SCHOOL *of*
GRADUATE STUDIES
EAST TENNESSEE STATE UNIVERSITY

**East Tennessee State University**
**Digital Commons @ East**
**Tennessee State University**

Electronic Theses and Dissertations

Student Works

8-2006

# Amended Estimators of Several Ratios for Categorical Data.

Dandan Chen
*East Tennessee State University*

Follow this and additional works at: https://dc.etsu.edu/etd

Part of the Categorical Data Analysis Commons

### Recommended Citation

Amended Estimators of Several Ratios for Categorical Data

————————

A thesis

presented to

the faculty of the Department of Mathematics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

————————

by

Dandan Chen

August, 2006

————————

Robert Price, Ph.D., Chair

Edith Seier, Ph.D.

Yali Liu, Ph.D.

Keywords: Contingency Table, Odds Ratio, Relative Risk, Poisson Ratio, Marginal Ratio, Bayesian Analysis

ABSTRACT

Amended Estimators of Several Ratios for Categorical Data

by

Dandan Chen

Point estimation of several association parameters in categorical data are presented. Typically, a constant is added to the frequency counts before the association measure is computed. We will study the accuracy of these adjusted point estimators based on frequentist and Bayesian methods respectively. In particular, amended estimators for the ratio of independent Poisson rates, relative risk, odds ratio, and the ratio of marginal binomial proportions will be examined in terms of bias and mean squared error.

3

# DEDICATION

I dedicate this thesis to my wife, June, and my son, Hao.

# ACKNOWLEDGMENTS

I would like to thank all of my professors for getting me to this point. I would especially like to thank my advisor Dr. Robert Price and other members of this committee, Dr. Edith Seier and Dr. Yali Liu for all of their advice, patience, and support during the preparation of this thesis.

CONTENTS

7

LIST OF TABLES

# 1  INTRODUCTION

Ratios of parameters are some interesting measures that are widely estimated in the study of categorical data. In this thesis, we will consider the estimation of the ratio of independent Poisson rates, the relative risk under product binomial sampling, the odds ratio, and the ratio of marginal binomial proportions under multinomial sampling. Estimates of these ratios can be undefined if sampling zeros occur. To avoid such situations, it has been suggested to add a small constant, say .5, to each frequency count before computing the estimate of the desired ratio. In addition, adding a small constant to the frequency counts may also reduce the bias and mean squared error of the estimator. In this thesis, we will try to determine the correction constant that minimizes bias and mean squared error for these ratios based on frequentist and Bayesian methods.

## 1.1  Contingency Tables

A *categorical variable* has a measurement scale consisting of two or more categories, and there is no intrinsic ordering to the categories. The joint distribution between two categorical variables determines their relationship. This distribution also determines the marginal and conditional distribution.

Let $X$ and $Y$ denote two categorical response variables, $X$ with $I$ categories and $Y$ with $J$ categories. Classifications of subjects on both variables have $IJ$ possible combinations. The responses $(X, Y)$ of a subject chosen randomly from some population have a probability distribution. A rectangular table having $I$ rows for categories of $X$ and $Y$ columns for categories of $Y$ displays this distribution. The *cell* of the

table represents the $IJ$ possible outcomes. When the cells contain frequency counts of outcomes for a sample, the table is called a *contingency table*, or *cross-classification table*.

We will consider the contingency table with 2 rows and 2 columns as shown in Table 1, which is called $2 \times 2$ (or 2-by-2) table.

Table 1: Observations in a $2 \times 2$ contingency table

|  | Column 1 | Column 2 | Total |
|---|---|---|---|
| Row 1 | $y_{11}$ | $y_{12}$ | $y_{1+}$ |
| Row 2 | $y_{21}$ | $y_{22}$ | $y_{2+}$ |
| Total | $y_{+1}$ | $y_{+2}$ | n |

For a $2 \times 2$ table, Table 2 displays the notation for the joint and marginal distributions. Let $\pi_{ij}$ denote the probability that $(X, Y)$ occurs in the cell in row $i$ and column $j$. The probability distribution $\{\pi_{ij}\}$ is the *joint distribution* of $X$ and $Y$. The *marginal distributions* are the row and column totals that result from summing the joint probabilities. We denote these by $\{\pi_{i+}\}$ for the row variable and $\{\pi_{+j}\}$ for the column variable, where the subscript "+" denotes the sum over that index; that is,

$$\pi_{i+} = \sum_j \pi_{ij} \text{ and } \pi_{+j} = \sum_i \pi_{ij}.$$

These marginal distributions satisfy $\sum_i \pi_{i+} = \sum_j \pi_{+j} = \sum_i \sum_j \pi_{ij} = 1.0$, and provide single-variable information.

In most contingency tables, one variable, say $Y$, is a response variable and the other $(X)$ is an explanatory variable. When $X$ is fixed rather than random, the notion of a joint distribution for $X$ and $Y$ is no longer meaningful. However, for a

Table 2: Joint and marginal probabilities

|  | Column 1 | Column 2 | Total |
|---|---|---|---|
| Row 1 | $\pi_{11}$ | $\pi_{12}$ | $\pi_{1+}$ |
| Row 2 | $\pi_{21}$ | $\pi_{22}$ | $\pi_{2+}$ |
| Total | $\pi_{+1}$ | $\pi_{+2}$ | 1 |

fixed category of $X$, $Y$ has a probability distribution. Given that a subject is classified in row $i$ of $X$, $\pi_{j|i}$ denotes the probability of classification in column $j$ of $Y$. Note that $\sum_j \pi_{j|i} = 1$. The probabilities $\{\pi_{1|i}, ..., \pi_{J|i}\}$ form the *conditional distribution* of $Y$ at category $i$ of $X$. Table 3 displays the notation for the conditional distributions in the $2 \times 2$ table.

Table 3: Conditional probabilities

|  | Column 1 | Column 2 | Total |
|---|---|---|---|
| Row 1 | $\pi_{1|1}$ | $\pi_{1|2}$ | 1 |
| Row 2 | $\pi_{2|1}$ | $\pi_{2|2}$ | 1 |

When both variables are response variables, descriptions of the association can use their joint distribution or conditional distribution, and

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} \qquad \text{for all } i \text{ and } j.$$

Two categorical response variables are defined to be *independent* if all joint probabilities equal the product of their marginal probabilities,

$$\pi_{ij} = \pi_{i+}\pi_{+j} \qquad \text{for } i = 1, 2 \text{ and } j = 1, 2. \tag{1}$$

When $Y$ is a response and $X$ is an explanatory variable, it is more natural to define

independence of $X$ and $Y$, if

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} = (\pi_{i+}\pi_{+j})/\pi_{i+} = \pi_{+i} \qquad \text{for } i = 1,\, 2.$$

## 1.2   Distributions for Categorical Data

Inferential data analyses require assumptions about the random mechanism that generated the data. Three common probability distributions for categorical responses are the binomial, multinomial, and Poisson. We describe each of these probability distributions next.

### 1.2.1   Binomial Distribution

Many experiments have a fixed number $n$ of binary observations. Let $y_1, y_2, ..., y_n$ denote responses for $n$ independent and identical trials such that $P(Y_i = 1) = \pi$ and $P(Y_i = 0) = 1 - \pi$. We use the generic labels "success" and "failure" for outcomes 1 and 0, respectively. *Identical trials* means that the probability of success $\pi$ is the same for each trial. *Independent trials* means that the $\{Y_i\}$ are independent random variables. These are often called *Bernoulli trials*. The total number of successes, $Y = \sum_{i=1}^{n} Y_i$, has the *binomial distribution* with index $n$ and parameter $\pi$, denoted by $\mathrm{Bin}(n, \pi)$.

The probability mass function for the possible outcomes $y$ for $Y$ is

$$p(y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}, \qquad y = 0, 1, 2, \ldots, n. \tag{2}$$

The mean and variance of $Y$ are

$$\mu = E(Y) = n\pi \text{ and } \sigma^2 = var(Y) = n\pi(1 - \pi),$$

respectively.

