Notes on a Bayesian Framework for Association Studies with Multiple Hypotheses

William H. Press

March 31, 2005

1 Introduction

There is a vast literature devoted to the question of deciding whether a contingency table with modest numbers of counts, say something like

	C_0	C_1
f_0	7	12
f_1	8	3
totals	15	15

(1)

shows a *significant* difference between a control sample C_0 and a diagnosed sample C_1 with respect to some feature or test with values f_0 and f_1. The modest count values require so-called "exact" (as opposed to asymptotic) methods; Agresti[1] surveys the (generally frequentist) literature.

The standard approach for a table like Table (1) is:

1. Choose a test statistic that quantifies the discrepancy with the null hypothesis that there is no association between (C_0, C_1) and (f_0, f_1). Popular choices are the Wald statistic and Rao's efficient score statistic.

2. Choose a method, for example "Fisher's exact test" or "Barnard's exact test". The choice of a method is equivalent to choosing a population of 2×2 tables against which to compare Table (1).

3. Compute the one-sided p-value, that is, the probability of finding in the population defined by the method a value of the test as extreme as that seen.

4. Reject the null hypothesis (no association) if $p < 0.05$ (say).

Within this paradigm, the single most popular (and therefore most standard) set of choices is probably Fisher's exact test with the Wald statistic.

There are widely recognized weaknesses in steps 1, 2, and 4 in this paradigm; only step 3 seems objectively above reproach!

Step 1 is open to the usual criticism of p-tests, namely that different choices of statistic can give quite different tail probabilities for a given data set. In practice, however, the differences are rarely large. The issue is common to all p-tests and therefore conveniently swept under the rug.

Step 4 is open to the criticism that, for small numbers of counts, p may take on well-spaced discrete values. Requiring $p < 0.05$ may be excessively conservative if consecutive discrete values lie at (say) 0.051 and 0.01. The inevitable advice given (and rarely followed) is to focus on the p value, and not the accept/reject decision.

Step 3 is by far the most problematic. The issue is that the null hypothesis contains an unknown parameter π_0, the common probability of seeing f_0 under both C_0 and C_1. Fisher's exact test eliminates π_0 by fixing a row marginal at its observed value (e.g., 19 in Table 1), and thus compares Table 1 only to the population of tables having all marginals identical to Table 1. Barnard's exact test instead chooses a specific value of π_0, essentially its maximum likelihood value given the observed data. (A clear pedagogical discussion is in [2].) The problem with any such choice is that, absent clairvoyance, it inevitably assumes a *wrong* value for π_0. Thus the p-values are not tail probabilities of the test statistic for the actual experiment, but only uncontrolled approximations. In no sense is it true that the null hypothesis will be incorrectly rejected only 1 time in 20.

For the counts in Table (1), using the Wald statistic, the Fisher exact test gives $p = 0.0641$, while the Barnard exact test gives $p = 0.0341$.[2] While the frequentist literature delights in disecting this kind of small difference, our interest here is in the more fundamental question of whether an association is established with anything like the implied signficance of these methods.

2 Comparing Hypotheses

In what follows, we will frequently use a basic Bayesian technique for comparing two hypotheses that are competing as possible explanations of a data set, by the ratio of their likelihoods (or, in more Bayesian terminology, their probabilities).

Suppose two hypotheses are H_A and H_B, and the data is D. Further suppose that H_A has n_A parameters, which we take as a vector $\boldsymbol{\lambda}_A$ and that H_B has n_B

