Note added September 18, 2023: This version of the note was deprecated on August 29 and superseded by Section 7.7 and Appendix of my lectures on the arxiv: https://arxiv.org/abs/1807.05996.

On Goodness-of-Fit Tests

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

March 6, 2016 (last updated June 11, 2016)

Abstract

This note reviews some basic concepts regarding *goodness-of-fit (g.o.f.) tests*, generally defined as hypothesis tests of a null hypothesis H_0 (typically with adjustable parameters) when no specific alternative hypothesis has been specified. Goodness-of-fit tests are quite useful, but one should be aware of the ill-posed nature of the problem, and thus how choosing a g.o.f. test implicitly selects some alternative hypotheses for good discrimination.

1 Introduction

A goodness-of-fit (g.o.f.) test is generally defined as a test of the null hypothesis (typically composite, i.e., having adjustable parameters) when an alternative hypothesis has not been explicitly specified. In a typical example, one has observed data $\vec{x}$, and the null hypothesis H_0 is that $\vec{x}$ is a random sample from a specific probability density function (pdf) $p_0(\vec{x};\vec{\alpha})$; here $\vec{\alpha}$ indicates parameters that are typically not specified in advance, but rather set to their "best-fit" values. A g.o.f. test is then used to test H_0. In this context, the given pdf p_0 is often called "the model" (short for "the statistical model"). In spite of the issues raised in this note, g.o.f. tests constitute an important step in data analysis; in fact the discussion here indicates that using more than one g.o.f. test is advisable.

The most common g.o.f. test is surely the chisquare g.o.f. test used in introductory lab classes, either for measurements of dependent variables x_i as a function of an independent variable (say current vs. voltage), or for binned data. As analyses with unbinned likelihood functions have become commonplace in HEP, the use of unbinned g.o.f. tests has increased as well. For unbinned measurements in one dimension (1D), the Kolmogorov-Smirnov (K-S) test is commonly used, probably because it has been readily available in HEP software packages (CERNLib and ROOT) for decades, and (like the chisquare test) the interpretation of the test statistic is asymptotically independent of the null hypothesis model p_0. However, as this note emphasizes, other tests may be more appropriate than the K-S test (for example when one is interested in departures from p_0 in the tails).

