Recommended Statistical Procedures for *BABAR* Summary of Recommendations

Roger Barlow, Robert Cahn, Glen Cowan, Francesca Di Lodovico,
William Ford, Gautier Hamel de Monchenault, David Hitlin,
David Kirkby, Francois Le Diberder, Gerry Lynch,
Frank Porter, Soeren Prell, Art Snyder,
Mike Sokoloff, Roland Waldi

May 7, 2002

General Recommendations

1. Whatever methodology is used, it should be clearly presented, so that the reader can understand what was done. If the technique is based on some published material, a reference should be given.

2. It should be anticipated that someone may wish to compare and/or combine the result with other information. Thus, enough information must be given to permit this, *e.g.*, systematic uncertainties should be quoted separately.

3. Whatever "physically relevant" quantity is quoted as a measurement, it is usually desirable to provide also the "actually measured" quantity.

4. It should be kept in mind that the sampling distribution is in general only approximately known, and there can be long "tails". Thus, robust methodologies are desirable, even at the cost of sufficiency or power.

5. It is useful to separate the presentation of "results", from the discussion of their "implications". "Results" means the information content of whatever measurement has been performed. "Implications" means the relevance to statements about physics.

6. An important general principle is that the "experiment design should precede the experiment". Resist the temptation to "tune" your analysis until some desired result is achieved. It is sometimes nearly impossible to do this in practice, but it should be kept in mind that not following this approach has sometimes led to disastrous results.

7. Avoid the potential bias introduced by basing the decision to publish on the result of an analysis.

8. The notation "χ^2" should not be employed as a substitute for a general $-2\ln\mathcal{L}$ quantity.

Confidence Intervals

1. The basic summary of the information in the experimental result will often be expressed as a best value, plus an interval to express the "resolution" of the measurement. This should be a two-sided 68% confidence interval.

2. It may also sometimes be desirable to provide some interpretation of the result in the form of a Bayes' interval.

3. If an interval is quoted without explicit definition, it should be a 68% confidence interval (or good enough approximation), in the frequency sense. Note that a one standard deviation interval is not necessarily a 68% confidence interval (the sampling distribution may not be normal).

4. "Two-sided" intervals should be quoted whenever possible. One-sided intervals ("limits") may optionally be quoted in addition. The confidence level of a one-sided interval should always be given; common practice is to quote a 90% confidence limit.

5. If the Gaussian approximation is valid, evaluating confidence intervals either by finding the points where $\Delta\chi^2 = 1$ or where $\Delta \ln \mathcal{L} = -1/2$ is recommended. If the Gaussian approximation is deemed unacceptable, then a calculation based on the sampling pdf should be made (perhaps with a Monte Carlo) in order to understand how to estimate an interval. A possible outcome is that the above methods may be applied, but with different values for the changes. A description should be given.

6. The "parabolic" errors produced by programs such as MINUIT may be used to estimate a 68% confidence interval if one is confident that they are a sufficiently accurate approximation. The same statement applies to use of the first-order formula for propagation of errors. In multi-variate situations, possible correlations should be incorporated.

7. Quoting an interval containing a given fraction (*e.g.*, 68%) of the area under the likelihood function is generally not recommended, although it can produce valid (in the frequency sense) confidence intervals in some circumstances. This method is, of course, the recommended procedure if a Bayes' interval with uniform prior is desired.

8. Methods which attempt to impose a constraint based on the possible true values of a parameter on a frequentist analysis (such as the Feldman-Cousins proposal) are not recommended.

Hypothesis Tests

1. The significance of a signal is best quoted as a probability for the observation to be consistent with no signal. It takes some judgement to decide what probability to evaluate for this, but in general the idea is to evaluate the probability that the no-signal model can give a "fluctuation" as large as, or greater than, that observed in the data.

2. If the normal distribution is a reasonable approximation, then quoting significance as a "number of sigma" (deviation from the no signal hypothesis) can also be used. This should not be done if the normal distribution is not a good approximation, as it can be misleading.

3. When there is a small number of events, the Poisson distribution should be used. The significance should be calculated as the Poisson probability for the background to fluctuate to the observed number of events, or more. This should not be applied blindly – first ask whether the background estimate is reliable or not. If the background estimate is deemed reliable, but the estimated value has a significant statistical uncertainty, then the probability distribution describing the background estimate should be folded into the computation of the significance of the "signal".

4. Another common case involves changes in a χ^2 variable or changes in $-2\ln\mathcal{L}$ to determine significance. To use these methods properly, it must either be known that the Gaussian approximation is valid, or a calculation (*e.g.*, by Monte Carlo) must be done to understand the distribution expected for these variables under the null (no-signal) hypothesis.

5. Methods involving integrals of the likelihood function should be first understood probabilistically, or relegated to the domain of an explicitly Bayesian analysis.

