

## GOODNESS OF FIT — WITH A VIEW TOWARDS PARTICLE PHYSICS

S. L. LAURITZEN

*Department of Statistics, University of Oxford  
Oxford, United Kingdom  
E-mail: steffen@stats.ox.ac.uk*

This article reviews aspects of significance testing. Problems of detection of a specific signal with background noise from observed Poisson counts of events is used as a basic example throughout. In particular we discuss issues of using alternative test-statistics, unbinned likelihood fits, and comparing unweighted and weighted histograms. We point at the possibility of adding simultaneous confidence intervals to the statistical toolbox normally used by particle physicists.

*Keywords:* Borel scales, Cournot's principle, power-divergence statistic, non-identifiability, simultaneous inference.

### 1. Significance testing

#### 1.1. General issues

Significance testing is a well-trodden area of theoretical statistics and it seems just about impossible to say anything about this topic which has not been said many times before. For an excellent discussion of almost every corner, see for example Ref. 1 or Ref. 2. Still, significance testing is causing much controversy between statisticians and it can be hard to find two statisticians who would be in complete agreement.

In the present article we give some brief remarks which primarily serve to set the scene and identify which corner is to be explored. It also briefly indicates the multitudes of issues involved, thus explaining why it may not even be helpful to treat these different situations in a completely unified way, hence giving some rationale for the persistence of disagreement.

**Decision vs. evidence** There are two related but different types of situation which may be approached by significance testing.

In the first of these, procedures for accepting or rejecting a hypothesis are established with the purpose of using them automatically and repeatedly in a number of similar if not virtually identical situations. Such cases occur for example in industrial quality control. A decision-theoretic framework<sup>3, 4</sup> describes this situation well, the formal Neyman–Pearson theory of significance testing is both appropriate and convincing. Within this theory a linear combination of the probabilities of taking an incorrect decision (type I and II errors) is minimized, often by holding

one of these fixed at a given *level of significance*.

The second situation, which forms the basis of this article, pertains to the case where a scientific theory needs to be examined in the light of a single or few related but different experimental results. The decision-theoretic approach seems here less appropriate as the acceptance or rejection of a scientific theory rarely will be a consequence of the experiment under study, but will involve numerous other ways of gaining and incorporating scientific knowledge about the phenomenon. This situation is closer to the Fisherian way of thinking about significance tests and would rather lead to an attempt to quantify the *evidence* in the experimental result for or against the validity of a specific theory, typically in the form of a so-called *p-value* or *significance probability*.

Much of the controversy<sup>5, 6</sup> between Fisher and Neyman on issues of significance testing was centered around these contrasting situations. The difference has probably been exaggerated in the sometimes very heated debate between the two. Most researchers would agree that it would be untenable to quantify the evidence in a given, unique situation in a way that would not have reasonable properties if used repeatedly in conceptual or similar situations. Indeed, the approaches of Neyman and Fisher appear to be less different than what first meets the eye<sup>7</sup>.

In any case, the point of view adopted here to analyse problems of goodness of fit is closest to the Fisherian as this seems to have more direct bearing on the context under discussion.

**Exploration vs. confirmation** The way a significance test is used depends very much on the stage

of scientific investigation. In an exploratory phase of a scientific enquiry, significance tests can play an important role in searching for abnormalities in an experimental result, the primary aim being to *identify potentially interesting phenomena* for future exploration and the planning of further experiments. Such cases seem to need a treatment quite different from those in a confirmatory phase, where the issue is to establish conclusive evidence for a given theory which is also *convincing to others*.

**Refutation vs. validation** Significance tests are used for a variety of different purposes. In some cases they are used in a Popperian quest for refuting a scientific theory, thereby paving the way for establishing alternative and improved theories. In other cases, the objective of the significance test is to validate a certain aspect of a model, to justify assumptions needed for further analysis.