*cousins@physics.ucla.edu

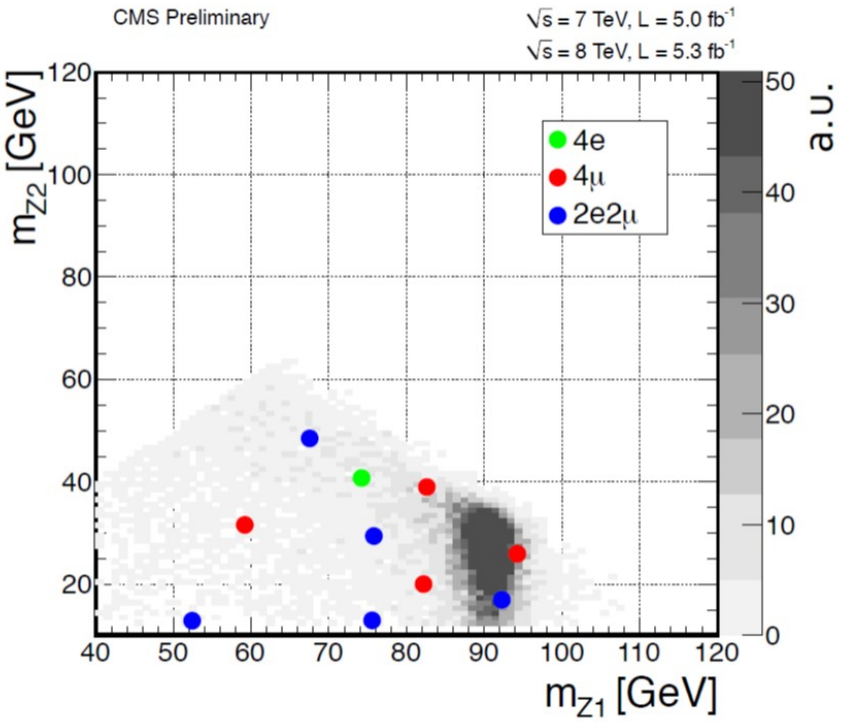

Figure 1: Scatter plot shown by Joe Incandela at the CMS talk on the discovery of the Higgs-like boson. For Higgs to ZZ^* candidates in the four-lepton final state, the horizontal axis is the mass of the higher-mass pair (Z_1), while the vertical axis is the mass of the lower-mass pair (Z_2). The large dots are the 10 events in data. The gray shading, peaking in the region around (90 GeV,30 GeV), is the expected pdf for a SM Higgs with mass of 126 GeV.

For g.o.f. tests of unbinned data in more than one dimension, there are no conventions in HEP, in spite of sporadic work in the last 15 years or more. An example of a potential application arose in the ZZ^* Higgs discovery channel, in which the events were expected to have one Z nearly on mass shell and the other off-shell. The CMS data presented at the discovery talk on 4 July 2012 included Figure 1, a scatter plot of the invariant mass of the lower-mass pair vs the higher-mass pair. In this case, the null hypothesis p_0 could be taken as SM Higgs production and decay, for which the pdf is shown in gray shading. To the eye there seemed to be fewer on-shell Z's than expected. This was noted and presumed to be a statistical fluctuation (borne out when more data was obtained). As far as I know, no unbinned 2D g.o.f. test of p_0 was attempted. One can imagine that such tests might be useful for future discovery plots in low-statistics regions.

As noted above, in common g.o.f. tests the model p_0 has parameters that are adjusted to obtain the "best fit", in which case the g.o.f. test is performed using the best-fit parameters. In such a case, it is important keep in mind that the g.o.f. test is a test of the complete model including the best-fit adjustable parameters. This is distinct from the separate inference problems of "measuring" the parameters and uncertainties on them (known to statisticians as point estimation and interval estimation, respectively).

In approaching the theory of g.o.f. tests, it is useful to understand first the the Neyman-

Pearson (N-P) theory of hypothesis testing, briefly described in Section 2. The N-P theory provides the language in which the limitations of g.o.f. tests become clear.

2 Brief outline of Neyman-Pearson hypothesis testing

As discussed in more detail in books such as Refs. [1, 2], J. Neyman and E.S. Pearson [3] considered the case when the null hypothesis H_0 is tested with an alternative hypothesis H_1 specified. E.g., if H_0 is the hypothesis that the data $\vec{x}$ are a random sample from pdf $p_0(\vec{x})$, then H_1 could be the hypothesis that the data $\vec{x}$ are a random sample from a different pdf $p_1(\vec{x})$ There is an important distinction between *simple* hypotheses, which have no adjustable parameters to be fit to the data, and *composite* hypotheses, which contain such parameters. We can however illustrate a key point using simple hypotheses.

For testing a simple "null" hypothesis H_0 against another simple "alternative" hypothesis H_1 , one constructs a *test statistic* T (function of the observed data $\vec{x}$) and rejects H_0 if T lies inside a region called the *critical region*; otherwise H_0 is accepted. The extent of the critical region is typically determined by the specification of the *significance level* of the test, which is the probability α of rejecting H_0 if it is true (the Type I error).

Conversely, H_0 is accepted if the test statistic lies inside the complementary region called the *acceptance region*. The probability β of accepting H_0 if it is false (the Type II error) depends on H_1, and is the probability given H_1 that the test statistic is in the acceptance region for H_0. The probability of accepting H_1 if it is true, $1-\beta$, is called the *power* of the test.

The Neyman-Pearson Lemma [3] states that for fixed α, the test statistic T that *maximizes the power* against H_1 is the *likelihood ratio*

$$\lambda = \mathcal{L}(H_0)/\mathcal{L}(H_1), \tag{1}$$

where $\mathcal{L}(H_0) = p_0(\vec{x})$ and $\mathcal{L}(H_1) = p_1(\vec{x})$. The critical region in the observation space $\vec{x}$ is determined from the requirement that

$$\lambda \leq \lambda_\alpha^{\rm cut}, \tag{2}$$

with $\lambda_\alpha^{\rm cut}$ adjusted to provide the desired α.

