"What Is Better Than Chi-Square?" and Related Koans

William H. Press

September 23, 2005

Introduction

A koan is an enigmatic, and often senseless, question posed as an aid to meditation in Zen Buddhism. By "chi-square", we mean Pearson's general scheme of a p-value, or tail, statistic that is constructed as a sum of independent random variables, each of which is the square of a normal deviate $N(0,1)$ (or at least approximately so). Such a statistic is distributed according to the familiar χ^2_I distribution, with I degress of freedom, whose tail values are tabulated or readily computable.

Chi-square tests are used to distinguish between two samples, under the null hypothesis that they are drawn from the same distribution. Generally, each sample has a number of subsamples or "bins" ($i = 1, \ldots, I$) that are more or less independent, and the chi-square statistic is a sum over the bins, with one squared normal deviate ($\sim \chi^2_1$) obtained from each bin. If there are a small number of linear constraints among the values in the bins, so that they are not independent, then it is well known [6] that the I in χ^2_I is reduced by the number of constraints, with the general scheme otherwise unchanged.

In this note, we are interested in the particular regime where the number of bins I is very large, and where the data in the bins are integer numbers of counts, m_i for the first sample and n_i for the second. Further, our interest is when the total numbers of counts is much larger than the number in any single bin, i.e.,

$$M \equiv \sum_i m_i \gg m_j, \qquad N \equiv \sum_i n_i \gg n_k \qquad \forall j,k \tag{1}$$

We have no restriction on whether the m_i's and n_i's are small or large, or whether they are the same order of magnitude as i varies. This regime of interest occurs in many bioformatics applications, where the counts may be (e.g.) the numbers of occurrences of every possible nucleotide subsequence of a given length (hence large I) in a large corpus of genomic sequences (hence large M and N). However, some subsequences may be rare in the corpus (hence no restriction on m_i and n_i).

We define r, $0 < r < 1$, such that

$$M = r(M+N), \qquad N = (1-r)(M+N) \tag{2}$$

so that the sample sizes M and N are in the proportions $r : 1 - r$. We are interested both in the case where r is $\lesssim 1$ and in the case where $r \ll 1$. In the former case, we may be comparing two corpuses. In the latter case, we may be comparing a single gene to a much larger corpus of genes.

In the regime described, we can state the null hypothesis regarding m_i and n_i precisely. From a single distribution, we, in effect, drew $M + N$ counts, and $m_i + n_i$ were found to be in bin i. Since M and N are large, it is irrelevant whether we view the value of r as exact or as a statistical estimator (i.e., whether M and N are fixed by the experimental design or are random variables). In either case r is effectively known. Thus the null hypothesis, independently for each i, is that

$$m_i \sim \text{Binomial}(m_i + n_i, r) \tag{3}$$

Hereafter, we denote the (frequently occuring) binomial probability distribution by

$$\text{bin}(m|t, r) = \binom{t}{m} r^m (1 - r)^{t-m} \tag{4}$$

Binomial-Derived Chi-Squares Are Not Exact

It is well known (e.g., [4]) that the conventional formula given for the two-sample chi-square statistic is not exact in the limit of small numbers of counts. In brief, dropping for now the index i, and defining $t \equiv m + n$, we form a statistic x from the difference between what is observed and its expectation,

$$x \equiv (1 - r)m + rn = m - rt \tag{5}$$

The relevant moments of x are

$$\begin{aligned} \langle x \rangle &= \sum_m (m - rt)\, \text{bin}(m|t, r) = 0 \\ \langle x^2 \rangle &= \sum_m (m - rt)^2\, \text{bin}(m|t, r) = r(1 - r)t \\ \langle x^4 \rangle &= \sum_m (m - rt)^4\, \text{bin}(m|t, r) = r(1 - r)t[3r(1 - r)(t - 2) + 1] \end{aligned} \tag{6}$$

Since $E(\chi^2_I) = I$, the chi-square contribution for one bin is obtained by squaring x and normalizing it appropriately,

$$\widehat{\chi^2} \equiv \frac{x^2}{r(1 - r)t} \tag{7}$$

Equation (7) is the formula conventionally given (e.g., [10]) for the two-sample binned chi-square test with unequal sample sizes, and is originally due to Pearson [8].

