CBX 98-73
Harry Nelson
December 4, 1998

Goodness of Fit
for
Poisson Distributed Data

see also: `http://hep.ucsb.edu/people/hnn/wrongd/goodness/`

Abstract

When a fit to binned data is performed, and the mean number of events in a typical bin is about 10 or smaller, the evaluation of a goodness of the fit becomes problematic. We discuss aspects of obtaining a useful goodness of fit test for this case.

When a fit to binned data is performed, and the mean number of events in a typical bin is about 10 or smaller, the evaluation of a goodness of the fit becomes problematic. The probability distribution function (*pdf*) that describes such data is usually a Poisson distribution. We discuss aspects of obtaining a useful goodness of fit test, when one assumes a Poisson distribution, and performs a certain type of maximum likelihood fit to the data. The two limits of plentiful data, where a χ^2 fit can be usefully employed, and that of unbinned data, are discussed and recovered.

The Poisson distribution for r events given a mean of μ is (see 1998 RPP[1] Table 28.1 p. 170):

$$f(r;\mu) = \frac{\mu^r e^{-\mu}}{r!}$$

The normalization is $\sum_{r=0}^{\infty} f(r;\mu) = 1$, for all values of μ; this normalization is exploited in what follows.

When fitting models to binned data, where a bin contains, generically, r events, one often selects among models by using them to predict μ for each bin, and choosing that model which maximizes the sum over all bins of the logarithms of the Poisson distribution, which we generically call here the 'Poisson likelihood technique'[2]:

$$\begin{aligned} \ln f(r;\mu) &= r\ln\mu - \mu - \ln r! \\ &= r\ln\mu - \mu - \ln\Gamma(r+1). \end{aligned} \tag{1}$$

Usually, a given model will contain a number of parameters that can be adjusted to smoothly vary μ for each bin. For concreteness, denote by m the number of such parameters. For that model, one can then make an estimation of the m parameters by varying them so as to maximize the sum over all bins of Eq. 1; we will denote this sum of Eq. 1 over all bins by ln L.

Concerning a practical matter: one part of Eq. 1 can be tedious to compute, $r!$. It is worth recalling that the Γ function, defined by $\Gamma(x) = \int_0^\infty dy e^{-y} y^{x-1}$, is related to $r!$ through $r! = \Gamma(r+1)$. Further, the function $\ln\Gamma(x)$ is well-conditioned, that is, it does not become so large or small that it causes underflows or overflows in `FORTRAN`, and, $\ln\Gamma(x)$ is available as `ALGAMA` and related functions, under `C304` in the CERN `MATHLIB`; in `PAW`, this function is available in `SIGMA` as `ALOGAM`.

The sum ln L, regarded as a function of the m parameters of the model, has at least two broad uses: 1) it can be used to make errors, or confidence intervals, on the m estimated parameters; 2) the particular value of the sum, evaluated with all m parameters set equal to their best fit values, that is, where the m parameters cause the maximum ln L, can be used to estimate the goodness of fit of the model, and of the particular choice of the model's m parameters, to the data.

For both uses, it is usual to change from maximization of ln L to the equivalent procedure of minimization of the sum over all bins of $-2\ln f(r;\mu)$; we will denote this sum over all bins by $-2\ln$ L. The change can be remembered, mnemonically, by recalling that for cases with large statistics, where one might employ the χ^2 technique, the likelihood is proportional to $e^{-\chi^2/2}$, and so $-2\ln e^{-\chi^2/2} = \chi^2$. The sum $-2\ln$ L behaves in some ways like the χ^2. In particular, the error on each of the m parameters is usually estimated by the variation in that parameter needed to increment $-2\ln$ L from its value at best fit by one unit. Indeed, the whole covariance matrix of the m parameters can be estimated in this manner. For the first use mentioned above, for the estimation of errors, the interpretation increments in $-2\ln$ L proceeds just like that of increments in χ^2, for a χ^2 fit.

The interpretation of the sum at best fit of $-2\ln$ L for goodness of fit is a little more complicated than is the interpretation of χ^2 in a χ^2 fit. There are at least two reasons why: first, the mean value of $-2\ln f(r;\mu)$, weighted by the Poisson distribution, which we denote $\langle -2\ln f(r;\mu)\rangle$, is not equal to 1 for each bin, and for all μ, in distinction to the case for a χ^2 fit. So, when the Poisson likelihood technique is used, the mean of the sum $-2\ln$ L will not be equal to the number of degrees of freedom, which is the number of bins N, minus the number of fit parameters, m, or $N-m$, as is the case for a χ^2 fit.