### 1.2. Paradigm

The (largely Fisherian) paradigm of significance testing used in the present article is outlined below:

- A null *hypothesis*  $H_0$  or theory is entertained or proposed and data  $X$  collected;
- A *test statistic*  $T = t(X)$  is constructed (possibly with an alternative theory in mind) in such a way that large values of  $T$  indicate deviations from  $H_0$ ;
- The *p-value*  $p = P(T \geq t_{\text{obs}} | H_0)$  is calculated, approximately or exactly;
- The *p-value* is interpreted by the fundamental principle:

*Events of small probability do not happen.*

This fundamental principle for relating probabilities to the real world has been termed *Cournot's principle*<sup>8, 9</sup>. Hence, *if  $p$  is sufficiently small*, say  $p \leq \varepsilon$ ,  $H_0$  is *untenable*. Emile Borel<sup>10, 11</sup> used the term “the single law of chance” for Cournot’s principle and set the following scales for probabilities to be small:

- l’échelle humaine:  $\varepsilon \sim 10^{-6}$
- l’échelle terrestre:  $\varepsilon \sim 10^{-15}$
- l’échelle cosmique:  $\varepsilon \sim 10^{-50}$

Modern statistical practice tends to use  $\varepsilon \sim 10^{-1}$ , but Particle Physics may well need different scales to allow for scientific progress and simultaneously prevent too many false discoveries.

Although the general issue of significance testing has a strong frequentist flavour, rules such as Cournot’s principle are also needed for subjectivist Bayesian probability to make a bridge to observable phenomena in the real world<sup>12</sup>.

### 1.3. Goodness of Fit

This term is used to describe particular types of significance tests, but it is used in many different ways and contexts<sup>13, 14</sup>, for example:

- Is a given distribution of a specified type?
- Any significance test without explicit specification of an alternative hypothesis;
- Any significance test used to validate, justify, or refute a postulated model.

To avoid the discussion to be too narrow, we will mostly adopt the latter, which conforms well with the application of Cournot’s principle.

## 2. Basic example

To avoid discussing the problems out of context, we will focus on variants of the following problem and setup, describing problems of detection of signal events in the presence of noise in the form of background events. More precisely, we consider the following:

- $X_i = x_i, i = 1, \dots, n$  are ‘binned’ counts of independent Poisson events, the  $i$ -th bin corresponding to events of mass or energy around  $m_i$ .
- The Poisson intensity  $\nu_i$  in bin  $i$  is given as

$$\nu_i = \nu_i(\theta) = \beta_i + \frac{\alpha}{\sigma} \phi \left( \frac{m_i - \mu}{\sigma} \right), \quad (1)$$

where  $\phi$  denotes the standard Gaussian density.

Here  $\beta_i$  is the intensity of *background* events whereas the second term is the intensity of the interesting *signal* events. The background intensity may depend on one or several unknown parameters  $\eta$  so  $\beta_i = \beta_i(\eta_i)$  and  $\theta = (\eta, \alpha, \mu, \sigma)$  denotes the vector of all of these parameters. The signal intensity may well be absent, corresponding to  $\alpha = 0$  and often the main issue of interest is to infer whether apparent signal events are just random artifacts.

It is quite critical what the exact status is concerning prior knowledge about the background intensity. Can the background intensity be assumed

known from other experiments and theory or must it be estimated? How can the background reasonably be modelled? Can the measurement error  $\sigma$  be considered as known or unknown? Is the position of the signal peak  $\mu$  known? The complexity of problems vary greatly according to circumstance as outlined above.

### 2.1. Standard practice

Standard practice<sup>15</sup> for tackling the situation can be briefly described as follows:

- Fit model to background intensity;
- Calculate goodness of fit statistics using either *the likelihood ratio statistic*  $G^2$

$$G^2 = -2 \log L(\hat{\theta}) \\ = 2 \sum_{i=1}^n \left\{ \nu_i(\hat{\theta}) - X_i + X_i \log \frac{X_i}{\nu_i(\hat{\theta})} \right\}$$

or its approximation, known as *Pearson's*  $\chi^2$

$$C^2 = \sum_{i=1}^n \frac{\{\nu_i(\hat{\theta}) - X_i\}^2}{\nu_i(\hat{\theta})}.$$

In some cases the latter is substituted with the *Wald statistic*

$$W^2 = \sum_{i=1}^n \frac{\{\nu_i(\hat{\theta}) - X_i\}^2}{X_i},$$

which can be computationally more convenient.

