JIIS manuscript No.
(will be inserted by the editor)

Bayesian Analysis of GUHA Hypotheses

Robert Piché · Marko Järvenpää · Esko Turunen · Milan Šimůnek

Abstract The LISp-Miner system for data mining and knowledge discovery uses the GUHA method to comb through a large data base and finds 2×2 contingency tables that satisfy a certain condition given by generalised quantifiers and thereby suggest the existence of possible relations between attributes. In this paper, we show how a more detailed interpretation of the data in the tables that were found by GUHA can be obtained using Bayesian statistical methods. Using a multinomial sampling model and Dirichlet prior, we derive posterior distributions for parameters that correspond to GUHA generalised quantifiers. Examples are presented illustrating the new Bayesian post-processing tools implemented in LISp-Miner. A statistical model for the analysis of contingency tables for data from two subpopulations is also presented.

Keywords Data mining · GUHA · contingency table · Bayesian statistics

Mathematics Subject Classification (2000) 62F15 · 62H17 · 62-07

1 Introduction

GUHA (General Unary Hypothesis Automaton) is one of the earliest data mining methods, introduced almost half a century ago (Hájek et al., 1966). An overview and history of its development can be found in (Hájek et al., 2010). GUHA is a tool for automated exploratory data analysis of large data sets. The mathematical structure underlying GUHA theory, based on the first order monoidal logic of finite models, allows software to identify "interesting" features in the data without exhaustive search. The theoretical foundation of the GUHA method is supported in many works; to name just a few papers concerning logic of association rules we refer to (Rauch, 2005, 2009, 2013). The most notable software implementation of GUHA method is the freely available LISp-Miner program developed at the University of Economics Prague (Rauch and Šimůnek, 2012; Šimůnek, 2003) .

In this paper we deal mainly with the `4ft-Miner` procedure which is a GUHA procedure implemented in the LISp-Miner system (Rauch and Šimůnek, 2005). The `4ft-Miner` procedure systematically generates both basic Boolean attributes such as

Age($>$50), Education(university), HasCar(yes),

and more complex Boolean attributes such as

Age($>$50) **and not** Education(university) **and** HasCar(yes) .

It outputs relations between pairs of attributes, called *hypotheses*, that are 'interesting'. For example, given a database of attributes of a set of married women, the procedure reports (among other things) that the attribute pair

$\varphi =$ ChildCount(0), $\psi =$ Contraceptive(no-use)

satisfies the *founded implication* relation, whereby at least 95% of women having attribute φ (are childless) have attribute ψ (are not using contraceptives), and at least 90 of the observed women are both φ and ψ. The founded implication relation is one of the many *generalized quantifier*

Robert Piché · Marko Järvenpää
Tampere University of Technology, Tampere, Finland, E-mail: {robert.piche,marko.jarvenpaa}@tut.fi

Esko Turunen
Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University, Prague, Czech Republic, E-mail: esko.turunen@tut.fi

Milan Šimůnek
University of Economics Prague, Czech Republic, E-mail: simunek@vse.cz

association rules that can be identified in LISp-Miner; others will be presented later.

GUHA is intended as a computationally effective tool for the first, exploratory stage of data analysis, when the aim is to get orientation in the domain of investigation. A full analysis normally requires some post-processing stages that would typically be too computationally demanding to be directly applied to the large data set. After a GUHA procedure has sifted through the data and has produced a list of hypotheses, the analyst needs to identify the most interesting hypotheses for further study. This further study could include, among other things, discussions with subject domain experts, additional data collection, and other kinds of data analysis.

As a first step, the analyst would typically just look at the actual attribute data of a hypothesis. This data has the form of a 2×2 double dichotomy contingency table. For example, the contingency table for the previously mentioned database is

	ψ	$\neg\psi$
φ	95	2
$\neg\varphi$	534	842

which says that 95 women have attribute $\varphi \wedge \psi$, i.e. are childless and are not using contraceptives, 2 women are $\varphi \wedge \neg\psi$, and so on. The LISp-Miner software has facilities for basic visualisation of contingency tables (Figure 1).

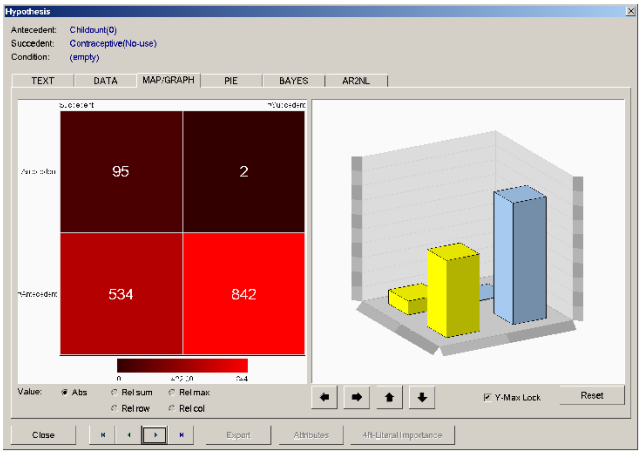

Fig. 1 *Graphical presentation of a contingency table in LISp-Miner.*

The analyst might be satisfied to let the numbers in the contingency table "speak for themselves". For example, for the above contingency table the analyst could simply report that "of the 97 married women who are childless, 95 do not use contraceptives". Clearly, these numbers support the conclusion that "most of the married women who are childless do not use contraceptives".

However, more advanced post-analyses of the hypothesis are possible, by making use of statistical methods that have been developed for analysis of contingency data. The LISp-Miner software includes facilities for conventional statistical hypothesis testing at given level of significance, including Fisher's exact test and the chi-squared test. Conventional statistical procedures have the advantage that they are well-established and are supported by an extensive literature. However, because mistakes in applying or interpreting classical hypothesis tests, p-values, and levels of significance are rather common in applied research (Šimundić and Nikolac, 2009), it seems fair to say that interpretation of data using classical statistics tools is not straightforward and requires extensive specialist training.

An alternative for the interpretation of contingency tables produced by a GUHA analysis is the use of Bayesian statistical methods. The use of Bayesian statistical inference is growing in many applications areas, in part because the comparative ease of modelling and interpretation. The result of Bayesian inference is a (subjective) *probability distribution* for the parameters of interest, and so can, in principle, be interpreted and understood by anyone with a basic knowledge of probability. In addition to an estimate of the parameter, the statistical analysis quantifies the uncertainty of the answer, and this information can be as valuable as the estimate itself. For example, we will show how a Bayesian analysis of the contingency table discussed above yields the statements "We are 86% certain that at least 95% of married women who are childless do not use contraceptives". and "We are 95% certain that the proportion of childless women not using contraceptives is in the interval 0.96 ± 0.03". Such statements express the idea of "most ψ are ϕ" in a way that not only reports the prevalence of the relation, but also a rigorous quantification of the degree of probability (credibility, belief) of the report. Our goal in this paper is to make this kind of analysis available as a postprocessing facility to users of GUHA data mining methods.

Initial results on Bayesian postprocessing of GUHA results were presented in a conference paper (Piché and Turunen, 2010). The present paper gives a more detailed presentation, including background theory and derivations. We derive some new exact and approximate formulas for posterior probabilities, and study some additional generalised quantifier rules. We also present preliminary results on the assessment of the difference between a pair of 2×2 contingency tables.

The paper is organised as follows. In section 2 we present some of the main ideas of GUHA method. In section 3 we recall some useful basic facts of probability and introduce the standard probability distribution functions that we will use in this paper. In section 4 we present some basic ideas of Bayesian inference and present the statistical model of 2×2 contingency tables. We derive the joint posterior dis-

tribution for the model's parameters, which is essentially a complete specification of the subjective state of knowledge about the model parameters in light of the data observed in the contingency table. In section 5 we show how to derive statements from this posterior that are quantified probabilistic versions of statements corresponding to the GUHA generalised quantifers defined in section 2. In section 6 we present some examples to illustrate the implementation of the Bayesian inference tools in the LISp-Miner software. In section 7 we present a statistical model for *pairs* of 2×2 contingency tables, and show how this model can be used to assess the *difference* between generalised quantifier parameters of disjoint sub-populations. Finally, section 8 closes the paper.

2 Data mining background

2.1 The GUHA method in data mining

Data is assumed to be in the form of a categorical array, where each of the m rows corresponds to an object (unit, subject) and each column corresponds to an object's property. The array's cells can contain arbitrary symbols, but before a GUHA data mining task can be carried out the data array must be discretized. In this preprocessing stage, multicategorical attributes such as Age $\in \mathbb{N}$ are transformed into a set of Boolean attributes such as

$$\text{Age}(<30),\ \text{Age}(30\text{–}39),\ \text{Age}(40\text{–}49),\ \text{Age}(\geq 50)$$

This preprocessing can be automated in various ways, see e.g. Rauch and Šimůnek (2005); Rauch (2013).