6. Whenever a "significance" is quoted, the definition of the quantity used should be given.

7. When performing a least-squares fit to binned data, beware of programs such as PAW when fitting to low-statistics data. PAW's default calculation is incorrect; it ignores bins with no events. When performing

such a fit to data which may have low bin contents, a safe procedure is to combine bins with especially low event counts, until a minimum count is reached. A reasonable minimum for Poisson data is to require at least seven events per bin. If this procedure is followed, the χ^2 goodness-of-fit statistic will generally be a good approximation.

8. If the statistics are low, a maximum likelihood fit to binned event data is preferable to a least-squares fit.

9. If it is important not to lose the information from combining bins (or from using bins which are wide compared to interesting structure), the least-squares method is not appropriate (or possibly even a binned likelihood fit). Instead, an unbinned fit to the individual event distribution should be performed, typically a maximum likelihood fit. Goodness-of-fit may be evaluated using a likelihood ratio statistic. A Monte Carlo may be required to understand the sampling distribution of this statistic.

10. A graphical display of a fit result (*i.e.*, a fitted curve superimposed on data) is highly recommended. This provides a way to visually check that a sensible result is being obtained.

11. When using the results of a fit, it is important not to lose sight of assumptions which might have been made, such as restrictions on background shape. A corresponding evaluation of systematic uncertainties should be made.

12. When quoting a χ^2 goodness-of-fit statistic, the χ^2 and the number of degrees of freedom should both be given, so that the reader can evaluate the confidence level. Or, the χ^2 confidence level may be directly given. Quoting χ^2/n_{DOF} is not recommended.

13. The basic method for checking consistency between correlated results (*e.g.*, related analyses on the same dataset) is described in the main report. A significance value of 0.05 is suggested for this test. If multiple tests are being made, then that should be properly accounted for in the evaluation of overall significance.

    (a) In testing for consistency, proper accounting for any correlations should be performed. If this is deemed too difficult, then limiting cases might be considered as a crude check. For example, if the sampling distributions are approximately Gaussian and the correlation coefficient is believed to be non-negative, then the difference can be compared with the range of standard deviations permitted: $|\sigma_{\hat{\theta}_1} - \sigma_{\hat{\theta}_2}|$ to $\sqrt{\sigma^2_{\hat{\theta}_1} + \sigma^2_{\hat{\theta}_2}}$.

(b) It may be possible to estimate the correlation by looking at the behavior in a related "control sample".

(c) The systematic uncertainties should be evaluated as appropriate in each analysis. The existence of more than one analysis may be used as a "check" for mistakes, but no new systematic uncertainty should be assigned to cover the difference between the two analyses.

(d) If there are systematic uncertainties which differ for the two analyses, these should be incorporated into the consistency test. If possible, any common systematic should be separated out, leaving only the "independent" systematics, call them s_{x_1} and s_{x_2}. Then it is reasonable to assign a systematic uncertainty of $s_{\Delta\theta} = \sqrt{s^2_{x_1} + s^2_{x_2}}$ to the difference, similarly to the result for uncorrelated statistical uncertainties. If it is too difficult to separate out the common systematics, then the best one can do is embark on a treatment similar to case for unknown correlations in the statistical errors.

Graphical Presentation

1. The most common graphical presentation of data is a histogram of the distribution of one variable $y = f(\vec{x})$. Each bin, labeled by the index j, has a full width, Δy_j, centered on y_j, and contains a non-negative integer number, n_j, of data elements. For this situation:

   (a) The horizontal axis label should describe the variable y (e.g., "Decay-Time Difference, Δt") and specify the units in which it is measured (*e.g.*, "ps").

   (b) The vertical axis label should specify the normalized basic unit of the data (*e.g.*, "Events/0.1 ps"). In the case of equal-width bins, the normalization Δy, should generally be to the common bin width.

   (c) The contents of each bin should be represented with a point at $(y_j, n_j \cdot (\Delta y / \Delta y_k))$ and an error bar that represents the "1-sigma" (68% confidence level) interval of the expected number of events given the observed number.

   (d) Horizontal error bars are discouraged for equally-spaced histograms, but recommended for unequal bin spacing. Horizontal error bars should cover the full width of the bin when used. In some cases, it may be useful to plot the point at the horizontal center-of-gravity of the data, with an error bar given by the rms variation. If this is done, it should be clearly explained in the caption.

2. The graphing of an appropriate error interval deserves attention. We assume that the probability density for observing n events in a bin as a function of the expected number, ν, is described by the Poisson distribution $P(n\,;\nu)$.