- Calculate  $p$ -values approximately or by Monte-Carlo methods.

### 3. Issues to be considered

The setup described raises a number of issues:

- Is one of the test statistics to be preferred?
- When is the  $\chi^2$  distribution appropriate for calculating  $p$ -values?
- When calculating  $p$ -values using a  $\chi^2$ -distribution, what is the appropriate number of degrees of freedom?
- If one fits the model with or without the signal component, can the difference between the two test statistics be used and what is its distribution?

Partial answers to these and other questions will be attempted in the following.

### 3.1. Power divergence statistics

It can be helpful to consider the one-parameter family of *power-divergence* statistics<sup>16</sup> given by

$$I_\lambda(X) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^n X_i \left[ \left\{ \frac{X_i}{\nu_i(\hat{\theta})} \right\}^\lambda - 1 \right]$$

for  $-\infty < \lambda < \infty$ . Provided  $\sum_i X_i = \sum_i \nu_i(\hat{\theta})$ , it follows that

$$I_1(X) = C^2, \quad \lim_{\lambda \rightarrow 0} I_\lambda(X) = G^2,$$

so the commonly used statistics mentioned above are special cases. Ref. 14 recommends  $\lambda = 2/3$ , which is 'between'  $C^2$  and  $G^2$ .

For  $\lambda = -1/2$ ,  $I_\lambda$  becomes the *Freeman-Tukey statistic*  $F^2$

$$F^2 = 4 \sum_i \left\{ \sqrt{X_i} - \sqrt{\nu_i(\hat{\theta})} \right\}^2.$$

The Freeman-Tukey statistic<sup>17</sup> is obviously based on the idea that for a Poisson variable with large mean  $\nu$ ,  $\sqrt{X}$  is approximately normally distributed:

$$\sqrt{X} \sim \mathcal{N}(\sqrt{\nu}, 1/4).$$

These statistics all have the same asymptotic  $\chi^2$  distribution under the null hypothesis, and each is optimal in some sense. My personal preference would be the likelihood ratio statistic  $G^2$ , as it is constructed to have maximal power at the most likely alternative, but it may well be a matter of taste.

For important issues it could be reasonable to calculate  $I_\lambda$  and the associated  $p$ -value for a range of different values of  $\lambda$ . It would not be desirable if the interpretation of an experiment depends critically on  $\lambda$ , so if the  $p$ -value is on different sides of the threshold for small probabilities as  $\lambda$  varies, the experiment may be considered inconclusive.

The use of  $W^2$  is mostly motivated by the convenience of computation, because its minimization is a direct weighted least squares, whereas the others might be computationally less easy to minimize. The statistic  $W^2$  is potentially less powerful than  $C^2$  against large deviations from the hypothesis, as a large and explicit signal with  $X_i > \nu_i(\hat{\theta})$  will yield

$$W_i^2 = \frac{\{\nu_i(\hat{\theta}) - X_i\}^2}{X_i} < \frac{\{\nu_i(\hat{\theta}) - X_i\}^2}{\nu_i(\hat{\theta})} = C_i^2.$$

There is some ambiguity about which of the above statistics is 'best'. Much effort has been used

to discuss which of them has a distribution closest to the  $\chi^2$ -distribution. Much of this depends both on the specific circumstances considered and how closeness is measured. Personally I would be less worried about getting an accurate calculation of the  $p$ -value than not detecting a signal because the test statistic is less powerful. Also, because effective Monte-Carlo methods are rapidly being developed, the use of the  $\chi^2$ -approximation is losing importance.

### 3.2. Is the $\chi^2$ distribution appropriate?

The derivation of the  $\chi^2$  distribution is based on the following two elements:

- For  $\nu_i$  large,  $X_i$  are approximately Gaussian  $\mathcal{N}(\nu_i, \nu_i)$ ;
- For  $\nu_i$  large, the model for the intensity  $\nu_i(\theta)$  is approximately linear in the unknown parameters within the likely area of variation of  $X_i$ . In particular, the fitting of  $\theta$  is approximately a linear least squares problem.