If x is normally distributed, then equation (7) is exactly $\sim \chi^2_1$, as desired. The premise becomes true in the limit of m and n both large, in which case

$$\text{Binomial}(t, r) \to N(rt, \sqrt{r(1 - r)t}) \tag{8}$$

In our regime of interest I is large, so χ^2_I is also approximately normal,

$$\chi^2_I \to N(I, \sqrt{2I}) \tag{9}$$

Therefore, even if m and n are not large, we can obtain a chi-square distributed statistic via the central limit theorem, if only $\widehat{\chi^2}$ has the desired expectation value (= 1) and variance (= 2). Does it? The expectation value is correct by the construction of equation (7). But the variance, from equation (6), is

$$\mathrm{Var}(\widehat{\chi^2}) = \frac{2r(1-r)(t-3)+1}{r(1-r)t} \tag{10}$$

which becomes 2 only in the limit of large rt and $(1-r)t$. Thus, when some m_i's or n_i's are small, the sum of the individual $\widehat{\chi^2}$'s is not χ^2_I distributed, even in the limit of large I, because the expectation value and variance are discrepant with respect to one another.

It is not hard to construct corrected chi-square statistics, for example by an affine scaling of $\widehat{\chi^2}$ (e.g., Lucy's Y^2 and Z^2 statistics [5]), possibly allowing also a correction to the number of degrees of freedom I, that restores exact agreement with χ^2_I, at least in the normal limit of equation (9). These are however *post hoc* fixes, and are not principled ways of dealing with the discrete binomial distribution of m_i and n_i when either is small.

A more straightforward (but not much different) approach might be simply to sum $\widehat{\chi^2}$, and also equation (10) over bins when the data are analyzed, giving values (say) χ^2 and V. One would then obtain p-values from the normal distribution $N(\chi^2, \sqrt{V})$. Below, we will refer to this as a "variance-by-hand" method.

Other proposed fixes, such as the likelihood ratio test [1], modified Neyman χ^2 [2], and chi-square-gamma statistic [7], seek not to restore an exact χ^2_I distribution, but to mitigate the effect of small number bins in other ways. These must also be viewed as *ad hoc* to varying degrees.

Chi-Square Is Not Optimal When Only a Few Bins Are Causally Different

There are deficiencies in chi-square that are unrelated to the issues of small number counts. One such is the power of the method to detect differences between two populations that may be causal in only a small number of bins.

As an example, consider the case where the two samples are first drawn from the same distribution, but are then perturbed in just a fraction f of the I bins by a change δr in the probability r_0. For simplicity, take $M = N$ so that $r_0 = 1/2$ before the perturbation, and suppose the number of counts in every bin is initially about n_0.

Can the chi-square test detect this perturbation? In order of magnitude, the change in chi-square is

$$\delta\chi^2 \sim (fI)\frac{(n_0\delta r)^2}{n_0} \tag{11}$$

which is detectable if it is greater than a few $\times\sqrt{I}$, implying

$$\delta r \gtrsim \frac{1}{(fn_0)^{1/2} I^{1/4}} \tag{12}$$

We see that the detectability gets better (δr gets smaller) as f increases, as it intuitively should.

A useful comparison, however, is with the "best bin" strategy of looking for the single most discrepant bin, and then applying a multiple hypothesis correction to the resulting p-value. In the normal limit, the maximum t-value seen will be

$$T \sim \frac{\delta r n_0}{\sqrt{n_0}} = \delta r n_0 \tag{13}$$

We now equate the implied tail probability (in asymptotic approximation for the normal distribution) to α/I, where α is the desired significance level. Up to logarithmic corrections, this gives the order of magnitude result,

$$\delta r \gtrsim \frac{1}{\sqrt{n_0}} \sqrt{\ln\left(\frac{I}{\alpha}\right)} \tag{14}$$

Comparing equations (12) and (14), we see that best-bin beats chi-square when $f \lesssim I^{-1/2}$. In the extreme case of $f = 1/I$, best-bin can detect a signal that is a factor $O(I^{1/4})$ smaller in counts. While the fourth-root dependence is modest, the effect can be devastating when, as we contemplate for bioinformatic applications, I is as large as $\sim 10^9$.

Bayes Factor Log Odds Approach

One might hope that an approach based on Bayes factors and using exact binomial probabilities rather than normal approximations might fix one or both of the above highlighted problems with chi-square.

