Forensic Science International xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Evidence evaluation for discrete data☆

Colin Aitken [a,*], Erica Gold [b]

[a] The School of Mathematics and Maxwell Institute, The University of Edinburgh, Edinburgh EH9 3JZ, United Kingdom
[b] Department of Language and Linguistic Science, University of York, YO10 5DD, United Kingdom

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Evidence evaluation
Likelihood ratio
Discrete data
Forensic phonetics

ABSTRACT

Methods for the evaluation of evidence in the form of measurements by means of the likelihood ratio are becoming more widespread. There is a paucity of methods for the evaluation of evidence in the form of counts by means of the likelihood ratio. Two suggestions for such methods are described. Examples of their performance are illustrated in the context of a problem in forensic phonetics. There is discussion of the problems particular to the evaluation of evidence for discrete data, with suggestions for further work.

1. Introduction

The interpretation of scientific evidence may be thought of as the assessment of a comparison. This comparison is that between evidential material whose source is known, *control* material (denote this by M_c) and evidential material whose source is unknown, *recovered* material (denote this by M_r). Denote the combination by $M = (M_c, M_r)$. Consider semen stains found on a rape victim. These stains are the recovered material M_r, they are assumed to come from the criminal but his identity is unknown. A suspect is identified. A sample is taken for the purposes of identifying his DNA. This sample is the control material M_c as its source is known. Alternatively, consider glass broken during the commission of a crime. M_c would be the fragments of glass (the source form of the material) found at the crime scene, M_r would be fragments of glass (the receptor or transferred particle form of the material) found on the clothing of a suspect and M would be the two sets of fragments.

Discrete data, such as those obtained from DNA profiles, or measurements, such as the refractive indices of the glass fragments are taken from M. These are observations of the evidential material, *evidential observations*, and comparisons are made of the control form and the recovered form. Denote these observations by E_c and E_r, for control and recovered observations, respectively, and let $E = (E_c, E_r)$ denote the combined set. Comparison of E_c and E_r is to be made and the assessment of this comparison has to be quantified. The totality of the evidence is denoted Ev and is such that $Ev = (M, E)$.

The evidence Ev is evaluated by its effect on the odds in favour of a proposition put forward by the prosecution H_p compared with a proposition put forward by the defence H_d. For current purposes, these are propositions at the source level; *i.e.*, they relate to the source of the evidence only and not to any possible related activity or offence. The odds form of Bayes' Theorem is

$$\frac{Pr(H_p|Ev)}{Pr(H_d|Ev)} = \frac{Pr(Ev|H_p)}{Pr(Ev|H_d)} \times \frac{Pr(H_p)}{Pr(H_d)}.$$

Consider the likelihood ratio $Pr(Ev|H_d)/Pr(Ev|H_d)$ further. This equals

$$\frac{Pr(E|H_p, M)}{Pr(E|H_d, M)} \times \frac{Pr(M|H_p)}{Pr(M|H_d)}.$$

The value of the second ratio, $Pr(M|H_p)/Pr(M|H_d)$, concerns the evidential material and is a matter for subjective judgement. It is not proposed to consider its determination further here. Instead, consideration will be concentrated on

$$\frac{Pr(E|H_p, M)}{Pr(E|H_d, M)}$$

and M will be omitted, for clarity of notation. Also, $Pr(H_p|Ev)/Pr(H_d|Ev)$ which can be written $Pr(H_p|(E, M)/Pr(H_d|(E, M))$, will be written as $Pr(H_p|E)/Pr(H_d|E)$, again for clarity of notation.

The observational evidence E is evaluated by its effect on the odds in favour of a proposition put forward by the prosecution H_p compared with a proposition put forward by the defence H_d. Thus:

☆ 6th European Academy of Forensic Science Conference (EAFS 2012), Guest-edited by Didier Meuwly.

* Corresponding author. Tel.: +44 131 650 4877.
E-mail address: c.g.g.aitken@ed.ac.uk (C. Aitken).

0379-0738/$ – see front matter © 2013 Published by Elsevier Ireland Ltd.
http://dx.doi.org/10.1016/j.forsciint.2013.02.042

$$\frac{Pr(H_p|E)}{Pr(H_d|E)} = \frac{Pr(E|H_p)}{Pr(E|H_d)} \times \frac{Pr(H_p)}{Pr(H_d)}.$$

The evaluation is made at the source level. No consideration is given to activity or offence levels of proposition.

The observational evidence E has been written as (E_c, E_r); denote E_c by X and E_r by Y. The control data X, is evidence whose source is known; the recovered data, Y, is evidence whose source is unknown. The statistic used to evaluate the evidence is the likelihood ratio

$$\mathrm{LR} = \frac{Pr(E|H_p)}{Pr(E|H_d)} = \frac{Pr(X,Y|H_p)}{Pr(X,Y|H_d)}.$$

Likelihood ratios greater than one support the prosecution proposition. The evidence is more likely if the prosecution's proposition is true than if the defence proposition is true. Note, no probabilistic statement is made about the truth of the prosecution or defence propositions.

There are normally two sources of variation. There is variation between items and variation within items. For example, for measurements in the form of elemental composition of glass fragments from windows there is variation between windows and within windows.

Consider a crime in which a window has been broken. Fragments of glass are found at the base of the window which are assumed to come from the window. Measurements of the elemental composition of the glass in these fragments may be denoted X as their source is known. A suspect is found. He has glass fragments on his clothing. Measurements of the elemental composition of the glass in these fragments may be denoted Y as their source is unknown. The fragments may have come from the window at the crime scene but they may have come from somewhere else. A characteristic of the glass from a particular window is the mean elemental composition of the glass; such a characteristic is an example of a parameter and is here denoted θ. This characteristic varies from window to window and it is this variation which is known as between-item variation. Variation in the measurements X and in the measurements Y is known as within-item variation.

Between-item variation in θ is represented by a probability density function $f(\theta)$. Within-item variation depends on the particular value of θ for the window in question and this dependency is reflected in the notation. Within-item variation for X and Y is represented by the probability density functions $f(x|\theta)$ and $f(y|\theta)$.

As the data are measurements, probabilities are replaced by probability density functions. The likelihood ratio is

$$\begin{aligned}\mathrm{LR} &= \frac{Pr(X,Y|H_p)}{Pr(X,Y|H_d)} \\ &= \frac{\int f(x,y|\theta)\, f(\theta) d\theta}{\int f(x|\theta)\, f(\theta) d\theta \int f(y|\theta)\, f(\theta) d\theta} \\ &= \frac{\int f(x|\theta)\, f(y|\theta)\, f(\theta) d\theta}{\int f(x|\theta)\, f(\theta) d\theta \int f(y|\theta)\, f(\theta) d\theta},\end{aligned} \quad (1)$$

since X and Y are independent if H_d is true as they come from different sources and are conditionally independent given θ if H_p true. See [1] for further details. Many methods have been developed for the likelihood ratio when the data are in the form of measurements. The methods enable assessments of rarity and similarity of the control and recovered evidence to be made simultaneously with the one statistic.

Methods based on significance testing are two-stage methods. Assess similarity first. If the control and recovered measurements are assessed as similar in some sense, such as with the use of a significance test, then rarity is assessed in a second stage. Similarity in measurements which are rare is taken to be stronger evidence in support of a common source then similarity in measurements which are common. However, this two-stage approach is difficult to interpret, its evaluation is difficult to combine with evaluations of other evidence and is subject to what is known as the effect of falling off a cliff: a comparison which is just significant, and hence leads to the discarding of the evidence has a very different outcome to a comparison which is just not significant and leads to progression to the assessment of rarity. In contrast, the likelihood ratio approach provides an evaluation which is easy to interpret, which is easily combined with evaluations from other evidence and provides a continuous measure of value with no cliff.

Other than for DNA profiling, there is a paucity of methods when the data X and Y are discrete. Often with continuous data it is possible to assume Normality. Theory then allows for multivariate continuous data to be modelled using the means and covariances only. Thus, for k variables there are k means, k variances and $k(k-1)/2$ covariances. This may be too many parameters to estimate robustly and methods exist based on graphical models to reduce the number of parameters to be estimated [2].