The GUHA method systematically generates more complex Boolean attributes such as

$$\text{Age}(\geq 50) \text{ and not Education(university) or HasCar(yes)}.$$

Any two attributes φ and ψ in the data can be represented by a 2×2 double dichotomy contingency table of the form

$$\mathbf{a} = \begin{array}{c|cc} & \psi & \neg\psi \\ \hline \varphi & a & b \\ \neg\varphi & c & d \end{array}, \tag{1}$$

where a is the number of objects having both attributes φ and ψ, b is the number of objects having attribute φ but not ψ, etc. We also have $a+b+c+d=m$, the number of objects described in the data set. We can write this as

$$\begin{aligned} a &= \#\{x \mid v(\varphi(x)) = v(\psi(x)) = \text{TRUE}\} \\ b &= \#\{x \mid v(\varphi(x)) = v(\neg\psi(x)) = \text{TRUE}\} \\ c &= \#\{x \mid v(\neg\varphi(x)) = v(\psi(x)) = \text{TRUE}\} \\ d &= \#\{x \mid v(\neg\varphi(x)) = v(\neg\psi(x)) = \text{TRUE}\}, \end{aligned}$$

where # denotes the number of elements in the set, v is a function that evaluates the truth condition `TRUE` or `FALSE` (also denoted 1 or 0), and x is the free variable related to the rows (objects) of the data array.

The aim of exploratory data analysis in GUHA is to identify from all the possible 2×2 tables the ones with 'interesting' relations between attributes φ and ψ. Relations that are `TRUE` are said to be *supported by the data*. Relations between the attributes that are not `TRUE` are `FALSE` and are said to be *not supported by the data*.

In particular, in the `4ft-Miner` procedure, outputs are relations between φ and ψ, called *hypotheses*. A hypothesis is an association rule and can be written as $\approx x(\varphi(x), \psi(x))$ (Hájek and Havránek, 1978), where $\approx$ is a *generalised* (or non-standard) quantifier. A simplified notation $\varphi \approx \psi$ is introduced in (Rauch, 2013). GUHA supports a wide range of semantically rich generalised quantifiers that correspond to 'interesting' relations. The GUHA analysis is computationally feasible because generalised quantifiers satisfy certain monotonicity conditions and each generalised quantifier has its characteristic truth definition. Also other kind of optimizations are used to avoid unnecessary exhaustive search, see (Rauch and Šimůnek, 2005).

Some generalized quantifiers are listed as follows.

Founded implication: $\varphi \Rightarrow_{p,\text{BASE}} \psi$, where BASE $\in \mathbb{N}$ and $p \in (0,1]$, means that at least $100p$ % of the objects that satisfy φ also satisfy ψ and the number of these objects is at least BASE. We then say that φ implies ψ with confidence p and support BASE. Formally:

$$\begin{aligned} &v(\varphi \Rightarrow_{p,BASE} \psi) = \text{TRUE} \\ &\quad \text{iff } \frac{a}{a+b} \geq p \text{ and } a \geq \text{BASE}. \end{aligned} \tag{2}$$

Founded equivalence: $\varphi \equiv_{p,\text{BASE}} \psi$, where BASE $\in \mathbb{N}$ and $p \in (0,1]$, means that attributes φ and ψ have the same truth values `TRUE` or `FALSE` in at least $100p$ % of all objects and the number of objects satisfying both φ and ψ is at least BASE. We then say that φ is equivalent to ψ with confidence p and support BASE. Formally:

$$\begin{aligned} &v(\varphi \equiv_{p,\text{BASE}} \psi) = \text{TRUE} \\ &\quad \text{iff } \frac{a+d}{a+b+c+d} \geq p \text{ and } a \geq \text{BASE}. \end{aligned} \tag{3}$$

Double implication: $\varphi \leftrightarrow_{p,\text{BASE}} \psi$, where BASE $\in \mathbb{N}$ and $p \in (0,1]$, means that at least $100p$ % of such objects that satisfy φ or ψ also satisfy both of them, and the number of objects satisfying both φ and ψ is at least BASE. Formally:

$$\begin{aligned} &v(\varphi \leftrightarrow_{p,\text{BASE}} \psi) = \text{TRUE} \\ &\quad \text{iff } \frac{a}{a+b+c} \geq p \text{ and } a \geq \text{BASE}. \end{aligned} \tag{4}$$

Above average: $\varphi \sim^{+}_{q,BASE} \psi$, where BASE $\in \mathbb{N}$ and $q > 0$, means that among the objects satisfying φ there are at least $100q$ % more objects satisfying ψ than there are objects satisfying ψ, and the number of objects satisfying both φ and ψ is at least BASE. Formally:

$$v(\varphi \sim^{+}_{q,\text{BASE}} \psi) = \text{TRUE} \quad \text{iff } \frac{a}{a+b} \geq \frac{(1+q)(a+c)}{a+b+c+d} \text{ and } a \geq \text{BASE}. \tag{5}$$

Simple association: $\varphi \sim \psi$ means that coincidence of φ and ψ predominates over difference. Formally:

$$v(\varphi \sim \psi) = \text{TRUE iff } ad > bc. \tag{6}$$

Then we say that 'ψ is more prevalent among φ than among $\neg\varphi$' and we also say that 'φ is more prevalent among ψ than among $\neg\psi$'. These interpretations follow from the fact that

$$ad > bc \;\Leftrightarrow\; \frac{a}{a+b} > \frac{c}{c+d} \;\Leftrightarrow\; \frac{a}{a+c} > \frac{b}{b+d}. \tag{7}$$

For more information and additional generalized quantifiers see (Eerola, 2009) and (Turunen, 2012).

After a GUHA procedure has generated compound attributes in the data and found the hypotheses that satisfy a given generalized quantifier, the analyst can proceed to a deeper study of the relations that he or she identifies as being interesting. The statistical methods presented in this paper are intended to serve as preliminary tools to aid in this identification.

2.2 The LISp-Miner system

The LISp-Miner system for Knowledge Discovery in Databases (KDD) is developed at the University of Economics Prague since 1996 and is used both for teaching and for research (Šimůnek, 2003). It is based on a long-term development of the GUHA method but it addresses also current trends in the KDD area, mainly incorporation of domain knowledge, and aims ultimately towards a semi-automated process of data-mining.

The LISp-Miner system consists currently of eight modules implementing GUHA-procedures: `4ft-Miner`, `CF-Miner`, `KL-Miner`, `ETree-Miner`, `SD4ft-Miner`, `SDCF-Miner`, `SDKL-Miner`, and `Ac4ft-Miner`. It also includes a machine learning procedure `KEx` and a general preprocessing module `LM-DataSource`.

The `4ft-Miner` module looks for *4ft-associational rules* with a richer syntax compared to the common *apriori* algorithm. Boolean attributes can be connected with conjuction, disjunctions, and logical negation, and there are many possibilities to define possible interesting patterns in terms of *4ft-quantifiers*. Implementation is very fast thanks to several optimization techniques, the most important of them being *strings of bits* for fast logical operation on all the records in the underlying data matrix at once.

In section 7 the `SD4ft-Miner` module is also considered. The module aims to find all interesting *differences* between pairs of 2×2 contingency tables. It uses a subset of generalized 4ft-quantifiers from the `4ft-Miner` module.

3 Probability distributions and their properties

We start by presenting some results on probability distributions and their properties that we apply later in this paper.

3.1 Distribution of a function of a random vector

We expect the reader is familiar with such elementary concepts of probability theory as random variable and random vector, independence, probability density function (pdf) and probability mass function (pmf), marginal and joint distributions, expectation (denoted $\mathsf{E}(\cdot)$), mean, variance (denoted $\mathsf{V}(\cdot)$, and covariance. These concepts, as well as the following fact, are covered in textbooks of mathematical statistics, e.g. (Roussas, 1997).

Fact 1 (Change of variables) Let x be a continuous random vector and let $T = \{x \in \mathbb{R}^n \mid p_x(x) > 0\}$. If h is a continuously differentiable bijection $T \to S = h(T)$ and if $h'(x)$ has full rank for all $x \in T$, then the random vector $y = h(x)$ has the pdf

$$p_y(y) = p_x(h^{-1}(y))|\det((h^{-1})'(y))| \quad \text{for } y \in S, \tag{8}$$

and zero elsewhere. This formula can be generalised to cases where h is not defined everywhere on T: it is enough that the set where h is not defined has zero measure.

To find the distribution of a random vector $y = g(x) \in \mathbb{R}^m$ that is a function of $x \in \mathbb{R}^n$ with $m < n$, one may proceed in the following way: First, introduce the random vector

$$z = h(x) = \begin{bmatrix} g(x) \\ g_{m+1}(x) \\ \vdots \\ g_n(x) \end{bmatrix},$$

where $g_{m+1}, \cdots, g_n$ are auxiliary functions chosen in such a way that h is a bijection. Then, compute the pdf of z using Fact 1, Finally, integrate over the variables $z_{m+1}, \cdots, z_n$ to obtain the distribution of y as a marginal distribution. In particular, the following result can be derived using this procedure.