It is straightforward to numerically evaluate, in `PAW` or a `FORTRAN` routine, the exact summation:

$$
\begin{aligned}
\langle -2\ln f(r;\mu)\rangle &= -2\sum_{r=0}^{\infty} f(r;\mu)\ln f(r;\mu) \\
&\rightarrow 2\mu(1-\ln\mu) \qquad\qquad\qquad\qquad \text{as } \mu\rightarrow 0 \\
&\rightarrow 1 + 2[\mu(1-\ln\mu) + \ln\Gamma(\mu+1)] \quad \text{as } \mu\rightarrow\infty
\end{aligned} \tag{2}
$$

The exact summation, as well as the two limiting cases, are shown in Fig. 1. The `PAW kumac` used to generate the curves in Fig. 1 is given in the Appendix. The limit as $\mu\rightarrow 0$ appears in the unbinned goodness of fit test described starting in Eq. 27 of CBX 98-61.[3] The limit as $\mu\rightarrow\infty$ is one unit greater than $-2\ln f(\mu;\mu)$, and thus, were $-2\ln f(\mu;\mu)$ to be subtracted off, would converge to one unit, per bin, similar to χ^2.

In a fit based on the Poisson log-likelihood, it is then straightforward to estimate the mean of the log-likelihood, per bin, at best fit from Eq. 2 and Fig. 1. Of course, one must sum up the contribution of each bin. In doing so, we do not yet incorporate the tendency of the estimation of the m parameters to 'pull' the sum $-2\ln$ L to a lower value. As long as the number of bins and events is large compared to m, neglect of the pulls should be OK; to lowest order, the effect of the pulls on the mean will be to reduce the sum $-2\ln$ L by m units, as long as sensitivity to the m parameters is reasonably spread out among the bins. A technique to treat the pulls more accurately is discussed at the end of this note; it consumes more computer time than the evaluation of the sum in Eq. 2.

Another approach toward treating goodness of fit is, for each bin, to subtract off $-2\ln f(\mu;\mu)$, which comes from the simple estimate of the most probable r, as being equal to μ.[4] Indeed, the mean value of $-2[\ln f(r;\mu) - \ln f(\mu;\mu)]$ does converge to 1 as $\mu\rightarrow\infty$, as shown in Fig. 2, where the curve is labeled 'Exact'. Also shown is the mean value of certain approximations[1, 6, 7] which give:

