Bayesian Evaluation of Non-Admissible Conditioning: The Case of Fisher Test

Article · January 1998

Source: RePEc

2 authors:

Michel G.M. Mouchart
Université Catholique de Louvain - UCLouvain
111 PUBLICATIONS **988** CITATIONS

SEE PROFILE

Eliana Scheihing
Universidad Austral de Chile
27 PUBLICATIONS **37** CITATIONS

SEE PROFILE

Bayesian Evaluation of Non-Admissible Conditioning: The Case of Fisher Test

Michel MOUCHART[1] **and Eliana SCHEIHING**[2]

March 17, 1998

Abstract

We first analyse the general problem of admissible conditioning and next consider the evaluation of the loss of information when a non-admissible conditioning is used as an approximation of the exact posterior distribution. Considering the case of Fisher test, we evaluate from a Bayesian point of view how much information is lost when the sampling process for a 2x2 contingency table is analysed conditionally on the two margins. This loss of information due to non-admissible conditioning is evaluated for different sampling models and with respect to the entropy divergence and to the Hellinger distance between the exact and the approximate posterior distributions and with respect to relative risks based on a quadratic loss function. The numerical results obtained through simulation indicate that for a specific range of parameters the loss of information increases with the sample size and decreases with the precision of the a priori distribution. Hence such an approximation is shown to be a non-asymptotic one.

Keywords: Approximate Bayesian Solutions, Admissible conditioning, Contingency tables, Fisher Test.
Acknowledgements : Research supports from the contract "Projet d'Actions de Recherche Concertées" No. 93/98-164, from the Belgian Program on Interuniversity Poles of Attraction (No. 26), from the Belgian OSTC Transport and Mobility (contract TRB3/007), from the Departmental fellowship of UCL and from DID-UACh are gratefully acknowledged. This version has benefited from many useful discussions with J.M. Rolin and E. San Martín.

[1] CORE and Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.
[2] Instituto de Informática, Universidad Austral de Chile, Valdivia, Chile

1 Introduction

This paper revisits an important problem in statistical inference, namely the loss of information in conditional inference. More specifically, let us consider that the data take the form of a multivariate X, to be analysed by means of a statistical model characterized by a family of sampling probabilities $\{ p(x \mid \theta) : \theta \in \Theta \}$. Let us assume that $\lambda = l(\theta)$ is a parameter of interest and that $\psi = f(\theta)$ represents a nuisance parameter such that θ is a function of (λ, ψ). Thus (λ, ψ) is a reparametrization of θ. Suppose now that we can construct a statistic $Z = g(X)$ such that the likelihood function $L_X(\theta)$ factorizes into:

$$L_X(\theta) = L_Z(\lambda, \psi) \cdot L_{X|Z}(\lambda) \qquad (1)$$

i.e. the conditional sampling distribution, given Z, depends on λ only whereas the marginal sampling distribution of Z depends on λ and ψ. Thus λ is a sufficient parameter for the conditional sampling process. Dawid(1980) would say that Z is "$\lambda-$inducing".

In the framework of the likelihood principle, a natural question is: Under which conditions may we forget the first term of the likelihood function, namely $L_Z(\lambda, \psi)$? When this is allowed this would mean that the data would be treated "as if" Z were not random. Some statisticians would then say that Z is considered as exogenous for the inference on λ.

This situation is rather frequent in statistics. Think, for instance, of the sampling distribution of S^2 independent of μ in the normal $N(\mu, \sigma^2)$ case (where the sampling distribution of $Z = \bar{X}$, independent of S^2, depends on both μ and σ^2) used for tests on σ^2 (see e.g. Barnard (1963)) or of likelihoods conditional on the p individual averages $Z = (\bar{X}_1, \cdots, \bar{X}_p)$ in fixed individual effect models as used for panel data (see e.g. Chamberlain(1980)).

Note also that a dual situation arises when the likelihood is factorizable as

$$L_X(\theta) = L_Z(\lambda) L_{X|Z}(\psi, \lambda) \qquad (2)$$

In such a case, a natural question is under which condition are allowed to simplify the likelihood function by considering only the first term, "forgetting", by so-doing, the information contained, say in Y when X has the form $X = (Y, Z)$?. An example of such a simplification may be found in Cox'estimator in the semi-parametric proportional hazard model which is based on the marginal distribution of the rank statistic. (see e.g. Cox(1972)).

Both situations when one wants to simplify the likelihood function by reduction to only one of the two factors, raise substantial problems for statistical inference. A frequently used, and heuristic, argument for justifying such a reduction runs around the idea that when ψ is a complex and/or large enough parameter, the statistical information contained in the term depending on ψ and λ would be "swallowed" by ψ leaving (almost) no residual information on λ. See for instance Section 10.8 in Kalbfleisch(1985) for an argumentation. In such a case, Z would be "approximately" ancillary (in (1) or sufficient (in (2)) for the inference on λ, and Barnard (1963) would add "in the absence of information on ψ".

In this paper we want to develop a general methodology (previously sketched in Mouchart and Scheihing (1993)) for evaluating such situations. The basic idea is to consider the inference on λ, the parameter of interest, based on the reduced likelihood as approximation to an exact inference based on the complete likelihood. The general approach is Bayesian; thus the inference on λ takes the form either of a posterior distribution (approximate or exact) or of a risk function of a Bayesian estimator relative to a specific loss function and to a posterior distribution (again, approximate or exact). This paper bears on methodology rather than on abstract theory. Thus we keep the paper within .reasonable size by treating explicitly the problem of conditioning only. Also we use Fisher test, i.e a test conditional on the two margins in a 2×2 contingency table, as a case study to exemplify how much information is lost in an approximate inference, based on $L_{X|Z}(\lambda)$ only, as compared to an exact inference, based on $L_X(\theta)$.

Let us now expose the main questions motivating the approach of this paper. A first question is to decide what is meant by "losing no information on λ". We face this question in the framework of the theory of reduction of Bayesian experiments in which a conditional model is a reduction of a complete model and a reduction is admissible if it leads to the same posterior distribution on the parameter of interest, namely λ, as the complete model. Note that this concept of admissibility (for inference) generalizes the usual concept of admissibility (in a decision-theoretic framework) in the sense that inferential admissibility implies decisional admissibility for any loss function, but inferential admissibility refers to the reductions of statistical experiments rather than to statistical procedures defined on a statistical experiment. For more details see, e.g. Florens, Mouchart and Rolin (1990a) and Heyer(1982).

A second question concerns the statement "in the absence of prior information on ψ". We face this question by taking a Bayesian approach with the idea that introducing explicitly a prior information on the parameters is a pragmatic way of evaluating the possible role of prior information, a so-called non-informative or reference prior providing a standard of comparison as a limit situation.

A third question concerns the role of the sampling model in the eventual non-admissibility of conditioning. Thus, in the case of a 2x2 contingency table, we first analyse two standard models, namely a multinomial one, conditional on the sample size, and a double independent binomial one conditional on the column margins, to show that in such cases the conditioning on the two margins is not admissible. We also show that if one is willing to insist that conditioning on the two margins does not lose information, the meaning of the parameter of interest may be deeply affected.

The last question we want to consider is the following. Once a conditioning is recognized as non-admissible, a natural question is "how much information is lost by a non-admissible conditioning?". We face this question by suggesting two kinds of measures. One is more inference-oriented and uses the predictive expectation of a divergence or distance between the exact and the approximate posterior distributions. The other one is more decision-oriented and compares the optimal risk functions based on the exact and on the approximate posterior distributions respectively.

Reasons for non-admissible conditioning may be numerous in actual statistical practice. One may be operationality: the exact posterior distribution may be much more complex than the approximate one and the available resources for computing can make the exact solution unavailable or too expensive, also the dimensionality of ψ may render the economy of specifying the prior distribution particularly attractive. Another one may be robustness against specification errors: inference, based on a partial specification of the data generating process, may be preferred to a more specific one. In all these cases, it is important to gain a better understanding of the nature of the approximation and to know what are the main factors explaining the quality of this approximation. Designing such a method is indeed the main objective of this paper.

These questions are first examined analytically both in general terms and in the particular case of conditioning on the margins of a 2×2 contingency table. Next we try to gain some quantitative insight on these questions by conducting a numerical experiment. In a companion paper, Mouchart and Scheihing (1994), we give details on the simulation study, namely on the design of the simulation, the numerical difficulties and the results. There we obtain two types of conclusions relative to the 2x2 contingency table. One is that the numerical problems raised by the comparison of risk functions are substantially more complex than those involved in the comparison of posterior distributions. Secondly the approximation relative to the conditioning on the two margins is a small sample one, in the sense that it improves when the prior information increases but deteriorates when the sample size increases.

The paper is organized as follows. In the next section, we analyse the problem of admissible conditioning in general terms and the evaluation of the loss of information when a non-admissible conditioning is specified. In section 3 we consider the problem in the particular case of the 2x2 contingency table, with special emphasis on the role of the model specification. Section 4 summarizes the results of a simulation study. Section 5 spells out some difficulties to be faced when insisting on the admissibility of conditioning on the margins. We present some closing remarks in Section 6. An appendix gives details on some analytical derivations.