Focussing on one bin, let H_0 be the hypothesis that the distributions are the same, that is, in the expected ratio of $r : (1-r)$; and let H_1 be the hypothesis that they are different, that is, have some other ratio $s : (1-s)$. The odds ratio is the ratio of the data likelihoods, integrating over the unknown value of s, and with appropriate priors.

$$\frac{P(H_0|m,n)}{P(H_1|m,n)} = \frac{P(H_0)}{P(H_1)} \frac{\mathrm{bin}(m|t,r)}{\int \mathrm{bin}(m|t,s)p(s)ds} \tag{15}$$

Taking the simplest noninformative prior $p(s) = 1$, and taking the logarithm, gives the result

$$\ln \mathcal{L}_i = \sum_i \ln(m_i + n_i + 1) + \ln \mathrm{bin}(m_i|m_i + n_i, r) + W_i \tag{16}$$

where the W_i's parametrize the priors $P(H_0)$ and $P(H_1)$ as any convenient set of values that sum to $\ln P(H_0) - \ln P(H_1)$.

It is not hard to see that, in the limit of large m_i and n_i, $-2\ln\mathcal{L}_i$ is basically equivalent to chi-square. In particular, with the notation of equation (7),

$$-2\ln\mathcal{L} = \sum_i \widehat{\chi^2} + \{\ln[2\pi r(1-r)] - \ln(m_i+n_i) - 2W_i\} \tag{17}$$

If we choose each prior W_i so as to make the term in braces equal to $-T\sqrt{2/I}$, then the equivalence to chi-square is exact, with the log-odds decision point of zero corresponding to a chi-square decision point with a t-value of T. This data-length dependendent prior may seem peculiar, but it is necessary if one wants Bayesian results that can equally well be interpreted as frequentist p-value (tail) tests, a desirable feature. The need for such priors is related to "Lindley's Paradox" [12] and has been discussed at greater length elsewhere [11]. Henceforth we refer to such priors as "p-value priors".

What we see is that the above Bayes log-odds method has exactly the same issue as chi-square regarding the detectability of a signal that is confined to a small number of bins. For the issue regarding small number counts, the Bayes log-odds method does avoid inappropriate use of equation (7) by using exact binomial probabilities. But the penalty is that, in order to choose p-value priors W_i that correspond to a specified p-value, one must do both "variance-by-hand" and "expectation-by-hand" calculations on equation (16). (Below, we will see that these calculations are in fact not computationally difficult.)

Bayes Factor with a Prior on the Probability of Causal Differences

Bayesian methods have a tendency to answer the question you asked, not the question that you *meant* to ask. The problem with the approach in the previous section is that the alternative hypothesis H_1 gave *all* bins their own, different, values of s. If we frame our alternative hypotheses with greater care, we can get a better answer to the question that we meant to ask.

Consider now the large number of alternative hypotheses $H_{\mathbf{f},\mathbf{v},\mathbf{s}}$ indexed by the vector quantities $\mathbf{f}$, $\mathbf{v}$, and $\mathbf{s}$, each having I components. The value $f_i \in [0,1]$ is the probability that bin i is causally different in the two samples. The binary value $v_i \in \{0,1\}$ indicates by a 1 value that a component is *actually* different, or by a 0 value that it is not. The component s_i is the value of the s-probability (as in the denominator of equation 15 when $v_i = 1$, or r, when $v_i = 0$.

We need a mixture prior on $\mathbf{f}$ that gives finite weight to the hypothesis that $\mathbf{f} = 0$ (as a vector), meaning that no bins are causally different. We can write this as

$$P(f) = w\prod_i \delta(f_i) + (1-w)\prod_i p(f_i) \tag{18}$$

where $p(f_i)$ is now the "reduced" prior on $0 < f_i \leq 1$. We will see that a sensible choice for $p(f)$ is important. As for priors on all the s_i's (when $v_i = 1$), we will take these to be uniform in $[0,1]$.

This setup, using a vector of f_i's instead of a single common value f, is the "variant" method (due to Kochanek) described in §3.4 of [9]. In that paper, the justification for using a common value of f was that the individual vector components represented experiments drawn from (at least notionally) a common standard of practice. Thus, evidence that one experiment was "bad" was relevant evidence about that common standard. For the bioinformatics applications we have in mind here, on the other hand, a difference of counts for one causal feature says nothing about whether other causal features ought to exist.

