Recommendations on Signal Significance
(Updated 22 August 2003)

Significance, as it is normally defined, is the frequentist probability of making an observation that is at least as inconsistent with the null hypothesis as the observation actually made. In the statistics literature, this is formally known as the "p-value" of the observation. We (and the Particle Data Group) recommend that this terminology be used. See Reference [1] for a more detailed introduction to this concept.

Some Facts About P-Values

There are a number of useful "facts" about p-values that assist in understanding how to use them:

1. A p-value expresses the probability for a given hypothesis, of obtaining data at least as extreme as ours. For example, if the hypothesized distribution is a Poisson of mean 2.9 and we have observed 10 events, the p-value is

   Sum(n=10-->infinity) exp(-2.9)*2.9^n/n!

   Small p-values imply that the data is unlikely for the given model (and the deviation is in the "interesting" direction).
2. In ideal situations, and assuming the hypothesis is correct, p-values will be uniformly distributed between 1 and zero. 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 equidistant in p, and do not have equal weights. The p-value distribution cannot be uniform in the sense of dn/dp being constant. However it is "as uniform as possible" for a discrete distribution, with Prob(observing p.le.c) = c, where c is the location of any p-value.
3. A p-value is a useful quantity. a) It measures the compatibility of the data with the given hypothesis. b) It enables p-values from different experiments to be combined (even though this procedure has some degree of arbitrariness associated with it). The combined p-value determines how consistent the collection of experiments are with the hypothesis. Assuming that the p-value distributions are uniform, p-values may be combined by using the formula given by Eq (13) in Reference [1]. A slightly unfortunate feature of this formula is that, when combining 3 p-values, the result can be different if all 3 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. c) See also point 4).
4. Measures of significance are also used in Hypothesis Testing, where a p-value is used to accept or reject a given hypothesis. One defines, before the measurement is performed, a significance level alpha and then uses a test statistic (like a measure of goodness of fit) to see whether the data are consistent with the hypothesis at this level, by checking whether p.le.alpha. The expected rate of

`Errors of the First Kind' (i.e. how often the hypothesis is rejected when it is in fact true) is then alpha, and not the p-value. The p-value may be reported but its actual value is not relevant to the statistical conclusion.

5. A p-value measures the probability of observing DATA at least as extreme or unlikely as ours, assuming the hypothesis is true. It does NOT measure the probability that the HYPOTHESIS IS TRUE, based on our data. (See point 10 for an example.) This is an example of the difference between the probability of data, given a hypothesis; and the probability of the hypothesis, given the data. In particular, the following inferences are both WRONG: I) If p=3%, the probability of rejecting a true hypothesis is 3%. This is determined by alpha, not p. II) If p=7%, the probability that the hypothesis is in fact correct is 7%. The p-value cannot say anything about the probability of the hypothesis being correct (that is not even a frequentist concept!).
6. P-values are often used to summarize measures of "Goodness of Fit," ie, where we are comparing data distributions to a given hypothesis. Such measures are not to be regarded as a test of the null hypothesis. Similarly, a single p-value does not provide a means of Hypothesis Testing, in which two hypotheses are compared. Thus, a p-value can be used to see whether data is consistent with the Standard Model. If the p-value is small, this in itself does not imply that the Standard Model should be rejected. A useful procedure would be to compare the quality of the fits of the data to the Standard Model and to an a priori credible alternative. That still doesn't prove that the Standard Model is correct though.
7. P-values are invariant with respect to monotonic transformations of the data variable. They are not invariant with respect to the choice of statistic.
8. A Composite Hypothesis is one which involves free parameters (Contrast a Simple Hypothesis, which is completely defined). To calculate the compatibility of data with a Composite Hypothesis, choices must be made about what to do for the free parameter(s). A simple case would involve fitting the parameters using as a statistic to be minimized such as the weighted sum of squared deviations between data and the hypothesis. The probability for observing this chi-squared value or a larger value, corresponding to N-f degrees of freedom [N and f are the numbers of data points and of free parameters] is a p-value for the hypothesis. This is equivalent to using as p-value the largest one (i.e. the best fit) as the parameter(s) are varied. In other 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. In determining the p value, Monte Carlo simulation is likely to be very useful. Because the parameters have been allowed to vary, this p-value may be biased upwards.
9. Nuisance parameters can cause complications. Possible ways of dealing with them are discussed briefly in the Appendix below.
10. Here is a simple example illustrating that p-values do NOT give the probability of the hypothesis being wrong: 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 1 and zero. Now suppose that muons result in a p-value distribution of 1 - 0.1*(p-0.5) i.e. not too different from that for pions (because the pion and muon masses are similar), but slightly more peaked at small p. For a

