

Statistical Issues in Particle Physics – A View from BaBar

Frank Porter

Caltech

(For the BaBar Collaboration)

Outline

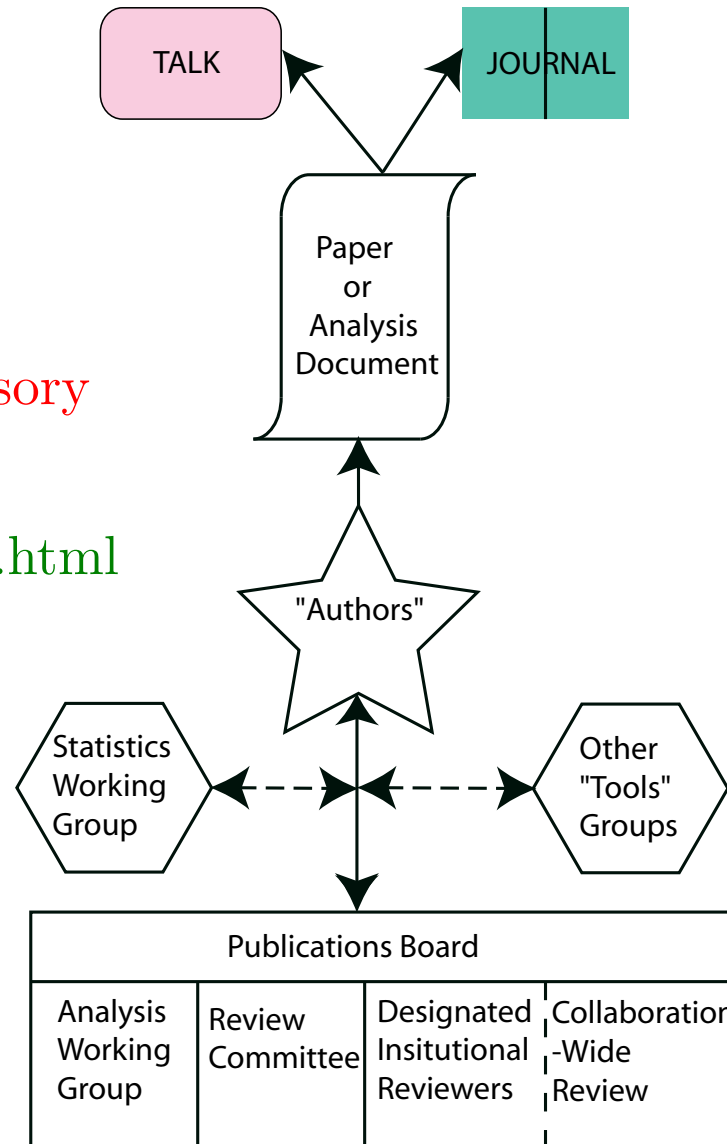
- ❑ BaBar Analysis Organization (1 slide)
- ❑ Statistics Philosophy (1 slide)
- ❑ Statistics Practice in BaBar (27 slides)
- ❑ Reflections (1 slide)

Organization – BaBar Analysis

Statistics Working Group is advisory

<http://www.slac.stanford.edu/>

BFROOT/www/Statistics/index.html



Philosophy

Basic approach is to know first what you are trying to accomplish.

Two broad domains:

- ❑ Summarizing information (“descriptive”)
 - This is viewed as **obligatory** (and **useful**).
 - Recommend **frequency** statistics.
 - Emphasis on clarity, ability to compare & combine with other results.
- ❑ Providing interpretation
 - Viewed as **optional**.
 - Recommend **Bayesian** approach.

Principle products of BaBar physics analyses:

- ❑ **Best estimates**
- ❑ **Interval estimates**
- ❑ **Significance levels** (eg, of a possible discovery)
- ❑ **Goodness-of-fit** (of data to some model)

Statistical Practice in BaBar

- ❑ Blind analysis
- ❑ Confidence intervals
- ❑ Significance
- ❑ Systematic Uncertainties
- ❑ Goodness of Fit
- ❑ Consistency of analyses

Left out: methods/tools for optimizing analyses; pattern recognition/data reduction/simulation

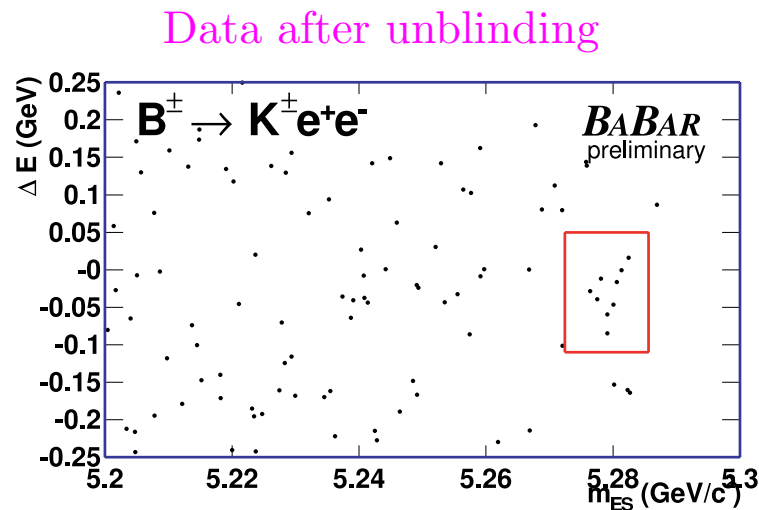
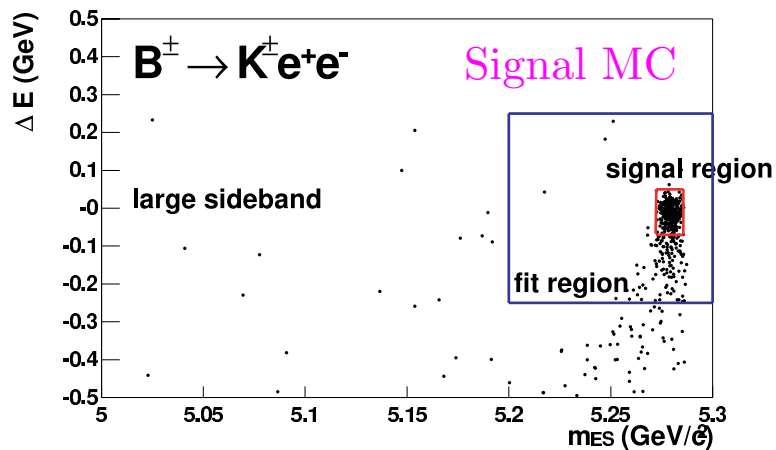
Blind Analysis