parameters $\boldsymbol{\lambda}_B$. Then we write (see, e.g., [3]),

$$\begin{aligned}\frac{\text{prob}(H_A|D)}{\text{prob}(H_B|D)} &= \frac{\text{prob}(D|H_A)}{\text{prob}(D|H_B)} \times \frac{\text{prob}(H_A)}{\text{prob}(H_B)} \\ &= \frac{\int \text{prob}(D, \boldsymbol{\lambda}_A|H_A)\ d^{n_A}\lambda}{\int \text{prob}(D, \boldsymbol{\lambda}_B|H_B)\ d^{n_B}\lambda} \times \frac{\text{prob}(H_A)}{\text{prob}(H_B)} \\ &= \frac{\int \text{prob}(D|H_A, \boldsymbol{\lambda}_A)\ \text{prob}(\boldsymbol{\lambda}_A|H_A)\ d^{n_A}\lambda}{\int \text{prob}(D|H_B, \boldsymbol{\lambda}_B)\ \text{prob}(\boldsymbol{\lambda}_B|H_B)\ d^{n_B}\lambda} \times \frac{\text{prob}(H_A)}{\text{prob}(H_B)}\end{aligned} \tag{2}$$

Here $d^{n_A}\lambda$ signifies integration over all of H_A's parameters, and similarly for H_B. Absent other information, we might in some cases take the ratio of prior probabilities for the two hypotheses, $\text{prob}(H_A)/\text{prob}(H_B)$, as unity, if we have no reason to favor one hypothesis over the other; but we might in other cases take the ratio to be a small number, especially if H_A is just one of many possible hypotheses that we are testing, each one unlikely by itself. Within each hypothesis, we also need prior probabilities on its parameters $\boldsymbol{\lambda}$, an issue that we address below.

An important special case of equation (2) is where H_A and H_B differ only in that one of them (H_B, say) requires one or more constraints among its parameters $\boldsymbol{\lambda}_B$. An example might be requiring the equality of two of the λ's, say λ_i and λ_j. For this example, and with the assumption on priors already mentioned, equation (2) becomes

$$\frac{\text{prob}(H_A|D)}{\text{prob}(H_B|D)} = \frac{\int \text{prob}(D|H_A, \boldsymbol{\lambda}_A)\ \text{prob}(\boldsymbol{\lambda}_A|H_A)\ d^{n_A}\lambda}{\int \text{prob}(D|H_B, \boldsymbol{\lambda}_B)\ \text{prob}(\boldsymbol{\lambda}_B|H_B)\ \delta(\lambda_i - \lambda_j)\ d^{n_B}\lambda} \tag{3}$$

where δ is a Dirac delta function. That is, we integrate over the *allowed* space of parameters.

3 Contingency Tables of Probabilities

In situations of interest, we often have a set of mutually exclusive diagnoses or conditions, denoted $C_1, C_2, \ldots$, and also the absence of any of the diagnoses – that is, the control group – which we denote C_0. Each diagnosis is a column in a contingency table.

We are also given a set of (again mutually exclusive) feature vector values, $f_0, f_1, f_2, \ldots$, that characterize individuals in all the condition groups. By feature vector, we mean the aggregate of results of *all* the test results or genetic markers under study. For example, if each test is either positive (denoted 1) or negative (denoted 0), and if there are 6 tests, then there are 2^6 f's, labeled in binary as $f_{000000}, \ldots, f_{111111}$. In other words, the f's maximally *deaggregate* the tested groups, just as the C's maximally deaggregate the diagnoses.

In principle, if we had an infinite population, we could assign probabilities p_{ij} to each of the f_i's under each column C_j, giving a table like

	C_0	C_1	C_2	$\cdots$
f_0	p_{00}	p_{01}	p_{02}	$\cdots$
f_1	p_{10}	p_{11}	p_{12}	$\cdots$
f_2	p_{20}	p_{21}	p_{22}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

(4)

The p_{ij}'s in each column sum to 1, by their definition.

If an undiagnosed patient presents with feature vector f_i, then the odds that he/she has a particular condition C_j are, from equation (2),

$$\begin{aligned} \frac{\text{prob}(C_j|f_i)}{\text{prob}(C_0|f_i)} &= \frac{\text{prob}(f_i|C_j)}{\text{prob}(f_i|C_0)} \times \frac{\text{prob}(C_j)}{\text{prob}(C_0)} \\ &= \frac{p_{ij}}{p_{i0}} \times \frac{\text{prob}(C_j)}{\text{prob}(C_0)} \end{aligned} \tag{5}$$

The ratio of priors is, essentially, the incidence of C_j in the population at large. The ratio p_{ij}/p_{i0} is the "evidence odds ratio", which we will often just call the *odds ratio*. A main goal of this note is to estimate such odds ratios in a meaningful way.