2 Admissible Conditioning in Bayesian Methods

2.1 Exact Admissibility

In this section we briefly present the Bayesian approach to the admissibility of a reduction by conditioning. This is the question, whether it is admissible in the sense of losing no information, to specify the sampling process only partially by treating part of the data "as if" it were not random. For the sake of completeness, we present here a short summary of a topic treated more systematically in chapter 3 of Florens, Mouchart and Rolin (1990a), see also, Florens and Mouchart

(1977,1985) and Dawid (1980).

As usual, we start by considering that a Bayesian model may be viewed as unique probability measure jointly on the parameters and the observations. If we write x for the observations and θ for the parameters, the Bayesian model may accordingly be represented in terms of densities, with respect to suitable measures, as follows:

$$q(x,\theta) = m(\theta)\, p(x \mid \theta) = p(x)\, m(\theta \mid x)$$

where $m(\theta)$ is assumed to be a density of a probability measure. This assumption is made to avoid marginalization paradoxes when integrating out nuisance parameters with respect to improper prior distributions: such a theme would not fall in the scope of this paper.

We decompose both the observations x and the parameter θ into two components: $x = (y, z)$ and $\theta = (\psi, \lambda)$, where ψ is a nuisance parameter and λ is a parameter of interest, that is the underlying loss function depends on λ only. Therefore we only have interest in the posterior distribution of λ which may be decomposed as:

$$m(\lambda \mid x) \quad \propto \quad m(\lambda)\, p(z \mid \lambda)\, p(y \mid z, \lambda) \tag{3}$$

We are interested in the following question: how far can we forget the term $p(z \mid \lambda)$ for the inference on λ (i.e. without modifying $m(\lambda \mid x)$?

Let us introduce three sufficient conditions in a decreasing order of generality.

If z and λ are *mutually ancillary*, that is z and λ are independent in probability:

$$z - \lambda \tag{4}$$

so that $p(z \mid \lambda) = p(z)$ and $m(\lambda \mid z) = m(\lambda)$, we have:

$$m(\lambda \mid x) \quad \propto \quad m(\lambda)\, p(y \mid z, \lambda) \tag{5}$$

Clearly, if z is ancillary, i.e. $z - \theta$, z is also mutually ancillary with any function of θ, but the converse is evidently false.

Note however that the use of formula (5) implies integrating the conditional sampling distribution of $(y \mid z)$ with respect to the conditional posterior distribution of ψ given z and λ, unless λ is a sufficient parameter of the conditional sampling distribution of $(y \mid z)$, that is

$$y - \psi \mid \lambda, z \tag{6}$$

in which case $p(y \mid z, \lambda) = p(y \mid z, \psi, \lambda)$ for any prior distribution. When both (4) and (6) are verified, λ and z are said to be *mutually exogenous*. Dawid (1980) would say that "z is D-ancillary for λ".

When λ is a sufficient parameter of the sampling distribution conditional on z, a rather natural way of obtaining the mutual exogeneity of λ and z, thus a stronger but also more operational condition than the mutual ancillarity condition, is through the structure of a *Bayesian cut* which means that along with (6), ψ is furthermore a sufficient parameter of the sampling marginal process of z, that is

$$z - \lambda \mid \psi \tag{7}$$

and is also, a priori independent of λ:

$$\psi - \lambda \tag{8}$$

this is so simply because (7) and (8) trivially imply (4). The advantage of that structure is that the two conditions ((6) and (7)) only depend on the sampling process and the last condition (8) only depends on the prior specification so that the structure of a Bayesian cut is robust with respect to any modification of the prior specification, provided it keeps the prior independence of ψ and λ. In a sampling theory framework, the corresponding concept of cut, see e.g. Barndorff-Nielsen (1978), replaces the prior independence between ψ and λ by the condition of being variation-free, that is the factorization of the parameter space: $(\psi, \lambda) \in \Psi \times \Lambda$. Note however that the two requirements of sufficiency of ψ and λ for the marginal (7) and the conditional (6) sampling processes respectively, are restrictive only under some condition of independence between ψ and λ (namely, prior independence or variation-free); otherwise these conditions could be trivial if the fact that λ be function of ψ is not excluded.

In summary, the three nested conditions of mutual ancillarity (4), mutual exogeneity ((4) and (6)) and Bayesian cut ((6),(7) and (8)) are each sufficient for ensuring the admissibility, for the inference on λ, of the conditioning on z. Note also that for a given sampling process $p(z \mid \psi, \lambda)$, the condition (4) crucially depends on the particular prior of the conditional prior distribution $m(\lambda \mid \psi)$, unless z were ancillary. Condition (6) is a property of the sampling process, under the condition (1) of the D.G.P. whereas for the Bayesian cut, under the conditions (6) and (7) of the sampling process, the prior specification is involved only in its prior independence structure but not in its particular analytical form.

2.2 Approximate Admissibility

Suppose now that z and λ are not mutually ancillary but that one decides, for some (good or bad) reason to "forget" $p(z \mid \lambda)$ in equation (3), that is using (5) as an "approximation" of (3). Thus let us write $m_A(\lambda \mid x)$, as an approximation to $m(\lambda \mid x)$, defined as:

$$m_A(\lambda \mid x) \propto m(\lambda)\, p(y \mid z, \lambda)$$

without having $z - \lambda$. A natural question is now "how much information do we lose by overlooking the term $p(z \mid \lambda)$?"

We will consider two ways of evaluating this loss of information. The first one, that is more inference-oriented, is based on the evaluation of the predictively expected divergence or distance between the posterior distribution of λ, $m(\lambda \mid x)$, and its approximation, $m_A(\lambda \mid x)$, more specifically one may measure a loss of information L as follows:

$$L = E[D(m_A(\lambda \mid x),\, m(\lambda \mid x))] = \int D(m_A(\lambda \mid x),\, m(\lambda \mid x))\, p(x)\, dx \tag{9}$$

where $D(\ ,\ )$ is a divergence (or a distance) between probability measures.

Note that $D(m_A(\lambda \mid x), m(\lambda \mid x))$ is a statistic, that is a function of the data only, and the expectation is taken accordingly with respect to the predictive distribution of the exact model. For the sake of interpretation it may be more convenient to use a divergence (or a distance) bounded in the interval [0,1].

The second way, that is more decision-oriented, compares the risk functions corresponding to the optimal decision rules under the exact posterior distribution of λ and under the approximate posterior distribution of λ. More specifically, consider a loss function $l(\lambda, a)$. Under the exact posterior distribution of λ, the optimal decision rule and the corresponding Bayesian risk function are:

$$\begin{aligned} a_E^*(x) &= arg \inf_a E[\, l(\lambda, a) \mid x\,] \\ \rho(a_E^*) &= E[\, l(\lambda, a_E^*(x))\,] \end{aligned}$$

Similarly, under the approximate posterior distribution of λ, the optimal decision rule and the corresponding Bayesian risk function are:

$$\begin{aligned} a_A^*(x) &= arg \inf_a E^A[\, l(\lambda, a) \mid x\,] \\ \rho(a_A^*) &= E[\, l(\lambda, a_A^*(x))\,] \end{aligned}$$

where

$$E^A[\, l(\lambda, a) \mid x\,] = \int l(\lambda, a)\, m_A(\lambda \mid x)\, d\lambda$$

Clearly

$$0 \leq \rho(a_E^*) \leq \rho(a_A^*)$$

Therefore we propose as a measure of the loss of information the following comparison between the two risk functions, which may be read as the "Relative Efficiency of the approximation m_A":

$$RelEff(m_A) = \frac{\rho(a_E^*)}{\rho(a_A^*)} \;\in\; [0, 1]$$

The value $\frac{1}{k}$ of this ratio means that the risk of a_A^* is k times greater than the risk of a_E^*.

As a particular case, consider the squared loss function $l(\lambda, a) = (a - \lambda)^2$; in such a case:

$$\begin{aligned} a_E^*(x) &= E(\lambda \mid x) \\ a_A^*(x) &= E^A(\lambda \mid x) \end{aligned}$$

and the corresponding risk functions are:

$$\begin{aligned}
\rho(a_E^*) &= E[\, var(\lambda \mid x)\,] \\
\rho(a_A^*) &= E[\,\{E^A(\lambda \mid x) - \lambda\}^2\,] \\
&= E[\, var(\lambda \mid x) + \{\, E(\lambda \mid x) - E^A(\lambda \mid x)\,\}^2\,]
\end{aligned}$$

Therefore:

$$RelEff(m_A) = \left(1 + \frac{E[\,\{\, E(\lambda \mid x) - E^A(\lambda \mid x)\,\}^2\,]}{E[\, var(\lambda \mid x)\,]}\right)^{-1} \tag{10}$$

which shows that under a quadratic loss function, an approximation will be evaluated according to the ratio of the (predictively) expected square bias of the approximate posterior expectation and the (predictively) expected exact posterior variance.