Fact 2 (Sum of independent random variables) If x and y are independent continuous univariate random variables then the density of $z = x + y$ is

$$\begin{aligned} p(z) &= \int_{-\infty}^{\infty} p_\xi(x) p_y(z-\xi)\,\mathrm{d}\xi \\ &= \int_{-\infty}^{\infty} p_x(z-\eta) p_y(\eta)\,\mathrm{d}\eta. \end{aligned} \tag{9}$$

3.2 Multinomial distribution

The multinomial model will be used in Section 4.2 to model the 2×2 contingency table.

Definition 1 (Multinomial distribution) Consider an experiment consisting of n independent identically distributed k-outcome trials, with θ_i being the probability of the ith outcome. Let $\theta = (\theta_1, \ldots, \theta_k)$, where $\sum_{i=1}^k \theta_i = 1$, and let x_i denote the number of trials that have the ith outcome. Then the random vector $x = (x_1, \ldots, x_k)$ is multinomially distributed with parameters θ and n, denoted $x \sim \text{Multinomial}(\theta, n)$ or

$$(x_1, \ldots, x_k) \sim \text{Multinomial}(\theta_1, \ldots, \theta_k, n).$$

Here, and in the following, we use the convention that $0^0 = 1$.

The following properties of the multinomial distribution are derived in (Balakrishnan and Nevzorov, 2003).

Fact 3 (Multinomial pmf) If $x \sim \text{Multinomial}(\theta, n)$ then

$$\begin{aligned} &\text{Multinomial}(z; \theta, n) := \text{Prob}(x = z) \\ &= \begin{cases} \frac{n!}{z_1! z_2! \cdots z_k!} \theta_1^{z_1} \cdots \theta_k^{z_k}, \text{ if } z \in \{0, \ldots, n\}^k, \sum_{i=1}^k z_i = n, \\ 0 \qquad\qquad\qquad\qquad \text{otherwise.} \end{cases} \end{aligned} \tag{10}$$

Note that the probability mass lies in a $k-1$ dimensional linear subspace of $\mathbb{R}^k$, because $\text{Prob}(\sum_{i=1}^k x_i = n) = 1$.

Fact 4 (Multinomial moments) If $x \sim \text{Multinomial}(\theta, n)$ then

$$\begin{aligned} &\mathsf{E}(x_i) = n\theta_i, \qquad \mathsf{V}(x_i) = n\theta_i(1-\theta_i), \quad i = 1, \ldots, k, \\ &\text{cov}(x_i, x_j) = -n\theta_i\theta_j, \quad i \neq j,\ i, j = 1, \ldots, k. \end{aligned}$$

3.3 Beta distribution

Definition 2 (Gamma function) The Gamma function is defined as $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\,\mathrm{d}t$. It satisfies the recursion

$$\Gamma(z) = (z-1)\Gamma(z-1)$$

with $\Gamma(1) = 1$, and so $\Gamma(n) = (n-1)!$ for positive integer n.

Definition 3 (Beta distribution) A random variable θ with values in $[0,1]$ is beta distributed with parameters $\alpha > 0$ and $\beta > 0$, denoted $\theta \sim \text{Beta}(\alpha, \beta)$, if it has the density

$$\text{Beta}(t; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1}, \quad t \in [0,1]. \tag{11}$$

Theorem 1 (Beta moments) *If* $\theta \sim \text{Beta}(\alpha, \beta)$ *then*

$$\mathsf{E}(\theta) = \frac{\alpha}{\alpha+\beta} \quad \textit{and} \quad \mathsf{V}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{12}$$

Proof First we derive the formula for the kth moment:

$$\begin{aligned} \mathsf{E}(\theta^k) &= \int_0^1 \theta^k \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\,\mathrm{d}\theta \\ &= \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(\alpha+\beta+k)} \int_0^1 \text{Beta}(\theta; \alpha+k, \beta)\,\mathrm{d}\theta \\ &= \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(\alpha+\beta+k)}. \end{aligned}$$

The moments thus follow the recursion

$$\begin{aligned} \mathsf{E}(\theta^k) &= \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+k-1)}{\Gamma(\alpha)\Gamma(\alpha+\beta+k-1)} \frac{\alpha+k-1}{\alpha+\beta+k-1} \\ &= \frac{\alpha+k-1}{\alpha+\beta+k-1} \mathsf{E}(\theta^{k-1}). \end{aligned}$$

The first two moments are thus

$$\begin{aligned} \mathsf{E}(\theta) &= \frac{\alpha+0}{\alpha+\beta+0} \mathsf{E}(\theta^0) = \frac{\alpha}{\alpha+\beta}, \\ \mathsf{E}(\theta^2) &= \frac{\alpha+1}{\alpha+\beta+1} \mathsf{E}(\theta) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}, \end{aligned}$$

and so

$$\mathsf{V}(\theta) = \mathsf{E}(\theta^2) - (\mathsf{E}(\theta))^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

□

Theorem 2 (Beta mode) *If* $\theta \sim \text{Beta}(\alpha, \beta)$ *with* $\alpha > 1$ *and* $\beta > 1$ *then*

$$\text{mode}(\theta) := \max_{0 \le t \le 1} \text{Beta}(t; \alpha, \beta) = \frac{\alpha-1}{\alpha+\beta-2}. \tag{13}$$

Proof The mode can be found by finding the maximum of the logarithm of the pdf. Denoting $p(t) = \text{Beta}(t; \alpha, \beta)$, we have

$$\begin{aligned} &\log(p(t)) = (\alpha-1)\log t + (\beta-1)\log(1-t) + \text{const.}, \\ &\frac{\partial \log(p(t))}{\partial \theta} = \frac{\alpha-1}{t} - \frac{\beta-1}{1-t}. \end{aligned}$$

This derivative is zero at $t = \frac{\alpha-1}{\alpha+\beta-2}$. By computing the second derivative it can easily be verified that the extremum is indeed a maximum when $\alpha > 1$ and $\beta > 1$. □

3.4 Dirichlet distribution

Information on the Dirichlet distribution can be found in (Kotz et al., 2000; Balakrishnan and Nevzorov, 2003; Devroye, 1986; Frigyik et al., 2010; Ng et al., 2011).

Definition 4 (Dirichlet distribution) A random vector $\theta = (\theta_1, \ldots, \theta_k)$ with values inside the region

$$R_k = \{t \in \mathbb{R}^k \mid t_1 \geq 0, \ldots, t_k \geq 0, \sum_{i=1}^{k} t_i \leq 1\}$$

has a Dirichlet distribution with positive parameters $\alpha = (\alpha_1, \ldots, \alpha_{k+1})$, denoted $\theta \sim \text{Dirichlet}(\alpha)$, if it has the density

$$\text{Dirichlet}(t;\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} t_i^{\alpha_i - 1} (1 - \sum_{i=1}^{k} t_i)^{\alpha_{k+1} - 1}, \quad t \in R_k. \tag{14}$$

where

$$B(\alpha) = \frac{\prod_{i=1}^{k+1} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k+1} \alpha_i)}$$

is the multinomial beta function. A more symmetric formula is obtained by introducing the slack variable $\theta_{k+1} = 1 - \sum_{i=1}^{k} \theta_i$, as follows. The random vector $\theta = (\theta_1, \ldots, \theta_{k+1})$ with values in the simplex (a k-dimensional subset of $\mathbb{R}^{k+1}$)

$$S_k = \{t \in \mathbb{R}^{k+1} \mid t_1 \geq 0, \ldots, t_{k+1} \geq 0, \sum_{i=1}^{k+1} t_i = 1\}$$

has a Dirichlet(α) distribution if its density is

$$\text{Dirichlet}(t;\alpha) \propto \prod_{i=1}^{k+1} t_i^{\alpha_i - 1}, \quad t \in S_k. \tag{15}$$

Note that the beta distribution is a Dirichlet distribution with $k = 1$, that is, $\text{Beta}(\alpha_1, \alpha_2) = \text{Dirichlet}((\alpha_1, \alpha_2))$.

