p-values for Model Evaluation

Frederik Beaujean[1*], *Allen Caldwell*[1], *D. Kollár*[2], *K. Kröninger*[3]
[1]Max-Planck-Institut für Physik, München, Germany
[2]CERN, Geneva, Switzerland
[3]II Physikalisches Institut, Universität Göttingen, Germany
* Corresponding author

Abstract
A quantitative procedure to decide whether a model provides a good description of data is often based on a specific test statistic and a p-value summarizing both the data and the statistic's sampling distribution. We provide a Bayesian motivation for using p-values in the goodness-of-fit problem with no explicit alternative models considered. Some typical pitfalls encountered with common statistics are reviewed for Poisson and Gaussian uncertainties. Finally, we present a new test statistic for ordered Gaussian data, the *runs* statistic.

1 Introduction

Progress in science is the result of an interplay between model building and the testing of models with experimental data. In this paper, we discuss model evaluation and focus primarily on situations where a statement is desired on the validity of a model without explicit reference to other models. We introduce different *discrepancy variables* [1] (an extension of classical test statistics to allow possible dependence on unknown (nuisance) parameters) for this purpose and define p-values based on these. p-values have been discussed extensively in the literature [2, 3], in particular also at previous PHYSTAT conferences [4, 5].

Following a Bayesian motivation for p-values in Section 2, we introduce an example fit problem in Section 3. Next, we explore some common pitfalls in p-value calculations with Gaussian uncertainties in Section 4 and study the usefulness of p-values despite approximations for the Poisson case in Section 5. Finally, we present a new discrepancy variable based on runs for ordered Gaussian data in Section 6.

In general, any discrepancy variable which can be calculated for the observations can be used to define a p-value. We use $R(\vec{x}|\vec{\theta}, M)$ and $R(\vec{D}|\vec{\theta}, M)$ to denote discrepancy variables evaluated with a possible set of observations $\vec{x}$ for given model M and parameter values $\vec{\theta}$, and for the observed data, $\vec{x} = \vec{D}$, respectively. To simplify the notation, we will occasionally drop the arguments on R and use R^D to denote the value of the discrepancy variable found from the data set at hand. R can be interpreted as a random variable, whereas R^D has a fixed value.

Assuming that smaller values of R imply better agreement between the data and model predictions, the definition of p (for continuous frequency distribution of R) is written as:

$$p = \int_{R>R^D} P(R|\vec{\theta}, M)\mathrm{d}R \ . \tag{1}$$

The basic fact used in interpreting p-values is the following: under M with the value of $\vec{\theta}$ fixed before data is analyzed, p is a random variable with uniform distribution on $[0, 1]$; i.e. $p \sim U[0, 1]$.

However, in most practical examples the value of $\vec{\theta}$ is not fixed a-priori. The choice of parameter values from fitting the data set, $\vec{\theta}_{fit}$, affects the distribution $P(R|\vec{\theta}, M)$ typically in an unknown way[1]. Hence, a p-value based on the distribution $P(R|\vec{\theta} = \vec{\theta}_{fit}, M)$ assuming fixed $\vec{\theta}$ is in general not $U[0, 1]$. This introduces confusion in interpreting p, as Berger put it: "being $U[0, 1]$ defines a proper p-value, allowing for its common interpretation across problems. Statistical measures that lack a common interpretation across problems are simply not very useful" [2].

[1]The one notable exception, χ^2, is discussed in Sec. 4

2 Bayesian motivation for p-values

When using a p-value to claim discovery of, say, a new particle at a high energy physics experiment, it is indispensable to take systematic effects of the detector into account. However, the correct distribution of the data fluctuations, including systematic effects, is often not known and best guesses are used. These guesses introduce a degree of subjectivity that affect p, no matter if the distribution of R is estimated from simulating data sets or approximated using simple closed-form expressions as in Sections 4, 5. This inherent vagueness should always be remembered when interpreting p-values.