3 Conditioning on the Two Margins

3.1 Introduction

In this section we use the exact Fisher test as a case study for the general theory developped in Section 2. Thus we now consider that X, the data, are in the form of a 2x2 contingency table, *viz.* $X = (X_{11}\, X_{12}\, X_{21}\, X_{22})$. We shall make use of the dot notation for the summation, for instance $X_{1.} = X_{11} + X_{12}$, or $X_{..} = X_{11} + X_{12} + X_{21} + X_{22}$. We want to examine the question how far it is admissible to conduct the inference conditionally on the margins, that is, in the notation of previous section, to assume the exogeneity of $Z = (X_{1.}, X_{.1})$. In a sampling-theory framework, this problem has received considerable attention for being considered as a basic example of substantial methodological issues, see e.g. Fisher (1935a,b), Plackett(1974, 1977), Haber(1989) and for a recent survey Agresti(1992). It should be pointed out, in particular, that Plackett (1977) enlarges the problem of the Fisher exact test and considers the more general issue of the inference on λ, the parameter characterizing the non-independence. He evaluates the information on λ contained in the likelihood function of the marginal totals by looking at its maximizing value. He accordingly concludes that the marginal totals contain no information on λ by examining the solution to the normal equation only, namely $\lambda = 0$ or $\lambda = \pm\infty$.

In a Bayesian framework, Altham(1969, 1971) and Lindley(1964), among others, paid attention to the same problem. This literature shows in particular that careful attention should be paid to the specification of the sampling process and to the meaning of the parameter of interest. We shall deal successively with these problems.

It is usual, in the analysis of the contingency tables, to exogeneize the grand total $X_{..}$ on the ground of a non-informative stopping rule for the sample size. Formally, this may be obtained as follows. Let α_0 and θ be sufficient parameters for the sampling distribution of $X_{..}$ and of $(X \mid X_{..})$ respectively, that is

$p(X \mid \theta, \alpha_0) = p(X_{..} \mid \alpha_0)p(X \mid X_{..}, \theta)$. In the sequel we shall always assume that $(\alpha_0, \theta, X_{..})$ operates a Bayesian cut on the model, that is θ — α_0, implying therefore that once the parameter of interest is a function of θ, the inference may proceed conditionally on $X_{..}$. In other words, in what follows, we never specify the way the sample size is generated, we only assume that any unknown parameter embodied in this process is a priori independent of the parameter characterizing the process generating the table conditionally on the sample size. We shall see later that this assumption might be less innocent than first appearance might suggest.

In order to evaluate the exogeneity of the margins $(X_{1.}, X_{.1})$ in the process conditional on $X_{..}$ one has to be more specific on the sampling process and on the parameter of interest. In what follows, the parameter of interest, denoted $\lambda = l(\theta)$, is a parameter characterizing the statistical association between the row and column criteria, a particular value of which would imply independence. Thus, at variance with the exact Fisher test, we are interested in the complete posterior distribution of λ rather than in the more simple question whether λ is equal or different from a particular value, say, 1. We now consider successively two sampling alternatives conditional on $X_{..}$.

3.2 Multinomial sampling

Suppose that $(X_{ij})_{1 \leq i,j \leq 2}$ conditionally on $X_{..}$ is generated by a multinomial sampling distribution:

$$p(X = x \mid X_{..} = n, \theta) = \frac{n!}{\displaystyle\prod_{1 \leq i,j \leq 2} x_{ij}!} \prod_{1 \leq i,j \leq 2} \theta_{ij}^{x_{ij}}$$

The parameter of interest is defined as the usual cross product ratio:

$$\lambda = \frac{\theta_{11}\theta_{22}}{\theta_{12}\theta_{21}} \quad , \tag{11}$$

the value 1 of which characterizes the row-column independence.

In order to evaluate the admissibility of the conditioning on $Z = (X_{1.}, X_{.1})$ for the inference on λ, we notice that $(X_{11}, Z, X_{..})$ characterizes the table. The sampling distribution conditional on Z and $X_{..}$ boils therefore down to the sampling distribution of $(X_{11} \mid Z, X_{..})$, a non central hypergeometric distribution of parameter λ, see e.g. Agresti (1992), Cornfield (1956), Fisher (1935a):

$$p(X_{11} = x_{11} \mid x_{1.}, x_{.1}, x_{..}, \theta) = \frac{\lambda^{x_{11}}}{\displaystyle\prod_{i,j} x_{ij}!} \left(\sum_{x_{11} \in V} \frac{\lambda^{x_{11}}}{\displaystyle\prod_{i,j} x_{ij}!} \right)^{-1} \tag{12}$$

$$= \lambda^{x_{11}} \begin{pmatrix} x_{1.} \\ x_{11} \end{pmatrix} \begin{pmatrix} x_{..} - x_{1.} \\ x_{.1} - x_{11} \end{pmatrix} \left[\sum_{u \in V} \lambda^u \begin{pmatrix} x_{1.} \\ u \end{pmatrix} \begin{pmatrix} x_{..} - x_{1.} \\ x_{.1} - u \end{pmatrix} \right]^{-1} \qquad x_{11} \in V$$

where

$$V = \{x \in \mathbb{N} \mid max\{0, x_{1.} + x_{.1} - x_{..}\} \leq x \leq min\{x_{1.}, x_{.1}\}\}$$

Thus λ is a sufficient parameter for the distribution of the table conditional on the grand total $X_{..}$ and the margins $Z = (X_{1.}, X_{.1})$:

$$X \text{ — } \theta \mid \lambda, Z, X_{..} \tag{13}$$

However, conditionally on $X_{..}$, we cannot construct ψ such that (ψ, λ, Z) operates a Bayesian cut on the model because the sampling distribution of the margins identifies *all* the parameters θ. This feature may indeed be verified by means of the following reparametrization:

$$\begin{array}{rcl} g : \Theta & \rightarrow & \Psi \times \mathbb{R}^+ \\ (\theta_{22}, \theta_{21}, \theta_{12}) & \rightarrow & (\psi, \lambda) \end{array} \tag{14}$$

where $\psi_1 = \theta_{22}$, $\psi_2 = \frac{\theta_{21}}{\theta_{21}+\theta_{22}}$, λ is defined by (11) and therefore

$$\begin{array}{rcl} \Theta & = & \{(\theta_{22}, \theta_{21}, \theta_{12}) \in (0,1)^3 \mid \theta_{22} + \theta_{21} + \theta_{12} < 1\} \\ \Psi & = & \{\psi = (\psi_1, \psi_2) \in (0,1)^2 \mid \psi_1 + \psi_2 < 1\} \end{array}$$

The sampling distribution of $(X_{1.}, X_{.1} \mid X_{..})$ may accordingly be written as:

$$\begin{array}{rcl} p(Z = (x_{1.}, x_{.1}) \mid X_{..}, \psi, \lambda) & = & X_{..}! \displaystyle\sum_{x_{11} \in V} \frac{\lambda^{x_{11}}}{\displaystyle\prod_{i,j} x_{ij}!} \\ & \cdot & \left(\dfrac{1 - \psi_1 - \psi_2}{1 - \psi_2 + \lambda\psi_2}\right)^{x_{1.}} \left(\dfrac{\psi_2}{1 - \psi_2}\right)^{x_{.1}} \psi_1^{x_{2.}} \end{array} \tag{15}$$

see also Plackett (1974). Therefore θ, or equivalently (ψ, λ), is a minimal sufficient parameter of the process generating $(Z \mid X_{..})$ and no prior distribution can satisfy the condition of a conditional Bayesian cut, namely $\lambda \text{ — } \theta \mid X_{..}$, because λ is a function of θ.

As the Bayesian cut is not possible, one may ask whether there exists some prior distribution that would render Z and λ to be mutually ancillary conditionally on $X_{..}$, that is:

$$Z \text{ — } \lambda \mid X_{..} \tag{16}$$

which means that

$$p(Z \mid \lambda, X_{..}) = \int p(Z \mid X_{..}, \psi, \lambda)\, m(\psi \mid \lambda) d\psi \tag{17}$$

would not depend on λ, where $p(Z \mid X_{..}, \psi, \lambda)$ is given in (15).

The relevance of this question may be appreciated from the following arguments. Note first that in such a case, z and λ become mutually exogenous conditionally on $X_{..}$ because (13) holds for any prior specification. Dawid (1980)

Example 7.11, gives an example of mutual exogeneity without a Bayesian cut because of a lack of prior independence between, in our notation, ψ and λ. Florens et al.(1990a) Proposition 3.4.7(i) and Theorem 3.4.6 shows that a Bayesian cut is equivalent to mutual exogeneity along with the posterior independence of ψ and λ.