Fact 5 (Dirichlet moments and mode) If

$$\theta = (\theta_1, \ldots, \theta_{k+1}) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_{k+1})$$

and $A = \sum_{i=1}^{k+1} \alpha_i$ then

$$\mathsf{E}(\theta_i) = \frac{\alpha_i}{A}, \quad i = 1, \ldots, k+1$$

$$\mathsf{V}(\theta_i) = \frac{\alpha_i(A - \alpha_i)}{A^2(A+1)}, \quad i = 1, \ldots, k+1,$$

$$\text{cov}(\theta_i, \theta_j) = \frac{-\alpha_i \alpha_j}{A^2(A+1)}, \quad i \neq j,\ i, j = 1, \ldots, k+1.$$

Moreover, if $\alpha_i \geq 1$ for all $i \in \{1, \ldots, k+1\}$ then

$$\text{mode}(\theta) = \frac{\alpha_1 - 1}{A - (k+1)}, \quad i = 1, \ldots, k+1.$$

Fact 6 (Dirichlet aggregation) Let

$$x = (x_1, \ldots, x_{k+1}) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_{k+1}),$$

and let x' be obtained by omitting the jth component and replacing the ith component with the sum of the ith and jth components. Then

$$x' \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_{k+1}).$$

In particular, the marginal distributions are

$$(x_1, \ldots, x_i) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_i, \sum_{j=i+1}^{k+1} \alpha_j), \quad i < k,$$

and

$$x_i \sim \text{Beta}(\alpha_i, \sum_{j=1, j \neq i}^{k+1} \alpha_j), \quad i \in \{1, \ldots, k+1\}.$$

Fact 7 (Dirichlet from beta)

$$x = (x_1, \ldots, x_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_{k+1})$$

iff

$$y_i = \frac{x_i}{1 - \sum_{j=1}^{i-1} x_j}, \quad i = 1, \ldots, k \tag{16}$$

where y_i are independent $\text{Beta}(\alpha_i, \sum_{j=i+1}^{k+1} \alpha_j)$ random variables.

3.5 Multivariate normal distribution

Definition 5 A p-variate random vector x is said to be normally distributed with parameters $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$ (symmetric positive-definite), denoted $x \sim \text{Normal}(\mu, \Sigma)$, if its joint pdf is

$$p(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}, \quad x \in \mathbb{R}^p. \tag{17}$$

Fact 8 (Normal moments of normal) If $x \sim \text{Normal}(\mu, \Sigma)$ then $\mathsf{E}(x) = \mu$ and its variance-covariance matrix is Σ.

4 Bayesian analysis of 2×2 contingency tables

4.1 Basics of Bayesian statistics

This section presents a very abridged outline of Bayesian statistics. For elementary introductions to Bayesian statistics see (Bolstad, 2007; Lee, 2012; Berry, 1996).

Statistical inference can be considered as an "inverse problem" in the following sense. A statistical model specifies the probability distribution of possible observations (in vector y), and the model includes some parameters (vector θ). The statistical inference problem is to describe θ when y

is given. In Bayesian statistics, the unknown parameters are modeled as random variables. That is, the probability distribution of θ serves as a model of the analyst's uncertainty about the parameters. The "inverse problem" is solved by applying the formula from probability theory that is known as Bayes' rule:

$$p(\theta \,|\, y) \propto p(\theta) p(y \,|\, \theta), \tag{18}$$

In the Bayesian statistics setting, $p(\theta)$ is the pdf describing the analyst's state of knowledge about the parameter θ before the data is processed; it is called the *prior density*, or simply *the prior*. The conditional pdf $p(y \,|\, \theta)$ is the probabilistic model for the observation, known as the *likelihood*. The pdf $p(\theta \,|\, y)$, called the *posterior density* or the *posterior*, describes the analyst's state of knowledge about θ after the data y has been obtained. The proportionality constant in (18) is $1/\int p(\theta) p(y \,|\, \theta)\, \mathrm{d}\theta$; this is the scaling factor that ensures that the right hand side to integrates to 1.

When we have little or no knowledge about the parameter, we can use a prior distribution with a large dispersion. Generally speaking, the more data we have, and the larger the prior's dispersion, the less the prior affects the inference result (i.e. the posterior). For computational convenience we often select a prior pdf that is "conjugate" to the data model (likelihood), in the sense that the posterior pdf belongs to the same family of distributions as the prior. Then, when the distributions in the family are standard statistical distributions, specific properties of the posterior are easily obtained. More advanced computational procedures such as Markov Chain Monte Carlo algorithms have been introduced in the last two decades to handle a wide range of Bayesian statistical models, but in this paper we restrict ourselves to models with conjugate priors, for which the computations are straightforward.

The posterior distribution is essentially a complete specification of the state of knowledge about the model parameters in light of the observed data. A good first step in getting acquainted with a posterior is to exploit human visual intelligence by plotting density functions of univariate marginals. A graph with a single narrow peak indicates that the mean or mode value is a good estimate of the parameter's value; the width of the peak gives an indication about the uncertainty associated with this estimate. One can then go on to compute other quantitative indicators such as the 95% credibility interval, that is, a parameter interval containing 95% of the probability.

Hypotheses can also be tested. In Bayesian hypothesis testing one computes the actual (posterior) probability that the hypothesis (a statement about the parameter vector) is true, and this is one minus the probability that the hypothesis is false. In contrast, the conceptual bases of classical Neyman-Pearson hypothesis testing and Fisher significance testing are considerably more complex (and mutually incompatible, see Hubbard (2011)), and consequently are often incorrectly applied and interpreted.

4.2 Multinomial model for 2×2 contingency table

We start by describing a standard statistical model for 2×2 contingency table (1) with unconstrained row and column sums. Given two attributes φ and ψ, there are four possible disjoint attribute combinations:

$$Y_i \in \{\underbrace{\varphi \wedge \psi}_{X_1}, \underbrace{\varphi \wedge \neg\psi}_{X_2}, \underbrace{\neg\varphi \wedge \psi}_{X_3}, \underbrace{\neg\varphi \wedge \neg\psi}_{X_4}\}. \tag{19}$$

To each attribute combination X_j in (19) we associate a parameter θ_j that represents its probability of occurence, conditional on the values of the parameters $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$:

$$\mathrm{Prob}(Y_i = X_j \,|\, \theta) = \theta_j, \quad i \in \{1, \ldots, m\}, j \in \{1,2,3,4\}. \tag{20}$$

The parameters satisfy $\theta_j \geq 0$ and $\sum_{j=1}^{4} \theta_j = 1$.

For a set of observations $Y_1, \ldots, Y_m$ that are independent given a vector θ, the pmf is

$$p(Y_1, \cdots, Y_m \,|\, \theta) = \theta_1^a \cdot \theta_2^b \cdot \theta_3^c \cdot \theta_4^d, \tag{21}$$

where a is the number of observations whose value is X_1, b is the number of observations whose value is X_2, and so on. We also have $a+b+c+d = m$. Thus, the pmf for the contingency table $\mathbf{a} = (a,b,c,d)$ is $\mathbf{a} \,|\, \theta \sim \mathrm{Multinomial}(\theta, m)$, that is

$$p(\mathbf{a} \,|\, \theta) = \frac{m!}{a!b!c!d!} \theta_1^a \cdot \theta_2^b \cdot \theta_3^c \cdot \theta_4^d \propto \theta_1^a \cdot \theta_2^b \cdot \theta_3^c \cdot \theta_4^d. \tag{22}$$

This is the "likelihood" for statistical inference, that is, a model of how the data could be produced by a random number generator, given the parameters.

Next we need to specify a prior distribution. It is convenient to choose a Dirichlet prior, because the Dirichlet distribution is conjugate to the multinomial distribution. Thus, we assume a prior of the form

$$\begin{aligned} &\theta \sim \mathrm{Dirichlet}(\alpha', \beta', \gamma', \delta'), \\ &p(\theta) \propto \theta_1^{\alpha'-1} \theta_2^{\beta'-1} \theta_3^{\gamma'-1} \theta_4^{\delta'-1}, \end{aligned} \tag{23}$$

where $\alpha', \beta', \gamma', \delta'$ are positive numbers that are chosen in such a way that the distribution (23) is a reasonable representation of our state of knowledge about θ prior to processing the observations in the contingency table. In the illustrative examples we shall generally use the prior parameters $\alpha' = \beta' = \gamma' = \delta' = 1$. This choice gives a uniform density over the θ-simplex, and this can be considered to be a "vague" prior.

Substituting the likelihood (22) and the prior (23) into Bayes' rule (18), we obtain the posterior

$$p(\theta \,|\, \mathbf{a}) \propto p(\theta) p(\mathbf{a} \,|\, \theta) \propto \theta_1^{a+\alpha'-1} \theta_2^{b+\beta'-1} \theta_3^{c+\gamma'-1} \theta_4^{d+\delta'-1}. \quad (24)$$

This is recognized as also being a Dirichlet distribution, and the posterior can be written

$$\theta \,|\, \mathbf{a} \sim \text{Dirichlet}(\alpha, \beta, \gamma, \delta), \quad (25)$$

where $\alpha = a + \alpha', \beta = b + \beta', \gamma = c + \gamma', \delta = d + \delta'$.

