Pitfalls of Goodness-of-Fit from Likelihood

Joel Heinrich
University of Pennsylvania, Philadelphia, PA 19104, USA

The value of the likelihood is occasionally used by high energy physicists as a statistic to measure goodness-of-fit in unbinned maximum likelihood fits. Simple examples are presented that illustrate why this (seemingly intuitive) method fails in practice to achieve the desired goal.

1. INTRODUCTION

> For every complex problem, there is a solution that is simple, neat, and wrong. *H.L. Mencken*

The complex problem considered here is goodness-of-fit (g.o.f.) for unbinned maximum likelihood fits in cases when binned g.o.f. methods and Kolmogorov-Smirnov are not well suited:

A physicist, having fit a complicated model to his multi dimensional data to obtain estimates of the values of certain parameters, is also expected to check how well the data match his model. In the sections that follow, we discuss a g.o.f. method, still occasionally used in high energy physics (HEP), that is simple, neat, and wrong.

2. THE SNW[1] METHOD

We start with a brief description of the method. (A true derivation, for obvious reasons, is not available.)

observation: Maximum likelihood fits are performed by maximizing the likelihood $L(\vec{\theta},\vec{x})$ with respect to the (unknown) parameters $\vec{\theta}$ for fixed data $\vec{x}$.

faulty intuition: Thus, the value of the likelihood provides the g.o.f. between the data and the probability density function (p.d.f.): The value of the likelihood at the maximum,

$$L_{\max} = L(\hat{\vec{\theta}},\vec{x})$$

corresponds to the best fit—the smaller the likelihood, the worse the g.o.f., ...

obstacle: To calculate this "g.o.f." P-value, we need the distribution of $L_{\max}$ for an ensemble of random $\vec{x}$ deviates from the p.d.f. using the true (but unknown) parameters $\vec{\Theta}$.

faulty resolution: We approximate this by replacing $\vec{\Theta}$ with the parameter estimate obtained from the fit to the actual data.

This method has a long history of use in high energy physics. It's recommended by several excellent statistical data analysis texts written by (and for) high energy particle physicists. Consequently, and because the method is "obvious", it's still being used in (some) HEP analyses.

Reference [1], written by a statistician and four physicists, describes the method, but criticizes:

> The likelihood of the data would appear to be a good [g.o.f.] candidate at first sight. Unfortunately, this carries little information as a test statistic, as we shall see...

Since this was ignored, maybe its warning was not strong enough. I have found no mention of the method in texts written (solely) by statisticians.

3. A SIMPLE TEST OF THE METHOD

> Always test your general reasoning against simple models. *John S. Bell*

Reference [2], following the above advice, tests the method against the p.d.f.

$$\frac{1}{\tau}e^{-t/\tau} \qquad (t \geq 0)$$

where t (we have in mind the decay-time of a particle) follows an exponential distribution, and τ (the mean lifetime) is a parameter whose value, being unknown, is estimated from data. The likelihood for N observations t_i is given by

$$-\ln L = \sum_{i=1}^{N}\left[\ln\tau + \frac{t_i}{\tau}\right]$$

The value ($\hat{\tau}$) of τ that maximizes the likelihood, and the value ($L_{\max}$) of the likelihood at its maximum, are given by

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N} t_i \qquad -\ln L_{\max} = N(1+\ln\hat{\tau})$$

[1] Simple, Neat, Wrong.

3.1. The First Surprise

The value of the likelihood at its maximum (in this test case) is just a simple function of $\hat{\tau}$—all samples with the same mean obtain the same "g.o.f." value. This is a disaster for g.o.f. Even if the true value of τ—call it $\mathcal{T}$—were known in advance, so that we could calculate the P-value associated with the observed $\hat{\tau}$, merely comparing the $\hat{\tau}$ of the data with $\mathcal{T}$ is not sufficient to show that the observed data are modeled well by the exponential distribution.