Let us consider that θ is distributed a priori as a Dirichlet distribution with parameter $a = (a_{ij})$ that is

$$m(\theta) = \frac{?\,(a_{..})}{\prod_{i,j} ?\,(a_{ij})} \prod_{i,j} \theta_{ij}^{a_{ij}-1} \mathbf{1}_{\{\theta \in \{(\theta_{ij}) \in [0,1]^4 | \sum \theta_{ij} = 1\}}$$

It is shown, in Proposition 1 of the appendix, that ψ and λ are not a priori independent and that:

$$p(Z = (x_{1.}, x_{.1}) \mid \lambda, X_{..}) = c\, X_{..}!\, \frac{E[(1 - Y + \lambda Y)^{-(a_{1.} + x_{1.})}]}{E[(1 - W + \lambda W)^{-a_{1.}}]} \sum_{x_{11} \in V} \frac{\lambda^{x_{11}}}{\prod_{i,j} x_{ij}!} \qquad (18)$$

where

$$\begin{aligned} c &= \frac{?^{\,2}(a_{..})}{?^{\,2}(a_{..} + x_{..})} \frac{?\,(a_{1.} + x_{1.})?\,(a_{2.} + x_{2.})?\,(a_{.1} + x_{.1})?\,(a_{.2} + x_{.2})}{?\,(a_{1.})?\,(a_{2.})?\,(a_{.1})?\,(a_{.2})} \\ Y &\sim \beta(a_{.1} + x_{.1}, a_{.2} + x_{.2}) \\ W &\sim \beta(a_{.1}, a_{.2}) \end{aligned}$$

For a fixed $X_{..}$, let us write, for simplicity, $f(\lambda, a) = P(Z \mid \lambda, X_{..})$, as given in (18). Clearly $f(0, a)$ and $f(1, a)$ represents non degenerate distributions of Z. Furthermore

$$f(1, a) = cX_{..}! \sum_{x_{11} \in V} \frac{1}{\prod_{i,j} x_{ij}!}$$

does not depend on a whereas

$$f(0, a) = c \frac{X_{..}!}{x_{.1}! x_{1.}! (X_{..} - x_{.1} - x_{1.})!} \frac{E[(1 - Y)^{-(a_{1.} + x_{1.})}]}{E[(1 - W)^{-a_{1.}}]}$$

does depend on a. Thus, because $f(\lambda, a)$ is continuous in λ on $I\!\!R_+$, there is no choice of a which makes $f(\lambda, a)$ independent of λ. The conclusion is therefore that under multinomial sampling, conditioning on the margins is generally not an admissible reduction.

3.3 Independent Binomial Samplings

In this case, the 2x2 contingency table is generated by two independent binomial samplings corresponding, say, to each column. More specifically, we again assume the first cut making $X_{..}$ exogenous for the inference on θ, the parameter sufficient for the process generating the 2x2 contingency table, conditionally on $X_{..}$. Let us decompose now θ as $\theta = (\alpha, \beta)$, where α is the parameter sufficient for the process generating $(X_{.1} \mid X_{..})$, that is:

$$X_{.1} - \theta \mid \alpha, X_{..} \qquad (19)$$

and β is the parameter sufficient for the process generating the table conditionally on $(X_{.1}, X_{..})$, that is:

$$(X_{11}, X_{12}) - \theta \mid \beta, X_{.1}, X_{..} \qquad (20)$$

Finally, we shall also assume the prior independence of α and β:

$$\alpha - \beta \qquad (21)$$

or, equivalently under the first cut

$$\alpha - \beta \mid X_{..} \qquad (22)$$

These hypotheses ensure that $(\alpha, \beta, X_{.1})$ operates a Bayesian cut on the model conditionally on $X_{..}$. Thus for an inference on any function of β, $(X_{.1}, X_{..})$ may be considered as exogenous. The first cut, along with (20) and (21), also implies $\beta - (X_{.1}, X_{..})$. The independent sampling of columns may now be written as:

$$X_{11} - X_{12} \mid X_{.1}, X_{..}, \theta$$

Let now $\beta = (\beta_1, \beta_2)$ where β_j, $j = 1, 2$ are the sufficient parameters for the process generating the jth column conditionally on its total, more specifically the β_j's are defined by the conditions:

$$X_{1j} - \theta \mid X_{.1}, X_{..}, \beta_j \qquad j = 1, 2$$

Finally, we assume that each column is generated as an independent binomial sampling:

$$(X_{1j} \mid X_{.j}, X_{..}, \theta) \quad \sim \quad Bi(X_{.j}, \beta_j) \qquad j = 1, 2$$

that is

$$P(X_{1j} = x_{1j} \mid X_{.j}, X_{..}, \theta) = \left(\begin{array}{c} X_{.j} \\ x_{1j} \end{array} \right) \beta_j^{x_{1j}} (1 - \beta_j)^{X_{.j} - x_{1j}} \qquad j = 1, 2$$

where $X_{.2} = X_{..} - X_{.1}$.

Remark: If furthermore $(X_{.1} \mid X_{..}, \alpha)$ were also assumed to follow a binomial distribution, then $(X \mid X_{..}, \theta)$ would also follow a multinomial distribution as in the previous section, but this is not necessary for what follows.

The parameter of interest is now defined as

$$\lambda = \frac{\beta_1(1-\beta_2)}{\beta_2(1-\beta_1)} \tag{23}$$

the value $\lambda = 1$, meaning $\beta_1 = \beta_2$, is often read as a condition of "homogeneity" and implies, again, a row-column independence of the 2x2 table for whatever process generating the column totals, that is for any distribution of $(X_{.1} \mid X_{..}, \alpha)$.

The problem is again to check whether the inference on λ, a function of β only, may be done, without loss of information, conditionally on the two totals $Z = (X_{.1}, X_{1.})$. We follow the same route as in the previous section and first check whether the sampling distribution generating $(X_{11} \mid Z, X_{..}, \theta)$ depends on λ only. This is easily seen by noticing that the sampling distribution of the 2x2 table conditionally on $(X_{.1}, X_{..})$ is an independent product of two binomial distributions with parameters $(\beta_1, X_{.1})$ and $(\beta_2, X_{..} - X_{.1})$ respectively, from which one concludes that the sampling distribution of the 2x2 table conditionally on $(Z, X_{..})$ is again a non-central hypergeometric distribution of parameter λ, exactly as in (12); thus condition (13) is again satisfied even if the column total $(X_{.1} \mid X_{..})$ is not binomially generated.

Conditionally on $(X_{.1}, X_{..})$, we cannot construct ψ such that (ψ, λ, Z) operates a Bayesian cut on the model. Indeed, the sampling distribution generating $(X_{1.} \mid X_{.1}, X_{..})$ depends on the full vector $\beta = (\beta_1, \beta_2)$. This feature may indeed be verified by means of the reparametrization $(\beta_1, \beta_2) \rightarrow (\lambda, \beta_2)$, where λ is defined by (23). The sampling distribution of $(X_{1.} \mid X_{.1}, X_{..})$ may accordingly be written as:

$$\begin{aligned} p(X_{1.} = x_{1.} \mid X_{..}, X_{.1}, \beta_2, \lambda) &= X_{.1}! X_{.2}! \sum_{x_{11} \in V} \frac{\lambda^{x_{11}}}{\prod_{i,j} x_{ij}!} \left(\frac{1-\beta_2}{1-\beta_2+\lambda\beta_2} \right)^{X_{.1}} \\ &\cdot \quad \beta_2^{x_{1.}} (1-\beta_2)^{X_{..} - X_{.1} - x_{1.}} \end{aligned} \tag{24}$$

Thus, the distribution of the margins, $Z = (X_{1.}, X_{.1})$ (given the grand total $X_{..}$) identifies both β_1 and β_2, so that $\theta = (\alpha, \beta)$ is a minimal sufficient parameter of the process generating $(Z \mid X_{..})$ and no prior distribution can satisfy the condition of a Bayesian cut, namely $\lambda - \theta \mid X_{..}$, because λ is a function of β and therefore of θ.

Given the impossibility of a Bayesian cut, one may still ask whether some prior specification could allow Z and λ to be mutually ancillary conditionally on $X_{..}$, that is equation (16). Note that in such a case, Z and λ become mutually exogenous conditionally on $X_{..}$ because (13) holds for any prior specification. Now (17) may be decomposed as follows

$$p(Z \mid \lambda, X_{..}) = p(X_{.1} \mid \lambda, X_{..}) p(X_{1.} \mid \lambda, X_{..}, X_{.1})$$

The question whether $p(Z \mid \lambda, X_{..})$ can possibly be independent of λ may be considered in two steps. The first term $p(X_{.1} \mid \lambda, X_{..})$ does not depend on λ under

(19) and (21) because λ is a function of β only, and (19) along with (22) jointly imply $X_{.1} - \lambda \mid X_{..}$. Let us now consider the second term

$$p(X_{1.} \mid \lambda, X_{..}, X_{.1}) = \int p(X_{1.} \mid \lambda, \beta_2, X_{..}, X_{.1}) m(\beta_2 \mid \lambda) d\beta_2$$

Similarly to the multinomial case, let us consider that $\beta_j' s$ are a priori distributed as independent Beta variates with parameters (a_j, b_j) (as is the case when θ follows a priori a Dirichlet distribution). It is shown in Proposition 2 of the appendix that

$$\begin{aligned} p(X_{1.} = x_{1.} \mid \lambda, X_{..}, X_{.1}) &= c\, X_{.1}! X_{.2}! \sum_{x_{11} \in V} \frac{\lambda^{x_{11}}}{\prod_{i,j} x_{ij}!} \\ &\quad \cdot \frac{E[(1 - Y + \lambda Y)^{-(a_2 + b_2 + X_{.2})}]}{E[(1 - W + \lambda W)^{-(a_2 + b_2)}]} \end{aligned}$$

where

$$\begin{aligned} c &= \frac{?\,(a_. + b_.)}{?\,(a_.)?\,(b_.)} \frac{?\,(a_. + x_{1.})?\,(b_. + x_{2.})}{?\,(a_. + b_. + X_{..})} \\ Y &\sim \beta(a_. + x_{1.}, b_. + x_{2.}) \\ W &\sim \beta(a_., b_.) \end{aligned}$$

For the same reason as for the multinomial sampling, one may conclude that under independent binomial sampling, conditioning on the margins is generally not an admissible reduction under the considered prior specification.