In the Bayesian statistics framework, the posterior distribution is a comprehensive specification of our state of knowledge about the underlying parameters of the contingency table. As the occurrence counts a, b, c, d grow, the pdf forms a peak around $(\frac{a}{m}, \frac{b}{m}, \frac{c}{m}, \frac{d}{m})$, in agreement with our intuition that the relative frequencies should approximate the probabilities. The dispersion of the pdf around the peak describes the degree of uncertaintly associated with this estimate.

5 Posterior probability distributions of generalized quantifier parameters

Because it is defined in a simplex in 4-dimensional parameter space, it is difficult to visualize the full posterior distribution (24) and to appraise the uncertainties that it models. It is easier to study a univariate marginal distribution, that is, the probability distribution of a scalar-valued function of the parameters. In this section we present some marginal distributions that correspond to the GUHA generalized quantifiers presented in section 2.

5.1 Founded implication

The GUHA procedure for the founded implication quantifier seeks attributes such that $\frac{a}{a+b}$, the ratio of the number of occurrences of $\varphi \wedge \psi$ to the number of occurrences of φ, is large. In our statistical model, the proportions of $\varphi \wedge \psi$ and of φ are $\theta_1 + \theta_2$ and θ_1, respectively, so the proportion of ψ among the φ is

$$\theta_{\text{fi}} := \frac{\theta_1}{\theta_1 + \theta_2},$$

which we call the *founded implication parameter*. The statistical inference question is to assess whether, or to what degree, the unknown parameter θ_{fi} can be said to be "large".

Theorem 3 *The posterior distribution of the founded implication parameter is*

$$\theta_{\text{fi}} \,|\, \mathbf{a} \sim \text{Beta}(\alpha, \beta). \quad (26)$$

Proof From

$$(\theta_3, \theta_4, \theta_1, \theta_2) \,|\, \mathbf{a} \sim \text{Dirichlet}(\gamma, \delta, \alpha, \beta)$$

and Theorem 7 with $i = 3$ we have

$$\theta_{\text{fi}} \,|\, \mathbf{a} = \frac{\theta_1}{\theta_1 + \theta_2} \,|\, \mathbf{a} = \frac{\theta_1}{1 - \theta_3 - \theta_4} \,|\, \mathbf{a} \sim \text{Beta}(\alpha, \beta).$$

□

Formulas for the posterior mean, variance and mode for θ_{fi} are given in Theorems 1 and 2. For large a and b, the mean and mode are both approximately equal to the data ratio $\frac{a}{a+b}$. Indeed, for the standard vague prior $\alpha' = \beta' = 1$, where the prior distribution of θ_{fi} is uniform, the mode coincides with the data ratio $\frac{a}{a+b}$.

The posterior distribution can be used to assess the validity of the statement "θ_{fi} is large" in various ways:

Plot the pdf: The analyst can look at a plot of the density function to evaluate whether most of the probability lies near 1.

Probability of $\theta_{\text{fi}} > p$: The posterior probability that the parameter of founded implication is larger than some given value p (say, $p = 95\%$) can be computed using the formula $1 - \text{BetaCDF}(p; \alpha, \beta)$, where BetaCDF is the cumulative distribution function (cdf).

Credibility interval: The inverse cdf can be used to find a "credibility interval" for the parameter: $q\%$ of the probability is contained in the interval

$$[\text{BetaCDF}^{-1}(\tfrac{1-q}{2}; \alpha, \beta), \text{BetaCDF}^{-1}(1 - \tfrac{1-q}{2}; \alpha, \beta)].$$

If computation of BetaCDF^{-1} is unavailable or too slow, the beta distribution can be approximated by a normal distribution having the same mean and variance. Then the 95% credibility interval is approximately

$$\frac{\alpha}{\alpha + \beta} \pm 1.96 \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}.$$

5.2 Founded equivalence

The GUHA procedure for the founded equivalence quantifier seeks attributes such that $\frac{a+d}{m}$, the proportion of objects having equal φ and ψ truth-values, is large. In our statistical model, the proportion of $(\varphi \wedge \psi) \vee (\neg\varphi \wedge \neg\psi)$ is

$$\theta_{\text{fe}} := \theta_1 + \theta_4,$$

which we call the *founded equivalence parameter*. The statistical inference question here is to assess whether, or to what degree, the unknown parameter θ_{fe} can be said to be "large".

Theorem 4 *The posterior distribution of the founded equivalence parameter is*

$$\theta_{\text{fe}} \,|\, \mathbf{a} \sim \text{Beta}(\alpha+\delta, \beta+\gamma). \tag{27}$$

Proof The result follows directly from

$$(\theta_1+\theta_4, \theta_2, \theta_3) \,|\, \mathbf{a} \sim \text{Dirichlet}(\alpha+\delta, \beta, \gamma)$$

and Theorem 6. □

From Theorems 1 and 2 we have

$$\mathsf{E}(\theta_{\text{fe}} \,|\, \mathbf{a}) = \frac{\alpha+\delta}{A},\ \mathsf{V}(\theta_{\text{fe}} \,|\, \mathbf{a}) = \frac{(\alpha+\delta)(\beta+\gamma)}{A^2(A+1)},$$
$$\text{mode}(\theta_{\text{fe}} \,|\, \mathbf{a}) = \frac{\alpha+\delta-1}{A-2}$$

where $A := \alpha+\beta+\gamma+\delta$. For large occurrence count values, the mean and mode are both approximately equal to the data ratio $\frac{a+d}{m}$.

Again, the validity of the statement "θ_{fe} is large" can be assessed by plotting the pdf, computing the posterior probability that $\theta_{\text{fe}} > p$, and/or computing a credibility interval for θ_{fe}.

5.3 Double implication

The GUHA procedure for the double implication quantifier seeks attributes such that $\frac{a}{a+b+c}$, the ratio of the number of occurrences of $\varphi \wedge \psi$ to the number of occurrences of $\varphi \vee \psi$, is large. In our statistical model, the proportions of $\varphi \wedge \psi$ and $\varphi \vee \psi = \neg(\neg\varphi \wedge \neg\psi)$ are θ_1 and $1-\theta_4 = \theta_1+\theta_2+\theta_3$, respectively, so the proportion of $\varphi \wedge \psi$ among $\varphi \vee \psi$ is

$$\theta_{\text{di}} := \frac{\theta_1}{\theta_1+\theta_2+\theta_3},$$

which we call the *double implication parameter*. The statistical inference question here is to assess whether, or to what degree, the unknown parameter θ_{di} can be said to be "large".

Theorem 5 *The posterior distribution of the double implication parameter is*

$$\theta_{\text{fe}} \,|\, \mathbf{a} \sim \text{Beta}(\alpha, \beta+\gamma). \tag{28}$$

Proof The result follows directly from

$$(\theta_4, \theta_1, \theta_2, \theta_3) \,|\, \mathbf{a} \sim \text{Dirichlet}(\delta, \alpha, \beta, \gamma)$$

and Theorem 7 with $i = 2$. □

From Theorems 1 and 2 we have

$$\mathsf{E}(\theta_{\text{di}} \,|\, \mathbf{a}) = \frac{\alpha}{\alpha+\beta+\gamma},$$
$$\mathsf{V}(\theta_{\text{di}} \,|\, \mathbf{a}) = \frac{(\alpha)(\beta+\gamma)}{(\alpha+\beta+\gamma)^2(\alpha+\beta+\gamma+1)},$$
$$\text{mode}(\theta_{\text{di}} \,|\, \mathbf{a}) = \frac{\alpha-1}{\alpha+\beta+\gamma-2}.$$

For large occurrence count values, the mean and mode are both approximately equal to the data ratio $\frac{a}{a+b+c}$.

Again, the validity of the statement "θ_{di} is large" can be assessed by plotting the pdf, computing the posterior probability that $\theta_{\text{di}} > p$, and/or computing a credibility interval for θ_{di}.

5.4 Above average

The GUHA procedure for the above average quantifier seeks attributes such that $\frac{a}{a+b} / \frac{a+c}{m}$, the ratio of the fraction of ψ occurrences among the φ occurrences to the overall proportion of ψ objects, is large. In our statistical model, the proportion of all ψ is $\theta_1+\theta_3$ and the proportion of ψ among φ is $\frac{\theta_1}{\theta_1+\theta_2}$, and their ratio is

$$\theta_{\text{aa}} := \frac{\theta_1}{\theta_1+\theta_2} / (\theta_1+\theta_3) = \frac{\theta_1}{(\theta_1+\theta_2)(\theta_1+\theta_3)},$$

which we call the *above average parameter*.