We now calculate,

$$\begin{aligned} P(H_{\mathbf{f},\mathbf{v},\mathbf{s}} \,|\, \mathbf{m},\mathbf{n}) &\propto P(\mathbf{m},\mathbf{n} \,|\, \mathbf{f},\mathbf{v},\mathbf{s}) P(\mathbf{f}) P(\mathbf{v}|\mathbf{f}) P(\mathbf{s}|\mathbf{f},\mathbf{v}) \\ &= \left[\prod_{i:v_i=0} (1-f_i)\mathrm{bin}(m_i|t_i,r)\right] \left[\prod_{i:v_i=1} f_i \,\mathrm{bin}(m_i|t_i,s_i)\right] \\ &\quad \times \left[w\prod_i \delta(f_i) + (1-w)\prod_i p(f_i)\right] \end{aligned} \tag{19}$$

Here we have used the uniform priors on the s_i's, and also that $P(\mathbf{v}|f)$ is the product of a factor f for each $v_i = 1$ and a factor $(1-f)$ for each $v_i = 0$.

We now marginalize (integrate) over the nuisance variables s_i, using

$$\int \mathrm{bin}(m|t,s)ds = (t+1)^{-1} \qquad \text{(independent of } m\text{)} \tag{20}$$

Next, we further marginalize by summing over all 2^I possible $\mathbf{v}$'s, which can be done by inspection as in [9] (cf. [3]). The result is a posterior distribution on $\mathbf{f}$ alone:

$$\begin{aligned} P(\mathbf{f} \,|\, \mathbf{m},\mathbf{n}) \propto \prod_i &\left[(1-f_i)\mathrm{bin}(m_i|t_i,r) + \frac{f_i}{t_i+1}\right] \\ &\times \left[w\prod_i \delta(f_i) + (1-w)\prod_i p(f_i)\right] \end{aligned} \tag{21}$$

Integrating respectively over an infinitesimal region near $\mathbf{f} = 0$ and the rest of $d^I\mathbf{f}$ (each f_i from 0 to 1) gives the odds ratio

$$\frac{P(f=0)}{P(f\neq 0)} = \left(\frac{w}{1-w}\right) \frac{\prod_i \mathrm{bin}(m_i|t_i,r)}{\prod_i \left[(1-\overline{f})\mathrm{bin}(m_i|t_i,r) + \overline{f}/(t_i+1)\right]} \tag{22}$$

where

$$\overline{f} = \int_0^1 p(f_i)\, f_i \, df_i \tag{23}$$

which is independent of i because we will take the same prior for all the f_i's.

How shall we choose $p(f)$ and hence $\overline{f}$? A natural choice is that which gives equal prior probability to every logarithmic range of f from $1/I$ (signal in a

single bin) to 1 (signal in all bins, as assumed by chi square tests). This gives

$$\overline{f} = \frac{1}{\ln I}\left(1 - \frac{1}{I}\right) \approx \frac{1}{\ln I} \tag{24}$$

Equation (22) implies the log-odds formula

$$\begin{aligned} \ln \mathcal{L} = \sum_i \{ & \ln \mathrm{bin}(m_i | m_i + n_i, r) \\ & - \ln \left[(1 - \overline{f}) \mathrm{bin}(m_i | m_i + n_i, r) + \frac{\overline{f}}{m_i + n_i + 1} \right] \Big\} + W \end{aligned} \tag{25}$$

where W's will be chosen to be a p-value prior (see discussion following equation 16). (W is just a reparametrization of w.) Equation (25) is the main analytic result of this paper. Note that it is equivalent to equation (16) when $\overline{f} = 1$.

It is informative to see how equation (25) works qualitatively when $\overline{f}$ is small: If a bin has

$$\mathrm{bin}(m_i | m_i + n_i, r) \gg \frac{\overline{f}}{m_i + n_i + 1} \tag{26}$$

Then a small positive amount, about $\overline{f}$, is added to the log odds score (in favor of the null hypothesis). If the inequality is strongly the other way, then a possibly much larger value, $\sim \ln \overline{f} / [(t_i + 1)\mathrm{bin}(m_i | t_i, r)]$ is instead subtracted (in favor of that bin's being causally different). In other words, the upside evidence is limited, since it generally reflects only the expected large fraction of not-different bins, while the downside evidence is (nearly) unlimited, as it should be, since it must be sensitive to a large signal from a small number of bins.

