

GOODNESS-OF-FIT FOR SPARSE DISTRIBUTIONS IN HIGH ENERGY PHYSICS

BRUCE YABSLEY

*High Energy Physics Department, School of Physics
University of Sydney, NSW 2006 Australia
E-mail: b.yabsley@physics.usyd.edu.au*

We consider Pearson's chi-square X^2 , the likelihood ratio G^2 , and Zelterman's D^2 as goodness-of-fit statistics for high energy physics problems in several dimensions, where the data are sparse. There is a fundamental obstacle in the "ultrasparse" case where all bins have at most one entry ($n_i = 0, 1 \forall i$). A condition for avoiding this regime is derived; the allowed number of bins k rises faster than the total number of events n : $k_{\max} = 0.4 \times n^{1.4}$. Reasonable binning in many dimensions may thus be possible for modest datasets $n > O(100)$, although special treatment is required to derive p -values. Results for an initial trial problem are encouraging; further studies are underway.

1. Motivation / Historical note

The talk of the Durham meeting was Heinrich's demonstration¹ that the likelihood cannot be used to test goodness-of-fit (g.o.f.) for unbinned maximum likelihood fits (see also Refs 2 and 3). This presents a problem for high energy physics, where the data are often characterised by several variables, leading to the use of unbinned fits to small samples. Due to the importance of such fits at the B-factories, Kay Kinoshita and I both pursued the matter in the following year, considering binning-free tests based on the random walk⁴ and the energy test,⁵ with inconclusive results. During discussion at PHYSTAT2003, an alternative approach was suggested:⁶

Conventional binned g.o.f. tests rely on results from the asymptotic limit where the number of bins k is fixed, and the number of events $n \rightarrow \infty$. This is one reason behind the conventional wisdom that fits should have $n_i \geq 5$ events in each bin. However an alternative limit, where $k \rightarrow \infty$ but the ratio of events to bins n/k remains finite, has been studied: see for example Ref. 7. There is considerable statistical literature on g.o.f. in this regime, mostly considering problems in the social sciences (for example, Ref. 8). Here I report the status of an attempt to appropriate this work for use in high energy physics.

2. Adapting a social science example

As a starting point, Zelterman⁹ considers a 2D histogram from an employment survey,¹⁰ with $n = 129$ events and $k = 899$ cells: well outside the conventional regime (Fig. 1). The null hypothesis is "independence of the rows [monthly salary] and columns [years since first degree] by using multinomial sam-

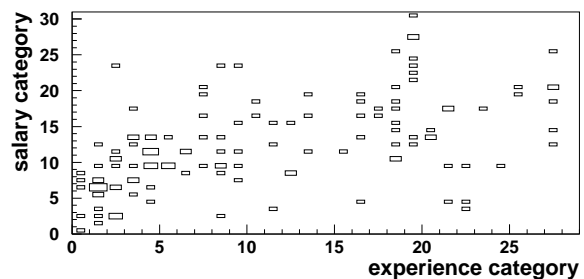


Fig. 1. The sparse histogram used as an example by Zelterman,⁹ plotting salary *vs.* years of experience; the smallest squares show one event/bin. The data are taken from an employment survey in Ref. 10.

pling, conditional on the marginal totals";⁹ the alternative hypothesis is that a correlation exists. By inspection, confirmed by linear regression, the data of course *are* correlated. Tests based on

$$X^2 = \sum_i (n_i - \lambda_i)^2 / \lambda_i, \quad (1)$$

$$G^2 = 2 \sum_i n_i \log n_i / \lambda_i, \text{ and} \quad (2)$$

$$D^2 = \sum_i [(n_i - \lambda_i)^2 - n_i] / \lambda_i \quad (3)$$

where n_i is the actual and λ_i the predicted number of events in bin i , find various results: X^2 (Pearson's χ^2) fails to reject the null hypothesis; the likelihood ratio statistic G^2 and D^2 both reject it at extreme significance. D^2 , which is outside the family of Cressie and Read,¹¹ was introduced by Zelterman for use in the case of sparse data.⁹ Both this and G^2 seem suitable to our purpose, based on this example.