### 1.2.2 Multinomial Distribution

Some trials have more than two possible outcomes. Suppose that each of $n$ independent, identical trials can have outcome in any of $m$ categories. Let $y_{ij} = 1$ if trial $i$ has outcome in category $j$ and $y_{ij} = 0$ otherwise. Then $y_i = (y_{i1}, y_{i2}, ..., y_{im})$ represents a multinomial trial, with $\sum_j y_{ij} = 1$; for instance, $(0,0,1,0)$ denotes outcome in category 3 of four possible categories. Note that $y_{im}$ is redundant, being linearly dependent on the others. Let $n_j = \sum_j y_{ij}$ denote the number of trials having outcome in category $j$. The counts $(n_1, n_2, ..., n_m)$ have the *multinomial distribution.*

Let $\pi_j = P(Y_{ij} = 1)$ denote the probability of outcome in category $j$ for each trial. The multinomial probability mass function is

$$ p(n_1, n_2, ..., n_m) = \frac{n!}{n_1! n_2! \cdots n_m!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_m^{n_m}. \tag{3} $$

For the multinomial distribution the mean, variance and covariance are

$$ E(n_j) = n\pi_j, \, var(n_j) = n\pi_j(1 - \pi_j) \text{ and } cov(n_i, n_j) = -n\pi_i\pi_j, $$

respectively. We note that the marginal distribution of each $n_j$ has a binomial distribution.

### 1.2.3 Poisson Distribution

Sometimes count data do not result from a fixed number of trials. For instance, if $y$ is equal to the number of phone calls arriving at a switchboard in the next minute, there is no fixed upper limit $n$ for $y$. Since $y$ must be a nonnegative integer, its distribution should place its mass on that range.

The simplest such distribution is the *Poisson*. The Poisson distribution is used for counts of events that occur randomly over time or space when outcomes in disjoint periods or regions are independent. Its probabilities depend on a single parameter, the mean $\lambda$. The Poisson probability mass function is given by

$$p(y) = \frac{e^{-\lambda}\lambda^y}{y!}, \qquad y = 0, 1, 2, \ldots \tag{4}$$

The Poisson distribution has the property that the $E(Y) = var(Y) = \lambda$.

## 1.3   Some Ratios in $2 \times 2$ Contingency Table

We will study the estimation of the ratio of independent Poisson rates, relative risk, odds ratio, and the ratio of marginal binomial proportions in the $2 \times 2$ contingency table. For small n, the distributions of these ratios are highly skewed. When the ratio equals 1, for instance, the estimator cannot be much smaller than the real ratio (since it should be greater than or equal to 0), but it could be much larger with nonnegligible probability. The log transformation of these ratios, having an additive rather than multiplicative structure, converges more rapidly to a normal distribution. So we will focus on these log ratios in this thesis.

### 1.3.1   Relative Risk

Consider a study to compare two groups on a binary response variable. Let the random variable $Y$ have two categories, such as (success, failure) for the outcome of a treatment. With only two possible outcomes, $\pi_{2|i} = 1 - \pi_{1|i}$, we will use the simpler notation $\pi_i$ for $\pi_{1|i}$. The *difference of proportions* of successes, $\pi_1 - \pi_2$, is a basic comparison of the two rows. A value $\pi_1 - \pi_2$ of fixed size may have greater importance when both $\pi_i$ are closed to 0 or 1 than when they are not. For a medical study comparing two treatments on the proportion of subjects who die, the difference between 0.001 and 0.010 may be more noteworthy than the difference between 0.401 and 0.410, even though both are 0.009. In such cases, the ratio of proportions is informative.

The *relative risk* is defined to be the ratio

17

$$\theta_r = \pi_1/\pi_2. \tag{5}$$

It can be any nonnegative real number. A relative risk of 1.0 corresponds to independence. For the proportions just given, the relative risks are $0.010/0.001 = 10.0$ and $0.410/0.401 = 1.02$.

### 1.3.2 Odds Ratio

For a probability $\pi$ of success, the *odds* are defined to be

$$\Omega = \pi/(1 - \pi).$$

The odds are nonnegative, with $\Omega > 1.0$ when a success is more likely than a failure.

For the $2 \times 2$ table designed to compare two groups on a binary response variable, within row $i$, the odds of success instead of failure are $\Omega_i = \pi_i/(1 - \pi_i)$ . The *odds ratio* is defined as the ratio of the odds $\Omega_1$ and $\Omega_2$ in the two rows,

$$\theta_o = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}. \tag{6}$$

For joint distributions, the equivalent definition for the odds in row $i$ is $\Omega_i = \pi_{i1}/\pi_{i2}$, $i = 1, 2$ and hence the odds ratio is defined as

$$\theta_o = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}. \tag{7}$$

Note that the odds ratio is also called the *cross product ratio*. We will discuss an adjusted estimator for the odds ratio for joint distributions in this thesis.

### 1.3.3  Ratio of Marginal Binomial Proportions

Refer to Table 2, $\pi_{1+} = \pi_{11} + \pi_{12}$ and $\pi_{+1} = \pi_{11} + \pi_{21}$ denote the marginal probabilities of interest. In some applications, a ratio of marginal proportions $\pi_{1+}/\pi_{+1}$ may be more interesting than a difference. For instance, if the difference between $\pi_{1+} = 0.01$ and $\pi_{+1} = 0.10$ is considered to be more important than the difference between $\pi_{1+} = 0.51$ and $\pi_{+1} = 0.60$, then a ratio of proportions may be preferred to a difference of proportions. The ratio of marginal binomial proportions is defined as

$$\theta_m = \frac{\pi_{1+}}{\pi_{+1}} = \frac{\pi_{11} + \pi_{12}}{\pi_{11} + \pi_{21}}. \tag{8}$$

### 1.3.4  Ratio of Poisson Rates

If cell counts $\{Y_{ij}\}$ in a contingency table are treated as independent Poisson random variables, it may be of interest to estimate the ratio of two Poisson means. Let $Y_1, Y_2$ be two independent Poisson variables, we have $E(Y_j) = VAR(Y_j) = \lambda_j$ for $j = 1, 2$, where $\lambda_j$ are the means of two Poisson distributions. The Poisson ratio is defined to be

$$\theta_p = \frac{\lambda_1}{\lambda_2}. \tag{9}$$

## 1.4   Point Estimation

We now consider the estimation of the ratios for a given probability mass function (p.m.f.) In certain instances, the experimenter needs a point estimate of the parameter $\theta$, namely the value of the parameter that corresponds to the selected p.m.f. Assume that $\theta$ is a real valued parameter and that $\hat{\theta}$ is an estimator of $\theta$. The probability distribution of an estimator $\hat{\theta}$ is often referred to as the sampling distribution of $\hat{\theta}$. Ideally, we would like the sampling distribution of $\hat{\theta}$ to be concentrated closely around the true value of the parameter $\theta$.

There are several theoretical approaches for finding frequentist estimators, such as maximum likelihood estimation (MLE) and uniformly minimum variance unbiased estimation (UMVUE). We will use the MLE to find the frequentist estimator.

Given outcomes $n_i$, $i = 1, ...N$, the *likelihood function* under that sampling model, is the probability of $n_i$, treated as a function of the unknown parameters. The maximum likelihood (ML) estimates are the parameter values that maximize this function. Under these parameter values, the observed data would have had highest probability of occurrence.

From a Bayesian perspective, point estimation means that we would use a single statistic to summarize the posterior distribution. The most important number summarizing a distribution would be the location. The posterior mean, or the posterior median would be good candidates. We will use the posterior median as the Bayesian estimate since it is more robust.

We will use frequentist criterias for evaluating our estimators. There are several simple measures of the quality of an estimator based on its sampling distribution. The

properties of the estimators that we will be considering are bias and mean squared error.