2.1 Relevance to g.o.f. tests

We see that in choosing a test statistic T for testing H_0, the power of the test depends on the alternative H_1, and the choice of T that gives maximum power also depends explicitly on H_1 in via likelihood ratio. By definition, H_1 has not been specified in a g.o.f. test. Thus the problem of choosing a "best" T for a g.o.f. test of H_0 is not well-posed! There typically exist many possible choices for T in a g.o.f. test. Some may be particularly popular (e.g. the ubiquitous chi-square g.o.f. test), but that does not mean that they are "best" in any general sense.

In fact, it follows from the Neyman-Pearson Lemma that for any choice of T used for a g.o.f. test, that choice will tend to have high power against some alternative H_1 if it happens

that T is approximately monotonic with the likelihood ratio $\lambda = \mathcal{L}(H_0)/\mathcal{L}(H_1)$, while the choice is vulnerable to having poor power against alternative hypotheses from which T bears no relationship to λ. Thus a choice of g.o.f. test statistic T picks out alternatives to H_0 (sometimes called *directions* of deviations from H_0) for which T has higher discrimination power. This is true whether or not one is aware of it!

Although the strict superiority of λ as test statistic T no longer holds with composite hypotheses, the lesson remains that the choice of T defines directions of high discrimination power.

3 Prototype problem: test of uniform density on (0,1)

A prototype g.o.f. test is the following: suppose that $\vec{y}$ consists of a set of N numbers $\{y_i\}$ between 0 and 1, and you want to test the hypothesis that they were obtained by sampling N times from a uniform density on (0,1). While this may seem to be an artificial special case (useful for testing the validity of a pseudorandom number generator), in fact many models can be re-cast into this form without loss of generality by using the *probability integral transform* as follows.

We consider the more general null hypothesis H_0 as the model in which each number x_i in the set of N numbers $\{x_i\}$ is an independent random sample from a general pdf $p_0(x)$ defined in the domain $a < x < b$, with x continuous. The probability integral transform then defines $y(x)$ via

$$y(x) = \int_a^x p_0(x')dx'. \tag{3}$$

The pdf $p(y)$ for y is then easily obtained using the transformation rule $p(y) = p(x)/|dy/dx|$ and the chain rule, showing that $p(y)$ is uniform on (0,1). Thus without loss of generality, the g.o.f. test of the null hypothesis H_0 for the model $p_0(x)$ is re-cast as the hypothesis that the set $\{y_i\} = \{y(x_i)\}$ is a random sample from the uniform density on (0,1). (If x is discrete, there are complications which are discussed in some of the cited papers.)

Over the last century, a plethora of tests have been invented to test for uniformity of $\vec{y}$ on (0,1), and hence test for $\vec{x}$ drawn from any given continuous $p_0(x)$. The book by D'Agostino and Stephens [4] contains a comprehensive discussion. The more recent article by Marhuenda et al. [5] defines and compares a plethora of tests against standard sets pf parameterized alternatives (higher or lower density near 0 or 1, or both, etc.). Subsets of tests in common use in HEP are discussed by F. James [2], and a few are implemented in ROOT.

3.1 Tests based on the empirical distribution function

The most widely used methods in HEP are based on the *cumulative distribution function* (CDF) and *empirical distribution function* (EDF). While in science "distribution" is often used synonymously with probability density function (pdf), in statistics "distribution function" is often short for *cumulative distribution function*, which is an *integral* of a pdf. Often $f(x)$ is used for a pdf, and upper case $F(x)$ is used for its CDF:

$$F(x) = \int_{-\infty}^x f(x')dx'. \tag{4}$$

Given N observed values x_i as above, the EDF is

$$F_N(x) = \frac{\text{number of observed values } \leq x}{N}. \tag{5}$$

Thus $F_N(x)$ is an increasing piecewise-constant function, starting from 0 for x less than the smallest observed x_i, increasing by $1/N$ at every observed value, and obtaining unity above the highest observed value of x_i.