We contend that the (frequentist) use of p-values for evaluation of models is essentially Bayesian in character. Assume that the p-value probability density for a good model, M_0, is uniform, $P(p|M_0) = 1$, and for poor models, $M_i \;\; (i = 1, k)$, can be represented by

$$P(p|M_i) \approx \lambda_i e^{-\lambda_i p} \tag{2}$$

where $\lambda_i \gg 1$ so that the distribution is strongly peaked at 0 and approximately normalized to 1. Using Bayes' theorem, we update the prior *degree-of-belief* (DoB) in model M_0, $P_0(M_0)$, to the posterior DoB, $P(M_0|p)$, after finding a particular p-value

$$P(M_0|p) = \frac{P(p|M_0)P_0(M_0)}{P(p|M_0)P_0(M_0) + \sum_{i=1}^{k} P(p|M_i)P_0(M_i)} \;\; . \tag{3}$$

If we take all models to have similar prior DoBs, $P_0(M_0) \approx P_0(M_i)$, then

$$P(M_0|p) \approx \frac{P(p|M_0)}{P(p|M_0) + \sum_{i=1}^{k} P(p|M_i)} \;\; . \tag{4}$$

In the limit $p \to 0$, we have

$$P(M_0|p) \approx \frac{1}{1 + \sum_{i=1}^{k} \lambda_i} \ll 1 \;\; , \tag{5}$$

while for $\lambda_i p \gg 1 \;\; \forall i$ we have $P(M_0|p) \approx 1$, ruling out any alternative to M_0.

Although this formulation in principle allows for a ranking of models, the vague nature of this procedure indicates that any model which can be constructed to yield a reasonable p-value should be retained. Effectively, the posterior $P(M_0|p)$ depends on the data only indirectly through $p = p(D)$. Clearly, if p is not a sufficient statistic, valuable information is not used.

3 Example fit problem

In the following, we test the usefulness of different discrepancy variables R by looking at the respective p-value distributions for an example typical of high energy physics. We first consider a data set which consists of a background known to be smoothly rising and, in addition to the background, a possible signal. This could correspond for example to an enhancement in a mass spectrum from the presence of a new resonance. The width of the resonance is not known, so that a wide range of widths must be allowed for. Also, the shape of the background is not well known. We do not have an exhaustive set of models to compare and want to look at GoF's for models individually to make decisions; direct model comparison is outside the scope of this paper. In Sections 4 and 6 we model fluctuations of the data relative to expectations with Gaussian distributions. We also consider the same problem in Section 5 with small event numbers, so that Poisson statistics are appropriate. These examples are discussed in more detail in [6, 7]. Typical data sets are shown in Fig. 1 for $N = 25$ data points (Poisson: bin contents), generated from the function

$$f(x_i) = A + B\,x_i + C\,x_i^2 + \frac{D}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \;\; , \tag{6}$$

with parameter values ($A = 0$, $B = 0.5$, $C = 0.02$, $D = 15$, $\sigma = 0.5$, $\mu = 5.0$). The y_i are generated from $f(x_i)$ as $y_i = f(x_i) + z_i$ where z_i is sampled according to $\mathcal{N}(0, 4)$. We fit the following four models to the data:

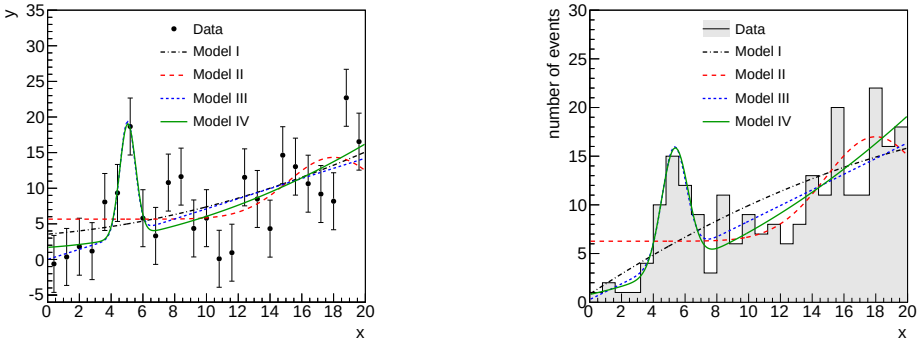

Fig. 1: Example data set for the case $N = 25$ with Gaussian (left) and Poissonian (right) fluctuations. The fits of the four models are superimposed on the data.