Many BaBar results are obtained in “blind analyses”. Purpose is to avoid introduction of bias. [More than one approach; See Aaron Roodman’s parallel session talk]

Not all: Exploratory analyses not blind, eg, discovery of the $D_{sJ}^*(2317)^+$.

Example of a blind analysis: $B^\pm \rightarrow K^\pm e^+ e^-$

After event selection, look for signal in distribution of two kinematic variables “ ΔE ” and “ m_{ES} ”. Monte Carlo, control sample (usually a type of data resembling signal) studies of entire sideband and fit region. Data only looked at in large sideband region prior to unblinding.



Blind Analysis (continued)

Issue: Updating a blind analysis with additional data.

- ❑ Simply add data, no change in analysis.
 - May be impractical, undesirable, eg, re-reconstruction of entire dataset; improvements in tools such as particle identification.
 - Sometimes would like to work harder to optimize analysis, or optimize on different criterion (eg, for most precision instead of most sensitivity).
- ❑ Notion of “reblinding” and re-optimization.
 - Considered safe to use variables which have not been inspected too carefully in blind region.
 - Not called a blind analysis.

Confidence intervals

Use frequency statistics for summarizing information.

Goal is to describe what we observe, with properties:

- Simple, coherent interpretation
- Facilitate combination with other results

We think it can be counter-productive to impose “physical” constraints. No reason to obscure observation of an “unlikely” result. Imposing constraint may complicate combination.

- Generally, recommendation is to quote two-sided 68% confidence intervals as primary result.
- Check for frequency validity (coverage).
- Technical complication, eg, in low statistics domain.

Confidence Intervals (continued)

Example in 2 Dimensions: D mixing and DCSD

Two mixing parameters to be determined:

$$x' \equiv \frac{\Delta m}{\Gamma} \cos \delta + \frac{\Delta \Gamma}{2\Gamma} \sin \delta,$$

$$y' \equiv \frac{\Delta \Gamma}{2\Gamma} \cos \delta - \frac{\Delta m}{\Gamma} \sin \delta,$$

where δ is an unknown strong phase (between Cabibbo-favored and doubly Cabibbo-suppressed amplitudes). Only sensitive to x'^2, y' , maximum of likelihood may occur at $x'^2 < 0$ (“unphysical” region).

In standard model, should see $x'^2 = y' = 0$, at current sensitivity.

Example in 2 Dimensions: D mixing and DCSD (continued)

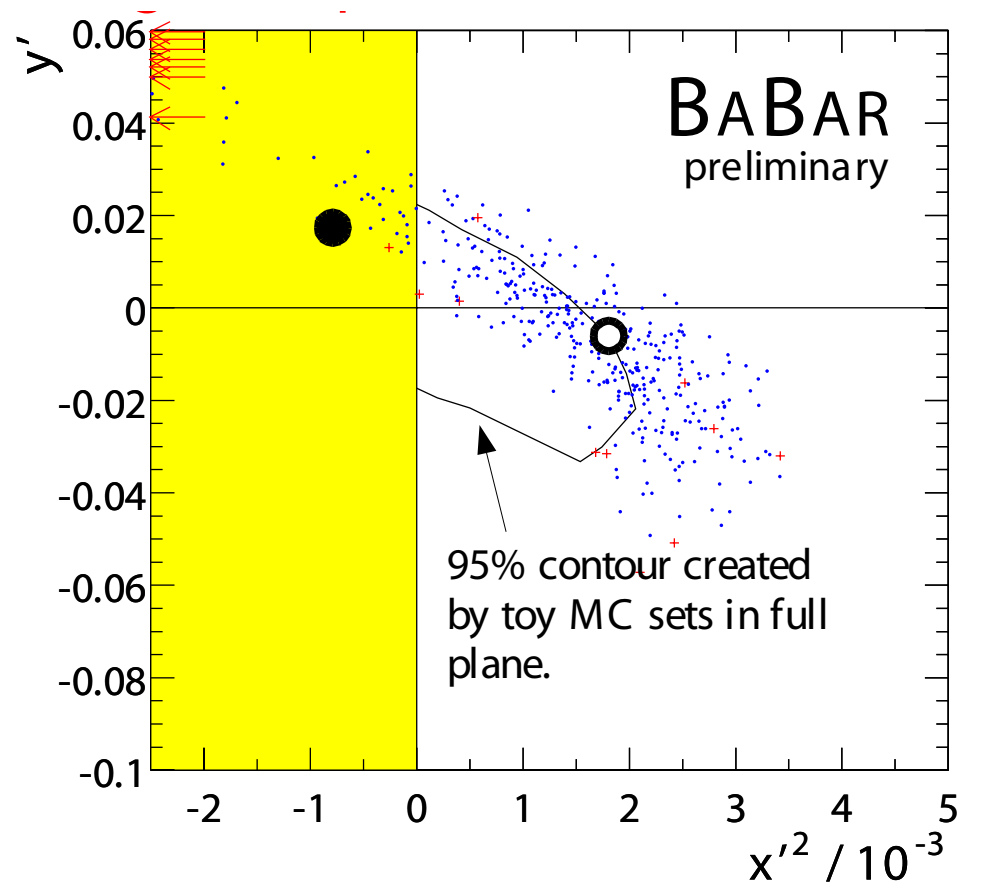
Frequentist approach to 95% C.L. contour:

