Generalization of Chisquare Goodness-of-Fit Test for Binned Data Using Saturated Models, with Application to Histograms

Robert D. Cousins*
Dept. of Physics and Astronomy
University of California, Los Angeles, California USA

April 29, 2010; revised March 3, 2013

Abstract

This note is a quick review of a generalization of the chisquare goodness-of-fit test for the situation when the data are not Gaussian (as for example histogram bin contents). The generalization, already in use for many years, is based on the likelihood ratio test in which one uses in the denominator a *saturated model*, i.e., a model that fits the data exactly. As with the Gaussian test (and in fact any goodness-of-fit test), the power of the more general test can depend strongly on the alternative hypothesis.

1 Introduction

A goodness-of-fit (GOF) test is a test of the null hypothesis when the alternative hypothesis is not specified. Since Neyman and Pearson taught us that (even for simple hypotheses) the best test of the null hypothesis depends on the alternative, there is no universally best GOF test. Nonetheless, the ubiquity of the chisquare GOF test attests to its utility, at least for picking up certain departures from the null. In its usual form for uncorrelated Gaussian (normal) distributed data, one has

$$\chi^2 = \sum_i \frac{(d_i - f_i)^2}{\sigma_i^2}, \tag{1}$$

where $d_i \pm \sigma_i$ is the ith measured data point with rms deviation σ_i (each assumed to be a known constant), and f_i is the model prediction, perhaps with parameters. (If σ_i is not a known constant at each i, but depends on the unknown true value of the model at i, then there are subtleties beyond the scope of this short

*cousins@physics.ucla.edu

note.) Considered as a random variable (since a function of the random data), in many applications this test statistic, χ^2, has a probability density [1] which is also frequently called the chisquare function, with the potentially confusing ambiguity in multiple meanings usually resolvable by context. This can/should also be avoided by using another name, such as S^2 or Q^2 for the left hand side of Eqn. 1, but I yield to common practice in this note.

In HEP, we can also encounter situations in which the so-called "regularity conditions" are not met, so that the distribution of the test statistic in Eqn. 1 is not a chisquare function; two common cases are when the true or best-fit values of parameters are on the boundary (physical constraint such as non-negativity), and when there are issues with degrees of freedom not being well-defined. Again, these issues are beyond the scope of this note, which focuses on a single issue, that concerning the generalization of this usual chisquare to the case where the data are not Gaussian.

2 Likelihood ratios and the saturated model

For the same data and model as above, the likelihood is:

$$\mathcal{L} = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-(d_i - f_i)^2/2\sigma_i^2\right), \tag{2}$$

leading to the common statement that $-2\ln\mathcal{L}$ is equal to χ^2; this is of course not correct since the former quantity has extra constants. Understanding why the extra terms "disappear" in the GOF test is an important point of this note.

Wilks [2] taught us that certain likelihood *ratios* obeying important regularity conditions have asymptotic probability densities which also follow the chisquare probability density. Likelihoods have the appealing property of being independent of the metric in which parameters are described. Likelihood ratios inherit this property and furthermore are invariant under change of metrics in which the data are described. For deep reasons such as the Neyman-Pearson Lemma, likelihood ratios are generally useful for comparing two hypotheses. In this note, likelihood ratios are denoted by λ, and are the basis for a generalization of Eqn. 1.

Given only a null hypothesis and the data, one can invent an alternative hypothesis for which f_i is equal to the data d_i at every measured value. Such a model, which typically needs as many parameters are there are data points, is called a *saturated model* [3]. In some circumstances, saturated models can be useful for comparisons with the null hypothesis, and in particular for providing a denominator in the likelihood ratio.

For the Gaussian data above, the saturated model sets $f_i = d_i$, so that the likelihood of the data in the saturated model is (since $\exp(0) = 1$)

$$\mathcal{L}_{\text{saturated}} = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}}. \tag{3}$$

The ratio of the two likelihoods above is then

$$\lambda = \prod_i \exp\left(-(d_i - f_i)^2/2\sigma_i^2\right), \tag{4}$$

and thus, importantly,

$$\chi^2 = -2\ln\lambda. \tag{5}$$

From this point of view (there are other paths to Eqn. 1 not as relevant to this note), the constants in $-2\ln\mathcal{L}$ were not just ignored; they were canceled when a ratio was formed. Since the saturated model does not depend on the parameters of the original model, the maximum of λ is of course at those parameters that maximize the original $\mathcal{L}$.