The *bias* of an estimator $\hat{\theta}$ is defined to be

$$b_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta. \tag{10}$$

An estimator is said to be unbiased if $b_\theta(\hat{\theta}) = 0$, that is $E_\theta(\hat{\theta}) = \theta$. The *mean squared error* (MSE) of $\hat{\theta}$ is defined to be

$$MSE_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2] \tag{11}$$

It is easy to show that $MSE_\theta(\hat{\theta}) = Var_\theta(\hat{\theta}) + [b_\theta(\hat{\theta})]^2$.

The bias of $\hat{\theta}$ gives some indication of whether the sampling distribution is centered around $\theta$ while $MSE_\theta(\hat{\theta})$ is a measure of the dispersion of the sampling distribution of $\hat{\theta}$ around $\theta$, it is natural to prefer estimators with small MSE.

## 2    FREQUENTIST METHOD

Poisson ratio, relative risk, odds ratio, and ratio of marginal binomial proportions have been studied for a long period of time. Mainly, people focused on interval estimation of these ratios using amended estimators but not much has been said about the properties of these point estimators [2], [5].

### 2.1    Estimating the Odds Ratio

For a multinomial distribution, the MLE of $\pi_{ij}$ is $y_{ij}/n$. Gart and Zweiful [4] showed that the amended estimators

$$\tilde{\theta}_o = \frac{(n_{11} + \frac{1}{2})(n_{22} + \frac{1}{2})}{(n_{21} + \frac{1}{2})(n_{12} + \frac{1}{2})} \tag{12}$$

and $\log \tilde{\theta}_o$ behave well in terms of bias and MSE.

For a multidimensional grid of points of $\pi_{11}$, $\pi_{12}$ and $\pi_{21}$, we calculated the mean bias and square root of MSE for estimators at all points in the grid. Figure 1 shows the result when $n = 20$. For a small sample size, we can get the same result as Gart [4]. But when sample size increases, the correction constant can be greater than 0.5.

### 2.2    Estimating the Relative Risk

Relative risk compares the parameters of two independent binomial distributions in a $2 \times 2$ table, and the joint probability function for the entire data set is the product binomial probability function. The MLE for $\pi$ of the binomial distribution is $y/n$ where $y$ is the count and $n$ is the number of total trails. To avoid certain
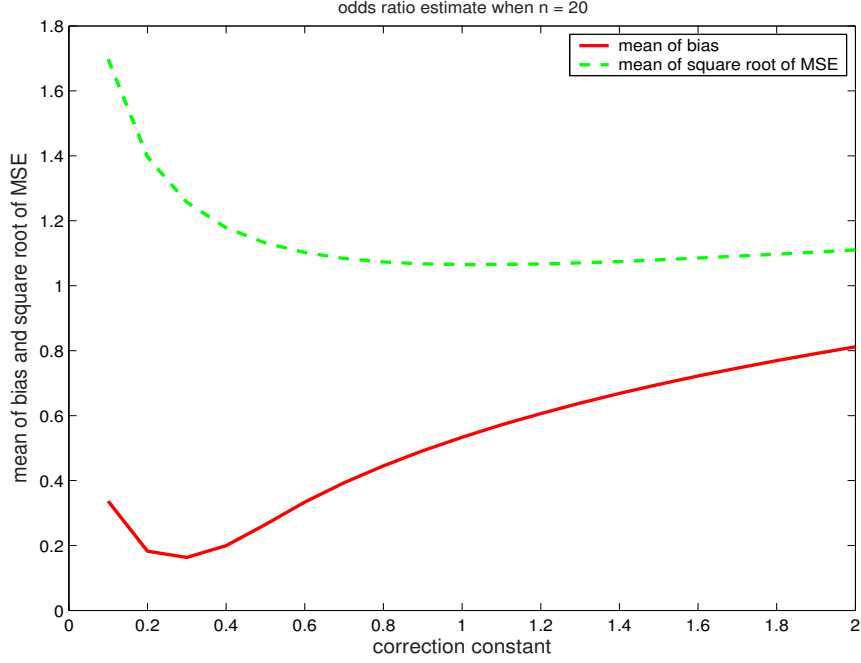
Figure 1: Log odds ratio estimate n=20

undefined situations, Walter [9] suggested a less biased estimator of the log relative risk $[log(\pi_{1|1}) - log(\pi_{1|2})]$,

$$log \, \tilde{\theta}_r = log\left(\frac{y_{11} + \frac{1}{2}}{y_{1+} + \frac{1}{2}}\right) - log\left(\frac{y_{21} + \frac{1}{2}}{y_{2+} + \frac{1}{2}}\right). \tag{13}$$

For a bidimensional grid of points of $\pi_1$ and $\pi_2$, we compared the mean of bias and square root of MSE for estimators at every point in the grid. When $y_{1+} = 20, y_{2+} = 32$, as shown in figure 2, we would choose 0.5 as our correction constant. If the sample size is smaller, this value could be as small as 0.3.

Agresti [1] suggested adding the same correction constant to $n_{i+}$ as that to $n_{i1}$, after comparing the effects under different sample size based on the mean bias and squared root of MSE, we would say that adding the same correction constant to $n_{i+}$
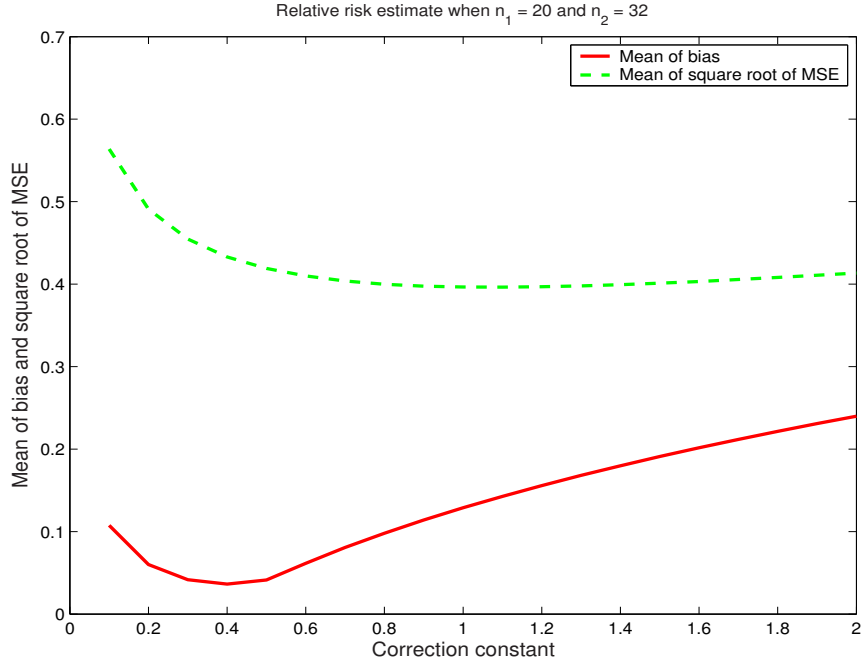
Figure 2: Log relative risk estimate $n_1 = 20, n_2 = 32$

as that to $n_{i1}$ is slightly better than adding twice the value to $n_{i+}$ as that to $n_{i1}$.