No exact mathematical relationship between the quantities in Fig. 1 is expected; they can thus be grouped into categories — binned — according to

convenience. Typical high energy physics data are different: the variables are invariant masses, momenta, angles, *etc.*, and the underlying processes are intrinsically simple. Formulae relating various quantities can be derived for some hypotheses, such as decay of a particle with given properties, and for others (*e.g.* combinatorial backgrounds) functions with few parameters are found to fit the data well. It would thus be attractive to choose bins fine enough to distinguish different (possibly correlated) distributions in each quantity: in many dimensions this can lead to arbitrarily small numbers of events-per-bin.

This approach will fail in the limit where $n_i = 0, 1 \forall$ bins i : a test statistic $I = \sum_i f(n_i, \lambda_i)$ cannot in general distinguish between regions of low and high event density. D^2 collapses to a unique value in this case, for any $\{\lambda_i\}$; all statistics I collapse to the unique value $n \cdot f(1, \lambda) + (k - n) \cdot f(0, \lambda)$ if the distribution is flat, $\lambda_i = \lambda \forall i$. (Note that the form given for I is general, including the family of Cressie and Read,¹¹ D^2 , and other statistics.) Since we will typically fit data with floating shape parameters, the limitation is fatal.

To avoid this “ultrasparse” regime, consider the following condition: Let m_j be the number of bins i where $n_i = j$ (so that there are m_1 bins with one entry, *etc.*), and find the number of bins $k = k_{\max}$ such that

$$P\left(m_2 \leq \frac{1}{10}n; m_j = 0 \forall j > 2\right) < 0.01 \quad (4)$$

in the case where the expected bin populations are equal, $\lambda_i = n/k$. If this condition is met, then the majority of datasets will have a significant number of bins with $n_i = 2, 3, \dots$, even though the average bin population may be low: at most 1% will be dominated by bins with one entry, $n_i = 1$. (We consider datasets with only a small number of $n_i = 2$ bins $m_2 \leq n/10$, and no bins with $n_i > 2$, to be dominated by their single-entry bins.)

For given n and k , the probability of any particular set of counts $(m_0, m_1, m_2, m_3, \dots)$ is

$$P(\{m_j\}|n, k) = \frac{n!}{\prod_j (j!)^{m_j}} \cdot \frac{k!}{\prod_j m_j!} \cdot \left(\frac{1}{k}\right)^n, \quad (5)$$

based on multinomial statistics and simple counting. Using (5) and an arbitrary-precision calculator, it is straightforward to solve (4) for k_{\max} . Results are shown in Fig. 2, together with a fit to the power law

$$k_{\max} = 0.4 \times n^{1.4}. \quad (6)$$

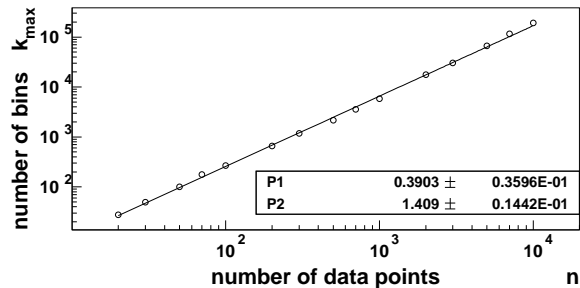


Fig. 2. The largest number of bins k_{\max} satisfying the condition (4), as a function of the number of events n . The results of a fit to the power-law $k_{\max} = P1 \cdot n^{P2}$ are also shown.

This suggests that binnings with $n/k \ll 1$ may be possible for moderate $n > O(100)$, enabling bins to be chosen in each of many dimensions, as required.

3. Example: $X(3872) \rightarrow \pi^+\pi^-J/\psi$

As an example, we consider angular analysis of decays $X(3872) \rightarrow \pi^+\pi^-J/\psi$, to determine the quantum numbers J^{PC} of the state; the sample was 58 events including ≈ 11 background.¹² Various hypotheses were tested using 1D histograms chosen with typical $n_i \geq 5$, but regions where $\lambda_i \lesssim 1$ on the null hypothesis: see Fig. 2 of Ref. 12. Event counts $n_i \gg 1$ in these bins disfavour the null.