Theorem 6 *The posterior pdf of the above average parameter is*

$$p(\theta_{\text{aa}} \,|\, \mathbf{a}) = \int_0^1 \int_0^1 q(\theta_{\text{aa}}, y, z)\, \mathrm{d}z\, \mathrm{d}y$$

where

$$q(x,y,z) = \frac{1}{B(\alpha)} \Big((xyw)^{\alpha-1} (y-xyw)^{\beta-1} (w-xy)^{\gamma-1} (1-y-w+xyw)^{\delta-1} yw z^{-1} \Big)$$

for $0 < x < 1$, *where* $w = \frac{1-y}{1-xy} z$, *and*

$$q(x,y,z) = \frac{1}{B(\alpha)} \Big(x^{-A} (yz)^{\alpha+1} (y-yz)^{\beta-1} (z-yz)^{\gamma-1} (x-y-z+yz)^{\delta-1} \Big)$$

for $x > 1$.

Proof The posterior pdf (25) can be written

$$p(\theta_1, \theta_2, \theta_3 \,|\, \mathbf{a}) = \frac{\theta_1^{\alpha-1} \theta_2^{\beta-1} \theta_3^{\gamma-1} (1-\theta_1-\theta_2-\theta_3)^{\delta-1}}{B(\alpha)} \tag{29}$$

on the simplex

$$(\theta_1,\theta_2,\theta_3)\in S=\{\theta\in\mathbb{R}^3:\theta_i\geq 0,\sum_{i=1}^{3}\theta_i\leq 1\}.$$

Consider the transformation

$$(x,y,z)=h(\theta)=\begin{bmatrix}\frac{\theta_1}{(\theta_1+\theta_2)(\theta_1+\theta_3)}\\ \theta_1+\theta_2\\ \theta_1+\theta_3\end{bmatrix},$$

and its inverse

$$\theta=h^{-1}(x)=\begin{bmatrix}xyz\\ y-xyz\\ z-xyz\end{bmatrix}.$$

The Jacobian matrix of the inverse transformation

$$(h^{-1})'(x)=\begin{bmatrix}yz & xz & xy\\ -yz & 1-xz & -xy\\ -yz & -xz & 1-xy\end{bmatrix}$$

has determinant yz, so the posterior pdf is transformed to

$$\begin{aligned}p(x,y,z\,|\,\mathbf{a})&=p_{\theta\,|\,y}(h^{-1}(x)|yz|\\ &=\frac{1}{B(\alpha)}\Big(xyz)^{\alpha-1}(y-xyz)^{\beta-1}(z-xyz)^{\gamma-1}\cdot\\ &\qquad(1+xyz-y-z)^{\delta-1}\Big)|yz|\end{aligned}$$

on the domain

$$T=\{(x,y,z):0\leq x\leq 1,\ 0\leq y\leq 1,\ 0\leq z\leq\frac{1-y}{1-xy}$$
$$\text{or } x\geq 1,\ 0\leq y\leq\frac{1}{x},\ 0\leq z\leq\frac{1}{x}\}$$

The marginal pdf is

$$p(x\,|\,\mathbf{a})=\begin{cases}\int_0^1\int_0^{\frac{1-y}{1-xy}}p(x,y,z\,|\,\mathbf{a})\,\mathrm{d}z\,\mathrm{d}y & \text{if } 0<x<1,\\ \int_0^{1/x}\int_0^{1/x}p(x,y,z\,|\,\mathbf{a})\,\mathrm{d}z\,\mathrm{d}y & \text{if } x\geq 1.\end{cases}$$

Using the change of variables $z'=\frac{1-y}{1-xy}z$ in the first integral and $y'=\frac{y}{x}$, $z'=\frac{z}{y}$ in the second one, we obtain the formula given in the Theorem. □

In this case, because we can't specify the posterior distribution of θ_{aa} using a standard probability distribution, elucidating the properties of this parameter is not so easy as for the parameters of generalized quantifiers considered earlier. The pdf of θ_{aa} can be plotted using the formula of Theorem 6 and numerical cubature software. This however requires computing a double integral for each point at which the density is evaluated, which can be slow.

A somewhat faster alternative is to approximate the posterior by a normal distribution having the same mean and variance. The posterior mean can be computed by numerical cubature over the simplex:

$$\mathsf{E}(\theta_{\text{aa}}\,|\,\mathbf{a})=\int_{S_3}\frac{\theta_1}{(\theta_1+\theta_2)(\theta_1+\theta_3)}\,\text{Dirichlet}(\theta;\alpha,\beta,\gamma,\delta)\,\mathrm{d}\theta$$

using, for example, the formulas in (Cools, 2003). The posterior variance can be computed similarly. The normal pdf can then be used to evaluate the "largeness" of θ_{aa} by plotting the pdf, computing the probability of $\theta_{\text{aa}}>p$, or computing a credibility interval.

The easiest alternative is to use Monte Carlo simulation. Samples from the full posterior (25) can be generated using standard algorithms (see e.g. (Devroye, 1986)). A normalised histogram of these samples can serve as a simple approximation of the pdf; a better approximation can be obtained using kernel density estimation. The samples can also be used to compute a credibility interval or the probability of $\theta_{\text{aa}}>p$ in a straightforward way.

5.5 Simple association quantifier

The GUHA procedure for the simple association quantifier seeks attributes such that $ad>bc$. As noted in section 2.1, this inequality is equivalent to the inequality $\frac{a}{a+b}>\frac{c}{c+d}$, that is, the fraction of ψ among φ is larger than the fraction of ψ among $\neg\varphi$. In our statistical model, the proportion of ψ among φ is $\frac{\theta_1}{\theta_1+\theta_2}$, the proportion of ψ among $\neg\varphi$ is $\frac{\theta_3}{\theta_3+\theta_4}$, and their ratio is

$$\theta_{\text{sa}}:=\frac{\theta_1}{\theta_1+\theta_2}\Big/\frac{\theta_3}{\theta_3+\theta_4}.$$

We call this the *simple association parameter*.

As in the case of the above average parameter, the posterior distribution of θ_{sa} is not easy to deal with analytically, but can be studied using numerical or Monte Carlo methods.

6 Implementation in LISp-Miner

The `4ftResult` module of the `4ft-Miner` procedure offers a number of tools for displaying 4ft-association rules that the procedure has found to be true in the data, including tables and basic graphical representations. In this section some examples are presented to illustrate the newly-implemented tools for Bayesian interpretation of the results.

The data set alluded to in the introduction is based on Tjen-Sien Lim's publicly available benchmark data test set (Myllymäki et al., 2002) from the 1987 National Indonesia Contraceptive Prevalence Survey. These are the responses from interviews of $m=1473$ married women who were not (as far as they knew) pregnant at the time of interview. The challenge is to predict a woman's contraceptive method from

knowledge about her demographic and socioeconomic characteristics.

The 10 survey response variables and their types are

Age	integer 16–49
Education	4 categories
Husband's education	4 categories
Number of children borne	integer 0–15
Islamic	binary (yes/no)
Working	binary (yes/no)
Husband's occupation	4 categories
Standard of living	4 categories
Good media exposure	binary (yes/no)
Contraceptive method used	3 categories

The data was automatically processed into binary form as follows. The three binary variables need no processing. The 3-category variable ("contraceptive method used") is divided into three binary properties, one for each category; each of the four 4-category variables is similarly divided into four binary properties. The age variable is divided into 118 properties: 31 3-year ranges (16–18, 17–19, ..., 47–49), 30 4-year ranges (16–19, ..., 46–49), 29 5-year ranges, and 20 6-year ranges. Similarly, the number-of-children variable is divided into 58 properties: 16 singletons (0, 1,..., 15), 15 two-unit ranges (0–1, 1–2, ..., 14–15), 14 3-unit ranges (0–2, ..., 13–15), and 13 4-unit ranges (0–3, ..., 12–15). Altogether, there were 198 binary properties.

In the first LISp-Miner run, the system was set the task of finding "well-founded implication" relations $\phi \Rightarrow_{0.95,50} \psi$ with the Contraceptive method properties as ψ and all possible conjunctions of length 1–9 of the remaining properties as ϕ. In 7 seconds, after explicitly testing 179 447 tables, 9 contingency tables satisfying the relation were found. One of them was

	ψ	$\neg\psi$
ϕ	95	2
$\neg\phi$	534	842

where ϕ="no children" and ψ ="not using contraceptives". From the table it can be read that, of the 97 married women who are childless, 95 do not use contraceptives. Figure 1 shows how the table is visualised in `4ftResult`.

The Bayesian analysis of section 4 is applied to this data with the vague (uniform distribution) prior distribution $\theta \propto 1$, that is, $\alpha' = \beta' = \gamma' = \delta' = 1$. The posterior probability distribution of the founded implication parameter is then, by Theorem 3,

$$\theta_{\text{fi}} \,|\, \mathbf{a} \sim \text{Beta}(96,3)$$

Figure 2 shows the plot of the pdf of this distribution that is produced by the `4ftResult` module. It can be seen that most of the probability is concentrated around the posterior mean $\frac{96}{96+3} = 0.9697$. More precisely: 95% of the probability is in the interval

$$[\text{BetaCDF}^{-1}(0.025;96,3), \text{BetaCDF}^{-1}(0.975;96,3)]$$
$$= [0.9294, 0.9936] = 0.9615 \pm 0.0321.$$

The posterior probability that $\theta_{\text{fi}} > 0.95$ is

$$1 - \text{BetaCDF}(0.95;96,3) = 0.8732,$$

that is, we are 87% sure that at least 95 % of married childless women are not using contraceptives.