If the observed $\vec{x}$ is drawn from $p_0(x)$, then we expect the EDF $F_N(x)$ to be similar to the CDF calculated from $p_0(x)$, which we call $F_0(x)$. A variety of g.o.f. tests are thus based on various definitions of the "distance" between $F_0(x)$ and $F_N(x)$. These tests include, for example, the Kolmogorov-Smirnov test (based on extremum of $|F_0(x) - F_N(x)|$), and the Cramér-von Mises family based on integrals weighted by a function $w(x)$ of the squared difference:

$$N \int_{-\infty}^{+\infty} (F_0(x) - F_N(x))^2 \, w(x) dF(x). \tag{6}$$

The unweighted case (i.e. $w(x) = 1$) corresponds to the classic Cramér-von Mises statistic. The weight function $w(x)$ can be chosen to give more power against certain deviations. For example, the Anderson-Darling (AD) test is designed to have more power against deviations at both ends of the distribution, with a weight function

$$w(x) = \frac{1}{F(x)(1 - F(x)}. \tag{7}$$

Since departures from assumed Gaussianity are often in the tails, AD is reputed to be useful for testing Gaussianity, as well as being generally useful. There are also versions that emphasize only one tail [4]. On the other hand, both the Kolmogorov-Smirnov family and the Cramér-von Mises family have variations that make the endpoints *not* special by posing the problem on a circle.

Keeping in mind the probability integral transform and the fact that Eqn. 6 is defined in terms of $dF(x)$ rather than dx, one can see that the distribution of such test statistics under the null hypothesis does not depend on the specific form p_0 in the null. Such tests are called distribution-free and are popular since standard tables can be computed and used. Nowadays, with the ability to simulate data sets and obtain the null distribution directly, it can be worth exploring the use of more powerful tests that do not have this property.

3.2 Other families of tests of uniformity on (0,1)

Refs. [4, 5] describe multiple classes of tests in addition to those based on the EDF. These include tests based on the "ordering statistics", i.e., on the ordered set of the observed points y_i (testing mean values of ith points, moments of differences $y_j - y_i$, etc.). Among the many tests, one can attempt to choose "omnibus" tests that perform reasonably well against a number of different alternatives. The conclusion of Ref. [5] is that a member of the class of tests called "Neyman smooth tests" is unique in being in the top-10 most powerful tests for all the alternatives that they considered. Neyman smooth tests and more general "smooth" tests seem to have had little use in HEP thus far, in spite of popularity in the statistics literature [6, 14]. They are tests where the alternative hypothesis is constructed

by fitting the data to a sum of Legendre polynomials. The variant that Ref. [5] studies uses Schwarz's famous Bayesian information criterion [7] to choose the order of polynomials. This spirit of using the data to construct the alternative hypothesis used in a likelihood ratio (or asymptotic equivalent) is similar to that of the saturated model discussed below for binned data.

4 Chisquare g.o.f. and variants

As noted above, Neyman and Pearson taught us that (even for simple hypotheses) the best test of the null hypothesis depends on the alternative, and hence there is no universally best g.o.f. test. Nonetheless, the ubiquity of the chisquare g.o.f. 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

$$\chi^2 = \sum_i \frac{(d_i - f_i)^2}{\sigma_i^2}, \tag{8}$$

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) to be compared with d_i. (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 note.) Since the test statistic χ^2 is a function of the random data, it is itself a random variable, and in unbounded Gaussian applications it has a probability density function [8] which is itself also frequently called chisquare. The potentially confusing ambiguity in multiple meanings is usually resolvable by context. This could also be avoided by using another name, such as S^2 or Q^2 for the left hand side of Eqn. 8, 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. 8 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.

More information regarding the chisquare test, including the generalization of Eqn. 8 to include correlations, is in the PDG RPP [8].

4.1 Gaussian chisquare g.o.f. is a likelihood ratio using the saturated model

For the same data and model as above, the likelihood for the null hypothesis H_0 is:

$$\mathcal{L}(H_0) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-(d_i - f_i)^2/2\sigma_i^2\right). \tag{9}$$

It is sometimes said that $-2\ln\mathcal{L}$ is equal to χ^2; this is clearly not correct since the former quantity has extra constant terms. Understanding how the extra terms "disappear" in the g.o.f. test (from the point of view of N-P testing) is enlightening.