$$-2[\ln f(r;\mu) - \ln f(\mu;\mu)] \approx 2(\mu - r) + 2r\ln(r/\mu), \tag{3}$$

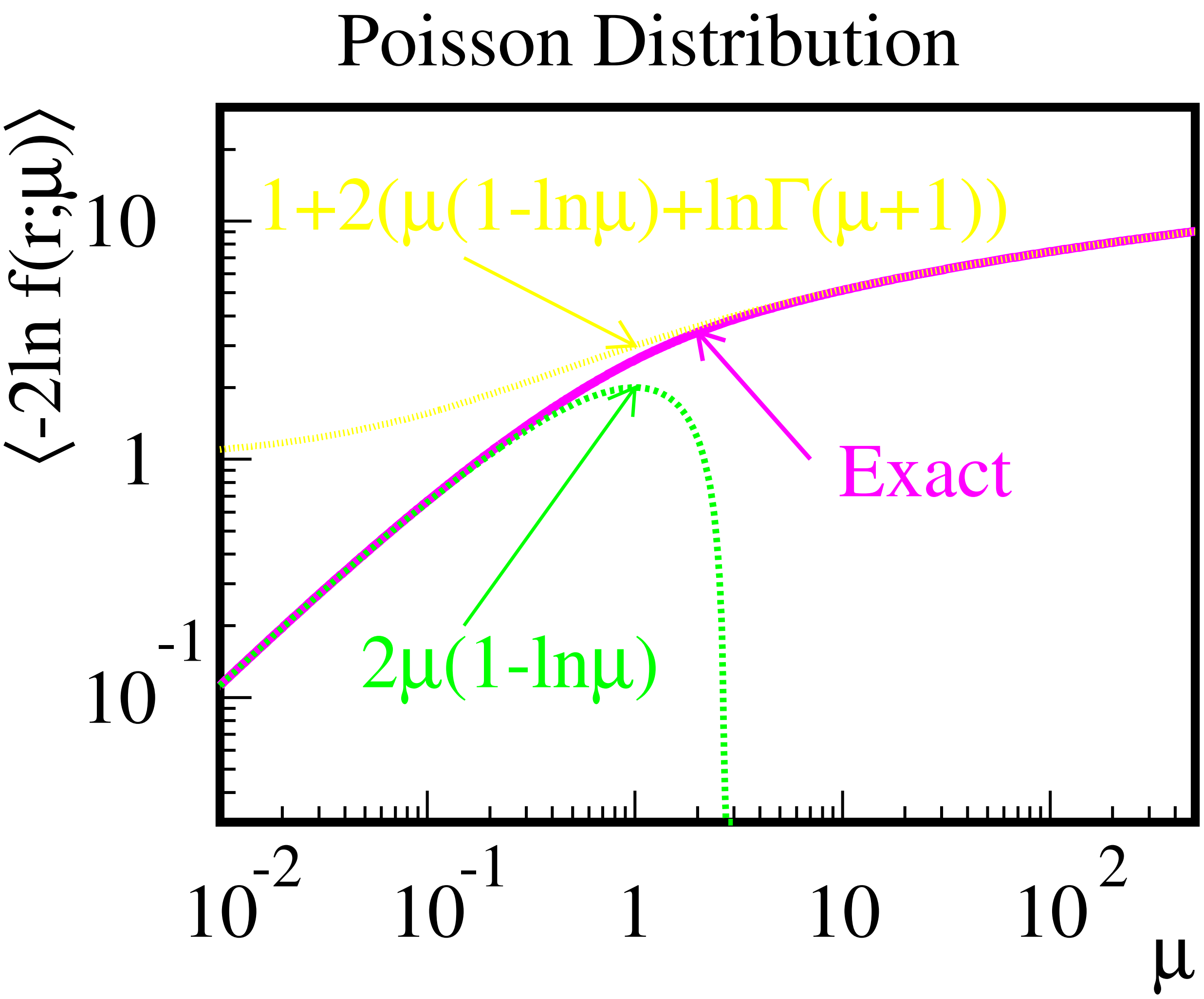

Figure 1: The mean value of the logarithm of the Poisson likelihood as a function of the mean number of entries, μ; described by Eq. 2.

as well as the mean value of the ("Pearson's") 'χ^2' technique for estimation of $-2[\ln f(r;\mu) - \ln f(\mu;\mu)]$,

$$-2[\ln f(r;\mu) - \ln f(\mu;\mu)] \approx \chi^2 = \frac{(r-\mu)^2}{\mu}, \tag{4}$$

which, by design, is 1, independent of μ, since the variance of the Poisson distribution is equal to μ.[1]

However, when μ becomes less than about 10, the per-bin mean of $-2[\ln f(r;\mu) - \ln f(\mu;\mu)]$ diverges from one for two of the three approximations, as one can see from Fig. 2. The 'Exact' calculation tends to always decrease from one, and the approximation from Ref. [1, 6] goes first above one, and then crosses over to below one at $\mu \approx 0.5$. The excursion of this approximation above one might be the explanation for the low confidence levels observed in both the $B \to D^*\pi$[8], and early fits to $D^0 \to K^+\pi^-$.

The χ^2 technique has a per-bin mean that stays at one, by definition, but has yet a different odd feature, to be described presently.

In any case, one must adjust the expectation for the mean of the sum $-2\ln \mathrm{L}$, which we will denote $\langle -2\ln \mathrm{L}\rangle$, to account for the tendencies described in Fig. 1 and Fig. 2, depending on what technique one uses.

A possible reason *not* to use any technique that involves the subtraction of $-2\ln f(\mu;\mu)$: this change qualitatively corresponds to a change in the 'prior density function' used to describe the distribution of the data in each bin, which may or may not influence the fair estimation of the m parameters, depending on one's philosophy, and on the strength of one's belief that the Poisson distribution describes the binned data.