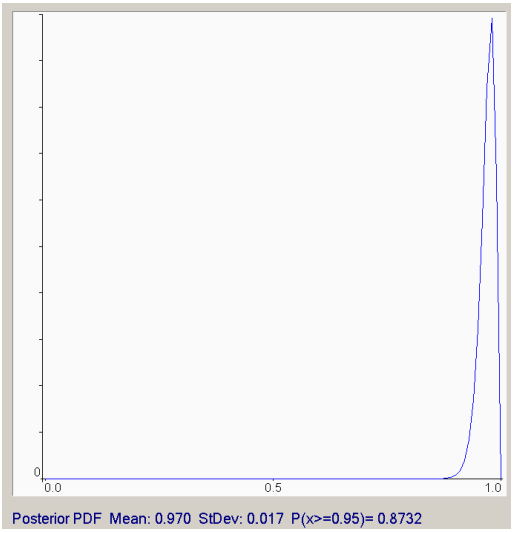

Fig. 2 Posterior probability density function of the founded implication parameter for the contingency table in Figure 1.

In the second LISp-Miner run, the system was set the task of finding "above-average" relations $\phi \sim^{+}_{3,15} \psi$, with the Contraceptive method properties as ψ and all possible conjunctions of length 1–9 of the remaining properties as ϕ. In 3 minutes 17 seconds, after explicitly testing 4 888 398 tables, 14 contingency tables satisfying the relation were found. One of them was

	ψ	$\neg\psi$
ϕ	21	2
$\neg\phi$	312	1138

, (30)

where ϕ ="Age 37–45 and Children 4 and Husband highly educated and Living standard high", and ψ ="Using long-term contraception method". Figure 3 shows how this data is visualised as a pie chart in `4ftResult`.

For the Bayesian model with the vague prior ($\theta \propto 1$) the posterior distribution for the full parameter set is

$$\theta \,|\, y \sim \text{Dirichlet}(22,3,313,1139). \quad (31)$$

Sampling is used to visualise the probability distribution for the above-average parameter. Figure 4 shows the histogram

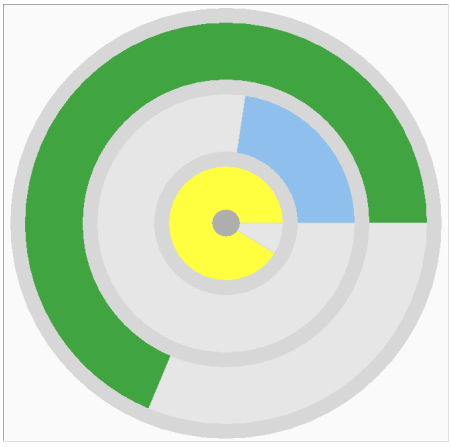

Fig. 3 Pie chart representation of the contingency table (30). The innermost (yellow) band shows $a/(a+b)$, the proportion of ψ among the φ. The central (blue) band shows $(a+c)/(a+b+c+d)$, the proportion of ψ in the whole population. The outer (green) band shows the "lift" $a/(a+b)-(a+c)/(a+b+c+d)$, which indicates how much more frequent ψ is within φ than in general.

of θ_{aa} obtained from 10^4 samples generated from the full posterior (31). The Dirichlet distribution's samples are generated using an algorithm based on Gamma distribution sample generation (Devroye, 1986), with uniform random variates computed using the standard libraries of Visual Studio 2010. This Monte Carlo computation requires less than 0.1 s on a laptop.

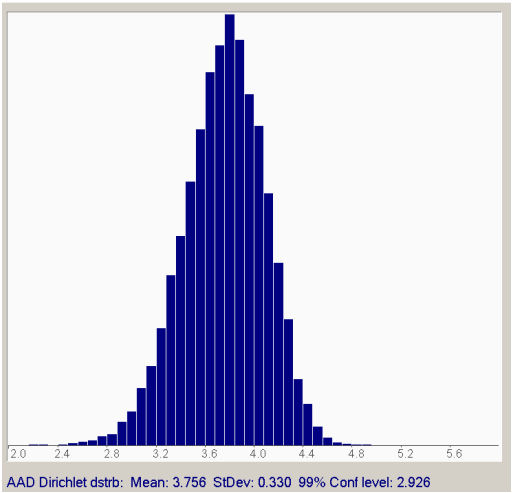

Fig. 4 Histogram of samples from the posterior distribution of the above average parameter for the contingency table (30).

Note that, although the contingency table satisfies the generalised quantifier for the statement "ψ is over 4 times more prevalent among ϕ than in general", the statistical model indicates that the actual factor may be somewhere between 2.4 and 4.8. Because 99% of the samples satisfy $\theta_{\text{aa}} >$ 2.926, we can say that we are 99% certain that the use of long-term contraceptives is at least 2.9 times more prevalent among rich women aged 37–45 with 4 children and a highly educated husband than among married women in general.

7 Analysis of pairs of contingency tables

Up to this point, we have focused on the analysis of single 2×2 contingency tables that are found in a data set by the `4ft-Miner` module of the LISp-Miner system. In this section, we propose statistical models to interpret the output of the LISp-Miner system's `SD4ft-Miner` procedure, which finds *pairs* of contingency tables from two sub-populations of the data. Sub-populations are disjoint sets, for example 'young people' and 'old people' in a database of customer information. Attributes φ and ψ in the two sub-populations can be represented by a pair of contingency tables of the form

$$\mathbf{a}_1 = \begin{array}{c|cc} & \psi & \neg\psi \\ \hline \varphi & a_1 & b_1 \\ \neg\varphi & c_1 & d_1 \end{array}, \qquad \mathbf{a}_2 = \begin{array}{c|cc} & \psi & \neg\psi \\ \hline \varphi & a_2 & b_2 \\ \neg\varphi & c_2 & d_2 \end{array}. \tag{32}$$

The subpopulation sizes are $m_1 = a_1+b_1+c_1+d_1$ and $m_2 = a_2+b_2+c_2+d_2$.

The `SD4ft-Miner` procedure finds pairs of contingency tables that show significant *differences* in generalised quantifiers. The current version of `SD4ft-Miner` supports four of the generalised quantifiers described in section 2: founded implication, double implication, founded equivalence, and above average. In particular, in the procedure for the difference of founded implication quantifiers, a hypothesis related to the subpopulations is labeled `TRUE` if

$$\left|\frac{a_1}{a_1+b_1}-\frac{a_2}{a_2+b_2}\right| \geq p \text{ and } a_1 \geq \text{BASE}_1 \text{ and } a_2 \geq \text{BASE}_2, \tag{33}$$

where $0 < p \leq 1$, $\text{BASE}_1 > 0$, and $\text{BASE}_2 > 0$. The interpretation is then something like "the proportion of ψ among ϕ in subpopulation 1 differs from the proportion in subpopulation 2". Simple tutorial examples can be found in (Rauch and Šimůnek, 2009, 2012).

The statistical model for single contingency tables of section 4 is readily extended to apply to subpopulations. We introduce parameter sets $\theta^k = (\theta^k_1, \theta^k_2, \theta^k_3, \theta^k_4)$ for both subpopulations $k \in \{1,2\}$, such that the conditional probability of occurrence of an attribute combination X_j in an observation Y^k_i is

$$\text{Prob}(Y^k_i = X_j \,|\, \theta) = \theta^k_j,$$
$$i \in \{1,\ldots,m\},\ j \in \{1,2,3,4\},\ k \in \{1,2\}.$$

The parameters satisfy $\theta_j^k \geq 0$ and $\sum_{j=1}^4 \theta_j^k = 1$ for both subpopulations $k \in \{1,2\}$. Assuming the observations to be independent given the θ's, the sampling model (likelihood) for the contingency tables is

$$p(\mathbf{a}_1, \mathbf{a}_2 \,|\, \theta^1, \theta^2) = p(\mathbf{a}_1 \,|\, \theta^1) p(\mathbf{a}_2 \,|\, \theta^2),$$
$$\mathbf{a}_1 \,|\, \theta^1 \sim \text{Multinomial}(\theta^1, m_1),$$
$$\mathbf{a}_2 \,|\, \theta^2 \sim \text{Multinomial}(\theta^2, m_2).$$

Assuming independent priors of the form

$$\theta^k \sim \text{Dirichlet}(\alpha_k', \beta_k', \gamma_k', \delta_k'),$$

the complete posterior is obtained by Bayes' rule as

$$p(\theta^1, \theta^2 \,|\, \mathbf{a}_1, \mathbf{a}_2) = p(\theta^1 \,|\, \mathbf{a}_1) p(\theta^2 \,|\, \mathbf{a}_2),$$
$$\theta^1 \,|\, \mathbf{a}_1 \sim \text{Dirichlet}(\alpha_1, \beta_1, \gamma_1, \delta_1),$$
$$\theta^2 \,|\, \mathbf{a}_2 \sim \text{Dirichlet}(\alpha_2, \beta_2, \gamma_2, \delta_2),$$

where $\alpha_k = a_k + \alpha_k'$, $\beta_k = b_k + \beta_k'$, $\gamma_k = c_k + \gamma_k'$, $\delta_k = d_k + \delta_k'$.