In the following some cases where there is trouble will be discussed.

#### 3.2.1. Unbinned fit

If  $k$  unknown parameters have been fitted based on unbinned data and  $G^2$  is calculated from binned data, the asymptotic distribution of  $G^2$  (or any of the other statistics) is *not*  $\chi^2$  with  $n - k - 1$  degrees of freedom.

Fortunately, its correct asymptotic distribution is well understood. It approximately holds<sup>18</sup> that

$$G^2 = A^2 + \sum_{j=1}^k \zeta_j B_j^2,$$

where  $A^2$  is  $\chi^2(n - k - 1)$  and independent of  $B_j^2$ ,  $j = 1, \dots, k$ , with each  $B_j^2$  distributed as  $\chi^2(1)$  and  $0 \leq \zeta_j \leq 1$ . In particular it holds (approximately) that

$$A^2 < G^2 < A^2 + \sum_{j=1}^k B_j^2,$$

where the lower bound is  $\chi^2(n - k - 1)$  and the upper bound is  $\chi^2(n - 1)$ .

This yields a simple practical way of guarding against problems of this kind: Asymptotically *the correct  $p$ -value is between those based on  $\chi^2(n - 1)$  and  $\chi^2(n - k - 1)$* . One can just calculate each of them

and this will usually be precise enough to identify whether the correct  $p$ -value is extremely small.

This result also holds for the other test statistics in the power divergence family<sup>14</sup> and for  $W^2$ .

#### 3.2.2. Parameter singularity

One specific example where the difficulty in using the  $\chi^2$  approximation is due to intrinsic non-linearity of the testing problem is exactly in the case of signal with background noise, as in (1). If the location  $\mu$  of the peak or the measurement uncertainty  $\sigma$  are not known, a singularity arises because under the null hypothesis  $\alpha = 0$ ,  $\mu$  and  $\sigma$  do not make sense.

The following method to tackle this problem has been developed by Ref. 19. First proceed as if  $\mu$  and  $\sigma$  were known, and calculate the usual test statistic for the hypothesis  $\alpha = 0$ . When  $\mu$  and  $\sigma$  are known, the hypothesis is a simple, linear hypothesis. Denote the corresponding test statistic as

$$T_{\mu, \sigma} = t_{\mu, \sigma}(X).$$

Each of these follows a  $\chi^2$  distribution under the null hypothesis. We now use the test statistic

$$T^* = \sup_{(\mu, \sigma) \in R} T_{\mu, \sigma}$$

where  $R$  is a *plausible region* for  $(\mu, \sigma)$ .

The approximate distribution of  $T^*$  is that of the maximum of related  $\chi^2$  statistics. The corresponding  $p$ -value is not known exactly, but approximate Monte-Carlo methods using the  $\chi^2$  distribution for the individual statistics have been developed<sup>19</sup>.

The method is somewhat involved, but not unusable, in particular because in many cases,  $\mu$  is known and  $\sigma$  is approximately known, so the plausible region  $R$  can be quite small.

Recently, Ref. 20 has extended and refined the method so that it becomes more accurate and more generally usable. It seems worthwhile to explore the possibility of exploiting this method.

### 3.3. Validating the model

The  $\chi^2$ -distribution used in the case just discussed would typically be the *difference* between  $G^2$  assuming only background and  $G^2$  when also the peak is fitted.

For the  $\chi^2$  distribution to be valid it is important that the model is properly established, in particular

that the background intensity is not incorrectly specified.

Thus it must at least have a non-significant  $G^2$  value when the peak is fitted, to document that the data indeed can be explained in terms of background plus peak.

In addition a careful *residual analysis* should be made to detect systematic or too large deviations from the model (1).

#### 4. Comparing weighted and unweighted histograms

In some cases the information about the background intensity  $\beta_i$  is obtained from an independent experiment with Poisson counts  $Y_i$  with intensities  $c\rho_i\beta_i$ , where  $\rho_i$  are known factors and  $c$  is a constant determining the total intensity of events in the auxiliary experiments. In other words, the auxiliary experiment has only background events, but may not have the same background rates.