### 2.3   Estimating the Ratio of Marginal Binomial Proportions

The joint probability function for the entire data set is the multinomial distribution. The MLE of the marginal ratio is the sample marginal ratio. If counts in the first column of $2 \times 2$ table equal to 0, the sample marginal ratio is undefined. We have to add a constant $c$ to each cell

$$log\, \tilde{\theta}_m = log[(n_{11} + c) + (n_{12} + c)] - log[(n_{11} + c) + (n_{21} + c)] \qquad (14)$$

For a multidimensional grid of points of $\pi_{11}$, $\pi_{12}$ and $\pi_{21}$, we calculated the mean bias and square root of MSE for the estimator at all points in the grid. Figure 3

24
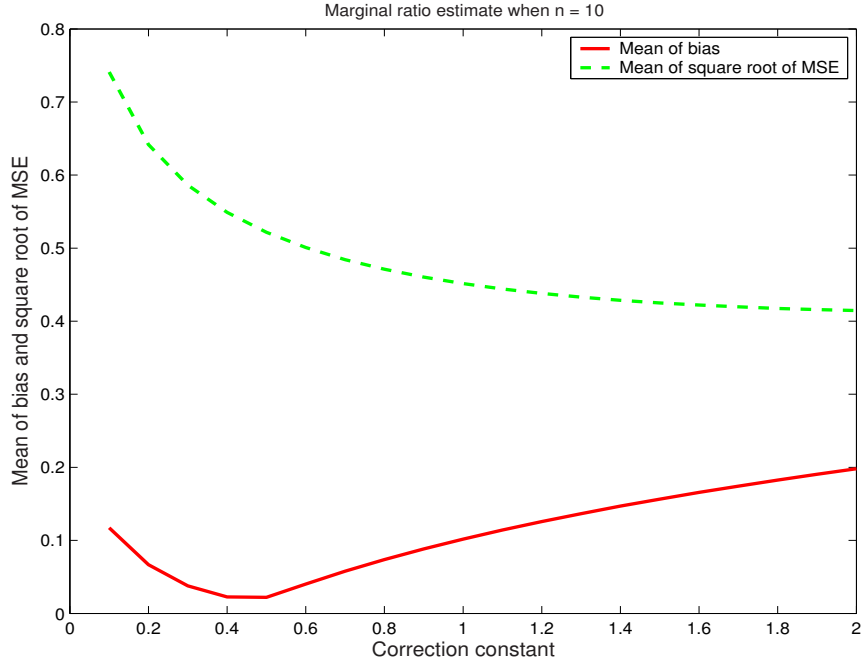
Figure 3: Log marginal ratio estimate n= 10

shows the result when $n = 10$. Similar results occur when $n = 20$. This indicates that $c = 0.5$ is a good generalized correction constant for $\log \tilde{\theta}_m$.

## 2.4  Estimating the Ratio of Poisson Rates

For a Poisson random variable $Y$, with mean $\lambda$, given a sample with counts $n > 0$, the MLE of $\lambda$ is $\hat{\lambda} = n$ [7]. It is easy to show that

$$E[log(Y + c)] = log(\lambda) + (c - \frac{1}{2})/\lambda + O(\lambda^{-2}).$$

Thus, for two independent Poisson random variables, Agresti [1] showed that using the sample log odds ratio after adding $\frac{1}{2}$ to each cell can reduce bias in estimating the log Poisson ratio.

Figure 4: Log Poisson ratio estimate

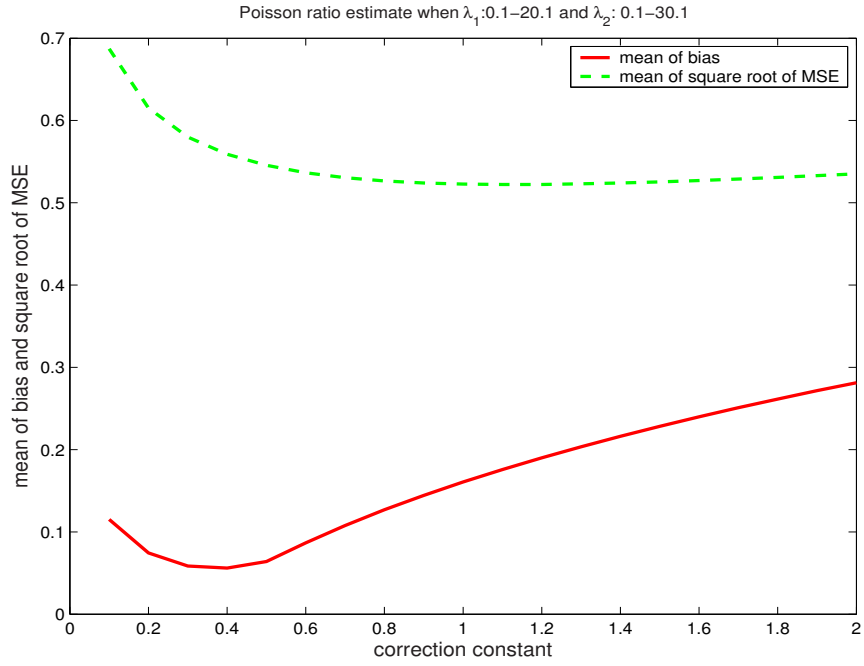By considering 10 times the real mean as the possible maximum outcome, with two independent Poissons with means ranging from 0.1 to 20.1 and 0.1 to 30.1, one can compare the effects of different amended estimators. According to Figure 4, 0.4 to 0.6 is an acceptable correction constant to be added to each cell while estimating the log Poisson ratio based on the mean bias and square root of MSE.

# 3 BAYESIAN METHODS

Bayesian data analysis is a practical method for making inferences from data using probability models for quantities we observe and for quantities about which we wish to learn. It facilitates a common-sense interpretation of statistical conclusions. The process of Bayesian data analysis can be idealized by dividing it into the following three steps [6]:

1. Setting up a *full probability model.*

2. Conditioning on observed data: calculating and interpreting the appropriate *posterior distribution.*

3. Evaluating the fit of the model and the implications of the resulting posterior distribution.

In order to make a probability statement about $\theta$ given $y$, we must begin with a *model* providing a *joint probability distribution* for $\theta$ and $y$. The joint probability mass or density function can be written as a product of two densities that are often referred to as the *prior distribution $p(\theta)$* and the *sampling distribution* (or *data distribution*) $p(y|\theta)$ respectively:

$$p(\theta, y) = p(\theta)p(y|\theta). \tag{15}$$

Simply conditioning on the known value of the data $y$, using the basic property of conditional probability known as Bayes' rule, the *posterior* density is

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{16}$$

whereby the formula of conditional probability, we have $p(y) = \sum_\theta p(\theta)p(y|\theta)$, and the sum is over all possible values of $\theta$ (or $p(y) = \int p(y)p(y|\theta)d\theta$ for continuous $\theta$). For the sake of computational convenience, the posterior density is usually expressed in the unnormalized form,

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{17}$$

where "$\propto$" stands for "proportional to." Note that the probability function $p(y|\theta)$ is often referred to as the likelihood function. Thus, *Bayes' formula* can be simply interpreted in words by the statement that the posterior density is proportional to the likelihood function times the prior density. This implies that Bayesian inference, like the traditional methodologies of statistics, also complies with the likelihood principle. Nevertheless, a particular specification of the prior density $p(\theta)$ will make the posterior distribution not dependent on the data only.

Simulation forms a central part in most applied Bayesian analysis, since it is relatively easy to generate samples from a probability distribution even when the density function cannot be explicitly integrated. In performing simulations, it is helpful to consider the duality between a probability density function and a histogram of a set of random draws from the distribution: given a large enough sample, the histogram can provide near complete information about the density. From these simulated values, we can estimate the posterior distribution of any quantity of interest.

## 3.1 Using the Posterior to Estimate the Odds Ratio

It is natural to consider the sampling distribution as a multinomial distribution. If $y$ is the vector of counts then

$$p(y|\theta) \propto \prod_{i=1}^{4} \theta_i^{y_i}, \tag{18}$$

where the sum of the probabilities, $\sum_{i=1}^{4} \theta_i$, is 1 and the distribution is implicitly conditioning on the number of observations, $\sum_{i=1}^{4} y_i = n$.

