The Significance of HEP Observations

My charge:

"...a presentation on standard treatments of significance :

'The significance of HEP Observations'

The talk will cover all related statistic concepts, and some methods which are being used by BaBarians, as well as blinding analysis and unbiased measurements."

❏ Significance as hypothesis test

❏ Pitfalls

❏ Avoiding pitfalls

❏ Conclusions

Perspective

NOTE!!!

Context is frequency statistics!

Domain of descriptive statistics.

Significance as Hypothesis Test

When asking for the “significance” of an observation (of, perhaps a new effect), you ask for a test of the hypotheses:

Null hypothesis H_0 : There is no new effect;

against

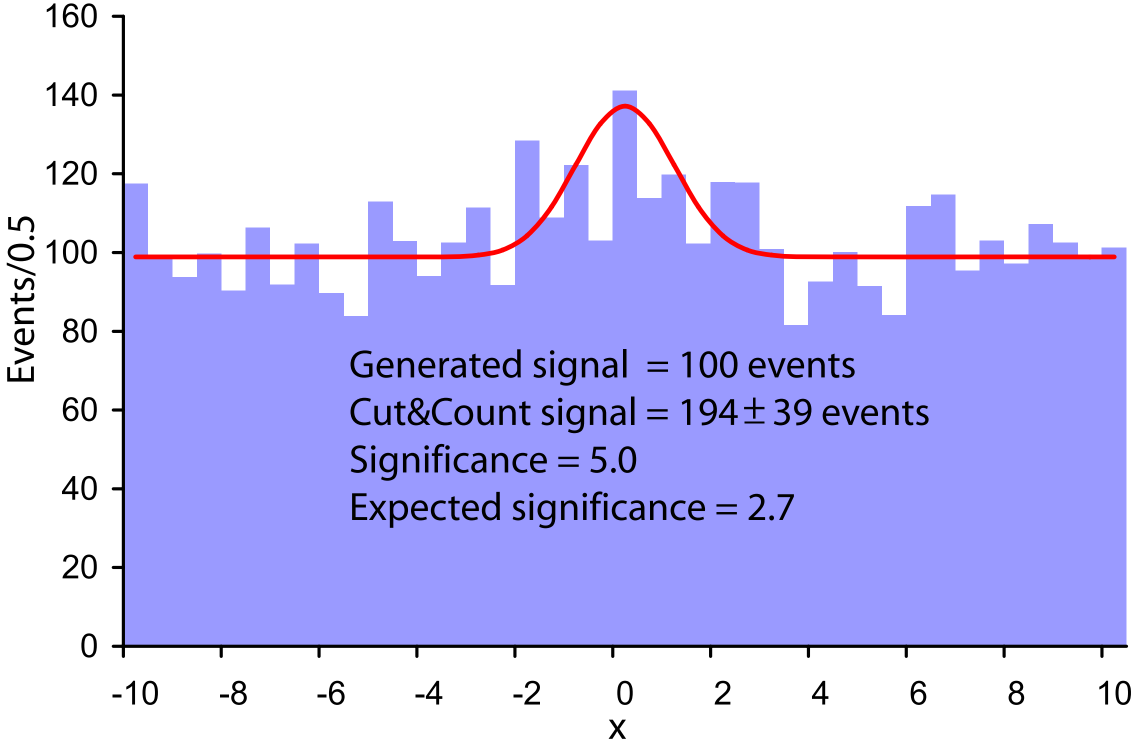
Alternative hypothesis H_1 : There is a new effect.

Reject the null hypothesis (that is, claim a new effect) if the observation falls in a region that is “unlikely” if the null hypothesis is correct. “Significance” (as typically used in HEP) is the probabilty that we erroneously reject the null hypothesis. Also called “confidence level”, or “P-value”, or the probability of a “Type I error”.

In HEP practice, we usually define the rejection region based on the observation, by taking it to be the region for which an observation is no more likely than the actual observation.

Computation of significance: Example

Flat normal background (avg 100/bin) + gaussian signal (100 events) with mean 0, $\sigma = 1$



Distribution is normal here, so $P = 5.7 \times 10^{-7}$ (two-tail).

Note: H_0 is flat distribution with no signal.

H_1 is $N_{\text{sig}} \neq 0$.