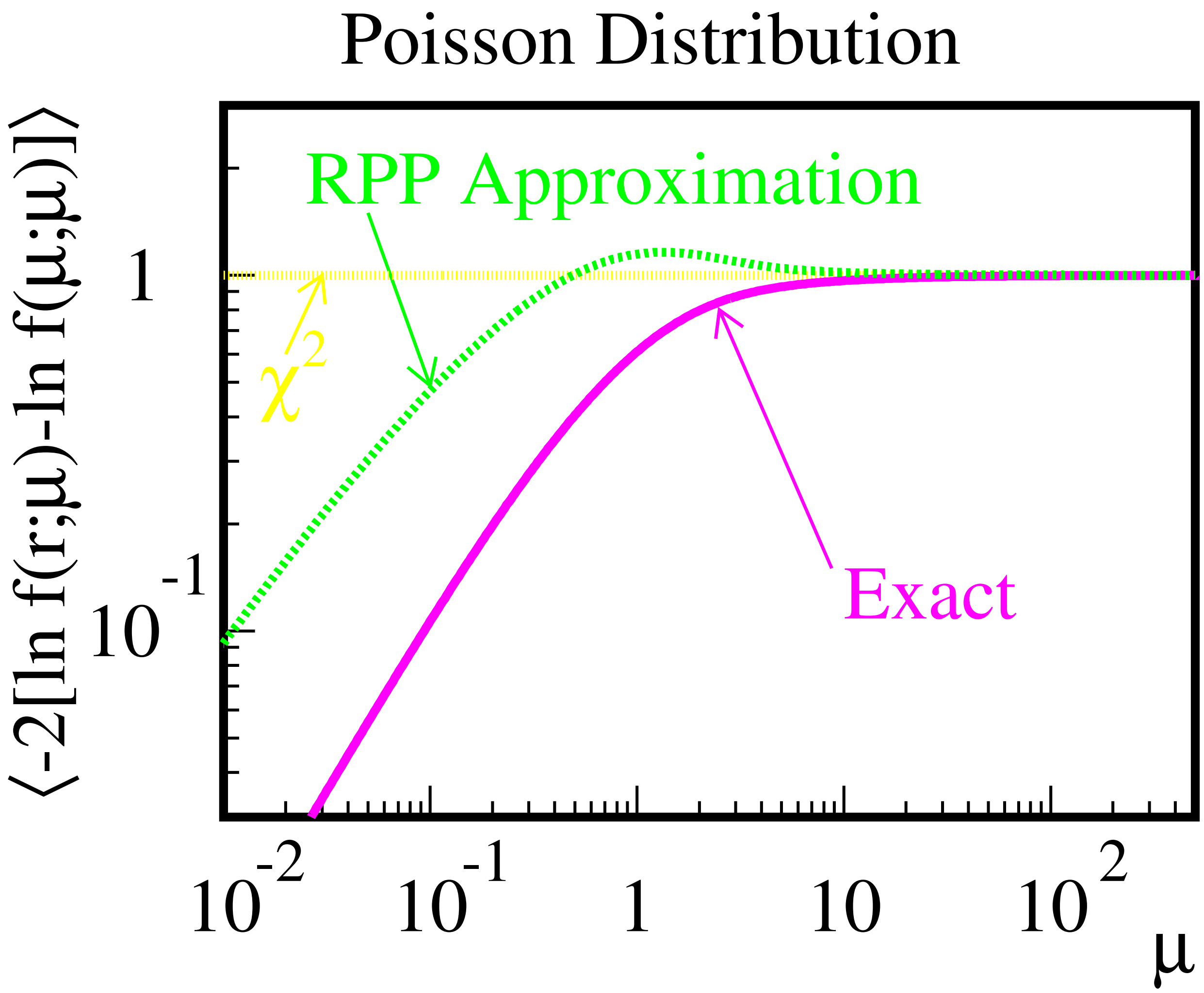

Figure 2: The mean value of the logarithm of the Poisson likelihood, relative to the most probable value, in three approximations: use of Eq. 1 for $\ln f(\mu;\mu)$, and Eq. 2, labeled 'Exact'; use of the expression Eq. 3, labeled 'RPP Approximation'; and use of χ^2, as given in Eq. 4.