4 Numerical Results

In this section we briefly summarize the main results of a numerical experiment aimed at evaluating numerically how much information is actually lost when conditioning on the two margins of a 2x2 contingency table. More extensive presentation of those results is given in Mouchart and Scheihing (1994).

Let us first present the experiment for the expected discrepancy between the posterior distribution of λ, $m(\lambda \mid x)$ and its approximation $m_A(\lambda \mid x)$, as written in (9). One might consider two specifications for $D(\cdot, \cdot)$, namely, the entropy divergence:

$$\begin{aligned} D_E(m_A(\lambda \mid x), m(\lambda \mid x))] &= \int ln(\frac{m(\lambda \mid x)}{m_A(\lambda \mid x)})\, m(\lambda \mid x) d\lambda \\ &\in (0, \infty) \end{aligned}$$

or the Hellinger distance:

$$D_H(m_A(\lambda \mid x), m(\lambda \mid x))] = \frac{1}{2} \int (\sqrt{m(\lambda \mid x)} - \sqrt{m_A(\lambda \mid x)})^2\, d\lambda$$

$$\begin{aligned} &= 1 - \int \sqrt{m(\lambda \mid x)\, m_A(\lambda \mid x)}\, d\lambda \\ &\in [0,1] \end{aligned}$$

One motivation for being interested in those quantities is the following. From a decisional point of view, the weak topology among probability measures corresponds to the convergence of the integral of bounded continuous functions, an attractive space for utility functions. That topology is metrized by the Prohorov distance and by the Levy distances ; these are numerically not very convenient for large scale simulations but they are dominated by the total variation distance which in turns is dominated by twice the entropy divergence and by four times the Hellinger distance and these last two are far easier to evaluate numerically.For details, see e.g. chap.3 in Devroye et al.(1991).

Secondly, for computational purposes, the Relative Efficiency relative to the square loss function written in (10) may be rewritten as:

$$RelEff(m_A) = \left(1 + \frac{E_x[(\,E(\lambda \mid x) - E^A(\lambda \mid x)\,)^2]}{E(\lambda^2) - E_x[(\,E(\lambda \mid x)\,)^2]}\right)^{-1}$$

For all these measures of comparison, we consider their behaviour in relation to the values of two parameters: n, the sample size and n_0, the relative precision of the prior information. We simulate both cases of a multinomial sampling with Dirichlet prior:

$$\theta \sim Di(n_0, p_0) \text{ with } n_0 \in \{10, 15, 20, \cdots, 50\} \text{ and } p_0 = 0.25(1,1,1,1)'$$

and of two independent binomial samplings with two independent Beta priors:

$$\beta_j \sim \beta(n_0, p_0) \text{ with } n_0 \in \{5, 10, 15, \cdots, 30, 40, 50\} \text{ and } p_0 = 0.5(1,1)'$$

For the sample sizes n, we use the same values as for n_0.

For both the multinomial sampling and the independent binomial samplings, the evaluation of the Hellinger distance and the Entropy divergence gave roughly similar results for the range of sample sizes and precision of prior information we have simulated. As an example, let us consider the case of the Hellinger distance for the multinomial sampling. Writing the loss of information $L(n, n_0)$, defined in (9), as a function of n, the sample size, and n_0, the precision of the prior distribution, Figure 1(b) suggests that $L(n, n_0)$ behaves approximately as:

$$L(n, n_0) \sim h(\frac{n}{n_0})$$

where $h : I\!\!R_+ \to I\!\!R_+$ is such that $h' > 0$ or equivalently: there exists a function $\alpha : I\!\!R_+ \to I\!\!R_+$ such that $\alpha' > 0$ and $L(n, \alpha(c) n_0) = c$. This suggests that $L_n^{'} > 0$ and $L_{n_0}^{'} < 0$, implying in particular that the loss of information does not vanish asymptotically (in n). The graphs also suggest that $L_{n,n}^{''} < 0$, that is for any n_0,

$L(n, n_0)$ is concave in n (Figures 1(c)) and $L''_{n_0,n_0} > 0$ that is for any n, $L(n, n_0)$ is convex in n_0 (Figures 1(d)) and eventually tends to 0 when $n_0 \to \infty$, in which case the prior distribution tends to become dogmatic and remains unrevised by the sample, whether the revision is "exact" or "approximate". Thus, for the range of examined values of n and n_0, the error of approximation tends to vanish when $\frac{n}{n_0} \to 0$ or when n_0 increases for fixed n whereas this error does not tend to zero, but rather tends to increase when the sample size n increases with fixed prior precision n_0.

Results on the relative efficiency, written $RE(n, n_0)$, were roughly similar, comforting the fact that the approximation by non-admissible conditioning is actually a small sample one. With respect to the qualitative structure of the function $RE(n, n_0)$, the relative roles of n and n_0 are more complex and deserve a more detailed analysis. Furthermore numerical problems arising from simulating $RE(n, n_0)$ are substantially more involved than for $L(n, n_0)$. Those aspect are discussed more in details in Mouchart and Scheihing (1994).

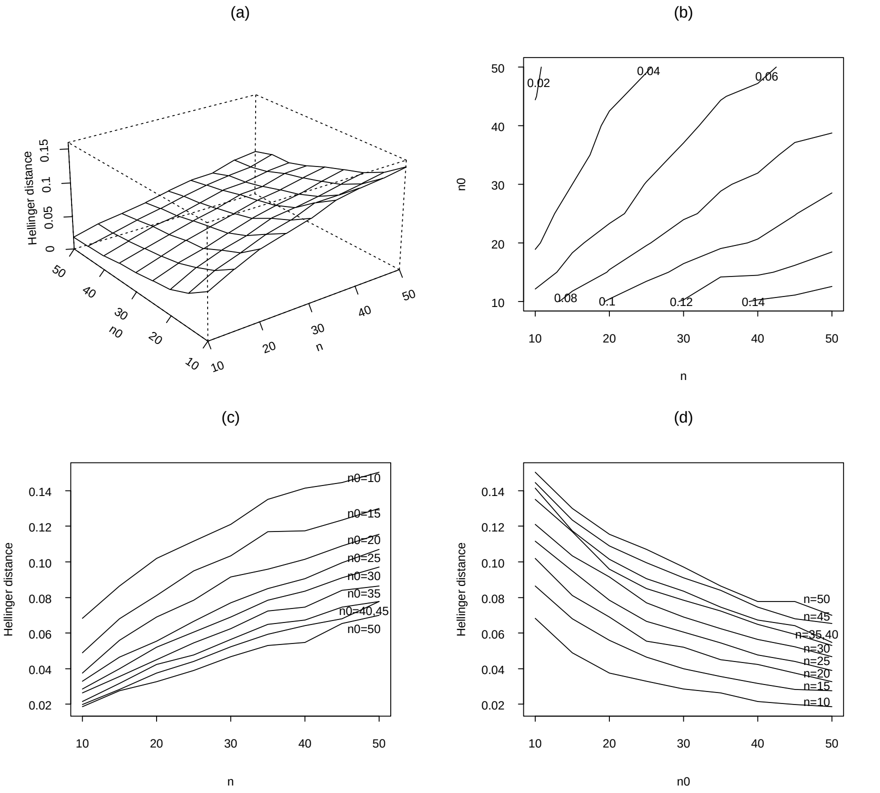

Figure 1: Expectation of the Hellinger distance in the multinomial sampling as a function of the size sample n and the a priori distribution weight n_0. Perspective (fig. (a)), contour (fig. (b)) and one-dimensional plots (figs. (c) and (d)).

5 Admissible Conditioning on the two margins

Justifying for an inference on λ the conditioning of the two margins may be done on two different grounds. One is that, the configuration of n and n_0 is such that, for instance, $L(n, n_0)$ is "close to zero" : this is the situation of quasi-admissibility. A quite different justification is that the conditioning is actually admissible because the sampling, conditional on $X_{..}$ is neither multinomial nor independent binomial.

In this section we give a closer look to this second justification and draw the attention to possible difficulties when building such models. More specifically, let us consider a sampling model where the table is actually generated conditionally on the two margins $Z = (X_{.1}, X_{1.})$. This would ensure the admissibility of conditioning on the two margins provided the prior independence (or variation free character) of the parameters characterizing the two sampling processes (marginal on Z and conditional on Z).