The founded implication parameters for the two subpopulations have, by Theorem 3, the posterior distributions

$$\theta_{\text{fi}}^k \,|\, \mathbf{a}_{1:2} \sim \text{Beta}(\alpha_k, \beta_k) \qquad (k \in \{1,2\}).$$

The posterior distribution for the difference $\theta_{\text{fi}}^1 - \theta_{\text{fi}}^2$ is therefore the distribution of the difference of independent beta random variables. The probability density function of this difference can be computed using the convolution integral of Fact 2, or using the hypergeometric function formulas in (Pham-Gia et al., 1993). The probability

$$\text{Prob}(\theta_{\text{fi}}^1 \geq \theta_{\text{fi}}^2 \,|\, \mathbf{a}_{1:2})$$

can be evaluated using the closed-form formulas in (Cook, 2009). Alternatively, approximate values of pdf, probability, or credibility intervals can be rapidly computed using a normal approximation or using Monte Carlo sampling.

8 Conclusions

In this paper we have presented Bayesian statistical methods that can be used to help in the interpretation and presentation of the results of a GUHA data mining analysis. We showed how the truth values of generalised quantifiers can be related to statements about statistical parameters in a sampling model, and presented detailed derivations of the posterior distributions of these parameters. In some cases, we could express the posterior distribution in closed form using standard beta distributions, but in all cases statistical analysis (plotting the pdf, computing the probability of a one-sided hypothesis, computing credibility intervals) can be rapidly computed using straightforward Monte Carlo sampling algorithms.

The principal value of this new post-processing tool is the ability to quantify the *credibility* of an inference that is made from contingency tables. An analyst with basic understanding of probability concepts can readily interpret a plot of a probability density function for a parameter that describes the prevalence of some attribute: the peak shows the commonest value, and the width of the peak gives an idea of the possible variability in the estimate. Similarly, credibility intervals (also known as"error bars") express the value *and* the extent of uncertainly of this value.

The paper also presents some initial results for the interpretation of GUHA results for two subpopulations. Further work in this area could be done in developing statistical models corresponding to more advanced data mining procedures for comparing subpopulations. The `Ac4ft-Miner` module of LISp-Miner uses the concept of *actions*, in which a deliberate change in one property or properties leads to a desirable change in another property which could not be influenced directly. For example, "Pro-active lowering of monthly payments of loans" could imply "The number of payment delinquencies decreases". For details see (Dardzinska, 2013; Ras and Wieczorkowska, 2000).

References

N. Balakrishnan and V. B. Nevzorov. *A Primer on Statistical Distributions*. John Wiley & Sons, Inc, 2003.

D. A. Berry. *Statistics: A Bayesian Perspective*. Duxberry Press, 1996.

W. Bolstad. *Introduction to Bayesian Statistics*. John Wiley & Sons, Inc, 2nd edition, 2007.

J. D. Cook. Exact calculation of beta inequalities. Technical Report 54, University of Texax M. D. Anderson Cancer Center Department of Biostatistics, 2009. `http://biostats.bepress.com/mdandersonbiostat/paper54`.

R. Cools. An encyclopaedia of cubature formulas. *J. Complexity*, 19:445–453, 2003.

A. Dardzinska. *Action Rules Mining*, volume 468 of *Studies in Computational Intelligence*. Springer, 2013.

L. Devroye. *Non-Uniform Random Variate Generation*. Springer, New York, 1986. Web Edition `http://www.nrbook.com/devroye/`.

H. Eerola. Lääketieteellisen datan analysointia GUHA-tiedonlouhintamenetelmällä (in Finnish). Master's thesis, Tampere University of Technology, 2009.

B. Frigyik, A. Kapila, and M. Gupta. Introduction to the Dirichlet distribution and related processes. Technical Report UWEETR-2010-0006, University of Washington Information Design Lab, 2010. `http://ee.washington.edu/research/guptalab/publications/UWEETR-2010-0006.pdf`.

P. Hájek and T. Havránek. *Mechanizing hypothesis formation: mathematical foundations for a general theory*. Springer, 1978. http://www.cs.cas.cz/hajek/guhabook/.

P. Hájek, I. Havel, and M. Chytil. The GUHA method of automatic hypotheses determination. *Computing*, 1:293–308, 1966. ISSN 0010-485X. doi: 10.1007/BF02345483.

P. Hájek, M. Holeňa, and J. Rauch. The GUHA method and its meaning for data mining. *Journal of Computer and System Sciences*, 76(1):34–48, 2010. ISSN 0022-0000. doi: 10.1016/j.jcss.2009.05.004.

R. Hubbard. The widespread misinterpretation of *p*-values as error probabilities. *Journal of Applied Statistics*, 38 (11):2617–2626, Nov. 2011. ISSN 0266-4763 (print), 1360-0532 (electronic). doi: http://dx.doi.org/10.1080/02664763.2011.567245.

S. Kotz, N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions, Volume 1: Models and Applications*. John Wiley & Sons, Inc, second edition, 2000.

P. M. Lee. *Bayesian Statistics: An Introduction*. Wiley, 2012.

P. Myllymäki, T. Silander, H. Tirri, and P. Uronen. B-course contraceptive method choice dataset, 2002. http://b-course.cs.helsinki.fi/obc/cmcexpl.html.

K. W. Ng, G. Tian, and M. Tang. *Dirichlet and Related Distributions*. John Wiley & Sons, Ltd, 2011.

T. Pham-Gia, N. Turkkan, and P. Eng. Bayesian analysis of the difference of two proportions. *Communications in Statistics - Theory and Methods*, 22(6):1755–1771, 1993.

R. Piché and E. Turunen. Bayesian assaying of GUHA nuggets. In E. Hüllermeier, R. Kruse, and F. Hoffmann, editors, *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods*, volume 80 of *Communications in Computer and Information Science*, pages 348–355, 2010. doi: 10.1007/978-3-642-14055-6.

Z. Ras and A. Wieczorkowska. Action-rules: How to increase profit of a company. In D. Zighed, J. Komorowski, and J. Zytkow, editors, *Principles of Data Mining and Knowledge Discovery*, volume 1910 of *Lecture Notes in Computer Science*, pages 75–116. Springer, 2000. ISBN 978-3-540-41066-9. doi: 10.1007/3-540-45372-5_70.

J. Rauch. Logic of association rules. *Applied Intelligence*, 22:9–28, 2005.

J. Rauch. Considerations on logical calculi for dealing with knowledge in data mining online. *Applied Intelligence*, 22:177–201, 2009.

J. Rauch. *Observational Calculi and Association Rules*. Studies in Computational Intelligence. Springer, 2013.

J. Rauch and M. Šimůnek. An alternative approach to mining association rules. In T. Young Lin, S. Ohsuga, C.-J. Liau, X. Hu, and S. Tsumoto, editors, *Foundations of Data Mining and Knowledge Discovery*, volume 6 of *Studies in Computational Intelligence*, pages 211–231. Springer, 2005. ISBN 978-3-540-26257-2. doi: 10.1007/11498186_13.

J. Rauch and M. Šimůnek. Dealing with background knowledge in the sewebar project. In B. Berendt, D. Mladenic, M. de Gemmis, G. Semeraro, M. Spiliopoulou, G. Stumme, V. Svatek, and F. Železnỳ, editors, *Knowledge Discovery Enhanced with Semantic and Social Information*, pages 89–106. Springer, 2009.

J. Rauch and M. Šimůnek. LISp-Miner project homepage, 2012. URL http://lispminer.vse.cz/. [Online; accessed 21-Sep-2012].

G. Roussas. *A Course in Mathematical Statistics*. Academic Press, second edition, 1997.

E. Turunen. The GUHA method in data mining. Lecture notes, Tampere University of Technology, 2012. http://URN.fi/URN:NBN:fi:tty-201209261292.

M. Šimůnek. Academic KDD project LISp-Miner. In A. Abraham, K. Franke, and K. Koppen, editors, *Intelligent Systems Design and Applications*, Advances in Soft Computing, pages 263–272. Springer, 2003.

A.-M. Šimundić and N. Nikolac. Statistical errors in manuscripts submitted to biochemia medica journal. *Biechemia Medica*, 19(3):294–300, 2009.