An important point is that, for realistic situations, there should be many instances where p_{ij} is almost exactly equal to p_{i0}, meaning that the feature f_i has virtually no predictive value in diagnosing condition C_j. We say "almost" and "virtually" merely to cover a debater's possible objection that "everything is causally related to everything, at least in some tiny high-order way". In practical situations, such high-order correlations vanish so rapidly that it is meaningful to classify p_{ij}'s as either *equal* to the control population's p_{i0} or *different* from it. What makes this assertion not tautological is that it will lead us to assign a finite prior probability to the case of "equal", sometimes approaching unity. In other words, "equal" is not a set of measure zero in this application, but is indeed often the (prior) most likely case, since we expect most features studied to prove irrelevant to most diagnoses.

One may well worry that with maximally deaggregated multiple tests and multiple diagnoses, the number of individuals in any single box in equation (4) will be so small that no meaningful p_{ij} can be estimated. That is precisely the subject of the rest of this note. The framework that we develop will (i) give meaningful estimates of whether p_{ij}'s should be classified as "different" (having predictive value) or "equal" (having no predictive value), (ii) estimate the resulting odds ratios, taking into account the possibility of having small numbers of counts, and (iii) provide a methodology for comparing more- or

less-aggregated hypothesis spaces, so that features and/or diagnoses will be aggregated in such a way as to provide odds ratios with themselves a high probability of being meaningful.

4 Contingency Tables of Counts and the Poisson Assumption

A given experiment yields, of course, counts n_{ij}, not probabilities p_{ij}:

	C_0	C_1	C_2	$\cdots$
f_0	n_{00}	n_{01}	n_{02}	$\cdots$
f_1	n_{10}	n_{11}	n_{12}	$\cdots$
f_2	n_{20}	n_{21}	n_{22}	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$
totals	$n_{.0}$	$n_{.1}$	$n_{.2}$	$\cdots$

(6)

Here n_{ij} is the number of counts with feature vector f_i and diagnosis C_j, while $n_{.j}$ denotes the marginals by diagnosis, that is, the number of patients studied with each condition C_j (including the control group C_0).

We next assume an independent Poisson model for the number of counts n_{ij} in each box. That is, n_{ij} is assumed to be drawn from an independent Poisson distribution with (its own) rate parameter λ_{ij},

$$\begin{aligned} n_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \text{prob}(n_{ij}|\lambda_{ij}) &= \frac{1}{\Gamma(n_{ij}+1)} \lambda_{ij}^{n_{ij}} e^{-\lambda_{ij}} \end{aligned} \tag{7}$$

Since much will follow from this assumption, a few words on its validity are in order.

The Poisson distribution is the limiting form of a binomial or multinomial process when the probability of selection is small. Furthermore, and more interestingly, once a finite sample has been selected via an independent Poisson process, any subdivision of it into subsamples by a binomial or multinomial process (that is, with probabilities summing to unity) produces subsamples that are not only Poisson distributed, but *independently* Poisson distributed. This follows from the fact that any such subsample could have been identically drawn from the original population by applying the compound selection criterion from

the start; the winnowing into sample and then subsample is purely procedural and does not affect the outcome.

Thus, if we can identify a Poisson process anywhere "above" a single count n_{ij}, it will follow that n_{ij} itself is independently Poisson. In many studies, the covering Poisson process is inherent in the experimental design: Criteria for patient or control group participation are established, and all individuals who "walk in the door" meeting these criteria (and other implicit criteria, such as being aware of the study) are accepted.