The second reason that the interpretation of the sum $-2\ln \mathrm{L}$ at best fit for goodness of fit is a little more complicated than in the case of the χ^2 fit is that the variance of $-2\ln f(r;\mu)$, weighted by the Poisson distribution, is not equal to 2 for each bin, as is the case for the χ^2 fit.[1]

For a χ^2 fit then, where the mean of χ^2 will be $N{-}m$, the variance will be $2(N{-}m)$; the subtraction of m originates in the 'pull' of the estimation of the m parameters. When the *total* number of events in the fit is large (in distinction to the number of events *per bin*), the probability distribution function (*pdf*) of χ^2 can be reasonably approximated by invocation of the central limit theorem, which would describe the *pdf* of χ^2 with a gaussian distribution with mean of $N-m$, and variance $2(N-m)$. In treating the goodness of fit for the Poisson likelihood technique, we will retain use of the central limit theorem, which will tend to be helpful when the *total* number of events in the data is large, but we will recommend use of the appropriate per bin means, given by Eq. 2, in replacement of the χ^2 result of one per bin, and the appropriate per bin variances, in replacement of the χ^2 result of two per bin; we discuss the per bin variances for Poisson distributed data presently.

The sum $-2\ln \mathrm{L}$, for a given fit, and prior to consideration of the 'pulls' of the parameters to be estimated by the fit, should have a *mean* of the sum over bins of $\langle -2\ln f(r;\mu)\rangle$. The *pdf* of the of the best fit sum $-2\ln \mathrm{L}$, by the central limit theorem, will be gaussian, with mean given by the sum over bins of $\langle -2\ln f(r;\mu)\rangle$, and a variance given by the sum over bins of $\langle[-2\ln f(r;\mu) - \langle -2\ln f(r;\mu)\rangle]^2\rangle$ $=\langle[-2\ln f(r;\mu)]^2\rangle - \langle -2\ln f(r;\mu)\rangle^2$. The effect of pulls will be, to lowest order, a reduction of the mean sum $\langle -2\ln \mathrm{L}\rangle$ by m units, and a reduction of the variance of $-2\ln \mathrm{L}$ by $2m$ units.

The variance from one bin is given by:

$$\begin{aligned}\langle[-2\ln f(r;\mu)]^2\rangle - \langle -2\ln f(r;\mu)\rangle^2 &= \sum_{r=0}^{\infty} f(r;\mu)[-2\ln f(r;\mu) - \langle -2\ln f(r;\mu)\rangle]^2 \\ &\rightarrow 4\mu\ln^2\mu \quad \text{as} \;\; \mu\rightarrow 0 \\ &\rightarrow 2 \qquad\quad\;\; \text{as} \;\; \mu\rightarrow\infty \end{aligned} \tag{5}$$

The exact summation, as well as the two limiting cases, are shown in Fig. 3. The `PAW kumac` used to generate the curves in Fig. 3 is given in the Appendix. The limit as $\mu \rightarrow 0$ is equivalent to the unbinned goodness of fit test described in CBX 98-61[3]. The limit of the per bin variance as $\mu\rightarrow\infty$ is 2, which is the variance (per degree of freedom) of the χ^2 distribution; see again Table 28.1 of the 1998 RPP, page 170[1].

