CDF/MEMO/STATISTICS/PUBLIC/8023
Version 1.01
January 5, 2006

# Simple Facts about $P$-Values

Craig Blocker[a], John Conway[b], Luc Demortier[c], Joel Heinrich[d], Tom Junk[e], Louis Lyons[f], Giovanni Punzi[g]

(CDF Statistics Committee)

[a] *Brandeis University, Waltham, Massachusetts 02254*
[b] *University of California, Davis, Davis, California 95616*
[c] *The Rockefeller University, New York, New York 10021*
[d] *University of Pennsylvania, Philadelphia, Pennsylvania 19104*
[e] *University of Illinois, Urbana, Illinois 61801*
[f] *University of Oxford, Oxford OX1 3RH, United Kingdom*
[g] *I.N.F.N.-Sezione di Pisa, Largo B. Pontecorvo 3, 56100 Pisa, Italy*

## 1 Introduction

Probably one of the most common questions addressed to the statistics committee concerns the calculation of significances, i.e. "$p$-values". In early 2003, the committee decided to start compiling a list of "simple facts about $p$-values," with the hope of providing some useful guidelines for the correct interpretation and handling of these objects. This note is the final result of the compilation; for ease of reference, the "simple facts" are presented as answers to a set of "simple questions".

## 2 What are $p$-values?

For a given hypothesis, a $p$-value expresses the probability of obtaining data at least as extreme as ours. For example, if the hypothesised distribution ("null hypothesis") is a Poisson of mean 2.9, we have observed 10 events, and we wish to investigate the possibility that the Poisson mean could be *higher* due to contributions from a signal process, then the $p$-value is defined as the following null hypothesis tail probability:

$$\sum_{n=10}^{\infty} \frac{2.9^n \, e^{-2.9}}{n!} \tag{2.1}$$

Small $p$-values imply that the data is unlikely for the given model (and the deviation is in the 'interesting' direction).

In the example above we used the number of events $n$ to test the null hypothesis: $n$ is called the test statistic. In this simple case, the test statistic is equal to the data variable. In cases involving more than one data variable, the test statistic can be a function of all the data variables or only a subset of them. A well-known example is the chisquared test, where the test statistic is the sum of the squares of the deviations between data and theory, and the corresponding $p$-value is the tail area under a chisquared distribution.

# 3 How are $p$-values distributed?

In ideal situations, with a continuous test statistic and no systematic uncertainties, and assuming the null hypothesis $H_0$ is correct, $p$-values will be uniformly distributed between 0 and 1. In contrast, when the data is discrete rather than continuous (e.g. for a Poisson distribution, where the data values are only integers), the possible $p$-values are also discrete, are not uniformly spaced in $p$, and do not have equal weights. The $p$-distribution cannot be uniform in the sense of $dn/dp$ being constant. However it is 'as uniform as possible' for a discrete distribution, with $\mathbb{P}\text{r}(p \leq c \,|\, H_0) = c$ if $c$ is the location of an observable $p$-value, and $\mathbb{P}\text{r}(p \leq c \,|\, H_0) < c$ otherwise.

A $p$-value with the property that $\mathbb{P}\text{r}(p \leq c \,|\, H_0) = c$ is called exact. A $p$-value with the property that $\mathbb{P}\text{r}(p \leq c \,|\, H_0) \leq c$ is called valid.

# 4 How can $p$-values be used?

$P$-values are used for testing different types of hypotheses:

1. Significance tests
   These are tests where one tries to reject a given hypothesis, called null hypothesis. In a search for the Higgs boson for example, the null hypothesis is that the data arose from background processes only. The $p$-value should be defined in such a way that it is small if a Higgs signal is present. A discovery would be claimed if the $p$-value is small enough, say below a fixed significance level $\alpha$. This level $\alpha$ is also referred to as the Type-I error rate, for the following reason. In a large number of independent significance tests using the same value of $\alpha$, and for which the null hypothesis $H_0$ is true and the $p$-values are exact (see section 3), the fraction of tests which (incorrectly) reject $H_0$ will tend to $\alpha$ as the number of tests increases.

   In a significance test one reports both $\alpha$ and the $p$-value. The interpretation of the $p$-value is that it is equal to the smallest Type-I error rate for which the null hypothesis would be rejected. Note that the $p$-value itself is not a Type-I error rate, because the latter must be fixed *before* doing the experiment.

2. Goodness-of-fit tests
   In contrast with significance tests, goodness-of-fit tests are meant to evaluate the

evidence *in favor* of a hypothesis. Suppose for example that one is interested in checking how well a Monte Carlo simulation models the data in a control sample. One way to do this is to compare one or more Monte Carlo histograms to their data counterparts by performing chisquared test(s). If none of the resulting $p$-values is too small, one would fail to reject the simulation as a valid model for the data. Note however that the size of these $p$-values is *not* a measure of the agreement between the simulation and the data. This is because, under the null hypothesis that the Monte Carlo simulation is a correct model for the data, the $p$-values are uniformly distributed between zero and one. Hence the terminology that "we failed to reject the null hypothesis," rather than "we accept the null hypothesis."