Such a summary of the distribution is not possible with discrete data. Each variable with a discrete response, may have several levels of response. Assume there are k variables and that the ith variable has m_i levels of response with corresponding probabilities p_{i_1} to $p_{i_{m_i}}$ (which sum to 1). There are then $\prod_{i=1}^{k}(m_i - 1)$ marginal probabilities to estimate along with interactions which may be pair-wise, or in triples or of higher-order. If the variables are independent, then relatively elementary models for counts may be applied to each variable and the overall probability taken to be the product of the individual probabilities but such simplicity is not possible with dependency amongst the variables or observations.

Another set of problems arises if there is autocorrelation between adjacent observations. Autocorrelation arises in a financial context, for example, when the value of an economic indicator of performance at a particular time point may depend on (be correlated with) the value it took at the immediately previous time point (a dependence which is said to be an autocorrelation with lag 1). If there is also dependence on the value at the time before that, the autocorrelation is at lag 2, with a continuation back with higher lags, depending on the phenomenon which generated the data. Autocorrelation also arises in a forensic scientific context. Experiments carried out to analyse drugs on banknotes indicated that it was possible for drug traces to pass from one contaminated banknote to an adjacent one [3]. For continuous data, a single parameter may be used to model autocorrelation at a particular lag. For discrete data, the procedure is not so simple. For each value taken by the discrete random variable there may be a separate probability for its dependence on the value taken at the previous time point. Thus, if there are m categories for a variable, there may be $m(m+1)/2$ probabilities to consider, a considerably more daunting estimation problem than with continuous data.

The methods described here are motivated by a problem in forensic phonetics being investigated in The University of York under the aegis of the Bayesian Biometrics for Forensics Network (BBFOR2).[1] The data are the number of 'clicks' (a parameter that can be analysed in speech) in each of a succession of minutes ranging from four to six. A click is defined as 'a stop made with an ingressive velaric airstream, such as Zulu [|||]' [4]. Further, 'a major ingredient in the production of the airstream [for clicks] is a complete closure made by the back of the tongue against the

[1] BBFOR2 is an FP7 Marie Curie Initial Training Network that is working in the areas of speaker recognition (comparison), face recognition and fingerprint recognition. These disciplines are being studied both individually as well as in combination.

Please cite this article in press as: C. Aitken, E. Gold, Evidence evaluation for discrete data, Forensic Sci. Int. (2013), http://dx.doi.org/10.1016/j.forsciint.2013.02.042

velum. A second closure is also made, further forward in the mouth, either by the tip, blade or front of the tongue, or by the lips' [5]. The closure at the back of the tongue remains in contact with the soft palate, and the front portion of the tongue is then drawn downwards. This process increases the volume of air trapped in between the two closures 'rarefying the intra-oral air-pressure. When the more forward of the two closures is released, the outside air at atmospheric pressure flows in to fill the partial vacuum' [5]. It is at this point that a click is realised.

Clicks are most commonly recognized for their existence in a number of African languages [4, p. 139] where they are used phonemically [5, p. 174]. However, clicks are also found in English, but unlike those in African languages they are not used phonemically. In English, clicks can be analysed in speech on a discourse or paralinguistic level for forensic speaker comparison purposes.

The research uses speech material from the DyViS (Dynamic Variability in Speech) database [6]. The corpus is comprised of one hundred males of Standard Southern British English ranging in age from 18 to 25 years old. The database includes different speaking styles that simulate forensic conditions, and the current study is concerned with spontaneous speech recorded from the speakers during simulated mock phone calls with an accomplice (the same interlocutor was used for all 100 recordings).

The data are counts but they are not Poisson as the means and variances for the responses of the 100 subjects in the experiment are not equal. There are many zeros, minutes in which there were no clicks. Also, the responses from one time period to the next time period are not independent; the number of clicks in one minute can easily be shown to be dependent on the number of clicks in the immediately previous minute. The forensic example given here (speech) provides data which do not act in the same way as other forensic disciplines. Statistical models are required which are additional to those that are currently being used for continuous data. Forensic speech science works with qualitative data. The context is used to build a foundation on which basic models for discrete data may be developed and hence motivate further research.

The methods described here are not simple in themselves but are designed for simple situations. It is not expected that appropriate data for these situations will be found in most forensic science problems. Statistical models are required additional to those that are currently being used for continuous data.

Two simple situations are described to show how evidence in the form of discrete data may be evaluated. The discussion is somewhat abstract and it is intended to illustrate principles against which particular problems may be tested. The first situation assumes independence between observations on a particular item of evidence and a Poisson model with the same mean, λ, for each observation on this item. The formula for the Poisson distribution is given in Appendix A. Data are in the form of counts. Consider control evidence which has a succession of independent observations with a Poisson model, with mean λ_c (c for control) on a particular item. There is recovered evidence which also has a succession of independent observations with a Poisson model with mean λ_r (r for recovered) on a particular item. If these two items come from the same source, an assumption which is normally the prosecution proposition H_p, $\lambda_c = \lambda_r$. If these two items come from different sources, an assumption which is normally the defence proposition H_d, λ_c may or may not equal λ_r. There is a population of relevant items ('relevant' in some forensic sense), an example of which would be the speech of different people in phonetics, and each of these items has a mean λ associated with it. There is variability in λ across the population; the mean number λ of counts for an item in the population varies from item to item. If there were N items, indexed by i with $i = 1, \ldots, N$ then item i may be said to have mean λ_i, $i = 1, \ldots, N$. In the simple situation described here, the variation in λ across the population is taken to have a gamma distribution. The gamma distribution is characterised by two parameters, conventionally denoted α and β. The probability density function of the gamma distribution is given in Appendix A.1. Various forms of the gamma probability density function are illustrated in Fig. 1. The mean of the distribution is α/β and the variance α/β^2. Given values (perhaps estimates) of the means $\{\lambda_i, i = 1, \ldots, N\}$ in the individual items in the population, α and β may be estimated by equating the overall mean of the λ_i's to α/β and the overall variance of the λ_i's to α/β^2, a procedure known as the method of moments. In the example given here data are not available for such a procedure. Instead, subjective choices are made for α and β and some commentary is given as to how the values chosen reflect certain interpretations in the variation of the characteristic across items in the population.

The second simple situation assumes a dependency between adjacent observations. The observations are taken to be binary in nature: presence (*e.g.*, at least one click) or absence of a characteristic normally denoted 1 and 0, respectively. Thus, there are 2 categories, hence $m = 2$ and there are $m(m + 1)/2 = 3$ probabilities to consider. These are

(a) the probability of the presence of a characteristic in the first member of a pair;
(b) the probability of the presence of a characteristic in the second member of a pair given the presence of the characteristic in the first member of the pair;
(c) the probability of the presence of a characteristic in the second member of a pair given the absence of the characteristic in the first member of the pair.

The characteristic is a binary variable, the responses are presence and absence. Thus if the probability of a presence is known, the probability of an absence is also known since the sum of the two probabilities is equal to 1.

With an increase of only one additional category to $c = 3$ there would be an increase to 6 probabilities to consider. Two pairs of observations per source are considered. There is control evidence which has two pairs of independent observations with a bivariate binary model with parameter θ_c (c for control) on a particular item. There is recovered evidence which has two pairs of independent observations with a bivariate binary model with parameter θ_r (r for recovered) on a particular item. If these two items come from the same source, normally the prosecution proposition H_p, $\theta_c = \theta_r$. If these two items come from different sources, normally the defence proposition H_d, θ_c may or may not equal θ_r. The parameter θ has three components to it, one for each of the three probabilities.