Given only the null hypothesis and the data, one can use the data to invent an alternative hypothesis for which the model 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* [9].

For the Gaussian data above, the saturated model sets $f_i = d_i$, so that the likelihood of the data in the alternative hypothesis H_1 of the saturated model is (since $\exp(0) = 1$)

$$\mathcal{L}(H_1) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}}. \tag{10}$$

The ratio of the two likelihoods above is then

$$\lambda = \mathcal{L}(H_0)/\mathcal{L}(H_1) = \prod_i \exp\left(-(d_i - f_i)^2/2\sigma_i^2\right), \tag{11}$$

and thus

$$\chi^2 = -2\ln\lambda. \tag{12}$$

From this point of view, the constants in $-2\ln\mathcal{L}$ were not just ignored; they were canceled when a ratio was formed. (There are other paths to Eqn. 1 not as relevant to this note.) 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 $\mathcal{L}$.

4.2 Pearson's chisquare for binned histograms

The original chisquare variable is that of Karl Pearson in 1900, designed for histograms ("frequency tables") with multinomial-distributed bin contents, and defined as

$$\chi^2 = \sum_i \frac{(d_i - f_i)^2}{f_i}. \tag{13}$$

(In a multinomial histogram, the total number of events is fixed by design, so there is one fewer independent bin content than in Poisson data.) This is very similar to Eqn. 8, since in multinomial and Poisson data, f_i can be a proxy for σ_i^2.

4.3 Use of saturated model to construct improved chisquare g.o.f. for binned histograms

In 1983, Baker and Cousins [10] reviewed construction of likelihood ratios as in Eqn. 1 using saturated models for testing g.o.f. for Poisson and multinomial data in histograms. (They did not call them saturated models, instead citing a 1928 paper by Neyman and Pearson.) By that time, it had been widely noted that using a Poisson likelihood model for fitting typical HEP histograms cured a defect of fits using Pearson's chisquare, namely that the area under the fitted curve was not equal to the observed number of events. It was less well known how to construct a g.o.f. test consistent with such a likelihood fit. For histograms with independent Poisson-distributed bin contents d_i, the result for the g.o.f. test statistic,

$$-2\ln\lambda = 2\sum_i f_i - d_i + d_i\ln(d_i/f_i), \tag{14}$$

was called $\chi^2_{\lambda,p}$ since asymptotically its pdf tends to the chisquare distribution. This Poisson form is mentioned in the PDG's Review of Particle Properties [8]; some time ago it was 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 noted in the Baker-Cousins paper, probably the safest thing to do is to study the distribution by Monte Carlo.

Heinrich [11] 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. More generally, how to bin data is another complex problem since binning discards information on the location of events within the bin, and suppresses the ability to observe high-frequency components. Thus the robustness of results (or lack thereof) to the binning chosen should be understood and communicated.

4.4 Tests complementary to the chisquare g.o.f. test

The tests based on chisquare or asymptotic equivalents discard all or most of the information on the ordering of the deviations as well as the signs of the deviations. Hence they can easily miss a trend in the deviations that is obvious to the eye (e.g., all deviations of the same sign, or a trend due to a slope not present in the null hypothesis). Thus, authors such James [2] suggest complementing the chisquare test with a "runs test", often called the Wald-Wolfowitz runs test. Performing more than one g.o.f. test raises the issue of whether or not the results can be combined into one overall g.o.f. summary. Doing so might be useful in some contexts, but one should be aware that combining p-values is itself fraught with ambiguity, as discussed in Ref. [13].

5 Caution against absolute likelihood as g.o.f. test statistic

Occasionally one finds the recommendation to use as a g.o.f. 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 is a flawed concept: unlike the likelihood *ratio*, such a g.o.f. 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. For binned data, 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 [12] discusses the pitfalls of using the absolute likelihood as a g.o.f. statistic. For unbinned likelihoods, which are common in HEP, the problem is exacerbated. One might hope that one could just take binned g.o.f. 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.