Note: The significance is not obtained by dividing the signal estimate (194) by the uncertainty in the signal (39), $194/39 = 5.0$. That would be akin to asking how likely a signal of the estimated size would be to fluctuate to zero. It is a good approximation in this example, however, because B/S is large.

What about Systematic Uncertainties?

$$B(\text{Nobel prize}) = 10 \pm 1 \pm 5$$

❑ They may be important!

- Maybe the ± 5 is a systematic uncertainty in the estimate of the background expectation. A “10σ” statistical significance is really only a “2σ” effect.

❑ They may be irrelevant

- Maybe the ± 5 is a systematic uncertainty on the efficiency, entering as a multiplicative factor. It makes no difference to the significance whether the result is 10 ± 1 or 5 ± 0.5.

❑ They may be “fuzzy”

- E.g., how is the background expectation estimated? What is the sampling distribution?
- E.g., Are “theoretical” uncertainties present?

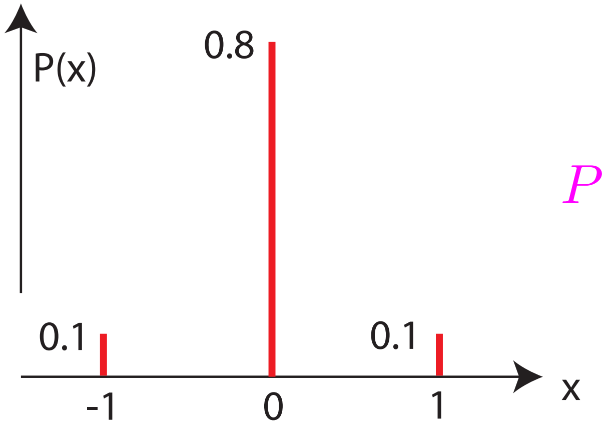
Aside: Significance as "$n\sigma$"

HEP parlance is to say an effect has, e.g., "5σ" significance.

At face value, this means the observation is "5 standard deviations" away from the mean:

$$\sigma \equiv \langle (x - \bar{x})^2 \rangle.$$

But we often don't really mean this. Note that a 5σ effect of this sort may not be improbable:



$$P(|x - \bar{x}| = 5\sigma) = 20\% \ !$$

Aside: Significance as "$n\sigma$" (continued)

Instead, we often mean that the probability (P-value) for the effect is given by the probability of a fluctuation in a normal distribution 5σ from the mean, i.e.,

$$\begin{aligned} P &= P(|x| > 5),\ \text{for } x \in N(0,1) \\ &= 5.7 \times 10^{-7} \end{aligned}$$

(two-tailed probability).

Also now popular to call $-2\Delta \ln \mathcal{L}$ the "n" in "$n\sigma$".

But sometimes we really do mean 5σ, usually presuming that the sampling distribution is approximately normal. [This may not be an accurate presumption when far out in the tails!]

Desirable to be more concise by quoting probabilities, or "P-values" as is common in the statistics world. At least say what you mean!

Pitfalls

What are the dangers? In a nutshell:

Unknown or unknowable sampling distributions

Ways to not know the distribution:

- The Improbable Tails
- Systematic Unknowns
- The Stopping Problem
- The Bump Hunt Conundrum

The Improbable Tails

Our earlier example was known to be normal sampling.

Often, this is true approximately (central limit theorem).

But for significance, often interested in distribution far into the tails. The normal approximation may be very bad here!

If there is any doubt, need to compute the actual distribution. Typically this is done with a "toy Monte Carlo" to simulate the distribution of the significance statistic.

To get to the tails, this may require a fair amount of computing time.

Still need to be wary of pushing calculation beyond its validity as a model of the actual distribution.

BaBar (and others) now routinely performs these calculations.

Systematic Unknowns

❑ Nuisance parameters

- Unknown, but relevant parameters. Estimated somehow, but with some uncertainty. Even if sampling distribution is known, cannot in general derive exact P-values in lower dimensional parameter space.
- Central Limit Theorem is our friend.
- Can try other values besides best estimate of nuisance parameter.

❑ "Theoretical" systematic uncertainties. Guesses, no sampling distribution.

- Use worst case values when evaluating significance.
- Or, give the dependence, e.g., as a range.

The Stopping Problem

There is a strong tendency to work on an analysis until we are convinced that we got it "right", then we stop.

Simple example: "Keep sampling" until we are satisfied.