There is a population of relevant items (again, relevant in some forensic sense) and each of these items has a parameter θ associated with it. For example, if there were N items in the population, indexed by $i = 1, \ldots, N$, the parameter for the ith item would be denoted θ_i. The parameters $\{\theta_i, i = 1, \ldots, N\}$ vary across the population. In the simple situation described here, the variations in the three components of θ across the population are taken to have beta distributions. The beta distribution, like the gamma distribution, is characterised by two parameters, and, like the gamma distribution, conventionally denoted α and β. The probability function of the distribution is given in Appendix A.2. Various forms of the beta probability density function are illustrated in Fig. 2. The mean of the distribution is $\alpha/(\alpha + \beta)$ and the variance $\alpha\beta/\{(\alpha + \beta)^2(\alpha + \beta + 1)\}$. In the example given here data are not available for estimation of (α, β) by the method of moments. Instead, as with the gamma distribution previously, subjective choices are made for α and β and, as before, some commentary is given as to how the values chosen reflect certain interpretations in the variation of the characteristic across items in the population.

Please cite this article in press as: C. Aitken, E. Gold, Evidence evaluation for discrete data, Forensic Sci. Int. (2013), http://dx.doi.org/10.1016/j.forsciint.2013.02.042

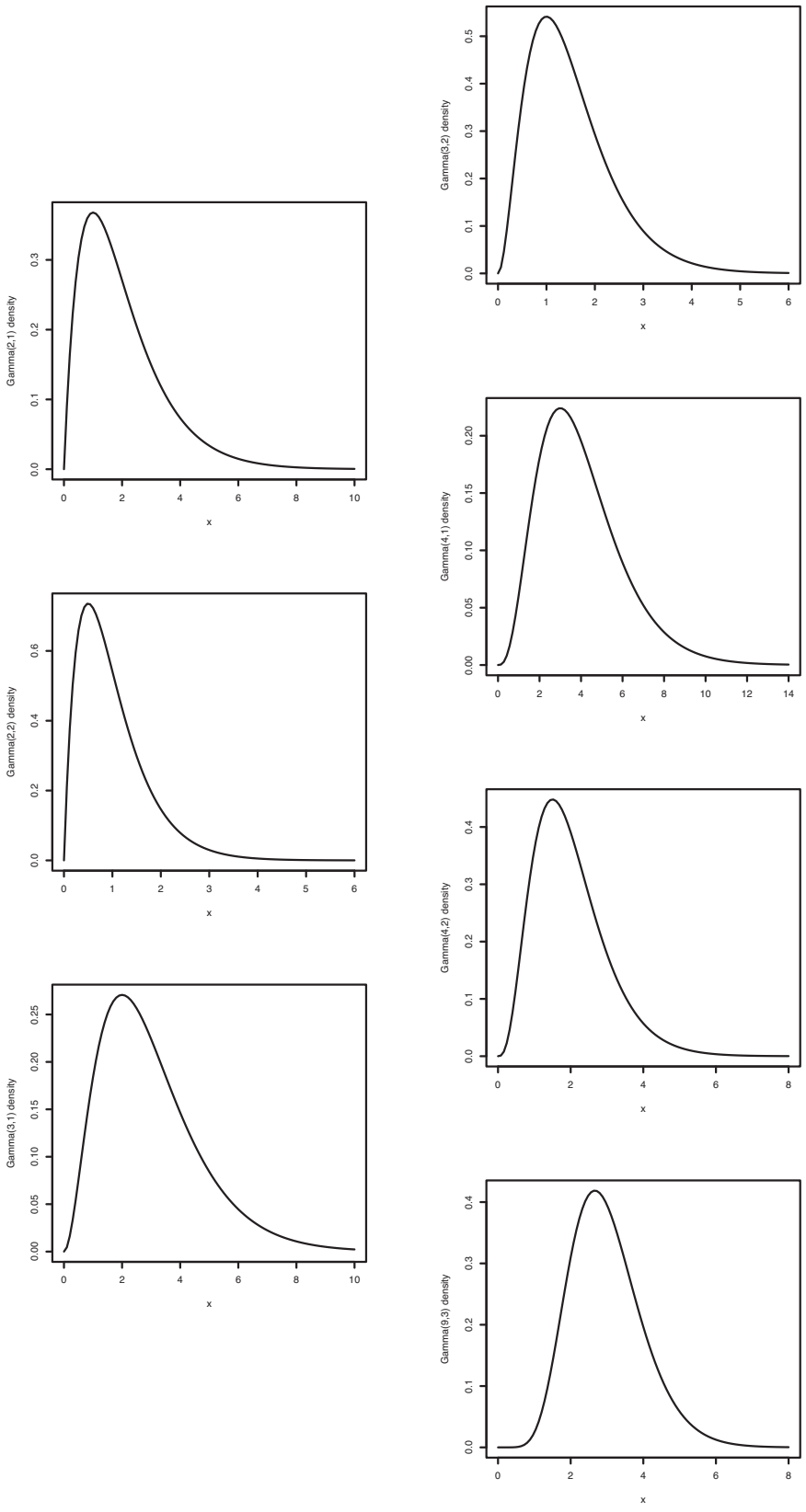

Fig. 1. Probability density functions for the gamma distribution for possible variation in the mean number of counts *x* in a characteristic. Parameter values are shown as a pair of digits on the vertical axis. The overall mean in *x* is the ratio of the first member of the pair to the second member of the pair, the overall variance in *x* is the ratio of the first member to the square of the second member. Note the different scales in the different graphs.

Please cite this article in press as: C. Aitken, E. Gold, Evidence evaluation for discrete data, Forensic Sci. Int. (2013), http://dx.doi.org/10.1016/j.forsciint.2013.02.042