Next best, even if a study accepts a predetermined number of patients overall, the column marginals $n_{\cdot j}$ may be accurately enough Poisson if the number of patients with each diagnosis is not fixed a priori.

Next best after that, even if the experimental design fixes the column marginals a priori, independent Poisson will still be an adequate approximation as long as all the ratios $n_{ij}/n_{\cdot j}$ are small, i.e., there are many features with comparable probabilities.

Finally, although independent Poisson is a computational convenience, a number of our results are actually more general and follow in the case of a multinomial distribution as well. To summarize: assuming independent Poisson counts is not likely to get us into trouble!

An important fact about independent Poisson processes is that any sum of their deviates is itself Poisson, and drawn from a process whose rate is the sum of the individual rates. This property allows us to aggregate cells at will by summing their counts and λ's, and we use it often below.

Given a cell with counts n (we drop the subscripts when they are not relevant), the probability distribution of the underlying λ is given by

$$\text{prob}(\lambda|n) \propto \text{prob}(n|\lambda)\ \text{prob}(\lambda) \tag{8}$$

As always, we need a prior, in this case $\text{prob}(\lambda)$. Two natural priors, each scale-free in a certain way, are the uniform prior,

$$\text{prob}(\lambda) \propto d\lambda \tag{9}$$

and the log-uniform prior

$$\text{prob}(\lambda) \propto \frac{d\lambda}{\lambda} \tag{10}$$

(Both of these are improper, in the sense of being not integrable, but this will never matter in the calculations below.)

One way of distinguishing between these (or any other) priors, is to see what they predict for the mean number of counts $\langle n \rangle$ in a cell whose observed number of counts is n. Since the mean of a Poisson process with rate λ is equal to λ, we need just calculate

$$\langle n \rangle = \int_0^\infty \lambda\ \text{prob}(\lambda|n) d\lambda = \frac{\int_0^\infty \lambda\ \text{prob}(n|\lambda)\ \text{prob}(\lambda) d\lambda}{\int_0^\infty \text{prob}(n|\lambda)\ \text{prob}(\lambda) d\lambda} \tag{11}$$

The integrals are straightforward, and give the result $n + 1$ in the case of the uniform prior and n in the case of the log-uniform prior.

While the log-uniform prior thus seems "unbiased", its choice has rather severe consequences for any cell with observed $n = 0$. The fact that the mean $\langle n \rangle$ is then also zero implies that, with a log-uniform prior, a zero cell may *never* take on a nonzero count. Clearly this is unreasonable, as is any prior that gives zero probability to a case that can actually occur. We therefore adopt as our standard the uniform prior which, *on average*, credits every cell with one extra count that was not (yet) seen. This is a conservative assumption, in that it tends to nudge cases with small numbers of counts towards the (uninformative) uniform distribution. (One could of course consider other priors with intermediate behavior, for example $\lambda^{-1/2}$; the effect on any of our results would be slight.)

With a uniform prior on λ, the normalized density $\text{prob}(\lambda|n)$ is

$$\text{prob}(\lambda|n) = \frac{1}{\Gamma(n+1)} \lambda^n e^{-\lambda} \tag{12}$$

5 Probability from Counts

Let us focus attention on a single cell and its column marginal, namely,

	$\cdots$	C_j	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
f_i	$\cdots$	n	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
totals	$\cdots$	N	$\cdots$

(13)

The probability distribution associated with this pattern of data is the joint distribution of two λ's, one for the observed n, and the other for the observed $N - n$ as the sum of all the other independent Poisson cells in the column. We denote the two rates λ_n and λ_{N-n}. Their joint distribution, from equation (12), and the fact they are independent, is

$$\text{prob}(\lambda_n, \lambda_{N-n}|n, N) = \frac{1}{\Gamma(n+1)\Gamma(N-n+1)} {\lambda_n}^n {\lambda_{N-n}}^{N-n} \exp(-\lambda_n - \lambda_{N-n}) d\lambda_n d\lambda_{N-n} \tag{14}$$