The conjugate prior distribution is a multivariate generalization of the beta distribution known as the Dirichlet,

$$p(\theta|\alpha) \propto \prod_{i=1}^{4} \theta_i^{\alpha_i - 1} \tag{19}$$

The resulting posterior distribution for the $\theta_i$'s is a Dirichlet with parameters $\alpha_i + y_i + c_1$ if we add a correction constant $c_1$ to each $y_i$. Since our target is to get a general parameter for all possible outcomes of cells, it is reasonable to let $\alpha_i = c_2$ for $i = 1, 2, 3, 4$. If $c = c_1 + c_2$, the posterior density is given by

$$p(\theta|y) \propto \prod_{i=1}^{4} \theta_i^{y_i + c - 1}. \tag{20}$$

We will use simulation to get our point estimator. Given a sample size $n$ and $c$ value, for each possible outcome, draw $x_1, \ldots, x_4$ from independent gamma distributions with common scale 1 and shape parameter $y_1 + c, y_2 + c, y_3 + c, y_4 + c$. For each $i$, let $\theta_i = x_i / \sum_{i=1}^{k} x_i$, $log \, \tilde{\theta}_r = log(\theta_1 \theta_4) - log(\theta_2 \theta_3)$. Repeating this procedure 1000 times produces 1000 draws of the posterior odds ratio. We will then take the median of this posterior as our point estimate.
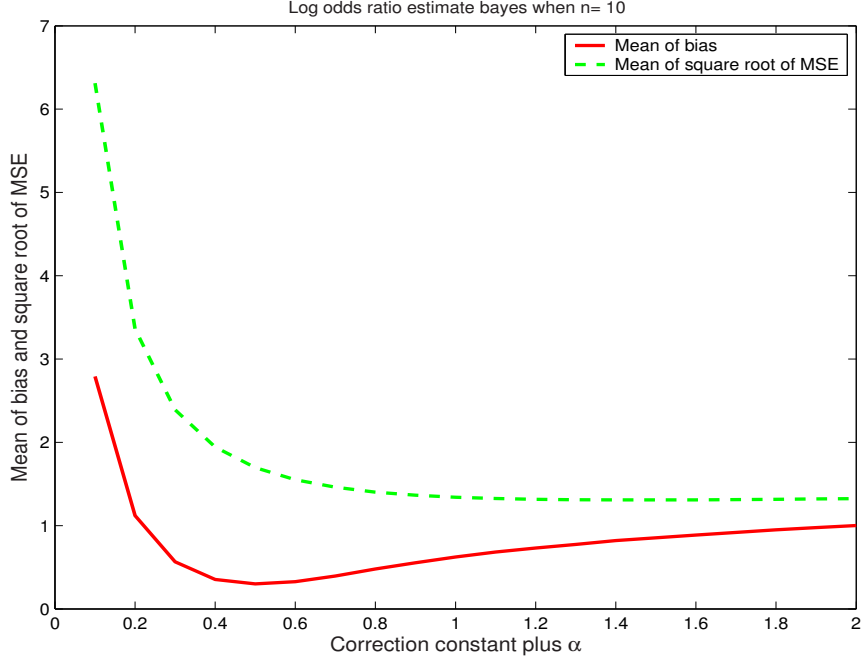
Figure 5: Log odds ratio estimate (Bayesian) n= 10

If the real parameters $p_1, p_2, p_3,$ and $p_4$ are given then the log odds ratio is $log\,\theta_r = log(p_1p_4) - log(p_2p_3)$. The mean squared error of our estimator $log\,\tilde{\theta}_r$ is

$$MSE(log\,\tilde{\theta}_r) = E(log\,\tilde{\theta}_r - log\,\theta_r)^2 = \sum_{y \in Y}((log\,\tilde{\theta}_r - log\,\theta_r)^2 p(y|\theta) \qquad (21)$$

where Y is the space of all possible outcomes. The bias of our estimator is

$$b(log\,\tilde{\theta}_r) = E(log\,\tilde{\theta}_r) - log\,\theta_r = \sum_{y \in Y}(log\,\tilde{\theta}_r - log\,\theta_r)p(y|\theta) \qquad (22)$$

Since the real parameters can be any value on the interval (0,1), we grid each parameter from .01 to .97 in increments of .01. For each point in this multidimensional grid, we can get a bias and MSE for posterior estimator given the total count $n$ and $c$ value. Thus we use mean bias and MSE as our criterion to find the optimized $c$ value. As shown in Figure 5, $c = .5$ is good value when the sample is small.

30

## 3.2    Using the Posterior to Estimate the Relative Risk

Relative risk is usually used to compare two groups on a binary response variable. The sampling distribution is product binomial,

$$p(y|\theta, n_1, n_2) = \binom{n_1}{y_1}\theta_1^{y_1}(1-\theta_1)^{n_1-y_1}\binom{n_2}{y_2}\theta_2^{y_2}(1-\theta_2)^{n_2-y_2}. \tag{23}$$

Because of the vague prior information, we choose a conjugate prior for both $\theta_1$ and $\theta_2$,

$$\theta_i \sim Beta(c_2, c_2), \text{ for } i = 1, 2.$$

After adding a correction constant $c_1$ to each $y_i$, the posterior distribution is

$$\theta_i|y \sim Beta(y_i + c - 1, n_i - y_i + c - 1) \tag{24}$$

where $c = c_1 + c_2$.

Draw $x_1, x_2$ from independent gamma distributions 1000 times with common scale 1 and shape parameters $y_1 + c$, $n_1 - y_1 + c$, and then let $\theta_1 = x_1/(x_1 + x_2)$. Similarly, draw $x_3, x_4$ from independent gamma distributions 1000 times with common scale 1 and shape parameters $y_2 + c$, $n_2 - y_2 + c$, and then let $\theta_2 = x_3/(x_3 + x_4)$. For each pair $\theta_1$ and $\theta_2$, let $log\,\tilde{\theta}_r = log\,\theta_1 - log\,\theta_2$. We take the median of these 1000 $log\,\tilde{\theta}_r$ as the posterior estimator of the log relative risk.

Since the conditional probabilities of the cells range from 0 to 1.0, we grid all cell probabilities from .01 to .99 in increments of .01. For each point in this multidimensional grid, we can get a bias and MSE for the posterior estimator given $n1, n2$, and $c$. By comparing the mean of bias and MSE of all grid points with different $c$ values, one can get the optimized correction constant. For example, when $n_1 = 5$ and $n_2 = 10$, Figure 6 shows that .5 is a good value for $c$.
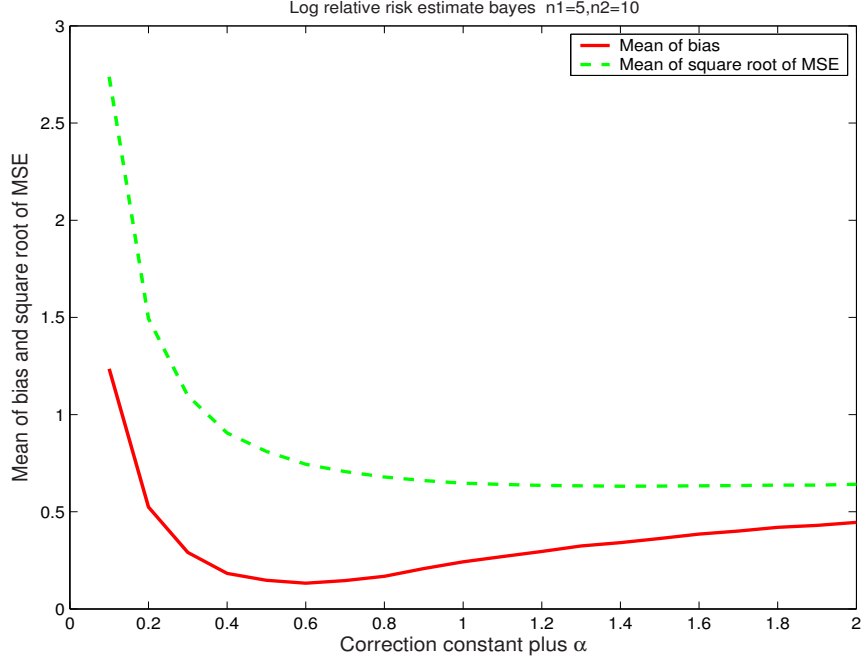
Figure 6: Log relative risk estimate (Bayesian) with $n_1 = 5$ and $n_2 = 10$