- Map contour by generating ensembles of MC experiments at many points $(x_0'^2, y_0')$ in (x'^2, y') space.
- A point is inside the contour iff at least 95% of MC experiments for that point have a likelihood ratio which is bigger than the data:

$$\lambda_{\text{MC}} = \frac{\mathcal{L}_{\text{max}}(\text{MC})}{\mathcal{L}_{(x_0'^2, y_0')}(\text{MC})}$$

$$\lambda_{\text{Data}} = \frac{\mathcal{L}_{\text{max}}(\text{Data})}{\mathcal{L}_{(x_0'^2, y_0')}(\text{Data})}$$

If $P(\lambda_{\text{MC}} > \lambda_{\text{Data}}) \geq 0.95$, then $(x_0'^2, y_0')$ is inside (or on) the contour.



- Converged point for fit to data.
- Test point of toy Monte Carlo set.
- Red: $\lambda_{\text{MC}} < \lambda_{\text{Data}}$
- Blue: $\lambda_{\text{MC}} > \lambda_{\text{Data}}$

(U. Egede, International Workshop on Frontier Science, Frascati, October 6-11, 2002)

Frank Porter, 10 September 2003, PHYSTAT2003

Confidence Intervals: Low Statistics Issues

Low statistics \Leftrightarrow Non-normal sampling

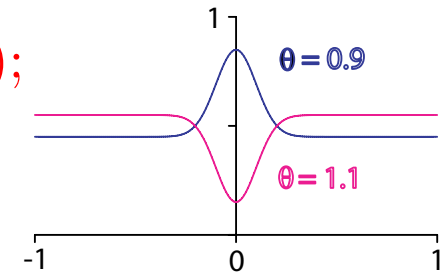
- ❑ Technical issue: Hitting a mathematical boundary.
- ❑ Frequency validity (coverage).
 - Incorporating uncertainty in background.
 - Incorporating scale uncertainty.

Hitting a Math Boundary

Consider a maximum likelihood fit of a set of events to some distribution, depending on parameters of interest.

Example: $p(x; \theta) = \frac{\theta}{2} + \frac{1 - \theta}{A\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}, x \in (-1, 1);$

$$\mathcal{L}(\theta; \{x_i, i = 1 \dots, N\}) = \prod_{i=1}^N p(x_i; \theta).$$



- Maximum wrt θ may be outside of region where PDF is defined.
 - The function $p(x; \theta)$ may become negative in some regions of x .
 - If there are no events in these regions, the likelihood is still “well-behaved”.
 - However, the resulting fit, as a description of the data, will typically look poor even in region of positive PDF. This is considered unacceptable.

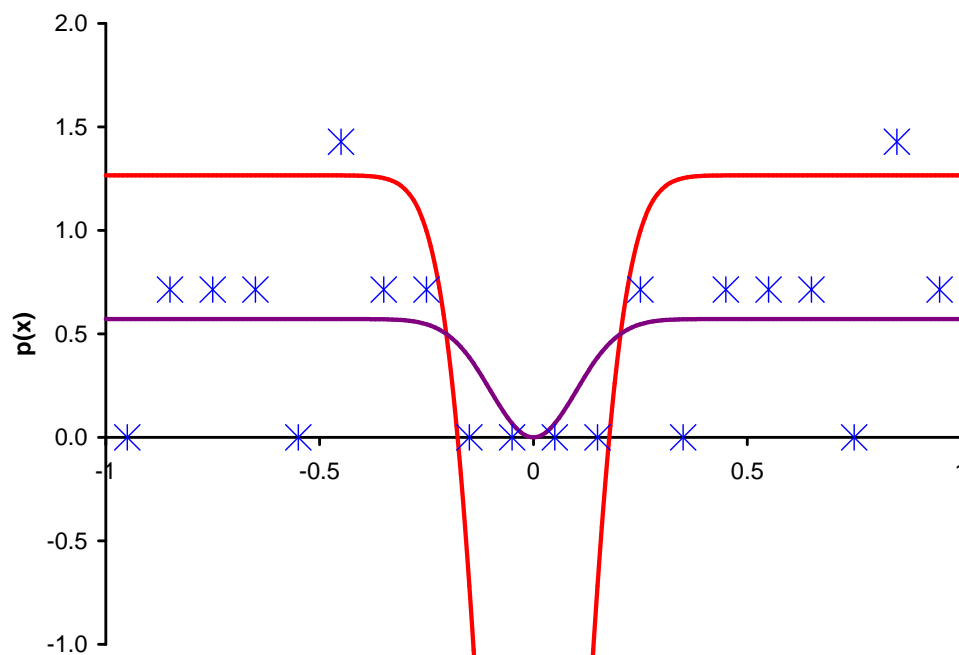
Hitting a Math Boundary (continued)

$$p(x; \theta) = \frac{\theta}{2} + \frac{1-\theta}{A\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}}, \quad x \in (-1, 1).$$

Points: Histogrammed “data”.