Calculation of the p-value Priors

If we write

$$\begin{aligned} Q_i &\equiv \ln \mathrm{bin}(m_i | t_i, r) - \ln \left[(1 - \overline{f}) \mathrm{bin}(m_i | t_i, r) + \frac{\overline{f}}{t_i + 1} \right] \\ \ln \mathcal{L} &= \sum_i Q_i + W_i \end{aligned} \tag{27}$$

then, to make zero log odds equivalent to a one-sided normal t-value of T, we must have

$$\left\langle \sum_i Q_i \right\rangle + W = T \sqrt{\mathrm{Var}\left(\sum_i Q_i \right)} \tag{28}$$

implying

$$\begin{aligned} W &= -\sum_i \langle Q_i \rangle + T \sqrt{\sum_i \mathrm{Var}(Q_i)} \\ &= -\sum_i G_1 + T \sqrt{\sum_i (G_2 - G_1^2)} \end{aligned} \tag{29}$$

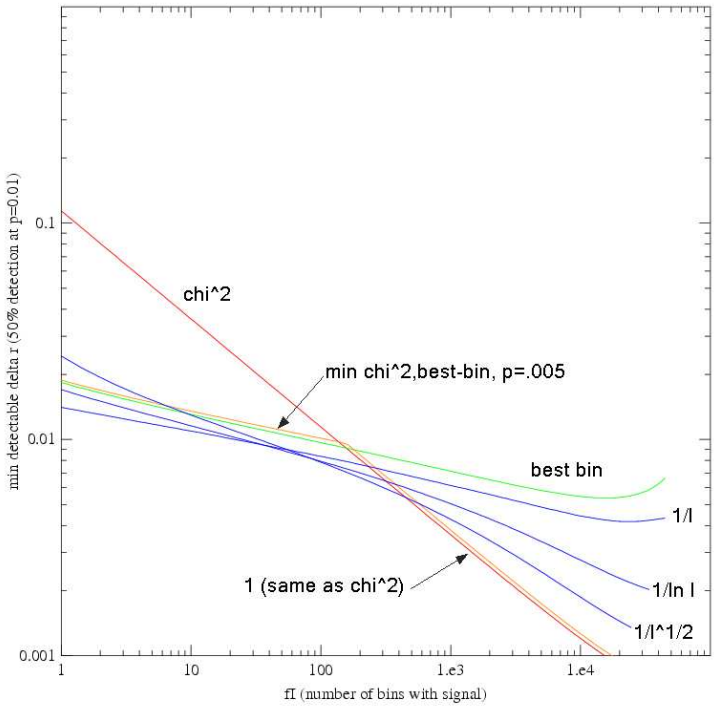
Here $G_{1,2}$ denotes the moments of Q over the binomial distribution,

$$G_{1,2}(r,t) = \sum_{m=0}^{t} \mathrm{bin}(m|t,r)[Q_i(m,t,r)]^{1,2} \tag{30}$$

These are not difficult to calculate numerically: for small m (< 40, say), the direct sum over m can be organized to have only a few updating operations per term in the sum, while for larger m, $\mathrm{bin}(m|t,r)$ can be approximated by a Gaussian, and Gauss-Hermite integration can be then applied. These calculations, yielding W_i, are done for every data set and every value of i.

Bayes Factor Results

The figure shows results for the minimum detectable δr when a fraction f bins are perturbed. Here $I = 10^5$ and $n_0 = 10^4$; for other values the shapes of the curves are nearly identical, except for the scalings implied by equations (12) and (14). The blue curves are labeled by the prior $\overline{f}$. (For $\overline{f} = 1$, the blue curve falls exactly on the χ^2 curve.)



Perhaps disappointingly, there is no single value of $\overline{f}$ that does better than both chi-square and best-bin. Nor does any value beat a combination of chi-

square and best-bin (including the multiple hypothesis correction for two tests) by more than a factor of 0.6 in minimum detectable δr.

A Better Tail Test

As disappointed Bayesians, we turn for solace back to a conventional frequentist approach. We will fix chi-square's deficiencies essentially by brute force.

First, we fix the problem of small m_i or n_i. For each i, we form the quantity

$$u_i \equiv \text{bincdf}(r, m_i + n_i, m_i) + \text{ran}(0,1)\text{bin}(r, m_i + n_i, m_i) \tag{31}$$

Here bincdf is the exact cumulative distribution function of the binomial distribution (which is readily calculable as an incomplete beta function), and ran is a uniform random deviate. The additional randomization is needed to convert the "steps" of the binomial's cdf into an *exactly* uniform p-value (in the case of the null hypothesis).