### 3.3 Using the Posterior to Estimate the Ratio of Marginal Binomial Proportions

Similar to the odds ratio, we consider the sampling distribution to be multinomial. Let $y$ be the vector of counts. We add the same correction constant $c_1$ to each $y_i$, and let $\alpha_i = c_2$ for $i = 1, 2, 3, 4$, let $c = c_1 + c_2$, then the posterior is

$$p(\theta|y) \propto \prod_{i=1}^{4} \theta_i^{y_i+c-1}. \tag{25}$$

Given a sample size $n$ and $c$ value, for each possible outcome, draw $x_1, \ldots, x_4$ from independent gamma distributions 1000 times with common scale 1 and shape parameters $y_1 + c, y_2 + c, y_3 + c, y_4 + c$, and for each $i$, let $\theta_i = x_i / \sum_{i=1}^{4} x_i$, and $log\,\tilde{\theta}_m = log(\theta_1 + \theta_2) - log(\theta_1 + \theta_3)$. From these 1000 $\tilde{\theta}_m$'s, we take the median as our estimator.
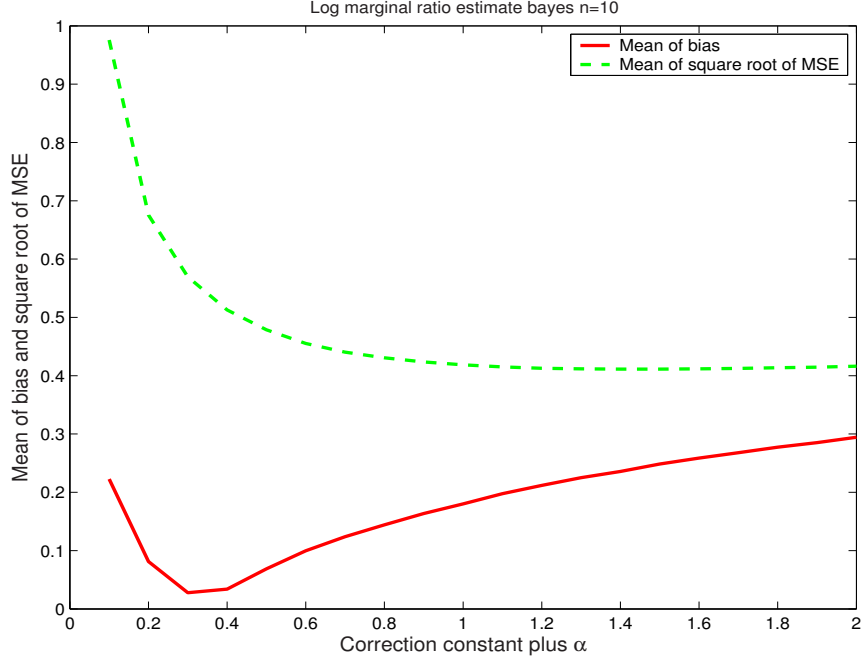
Figure 7: Log marginal ratio estimate (Bayesian) n= 10

If the real parameters $p_1, p_2, p_3$, and $p_4$ are given, the log marginal ratio $log\, \theta_m = log(p_1 + p_2) - log(p_1 + p_3)$. The mean squared error of our estimator $log\, \tilde{\theta}_m$ is

$$MSE(log\, \tilde{\theta}_m) = E(log\, \tilde{\theta}_m - log\, \theta_m)^2 = \sum_{y \in Y}((log\, \tilde{\theta}_m - log\, \theta_m)^2 p(y|\theta) \qquad (26)$$

where Y is the space of all possible outcomes. The bias is

$$b(log\, \tilde{\theta}_m) = E(log\, \tilde{\theta}_m) - log\, \theta_m = \sum_{y \in Y}(log\, \tilde{\theta}_m - \theta_m)p(y|\theta_m). \qquad (27)$$

We grid each parameter from .01 to .97 in increments of .01. For each point in this multidimensional grid, we can get a bias and MSE for the posterior estimator given the total count $n$ and $c$ value. Thus we use the mean of the bias and MSE as our criterion to find that an acceptable $c$ value is .5 when $n = 10$ (see Figure 7).

33

## 3.4  Using the Posterior to Estimate the Ratio of Poisson Rates

The sampling distribution of independent Poisson random variables is

$$p(y_1, y_2|\theta_1, \theta_2) = \frac{e^{-\theta_1}\theta_1^{y_1}}{y_1!} \frac{e^{-\theta_2}\theta_2^{y_2}}{y_2!}. \tag{28}$$

The conjugate prior for the Poisson model is a gamma density,

$$p(\theta) \propto e^{-\beta\theta}\theta^{\alpha-1}.$$

We choose $Gamma(c_2, 0)$ as our prior for both $\theta_1$ and $\theta_2$. With this conjugate prior, after adding a correction constant $c_1$ to each $y_i$, the posterior distribution is

$$\theta_i|y \sim Gamma(y_i + c, 1)$$

where $c = c_1 + c_2$ and $i = 1, 2$. Draw $\theta_1, \theta_2$ from their posterior Gamma distribution 1000 times respectively. Let $log\,\tilde{\theta}_p = log\,\theta_1 - log\,\theta_2$. We again take the median of these 1000 posterior $log\,\tilde{\theta}_p$'s as the estimator.

By considering 10 times the real mean as the possible maximum outcome, with two Poisson means ranging from 0.1 to 4.1 and 0.1 to 7.1, one can compare the effects of different posterior estimators. According to Figure 8, 0.5 is an acceptable $c$ value while estimating the log Poisson ratio based on the mean of the bias and the square root of MSE.
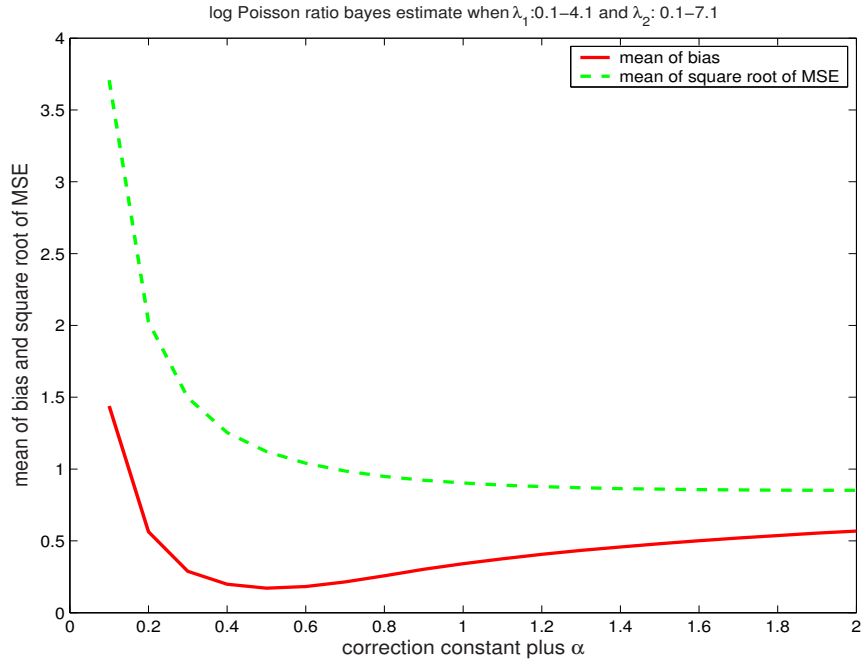
Figure 8: Log Poisson ratio estimate (Bayesian)

## 4 CONCLUSION

If we evaluate the amended log ratios in terms of bias and MSE, 0.5 is an ideal correction constant for the following measures of association: the odds ratio, the relative risk, the ratio of marginal binomial proportions and the ratio of independent Poisson rates. As for the Bayesian method, the selection of an optimized correction constant depends on the prior distribution. If we choose the prior distribution with parameter $\alpha = 0.5$, the correction constant is no longer needed.

# BIBLIOGRAPHY

[1] Alan Agresti (1990), *Categorical data analysis*, John Wiley & Sons.