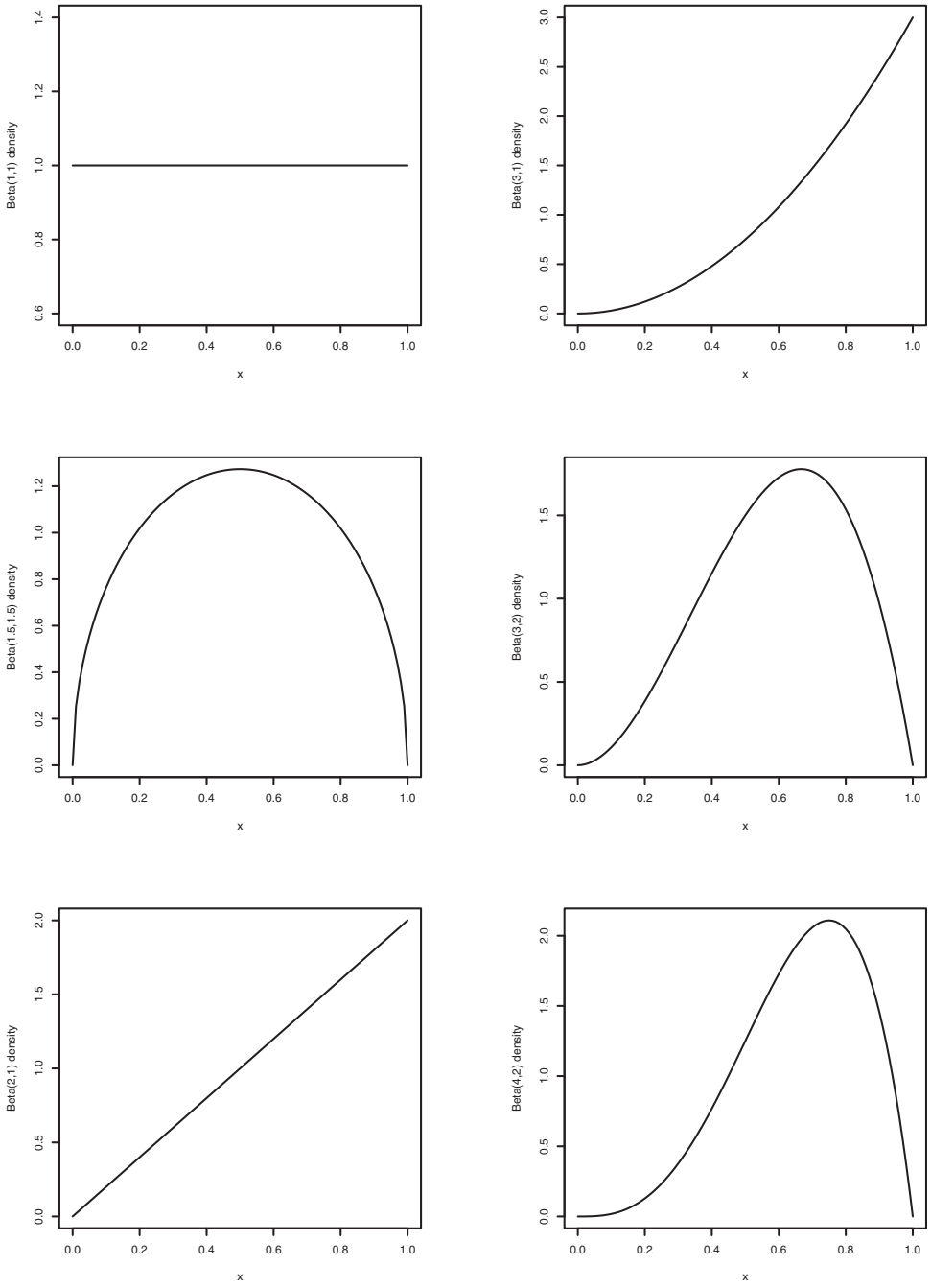

Fig. 2. Probability density functions for the beta distribution. The parameter values are given in pairs in the labels for the vertical axis. Let α and β denote the two digits. The variable x is the probability of a zero. The mean probability of a zero is the ratio of the first parameter value (α) to the sum of the two parameter values ($\alpha + \beta$). The variance in the probability is $\alpha\beta/\{(\alpha + \beta)^2(\alpha + \beta + 1)\}$.

Note that in both the situations described above, if investigation of the values of λ or θ show that the gamma or beta distributions for λ or the components of θ are not appropriate then some suitable non-parametric approach could be investigated.

2. Materials and methods

The formulae for the likelihood ratio of evidence in a form applicable to data of the two forms described above are given below. The technical details of the derivation of the formulae are given in Appendix A.

2.1. Data of independent counts with a Poisson distribution

Consider a crime in which a piece of recorded speech is of importance. A characteristic, S, of the speech is noted. The number of occurrences of S in each of a succession of consecutive time periods, k_y in total, in their speech is noted. It is assumed that these characteristics are independent between time periods and follow a Poisson distribution. These are the recovered data. A suspect is identified and the number of occurrences of S in each of a succession of consecutive time periods, k_x in total, in their speech is noted. These are the control data.

Assume the time periods are minutes. Let the number of occurrences of S per minute for the control speech be $\mathbf{x} = (x_1, \ldots, x_{k_x})$ and for the recovered speech be $\mathbf{y} = (x_1, \ldots, y_{k_y})$.

Let $t_x = \sum_{i=1}^{k_x} x_i$ and $t_y = \sum_{i=1}^{k_y} y_i$. Then it is shown in Appendix A.1 that the likelihood ratio for the value, V, of the evidence $\mathbf{x}$ and $\mathbf{y}$ is

$$V = \frac{\Gamma(\alpha + t_x + t_y)\Gamma(\alpha)}{\Gamma(\alpha + t_x)\Gamma(\alpha + t_y)} \times \frac{(\beta + k_x)^{\alpha + t_x}(\beta + k_y)^{\alpha + t_y}}{\beta^{\alpha}(\beta + k_x + k_y)^{\alpha + t_x + t_y}}. \quad (2)$$

Note that the numbers of minutes of recording, k_x and k_y, and the total number of occurrences of S, t_x and t_y, are all the data from the evidence that are required. The individual records $(x_1, \ldots, x_{k_x})$ and $(y_1, \ldots, y_{k_y})$ are not needed for evaluation of the

Please cite this article in press as: C. Aitken, E. Gold, Evidence evaluation for discrete data, Forensic Sci. Int. (2013), http://dx.doi.org/10.1016/j.forsciint.2013.02.042

Table 1
Expression for the component terms in the likelihood ratio (3) based on the bivariate Bernoulli model. The control data, whose source is known, are $\mathbf{x} = (x_{i1}, x_{i2})$, $i = 1, 2$, for independent periods $i = 1, 2$ where x_{ij} ($i = 1, 2, j = 1, 2$) = 0 or 1 according as whether the characteristic is absent or present. The recovered data, whose source is unknown, are $\mathbf{y} = (y_{i1}, y_{i2})$, for independent periods $i = 1, 2$ where y_{ij}($i = 1, 2, j = 1, 2$) = 0 or 1 according as whether the characteristic is absent or present. The probability of an absence is denoted θ and the probability of a presence is $(1 - \theta)$. Subscripts are introduced to indicate the particular circumstance of the absence (0) or presence (1) of a click and θ denotes $(\theta_0, \theta_{00}, \theta_{10})$. Independent beta$(\alpha, \beta)$ distributions are assumed for $\theta_0, \theta_{00}, \theta_{10}$; the subscripts indicate the particular circumstance for the prior. The parameters are (α_0, β_0), $(\alpha_{00}, \beta_{00})$, $(\alpha_{10}, \beta_{10})$, respectively.

const	$\Gamma(\alpha_0+\beta_0)\Gamma(\alpha_{00}+\beta_{00})\Gamma(\alpha_{10}+\beta_{10})/\{\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma(\alpha_{00})\Gamma(\beta_{00})\Gamma(\alpha_{10})\Gamma(\beta_{10})\}$
n_0	$\Gamma(4+\alpha_0-x_{11}-x_{21}-y_{11}-y_{21})\Gamma(\beta_0+x_{11}+x_{21}+y_{11}+y_{21})/\{\Gamma(4+\alpha_0+\beta_0)\}$
n_{00}	$\{\Gamma(\alpha_{00}+(1-x_{11})(1-x_{12})+(1-x_{21})(1-x_{22})+(1-y_{11})(1-y_{12})+(1-y_{21})(1-y_{22}))\Gamma(\beta_{00}+(1-x_{11})x_{12}+(1-x_{21})x_{22}+(1-y_{11})y_{12}+(1-y_{21})y_{22})\}/\{\Gamma(4+\alpha_{00}+\beta_{00}-x_{11}-x_{21}-y_{11}-y_{21})\}$
n_{10}	$\{\Gamma(\alpha_{10}+x_{11}(1-x_{12})+x_{21}(1-x_{22})+y_{11}(1-y_{12})+y_{21}(1-y_{22}))\Gamma(\beta_{10}+x_{11}x_{12}+x_{21}x_{22}+y_{11}y_{12}+y_{21}y_{22})\}/\{\Gamma(\alpha_{10}+\beta_{10}+x_{11}+x_{21}+y_{11}+y_{21})\}$
d_{0c}	$\Gamma(2+\alpha_0-x_{11}-x_{21})\Gamma(\beta_0+x_{11}+x_{21})/\{\Gamma(2+\alpha_0+\beta_0)\}$
d_{00c}	$\{\Gamma(\alpha_{00}+(1-x_{11})(1-x_{12})+(1-x_{21})(1-x_{22}))\Gamma(\beta_{00}+(1-x_{11})x_{12}+(1-x_{21})x_{22})\}/\{\Gamma(2+\alpha_{00}+\beta_{00}-x_{11}-x_{21})\}$
d_{10c}	$\Gamma(\alpha_{10}+x_{11}(1-x_{12})+x_{21}(1-x_{22}))\Gamma(\beta_{10}+x_{11}x_{12}+x_{21}x_{22})/\{\Gamma(\alpha_{10}+\beta_{10}+x_{11}+x_{21})\}$
d_{0r}	$\Gamma(2+\alpha_0-y_{11}-y_{21})\Gamma(\beta_0+y_{11}+y_{21})/\{\Gamma(2+\alpha_0+\beta_0)\}$
d_{00r}	$\{\Gamma(\alpha_{00}+(1-y_{11})(1-y_{12})+(1-y_{21})(1-y_{22}))\Gamma(\beta_{00}+(1-y_{11})y_{12}+(1-y_{21})y_{22})\}/\{\Gamma(2+\alpha_{00}+\beta_{00}-y_{11}-y_{21})\}$
d_{10r}	$\Gamma(\alpha_{10}+y_{11}(1-y_{12})+y_{21}(1-y_{22}))\Gamma(\beta_{10}+y_{11}y_{12}+y_{21}y_{22})/\{\Gamma(\alpha_{10}+\beta_{10}+y_{11}+y_{21})\}$

likelihood ratio. This is a consequence of the assumptions of independence between successive time periods and of a Poisson model for the observations.