3.2. The Second Surprise

Since under this method, our P-value ensemble is actually based on the value of $\hat{\tau}$ computed from the data (not knowing the true value $\mathcal{T}$), we *always* obtain a P-value of about 50%, *for any data whatsoever*. This is a second disaster for g.o.f. By construction, the distribution of $L_{\max}$ from our ensemble of N-event pseudo experiments tracks the $L_{\max}$ observed from the data.

The fact that the method yields "reasonable" P-values has undoubtedly contributed to its longevity in practice: P-values very near 0 or 100% would have triggered further investigation.

3.3. Lessons Learned

In this example, g.o.f. is equivalent to testing the single hypothesis: "The data are from an exponential distribution of unspecified mean." $L_{\max}$ provided no information with respect to this hypothesis.

What went wrong? In our test case, the likelihood could be expressed as a function of just the parameter and its maximum likelihood estimator (m.l.e.): $L(\tau;\hat{\tau})$. *All* data samples with the same m.l.e. gave the same "g.o.f."

Exactly the same thing happens in the Gaussian (normal) case—the likelihood can be written using solely the 2 parameters and their estimators: $L(\mu,\sigma;\hat{\mu},\hat{\sigma})$.

Other "textbook" distributions—scaled gamma, beta, log-normal, geometric—also fail in the same way. Geometric is a discrete distribution, so the problem is not restricted to the continuous case.

4. MORE TROUBLE: NON INVARIANCE

Returning to our exponential example, suppose we make the substitution $t = x^2$. The p.d.f. transforms as

$$\frac{1}{\tau}e^{-t/\tau}dt = \frac{2x}{\tau}e^{-x^2/\tau}dx$$

and the g.o.f. statistic is now calculated as

$$-\ln L = \sum_{i=1}^{N}\left[\ln\tau + \frac{x_i^2}{\tau} - \ln(2x_i)\right] \qquad \hat{\tau} = \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

$$\begin{aligned} -\ln L_{\max} &= N(1+\ln\hat{\tau}) - \sum_{i=1}^{N}\ln(2x_i) \\ &= N(1+\ln\hat{\tau}) - \sum_{i=1}^{N}\ln(2\sqrt{t_i}) \end{aligned}$$

That is, the "g.o.f." statistic is not invariant under change of variable in the continuous p.d.f. case. (The value of the m.l.e. is, of course, invariant.)

Under change of variable, the "g.o.f." statistic picks up an extra term from the Jacobian—an extra function of the data. We're free to choose any transformation, so we can make the "g.o.f." statistic more or less anything at all—a serious pathology.

At this point, experts point out that *ratios* of likelihoods have the desired invariance under change of variable, but, while the likelihood ratio is a useful test statistic in certain special cases, it is not at all clear how to obtain a useful g.o.f. statistic from the likelihood ratio in the general, unbinned, case.

5. A REPLACEMENT MODEL

Since we now lack an intuitive understanding, we need a replacement intuition for what is going on. I propose this model:

Denote by H_0 the hypothesis that the data are from the p.d.f. in question. Specify an alternative hypothesis H_1 that the data are from a uniform p.d.f. (flat in the variables that we happen to have chosen). At least, the H_1 p.d.f. is flat over the region where we have data—outside that region it can be cut off.

Performing a classic Neyman-Pearson hypothesis test of H_0 vs H_1, we use the ratio of their likelihoods as our test statistic:

$$\lambda(\vec{x}) = \frac{L(\vec{x}|H_0)}{L(\vec{x}|H_1)} = \frac{L(\vec{x})}{\text{constant}}$$

So, the "g.o.f." statistic can be re-interpreted as suitable for a hypothesis test that indicates which of H_0 (our p.d.f.) and H_1 (a flat p.d.f.) is more favored by the data—a well established statistical practice.

The benefit of the new interpretation is that it explains behaviors that were baffling under the g.o.f. interpretation: Neyman-Pearson hypothesis tests and g.o.f. tests behave quite differently.