It is then common to form a *weighted histogram* with weights  $W_i$  in the  $i$ th bin, where

$$W_i = Y_i/\rho_i, \quad i = 1, \dots, n$$

and compare the histogram so obtained with the histogram based on  $X_i$ , containing a potential signal peak.

The exact distribution of associated test statistics, calculated as if the weighted events were indeed proper events, cannot be described in simple terms and the asymptotic results cannot be immediately applied to this more complex situation. An alternative would be to compare the histograms with a proper significance test as follows.

Under the null hypothesis  $H_0 : \alpha = 0$ , the likelihood function in terms of the original observations  $X_i$  and  $Y_i$  is

$$\begin{aligned} L(c, \beta) &\propto \prod_{i=1}^n \beta_i^{x_i+y_i} c^{y_i} e^{-\beta_i(1+c\rho_i)} \\ &= c^{\sum y_i} e^{-\sum \beta_i - c \sum \rho_i \beta_i} \prod_{i=1}^n \beta_i^{z_i}, \end{aligned}$$

where we have let  $\beta = (\beta_1, \dots, \beta_n)$  be the unknown background intensities and  $Z_i = X_i + Y_i$  the combined number of events in bin  $i$ .

Under  $H_0$ ,  $Z_i$  and the total number of events  $T = \sum Y_i$  in the auxiliary experiment are sufficient statistics and the likelihood function is maximized by

solving the system of equations which equate their observed values to their expectations:

$$\begin{aligned} t &= \sum_{i=1}^n y_i = c \sum_{i=1}^n \rho_i \beta_i \\ z_i &= \beta_i(1 + c\rho_i), \quad i = 1, \dots, n. \end{aligned}$$

These equations can be solved iteratively, for example by using starting values  $c = \beta_1 = \dots = \beta_n = 1$  and repeating

$$\begin{aligned} c &\leftarrow \sum_{i=1}^n \rho_i \beta_i / t \\ \beta_i &\leftarrow (1 + c\rho_i) / z_i, \quad i = 1, \dots, n. \end{aligned}$$

This iteration is convergent as it can be seen to be a special instance of the algorithm known as *Iterative Proportional Scaling* or *Iterative Proportional Fitting*<sup>21</sup>. It provides maximum likelihood estimates  $\hat{c}$  and  $\hat{\beta}_i$  under the null hypothesis. The log-likelihood ratio statistic becomes

$$D = -2 \log \frac{L(\hat{c}, \hat{\beta})}{L(\hat{\nu})}$$

where  $\hat{\nu} = (\hat{\nu}_1, \dots, \hat{\nu}_n)$  is the maximum likelihood estimate under an alternative hypothesis, but many other reasonable test statistics could be used, for example the analogue of  $C^2$

$$\tilde{D} = \sum_i \frac{(x_i - \hat{\beta}_i)^2}{\hat{\beta}_i} + \sum_i \frac{(y_i - \hat{c}\rho_i\hat{\beta}_i)^2}{\hat{c}\rho_i\hat{\beta}_i},$$

or any other statistic from the power-divergence family. The  $p$ -value associated with any of these or other statistics can be calculated on the basis of the conditional distribution of the number of events, given the statistic which is sufficient under  $H_0$ , as the unknown parameters  $c$  and  $\beta_i$  do not enter into that distribution.

This distribution is very easy to simulate using the following Monte-Carlo procedure, which is a variant of Patefield's algorithm for simulating two-way contingency tables, conditional on the marginal totals<sup>22</sup>.