3. Fixed-level tests
   These tests are similar to significance tests: one defines a significance level $\alpha$ before the measurement is performed, and then sees whether the data are consistent with the hypothesis at this level, by checking whether $p \leq \alpha$. As already mentioned, the expected rate of 'Errors of the First Kind' (i.e. how often the hypothesis is rejected when it is in fact true) is $\alpha$; it is not the $p$-value. Fixed-level tests are particularly useful in situations where a given test is repeated a large number of times. When selecting events for a cross section measurement for instance, each event is tested with the hypothesis that it was produced by the signal process. This is done by subjecting the event to a set of selection cuts. The efficiency of these cuts is then equal to $1 - \alpha$, where $\alpha$ is the probability of rejecting a "good" event. In this situation there is no need to report each individual $p$-value, hence the distinction from significance tests.

# 5 What are $p$-values *not* meant to measure?

A $p$-value measures the probability of observing *data* at least as extreme as ours, assuming the hypothesis is true:

- It does *not* measure the probability that the *hypothesis is true*, based on our data (see example below). This is an example of the difference between the probability of data, given a hypothesis; and the probability of the hypothesis, given the data.
- It also does *not* measure the probability of rejecting the null hypothesis when it is in fact true. This type-I error probability is given by $\alpha$, not $p$.

In summary, it is wrong to think that:

1. *Wrong:* "If $p = 7\%$, the probability that the hypothesis is in fact correct is 7%." The probability of a hypothesis being true requires Bayes' theorem together with a choice of prior probability for the hypothesis.
2. *Wrong:* "If $p = 3\%$, the probability of rejecting a true hypothesis is 3%." This probability is determined by $\alpha$, not $p$.

A simple example illustrating that $p$-values are *not* hypothesis probabilities: Consider a particle identifier for pions, using $dE/dx$ or the Cherenkov ring angle. For the pion hypothesis, the $p$-value distribution should be flat between 0 and 1:

$$f(p \mid \pi) \;=\; 1. \tag{5.1}$$

Now suppose that muons result in the following $p$ distribution:

$$f(p \mid \mu) \;=\; 1 - 0.1 \times (p - 0.5), \tag{5.2}$$

which is not too different from that for pions (because the pion and muon masses are similar), but is slightly more peaked at small $p$. In a sample of tracks with equal numbers of pions and muons, tracks with $p$ close to 0.1 will have a pion to muon ratio of $f(0.1 \mid \pi)/f(0.1 \mid \mu) = 1/1.04$. In other words, any track with $p$ close to 0.1 in that sample will be a pion with probability 1/2.04, which is quite different from 0.1. With a perhaps more realistic particle composition of 100 times more pions than muons, the pion to muon ratio for tracks with $p$ close to 0.1 becomes 100/1.04, and the "pion hypothesis probability" for a given track will be 100/101.04, even more different from the $p$-value of 0.1. What we have actually done here is provide a Bayesian analysis of the problem, using as prior the (assumed known) particle composition of the track sample. In this particular example probabilities correspond to rates, allowing for a frequentist interpretation of hypothesis probabilities.

# 6 What invariance properties do $p$-values enjoy?

$P$-values are invariant with respect to 1-to-1 transformations of the data coordinates. However, they are not invariant with respect to the choice of test statistic.

# 7 How do $p$-values behave versus sample size?

For significance tests, a simple Bayesian argument shows that the evidence provided by a $p$-value against the null hypothesis decreases as the sample size increases. A good rule of thumb is that $p$-values should be rescaled by a factor of $\sqrt{n}$, with $n$ the sample size, when comparing significances obtained from samples of different sizes. See for example section 4.3 in I.J. Good, "The Bayes/Non-Bayes Compromise: A Brief Review," J. Amer. Statist. Assoc. **87**, 597 (1992).

A related problem is known as "sampling to a foregone conclusion," and is a consequence of the law of the iterated logarithm (LIL) in probability theory. Suppose that one accumulates data continuously, and that at regular time intervals one calculates a $p$-value using *all the data collected thus far*, in order to test a given null hypothesis $H_0$. Then, for any given significance level $\alpha$, and *even if* $H_0$ *is true*, one is guaranteed to reach a point where the $p$-value fluctuates to a value smaller than $\alpha$. This is a purely mathematical consequence of the LIL. One way to avoid it is, as above, to rescale the $p$-value by the square root of the sample size (although this tends to somewhat overcompensate).

## 8 How can systematic uncertainties be incorporated in $p$-value calculations?

Nuisance parameters (i.e. parameters such as energy scale, tracking efficiency, integrated luminosity, etc., which are of no physics interest but introduce systematic uncertainties in a measurement) can cause complications. Possible ways of dealing with them are discussed briefly in the Appendix.

## 9 What are composite hypotheses and how can they be dealt with?

A hypothesis is composite if it does not specify unique values for all the free parameters in the problem (contrast with simple hypotheses, in which everything is completely specified). The unspecified free parameters could be nuisance parameters, in which case they can be handled as described in the appendix, or they could be interesting physics parameters. A simple case would involve fitting the parameters using as statistic the weighted sum of squared deviations between data and the hypothesis. The $p$-value is the probability for obtaining this weighted sum or a larger one. In the asymptotic limit (lots of data), it can be calculated by referring the weighted sum to a chisquared distribution for $N - f$ degrees of freedom ($N$ and $f$ are the numbers of data points and fit parameters, respectively). This is equivalent to using as $p$-value the largest one as the parameter(s) are varied to obtain the best fit (see description of supremum $p$-value in appendix).