Motivate our example:

- Ample historical evidence that experimental measurements are sometimes biased by some preconception of what the answer "should be". For example, a preconception could be based on the result of another experiment, or on some theoretical prejudice.
- A model for such a biased experiment is that the experimenter works "hard" until s/he gets the expected result, and then quits. Consider a simple example of a distribution which could result from such a scenario.

Stopping Problem: Normal likelihood function example

- Consider an experiment in which a measurement of a parameter θ corresponds to sampling from a Gaussian distribution of standard deviation one:

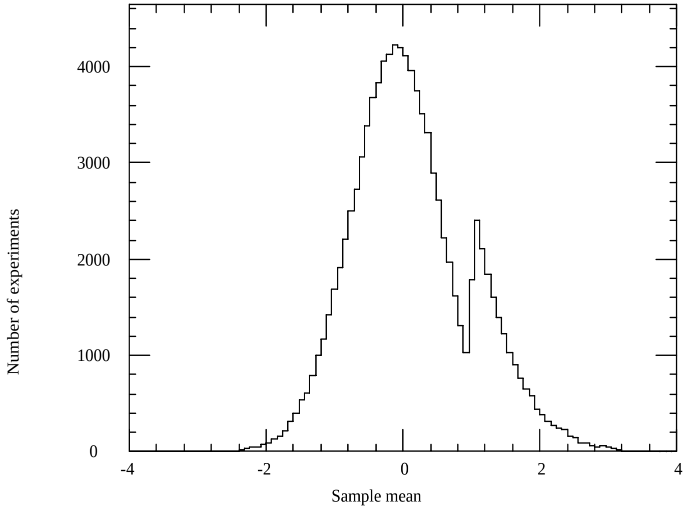
$$N(x;\theta,1)dx = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} dx.$$

- Suppose the experimenter has a prejudice that θ is greater than one.
- Subconsciously, he makes measurements until the sample mean, $m = \frac{1}{n}\sum_{i=1}^{n} x_i$, is greater than one, or until he becomes convinced (or tired) after a maximum of N measurements.
- The experimenter then uses the sample mean, m, to estimate θ.

Stopping Problem: Normal likelihood function example

For illustration, assume that $N = 2$. In terms of the random variables m and n, the pdf is:

$$f(m,n;\theta) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(m-\theta)^2}, & n=1,\ m>1 \\ 0, & n=1,\ m<1 \\ \frac{1}{\pi} e^{-(m-\theta)^2} \int_{-\infty}^{1} e^{-(x-m)^2}\, dx & n=2 \end{cases}$$

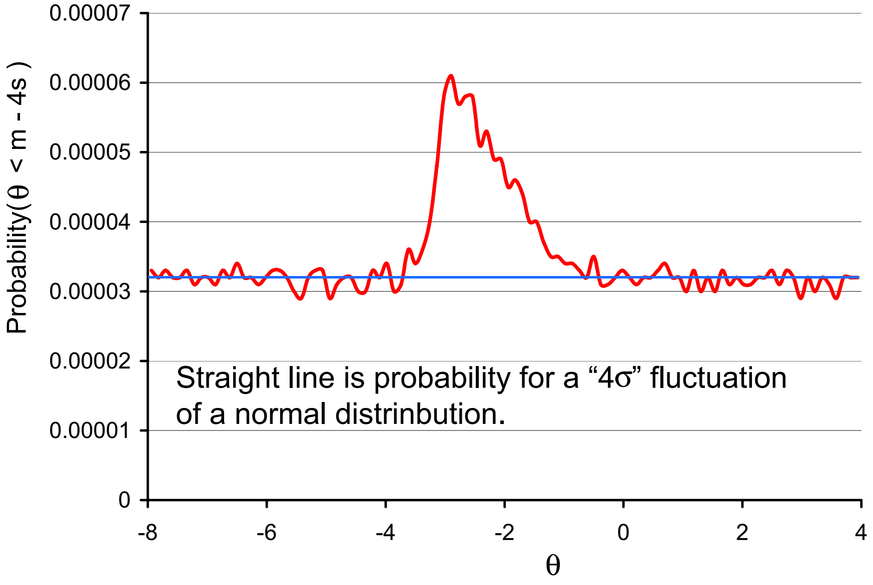


Histogram of sampling distribution for m, with pdf given by above equation, for $\theta = 0$.