A simple argument shows that the conditional distribution of  $(X_k, Y_k)$ , given  $Z_i, i = 1, \dots, n$ ,  $\sum_i Y_i = t$ , and  $(X_i, Y_i) = (x_i, y_i), i = 1, \dots, k-1$  is given as

$$p(x_k, y_k) = h_k(\rho_k) \rho_k^{y_k} \frac{\binom{s - \sum_1^{k-1} x_i}{x_k} \binom{t - \sum_1^{k-1} y_i}{y_k}}{\binom{s+t - \sum_1^{k-1} z_i}{x_k + y_k}}$$

for  $x_k + y_k = z_k$ , where  $s = \sum_i X_i = \sum_i Z_i - t$  and the expressions in brackets are binomial coefficients. This yields an obvious recursion for simulating from the correct distribution of any test-statistic in cases where the number of events in each bin is limited. For large event numbers, it may be easier to use asymptotic results.

### 5. Simultaneous confidence intervals

An alternative approach to the problem of assessing whether a peak is indeed present in the model (1) uses the idea of *simultaneous inference*<sup>23</sup>. This approach initially avoids fitting a model altogether and calculates a band within which the true Poisson intensity with high probability must be. If the band is sufficiently narrow, and displays an explicit peak, it might be immediately obvious that the data are inconsistent with any reasonably smooth background model.

Using the fact that the counts in separate bins are independent, it is possible to produce a *simultaneous confidence band* for the Poisson intensity, using that if

$$P(|X_i - \nu_i| > c) = \beta$$

for every bin  $i$ , then it follows that

$$P(\max_i |X_i - \nu_i| > c) = 1 - (1 - \beta)^n.$$

Hence, if a  $1 - \alpha$  confidence band is desired, we must just choose

$$\beta = 1 - (1 - \alpha)^{1/n}.$$

This now yields a band around the observed histogram within which the proposed background intensity should fit. If this is not possible, then this can be taken as evidence either against the model for background or against absence of the peak.

### References

1. D. R. Cox, *Scandinavian Journal of Statistics* **4**, 49 (1977).
2. D. R. Cox and D. V. Hinkley, *Theoretical Statistics* (Chapman and Hall, London, 1974).
3. A. Wald, *Statistical Decision Functions* (John Wiley and Sons, New York, 1950).
4. E. Sverdrup, *Revue de l'Institut International de Statistique* **34**, 309 (1966).
5. R. A. Fisher, *Journal of the Royal Statistical Society, Series B* **17**, 69 (1955).
6. J. Neyman, *Journal of the Operations Research Society of Japan* **3**, 145 (1961).
7. J. O. Berger, *Statistical Science* **18**, 1 (2003).
8. A. A. Cournot, *Exposition de la théorie des chances et des probabilités* (Hachette, Paris, 1843).
9. G. Shafer and V. V. Vovk, *Probability and Finance: It's Only a Game* (John Wiley and Sons, New York, 2001).
10. E. Borel, *Les Probabilités et la Vie* (Presses Universitaires de France, Paris, 1943).
11. E. Borel, *Probabilities and Life* (Dover Publications, New York, 1943). English translation of Borel (1943).
12. A. P. Dawid, *Statistical Science* **19**, 44 (2004).
13. R. B. D'Agostino and M. A. Stephens, *Goodness-of-Fit Techniques* (Marcel Dekker, New York, 1986).
14. T. R. C. Read and N. Cressie, *Goodness-of-Fit Statistics for Discrete Multivariate Data* (Springer-Verlag, New York, 1988).
15. G. Cowan, *Statistical Data Analysis* (Clarendon Press, Oxford, 1998).
16. N. Cressie and T. R. C. Read, *Journal of the Royal Statistical Society, Series B* **46**, 440 (1984).
17. M. F. Freeman and J. W. Tukey, *Annals of Mathematical Statistics* **21**, 607 (1950).
18. H. Chernoff and E. L. Lehmann, *Annals of Mathematical Statistics* **25**, 579 (1954).
19. R. B. Davies, *Biometrika* **74**, 33 (1987).
20. C. Ritz and I. M. Skovgaard, *Biometrika* **92**, 507 (2005).
21. J. N. Darroch and D. Ratcliff, *Annals of Mathematical Statistics* **43**, 1470 (1972).
22. W. M. Patefield, *Applied Statistics* **30**, 91 (1981).
23. R. Miller, *Simultaneous Statistical Inference*, 2nd edn. (Springer-Verlag, 1981).