A theoretical paper [7] describes a considerable extension to this model using a multivariate Poisson model with a component to allow for a number of zero observations greater than expected in such a model (so-called zero-inflation).

2.2. A bivariate Bernoulli model

The previous method based on the Poisson distribution assumes independence between observations. However, this may not always be the case. It may be that the number of occurrences of S in one minute by a particular speaker is dependent on the number of occurrences of S in the previous minute by the same speaker.

As an initial investigation of the effect of dependency, consider a simple model in which the data recorded are binary, presence (at least one occurrence) or absence of S in a minute.

Two independent periods of recording are taken, for example, of two minutes of speech in each period, for control and recovered speech. Let $\mathbf{x} = (x_{i1}, x_{i2})$, $i = 1, 2$ be the control data, whose source is known, for periods $i = 1, 2$ where x_{ij} ($i = 1, 2, j = 1, 2$) = 0 or 1 according as whether the characteristic (for example, click) is absent or present. Let $\mathbf{y} = (y_{i1}, y_{i2})$ be the recovered data, whose source is unknown, for periods $i = 1, 2$ where y_{ij} ($i = 1, 2, j = 1, 2$) = 0 or 1 according as whether S is absent or present. The probability of an absence is denoted θ and the probability of a presence is then $(1 - \theta)$. Subscripts are introduced to indicate the particular circumstance of the absence or presence of S.

Two independent periods for the control and recovered data are assumed in order to develop a model beyond one bivariate binary observation for each source. Thus

$$p(x_{i1} = 0) = p(y_{i1} = 0) = \theta_0, \qquad p(x_{i1} = 1) = p(y_{i1} = 1) = 1 - \theta_0, \quad i = 1, 2;$$

$$p(x_{i2} = 0|x_{i1} = 0) = p(y_{i2} = 0|y_{i1} = 0) = \theta_{00} \qquad i = 1, 2;$$

$$p(x_{i2} = 0|x_{i1} = 1) = p(y_{i2} = 0|y_{i1} = 1) = \theta_{10} \qquad i = 1, 2.$$

Denote $(\theta_0, \theta_{00}, \theta_{10})$ by $\boldsymbol{\theta}$. Assume independent beta(α, β) distributions for θ_0, θ_{00}, θ_{10} where again subscripts are introduced to indicate the particular circumstance for the prior. Thus the parameters are (α_0, β_0), $(\alpha_{00}, \beta_{00})$, $(\alpha_{10}, \beta_{10})$, respectively, denoted in general by $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Parameters (α, β) may be estimated by appropriate method of moments estimators from sample proportions and variances from some relevant population, using the formulae given in the introduction. Alternatively, in the absence of data as here in this artificial example, they may be chosen subjectively to indicate some personal belief in the probabilities of these various circumstances. The likelihood ratio, V, then has the form

$$V = \frac{n_0 \times n_{00} \times n_{10}}{\text{const} \times d_{0c} \times d_{00c} \times d_{10c} \times d_{0r} \times d_{00r} \times d_{10r}} \tag{3}$$

where details of the individual components of (3) are given in Table 1. An outline derivation of the likelihood ratio and the expressions for the component terms are given in Appendix A.2.

3. Results

3.1. Data of independent counts with a Poisson distribution

From (2), the following results are obtained for a comparison of the total number of outcomes for control and recovered data. Here only the sum of the number of observations in the total period is noted. Because of the independence of the counts in each of the intervals the ordering of them does not affect the value of the evidence. In forensic phonetics, the data could be the numbers of clicks in each of k_x and k_y minutes of recorded speech.

Results are given in Tables 2 ($k_x = k_y = 4$) and 3 ($k_x = k_y = 6$) for various values of the prior parameters (α, β) and for various totals $t_x = \sum_{i=1}^{k_x} x_i$ and $t_y = \sum_{i=1}^{k_y} y_i$.

The $E(X)$ and $Var(X)$ denote the mean and variance of the Poisson means over the population of interest. These parameters are derived from the values of α and β from the formulae $E(X) = \alpha/\beta$ and $Var(X) = \alpha/\beta^2$.

Very small values of the evidence, much less than one, occur when a control piece of speech with no clicks in six minutes is

Table 2
Values of evidence (2) for lengths of observations $k_x = k_y = 4$ for various numbers of outcomes of control x and recovered y evidence and various values of parameters (α, β) of the gamma prior distribution.

$t_x = \sum_{i=1}^{k_x} x_i$	$t_y = \sum_{i=1}^{k_y} y_i$	Value of the evidence V (2)						
		$\alpha = 3$	$\alpha = 2$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 4$	$\alpha = 9$
		$\beta = 1$	$\beta = 1$	$\beta = 2$	$\beta = 2$	$\beta = 1$	$\beta = 2$	$\beta = 3$
		$E(X) = 3$	$E(X) = 2$	$E(X) = 1$	$E(X) = 1.5$	$E(X) = 4$	$E(X) = 2$	$E(X) = 3$
		$Var(X) = 3$	$Var(X) = 2$	$Var(X) = 0.5$	$Var(X) = 0.75$	$Var(X) = 4$	$Var(X) = 1$	$Var(X) = 1$
0	0	21.4	7.72	3.24	5.83	59.5	10.5	35.1
4	4	2.72	1.76	1.37	1.37	5.1	1.7	3.5
8	8	1.71	1.71	2.47	1.60	2.2	1.4	1.5
0	4	2.04	0.74	0.42	0.76	5.7	1.4	5.8
0	8	0.19	0.07	0.05	0.10	0.5	0.2	0.9
0	12	0.01	0.007	0.007	0.01	0.05	0.02	0.15

Please cite this article in press as: C. Aitken, E. Gold, Evidence evaluation for discrete data, Forensic Sci. Int. (2013), http://dx.doi.org/10.1016/j.forsciint.2013.02.042

Table 3
Values of evidence (2) for lengths of observations $k_x = k_y = 6$ for various numbers of outcomes of control x and recovered y evidence and various values of parameters (α, β) of the gamma prior distribution.

$t_x = \sum_{i=1}^{k_x} x_i$	$t_y = \sum_{i=1}^{k_y} y_i$	Value of the evidence V (2)						
		$\alpha = 3$	$\alpha = 2$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 4$	$\alpha = 9$
		$\beta = 1$	$\beta = 1$	$\beta = 2$	$\beta = 2$	$\beta = 1$	$\beta = 2$	$\beta = 3$
		$E(X) = 3$	$E(X) = 2$	$E(X) = 1$	$E(X) = 1.5$	$E(X) = 4$	$E(X) = 2$	$E(X) = 3$
		$Var(X) = 3$	$Var(X) = 2$	$Var(X) = 0.5$	$Var(X) = 0.75$	$Var(X) = 4$	$Var(X) = 1$	$Var(X) = 1$
0	0	53.5	14.22	5.22	11.94	201.84	27.30	198.36
4	4	5.3	2.53	1.50	1.90	13.45	2.93	12.25
8	8	2.6	1.92	1.82	1.50	4.62	1.62	3.20
0	4	4.5	1.19	0.56	1.27	16.97	2.91	25.71
0	8	0.4	0.10	0.06	0.14	1.43	0.31	3.33
0	12	0.03	0.008	0.006	0.01	0.12	0.03	0.43

Table 4
Expectations $E(X)$ and variances $Var(X)$ of gamma distributions which may be considered to represent the associated verbal interpretations of the variation of items over members of a relevant population for investigation of a forensic characteristic with a within-item Poisson distribution and between-item gamma distribution. The parameters (α, β) of the associated gamma distributions are derived from the formulae $E(X) = \alpha/\beta$ and $Var(X) = \alpha/\beta^2$.