The likelihood function, as a function of θ, has the shape of a normal distribution, given any experimental result. The peak is at $\theta = m$, so m is the maximum likelihood estimator for θ.

Stopping Problem: Normal likelihood function example

In spite of the normal form of the likelihood function, the sample mean is not sampled from a normal distribution. The “4σ” tail is more probable (for some θ) than the experimenter thinks.



Straight line is probability for a “4σ” fluctuation of a normal distrinbution.

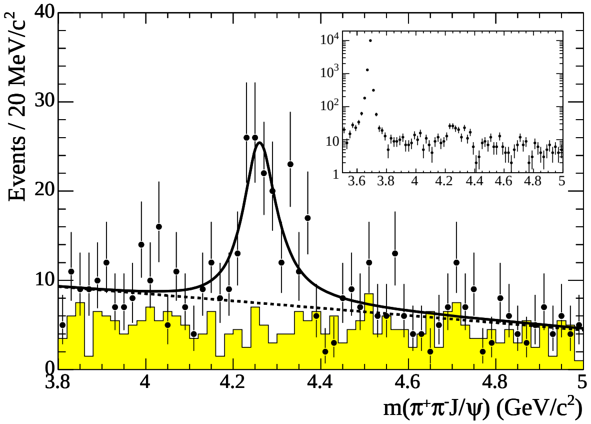
Stopping Problem: Normal likelihood function example

- The experimenter in this scenario thinks he is taking n samples from a normal distribution, and makes probability statements (e.g., about significance) according to a normal distribution.
- He gets an erroneous result because of the mistake in the distribution.
- If the experimenter realizes that sampling was actually from a non-normal distribution, he can do a more careful analysis to obtain more valid results.

Related effect: First "observations" tend to be biased high.
Example? BES early $f_D = 371$ vs $f_D \sim 220$ MeV more recently. Nothing "wrong" with this, but best estimates should average in earlier "null" data. Importance of reporting null results and quoting combinable results, e.g., two-sided confidence intervals.

The Bump Hunt Conundrum

- Context is the typical bump-hunting approach in HEP
- That is, analysis is developed concurrent with looking at the data
- We see a “bump”, how significant is it?



BaBar

$e^+e^- \to \gamma_{\mathrm{ISR}}\pi^+\pi^- J/\psi$

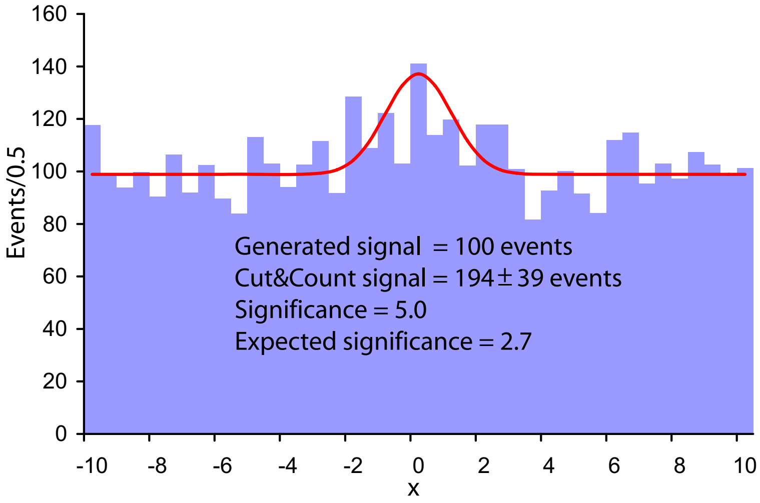
$P(H_0) \sim 10^{-11} - 10^{-16}$

- If the analysis is our typical bump-hunt approach, it is impossible to compute the significance!
- We don’t know the sampling distribution (under null hypothesis).
- Hence, we cannot compute probabilities (under null hypothesis, in particular).

Bump Hunting – Example I

Recall our earlier example:

Flat normal background (avg 100/bin) + gaussian signal (100 events) with mean 0, $\sigma = 1$



Distribution is normal here, so $P = 5.7 \times 10^{-7}$ (two-tail).