For example, a reasonable g.o.f. statistic should be at least approximately distribution independent, but $\lambda(\vec{x})$ is often highly correlated with the m.l.e.'s

(100% in our exponential case). This high correlation was confirmed in the example contributed by K. Kinoshita[3] to the 2002 Durham Conference. Not knowing the true value of the parameters then makes it difficult, or impossible, to use $\lambda(\vec{x})$ as g.o.f., since we don't know what $\lambda(\vec{x})$ *should* be.[2] The behavior of these correlations is natural and obvious in the hypothesis test picture: changing the parameters changes the "flatness" of the H_0 p.d.f., and $\lambda(\vec{x})$ reflects this.

Reference [1] pointed out that, with no unknown parameters, one can always transform the p.d.f. to a flat distribution. Then $\lambda(\vec{x})$ becomes constant independent of the data—bad news for g.o.f. In the hypothesis test picture, this becomes a comparison between two identical hypotheses, and the result is what we would expect.

6. TEST BIAS

Take the H_0 p.d.f. to be

$$e^{-t} \qquad (t \geq 0)$$

This distribution is fully specified—no unknown parameters. Our "g.o.f." statistic is then

$$-\ln L = N\hat{t}$$

whose mean is $\langle -\ln L \rangle = N$, and variance is $\mathrm{Var}(-\ln L) = N$, for an ensemble of data sets from the H_0 p.d.f. A data set with $\hat{t}$ close enough to 1 will be claimed to be a good fit to the H_0 p.d.f.

But say, unknown to us, the data are really from a triangular p.d.f.:

$$1 - |t - 1| \qquad (0 \leq t \leq 2)$$

The mean and variance of $N\hat{t}$ will be N and $N/6$ respectively, for data from the triangular distribution. So, although the exponential and triangular p.d.f.'s are quite different, the triangular data will be more likely to pass the g.o.f. test than exponential data for which it was intended. Statisticians refer to this situation as a case of "test bias".

We conclude that, even with no free parameters, the "g.o.f." test is biased: there exist "impostor" p.d.f.'s that should produce bad fits, but instead pass the "g.o.f." test with greater probability then the p.d.f. for which the test was designed. Reference [4] gives additional examples of this behavior.

From the hypothesis test point of view, this behavior makes sense. The exponential and triangular data have the same "distance" from the flat distribution, on the average, with the triangular data being less susceptible to fluctuations. The hypothesis test doesn't tell us when the data are inconsistent with both H_0 and H_1.

7. ANOTHER EXAMPLE

Here we try to find an example p.d.f. (with a free parameter) that the method in question can handle well. We use the insight provided by the hypothesis test picture. We want to keep the correlation between the free parameter and the g.o.f. statistic $L_{\max}$ to a minimum. In the hypothesis test picture, this is achieved when the "flatness" of the p.d.f. is independent of the parameter. A location parameter has this property. Additionally, we want the p.d.f. to be easily distinguishable from a flat p.d.f. So we choose the Gaussian

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-0.5(x-\mu)^2/\sigma^2}$$

where μ is unknown, but σ is specified in advance. The likelihood is given by

$$-\ln L = \sum_{i=1}^{N} \left[\ln\sqrt{2\pi} + \ln\sigma + \frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2 \right]$$

When μ and σ are both unknown, their m.l.e.'s are

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

Using these expressions, we can rewrite the likelihood in the form $L(\mu, \sigma; \hat{\mu}, \hat{\sigma})$:

$$-\ln L = \frac{N}{2}\left[\ln(2\pi) + \ln(\sigma^2) + \frac{\hat{\sigma}^2 + (\hat{\mu} - \mu)^2}{\sigma^2} \right]$$

When only μ is unspecified, its m.l.e. is $\hat{\mu}$ as above, and the value of the maximized likelihood is

$$-\ln L_{\max} = \frac{N}{2}\left[\ln(2\pi) + \ln(\sigma^2) + \frac{\hat{\sigma}^2}{\sigma^2} \right]$$