Second, because in the case of an actual signal both tails will be overpopulated, we "fold" the values u_i so that the two tails are both on the left. This transformation

$$u_i \to 2\text{min}(u_i, 1 - u_i) \tag{32}$$

exactly preserves the uniformity of the u_i's in the null hypothesis.

Third, we sort the u_i's.

Fourth: In the null hypothesis, the value of the (folded and sorted) quantity u_i is beta distributed,

$$u_i \sim \text{Beta}(i, I - i + 1) \tag{33}$$

with expectation and variance

$$\langle u_i \rangle = \frac{i}{I+1}, \qquad \text{Var}(u_i) = \frac{i(I-i+1)}{(I+1)^2(I+2)} \tag{34}$$

For any fixed i, this suggests the use of the t-value statistic

$$s_i \equiv \frac{(i - (I+1)u_i)\sqrt{I+2}}{\sqrt{i(I-i+1)}} \tag{35}$$

For any i, a large positive value of s_i indicates overpopulated tails in the binomial distributions for the i "most extreme" bins.

Fifth (ideal case), we take as our test statistic

$$S_{1,I} \equiv \max_{1 \le i \le I} s_i \tag{36}$$

This fixes by brute force the chi-square deficiency of insensitivity to a signal in only a few bins, since every i's t-value now gets an equal chance. The only problem with this is that its distribution (or quantile points) in the null distribution is hard to calculate. Furthermore, most values of i are irrelevant, since

successive s_i's are highly correlated (correlation distance $\sim i$). We therefore instead turn to...

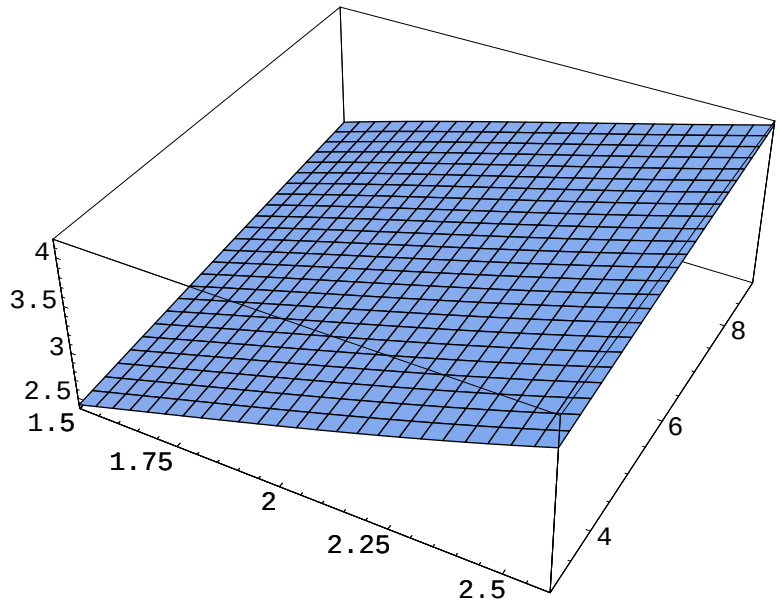
Fifth (actual case),

$$S_{1.05,I} \equiv \max_{i \in \text{seq}} s_i \tag{37}$$

where

$$\text{seq} \equiv 1, 2, 3, \ldots, i, \max(i+1, 1.05i), \ldots \qquad (\text{while } \leq I) \tag{38}$$

The quantile points of $S_{1.05,I}$ are readily computable by simulation, because we can use Beta deviates to jump forward to arbitrary values of i. Thus, each simulation requires only $O(\ln I)$ operations instead of $O(I)$. We have done 10^5 simulations for each of various values of I between 10^3 and 10^9. The resulting critical values $S_{1.05,I,p}$ are adquately fit by a smooth function of $log_{10} I$ and $(-\ln p)^{1/2}$, as shown in the following figure,