Red: $p(x; \theta)$ allowed to go negative in ML fit.

Purple: $p(x; \theta)$ constrained non-negative



- ❑ **Practical resolution:** Constrain fit to remain within bounds such that PDF is everywhere legitimate.
 - n.b., parameters may still be “unphysical”
 - Experience is that this gives fits which “look” like the data.
 - Applies in interval evaluation also (but check coverage, as always).

Interval Estimation in Poisson Sampling with Scale Factor and Background Subtraction

The Problem (eg): A “Cut and Count” analysis for a branching fraction B finds n events.

- The background estimate is $\hat{b} \pm \sigma_b$ events.
- The efficiency and parent sample are estimated to give a scaling factor $\hat{f} \pm \sigma_f$.

How do we determine a (frequency) Confidence Interval?

- Assume n is sampled from Poisson, $\mu = \langle n \rangle = fB + b$.
- Assume \hat{b} is sampled from normal $N(b, \sigma_b)$.
- Assume \hat{f} is sampled from normal $N(f, \sigma_f)$.

$$\mathcal{L}(n, \hat{b}, \hat{f}; B, b, f) = \frac{\mu^n e^{-\mu}}{n!} \frac{1}{2\pi\sigma_b\sigma_f} e^{-\frac{1}{2}\left(\frac{\hat{b}-b}{\sigma_b}\right)^2 - \frac{1}{2}\left(\frac{\hat{f}-f}{\sigma_f}\right)^2}.$$

Interval Estimation in Poisson Sampling (continued)

Variety of Approaches – Dealing With the Nuisance Parameters

See also Roger Barlow, “A Calculator for Confidence Intervals”, MAN/HEP/2001/04.

- ❑ Just give n , $\hat{b} \pm \sigma_b$, and $\hat{f} \pm \sigma_f$.
 - Provides “complete” summary.
 - Should be done anyway.
 - But it isn’t a confidence interval...
- ❑ Integrate over $N(\hat{f}, \sigma_f)$ “PDF” for f , $N(\hat{b}, \sigma_b)$ “PDF” for b . (variant: normal assumption in $1/f$).
 - Quasi-Bayesian (uniform prior for f , b (or, eg, for $1/f$)).
- ❑ Ad hoc: eg, Upper limit – Poisson statistics for n , but with scale, background shifted by uncertainty.
 - Easy
 - makeshift; extension to two-sided intervals?
- ❑ Here, investigate finding confidence intervals an old way.

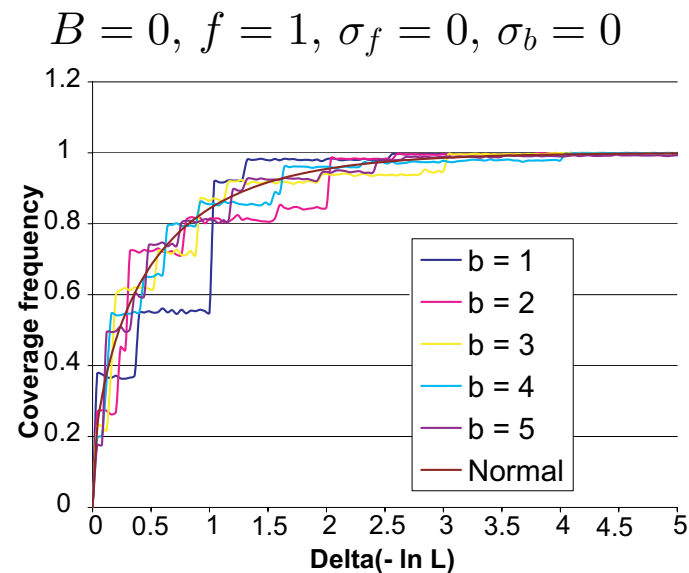
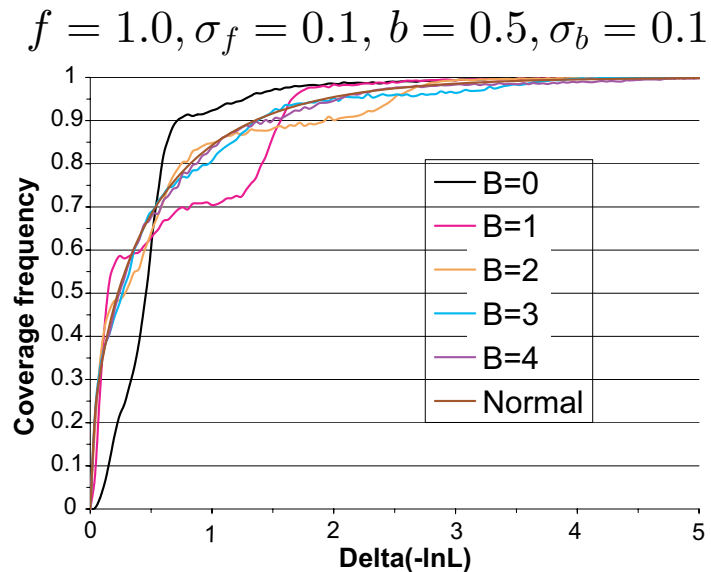
The Neglected(?) Method

1. Write down the likelihood function in all parameters.
2. Find the global maximum.
3. Search in B parameter for where $-\ln L$ increases from minimum by specified amount (e.g., $\Delta = 1/2$), re-optimizing with respect to f and b .