Our victory is that $L_{\max}$ only depends on $\hat{\sigma}$, which is an ancillary statistic for μ. That is, we don't need to know the true value of μ in order to calculate the distribution of our g.o.f. statistic in this carefully chosen example. In fact, a convenient form for the g.o.f. statistic is

$$N\frac{\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{N} \left(\frac{x_i - \hat{\mu}}{\sigma}\right)^2$$

[2] Small correlations are not fatal. For example, if the P-value of g.o.f. for the observed data in a particular case ranged only between, say, 20% and 30%, for different true values within $\pm 3\sigma$ of the estimated value of a parameter, one would be justified in concluding "good fit" (assuming the g.o.f. statistic used had the right properties in other respects).

which is well known to have the distribution (under the null hypothesis) of a χ^2 with $N-1$ degrees of freedom.

7.1. The Bad News

Before we declare that the method performs well in this example, there are several ugly facts to consider:

- Data that match the null hypothesis well yield $N\hat{\sigma}^2/\sigma^2 \simeq N$. Much larger or much smaller values of the g.o.f. statistic imply poor g.o.f. This is in contrast to Pearson's χ^2 (binned χ^2), for example, where smaller χ^2 is always better g.o.f. So we must interpret this statistic differently than how we are used to.
- The g.o.f. in this example simply reduces to a comparison between the sample variance and σ^2. Any distribution with variance approximately equal to σ^2 will usually generate data that "pass the test", even distributions that look nothing like a Gaussian. This is the same kind of problem that we first saw in section 3.1.
- A construction similar to that of section 6 will produce "impostor" p.d.f.'s that pass the "g.o.f." test with greater frequency than the null hypothesis. So, we have not eliminated the test bias problem.

In this example, the g.o.f. method in question will be able to flag some, but not all, of data samples that poorly match the null hypothesis. In answer to the question "Are the data from a Gaussian with unspecified mean, and variance equal to σ^2?", this g.o.f. method can only answer "No" or "Maybe": it checks the variance part of the question, but does nothing to check the Gaussian part.

8. CONCLUSIONS

- This "g.o.f." method is fatally flawed in the unbinned case. Don't use it. Complain when you see it used.
- With fixed p.d.f.'s, the method suffers from test bias, and is not invariant with respect to change of variables. These problems persist when there are floating parameters.
- With floating parameters, the method is often circular: "g.o.f." becomes a comparison between the measured values and the true (but unknown) values of the parameters...
- The misbehavior of this "g.o.f." statistic is understandable when reinterpreted as the ratio between the likelihood in question and a uniform likelihood, and used to distinguish between these two specific hypotheses. Dual-hypothesis tests are not g.o.f. tests.

Acknowledgments

I would like to thank Louis Lyons for several helpful discussions of the points raised here, and the organizers of the PHYSTAT2003 Conference for arranging a superb program.

References

[1] W.T. Eadie, D. Drijard, F.E. James, M. Roos, and B. Sadoulet, *Statistical Methods in Experimental Physics*, chapter 11, pages 268, 271, (North-Holland Publishing Co, Amsterdam, 1971).

[2] J.G. Heinrich, "Can the likelihood function be used to measure goodness-of-fit?", CDF Internal Note 5639, (2001).
`www-cdf.fnal.gov/publications/cdf5639_goodnessoffitv2.ps.gz`

[3] K. Kinoshita, "Evaluating quality of fit in unbinned maximum likelihood fitting", in *Proceedings of the Conference on Advanced Techniques in Particle Physics*, edited by M. Whalley and L. Lyons, p 176, (2002).
`www.ippp.dur.ac.uk/Workshops/02/statistics/proceedings/kinoshita.ps`

[4] J.G. Heinrich, "Unbinned likelihood as goodness-of-fit for fixed distributions: A critical review", CDF Internal Note 6123, (2002).
`www-cdf.fnal.gov/publications/cdf6123_gof_like_fixed.ps`