sample of tracks with equal numbers of pions and muons, those with p close to 0.1 for the pion hypothesis will have a pion/muon ratio of 1/1.04. With a perhaps more realistic particle composition of 100 times more pions than muons, the small p pion/muon ratio becomes 100/1.04. In neither case would the wrong rejection of the pion hypothesis be anywhere near 10%

Recommendations for the Care and Feeding of P-Values

The following recommendations should be considered when determining the p-value of an observation.

1. To estimate a p-value, one must first define how one classes all possible observations given a specific null hypothesis. For example, if one is looking for a signal for the production of a certain class of events, the statistic x could be the number of candidate events in each observation. In this case, a large number of candidate events above the expected background rate would be increasingly inconsistent with the null hypothesis (in general, the chosen statistic must be able to discriminate between a specific null hypothesis and the other classes of hypotheses that are of physics interest). The choice of x is not, however, unambiguous. For example, if one is comparing a data histogram to one predicted by a Monte Carlo calculation, one could use the chi-square statistic, or a binned Kolmogorov-Smirnov distance, or any number of other measures. The p-value will depend on the choice of statistic. See Reference [2] for a case study of multiple significance measures.
2. If one knows the frequentist probability density p(x) of the random variable x assuming the null hypothesis, and then makes an observation x_0, then the p-value would be the integral of p(x) from x_0 to infinity. This assumes that x is a one-sided statistic, with smaller values implying better agreement with the null hypothesis.
3. One often cannot analytically determine p(x). In that case, one can resort to a Monte Carlo calculation where one estimates p(x) from the distribution of x in the MC experiments. The Monte Carlo calculation should sample the complete ensemble of possible experimental outcomes given the null hypothesis (this principle also should be satisfied by p(x)). It should take into account uncertainties in the inputs into the Monte Carlo calculation. Given that significance is a frequentist concept without Bayesian counterpart [3], systematic uncertainties should be treated in a frequentist manner. For example, if one is looking for an excess of events over a background with a known Gaussian uncertainty, the common procedure whereby one fluctuates the mean of a Poisson random variable according to a Gaussian density is not correct from a frequentist point of view. The correct procedure, and a further discussion of ensembles, can be found in reference [4]. For an example that violates this, see Example D in the Appendix.
4. In the case where one makes several, possibly correlated, simultaneous observations of random variables, one must first categorize the outcomes according to some measure that determines their consistency with the null

hypothesis. This may be the joint probability of the observations assuming the null hypothesis (this may not be the most sensitive or optimal measure), or some other function of the random variables. If the random variables are totally uncorrelated, then the combined significance is given by Eq. (13) in reference [1].