In what follows we want to stress some guaranteed important issues to be dealt with when specifying such a sampling model and defining a parameter of interest deemed to characterize an absence or a presence of row-column independence. These issues are:

(i) such an independence (or lack of) should be a characteristic of the $(X_{11} \mid Z, X_{..}))$-DGP, where DGP stands for the actual Data Generating Process, but should hold irrespectively of the $(Z \mid X_{..})$-DGP

(ii) the null hypothesis of independence isis a simple one-point hypothesis in the space of the $(X_{11} \mid Z, X_{..})$-distributions but its alternative is context-specific.

Let us consider these questions more precisely. We again suppose a first cut making $X_{..}$ exogenous for the inference on any function of θ, the parameter minimal sufficient for the process generating $(X \mid X_{..})$. Conditioning on the two margins would be admissible under a second cut, namely under the condition

$$\psi - \lambda \mid X_{..} \tag{25}$$

or equivalently, under the first cut $\psi - \lambda$, where ψ is a parameter sufficient for the $(Z \mid X_{..})$-DGP: $Z - \theta \mid \psi, X_{..}$ and λ is a parameter sufficient, as in (12), for the $(X \mid Z, X_{..})$-DGP. Under these definitions, $\theta = (\lambda, \psi)$ but the analysis of the multinomial sampling and of the independent binomial samplings show that under (25), the sampling distribution of $(X \mid X_{..})$ may not be multinomial, and the sampling distributions of $(X_{1j} \mid X_{1.}, X_{..})$ may not be independent binomial. It is therefore fitting to ensure a proper understanding of the sampling procedure.

These difficulties will be illustrated in the context of two examples. A first one is the celebrated example of tea-tasting where a lady is tested for her ability to judge whether tea or milk was poured in the cup first, see Fisher (1935b). Let us consider, along with the tea-tasting experiment, another example. Let $X_{..}$ be the number of students in a classroom and let these students be classified into {failure, success} in a mathematics exam (row criterion) and into {type A, type B} for the schooling system followed the year before (column criterion). Here, the $(X_{1.} \mid X_{..})$-DGP would reflect a recruitment pattern of the school and the $(X_{.1} \mid X_{..})$-DGP would reflect the teacher's academic excellence policy. Note that in this case one could accept that $X_{1.}$ and $X_{.1}$ be independent conditionally on $X_{..}$ whereas in the tea-tasting experiment the joint $(X_{1.}, X_{.1} \mid X_{..})$-DGP would be degenerate, in the sense that $P(X_{1.} = X_{.1} \mid X_{..}) = 1$, once the tea-taster is told, before guessing, the number of cups of each type. Note also that in the classroom example, the exogeneity of $X_{..}$ for the parameter of the $(X_{.1}, X_{1.} \mid X_{..})$-DGP may be questionable whereas it would probably not to be questionable in the tea-tasting example.

In both examples, the row-column independence is characterized as a property of the $(X_{11} \mid X_{1.}, X_{.1}, X_{..})$-DGP without any restriction on the $(X_{1.}, X_{.1} \mid X_{..})$-DGP. In the $(X_{11} \mid X_{1.}, X_{.1}, X_{..})$-DGP , the condition of row-column independence is a condition that each individual is assigned to the row and to the column criterion as if he had been selected randomly, in the sense of a hypergeometric sampling. This is equation (12) with $\lambda = 1$, see Plackett (1974):

$$p(X_{11} = x_{11} \mid x_{1.}, x_{.1}, x_{..}, \theta) \quad = \quad \frac{\prod\limits_{i,j} \frac{1}{x_{ij}!}}{\sum\limits_{x_{11} \in V} \prod\limits_{i,j} \frac{1}{x_{ij}!}}$$

$$= \frac{\begin{pmatrix} x_{1.} \\ x_{11} \end{pmatrix} \begin{pmatrix} x_{2.} \\ x_{21} \end{pmatrix}}{\begin{pmatrix} x_{..} \\ x_{.1} \end{pmatrix}} = \frac{\begin{pmatrix} x_{.1} \\ x_{11} \end{pmatrix} \begin{pmatrix} x_{.2} \\ x_{12} \end{pmatrix}}{\begin{pmatrix} x_{..} \\ x_{1.} \end{pmatrix}}$$

The nature of this biconditional sampling is such that the joint process of the table cannot anymore be a counting process of independent individual units, as was the case of the multinomial or of the independent binomial samplings. For instance, in the tea-tasting case, a possible DGP for a tea-taster aware of his (her) absolute inability of recognizing which liquid has been poured first, would be to assign randomly each cup selected in the order of a random permutation until he faces the constraint $X_{.1} = X_{1.}$. Note that in such a case, which particular probability is used in the randomization is irrelevant as far as it is independent of the true type of the cup but the joint probability of all cups cannot be independent, if only because of the constraint $X_{.1} = X_{1.}$, that is the assignation of the $n - X_{1.}$ last ones is determined by the joint assignation of the $X_{1.}$ first ones.

Furthermore, the alternative hypothesis should be specific to any actual situation. For instance, the alternative hypothesis could model the (optimal control) behaviour of a tea-taster trying sequentially each cup with a view to exert her guessing ability or the teacher's policy when he would be willing to differentiate his evaluation according to the origin of his student. In such circumstances there is no reason to believe that the non-central hypergeometric distribution (12) would be the only reasonable alternative. Also, one could expect that the sampling model under the alternative hypothesis would be parameterized with more than one parameter.

6 Discussion

6.1 Summary

In this paper, we considered problems raised when simplifying a statistical inference by deleting one of the terms of the likelihood fonction. More specifically we focused the attention on the problem of conditioning by which part of the data are treated "as if" they were not random. We first considered conditions for admissible conditionning, in which case there is no loss of information for the inference on the parameter of interest. We next considered the problem of measuring how much information is lost when using a non-admissible conditioning.

For the case of non-admissible conditioning, an innovative feature of this paper, to the best of the authors knowledge, is to consider the simplified posterior distribution on the parameter of interest as an approximation to the exact posterior distribution and eventually to measure the loss of information by means of a predictively expected distance (or a divergence) between the exact and the approximate posterior distributions or by means of the ratio between the corresponding bayesian risks relative to a particular loss function.

This methodolody has been exemplified through a case study on an issue recognised as a benchmark when discussing basic problems of inferential methodology, namely the inference on the parameter describing a row-and-column association in a 2 × 2 contingency table operated conditionnally on the two margins; this is indeed the framework of the celebrated Fisher exact test. For this example, we first noticed that in the usual situations of multinomial or of independent binomial samplings conditioning on the two margins involves a loss of information which does not decrease when the sample size increases. The only situations where such a conditioning does not lose information are outside the framework of counting independent units (as is the case for the usual situations). This implies that the modeling of the non-independent alternative should be really context-specific; in particular the non-central hypergeometric distribution is likely not to be anymore a natural modeling of the $(X_{11} \mid X_{.}1, X_{1}., X_{..})$-DGP .

6.2 Some Remarks

When concluding this analysis some remarks seems to be in order.

1. An evaluation finer than the sole computation of the predictive expectation of $D(m_A(\lambda \mid x)\,, m(\lambda \mid x)$, the divergence (or distance) between the exact and the approximate posterior distributions, may consist of analysing, typically by numerical methods, the predictive distribution of that statistic.

Such an anlysis provides one with a natural background for a Bayesian testing of mutual exogeneity; on this approach, see, e.g. Florens and Mouchart (1989, 1993).

2. The existence of a conditional prior distribution $m(\psi \mid \lambda)$, producing property (10) might be investigated by considering (11) as an integral equation. But it would be rather difficult to justify such a (very particular) prior distribution on the grounds of making conditioning on the two margins admissible. Rather than the question of existence of such a prior distribution many statisticians may prefer to consider the question whether a "reasonable" prior specification may possibly produce property (10). Such a "reasonable" class a prior distributions is the Dirichlet one, a large enough class of distributions to accomodate also, in limit cases, non-informative prior distributions; our analysis showed that in this class there is no member producing property (10). Eventually, the methodology exposed above may easily be accomodated to other classes of prior distributions.

3. Let us sketch briefly how the methodology exposed above might be adapted to the second case, namely the problem of marginalizing on Z when λ is again the parameter of interest and the likelihood function displays the same structure as in (2). Equation (3) remains valid but now :

$$p(y \mid z, \lambda) = \int p(y \mid z, \lambda, \psi)\, m(\psi \mid z, \lambda)\, d\psi \qquad (26)$$

Thus the exact admissibility of marginalizing on Z, i.e. of ignoring the data Y under (2), is obtained if $p(y \mid z, \lambda)$ does not depend on λ, i.e. if :

$$\lambda \;-\; y \mid z$$

This condition, along with (2), i.e.

$$z \;-\; \psi \mid \lambda$$

is called a condition of *mutual sufficiency* (see e.g. Florens et al. (1990a)) , a condition more general than the condition of a Bayesian cut (equations (6), (7) and (8)). Although the general methodology for the case of conditioning may be easily adapted to the case of marginalization, the role of the prior information that would make the marginalization exactly or approximately admissible is somewhat different if only because equation (26) involves $m(\psi \mid z, \lambda)$, a distribution which is conditional not only on λ but also on Z. Furthermore, the property of mutual exogeneity may, in some cases, be obtained by invariance arguments (see, e.g. Florens et al. (1990a), Chapter 8 and Florens et al. (1990b)). Clearly, a general discussion of this issue falls out of the scope of this paper.