In some cases it is possible to use one statistic for determining the best values of the parameters, and another for measuring the discrepancy between data and prediction. For example, one could use an unbinned maximum likelihood method to determine the parameter values, and then a binned chisquared test to determine the goodness-of-fit. In this case the number of degrees of freedom to be used in the goodness-of-fit test is not well defined, and a Monte Carlo simulation is likely to be very useful.

## 10 Can one combine $p$-values from different experiments?

$P$-values from different experiments can be combined, even though this procedure has some degree of arbitrariness associated with it. The combined $p$-value allows one to test a given hypothesis using several independent measurements. Assuming that the $p$-value distributions are uniform, the combined $p$-value is

$$P \times \sum_{j=0}^{N-1} \frac{[-\ln(P)]^j}{j!} \tag{10.1}$$

where $P$ is the product of the $N$ individual $p$-values. A slightly unfortunate feature of this formula is that, when combining three $p$-values $p_1$, $p_2$, and $p_3$, the result can be different if all three are combined directly; if $p_1$ and $p_2$ are combined, and the result is then combined with $p_3$; if $p_2$ and $p_3$ are combined, and the result is then combined with $p_1$; etc.

# Appendix

## A Methods for dealing with nuisance parameters

1. The plug-in $p$-value

   Method: Replace the nuisance parameter by some estimate, for example the maximum-likelihood estimate.

   Comment: If the data to be tested is included in the estimate, this leads to double use of the data (once in the estimate, and once in the $p$-value); the resulting $p$-value will not be uniform. Furthermore, this $p$-value does not always account for the uncertainty on the estimate.

   Example: Suppose you observe a number of events $N$ from a Poisson process with unknown mean $\mu$, and a separate measurement provides a Gaussian estimate $m \pm u$ for $\mu$. If you include both $N$ and $m$ in a maximum-likelihood estimate of $\mu$, the resulting $p$-value will depend on the uncertainty $u$, but the double use of $N$ makes the $p$-value non-uniform. If you do not include $N$, and simply replace $\mu$ by $m$, then the $p$-value will not take the uncertainty u into account.

   Suppose you use a chisquared statistic to test whether a bunch of points lie on a straight line with unknown slope and intercept. You can use the points themselves to first estimate the slope and intercept by minimizing the chisquared, but then the resulting $p$-value will be non-uniform, unless you correct the chisquare by subtracting two degrees of freedom.

2. The supremum $p$-value

   Method: Calculate the $p$-value for all possible values of the nuisance parameter and keep the largest one.

   Comment: Does not necessarily yield a uniform $p$-value.

   Example: Chisquared statistic to test whether a bunch of points lie on a line with unknown slope and intercept. Vary the slope and intercept until you find the largest $p$-value. This does not yield a uniform $p$-value however.

3. The similar $p$-value

   Method: Assume there exists a sufficient statistic for the nuisance parameter. Then the conditional probability density of the data, given the sufficient statistic, does not depend on the nuisance parameter and can be used to calculate a $p$-value.

   Comment: Based on a proper probability computation, imbuing the end result with desirable properties, but a suitable sufficient statistic does not always exist.

Example: Observation of a number of events $N_1$ from a Poisson process with unknown mean $\mu$. An estimate of $\mu$ is available from another Poisson measurement $N_2$ (with a possibly non-trivial sensitivity reduction factor). The sufficient statistic is $N_1 + N_2$, and the density of $N_1$ given $N_1 + N_2$ is binomial and independent of $\mu$.

4. <u>The prior predictive $p$-value</u>

Method: Suppose you have a reasonable prior density for the nuisance parameter. Multiply the probability density for the data by this prior density and integrate out the nuisance parameter. Use the resulting density to calculate a $p$-value.

Comment: Based on a proper probability computation. It is only uniform in an average sense over the nuisance parameter. It depends on a prior and therefore requires that this dependence be checked. If prior dependence is a problem, it may be tempting to try a non-informative prior. However, non-informative priors are often improper, leading to divergent marginalization integrals.

Example: Observation $N$ from a Poisson process with unknown mean $\mu$ for which there exists an independent Gaussian measurement result $m \pm u$. Convolute a Poisson with mean $\mu$ with a Gaussian with mean $m$ and width $u$. The resulting distribution depends only on $m$ and $u$ and can be used to calculate the $p$-value of $N$.

5. <u>The posterior predictive $p$-value</u>

Method: This is similar to the prior predictive $p$-value, except that instead of integrating with respect to the prior, one integrates with respect to the posterior for the nuisance parameter. This posterior is calculated using the data to be tested.

Comment: Makes double use of the data, first to calculate the posterior and then to calculate the $p$-value. Generally works with improper non-informative priors, since the posterior will typically be proper.