5. In cases where one is seeking a signal in several different channels, a straight-forward way to estimate the p-value of the simultaneous observations is to combine all channels together into a single measure of the signal rate [5]. This may not be optimal if the channels have very different background rates.
6. Although it is common to see p-values quoted in terms of the equivalent number of standard deviations a measurement should be from the expected mean of a normal distribution, it is more straight-forward to quote the actual p-value (ie., probability) and state explicitly the technique and assumptions used to estimate it. If you do quote equivalent standard deviations, remember that an upper limit should be converted to a one-sided Gaussian p-value estimate.
7. The design of an experiment usually involves estimating the sensitivity of a particular approach. In cases where one is observing a number of signal events S and one expects a number of background events B, one often sees measurement techniques optimized on the basis of the ratio S/sqrt(B), or S/sqrt(S+B) (see Reference [6] for a thorough discussion). In both cases, one is in fact making the assumption that S and B are normally distributed distributions. These may result in misleading “optimal” strategies, especially in cases where S and or B have non-Gaussian probability densities (as is the case where they represent numbers less than of order 10 events).
8. Posteriori decisions on the random variable used to measure a signal (such as the selection criteria used to identify a candidate event sample) make it difficult if not impossible to accurately calculate a p-value for a given observation once the observation has been made. Blind analyses avoid this specific problem, and should be considered when a search for new phenomena is undertaken. See Reference [7] for a description of blind analyses.
9. When one uses binned data to search for a possible signal and the location of the expected signal is not known, the p-value will be larger than a simple Poisson probability calculation would predict. See reference [8] for more details on how to account for this effect.
10. Always completely document the technique used to determine the p-value for an observation. Do not assume that it is too trivial or is well-known. In our experience, neither assumption is correct. One may always refer to an earlier paper where a complete description of the technique has been provided.

References:

[1] P. Sinervo, “Signal Significance in Particle Physics,” CDF Note 6031 and hep-ex/0208005, (July 2002).

[2] L. Demortier, "Assessing the significance of a deviation in the tail of a distribution", CDF Note 3419 (November 1995).

[3] The closest Bayesian concept is the ``Bayes factor,'' which is a ratio of posterior Bayesian probabilities for two different hypotheses.

[4] L. Demortier, “Constructing Ensembles,” CDF Note 6125 (September 2002).
[5] R. Hollebeek, H.H. Williams and P. Sinervo, “The evaluation of upper limits for top quark production using combined measurements," CDF Note 1109 (January 1990).
[6] G. Punzi, “Sensitivity of Searches for New Signals and Its Optimization,” ArXiv:physics/0308063 (August 2003).
[7] P. Harrison, ``Blind Analyses'' in Proceedings of the Conference on Advanced Statistical Techniques in Particle Physics, M. Whalley and L. Lyons (ed.), IPPP/02/39 (July 2002), page 278. See also J. Heinrich, “The Benefits of Blind Analysis Techniques,” CDF Note 6576 (July 2003).
[8] P. Sinervo, In preparation.

APPENDIX: Methods of dealing with nuisance parameters

(A) 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. This p-value does not always account for the uncertainty on the estimate.

Example 1: 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+/-u for mu. If you include both N and m in a maximum-likelihood estimate of mu, the resulting p-value for the hypothesis that the observation arises from only the presence of background 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.

Example 2: Suppose you use a chi-square 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 chi-square, but then the resulting p-value will be non-uniform, unless you correct the chi-square to a probability by subtracting two degrees of freedom.

(B) The **supremum** p-value

Method: Calculate the p-value for all possible values of the nuisance parameter given a set of data and keep the largest one.

Comment: Generally does not result in a p-value with a uniform distribution. Biased toward larger p-values, so may be conservative if one wants to minimize the chance of rejecting a true hypothesis. It also has reduced power.

Example: Chisquare 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.

(C) 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. However, a suitable sufficient statistic may not always exist.

Example: Observation of a number of events N1 from a Poisson process with unknown mean mu. An estimate of mu is available from another Poisson measurement N2 (with a possibly non-trivial sensitivity reduction factor). The sufficient statistic for mu is N1+N2, and the density of N1 given N1+N2 is binomial and independent of mu.

(D) The **prior predictive** p-value

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 Bayesian probability computation. The p-value is only uniform in an average sense over the nuisance parameter. The p-value depends on a prior and therefore requires that this dependence be checked for sensitivity to the choice of prior. 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 (uniform priors over an infinite parameter range are improper too).

Example: Observation N from a Poisson process with unknown mean mu for which there exists an independent Gaussian measurement result m+/-u. Assume a uniform prior for mu and multiply this by the likelihood function for the data (a Poisson with mean mu) Convolute this probability density in 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.

(E) The **posterior predictive** p-value

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.