[2] Alan Agresti (1990), On Logit Confidence Intervals for the Odds Ratio with Small Samples. *Biometrics*, **55**, (597-602).

[3] Douglas G. Bonett and Robert M. Price (2006), Confidence Intervals for a Ratio of Binomial Proportions Based on Paired Data. *Statistics in Medicine* (IN PRESS).

[4] John J. Gart and James R. Zweifel (1967), On the Bias of Various Estimators of the Logit and Its Variance with Application to Quantal Bioassay. *Biometrika*, **54**, (181-187).

[5] John J. Gart; Jun-mo Nam (1988), Approximate Interval Estimation of the Ratio of Binomial Parameters: A Review and Corrections for Skewness . *Biometrics*, **44**, (323-338).

[6] Andrew Gelman, John B. Carlin, Hal S. Stern and Donald B. Rubin (2004), *Bayesian Data Analysis*, Chapman & Hall, 2nd Ed.

[7] Keith Knight(2000), *Mathematical Statistics*, Chapman & Hall.

[8] Robert M. Price and Douglas G. Bonett (2000), Estimating the Ratio of two Poisson rates. *Computational Statistics and Data Analysis*, **34**, 345-356.

[9] S. D. Walter(1975), The Distribution of Levin's Measure of Attributable Risk. *Biometrika*, **62**, (371-374).

# APPENDICES

## .1 Matlab code for frequentist Poisson ratio estimation

```
clear tic;

for c1 = .1: .1 :2

clear bs exps rmses

exps = []; % stores an empty matrix in A

bs = [];

rmses = [];


for lamda1 =.1:.2:20.1

    for lamda2 =.1:.2:30.1


n1 = round(10 * lamda1);

n2 = round(10 * lamda2);


x1 = (0:n1)';

x2 = (0:n2)';


n11 = kron(x1,ones(n2+1,1));

n21 = kron(ones(n1+1,1),x2);


ratio = log(lamda1) - log(lamda2);
```

```
prob=pdf('poiss',n11,lamda1).*pdf('poiss',n21,lamda2);

est=log(n11+ c1) - log(n21 + c1);


expvalue = est'*prob;

bias = abs(expvalue - ratio);


rmse = (((est-ratio).^2)'*prob)^.5;


exps = [exps; expvalue];

bs = [bs; bias];

rmses = [rmses; rmse];

    end

end

[c1, mean(bs), mean(rmses)]

end

toc;
```

## .2 Matlab code for frequentist relative risk estimation

```matlab
clear tic;

c1s = [];

mbs = [];

mrms = [];


for c1 = .1: .1 :2

clear bs exps rmses


exps = []; % stores an empty matrix in A

bs = [];

rmses = [];


c2 = 1* c1;

n1 =30;

n2 = 45;


x1 = (0:n1)';

x2 = (0:n2)';


n11 = kron(x1,ones(n2+1,1));

n21 = kron(ones(n1+1,1),x2);
```

```
for p1 = .01:.01:.99

    for p2 =.01:.01:.99

ratio = log(p1) - log(p2);


prob =pdf('bino',n11,n1,p1).*pdf('bino',n21,n2,p2);


est = log(n11 + c1)- log(n1 + c2) - log(n21 +c1) +log(n2 +c2);

expvalue = est'*prob;

bias = abs(expvalue - ratio);


rmse = (((est - ratio).^2)'*prob)^.5;

exps = [exps; expvalue];


bs= [bs; bias];

rmses = [rmses; rmse];

    end

end [c1,c2,n1,n2, mean(bs), mean(rmses)]

end

toc;
```

## .3 Matlab code for frequentist odds ratio estimation

```
clear tic;


for c1 = .1: .1 :2

 clear bs exps rmses


exps = []; % stores an empty matrix in A

bs = [];

rmses = [];


% Multinomial

n = 10;

 m = n + 1 ;

 x1 = (0:n)';

 x2 = (0:n)';

 x3 = (0:n)';

 x4 = (0:n)';


 n11 = kron(x1,ones(m^2,1));

 n12 = kron(ones(m,1),kron(x2,ones(m,1)));

 n21 = kron(ones(m^2,1),x3);

 F = [n11,n12,n21];
```

```
t = find(sum(F') <= n);

F = F(t,:);

n11 = F(:,1);

n12 = F(:,2);

n21 = F(:,3);

n22 = (n - sum(F'))' ;



% Multinomial

for p11 = .01:.01 :.97

    for p12 = .01 :.01:(.98- p11)

        for p21 = .01:.01:(.99-p11-p12)

            p22 = 1- p11- p12- p21;

            ratio =log(p11) +log(p22)-log(p21) -log(p12);
%
nk = gammaln(m) - gammaln(n11+1) - gammaln(n12+1) - gammaln(n21+1)

- gammaln(n22+1);
%
lnkp = nk + n11.*log(p11) + n12.*log(p12) + n21.*log(p21) +

n22.*log(p22);

 prob = exp(lnkp);


%est = ((n11 + c1)./(n1 + c2))./((n21 + c1)./(n2 + c2));
```

```
est = log(n11 + c1)- log(n12 + c1) - log(n21 +c1) +log(n22 +c1);

expvalue = est'*prob; bias = abs(expvalue - ratio); rmse = (((est

- ratio).^2)'*prob)^.5;


exps = [exps; expvalue];

bs = [bs; bias];

rmses = [rmses; rmse];

    end

end

end

[c1,n, mean(bs), mean(rmses)]
```

## .4 Matlab code for frequentist marginal ratio estimation

```
clear

for c1 = .3: .2 :2.1 tic; clear bs exps rmses

exps = []; % stores an empty matrix in A
bs = []; rmses = []; n = 20;



m = n + 1 ;
 x1 = (0:n)';
 x2 = (0:n)';
 x3 = (0:n)';
 x4 = (0:n)';
%
 n11 = kron(x1,ones(m^2,1));
 n12 = kron(ones(m,1),kron(x2,ones(m,1)));
 n21 = kron(ones(m^2,1),x3);
 F = [n11,n12,n21];
 t = find(sum(F') <= n);
 F = F(t,:);
 n11 = F(:,1);
```

```
 n12 = F(:,2);

 n21 = F(:,3);

 n22 = (n - sum(F'))' ;


% Multinomial

for p11 = .01:.01 :.97

    for p12 = .01 :.01:(.98- p11)

        for p21 = .01:.01:(.99-p11-p12)

            p22 = 1- p11- p12- p21;

            ratio = (p11 + p12)/ (p11 + p21);
%
nk =gammaln(m)-gammaln(n11+1)-gammaln(n12+1)-gammaln(n21+1)-
gammaln(n22+1);
%
lnkp = nk + n11.*log(p11) + n12.*log(p12) + n21.*log(p21) +
n22.*log(p22);
 prob = exp(lnkp);
 est = (n11+ n12+ c1) ./ (n11 + n21 + c1);


expvalue = est'*prob;

bias = abs(expvalue - ratio);

rmse = (((est

- ratio).^2)'*prob)^.5;
```

```
exps = [exps; expvalue];

bs = [bs; bias];

rmses = [rmses;rmse];

    end

end

end

[c1,n, mean(bs), mean(rmses)]

toc;

end
```

## .5 Matlab code for Bayesian Poisson ratio estimation

```matlab
clear

tic;

for c1 = .1: .1 :2

clear bs exps rmses thetas


exps = []; % stores an empty matrix in A

bs = [];

rmses = [];

thetas = [];


for lamda1 =.1:.2:4.1

    for lamda2 =.1:.2:7.1

clear thetas thetas = [];

n1 = round(10 * lamda1);

n2 = round(10 *lamda2);



x1 = (0:n1)';

x2 = (0:n2)';


n11 = kron(x1,ones(n2+1,1));

n21 = kron(ones(n1+1,1),x2);
```

```
j = length(n11) ;
 for i= 1:j
z1 = gamrnd(n11(i) + c1 ,1,[1000 1]);


z2 = gamrnd(n21(i) + c1 ,1,[1000 1]);

y = log(z1) - log(z2);

thetasimu = median(y);

thetas = [thetas; thetasimu];
 end


ratio = log(lamda1) - log(lamda2);

prob
=pdf('poiss',n11,lamda1).*pdf('poiss',n21,lamda2);

%est = log(n11 + c1) - log(n21 + c1);


expvalue = thetas' * prob;

bias = abs(expvalue - ratio);

rmse =
(((thetas - ratio).^2)'*prob)^.5;


exps = [exps; expvalue];

bs = [bs; bias];
```

```
rmses = [rmses; rmse];

    end

end

[c1, mean(bs), mean(rmses)]

end

toc;
```