6 Discussion, and reviews of g.o.f. in HEP

From the above discussion, it is clear that, for "g.o.f. test" defined in its purest form of no specified alternative, there is no unique "best" g.o.f. test. Therefore trying several can give some indication of how well the chosen statistics model approximates reality, and in which directions departures might be suspected (heavier tails, skewness, etc.). It is therefore useful to be aware of directions against one would like to have power, and to choose g.o.f. tests appropriately; simple toy MC simulations, such as those in the cited references, can help in this regard. For a g.o.f. test for unbinned data in one dimension, one has a variety of tests to choose among [4, 5], depending on roughly what sort of alternatives one wants power against. A recent monograph is by Thas [14]. If, on the other hand, there is a specific alternative of interest, then typically one leaves the world of generic g.o.f. tests and gains power by constructing alternative-specific likelihood ratio tests (as in the Higgs searches).

Generalization of these g.o.f. tests to higher dimensions, as in Fig. 1, remains a topic of research. Some comparisons were made by Aslan and Zech at Durham in 2002 [15], including their proposed *energy test* [16]. Williams [17] has provided a more recent review. One can of course resort to binning the data for g.o.f. tests even if the parameters are fit with an unbinned likelihood; in this case as noted above the robustness with respect to the binning chosen needs to be understood.

Acknowledgments

I thank members of the CMS Statistics Committee and participants in PhyStat meetings for past discussions. In addition Igor Volobouev pointed to Refs. [5, 14]. This note is based upon work supported by the U.S. Department of Energy under Award Number DE-SC0009937.

References

[1] A. Stuart, K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics*, Volume 2A, 6th ed., (London:Arnold, 1999), and earlier editions by Kendall and Stuart.

[2] Frederick James, *Statistical Methods in Experimental Physics*, 2nd Edition, (World Scientific, Singapore, 2006).

[3] J. Neyman and E.S. Pearson, "On the problem of the most efficient tests of statistical hypotheses", Phil. Trans. R. Soc. Series A, **231** (1933) 289, reprinted in *Breakthroughs in Statistics*, ed. by S. Kotz and N. Johnson, Vol. 1, (Springer-Verlag, 1992). See also discussions in G. Upton, I. Cook, *Oxford Dictionary of Statistics* (Oxford University Press, 2002) and Ref. [1].

[4] *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).

[5] Y. Marhuenda, D. Morales, and M. C. Pardo, "A comparison of uniformity tests," Statistics39 315 (2005), DOI:10.1080/02331880500178562 http://www.tandfonline.com/doi/abs/10.1080/02331880500178562.

[6] J.C.W. Rayner, O.Thas, and D.J. Best, *Smooth Tests of Goodness of Fit: Using R*, 2nd Ed., (New Jersey: Wiley, 2009).

[7] Gideon Schwarz, "Estimating the Dimension of a Model," Annals of Statistics 6 461 (1978) DOI:10.1214/aos/1176344136. http://projecteuclid.org/euclid.aos/1176344136

[8] K.A. Olive et al. (Particle Data Group), "The Review of Particle Physics," Chin. Phys. C, 38, 090001 (2014) and 2015 update. http://pdg.lbl.gov/2015/reviews/rpp2014-rev-probability.pdf, http://pdg.lbl.gov/2015/reviews/rpp2014-rev-statistics.pdf.

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

[10] 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.

[11] 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

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

[13] Robert D. Cousins, "Annotated Bibliography of Some Papers on Combining Significances or p-values," arXiv:0705.2209 (2007).

[14] Olivier Thas, *Comparing Distributions*, Springer Series in Statistics, (New YorK: Springer Science+Business Media, 2010). http://www.springer.com/us/book/9780387927091

[15] B. Aslan and G. Zech, "Comparison of different goodness-of-fit tests," arxiv math/0207300.

[16] B. Aslan and G. Zech, "A New class of binning free, multivariate goodness of fit tests: The Energy tests,", hep-ex/0203010.

[17] 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].