In 1983, Baker and Cousins [4] reviewed the use of the saturated model (although we did not call it that, instead citing a 1928 paper by Neyman and Pearson) when applied to multinomial and Poisson data in histograms. The result is a GOF statistic for each case, which we called (following some literature of the time) $\chi^2_{\lambda,m}$ and $\chi^2_{\lambda,p}$, respectively. The Poisson form is mentioned in the PDG's Review of Particle Properties [1]; some time ago we decided that it is best just to denote it by $-2\ln\lambda$ as some felt that calling it χ^2 might encourage people to forget that it only asymptotically follows the χ^2 distribution (and only if conditions are satisfied). As we said in our paper, probably the safest thing to do is to study the distribution by Monte Carlo. Heinrich [5] has studied the distribution and moments of λ for small statistics, and makes the point that for the asymptotic formulas to be valid, the contents of all bins must each be large.

I think that in some of the cases that we have had before the statistics committee, the expanded use of the saturated model might be useful. For example, it might be applied when comparing simultaneous predictions of a model to a set of several binomial problems (asymmetries or efficiencies). One needs to keep in mind that even the ungeneralized GOF statistic is insensitive to certain departures from the null (since it throws away sign and order information). The generalized version needs to be understood in the same spirit, in addition to checking the distribution under the null by M.C. But with these cautions, I would encourage the trial and study of GOF test statistics based on saturated models.

3 Caution against absolute likelihood as goodness of fit statistic

Occasionally one finds the recommendation to use as a goodness of fit test statistic the absolute likelihood at its maximum, as opposed to a *ratio* of likelihood maxima; typically Monte Carlo simulation is recommended as the way to get the null distribution of the test statistic. Although this might at first seem plausible, it ia a flawed concept: unlike the likelihood ratio, such a GOF statistic is without foundation, and power can vary arbitrarily with the metric. Simple examples can make clear that the value of the likelihood at its maximum must

be compared to something. For example, for Poisson counts, the probability of observing 100 events when $\mu = 100.0$ is much less than the probability of observing 1 event when $\mu = 1.0$, even though in both cases the data perfectly fit the theory. Thus the saturated model provides the reference of the largest value that the likelihood can be for that data (for *any* model), and hence provides a reasonable normalization for the maximum observed for a more constraining model.

Heinrich [6] discusses the pitfalls of using the absolute likelihood as a GOF statistic. For unbinned likelihoods, which are common in HEP, the problem is exacerbated. One might hope that one could just take binned GOF in the limit of small bins, but the answer depends on the way the limit is taken, i.e., in which metric the bin boundaries are equally spaced. Goodness of fit for unbinned data in one dimension has a variety of tests to choose among [7], depending on roughly what sort of alternatives one wants power against. In higher dimensions, the problem is yet more difficult [8].

Acknowledgments

I thank Louis Lyons and other members of the CMS Statistics Committee for corrections and comments on earlier drafts. This work was partially supported by the U.S. Department of Energy and the National Science Foundation.

References

[1] C. Amsler et al. (Particle Data Group), "Review of Particle Physics," Physics Letters **B667** (2008) 1. `http://pdg.lbl.gov/2009/reviews/rpp2009-rev-statistics.pdf`, Eqn. 32.12.

[2] S.S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," Annals of Math. Stat. **9** (1938) 60.

[3] J.K. Lindsey, *Parametric Statistical Inference* (New York: Oxford University Press), 1996.

[4] S. Baker and R. D. Cousins, "Clarification of the use of chi-square and likelihood functions in fits to histograms," Nucl. Instrum. Meth. **221** (1984) 437.

[5] Joel G. Heinrich, "The Log Likelihood Ratio of the Poisson Distribution for Small μ," CDF/MEMO/CDF/CDFR/5718, Version 2, `http://www-cdf.fnal.gov/physics/statistics/notes/cdf5718_loglikeratv2.ps.gz`

[6] Joel Heinrich, "Pitfalls of Goodness-of-Fit from Likelihood", talk at PhyStat 2003 (Stanford, CA), arXiv:physics/0310167

[7] *Goodness-of-Fit Techniques*, edited by Ralph B. D'Agostino and Michael A. Stephens, Vol. 68 of Statistics: Textbooks and Monograph. (New York, Marcel Dekker, 1986).

[8] Mike Williams, "How good are your fits? Unbinned multivariate goodness-of-fit tests in high energy physics," JINST 5 P09004 (2010), arXiv:1006.3019 [hep-ex].