Using toy Monte Carlo (MC) experiments with $n = 50$ and neglecting background, we study the power of tests on (1)–(3) to discriminate against $J^{PC} = 0^{-+}$ using binning in an increasing number of dimensions. Events are generated following Ref. 13, using a complete set of angles $(\theta, \phi, \psi, \chi, \phi_K)$. (Here (θ, ϕ, ψ) are as defined in Ref. 13 for the 0^{-+} case; χ is as defined for 0^{++} ; and ϕ_K is the azimuthal angle of the kaon from $B \rightarrow KX(3872)$ decay, in the system used to define ϕ .) To bin efficiently, we use non-equidistant bins $[0.0, 0.3]$, $[0.3, 0.6]$, $[0.6, 0.9]$, $[0.9, 1.0]$ in $|\cos \theta|$, where a $\sin^2 \theta = 1 - \cos^2 \theta$ distribution is expected on the null hypothesis (preserving the small expected population in the last bin,¹² but using fewer bins overall). Fig. 3 shows the value of such binning for discriminating between hypotheses.

Fig. 4 shows the power to reject the null, based on p -values taken from distributions of toy MC experiments. The X^2 and G^2 tests improve noticeably as more dimensions are added, up to case (e) with $k = 128$ bins, comparable to $k_{\max} \approx 95$ (from (6)) for $n = 50$. All tests lose power for binning (f), in the ultrasparse regime ($k = 512$), and inspection of test-

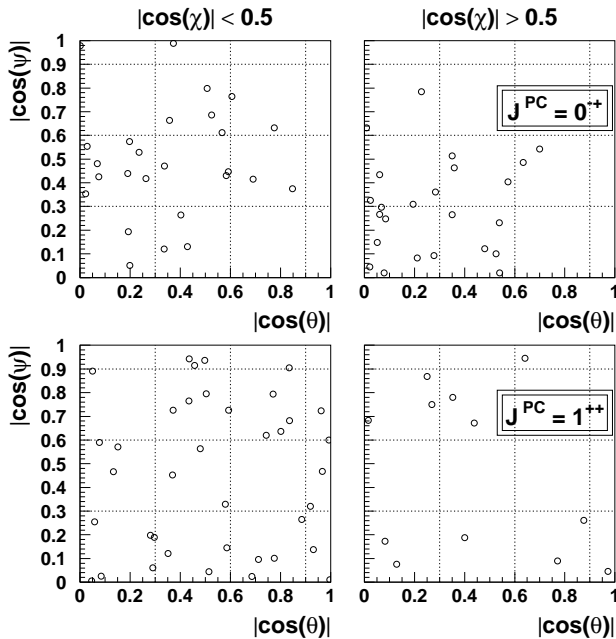


Fig. 3. Data from a single toy MC dataset, $n = 50$, for each of $J^{PC}(X(3872)) = 0^{++}$ (upper plots) and 1^{++} (lower). Dotted lines show 4×4 binning in $(|\cos \theta|, |\cos \psi|)$; left and right plots show $|\cos \chi| < 0.5$ and ≥ 0.5 bins respectively. For $J^{PC} = 0^{++}$, we expect $\sin^2 \theta \sin^2 \psi$ dependence and no dependence on $\cos \chi$. For $J^{PC} = 1^{++}$, the distribution $\propto \sin^2 \chi$; the dependence on the remaining angles $(\theta, \phi, \psi, \phi_K)$ is nontrivial.

statistic and p -value distributions shows pathologies such as domination by discrete values. X^2 generally shows higher power than G^2 , but loses power as $(1 - \alpha) \rightarrow 1.0$, where α is the significance: the mechanism needs to be studied. D^2 performs poorly in all cases, in marked contrast to the example of section 2.

Also shown is a test based on $\sum_l \ln \mathcal{L}_l$, where \mathcal{L}_l is the likelihood for the l^{th} event: $\propto \sin^2 \theta \sin^2 \psi$ for $J^{PC} = 0^{++}$. In this case, this test is discriminating, and more powerful than all of the binning-based tests shown. (This result is unlikely to be general, as there are known to be cases where \mathcal{L}_l -based tests fail to discriminate against certain alternative hypotheses, even where the null hypothesis is simple; the limitation is related to the failure of these tests to discriminate in the case of unbinned maximum likelihood fits. See Ref. 3.) Since \mathcal{L}_l is a function of $\cos \theta$ and $\cos \psi$ only, it is insensitive to variation in $\phi, \cos \chi$ or ϕ_K , not expected for $J^{PC} = 0^{++}$ but expected for some other hypotheses, in particular 1^{++} (see Fig. 3). Thus a test combining $\sum_l \ln \mathcal{L}_l$ with appropriate binning in $(\phi, \cos \chi, \phi_K)$ is presumably more powerful still: this remains to be studied.