Does it work? Investigate the frequency behavior of this algorithm.

- For large statistics (normal distribution), we know that for $\Delta = 1/2$ this produces a 68% confidence interval on B .
- How far can we trust it into the small statistics regime?

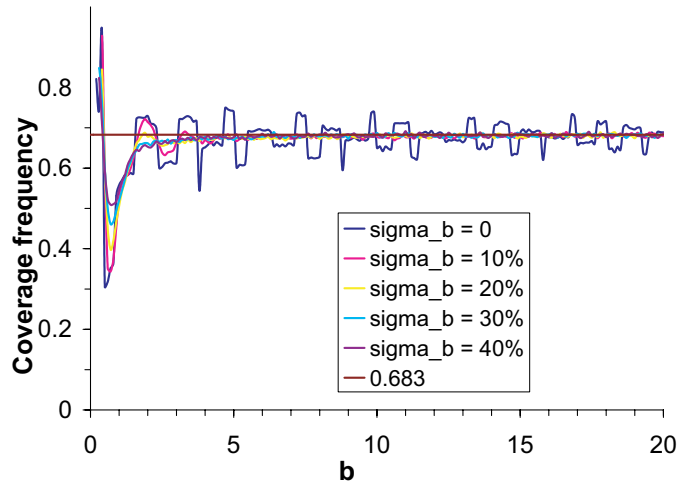
Method also applicable to unbinned analysis. (see also Alexander Bukin talk)



Study of coverage (continued)

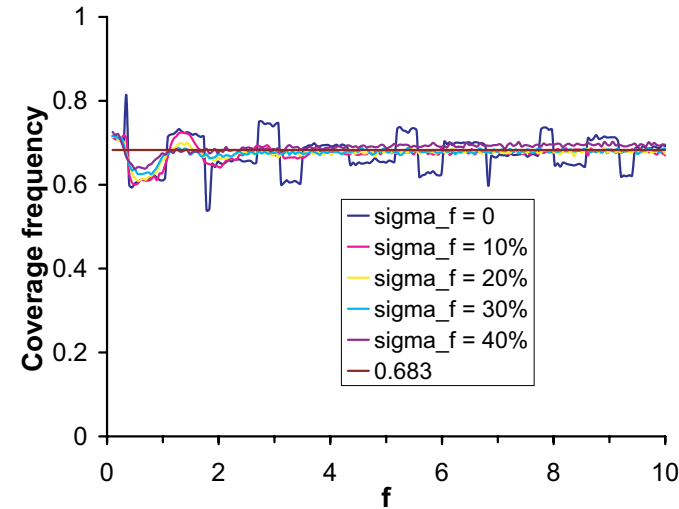
Dependence on b and σ_b

$B = 0, f = 1, \sigma_f = 0, \Delta = 1/2$



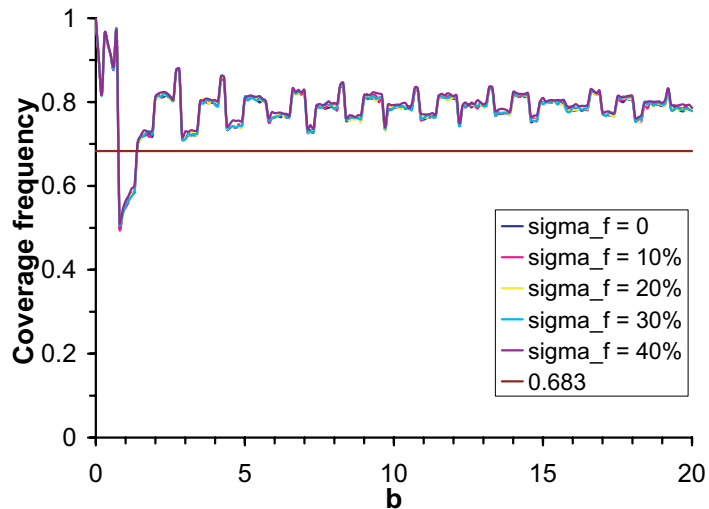
Dependence on f and σ_f for $B = 1$

$B = 1, b = 2, \sigma_b = 0, \Delta = 1/2$



Changing Δ

$B = 0, f = 1, \sigma_b = 0, \Delta = 0.8$



- Uncertainty in background and scale helps.
- Can increase Δ if want to put a floor on coverage.

Summary: Confidence Intervals with Low Statistics

- Always give n , $\hat{b} \pm \sigma_b$, and $\hat{f} \pm \sigma_f$.
- Justify chosen approach with computation of frequency.
- Likelihood method considered here works pretty well (Well enough?) even for rather low expected counts, for 68% confidence intervals. Uncertainty in b , f improves coverage.
- If $\sigma_b \approx b$ or $\sigma_f \approx f$, enter a regime not studied here. Normal assumption probably invalid.
- Could choose larger $\Delta(-\ln L)$ if want to insure at least 68%, or push to very low statistics.
- Good enough for 68% confidence interval doesn't mean good enough for significance test. If statistics is such that Gaussian intuition is misleading, should ensure this is understood.