7 Appendix: Proofs

Proposition 1: Consider the multinomial sampling with parameter $\theta = (\theta_{ij})$ for the 2x2 contingency table (section 3.2). Let $Z = (X_{1.}, X_{.1})$ be the total margins of the table and θ distributed a priori as a Dirichlet distribution with parameter $a = (a_{ij})$, then:

(i) ψ and λ are not independent

(ii)

$$p(Z = (x_{1.}, x_{.1}) \mid \lambda, X_{..}) = c\, X_{..}!\, \frac{E[(1 - Y + \lambda Y)^{-(a_{1.}+x_{1.})}]}{E[(1 - W + \lambda W)^{-a_{1.}}]} \sum_{x_{11} \in V} \frac{\lambda^{x_{11}}}{\prod_{i,j} x_{ij}!}$$

where

$$\begin{aligned} c &= \frac{?^2(a_{..})}{?^2(a_{..} + X_{..})} \frac{?(a_{1.} + x_{1.})?(a_{2.} + x_{2.})?(a_{.1} + x_{.1})?(a_{.2} + x_{.2})}{?(a_{1.})?(a_{2.})?(a_{.1})?(a_{.2})} \\ Y &\sim \beta(a_{.1} + x_{.1}, a_{.2} + x_{.2}) \\ W &\sim \beta(a_{.1}, a_{.2}) \end{aligned}$$

and λ is the usual cross ratio (equation (11)).

Proof: Consider the one-to-one transformation of parameter space defined in section 3.2, equation(14):

$$\begin{aligned} g : \Theta &\rightarrow \Psi \times \mathbb{R}^+ \\ (\theta_{22}, \theta_{21}, \theta_{12}) &\rightarrow (\psi, \lambda) \end{aligned}$$

where

$$\begin{aligned} \Theta &= \{(\theta_{22}, \theta_{21}, \theta_{12}) \in (0,1)^3 / \theta_{22} + \theta_{21} + \theta_{12} < 1\} \\ \Psi &= \{\psi = (\psi_1, \psi_2) \in (0,1)^2 / \psi_1 + \psi_2 < 1\} \end{aligned}$$

$$\begin{aligned} \psi_1 &= \theta_{22} \\ \psi_2 &= \frac{\theta_{21}}{\theta_{21} + \theta_{22}} \end{aligned}$$

and λ is defined by (11). Observe that

$$g^{-1}(\psi_1, \psi_2, \lambda) = (\psi_1, \frac{\psi_1 \psi_2}{1 - \psi_2}, \frac{1 - \psi_1 - \psi_2}{1 - \psi_2 + \lambda \psi_2}) \tag{27}$$

and

$$|Dg^{-1}| = \frac{\psi_1 \psi_2 (1 - \psi_1 - \psi_2)}{(1 - \psi_2)^2 (1 - \psi_2 + \lambda \psi_2)^2} \tag{28}$$

therefore

$$p(Z = (x_{1.}, x_{.1}) \mid \lambda, X_{..}) = \int_{\Psi} p(Z = (x_{1.}, x_{.1}) \mid X_{..}, \psi, \lambda)\, dm((\psi, \lambda) \mid \lambda) \tag{29}$$

Note that $p(Z = (x_{1.}, x_{.1}) \mid X_{..}, \psi, \lambda)$ is given by (15). Furthermore,

$$m((\psi, \lambda) \mid \lambda) = \frac{m(\psi_1, \psi_2, \lambda)}{\int_{\Psi} m(\psi_1, \psi_2, \lambda)\, d\psi_1 d\psi_2} \tag{30}$$

where

$$m(\psi_1, \psi_2, \lambda) \propto m(g^{-1}(\psi_1, \psi_2, \lambda))|Dg^{-1}|$$

if $\theta \sim Di(a)$, equations (27) and (28) imply

$$\begin{aligned} m(\psi_1, \psi_2, \lambda) &= \frac{\psi_1\psi_2(1-\psi_1-\psi_2)}{(1-\psi_2)^2(1-\psi_2+\lambda\psi_2)^2} \left(\frac{\lambda\psi_2(1-\psi_1-\psi_2)}{(1-\psi_2)(1-\psi_2+\lambda\psi_2)}\right)^{a_{11}-1} \\ &\cdot \left(\frac{1-\psi_1-\psi_2}{1-\psi_2+\lambda\psi_2}\right)^{a_{12}-1} \left(\frac{\psi_1\psi_2}{1-\psi_2}\right)^{a_{21}-1} \psi_1^{a_{22}-1} \\ &= \lambda^{a_{11}-1}\, \psi_1^{a_{2.}-1}\, (1-\psi_1-\psi_2)^{a_{1.}-1} \\ &\cdot \psi_2^{a_{.1}-1}\, (1-\psi_2)^{-a_{.1}}\, (1-\psi_2+\lambda\psi_2)^{-a_{1.}} \end{aligned} \tag{31}$$

Thus a priori λ and $\psi = (\psi_1, \psi_2)$ are not independent. From (31) we obtain:

$$I_1(\lambda) = \int_{\Delta} m(\psi_1, \psi_2, \lambda)\, d\psi_1 d\psi_2 \tag{32}$$

$$\begin{aligned} &= \lambda^{a_{11}-1} \int_0^1 \int_0^{1-\psi_2} \psi_1^{a_{2.}-1}\, ((1-\psi_2)-\psi_1)^{a_{1.}-1}\, d\psi_1 \\ &\cdot \psi_2^{a_{.1}-1}\, (1-\psi_2)^{-a_{.1}}\, (1-\psi_2+\lambda\psi_2)^{-a_{1.}}\, d\psi_2 \\ &= \lambda^{a_{11}-1}\, \frac{?\,(a_{1.})?\,(a_{2.})}{?\,(a_{..})} \\ &\cdot \int_0^1 \psi_2^{a_{.1}-1}\, (1-\psi_2)^{a_{.2}-1}\, (1-\psi_2+\lambda\psi_2)^{-a_{1.}}\, d\psi_2 \\ &= \lambda^{a_{11}-1}\, \frac{?\,(a_{1.})?\,(a_{2.})?\,(a_{.1})?\,(a_{.2})}{?\,^2(a_{..})}\, E[(1-W+\lambda W)^{-a_{1.}}] \end{aligned} \tag{33}$$

where $W \sim \beta(a_{.1}, a_{.2})$.
Furthermore, by the equations (29), (15), (30) and (32),

$$p(Z = (x_{1.}, x_{.1}) \mid \lambda, X_{..}) = X_{..}! \sum_{x_{11} \in V} \frac{\lambda^{x_{11}}}{\prod\limits_{i,j} x_{ij}!}\, \frac{I_2(\lambda)}{I_1(\lambda)} \tag{34}$$

where

$$\begin{aligned} I_2(\lambda) &= \lambda^{a_{11}-1} \int_{\Delta} \psi_1^{x_{2.}+a_{2.}-1}\, (1-\psi_1-\psi_2)^{x_{1.}+a_{1.}-1} \\ &\cdot \psi_2^{x_{.1}+a_{.1}-1}\, (1-\psi_2)^{-(x_{.1}+a_{.1})}\, (1-\psi_2+\lambda\psi_2)^{-(x_{1.}+a_{1.})}\, d\psi_1 d\psi_2 \end{aligned}$$

In the analogous way to (33)

$$
\begin{aligned}
I_2(\lambda) \quad &= \quad \frac{?\,(x_{1.} + a_{1.})?\,(x_{2.} + a_{2.})?\,(x_{.1} + a_{.1})?\,(x_{.2} + a_{.2})}{?^{\,2}(x_{..} + a_{..})} \\
&\cdot \quad E[(1 - Y + \lambda Y)^{-(x_{1.}+a_{1.})}]\,\lambda^{a_{11}-1} \qquad (35)
\end{aligned}
$$

where $Y \sim \beta(x_{.1} + a_{.1}, x_{.2} + a_{.2})$.
Thus the result follows from the equations (33),(34) and (35).