4. Further work

In addition to further study of the results presented here, the following extensions are planned:

- (1.) Varying n in the $X(3872)$ case, to see if Eq. (6) is a reliable guide to the breakdown of tests.
- (2.) Application to a basic compound-hypothesis problem: fitting for a possible signal in the presence of background (which may be mismodelled). The prototypical problem of this kind at the B-factories is a search for a rare B-decay.¹⁴
- (3.) A difficult compound-hypothesis problem: angular analysis of $B \rightarrow \phi K^*$ or similar decays¹⁵ to determine helicity amplitudes. This is a 3D problem with a few hundred events, and thus combines features of cases (1.) and (2.).

It would be desirable to also apply this method to the analysis in Ref. 16, with $O(100)$ events and the sensitivity due to fine structure in two dimensions. Unfortunately, based on Eq. (6), this is unrealistic.

5. Conclusion

Binned goodness-of-fit tests have been considered for sparse data, where typical bin populations $n_i \ll 5$. Such tests will fail in the “ultrasparse” case where all bins have $n_i = 0, 1$ only; the condition $k_{\text{max}} = 0.4 \times n^{1.4}$ defines a number of bins k_{max} that avoids this regime, for a given number of events n . For modest sample sizes, k_{max} is large enough to allow binning in many dimensions. For an angular analysis problem with $n = 50$, substantial improvement in the power of tests is found for careful binning in four dimensions, up to the expected limit $k \approx 100$. Further study using different n , and compound hypotheses, is underway. The (non- χ^2) distribution of test statistics in this regime remains to be studied.

Acknowledgments

I’m indebted to Jan de Leeuw for our discussion at PHYSTAT2003 on the problem of goodness-of-fit statistics for sparse data, and for providing a way into the statistical literature. I would like to thank Nancy Read, Bernard Silverman, and Mike Titterton for their advice and feedback at this meeting; and the conference organisers for providing a setting where these kinds of discussions can take place.