I. quadratic: $\vec{\theta} = (A, B, C)$ corresponding to the "standard model";

II. constant + Gaussian: $\vec{\theta} = (A, D, \mu, \sigma)$;

III. linear + Gaussian: $\vec{\theta} = (A, B, D, \mu, \sigma)$;

IV. quadratic + Gaussian: $\vec{\theta} = (A, B, C, D, \mu, \sigma)$ corresponding to the true function (6).

4 Revisiting the Gaussian case

For uncorrelated data assumed to follow Gaussian probability distributions relative to the model predictions, the discrepancy variable considered most often in high energy physics is the classic χ^2

$$R_G = \chi^2 = \sum_{i=1}^{N} \frac{\Big(y_i - f(x_i|\vec{\theta}, M)\Big)^2}{\sigma_i^2} \tag{7}$$

R_G is both fast to evaluate and, at first sight, easy to turn into a p-value (using ROOT's `TMath::Prob(...)`). However, in practical examples the conditions to do so are usually not satisfied. The frequency distribution of R_G is the celebrated χ^2-distribution with $(N - \dim\vec{\theta})$ degrees-of-freedom (DoF) if [8]

- the data fluctuations are Gaussian and the σ_i's are independent of the parameters,
- the function to be compared to the data depends linearly on the parameters, and
- the parameters are chosen such that R_G is at its *global* minimum.

In our example, the above conditions may be violated in two ways:

1. by construction: the predictions $f(x_i|\vec{\theta}, M)$ from (6) are non-linear in $\vec{\theta}$, or
2. for numerical reasons: the likelihood $P(\vec{x}|\vec{\theta}, M) \propto \exp(-R_G/2)$ has several modes.

Multimodality gives rise to technical issues: when using a gradient-based optimization algorithm like MIGRAD from the MINUIT package [9], it is critical to choose a good starting point in parameter space. If best-fit parameter values $\vec{\theta}_{loc}$ are chosen at a local minimum rather than at the global minimum $\vec{\theta}_{glob}$ such that $R_G(\vec{D}|\vec{\theta}_{loc}, M) > R_G(\vec{D}|\vec{\theta}_{glob}, M)$, then using the χ^2-distribution to turn $R_G(\vec{D}|\vec{\theta}_{loc}, M)$ into a p-value yields a p-value distribution that peaks at $p = 0$, significantly deviating from $p \sim U[0, 1]$. The physicist performing a fit often believes to have a good idea "where the best-fit parameters ought to be"

and takes that as a starting point. However, we have seen in our fits that even when we know the true value of $\vec{\theta}$, $\vec{\theta}_{true}$, starting MIGRAD there for some data sets doesn't lead to the global maximum. We thus recommend a different procedure: if there is any concern that several modes may exist, one should use a Monte Carlo sampling method (we used the implementation of the Metropolis-Hastings algorithm in BAT [7]) to explore the parameter space, and take the best-fit parameters encountered in the sampling to seed MIGRAD.

A further complication may arise when a Bayesian fit with non-uniform priors on $\vec{\theta}$ is performed; the maximum of the posterior doesn't coincide with the minimum of R_G. Choosing parameter ranges in a maximum-likelihood fit is (at least at the numerical level) equivalent to performing a Bayesian fit with uniform priors with compact support. Obviously different priors can lead to a different resulting p-value distribution. In our example we have used hypercubes in parameter space of a different size. Using the smaller volume, which contains $\vec{\theta}_{true}$, the distribution of p is biased towards $p = 0$ with a maximum deviation from uniformity of about 20%. On the other hand, with a much larger volume the distribution is now biased towards $p = 1$, again with a maximum deviation from uniformity of about 20%. The discrepancy between the two stems from the fact that the global optimum is in some cases outside of the smaller volume. For the larger volume, $p \sim U[0, 1]$ is expected, since the fit function is non-linear in $\vec{\theta}$. For plots and further details see Ref. [6], Chapter 4.

5 Revisiting the Poisson case

Similar to the previous section, we now fit the models I-IV to a histogram . We proceed in analogy to Baker&Cousins [10] and limit the discussion to three common discrepancy variables to judge the GoF. Suppose N is the number of bins, $\nu_i = \nu_i(\vec{\theta}, M)$ is the expected number of events in bin i, and n_i is the observed number of events. Then we define

$$R_P = \text{Pearson's } \chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i}, \quad R_N = \text{Neyman's } \chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{n_i} \quad . \tag{8}$$