$E(X)$	$Var(X)$	Interpretation
3	3	For some items the mean number of occurrences per unit period of time will be around 3 but there will be some variation in the item means.
2	2	The mean number of occurrences per unit period of time will be around 2 but there will be some variation in the item means though less than in the case above.
1	1/2	The mean number of occurrences per unit period of time will be around 1 with very little variation about this.
1 1/2	3/4	The mean number of occurrences per unit period of time will be around 1 1/2 with little variation about this.
4	4	The mean number of occurrences per unit period of time will be around 4 with a lot of variation about this.
2	1	The mean number of occurrences per unit period of time will be around 2 with not much variation about this.
3	1	The mean number of occurrences per unit period of time will be around 3 with not much variation about this.

compared with a recovered piece of speech with twelve clicks in six minutes. For example, $V = 0.007 \simeq 1/140$ when $t_x = 0$, $t_y = 12$; $k_x = k_y = 4$; $\alpha = 2$, $\beta = 1$, 2 and $V = 0.006 \simeq 1/170$ when $t_x = 0$, $t_y = 12$; $k_x = k_y = 6$; $\alpha = 2$, $\beta = 2$. All of these results provide support for the proposition of different sources for the speech: the evidence is 140 (170) times more likely if the two pieces of speech (**x** and **y**) were uttered by different people than by the same person.

In general, an increase in the number of intervals observed from 4 to 6 leads to an increase in the support for a proposition of the same source when there is a match $t_x = t_y$; compare rows 1, 2 and 3 of Tables 1 and 2. The highest values, 59.5 ($k_x = k_y = 4$) and 201.84 ($k_x = k_y = 6$), occur when there is complete match of rare occurrences, all zeros with a mean for occurrences per minute of 4, albeit with a high variance of 4. The next highest values, 35.1 ($k_x = k_y = 4$) and 198.36 ($k_x = k_y = 6$), occur when the mean is 3 and the variance 1. The lowest value, 5.22, for row 1 (all zeros in both the control and recovered speech) arises from the lowest value of $E(X)$, the mean, which equals 1. In such a situation, an absence of clicks is not unusual.

An interpretation of the expectations and variances from which the seven choices of parameter values were obtained is given in Table 4.

The corresponding graphs for these probability density functions are shown in Fig. 1.

3.2. Bivariate Bernoulli model

Results are given in Table 5 of applications of (3) for various combinations of bivariate Bernoulli models and prior parameters.

In Table 5, and a uniform prior (LR1) the results when all the observations are 0, row 1, and when all the observations are 1, row 2, are the same, 3.24 as there is no prior distinction between the probability of the presence or absence of a click.

The graphs of the beta distribution in Fig. 2 show how the probability distribution for θ changes as the values of the parameters change. A curve which increases from left to right as in Beta(2,1) and Beta(3,1) suggests a high belief in a high probability of a 0 observation (the variable x is the probability of a zero, absence of a characteristic). This results in a lower likelihood ratio when the data are all zeros, compared with the value obtained with a uniform prior, as a match in zeros is then more common in the former cases. Likelihood ratios less than 1

Table 5
Values of the likelihood ratio (3) for given control (x_{11}, x_{12}), (x_{21}, x_{22}) and recovered (y_{11}, y_{12}), (y_{21}, y_{22}) observations and for three different sets of prior parameter values. [LR1] Uniform priors: $\alpha_0 = \beta_0 = \alpha_{00} = \beta_{00} = \alpha_{10} = \beta_{10} = 1$.: no preference given to any particular set of values for the probability of a zero. [LR2] $\alpha_0 = 2$, $\beta_0 = 1$, $\alpha_{00} = 2$, $\beta_{00} = 1$, $\alpha_{10} = 1.5$, $\beta_{10} = 2.5$: more weight to zero in first place, to zero in second place given zero in first place and to one in first place given one in first place. [LR3] $\alpha_0 = 3$, $\beta_0 = 1$, $\alpha_{00} = 3$, $\beta_{00} = 1$, $\alpha_{10} = 1.5$, $\beta_{10} = 2.5$: more weight to zero in first place, to zero in second place given zero in first place and to one in first place given one in first place.

Row	(x_{11}, x_{12})	(x_{21}, x_{22})	(y_{11}, y_{12})	(y_{21}, y_{22})	Likelihood ratio values		
					LR1	LR2	LR3
1	(0, 0)	(0, 0)	(0, 0)	(0, 0)	3.24	1.78	1.42
2	(1, 1)	(1, 1)	(1, 1)	(1, 1)	3.24	3.23	3.35
3	(0, 0)	(1, 1)	(0, 0)	(1, 1)	2.13	1.51	1.52
4	(0, 0)	(0, 0)	(1, 1)	(1, 1)	0.30	0.40	0.48
5	(1, 0)	(0, 1)	(0, 0)	(1, 1)	0.53	0.72	0.81
6	(0, 0)	(0, 1)	(0, 0)	(0, 1)	2.16	1.60	0.94
7	(1, 0)	(0, 1)	(0, 0)	(0, 0)	0.53	0.48	0.53

Please cite this article in press as: C. Aitken, E. Gold, Evidence evaluation for discrete data, Forensic Sci. Int. (2013), http://dx.doi.org/10.1016/j.forsciint.2013.02.042

occur when there is a mismatch between outcomes as illustrated in rows 4, 5 and 7. Row 6 shows two values greater than one and one value less than one, illustrating the importance of prior values in situations with few data.

4. Discussion

The two models described here are basic models, the exact situations for which will rarely occur in practice. However, they illustrate issues that need to be considered in the analysis of discrete data and provide a foundation on which other models may be built.

The values obtained of the likelihood ratio are small but intuitively sensible. The size of the likelihood ratios is a function of the smallness of the artificial data sets used. The sets are deliberately small to enable the calculations to be done with very few lines of computer code, or in individual cases, with a pocket calculator.

The smallness of the datasets means that the choice of the prior parameters makes a big difference to the values of the likelihood ratio. The formulae (2) and (3) explain exactly how the values of the prior parameters affect the values of the likelihood ratios.

The model based on independent Poisson counts is likely to be more realisable in practice than the bivariate Bernoulli model. As can be seen here, given the complicated nature of (3), extension to more than two variables and more than two categories will lead to a more complicated model and a requirement to consider more prior parameters with an associated decrease in the robustness of the model.

More practical work is needed to collect data sets appropriate for analyses by these models, or extensions of them, and then to interpretation of the results.

Acknowledgement

Support for EG from the Bayesian Biometrics for Forensics Network, Marie Curie Actions EC Grant Agreement Number PITN-GA-2009-238803, is acknowledged.

Appendix A

A.1. Derivation of the likelihood ratio for independent Poisson counts with a gamma prior distribution

The recovered data are $\mathbf{Y} = \mathbf{y} = (y_1, \ldots, y_{k_y})$ where k_y are the number of separate items in the set; denote $\sum_{i=1}^{k_y} y_i$ by t_y. The control data are $\mathbf{X} = \mathbf{x} = (x_1, \ldots, x_{k_x})$ where k_x are the number of separate items in the set; denote $\sum_{i=1}^{k_x} x_i$ by t_x. It is assumed these data, $\mathbf{X}$ and $\mathbf{Y}$, are observations from a Poisson distribution. The mean λ of the Poisson distribution may be modelled by a gamma distribution with parameters (α, β).