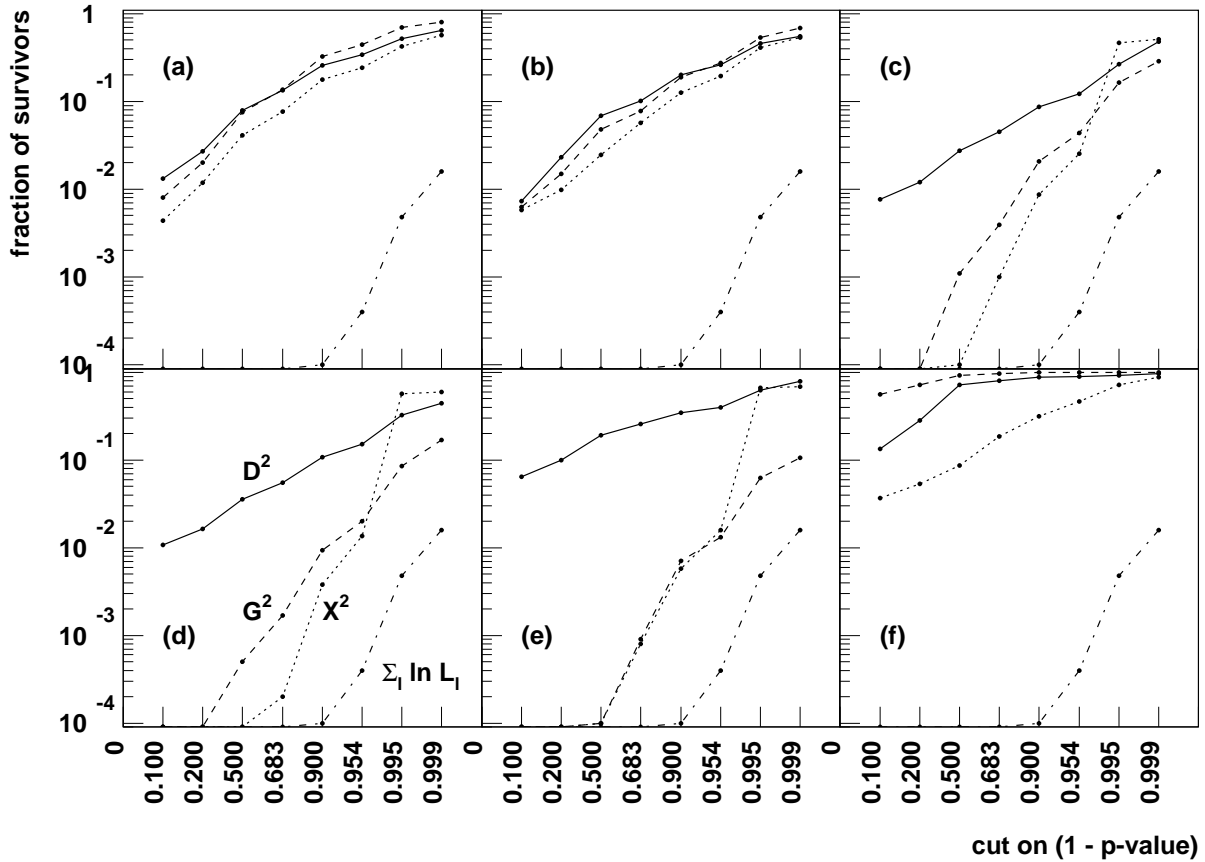


Fig. 4. Angular analysis of $X(3872) \rightarrow \pi^+\pi^-J/\psi$ (simulated): Power to reject the null hypothesis $J^{PC} = 0^{-+}$, in the case of 1^{++} data, for hypothesis tests based on D^2 (solid line), G^2 (dashed), X^2 (dotted), and the unbinned log-likelihood $\sum_i \ln \mathcal{L}_i$ (dot-dashed). The fraction of 1^{++} datasets that survive (*i.e.* $(1-\beta)$ where β is the power) is plotted against $(1-\alpha)$, where α is the significance: for tests A and B , A is more powerful if its curve lies below/right of the curve for B . Bins are chosen in a progressively larger number of variables: (a) 10 even bins and (b) 4 variable-width bins in $|\cos\theta|$, (c) 4×4 bins in $(|\cos\theta|, |\cos\psi|)$, (d) $4 \times 4 \times 2$ in $(|\cos\theta|, |\cos\psi|, |\cos\chi|)$, (e) $4 \times 4 \times 2 \times 4$ in $(|\cos\theta|, |\cos\psi|, |\cos\chi|, \phi_K)$, and (f) $4 \times 4 \times 2 \times 4 \times 4$ in $(|\cos\theta|, |\cos\psi|, |\cos\chi|, \phi_K, \phi)$.

References

- J. Heinrich, CDF Internal Note 5639 (2001). http://www-cdf.fnal.gov/publications/cdf5639_goodnessoffitv2.ps.gz
- K. Kinoshita, in *Proceedings of the Conference on Advanced Statistical Techniques in Particle Physics*, ed. M.R. Whalley, L. Lyons, 176–181 (2002).
- J. Heinrich, in SLAC-R-703, eConf C030908 (PHYSTAT2003 Proceedings), 52–55 (2003).
- K. Kinoshita, in SLAC-R-703, eConf C030908, Proceedings), 56–60 (2003). The writeup of my half of the talk is missing from the proceedings.
- B. Aslan and G. Zech, in SLAC-R-703, eConf C030908, (PHYSTAT'03 Proceedings), 97–100 (2003).
- J. de Leeuw, private communication.
- C. Morris, *Annals of Statistics* **3**, 165–188 (1975).
- S.E. Fienberg, *J. Royal Stat. Soc. B* **41**, 54–64 (1979).
- D. Zelterman, *J. Am. Stat. Assoc.* **82**, 624–629 (1987).
- G. Beatty, “Salary Survey of Mathematicians and Statisticians,” in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 743–747 (1983).
- N. Cressie and T.R.C. Read, *J. Royal Stat. Soc. B* **46**, 440–464 (1984).
- K. Abe *et al.*, [arXiv:hep-ex/0505038](https://arxiv.org/abs/hep-ex/0505038).
- J.L. Rosner, *Phys. Rev. D* **70**, 094023 (2004).
- Y. Chao, P. Chang *et al.*, *Phys. Rev. Lett.* **94**, 181803 (2005); K. Abe *et al.*, [arXiv:hep-ex/0506080](https://arxiv.org/abs/hep-ex/0506080).
- K.-F. Chen *et al.*, *Phys. Rev. Lett.* **94**, 221804 (2005).
- A. Poluektov *et al.*, *Phys. Rev. D.* **70**, 072003 (2004); K. Abe *et al.*, [arXiv:hep-ex/0411049](https://arxiv.org/abs/hep-ex/0411049) and [hep-ex/0504013](https://arxiv.org/abs/hep-ex/0504013).