

# **Goodness of Fit with a view to Particle Physics**

**Steffen Lauritzen, University of Oxford**

Jesus College, Oxford, September 2005

# General issues of significance testing

- Decision vs. evidence
- Justification vs. discovery
- $p$ -values vs. significance levels
- Goodness of fit for validation or the opposite.

# Paradigm

1. A null *hypothesis*  $H_0$  or theory is entertained or proposed and data  $X$  collected
2. A *test statistic*  $T = t(X)$  is constructed (possibly using an alternative theory) in such a way that large values of  $T$  indicate deviations from  $H_0$
3. The *p-value*  $p = P(T \geq t_{\text{obs}} \mid H_0)$  is calculated, approximately or exactly
4. The *p-value* is interpreted according to *Cournot's principle*:

*Events of small probability do not happen.*

Hence, *if  $p$  is sufficiently small*, say  $p \leq \varepsilon$ ,  *$H_0$  is untenable.*

## Borel's scales

Emile Borel (little after 1900) set the following scales for small probabilities:

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

It is interesting that modern statistical practice rather uses  $\varepsilon \sim 10^{-1}$ .

If a standard is needed in Particle Physics, it may be another?

# Goodness of Fit

This term is used in many different ways and contexts:

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

We will adopt the latter here.

## Basic Poisson model

To avoid discussion out of context, we will focus on a particular type of problem:

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

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

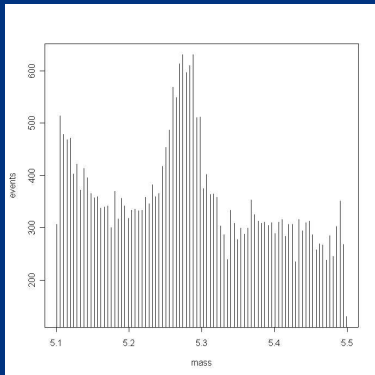
Here  $\beta_i$  is the intensity of *background* events whereas the second term is the intensity of the interesting *signal* events.

## Specific issues

- Are the signal events artifacts, i.e. is  $\alpha = 0$ ?
- How can the background intensity be modelled?
- Is the background intensity known or must it be fitted?
- Is the measurement error  $\sigma$  known or unknown?
- Is the position of the signal peak  $\mu$  known?

Complexity of problems vary according to circumstance.

# Specific example



*B* meson decay. Example provided by Jonas Rademacher.

## Standard practice

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

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

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

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

## Issues to be addressed

- 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 are the appropriate 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.

## Power divergence statistics

This one-parameter family of test statistics (Cressie and Read 1984) is 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 holds that

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

For  $\lambda = -1/2$  we get the *Freeman-Tukey statistic*  $F^2$

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

and Read and Cressie (1988) recommend  $\lambda = 2/3$ , which is 'between'  $X^2$  and  $G^2$ .

The statistics all have the same asymptotic  $\chi^2$  distribution and they all make sense in some way.

The Freeman-Tukey statistic (Freeman and Tukey 1950) is obviously based on the idea that for a Poisson variable with large mean  $\nu$ , it approximately holds that

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

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

The derivation of the  $\chi^2$  distribution is based on 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.

## Unbinned fit

If  $k$  unknown parameters have been fitted based on *unbinned* data,  $G^2$  (or any of the others) is *not*  $\chi^2$  with  $n - k - 1$  degrees of freedom.

Instead it approximately holds (Chernoff and Lehmann 1954) that

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

where  $A^2$  is  $\chi^2(n - k - 1)$ ,  $B_j^2$  are  $\chi^2(1)$  and  $0 \leq \zeta_j \leq 1$ , all random variables being independent.

Thus *the correct p-value is between those based on  $\chi^2(n - 1)$  and  $\chi^2(n - k - 1)$ .*

## Parameter singularity

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.

$$\nu_i = \beta_i + \frac{\alpha}{\sigma} \phi \left( \frac{m_i - \mu}{\sigma} \right).$$

A method developed by Davies (1987) is as follows:

First proceed as if  $\mu$  and  $\sigma$  were known and calculate the usual test statistic for the hypothesis  $\alpha = 0$ . Let

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

each of which follow 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  $p$ -value for this test statistic can now be calculated approximately by Monte-Carlo methods using the  $\chi^2$  distribution for the individual statistics.

The method is somewhat involved, but not unusable, in particular because in many cases,  $\mu$  is known and  $\sigma$  is approximately known.

## Validating the model

The  $\chi^2$ -distribution used in the previous example 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

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

is valid.

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 basic model.

## Simultaneous confidence intervals

Using the fact that the counts in separate bins are independent, it is possible to produce a *simultaneous confidence band* (Miller 1981) for the Poisson intensity, using that if

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

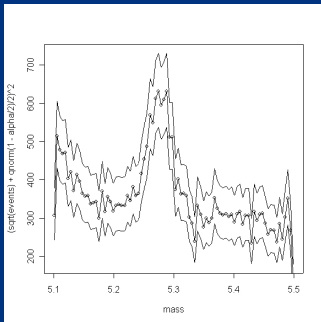
for every bin  $i$ , then it holds that

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

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

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

# Simultaneous confidence band



Simultaneous 99% confidence band using Gaussian approximation to  $\sqrt{X}$ .

## References

- Chernoff, H. and Lehmann, E. L. (1954). The use of maximum likelihood estimates in  $\chi^2$  tests for goodness of fit. *Annals of Mathematical Statistics*, **25**, 579–86.
- Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, **46**, 440–64.
- Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **74**, 33–42.
- Freeman, M. F. and Tukey, J. W. (1950). Transformations related to the angular and the square-root. *Annals of Mathematical Statistics*, **21**, 607–11.
- Miller, R. (1981). *Simultaneous Statistical Inference*, (2nd

edn). Springer-Verlag.

Read, T. R. C. and Cressie, N. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag, New York.