Consider (1). The prosecution proposition H_p is that the control and recovered items come from the same source (such as the person making the two pieces of speech). The defence proposition H_d is that they come from different sources (different people are making the two pieces of speech). Replace θ in (1) with λ. Also, the data are counts. Thus $f(x|\theta)$ and $f(y|\theta)$ are rewritten as $Pr(\mathbf{X} = \mathbf{x}|\lambda)$ and $Pr(\mathbf{Y} = \mathbf{y}|\lambda)$ where the bold notation is used to indicate there are several observations for $\mathbf{x}$ and $\mathbf{y}$. The probability density function $f(\theta)$ from (1) is replaced with $f(\lambda|\alpha, \beta)$. The likelihood ratio (1) may then be written as

$$\mathrm{LR} = \frac{\int Pr(\mathbf{X} = \mathbf{x}|\lambda) Pr(\mathbf{Y} = \mathbf{y}|\lambda) f(\lambda|\alpha, \beta) d\lambda}{\int Pr(\mathbf{X} = \mathbf{x}|\lambda) f(\lambda|\alpha, \beta) d\lambda \int Pr(\mathbf{Y} = \mathbf{y}|\lambda) f(\lambda|\alpha, \beta) d\lambda}.$$

Consider first $Pr(\mathbf{X} = \mathbf{x}|\lambda)$. As $\mathbf{x} = (x_1, \ldots, x_{k_x})$ are independent Poisson counts with mean λ, the probability function is

$$Pr(\mathbf{X} = \mathbf{x}|\lambda) = \prod_{i=1}^{k_x} Pr(X_i = x_i|\lambda) = \prod_{i=1}^{k_x} e^{-\lambda} \frac{\lambda^{x_i}}{y_i!} = e^{-k\lambda} \lambda^{t_x} / \prod_{i=1}^{k_x} x_i!.$$

The probability density function for the gamma distribution is

$$f(\lambda|\alpha, \beta) = \beta^\alpha \lambda^{\alpha-1} e^{\beta\lambda} / \Gamma(\alpha); \quad \alpha > 0, \beta > 0, \lambda > 0,$$

where Γ denotes the gamma function:

$$\begin{aligned} \Gamma(w) &= (w-1)! \,\text{for integer}\, w > 0, \\ \Gamma(1/2) &= \sqrt{\pi}. \end{aligned}$$

Thus

$$\begin{aligned} \int Pr(\mathbf{X} = \mathbf{x}|\lambda) f(\lambda|\alpha, \beta) d\lambda &= \int \prod_{i=1}^{k_x} Pr(X_i = x_i|\lambda) f(\lambda|\alpha, \beta) d\lambda \\ &= \frac{1}{\Gamma(\alpha)} \frac{1}{\prod_{i=1}^{k_x} x_i!} \int e^{-k_x\lambda} \lambda^{t_x} \beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda} d\lambda \\ &= \frac{\beta^\alpha}{\prod_{i=1}^{k_x} x_i! \Gamma(\alpha)} \frac{\Gamma(\alpha + t_x)}{(\beta + k_x)^{\alpha+t_x}}. \end{aligned}$$

Analogously,

$$\int Pr(\mathbf{Y} = \mathbf{y}|\lambda) f(\lambda|\alpha, \beta) d\lambda = \frac{\beta^\alpha}{\prod_{i=1}^{k_y} y_i! \Gamma(\alpha)} \frac{\Gamma(\alpha + t_y)}{(\beta + k_y)^{\alpha+t_y}}.$$

The numerator of the likelihood ratio is

$$\int Pr(\mathbf{X} = \mathbf{x}|\lambda) \; Pr(\mathbf{Y} = \mathbf{y}|\lambda) f(\lambda|\alpha, \beta) d\lambda.$$

This is

$$\int \frac{e^{-k_x\lambda} \lambda^{t_x}}{\prod_{i=1}^{k_x} x_i!} \times \frac{e^{-k_y\lambda} \lambda^{t_y}}{\prod_{i=1}^{k_y} y_i!} \times \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} d\lambda$$

which is

$$\frac{\beta^\alpha}{\Gamma(\alpha) \prod_{i=1}^{k_x} x_i! \prod_{i=1}^{k_y} y_i!} \times \frac{\Gamma(\alpha + t_x + t_y)}{(\beta + k_x + k_y)^{\alpha+t_x+t_y}}.$$

The likelihood ratio is then

$$\frac{\Gamma(\alpha + t_x + t_y)\Gamma(\alpha)}{\Gamma(\alpha + t_x)\Gamma(\alpha + t_y)} \times \frac{(\beta + k_x)^{\alpha+t_x} (\beta + k_y)^{\alpha+t_y}}{\beta^\alpha (\beta + k_x + k_y)^{\alpha+t_x+t_y}}.$$

A.2. Derivation of the likelihood ratio for a bivariate Bernoulli model

Let $\mathbf{x} = (x_{i1}, x_{i2})$, $i = 1, 2$ be the control data, whose source is known, for periods $i = 1, 2$ where $x_{ij} (i = 1, 2, j = 1, 2) = 0$ or 1 according as whether the characteristic is absent or present. Let $\mathbf{y} = (y_{i1}, y_{i2})$ be the recovered data, whose source is unknown, for periods $i = 1, 2$ where $y_{ij} (i = 1, 2, j = 1, 2) = 0$ or 1 according as whether the characteristic is absent or present. The probability of an absence is denoted θ and the probability of a presence is then $(1 - \theta)$. Subscripts are introduced to indicate the particular circumstance of the absence (0) or presence (1) of a click in the first period (a single subscript) and in the second period, given the presence or absence in the first period (a double subscript) and θ denotes $(\theta_0, \theta_{00}, \theta_{10})$. Thus, for the first period,

$$p(x_{i1}=0)=p(y_{i1}=0)=\theta_0,\quad p(x_{i1}=1)=p(y_{i1}=1)=1-\theta_0,\quad i=1,2;$$

and for the second period

$$p(x_{i2}=0|x_{i1}=0)=p(y_{i2}=0|y_{i1}=0)=\theta_{00}\quad i=1,2;$$

$$p(x_{i2}=0|x_{i1}=1)=p(y_{i2}=0|y_{i1}=1)=\theta_{10}\quad i=1,2.$$

Denote $(\theta_0, \theta_{00}, \theta_{10})$ by $\boldsymbol{\theta}$.

Note that

$$\begin{aligned}
p((0,0)|\theta_0,\theta_{00},\theta_{10}) &= \theta_0\theta_{00},\\
p((1,0)|\theta_0,\theta_{00},\theta_{10}) &= (1-\theta_0)\theta_{10},\\
p((0,1)|\theta_0,\theta_{00},\theta_{10}) &= \theta_0(1-\theta_{00}),\\
p((1,1)|\theta_0,\theta_{00},\theta_{10}) &= (1-\theta_0)(1-\theta_{10}).
\end{aligned}$$

Assume independent beta(α, β) distributions for θ_0, θ_{00}, θ_{10} where again subscripts are introduced to indicate the particular circumstance for the prior. Thus the parameters are (α_0, β_0), $(\alpha_{00}, \beta_{00})$, $(\alpha_{10}, \beta_{10})$, respectively, denoted in general by $(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

The prior distribution for θ, assuming the independence of the three components $\theta_0, \theta_{00}, \theta_{10}$ is the product of three beta distributions so that

$$\begin{aligned}
f(\theta_0,\theta_{00},\theta_{10}) = &\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\times\frac{\Gamma(\alpha_{00}+\beta_{00})}{\Gamma(\alpha_{00})\Gamma(\beta_{00})}\times\frac{\Gamma(\alpha_{10}+\beta_{10})}{\Gamma(\alpha_{10})\Gamma(\beta_{10})}\\
&\times\theta_0^{\alpha_0-1}(1-\theta_0)^{\beta_0-1}\theta_{00}^{\alpha_{00}-1}(1-\theta_{00})^{\beta_{00}-1}\theta_{10}^{\alpha_{10}-1}(1\\
&-\theta_{10})^{\beta_{10}-1}.
\end{aligned}$$

The numerator of the likelihood ratio is derived under the assumption of a common source for the control and recovered pieces of evidence. It is derived from

$$\begin{aligned}
Num &= \int_\Theta Pr(\mathbf{x}|\theta)Pr(\mathbf{y}|\theta)\,f(\theta)d\theta.\\
&= \int\int\int_\Theta Pr\{((x_{11},x_{12}),(x_{21},x_{22})|\theta_0,\theta_{00},\theta_{10}\})\\
&\quad\times Pr\{((y_{11},y_{12}),(y_{21},y_{22})|\theta_0,\theta_{00},\theta_{10}\})\,f(\theta_0,\theta_{00},\theta_{10})\\
&\quad d\theta_0 d\theta_{00} d\theta_{10}.
\end{aligned}$$