[See also Roger Barlow talk on asymmetric errors]

Confidence Intervals (continued)

In interpretation stage, Bayesian intervals may be given, as deemed useful to the consumer.

In BaBar practice, this is usually done when someone wants to give an upper limit, and is usually implemented with the **assumption of a uniform prior in the parameter of interest**.

BaBar recognizes the issue of choice of prior – recommend thought, and checks on how much it matters. Typically largely ignored.

Significance

- ❑ Defined as probability of observed deviation from null, under null hypothesis.
- ❑ Give frequentist result for significance.
 - A 68% confidence interval does not always tell you much about significance. **The tails may be non-normal.** A separate analysis is generally required, which models the tails appropriately.
- ❑ No recommendation on when to label result as “significant”. Label implies interpretation.
 - No uniform prescription seems to make sense, involves judgement. Eg, bizarre new particle vs. expected branching fraction.
 - Not our primary experimental role; up to reader ultimately to decide what they want to believe.

Perhaps the least-accepted of our points in BaBar: People insist on making qualitative statements (“observation of”, “evidence for”, “discovery of”, “not significant”, “consistent with”)

Code: “observation of” $\equiv > 4\sigma$, “evidence for” $\equiv > 3\sigma$

Significance (more semantics...)

From Physics Today, <http://www.physicstoday.org/pt/vol-54/iss-9/p19.html>
(coloring mine, references deleted)

nb: Just an observation on human nature, no criticism should be inferred.

“In March, back-to-back papers in Physical Review Letters reported the measurement of CP symmetry violation in the decay of neutral B mesons by groups in Japan and California. Now the word “**measurement**” has been replaced by “**observation**” in the titles of two new back-to-back reports by these same groups in the 27 August Physical Review Letters. That is to say, with a lot more data and improved event reconstruction, the BaBar collaboration at SLAC and the Belle collaboration at KEK in Japan have at last produced the **first compelling evidence** of CP violation in any system other than the neutral K mesons.”

Some people think a measurement should not be called a “**measurement**” unless the result is significantly different from zero. A senior Assistant Editor at a prominent journal has suggested that “**bounds on**” might be more appropriate than “**measurement**” in reference to a CP asymmetry angle which was observed as consistent with zero.

**Finding $\sin 2\beta = 0.00 \pm 0.01$ would be pretty exciting.
But it isn't a “**measurement**”?**

Significance – Two Issues

Issue: Many people mix question of “significance” with choice of interval (two-sided vs limit).

Algorithm of Feldman&Cousins designed to address this.

Our recommendation is to always give two-sided interval (if otherwise appropriate). Significance is quoted separately. One-sided intervals optional, usually regarded as part of interpretation (hence Bayesian approach suggested).

Typically followed, though perhaps not everyone buys into this yet...

Issue: Long ingrained tradition of quoting significance as “ $n\sigma$ ”. Meaning is ambiguous: sometimes it means n standard deviations, sometimes it means probability content of an $n\sigma$ fluctuation for a normal distribution.

Recommend to state probability directly if not a normal sampling.

Not much headway on this one...

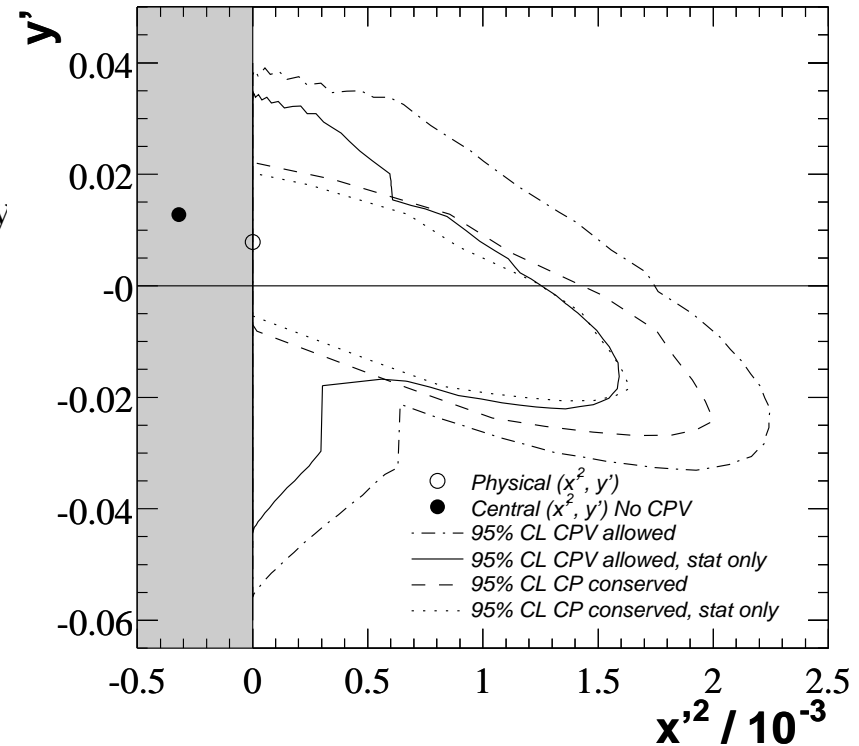
Systematic Uncertainties

- “Blind checks” and “educated checks”:
 - **Blind check**: testing for mistakes; no correction is expected. If pass test, **no contribution to systematic error**. Eg, divide data into chronological subsets and compare results.
 - **Educated check**: measuring biases, corrections. May affect quoted result. Always **contributes to systematic error**. Eg, dependence of efficiency on model.