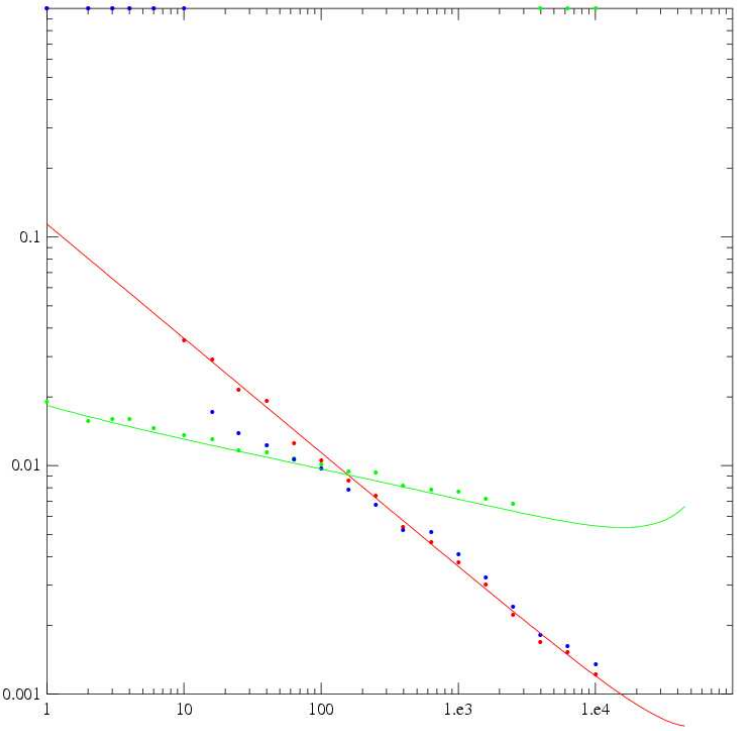


We refer to this test as the "S-max test".

Results for the S-max Test

These are harder simulations, so we'll do them sparsely and allow some scatter. We also compute chi-square and best-bin for the same simulations as a check. The results (same format as figure) are:

Gosh! Not at all what we wanted.



Summary and Advice

- Chi-square is not exact for small m_i and n_i. The best way to fix it is probably to compute, for each i, an exact t-value and then square it. You do this by inverting the binomial to an exact uniform (using the randomization trick above), then taking the inverse cdf of the normal $N(0,1)$. This guarantees *exact* chi-square probabilities for each bin, and is not an excessive amount of computation per bin.

- Chi-square lacks power when the signal is in a small number of bins. Despite the heroic efforts above, no fix for this seems better than the best-bin test. Best-bin works surprisingly well, even when the signal is in as many as $I^{1/2}$ bins. The exact test is: Compare the largest squared t-value obtained as above to the χ^2_1 distribution, and require a p-value of α/I, where α is the desired significance. This takes basically no time at all, and adds just a couple of lines to the chi-square calculation.

- A single test combining both chi-square and best-bin is to take the minimum of the two, requiring of each a significance level $\alpha/2$. (That is, multiply the smaller p-value by 2.)

References

[1] Baker, S. and Cousins, R.D. 1984, *Nucl. Instr. Meth. Phys. Res.* **221**, 437–442.

[2] Bevington, P.R. and Robinson, D.K. 1992 *Data Reduction and Error Analysis for the Physical Sciences*, 2nd ed. (New York: McGraw-Hill).

[3] Box, G.E.P. and Meyer, R.D. 1986, *Technometrics* **28**, 11–18.

[4] Eadie, W.T., Drijard, D., James, F.E., Roos, M., and Sadoulet, B., 1971, *Statistical Methods for Experimental Physics* (Amsterdam: North-Holland), p. 257.

[5] Lucy, L.B. 2000, *Mon. Not. Roy. Astron. Soc.* **318**, 92–100.

[6] Lupton, R. 1993, *Statistics in Theory and Practice* (Princeton: Princeton University Press), §4.2.

[7] Mighell, K.J. 1999, *Astrophys. J.* **518**, 380–393.

[8] Pearson, K. 1911, *Biometrika* **8**, 250–254.

[9] Press, W.H. 1997, in *Unsolved Problems in Astrophysics*, J.N. Bahcall and J.P. Ostriker, eds. (Princeton, NJ: Princeton University Press), 49–60.

[10] Press, W.H. et al. 2002, *Numerical Recipes in C++*, 2nd ed. (Cambridge, UK: Cambridge University Press).

[11] Press, W.H., 2005, "Note on the Significance of 2x2 Contingency Tables and Lindley's Paradox" (unpublished).

[12] Shafer, G. 1982, *J. Am. Stat. Assoc.* **77**, 325–334; also several commentaries following, *op. cit.*