Evidence evaluation for discrete data

Colin Aitken[1,3] and Erica Gold[2],
[1]The University of Edinburgh, Edinburgh, UK
[2] The University of York, York, UK
[3] Corresponding author: Colin Aitken, e-mail: c.g.g.aitken@ed.ac.uk

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. The outline of an empirical method based on relative frequencies that takes account of similarity and rarity is described. It is compared with two methods based on an assumption of independence of counts and one assuming dependence between adjacent Bernoulli variables. 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.

Keywords: bivariate Bernoulii, discrete data, likelihood ratio, forensic phonetics

Introduction

The interpretation of scientific evidence may be thought of as the assessment of a comparison. The 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:

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

The evidence E may be written as (X, Y) where X is the control data, evidence whose source is known and Y is the recovered data, evidence whose source is unknown. The statistic used to evaluate the evidence is the likelihood ratio

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

Other than for DNA profiling, there is a paucity of methods when the data X and Y are discrete. 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 [||]' (Ladefoged,

[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.

2006). The first two methods, a method assuming independence between counts and using a Poisson distribution and a method based on a bivariate Bernoulli model are described in Aitken and Gold (2013).

Data of independent counts with a Poisson distribution

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 variability in λ across a population; the mean number λ of counts for an item in the population varies from item to item. If there were N items then item i may be said to have mean $\lambda_i, i = 1, \ldots, N$. In the simple situation described here, the variation in λ across the population is taken to have a gamma distribution, characterised by two parameters, α and β. Subjective choices are made for α and β.

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, the likelihood ratio for the value, V, of the evidence $\mathbf{x}$ and $\mathbf{y}$ is

$$V = \frac{\Gamma(\alpha + t_x + t_y)\Gamma(\alpha)}{\Gamma(\alpha + t_x)\Gamma(\alpha + t_y)} \times \frac{(\beta + k_x)^{\alpha + t_x}(\beta + k_y)^{\alpha + t_y}}{\beta^{\alpha}(\beta + k_x + k_y)^{\alpha + t_x + t_y}}. \quad (1)$$

(Aitken and Gold, 2013)

Bivariate Bernoulli model

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.

Two pairs of observations per source are considered, for example, of two minutes of speech in each period, for control and recovered speech. The control evidence has two pairs of independent observations with a bivariate binary model with parameter θ_c (c for control) on a particular item. The recovered evidence 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. In the simple situation described here, the variations in the three components of θ across the population are taken to have beta distributions.

Let $\mathbf{x} = (x_{i1}, x_{i2}), i = 1, 2$ be the control data, whose source is known, for periods $i = 1, 2$ where $x_{ij} = 0$ or 1; $(i = 1, 2, j = 1, 2)$ 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 $y_{ij} = 0$ or 1; $(i = 1, 2, j = 1, 2)$ 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

$$\begin{aligned} p(x_{i1} = 0) = p(y_{i1} = 0) = \theta_0, \qquad & p(x_{i1} = 1) = p(y_{i1} = 1) = 1 - \theta_0, \ i = 1, 2; \\ p(x_{i2} = 0 \mid x_{i1} = 0) \ &= \ p(y_{i2} = 0 \mid y_{i1} = 0) = \theta_{00} \ \ i = 1, 2; \\ p(x_{i2} = 0 \mid x_{i1} = 1) \ &= \ p(y_{i2} = 0 \mid y_{i1} = 1) = \theta_{10} \ \ i = 1, 2. \end{aligned}$$

Assume independent beta(α, β) distributions for $\theta_0, \theta_{00}, \theta_{10}$ where again subscripts are introduced to indicate the particular circumstance for the prior. Thus the parameters are $(\alpha_0, \beta_0), (\alpha_{00}, \beta_{00}), (\alpha_{10}, \beta_{10})$, respectively, which may be estimated by appropriate method of moments estimators from sample proportions and variances from some relevant population. Alternatively, 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

$$V \ = \ \frac{n_0 \times n_{00} \times n_{10}}{const \times d_{0c} \times d_{00c} \times d_{10c} \times d_{0r} \times d_{00r} \times d_{10r}} \tag{2}$$

where, apart from $const$, the terms are functions of gamma functions with full expressions given in Aitken and Gold (2013).