- Quote systematic uncertainty separately from statistical
 - Systematic uncertainty may contain statistical components, eg, **MC statistics in evaluation of efficiency**.

Systematic Uncertainties Example: D mixing and DCSD revisited

- Want simple procedure. Willing to accept approximation.
- Scale statistical-only contour uniformly along ray from best-fit value.
- Factor is $\sqrt{1 + \sum m_i^2}$, where m_i is an estimate of the effect of systematic uncertainty i measured in units of the statistical uncertainty. This estimate is obtained by determining the effect of the systematic uncertainty on \hat{x}'^2, \hat{y}' .



Method conservative (\equiv lazy) in sense that scaling for a given systematic in one direction is applied uniformly in all directions. On the other hand, a linear approximation is being made.

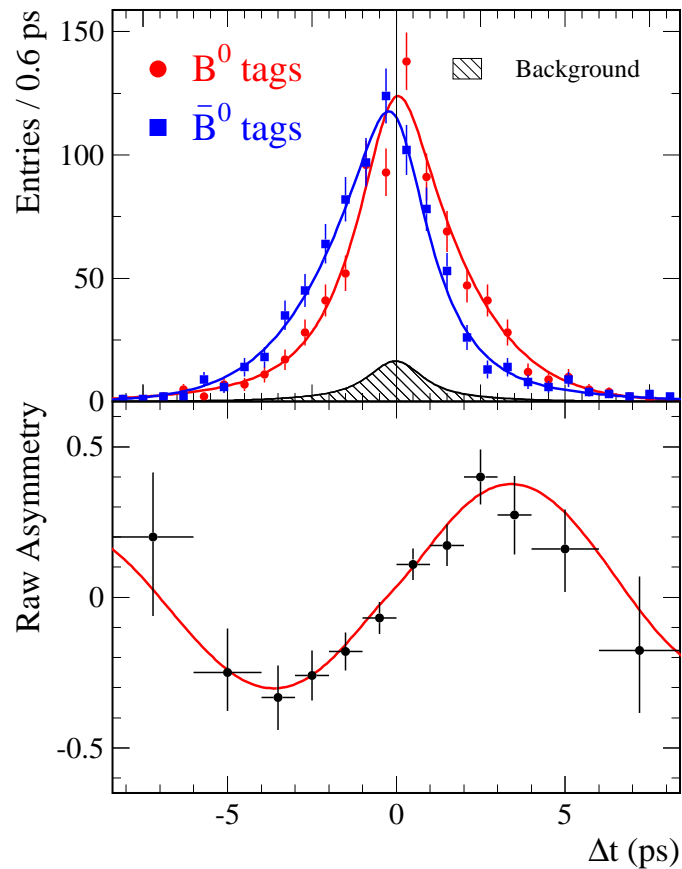
Goodness-of-Fit

- No perfect general goodness-of-fit test:
 - Given a dataset generated under null hypothesis, can usually find a test which rejects the null hypothesis (ie, choosing the test after you see the data is dangerous).
 - Given a dataset generated under alternative hypothesis, can usually find a test for which the null passes (ie, should think about what you want to test for).
- Nominal recommendation:
 - If you have a specific question, test for that.
 - χ^2 test when valid.
 - Consider more general likelihood ratio test, Kolmogorov-Smirnov, etc., otherwise.
 - Monte Carlo evaluation of distribution of test statistic.

[See also Ilya Narsky talk.]

Goodness-of-Fit (continued)

But recognize when test may not answer desired question, eg, in $\sin 2\beta$ analysis, a likelihood ratio (or a χ^2) test on the time distribution may have little sensitivity to testing goodness-of-fit of the asymmetry.



Consistency

BaBar has already encountered several times the question of whether a new analysis is consistent with an old analysis.

- ❑ Often, new analysis is a combination of additional data plus changed (improved...) analysis of original data.
- ❑ The stickiest issue is handling the correlation in testing for consistency in the overlapping data.
- ❑ People sometimes have difficulty understanding that statistical differences can arise even comparing results based on the same events.

Given a sampling $\hat{\theta}_1, \hat{\theta}_2$ from a bivariate normal distribution $N(\theta, \sigma_1, \sigma_2, \rho)$, with $\langle \hat{\theta}_1 \rangle = \langle \hat{\theta}_2 \rangle = \theta$, the difference $\Delta\theta \equiv \hat{\theta}_2 - \hat{\theta}_1$ is $N(0, \sigma)$ -distributed with $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$.

If the correlation is unknown, all we can say is that the variance of the difference is in the range $(\sigma_1 - \sigma_2)^2 \dots (\sigma_1 + \sigma_2)^2$. If we at least believe $\rho \geq 0$ then the maximum variance of the difference is $\sigma_1^2 + \sigma_2^2$.

Consistency – Simple example of two analyses on same events

Suppose we measure a neutrino mass, m , in a sample of $n = 10$ independent events. The measurements are $x_i, i = 1, \dots, 10$. Assume the sampling distribution for x_i is $N(m, \sigma_i)$.

We may form **unbiased** estimator, \hat{m}_1 , for m :

$$\hat{m}_1 = \frac{1}{n} \sum_{i=1}^n x_i \pm \sqrt{\frac{1}{n^2} \sum_{i=1}^n \sigma_i^2}.$$

The result (from a MC) is $\hat{m}_1 = 0.058 \pm 0.039$.