## .6    Matlab code for Bayesian odds ratio estimation

```
%bayesian


clear


for c1 =1.2:.1 :2 tic; clear bs exps rmses thetas


exps = []; % stores an empty matrix in A
bs = []; rmses = []; thetas = []; n = 10;



m = n + 1 ;
 x1 = (0:n)';
 x2 = (0:n)';
 x3 = (0:n)';
 x4 = (0:n)';
%
 n11 = kron(x1,ones(m^2,1));
 n12 = kron(ones(m,1),kron(x2,ones(m,1)));
 n21 = kron(ones(m^2,1),x3);
 F = [n11,n12,n21];
 t = find(sum(F') <= n);
 F = F(t,:);
```

```matlab
 n11 = F(:,1);

 n12 = F(:,2);

 n21 = F(:,3);

 n22 = (n - sum(F'))' ;


j = length(n11) ;
 for i= 1:j
y1 = gamrnd(n11(i) + c1 ,1,[1000 1]); y2 = gamrnd(n12(i) + c1

,1,[1000 1]); y3 = gamrnd(n21(i) + c1 ,1,[1000 1]); y4 =

gamrnd(n22(i) + c1 ,1,[1000 1]);

%sumx = x1+x2+x3+x4;

%x11 = x1 ./ sumx;

%x21 = x2 ./ sumx;

%x31 = x3 ./ sumx;

%x41 = x4 ./ sumx;


y = log(y1) +log(y4) - log(y2) -log(y3);

%x = min(y):50:1000;

%subplot(1,2,1)

%hist(y,x)

%title('Normally Distributed Random Numbers','FontSize',16)

thetasimu = median(y); thetas = [thetas; thetasimu]; end
```

```
% Multinomial

for p11 = .01:.01 :.97

    for p12 = .01 :.01:(.98- p11)

        for p21 = .01:.01:(.99-p11-p12)

            p22 = 1- p11- p12- p21;

            ratio =log(p11) +log(p22)-log(p21) -log(p12);
%
nk = gammaln(m) - gammaln(n11+1) - gammaln(n12+1) - gammaln(n21+1)

- gammaln(n22+1);
%
lnkp = nk + n11.*log(p11) + n12.*log(p12) + n21.*log(p21) +

n22.*log(p22);
 prob = exp(lnkp);
 %est = (n11+ n12+ c1) ./ (n11 + n21 + c1);


 expvalue = thetas' * prob;
bias = abs(expvalue - ratio); rmse = (((thetas-

ratio).^2)'*prob)^.5; exps = [exps; expvalue]; bs = [bs; bias];

rmses = [rmses; rmse];

    end

end end [c1,n, mean(bs), mean(rmses)] toc; end
```

## .7   Matlab code for Bayesian relative risk estimation

```
clear tic; c1s = []; mbs = []; mrms = [];




for c1 = .1: .1 :2 clear bs exps rmses thetas




exps = []; % stores an empty matrix in A

bs = []; rmses = []; thetas = [];

  c2 = 1* c1;

%c2 = .5;

n1 =5; n2 = 10;




x1 = (0:n1)'; x2 = (0:n2)';




n11 = kron(x1,ones(n2+1,1)); n21 = kron(ones(n1+1,1),x2); n12 = n1

-n11; n22 = n2 - n21; j = length(n11) ;

 for i= 1:j

z1 = gamrnd(n11(i) + c1 ,1,[1000 1]); z2 = gamrnd(n1 + c2 ,1,[1000

1]); z3 = gamrnd(n21(i) + c1 ,1,[1000 1]); z4 = gamrnd(n2 + c2

,1,[1000 1]);




y1 = z1 ./  z2; y2 = z3 ./ z4;
```

```
y = log(y1) - log(y2);

%x = min(y):50:1000;


thetasimu = median(y); thetas = [thetas; thetasimu]; end


for p1 = .01:.01:.99
    for p2 =.01:.01:.99
ratio = log(p1) - log(p2); prob =
pdf('bino',n11,n1,p1).*pdf('bino',n21,n2,p2);


expvalue = thetas'*prob; bias = abs(expvalue - ratio); rmse =
(((thetas - ratio).^2)'*prob)^.5;
%grid on;
%plot3(p1,p2,rmse);
exps = [exps; expvalue]; bs = [bs; bias]; rmses = [rmses; rmse];
    end
end


[c1,n1,n2, mean(bs), mean(rmses)] c1s = [c1s;c1]; mbs =
[mbs;mean(bs)]; mrms = [mrms;mean(rmses)];



end
```

```
toc;
```

## .8   Matlab code for Bayesian marginal ratio estimation

```
%bayesian


clear


for c1 =.4:.1 :.7 tic; clear bs exps rmses thetas


exps = []; % stores an empty matrix in A
bs = []; rmses = []; thetas = []; n = 10;



m = n + 1 ;
 x1 = (0:n)';
 x2 = (0:n)';
 x3 = (0:n)';
 x4 = (0:n)';
%
 n11 = kron(x1,ones(m^2,1));
 n12 = kron(ones(m,1),kron(x2,ones(m,1)));
 n21 = kron(ones(m^2,1),x3);
 F = [n11,n12,n21];
 t = find(sum(F') <= n);
 F = F(t,:);
```

```
n11 = F(:,1);

n12 = F(:,2);

n21 = F(:,3);

n22 = (n - sum(F'))' ;


j = length(n11) ;
 for i= 1:j
y1 = gamrnd(n11(i) + c1 ,1,[1000 1]); y2 = gamrnd(n12(i) + c1

,1,[1000 1]); y3 = gamrnd(n21(i) + c1 ,1,[1000 1]);



y = log(y1 + y2) - log(y1 + y3);



thetasimu = median(y); thetas = [thetas; thetasimu]; end


% Multinomial

for p11 = .01:.01 :.97

    for p12 = .01 :.01:(.98- p11)

        for p21 = .01:.01:(.99-p11-p12)

            p22 = 1- p11- p12- p21;

            ratio = log(p11 + p12)- log(p11 + p21);
%
nk = gammaln(m) - gammaln(n11+1) - gammaln(n12+1) - gammaln(n21+1)
```

```
- gammaln(n22+1);
%
lnkp = nk + n11.*log(p11) + n12.*log(p12) + n21.*log(p21) +
n22.*log(p22);
 prob = exp(lnkp);
 %est = (n11+ n12+ c1) ./ (n11 + n21 + c1);


 expvalue = thetas' * prob;
bias = abs(expvalue - ratio); rmse = (((thetas-
ratio).^2)'*prob)^.5; exps = [exps; expvalue]; bs = [bs; bias];
rmses = [rmses; rmse];
    end
end end [c1,n, mean(bs), mean(rmses)] toc; end
```

VITA

Dandan Chen

| | |
|---|---|
| Personal Data: | Date of Birth: November 9, 1970 |
| | Place of Birth: Zhejiang, China |
| | Marital Status: Married |
| | |
| Education: | B.S. in Automatic Control, Southeast University, |
| | Nanjing, China 1992 |
| | M.S. in Mathematics, East Tennessee State University, |
| | Johnson City, Tennessee 2006 |
| | |
| Professional Experience: | Graduate Assistant, East Tennessee State University, |
| | Department of Mathematics, 2004 -2006 |