Proposition 2: Consider the independent binomial sampling with parameter $\beta = (\beta_1, \beta_2)$ for the 2x2 contingency table (section 3.3). Let β_1 and β_2 be distributed a priori as two independent beta distribution with parameters (a_1, b_1) and (a_2, b_2) respectively, then

$$
p(X_{1.} = x_{1.} \mid \lambda, X_{..}, X_{.1}) \quad = \quad c\,X_{.1}!X_{.2}! \sum_{x_{11} \in V} \frac{\lambda^{x_{11}}}{\prod_{i,j} x_{ij}!} \, \frac{E[(1 - Y + \lambda Y)^{-(a_2+b_2+X_{.2})}]}{E[(1 - W + \lambda W)^{-(a_2+b_2)}]}
$$

where

$$
\begin{aligned}
c \quad &= \quad \frac{?\,(a_{.} + b_{.})}{?\,(a_{.})?\,(b_{.})} \, \frac{?\,(a_{.} + x_{1.})?\,(b_{.} + x_{2.})}{?\,(a_{.} + b_{.} + X_{..})} \\
Y \quad &\sim \quad \beta(a_{.} + x_{1.}, b_{.} + x_{2.}) \\
W \quad &\sim \quad \beta(a_{.}, b_{.})
\end{aligned}
$$

and λ is the usual cross ratio (equation (23)).

Proof: Consider the one-to-one transformation of parameter space defines in section 3.3:

$$
\begin{aligned}
g : [0,1]^2 \quad &\to \quad [0,1] \times \mathbb{R}^+ \\
(\beta_2, \beta_1) \quad &\to \quad (\psi, \lambda)
\end{aligned}
$$

where

$$
\begin{aligned}
\psi \quad &= \quad \beta_2 \\
\lambda \quad &= \quad \frac{\beta_1(1 - \beta_2)}{\beta_2(1 - \beta_1)}
\end{aligned}
$$

Observe that

$$
g^{-1}(\psi, \lambda) = (\psi, \frac{\lambda\psi}{1 - \psi + \lambda\psi}) \qquad (36)
$$

and

$$
|Dg^{-1}| = \frac{\psi(1 - \psi)}{(1 - \psi + \lambda\psi)^2} \qquad (37)
$$

therefore

$$p(X_{1.} = x_{1.} \mid \lambda, X_{..}, X_{.1}) = \int_0^1 p(X_{1.} = x_{1.} \mid \psi, \lambda, X_{..}, X_{.1})\, dm(\psi \mid \lambda) \tag{38}$$

Note that $p(X_{1.} = x_{1.} \mid \psi, \lambda, X_{..}, X_{.1})$is given by (24). Furthermore,

$$m(\psi \mid \lambda) = \frac{m(\psi, \lambda)}{\int_0^1 m(\psi, \lambda)\, d\psi} \tag{39}$$

where

$$m(\psi, \lambda) = m(g^{-1}(\psi, \lambda))|Dg^{-1}|$$

if $\beta_1 \sim \beta(a_1, b_1)$ and $\beta_2 \sim \beta(a_2, b_2)$, equations (36) and (37) imply

$$\begin{aligned} m(\psi, \lambda) &\propto \frac{\psi(1-\psi)}{(1-\psi+\lambda\psi)^2}\, \psi^{a_2-1}(1-\psi)^{b_2-1} \\ &\cdot \left(\frac{\lambda\psi}{1-\psi+\lambda\psi}\right)^{a_1-1} \left(\frac{1-\psi}{1-\psi+\lambda\psi}\right)^{b_1-1} \\ &= \lambda^{a_1-1}\, \psi^{a_{.}-1}(1-\psi)^{b_{.}-1}(1-\psi+\lambda\psi)^{-(a_1+b_1)} \end{aligned}$$

which implies

$$\begin{aligned} I_3(\lambda) &= \int_0^1 m(\psi, \lambda)\, d\psi & (40) \\ &= \lambda^{a_1-1} \int_0^1 \psi^{a_{.}-1}(1-\psi)^{b_{.}-1}(1-\psi+\lambda\psi)^{-(a_1+b_1)}\, d\psi \\ &= \lambda^{a_1-1} \frac{?\,(a_{.})?\,(b_{.})}{?\,(a_{.}+b_{.})} E[(1-W+\lambda W)^{-(a_1+b_1)}] & (41) \end{aligned}$$

where $W \sim \beta(a_{.}, b_{.})$.
Furthermore, by the equations (38), (24), (39) and (40),

$$p(X_{1.} = x_{1.} \mid \lambda, X_{..}, X_{.1}) = X_{.1}!X_{.2}! \sum_{x_{11} \in V} \frac{\lambda^{x_{11}}}{\prod_{i,j} x_{ij}!} \frac{I_4(\lambda)}{I_3(\lambda)} \tag{42}$$

where

$$I_4(\lambda) = \lambda^{a_1-1} \int_0^1 \psi^{a_{.}+x_{1.}-1}(1-\psi)^{b_{.}+x_{2.}-1}(1-\psi+\lambda\psi)^{-(a_1+b_1+X_{.1})}\, d\psi$$

In the analogous way to (41)

$$I_4(\lambda) = \lambda^{a_1-1} \frac{?\,(a_{.}+x_{1.})?\,(b_{.}+x_{2.})}{?\,(a_{.}+b_{.}+X_{..})} E[(1-Y+\lambda Y)^{-(a_1+b_1+X_{.1})}] \tag{43}$$

where $Y \sim \beta(a_{.}+x_{1.}, b_{.}+x_{2.})$.
Thus the result follows from the equations (41), (42) and (43).

References

Agresti A. (1992). A survey of exact inference for contingency tables *Statis. Sci.*, **7**(1), 131-177.

Altham, P.M.E.(1969). Exact Bayesian analysis of a 2x2 contingency table and Fisher's "exact" significance test. *J. Roy. Statist. Soc. Ser.B*, **31**, 261-269.

Altham, P.M.E.(1971). Exact Bayesian analysis of an intraclass 2x2 table *Biometrika* **84**, 679-680.

Barnard G.A. (1963). Some aspects of the fiducial argument *J. Roy. Statist. Soc. Ser.B*, **12**, 111-114.

Barndorff-Nielsen O. (1978). *Information and Exponential Families in Statistical Theory.* New York:John Wiley.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, **47**, 225-238.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. *Proc. Third Berkeley Symp. Math. Statist. Probab.*, **4**, 135-148, Univ. California Press, Berkeley.

Cox, D.R. (1972). Regression Models and Life Table, J.R.S.S., B, 187-220.

Dawid, A.P.(1980). A Bayesian look at the nuisance parameters. In: *Bayesian Statistics.* See J.M. Bernardo, M.H. de Groot, D.V. Lindley, and A.F.M Smith,(1980), 167-184.

Devroye, L., L. Gytrfi and G. Lugosi, 1991, *A Probabilistic Theory of Pattern Recognition*, New York : Springer Verlag.

Fisher, R.A. (1935a). The logic of inductive inference (with discussion). *J. Roy. Statist. Soc. A* **38**, 39-82.
Fisher R.A. (1935b). *The Design of Experiments* (8th ed. 1966). Edinburgh:Oliver and Boyd.

Florens, J.-P. and M. Mouchart (1977). Reduction of Bayesian experiments. CORE Discussion Paper 7737, Université Catholique de Louvain, Louvain-la-Neuve, Belgium(revised July 1979).

Florens, J.-P. and M. Mouchart (1985). Conditioning in dynamics models. *Journal of Time Series Analysis*, **53**(1), 15-35.

Florens, J.-P. and M. Mouchart (1989). Bayesian Specification Tests, in *Contributions to Operations Research and Economics* (B. Cornet and H. Tulkens, editors), Cambridge: The MIT Press, 467-490.

Florens, J.-P. and M. Mouchart (1993). Bayesian Testing and Testing Bayesians, *Handbook of Statistics* **11**, G.S. Maddala, C.R. Rao and H.D. Vinod (editors), Amsterdam : Elsevier Science Publishers B.V.

Florens, J.-P., M. Mouchart and J.-M. Rolin (1990a). *Elements of Bayesian Statistics*, New York:Marcel Dekker, inc.

Florens, J.-P., M. Mouchart and J.-M. Rolin (1990b). Invariance Arguments in Bayesian Statistics. Chap.16 in : *Economic Decision Making : Games, Econometrics and Optimisation* edited by J.Gabszewicz, J.-F. Richard and L.Wolsey, Amsterdam, Elsevier Science Publisher, 387-403.

Haber, M. (1989). Do the marginal totals of a 2x2 contingency table contain information regarding the table proportions? *Commun. Statis. B*, **18**(1) 147-156.

Heyer, H. (1982). *Theory of Statistical Experiments*. New York:Springer-Verlag.

Kalbfleisch, J.G. (1985). *Probability and Statistical Inference*. New York:Springer-Verlag.

Lindley, D. V. (1964). The Bayesian analysis of contingency tables. *Ann. Math. Statist.* **35**, 1622-1643.

Mouchart, M. and E. Scheihing (1993). Fisher test and inference in 2x2 contigency tables: A bayesian evaluation. *1993 Proceedings of the section on Bayesian Statistical Science*, Am. Statis. Assoc., 235-240.

Mouchart, M. and E. Scheihing (1994). Approximate Bayesian Inferences: Numerical experiences when measuring the loss of information. *1994 Proceedings of the section on Bayesian Statistical Science*, Am. Statis. Assoc., 221-226.

Plackett, R.L.,(1974). *The Analysis of Categorical Data*, Monograph **35**, London:Griffin.

Plackett, R.L.,(1977). The marginal totals of a 2x2 table. *Biometrika*, **64**, 37-42.