In cases where $n_i = 0$, practitioners of this approach set $n_i = 1$ in R_N's denominator to avoid divergence. Sometimes bins with $n_i = 0$ are ignored, which can lead to very misleading results since finding $n_i = 0$ is valuable information. Finally, we have the log likelihood ratio (sometimes called Cash statistic [11])

$$R_C = 2 \log \frac{P(\vec{x}|\nu_i = n_i)}{P(\vec{x}|\nu_i = \nu_i(\vec{\theta}))} = 2 \sum_{i=1}^{N_b} \left[\nu_i - n_i + n_i \log \frac{n_i}{\nu_i} \right] \quad , \tag{9}$$

where $P(\vec{x}|\nu_i)$ is the product of Poisson probabilities for each bin. Asymptotically, i.e. for $n_i \gg 1 \; \forall i$, R_P, R_N and R_C are χ^2-distributed with $(N - \dim \vec{\theta})$ DoF. But for "finite sample size ... general results are lacking" [10], and that is precisely the case of interest in physics. The situation is aggravated in our example, as some bins typically have few or even no events, see Figure 1. We have generated 10000 data sets with Poissonian fluctuations from (6) to estimate the p-value frequency distributions across 20 bins in Figure 2. For each data set and discrepancy variable R, we calculated a p-value using the χ^2-distribution with the value of $\vec{\theta}$ chosen to minimize the respective R.

Models I and II are ruled out by each R. By construction, models III and IV are very similar. R_P doesn't distinguish well between the two, while R_N's and R_C's distributions peak for III, but look more uniform for IV. If one is interested in setting a frequentist limit at the 95% confidence level, then the first bin of each distribution in Figure 2 is the relevant one. For model IV, the densities (R_P: 0.58, R_N: 1.85, R_C: 1.35) differ significantly from the desired value of 1, given a statistical uncertainty of $\mathcal{O}(5\%)$ obtained from a binomial model with uniform prior on the chance of ending up in this first bin. We also display model IV (true model) with the true parameter values to get a feeling for the quality of the approximation in using the asymptotic χ^2 distribution; here, only N DoF are used. R_N has a worrisome peak at $p = 0$, while R_P and R_C are fairly uniform. Based on this numerical study, we discourage the use

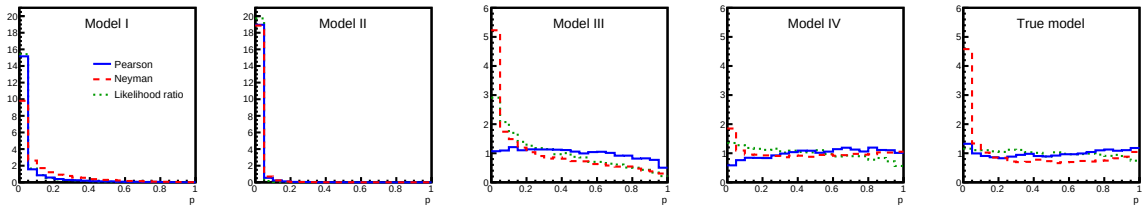

Fig. 2: p-value distributions based on R_P, R_N and R_C using the χ^2-distribution with $N - n$ degrees of freedom, where n is the number of fitted parameters.

of Neyman's χ^2 and recommend Pearson's χ^2 and the likelihood ratio, bearing in mind the inaccuracy due to finite sample size.

6 Runs statistic

When defining the p-value based on a statistic T, the usually multi-dimensional data is compressed into a single number T. Often, T is, by construction, insensitive to certain features of the data. While this is beneficial in some cases, it often represents a short coming. Returning to the Gaussian example of Section 4, R_G is merely a measure of the average distance of a single observation to its predicted value. But what if one is interested in checking that the *sequence* of data points agrees with model predictions? In high energy physics, data is often available in a 1-D ordering, e.g., the cross section y for a set of energies $x_i, i = 1 \ldots N$. Suppose there is a peak in the distribution $y(x)$ that is not predicted by the standard model. If the peak is localized in the sense that only a few y_i exceed the standard model predictions, then even for moderately large N, R_G will not detect a mismatch as the average deviation to the standard model is typically within accepted levels.

Recently, we have proposed the runs statistic [6,11] as a companion to R_G in order to gain sensitivity to local clustering of observations in the case of independent, Gaussian distributed samples. Assume the ordered set of N observations $\{(x_i, y_i)\}$ is partitioned into subsets containing the success and failure runs (defined as sequences of consecutive y_i above or below the expectation from the model, $f(x_i|\vec{\theta}, M)$, respectively).

Let A_j denote the subset of the observations of the j^{th} success run. The weight of the j^{th} success run is then taken to be

$$\chi^2_{\mathrm{run},j} = \sum_{i=j_1}^{j_1+N_j-1} \frac{\Big(y_i - f(x_i|\vec{\theta}, M)\Big)^2}{\sigma_i^2} \tag{10}$$

where the sum over i covers the $(x_i, y_i) \in A_j$ and N_j is the length of the run. The discrepancy variable is then the largest weight of any success run: $R_{sr} \equiv \max_j \chi^2_{\mathrm{run},j}$.