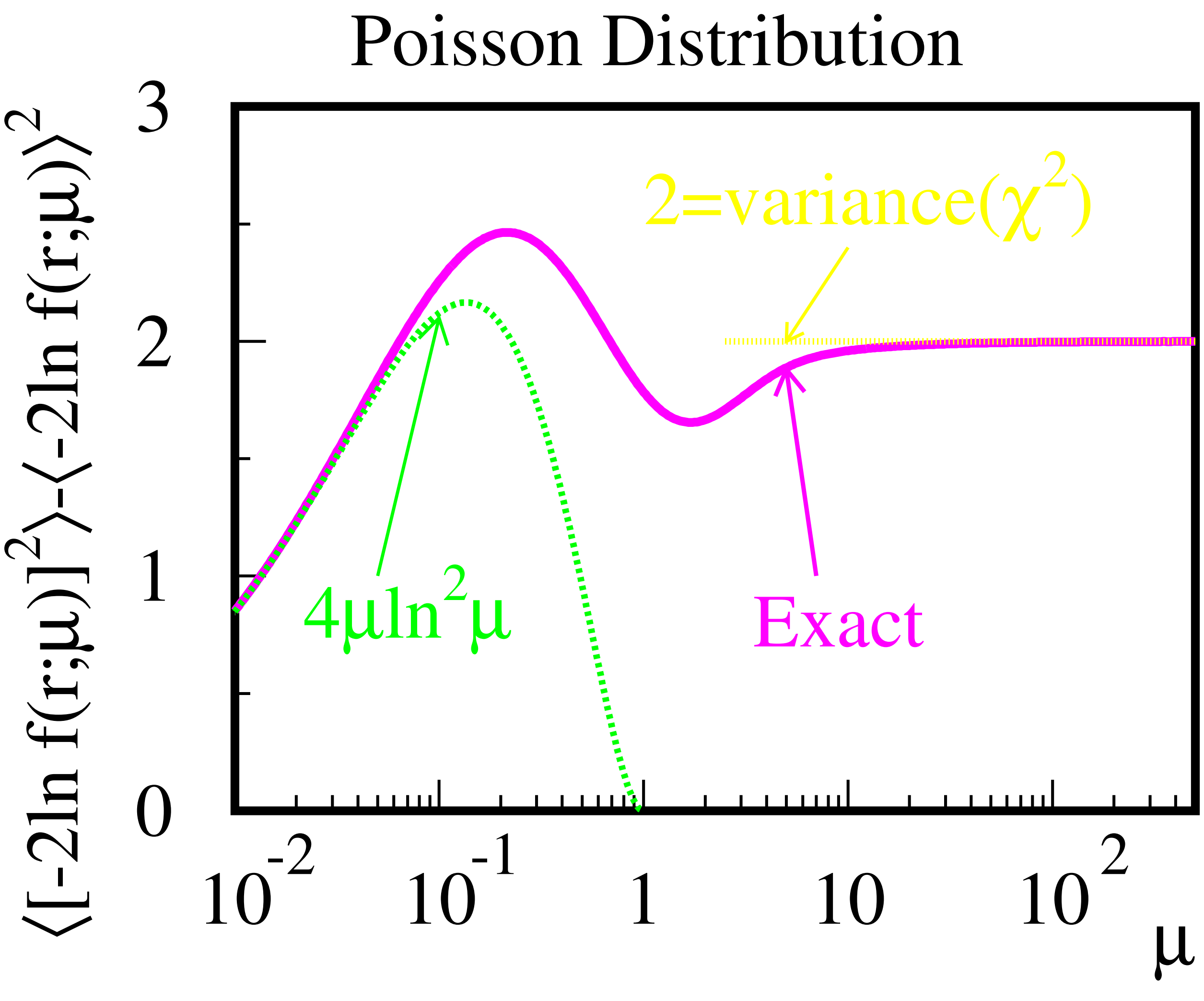

Figure 3: The variance of the logarithm of the Poisson likelihood as a function of the mean number of entries, μ; described by Eq. 5.

We have also evaluated the performance of the approximations described by Eqs. 3 and 4 for the per-bin variance; these are shown in Fig. 4. If one uses those approximations, then the variances depicted in Fig. 4 should be summed over bins to obtain the variance of the sum $-2\ln$ L. Note, that here, the χ^2 technique is qualitatively distinct from the other two approximations, and has the odd feature that the variance grows as $\mu \to 0$.

Qualitatively, Fig. 4 is the most interesting plot in this note. From it, one can deduce the influence of the various bins in driving the minimization, for the three techniques that appear in the plot. The χ^2 technique affords bins with small μ a relatively high influence, and the approximation from Eq. 3 affords bins with small μ a relatively low influence, when compared with the Poisson likelihood technique.

