Toecher, K. D. (1950). Extension of the Neyman-Pearson theory of tests to discontinuous variables. *Biometrika*, *37*, 130–144.

Vergnaud, G. (1983). Multiplicative structures. In R. Lesh and M. Landau (Eds.), *Acquisition of Mathematics Concepts and Processes*, New York: Academic Press, 127–174.

Walley, P. (1996). Inferences from multinomial data: Learning about a bag of marbles (with discussion). *Journal of the Royal Statistical Society B*, *58*, 3–57.

Walter, S. D. (1975). The distribution of Levin's measure of attributable risk. *Biometrika*, *63*, 371–374.

Walter, S. D. (1976). The estimation and interpretation of attributable risk in health research. *Biometrics*, *32*, 928–849.

Yule, G. U. (1900). On the association of attributes in statistics. *Philosophical Transactions of the Royal Society of London, Series A*, *54*, 257.

Notes

[1] A ^{TM}Windows interactive computer program, "LesImplications" (B. Lecoutre and J. Poitevineau), which calculates the probability statements involved in this paper is available upon request to the first author.

[2] The RIOC index, as defined by the authors, is in fact equal to H_{11} if $f_{.1} \geq f_{1.}$ and otherwise to H_{00}, which is again identical to Loevinger's (1947, 1948) coefficient of homogeneity.

[3] In this context other terms and interpretations have been proposed. References can be found in Gefeller (1992).

[4] Hildebrand, Laing, and Rosenthal (1977, p. 202) gave a close formula, but with $n-1$ instead of n in the denominator.

[5] The notion of implication hypotheses graphs used here has its origins in the work of Gras and Larher (1993).

Authors

BRUNO LECOUTRE is Directeur de Recherche C.N.R.S. at UPRESA 6085, Analyse et Modèles Stochastiques, C.N.R.S. and Université de Rouen, Mathématiques Site Colbert, 76821 Mont-Saint-Aignan Cedex, France; bruno.lecoutre@univ-rouen.fr. He specializes in experimental data analysis and Bayesian statistics.

CAMILO CHARRON is Maître de Conférences, at IUFM de Bretagne, Site de formation de Saint-Brieuc, 1 rue Théodule Ribot, BP 4502, 22045 Saint-Brieuc Cedex 2, France. He specializes in learning and developmental psychology. This research was carried out when he was at LaPsyDé, C.N.R.S. and Université René Descartes, Paris, France.