Empirical model

Consider a piece of speech from an unknown person (e.g., audio recording associated with a crime) (recovered speech). The number of minutes of speech are k and the number of clicks per minute are $\mathbf{y} = \{y_1, \ldots, y_k\}$. Consider a piece of speech from a known person (e.g., suspect) (control speech). The number of minutes of speech is chosen to be equal to that of the recovered speech. The number of clicks per minute are $\mathbf{x} = \{x_1, \ldots, x_k\}$. Let $p(\mathbf{x}) = p(x_1, \ldots, x_k)$ and $p(\mathbf{y}) = p(y_1, \ldots, y_k)$ be the probabilities of $\mathbf{x}$ and $\mathbf{y}$, respectively under some statistical model. Part of the problem is to determine the appropriate model. The following statistic is proposed for the likelihood ratio (LR):

$$\frac{\exp\{-\sum_{i=1}^{k}(x_i - y_i)^2\}}{p(x_1, \ldots, x_k) \times p(y_1, \ldots, y_k)}. \tag{3}$$

The numerator measures similarity. The more similar the control and recovered speech are in terms of numbers of clicks in each minute, the larger the value of the numerator and hence the larger the LR. The denominator measures rarity. The more rare the control and recovered speech are in terms of numbers of clicks in each minute, the smaller the value of the denominator and hence the larger the LR. Of course, multiplication of (3) by a constant

results in a statistic with the same properties as this one. Also, the probabilistic behaviour of the statistic has to be investigated. The absolute value of the statistic is not meaningful. However, relative values are meaningful. It is possible then to consider relative support of one pair of speech comparisons with another.

Results

Results are given in Table 1 (an extract from Aitken and Gold, 2013) of an application of (1) for various combinations of α and β. Very small values of the evidence, much less than one, occur when a control piece of speech with no clicks in six minutes is compared with a recovered piece of speech with twelve clicks in six minutes. For example, $V = 0.006 \simeq 1/170$ when $t_x = 0,\ t_y = 12;\ k_x = k_y = 6;\ \alpha = 2,\ \beta = 2$. This result provides support for the proposition of different sources for the speech: the evidence is 140 (170) times more likely if the two pieces of speech ($\mathbf{x}$ and $\mathbf{y}$) were uttered by different people than by the same person.

		Value of the evidence V (1)			
$t_x = \sum_{i=1}^{k_x} x_i$	$t_y = \sum_{i=1}^{k_y} y_i$	$\alpha = 3$ $\beta = 1$ $E(X) = 3$ $Var(X) = 3$	$\alpha = 2$ $\beta = 2$ $E(X) = 1$ $Var(X) = 0.5$	$\alpha = 4$ $\beta = 1$ $E(X) = 4$ $Var(X) = 4$	$\alpha = 9$ $\beta = 3$ $E(X) = 3$ $Var(X) = 1$
0	0	53.5	5.22	201.84	198.36
4	4	5.3	1.50	13.45	12.25
8	8	2.6	1.82	4.62	3.20
0	4	4.5	0.56	16.97	25.71
0	8	0.4	0.06	1.43	3.33
0	12	0.03	0.006	0.12	0.43

Table 1: Values of evidence (1) 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. Further results are available in Aitken and Gold (2013).

Results are given in Table 2 (an extract from Aitken and Gold, 2013) of applications of (2) for various combinations of bivariate Bernoulli models and prior parameters. A distribution such as beta(2,1) or 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 occur when there is a mismatch between outcomes as illustrated in 6. Rows 4 and 5 show two values greater than one and one value less than one, illustrating the importance of prior values in situations with few data.