The exact frequency distribution of R_{sr}, used to define the p-value, $p = P(R_{sr} > R^D_{sr}|N)$, is given in [11] for the case when $(\vec{\theta}, M)$ are fully specified (no fitting). A similar discrepancy variable can be defined for failure measurements, R_{fr}.

To illustrate the definition we present a simple example. Suppose $N = 5$ observations at x positions $(1, 2, 3, 4, 5)$ with standardized residuals $(y_i - f(x_i|\vec{\theta}, M))/\sigma_i$ given by $(0.3, -0.1, -0.8, 0.4, 0.2)$. Then there are two success runs $A_1 = \{(1, 0.3)\}$, $A_2 = \{(4, 0.4), (5, 0.2)\}$ and we find $R_{sr} = 0.16 + 0.04 = 0.2$ due to the second run. Similarly, for the single failure run, $R_{fr} = 0.65$.

For the example (6) from Section 3, the joint distribution of p-values for success and failure runs based on $P(R_{sr} > R^D_{sr}|N)$ for models I and IV is shown in 3. A cut in two dimensions allows for a clean separation, while from the marginal 1-D distributions the different models are much harder to separate.

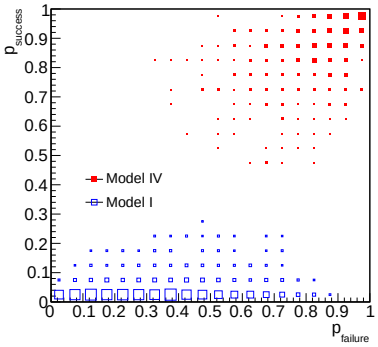

Fig. 3: Joint distribution of the p-values for success and failure runs. Bins with probability less than $3.5 \cdot 10^{-3}$ have been excluded from the plot for the purpose of clarity.

7 Discussion

In the examples which we studied it has become apparent that it is difficult to construct a p-value which is $U[0,1]$ when parameters are fitted. On the one hand, this is due to approximations of a discrepancy variable's frequency distribution. On the other hand, the numerical fitting procedure may have an impact if it finds only a local optimum. In the discussion at PHYSTAT2011, Kyle Cranmer stressed that he would prefer that a quantity defined as in (1) with non-uniform distribution should not be called p-value at all to avoid confusion in its interpretation. However, in our opinion $p \nsim U[0,1]$ is tolerable, as p-values should not be used in a simple accept/reject fashion, but merely as guidance as to whether a better model has to be constructed to explain the data. After all, the p-values displayed in Figure 2 serve that purpose: a physicist starting with model II only would be well advised to look further, and hopefully arrive at model III or IV.

References

[1] A. Gelman, X. L Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–759, 1996.

[2] M. J. Bayarri and James O. Berger. P values for composite null models. *Journal of the American Statistical Association*, 95(452):1127–1142, 2000.

[3] Mark J. Schervish. P values: What they are and what they are not. *The American Statistician*, 50(3):203–206, 1996.

[4] Luc Demortier. P values and nuisance parameters. In *PHYSTAT-LHC Workshop on Statistical Issues for LHC Physics" (2007)*, pages 23–33, 2007.

[5] I. V. Narsky. Goodness of fit. *PHYSTAT2003*, pages 70–74, 2003.

[6] F. Beaujean, A. Caldwell, D. Kollár, and K. Kröninger. p-values for model evaluation. *Physical Review D*, 83(1):012004, 2011.

[7] A. Caldwell, D. Kollár, and K. Kröninger. BAT- the Bayesian Analysis Toolkit. *Computer Physics Communications*, 180(11):2197–2209, 2009.

[8] W. T. Eadie, D. Drijard, F. E. James, M. Roos, and B. Sadoulet. *Statistical Methods in Experimental Physics*. North-Holland, Amsterdam, 1971.

[9] F. James. MINUIT - a system for function minimization and analysis of the parameter errors and correlations. *Computer Physics Communications*, 10:343–367, 1975.

[10] Steve Baker and Robert D. Cousins. Clarification of the use of χ^2 and likelihood functions in fits to histograms. *Nuclear Instruments and Methods in Physics Research*, 221(2):437–442, 1984.

[11] Frederik Beaujean and Allen Caldwell. A test statistic for weighted runs. *arxiv:1005.3233*, 2010.