   An error bar extending from n to $n \pm \delta n$ corresponds to a one-sided interval of confidence level

   $$\alpha_{\pm} = \pm \int_{n}^{n \pm \delta n} P(n\,;x)\,dx = \frac{\pm 1}{n!}\left(\Gamma(n+1, n \pm \delta n) - \Gamma(n+1, n)\right) .$$

   Ideally, we would like $\alpha_- = \alpha_+ \simeq 34\%$. This generally leads to error bars that extend further on the low side than the high side. However, for $n = 0$ or 1, this is not possible, since the total cumulative probability below n is less than 34%. In this case, we settle for $\alpha_- + \alpha_- \simeq 68\%$ by fixing the lower bound of the error interval at zero and integrating up to the desired total confidence level.

3. For Poisson bin contents, a treatment using Poisson statistics should be used if the bin content is less than 10 events. For larger contents, the Gaussian approximation may be used.

4. A graphical presentation of the model used in analysis is most useful when superimposed on a representation of the data used to determine the model's parameters. This combination provides a visual impression of the goodness of a fit.

5. A presentation that combines data and a fit model should be chosen to clearly and accurately convey the statistical goodness of fit in the parameters of interest.

6. A likelihood curve should be displayed as $\log(L/L_{max})$ where L_{max} is the maximum likelihood value corresponding to the best fit of the parameters to the data. This choice eliminates the somewhat arbitrary value of L_{max} from the plot and allows several related likelihood curves (with different values of L_{max}) to be directly compared.

7. The likelihood ratio should not be scaled by a factor of 2 since there is no reason to prefer a χ^2-style presentation when an unbinned likelihood fit is used.

8. In the case of a fit to several parameters, the likelihood ratio should be calculated as $\log(L_{max}(p)/L_{max})$ where $L_{max}(p)$ is the maximum likelihood obtained in a fit with p fixed and all other parameters varied.

Systematic Errors

1. Results should always be quoted with a separate statistical and systematic error. For presentation purposes in some cases, these may also be combined in quadrature and given as a single error. In some cases it may be appropriate to separate systematic errors into components, but this should not be general practice.

2. As many *BABAR* measurements will be limited by systematic errors, their calculation and manipulation should be done on the basis of sound practice, not folklore. Be able to justify the procedures used.

3. Errors in a result due to factors that contain uncertainties should be considered as systematic even if those uncertainties are in themselves basically statistical, provided the sample on which they are based is independent of the sample from which the result is obtained.

4. Errors in a result due to factors that contain statistical uncertainties, derived from the same data sample but not of primary interest, should typically be considerd as statistical.

5. Errors in a result due to factors that contain theoretical uncertainties may be manipulated in a manner similar to those above, but their evaluation is more uncertain. There will often be no universally-agreed 'correct' value. A full explanation of the methodology you choose to use should be given, with justifications as appropriate. The following techniques may be used in evaluating such 'theory errors'.

   (a) Errors cited by theorists may be used. At least two estimates should be obtained, and you should then use your own judgement in selecting a value, and should be free to increase or decrease it if it is believed there is reason to do so.

   (b) Several different models may be used, and the spread of results used to give an error. How you do this depends on your judgement of the credibility of individual models. It may be appropriate to quote different results for different models.

   (c) Taking the error as the difference between the results of two models divided by $\sqrt{12}$ should only be done if the two models represent opposite extremes of the effect in question, and if the effect can reasonably be expected to lie anywhere in between them.

In combining these uncertainties, the correct (or, if unavailable, best possible) estimate of the covariance matrix should be used.

6. Uncertainties in parameters that are incorporated in the fit in a complicated way are not amenable to treatment by algebraic manipulation. The errors that arise from them must be evaluated by altering the parameter values and repeating the fit. This should generally be done by taking half the difference between the result values from plus and minus one standard deviation of the parameter, unless there is reason or evidence to believe that the errors are asymmetric.

7. The likelihood function may be a good way to explore the effects of systematic errors, especially if event numbers are small. The errors quoted should be those at which the log likelihood falls by 0.5 (for 1 dimension); if these are inconsistent with those obtained by other means (e.g. half the values for which it falls by 2.0, or those obtained from a 68% integral) then the full likelihood function should be given.

8. The robustness of any result should be established by many checks. These would typically include

   (a) Inclusion/exclusion of appropriate parts of the data, depending on running conditions.

   (b) Variation of important cuts

   (c) Different fitting techniques

   (d) Performing a similar analysis for which the result is known beforehand, on logical grounds or from previous experiments.

   (e) A full consistency check on a simulated data sample.

   In this it is important to distinguish between *blind checks*, where no effect is expected and which are used to detect mistakes, and *educated checks* where an effect is plausible and which are used to detect biases and/or estimate corrections.

   The result of an educated check may lead to a correction for bias, and will always give a contribution to the systematic error.

   The result of a blind check should be considered carefully to ascertain whether it is significant. If it is not, then no further action need be taken – it should not contribute to the systematic error. If it is, this indicates a mistake in the analysis, which should be searched for and (ideally) located and corrected. Only if this is unsuccessful, and the effect is small in absolute terms, should this be incorporated in the systematic error. These arguments apply as well to a full consistency check on simulated data, which will usually be a blind check.