Then we notice that we have some further information which might be useful: we know the experimental resolutions, σ_i for each measurement. We form another **unbiased** estimator, \hat{m}_2 , for m :

$$\hat{m}_2 = \sum_{i=1}^n \frac{x_i}{\sigma_i^2} / \sum_{i=1}^n \frac{1}{\sigma_i^2} \pm 1 / \sqrt{\sum_{i=1}^n \frac{1}{\sigma_i^2}}.$$

The result (from the same MC) is $\hat{m}_2 = 0.000 \pm 0.016$.

The results are certainly correlated, so question of consistency arises (we know the error on the difference is between 0.023 and 0.055). In this example, the difference between the results is 0.058 ± 0.036 , where the 0.036 error includes the correlation ($\rho = 0.41$).

Consistency – Evaluating the Correlation

Art Snyder developed an approximate formula for evaluating the correlation in a comparison of maximum likelihood analyses (eg, in one-dimensional case).

Suppose we perform two maximum likelihood analysis, with event likelihoods \mathcal{L}_1 , \mathcal{L}_2 , on the same set of events [nb, **may use different information in each analysis**]. The results are estimators $\hat{\theta}_1$, $\hat{\theta}_2$ for parameter θ . The correlation coefficient ρ may be estimated according to:

$$\rho \approx \frac{\sum_{i=1}^N R_i \frac{d \ln \mathcal{L}_{1i}}{d\theta} \Big|_{\theta=\hat{\theta}_1} \frac{d \ln \mathcal{L}_{2i}}{d\theta} \Big|_{\theta=\hat{\theta}_2}}{\sqrt{\left(\sum_{i=1}^N \frac{d^2 \ln \mathcal{L}_{1i}}{d\theta^2} \Big|_{\theta=\theta_0} \right) \left(\sum_{i=1}^N \frac{d^2 \ln \mathcal{L}_{2i}}{d\theta^2} \Big|_{\theta=\theta_0} \right)}},$$

where (θ_0 is an expansion reference point)

$$R_i = \left[1 - (\hat{\theta}_1 - \theta_0) \frac{d^2 \ln \mathcal{L}_{1i}}{d\theta^2} \Big|_{\theta=\theta_0} / \frac{d \ln \mathcal{L}_{1i}}{d\theta} \Big|_{\theta=\hat{\theta}_1} \right] \left[1 - (\hat{\theta}_2 - \theta_0) \frac{d^2 \ln \mathcal{L}_{2i}}{d\theta^2} \Big|_{\theta=\theta_0} / \frac{d \ln \mathcal{L}_{2i}}{d\theta} \Big|_{\theta=\hat{\theta}_2} \right].$$

If $\theta_0 \approx \hat{\theta}_1 \approx \hat{\theta}_2$, then

$$\rho \approx \tilde{\sigma}_{\theta_1} \tilde{\sigma}_{\theta_2} \sum_{i=1}^N \frac{d \ln \mathcal{L}_{1i}}{d\theta} \Big|_{\theta=\hat{\theta}_0} \frac{d \ln \mathcal{L}_{2i}}{d\theta} \Big|_{\theta=\hat{\theta}_0},$$

where $\tilde{\sigma}_{\theta_k}^2 \equiv 1 / \sum_{i=1}^N \left(\frac{d \mathcal{L}_{ki}}{d\theta} \Big|_{\theta=\theta_0} \right)^2$

Consistency – Example: $\sin 2\beta$

- $32 \times 10^6 B\bar{B}$ pairs – PRL, vol 87, 27 August 2001:
 $\sin 2\beta = 0.59 \pm 0.14(\text{stat}) \pm 0.05(\text{syst})$
- $62 \times 10^6 B\bar{B}$ pairs – SLAC-PUB-9153, March 2002:
 $\sin 2\beta = 0.75 \pm 0.09(\text{stat}) \pm 0.04(\text{syst})$

Second result includes the earlier data, re-reconstructed. Analysis involves multivariate maximum likelihood fits; reprocessing changes, eg, relative likelihood for an event to be signal or background. Not simply counting events. **Question: are the two results statistically consistent?**

If these were independent data sets, a difference of 0.16 ± 0.17 would not be a worry. **The issue is the correlation.**

A specialized analysis deriving from the previous formula is performed on the events in common between the two analyses. **A correlation of $\rho = 0.87$ is deduced, yielding a difference of $\sim 2.2\sigma$.**

Consistency - Is BaBar Making Mistakes?

Answer: No compelling evidence for mistakes.

Question: But why is BaBar seeing these differences between old and updated results?

Answer: Because they should!

Question: Why is BaBar “different”?

Answer: Because they (almost) religiously use blind methodology.

No (little) opportunity to react to differences with further analysis...

Reflections

- ❑ Statistical sophistication in particle physics has grown substantially. (not specific to BaBar)
 - Bayesian vs Frequentist hang-ups persist, but now at least most people know the names.
 - Great sensitivity to issue of biases in analyses. BaBar relies heavily on blind analyses.
- ❑ BaBar adopts frequency statistics for describing results.
 - Much attention is devoted to MC validation & verifying coverage.
- ❑ Bayesian approach in HEP (including BaBar) still in infancy: no established methodology for choosing prior (other than defaulting on uniform). Justification is that it doesn't matter much, often...
- ❑ Even issues in frequency statistics (“physical region”, notion that backgrounds should “always” lead to higher limits). But extensive checking on frequency validity of adopted algorithms is typically done.
- ❑ Attempt to provide coherent documented approach in BaBar.
... A work in progress...