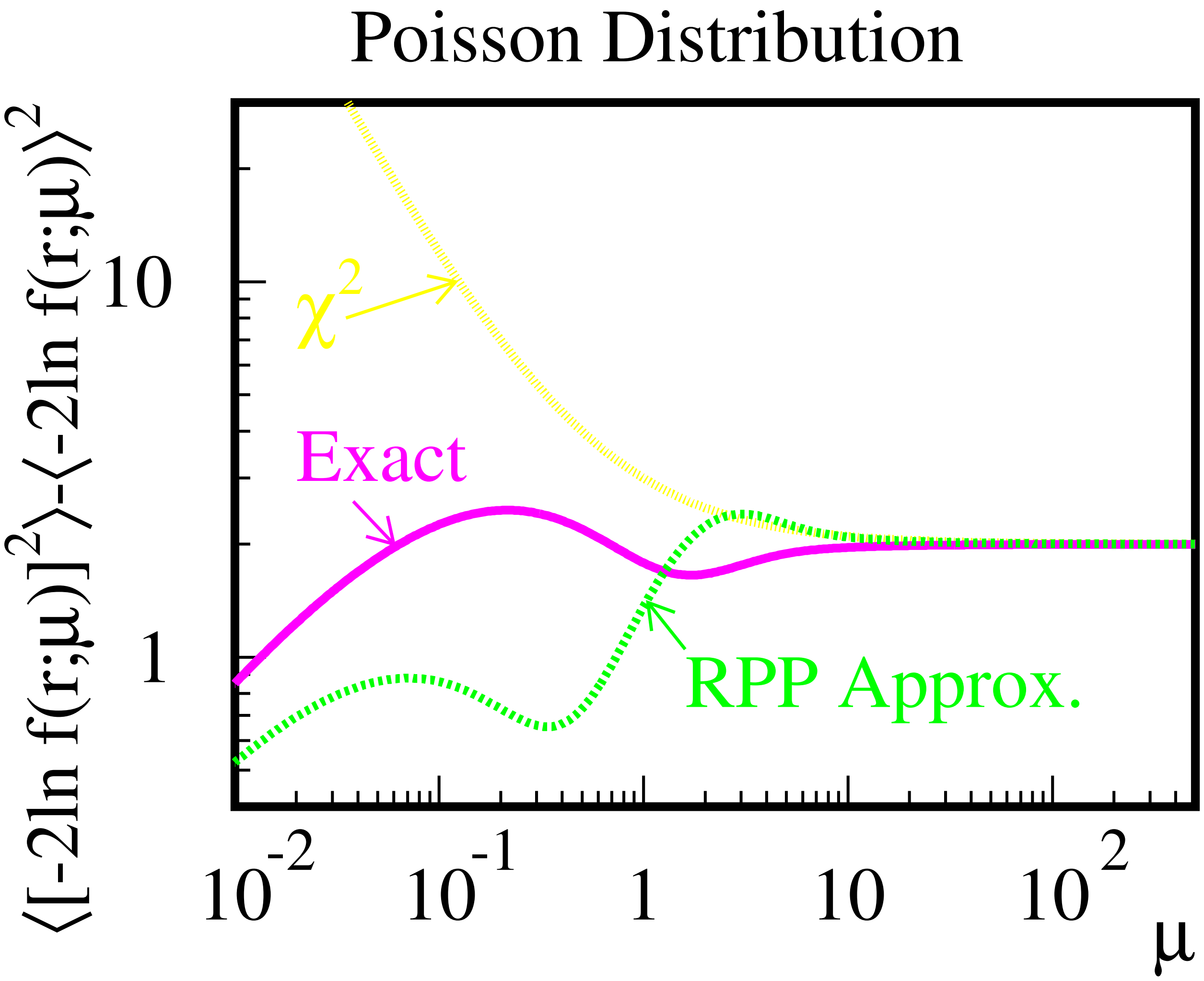

Figure 4: The variance of the logarithm of the Poisson likelihood relative to the most likely, as a function of the mean number of entries, μ, as described by Eqs. 3 and 4.

We have described how to obtain the mean and variance of the sum of the logarithm of the likelihood function, at best fit, for a Poisson likelihood fit, and recommended the subtraction of m from the mean, and $2m$ from the variance, to describe the effect of pulls from the fit. We have used to central limit theorem to estimate that the *pdf* of the logarithm of the likelihood function will be gaussian with the mean and variance so obtained.

Of course, the procedure above is just an estimate. A prescription to obtain a more accurate estimate of the goodness of fit (as well as confidence intervals for the m fit parameters), would be:

1. Perform the fit to the data to obtain the m parameters that best fit the data, and the sum $-2\ln \mathrm{L}$ at best fit.

2. Repeat the following procedure, many (probably thousands of) times:

   (a) For each bin, use the μ estimated from the m best-fit parameters to generate a random value for r, consistent with the Poisson distribution.

   (b) Perform a fit to the binned data so synthesized.

   (c) Histogram the sum $-2\ln \mathrm{L}$ of the fit, as well as each of the m parameters

The histograms so obtained can be used to directly obtain confidence levels for the goodness of fit, and for the m estimated parameters. Non-gaussian behavior will be modeled, as will the full effect of pulls.