9. In attempting to set a limit, the presence of a systematic uncertainty may complicate the analysis. We do not recommend using ad hoc approaches such as evaluating the limit using a correction term with a $\pm 1\sigma$ in the systematic. However, such methods are not necessarily "bad", and indeed can be used as a check on the robustness of the result – if the result shows great sensitivity to the methodology, this should be understood.

   (a) In the typical, well-behaved (*i.e.*, approximately Gaussian) case, add the statistical and systematic uncertainties in quadrature, according to the recommendation at the beginning of this section. Then treat the resulting uncertainty as a single normal error in obtaining the limit.

   (b) A more difficult, but still common situation is the low statistics case, where the statistical uncertainty should be treated with Poisson errors. This uncertainty will typically dominate over the systematic uncertainty, but the systematic should be incorporated into the limit, unless it is completely negligible. Assuming that a Gaussian approximation for the systematic is reasonable, the recommended procedure is to fold the Poisson distribution with a normal distribution describing the systematic uncertainty. A prescription for doing this is contained in the full report.

   (c) There may also be systematic uncertainties due to model uncertainties. If these are significant, then the dependence of the limit on these uncertainties should be presented.

Interpretation of Results

1. When it is desired to interpret a measurement in the context of a statement about physical "truth", the Bayesian methodology is recommended.

2. For the majority of measurements in *BABAR*, it will likely be appropriate to use the numerical frequentist results also as the interpretation. This is often equivalent to choosing a prior probability distribution ("prior") which is uniform in the unknown parameter of interest. This is readily justified for those results where the information from *BABAR* is good, that is, any conclusions are not highly dependent on the choice of smooth prior to express ignorance, and the measurement is substantially better than previous results.

3. However, there will undoubtably be situations where the conclusion is not so straightforward:

   (a) The measured result may be near a physical boundary. The prior should be zero for unphysical values of the parameter(s).

   (b) The measurement may be of comparable (or poorer) precision with previous knowledge. The prior should encapsulate the previous knowledge. There may be difficulty if it is not clear how to do this, *e.g.*, if a simple Gaussian approximation may not be valid, if the systematic uncertainties are substantial, if the information provided by the earlier experiment is insufficient to formulate a prior, or if there is disagreement with the methodology used by the earlier experiment. In such cases, the best one can do is formulate an educated consensus.

   (c) The measurement may have so little information that the choice of smooth prior makes an important difference. This should be pointed out in the discussion.

4. The general recommendation on the choice of a prior is as follows:

   (a) Express ignorance by using a prior which is zero in an "unphysical" region and a constant elsewhere.

(b) If the choice of how ignorance is expressed makes a difference to the result, say so, with an example. For example, if θ is a branching fraction, use prior densities of 1 and $1/\theta$ (for $\theta \geq 0$).

(c) Prior experimental information should be included according to the posterior from such experiments. Typically, this will take on the form of a normal distribution with standard deviation obtained by adding statistical and systematic errors in quadrature, as done by the Particle Data Group, including a scale factor if appropriate. Care must be taken, of course, to isolate common systematic uncertainties.

5. The question "Is there a signal?" may not be only a "statistical" one: We must be open to the possibility that we have made a mistake (an *a priori* more likely prospect the more startling the result). A critical examination of the experiment and analysis design, and of the available cross checks must be made to evaluate this possibility. Judgement is required; only general guidelines can be given:

   (a) Was the analysis pre-determined? That is, were the cuts established prior to looking at the data? It not, there may have been tuning of cuts, leading to a biased result.

   (b) How much "searching" was involved? If a lot of distributions were looked at, the "signal" may be just a large fluctuation that is not so improbable given the extent of the search. This again is better evaluated (*i.e.*, statistical methods can be used) if a careful design was originally carried out. If it was not pre-determined how much searching would go on, considerable judgement is required to evaluate the significance.

   (c) Does the purported signal show expected behavior? One check is to look at what is cut out, and see whether there is structure in the opposite direction from the signal in the discards. If so, the signal may be an artifact.

6. If it is supposed that the question of mistakes has been dealt with, the question of significance in terms of statistics may be asked. The significance may be computed either in frequentist terms or in Bayesian terms, but if the ultimate question is whether to claim a "significant" effect, some policy is required for an answer to be given.

It is recommended that the situation be quoted in such a way that the reader can decide. For example, instead of saying "We observe a signal for $B \to 3\pi$", say "We observe $B \to 3\pi$ at the 1% significance level." Or, instead of "We have observed CP violation in B decays", say "The measured value of $\sin 2\beta$ is different from zero at the 6% significance level." A corollary is that language in titles of the form "Observation of..." is discouraged, as well as adjectives such as "first".