The reason that we write the differentials $d\lambda_n d\lambda_{N-n}$ explicitly, is that we now want to change variables as follows:

$$\begin{aligned} \lambda_n &\equiv p\lambda \\ \lambda_{N-n} &\equiv (1-p)\lambda \end{aligned} \tag{15}$$

Although equation (15) is just a definition, it has the obvious interpretation that λ is the rate parameter characteristic of the whole column, while p is the

probability associated with the single cell of interest within the column. The Jacobian determinant is easily evaluated,

$$\frac{\partial(\lambda_n, \lambda_{N-n})}{\partial(p, \lambda)} = \begin{vmatrix} \frac{\partial \lambda_n}{\partial p} & \frac{\partial \lambda_n}{\partial \lambda} \\ \frac{\partial \lambda_{N-n}}{\partial p} & \frac{\partial \lambda_{N-n}}{\partial \lambda} \end{vmatrix} = \begin{vmatrix} \lambda & p \\ -\lambda & 1-p \end{vmatrix} = \lambda \tag{16}$$

giving

$$\text{prob}(p, \lambda | n, N) = \frac{1}{\Gamma(n+1)\Gamma(N-n+1)} (p\lambda)^n [(1-p)\lambda]^{N-n} \exp(-\lambda) \lambda \, d\lambda dp \tag{17}$$

The parameter λ is a nuisance variable, since it merely parameterizes the total number of counts in the column, an artifact of the experiment. We therefore integrate it out, giving

$$\text{prob}(p|n, N) = \int \text{prob}(p, \lambda | n, N) d\lambda = [B(n+1, N-n+1)]^{-1} p^n (1-p)^{N-n} dp \tag{18}$$

One can readily verify that equation (18) is properly normalized, i.e.,

$$\int \text{prob}(p|n, N) dp = 1 \tag{19}$$

Equation (18) would also follow from assuming a multinomial process in each column, with a uniform prior $\text{prob}(p) = 1$, instead of our Poisson assumption. Indeed, with either assumption, one can derive that the joint probability of any number of cells in a single column is Dirichlet-distributed,

$$\text{prob}(p_1, p_2, \ldots | n_1, n_2, \ldots) \propto p_1^{n_1} p_2^{n_2} \times \cdots \times (1 - p_1 - p_2 - \cdots)^{N - n_1 - n_2 - \cdots} dp_1 dp_2 \cdots \tag{20}$$

6 Equality of a Probability to the Control Group

We now generalize from the table (13) to

	C_0	$\cdots$	C_j	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
f_i	m	$\cdots$	n	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
totals	M	$\cdots$	N	$\cdots$

(21)

where the counts m and marginal M are for the control group. We want to compare the two hypotheses H_A, that the probability implied by m and M is

different from that implied by n and N, and H_B, that the two probabilities are equal. The general scheme is that of equation (**??**), with the delta-function constrant $p_0 = p_1$.

$$\frac{\text{prob}(H_A|D)}{\text{prob}(H_B|D)} = \frac{[B(m+1, M-m+1)]^{-1} \int p_0^m (1-p_0)^{M-m} dp_0 [B(n+1, N-n+1)]^{-1} \int p_1^n (1-p_1)^{N-n} d}{[B(m+1, M-m+1)]^{-1} [B(n+1, N-n+1)]^{-1} \int p^{m+n} (1-p)^{M+N-m-n} dp}$$
$$= \frac{B(m+n+1, M+N-m-n+1}{B(m+1, M-m+1) B(n+1, N-n+1)} \tag{22}$$

References

[1] Agresti, A. (1992) "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, vol. 7, pp. 131-177.

[2] Mehta, C.R. and Senchaudhuri, P. (2003) "Conditional versus Unconditional Exact Tests for Comparing Two Binomials", at www.cytel.com/Papers/twobinomials.pdf

[3] Sivia, D.S. (1996) *Data Analysis: A Bayesian Tutorial* (Oxford: Clarendon Press).

7 Acknowledgments

Thanks to