The empirical model (3) requires input from a data set. A data set provided by the BBFOR2 project records the numbers of clicks per minute for 100 speakers over periods of 4 to 6 minutes. The relative frequencies for each of the possible number of clicks from 0 to 17 are given in Table 3, where 1 was added to all observed frequencies to allow for zero entries (for numbers of clicks per minute less than the maximum observed) in the original data set. Two sample results are given in Table 4.

Discussion

The two models in Aitken and Gold (2013) are basic models, the exact situations for which will rarely occur in practice. The empirical model requires considerable further study to

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
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

Table 2: Values of the likelihood ratio (2) 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. Further results are available in Aitken and Gold (2013).

Clicks per minute	0	1	2	3	4	5
Relative frequency	0.563	0.233	0.086	0.037	0.035	0.016
Clicks per minute	6	7	9	11	15	17
Relative frequency	0.005	0.007	0.005	0.005	0.005	0.005

Table 3: Overall relative frequencies for the numbers of clicks per minute. If a previously unobserved number of clicks per second is observed in a particular case, record the frequency as 1/431 and adjust the other frequencies appropriately.

investigate its probabilistic properties. However, all models 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 small size of the 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 small size of the datasets means that the choice of the prior parameters makes a big difference to the values of the likelihood ratio.

The model based on independent Poisson counts is easier to implement than the bivariate Bernoulli model but has an unrealistic assumption of independence. An 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, care will be needed to avoid a decrease in the robustness of the model. Various issues need to be considered in extensions of this work.

- *Data collection:* More practical work is needed to collect data sets appropriate for analyses by these models, or extensions of them, and for interpretation of the results.
- *Autocorrelation:* The Poisson model assumes the data are independent. The bivariate Bernoulli model allows for correlation at a simple level of adjacent items with binary responses. A multivariate Dirichlet model provides an obvious extension to a bivariate Bernoulli variable when there is a multinomial response. However, the

x	**y**	$\sum(x_k - y_k)^2$	p_x	p_y	LR (3)
0000	0000	0	0.563^4	0.563^4	99.07
0000	1000	1	0.563^4	$0.563^3 \times 0.233$	88

Table 4: Comparison of two pairs of speech patterns, all over periods of $l = 4$ minutes. The first is where there are no clicks in either the control (x) or recovered speech (y), the second where there is a click in the first minute of the recovered speech. Relative frequencies p_x and p_y are given in Table 3.

example of forensic phonetics is concerned with counts within fixed time periods of one minute in length. Thus a Poisson distribution with correlated responses would be more appropriate.

An alternative approach would be to record the times at which the clicks were made and use a Poisson point process. This could also be extended to a point process with autocorrelation.

- *Variability between and within items:* Current results are based on prior assumptions about speech variation within and between speakers as the available data are for 100 speakers, each making one utterance. There is no measure of within-speaker variation. Ideally, each speaker should repeat the same piece of speech several times. These would be realisations of a multivariate discrete random variable as a model for the within-speaker variation and whose distribution could be estimated from these data. Summary statistics for each speaker could be derived from which between-speaker variation could be determined.

- *Nonparametric distributions:* It may be that a multivariate Poisson distribution which allows for correlated responses may not be appropriate. A nonparametric approach may overcome this problem. A discrete kernel probability mass function would allow for distributions which did not fit more standard distributions such as the Poisson or negative binomial when the variance was larger than the mean.

- *Temperament:* Some standardisation of performance will be required to allow for different levels of stress in the speaker. Stressful situations include committing a crime and being interviewed as a suspect for a crime. In both situations, the characteristics of speech will differ from when the speaker is relaxed.

- *Relevant population:* Characteristics of speech are very dependent on the population from which the speaker comes. Care will be needed in the evaluation of evidence based on speech that the relevant population is determined by what is known about the criminal rather than what is known about a suspect.

References

Aitken,C.G.G. and Gold,E. (2013) Evidence evaluation for discrete data. *Forensic Science International*, in press.

Ladefoged,P. (2006) *A Course in Phonetics*, 5th edition, Wadsworth Cengage Learning, Boston, 2006.

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