This expression for the numerator can then be integrated to remove the dependency on θ. The result is a product of four terms given below. Whilst the expressions are rather daunting, it should be remembered that many of the terms will equal zero as the **x** and **y** terms only take the values 1 or 0. The four terms for the numerator are:

$$\begin{aligned}
\text{const} &= \frac{\Gamma(\alpha_0+\beta_0)\Gamma(\alpha_{00}+\beta_{00})\Gamma(\alpha_{10}+\beta 10)}{\Gamma(\alpha_0)\Gamma(\beta_0)\Gamma(\alpha_{00})\Gamma(\beta_{00})\Gamma(\alpha_{10})\Gamma(\beta_{10})};\\
n_0 &= \frac{\Gamma(4+\alpha_0-x_{11}-x_{21}-y_{11}-y_{21})\Gamma(\beta_0+x_{11}+x_{21}+y_{11}+y_{21})}{\Gamma(4+\alpha_0+\beta_0)};\\
n_{00} &= \Gamma(\alpha_{00}+(1-x_{11})(1-x_{12})+(1-x_{21})(1-x_{22})+(1-y_{11})(1-y_{12})\\
&\quad+(1-y_{21})(1-y_{22}))\\
&\quad\times\frac{\Gamma(\beta_{00}+(1-x_{11})x_{12}+(1-x_{21})x_{22}+(1-y_{11})y_{12}+(1-y_{21})y_{22})}{\Gamma(4+\alpha_{00}+\beta_{00}-x_{11}-x_{21}-y_{11}-y_{21})};\\
n_{10} &= \Gamma(\alpha_{10}+x_{11}(1-x_{12})+x_{21}(1-x_{22})+y_{11}(1-y_{12})+y_{21}(1-y_{22}))\\
&\quad\times\frac{\Gamma(\beta_{10}+x_{11}x_{12}+x_{21}x_{22}+y_{11}y_{12}+y_{21}y_{22})}{\Gamma(\alpha_{10}+\beta_{10}+x_{11}+x_{21}+y_{11}+y_{21})};
\end{aligned}$$

The numerator is then const $\times\, n_0 \times n_{00} \times n_{10}$. The digit 4 in the numerator (see, for example, the expressions for n_0 and n_{00}) is a consequence of the total number of items in the control and recovered group.

The term in the denominator of the likelihood ratio for the control pieces of evidence is derived as follows

$$\begin{aligned}
Den_C &= \int_\Theta Pr(\mathbf{x}|\theta)\,f(\theta)d\theta\\
&= \int\int\int_\Theta Pr\{((x_{11},x_{12}),(x_{21},x_{22})|\theta_0,\theta_{00},\theta_{10}\})\\
&\quad\times f(\theta_0,\theta_{00},\theta_{10})d\theta_0 d\theta_{00} d\theta_{10}.
\end{aligned}$$

This expression for the control term in the denominator can then be integrated to remove the dependency on θ. The result is a product of four terms, one of which is the const term given above. The other three terms for the control term in the denominator are:

$$\begin{aligned}
d_{0c} &= \frac{\Gamma(2+\alpha_0-x_{11}-x_{21})\Gamma(\beta_0+x_{11}+x_{21})}{\Gamma(2+\alpha_0+\beta_0)};\\
d_{00c} &= \Gamma(\alpha_{00}+(1-x_{11})(1-x_{12})+(1-x_{21})(1-x_{22}))\\
&\quad\times\frac{\Gamma(\beta_{00}+(1-x_{11})x_{12}+(1-x_{21})x_{22})}{\Gamma(2+\alpha_{00}+\beta_{00}-x_{11}-x_{21})};\\
d_{10c} &= \Gamma(\alpha_{10}+x_{11}(1-x_{12})+x_{21}(1-x_{22}))\\
&\quad\times\frac{\Gamma(\beta_{10}+x_{11}x_{12}+x_{21}x_{22})}{\Gamma(\alpha_{10}+\beta_{10}+x_{11}+x_{21})};
\end{aligned}$$

Den_C is then const $\times\, d_{0c} \times d_{00c} \times d_{10c}$.

The term in the denominator of the likelihood ratio for the recovered pieces of evidence is then derived from the control term with the replacement of terms in x by the corresponding terms in y. This gives

$$\begin{aligned}
Den_R &= \int_\Theta Pr(\mathbf{y}|\theta)\,f(\theta)d\theta\\
&= \int\int\int_\Theta Pr\{((y_{11},y_{12}),(y_{21},y_{22})|\theta_0,\theta_{00},\theta_{10}\})\\
&\quad\times f(\theta_0,\theta_{00},\theta_{10})d\theta_0 d\theta_{00} d\theta_{10}.
\end{aligned}$$

This expression for the recovered term in the denominator can then be integrated to remove the dependency on θ. The result is a product of four terms, one of which is the const term given above. The other three terms for the recovered term in the denominator are:

$$\begin{aligned}
d_{0r} &= \frac{\Gamma(2+\alpha_0-y_{11}-y_{21})\Gamma(\beta_0+y_{11}+y_{21})}{\Gamma(2+\alpha_0+\beta_0)};\\
d_{00r} &= \Gamma(\alpha_{00}+(1-y_{11})(1-y_{12})+(1-y_{21})(1-y_{22}))\\
&\quad\times\frac{\Gamma(\beta_{00}+(1-y_{11})y_{12}+(1-y_{21})y_{22})}{\Gamma(2+\alpha_{00}+\beta_{00}-y_{11}-y_{21})};\\
d_{10r} &= \Gamma(\alpha_{10}+y_{11}(1-y_{12})+y_{21}(1-y_{22}))\\
&\quad\times\frac{\Gamma(\beta_{10}+y_{11}y_{12}+y_{21}y_{22})}{\Gamma(\alpha_{10}+\beta_{10}+y_{11}+y_{21})};
\end{aligned}$$

Den_R is then const $\times\, d_{0r} \times d_{00r} \times d_{10r}$.

As with the numerator, the digit 2 is a consequence of the two items in the control group for Den_C and the two items in the recovered group for Den_R.

The likelihood ratio, V, is, as in (3), then

$$V = \frac{n_0 \times n_{00} \times n_{10}}{\text{const} \times d_{0c} \times d_{00c} \times d_{10c} \times d_{0r} \times d_{00r} \times d_{10r}}.$$

It is a product and quotient of gamma functions with arguments $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{x}, \mathbf{y})$ and is symmetric in **x** and **y**.

References

[1] C.G.G. Aitken, F. Taroni, Statistics and the Evaluation of Evidence for Forensic Scientists, 2nd ed., John Wiley and Sons, Ltd., Chichester, 2004.
[2] C.G.G. Aitken, D. Lucy, G. Zadora, J.M. Curran, Evaluation of trace evidence for three-level multivariate data with the use of graphical models, Comput. Stat. Data Anal. 50 (2006) 2571–2588. http://dx.doi.org/10.1016/j.csda.2005.04.005.
[3] K.A. Ebejer, J. Winn, J.F. Carter, R. Sleeman, J. Parker, F. Körber, The difference between drug money and a "lifetime's savings", Forensic Sci. Int. 167 (2–3) (2007) 94–101.
[4] P. Ladefoged, A Course in Phonetics, 5th ed., Wadsworth Cengage Learning, Boston, 2006
[5] J. Laver, Principles of Phonetics, Cambridge University Press, Cambridge, 1994.
[6] F. Nolan, K. McDougall, G. de Jong, T. Hudson, The DyViS database: style controlled recordings of 100 homogeneous speakers for forensic phonetic research., Int. J. Speech Lang. Law 16 (1) (2009) 31–57.
[7] C.-S. Li, J.-C. Lu, J. Park, K. Kim, P.A. Brinkley, J.P. Peterson, Multivariate zero-inflated Poisson models and their applications, Technometrics 41 (1999) 29–38.

Please cite this article in press as: C. Aitken, E. Gold, Evidence evaluation for discrete data, Forensic Sci. Int. (2013), http://dx.doi.org/10.1016/j.forsciint.2013.02.042