In practice, hundreds of test fits are done in an analysis, and it would be onerous to repeat the procedure just described for all of them. The procedure based on the central limit theorem is a reasonable intermediate strategy, for both qualitative understanding, and for quick turnaround, until a near-to-final fit is ready.

For the final fit, the procedure enumerated above is surely worth the effort, as is the comparison with the gaussian *pdf*'s that one would surmise from the central limit theorem.

Acknowledgments

Most of the ideas described here grew out of discussions with David Asner, Claudio Campagnari, Tony Hill, David Lange, Tim Nelson, and Mike Witherell, whom I acknowledge, and thank.

Appendix

This is the `PAW kumac` used to make the curves presented in Figures 1 and 3.

```
* fin.kumac, for PAW
* Poisson goodness of fit (Harry Nelson 11/20/98)  hnn@charm.physics.ucsb.edu
* array from 0 to nt; nt is the maximum considered number of events in a bin
nt=2000
ntp1=[nt]+1
sigma a=array([ntp1],0#[nt])
* loop over some mus
* keep lower limit higher than zero...
do oom=-5.,2.7,0.01
 sigma emu=[oom]
 sigma mu=10.**emu
* make up the log of the poisson, then exponentiate
 sigma aa=a*log(mu)-mu-alogam(a+1.)
 sigma pois=exp(aa)
 sigma bb=-2.*aa
 sigma cc=-2.*mu*(log(mu)-1.)
 sigma dd=-2.*(mu*log(mu)-mu-alogam(mu+1.))
 sigma xn=vsum(pois)
 sigma xml=vsum(bb*pois)/xn
```

```
 sigma xml2=vsum(((bb-xml)**2)*pois)/xn
 sigma gml2=4.*mu*(log(mu))**2
** columns are: (1)mu, (2)exact mean -2lnL,
*** (3)mean -2lnL, small mu approximation (unbinned)
*** (4)mean -2lnL - 1, large mu approximation, (5)variance per bin, exact
*** (6)variance per bin, small mu approximation (unbinned)
 v/wr mu,xml,cc,dd,xml2,gml2 fin.data '6(1X,G11.5)' O
enddo
```

References

[1] C. Caso *et al.*, *European Physical Journal C* **3** (1998) 1.

[2] Actually, inclusion of the '$-\mu$' in Eq. 1 identifies this as the 'Extended' Maximum Likelihood technique, see R. Barlow *Nucl. Instr. and Meth.*, **A297**, 496 (1990); this technique preserves the area, like that of Ref. [6].

[3] H. Nelson and T. Nelson, 'A Guide to Unbinned Maximum Likelihood Fits,' CLEO Collaboration CBX 98-61, October 1998 (unpublished).

[4] Treating r as a continuous variable, differentiation of Eq. 1 with respect to r yields the equation for the most probable r, r_{mp}, given by: $\psi(1 + r_{mp}) = \ln\mu$. Here, the Psi or Digamma function is defined as $\psi(x) = d\ln\Gamma(x)/dx$. [5]. The value of r_{mp} tends to be less than μ, and r_{mp} is zero for $\mu < e^{-\gamma} \approx 0.561$, where $\gamma \approx 0.577$ is Euler's constant. If one respects the fact that r is actually discrete, then the most probable value of r is the integer nearest to, but less than, μ.

[5] *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, Edited by M. Abramowitz and I.A. Stegun, United States Department of Commerce, National Bureau of Standards, Washington, D.C. (1964).

[6] S. Baker and R. Cousins, *Nucl. Instr. and Meth.* **221**, 437 (1984).

[7] The discussion appears in Section 29.3.3, page 173 of the 1998 RPP; the approximation is Stirling's, that $\ln x! =$ constant $+ x\ln x - x$, so that $-2[\ln f(r;\mu) - \ln f(\mu;\mu)] = -2([r-\mu]\ln\mu - r\ln r + r + \mu\ln\mu - \mu) = 2(\mu - r) + 2r\ln(r/\mu)$.

[8] G. Brandenburg *et. al.*, 'A New Measurement of $B \to D^*\pi$,' CLNS 97/1485, CLEO 97-11, March 1998, published as *Phys. Rev. Lett.* **80** 2762 (1998).