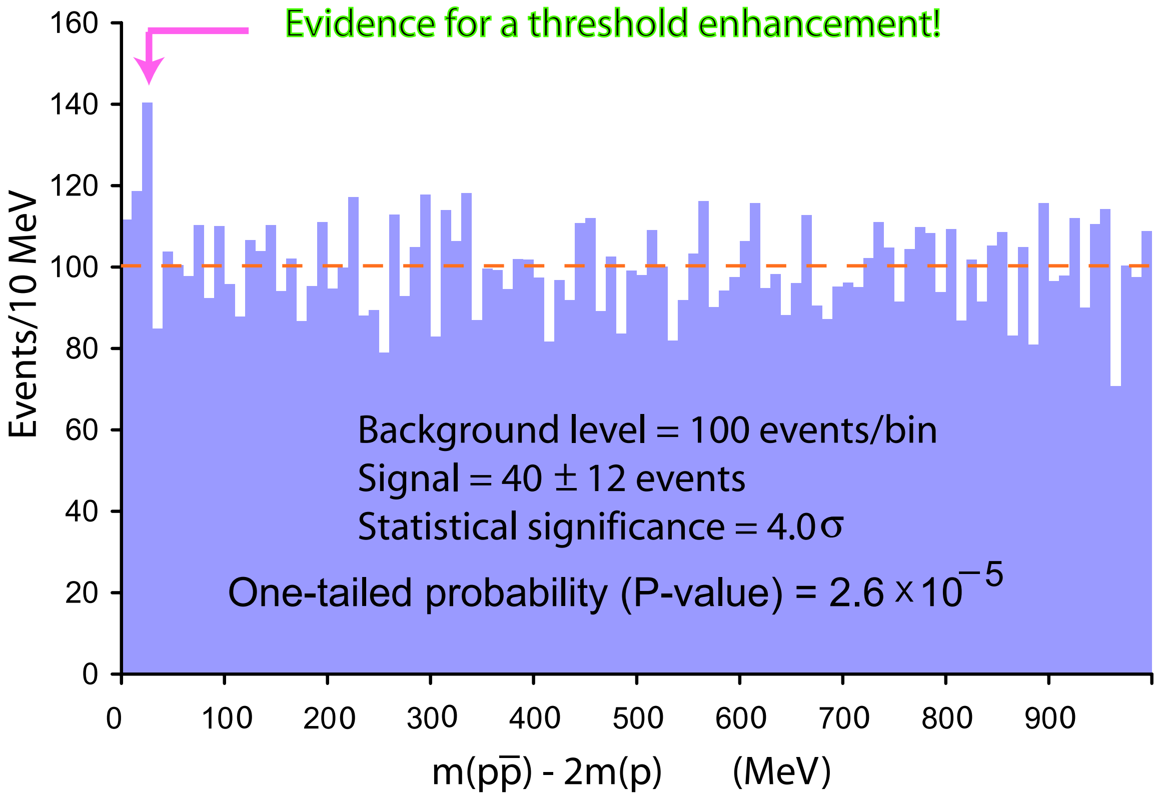
Note: H_0 is flat distribution with no signal.

H_1 is $N_{\text{sig}} \neq 0$.

As long as we knew at the outset that we were looking for an effect with mean 0 and $\sigma = 1$, we make correct inferences.

Bump Hunting – Example II

Looking for a mass bump.



Evidence for a threshold enhancement!

Background level = 100 events/bin
Signal = 40 ± 12 events
Statistical significance = 4.0σ

One-tailed probability (P-value) = 2.6×10^{-5}

So, what does it mean?

- ❏ Even though we cannot estimate probabilities, we still quote numbers intended to relay an idea of significance, usually quoted as some number of “sigma”.
- ❏ What we (usually) mean by this is:
  “If I had done the analysis in a controlled manner, and had been interested in the observed value for the mean and width, then the null hypothesis would require a fluctuation of this number of standard deviations of a Gaussian distribution to produce a bump as large as I see”

Is it useful?

- ❏ With this understanding of the meaning, it is perhaps not completely useless, since we can interpret it in the context of our experience.
- ❏ However, our experience is that we are sometimes fooled!
  pentaquarks example
- ❏ Life is short, we have also made great discoveries with this approach.
  Just remember that quoted significances are highly misleading.

Note: The "threshold enhancement" in Example II is a statistical fluctuation! The sampling distribution was the same $N(100, 10)$ for all bins.

Avoiding pitfalls

❑ Model those tails!

❑ “Conservatism”

  – Based on our experience with the failings of our methodology, we don’t claim new discoveries lightly.

❑ Do it better: Be Blind

It can be done “better”

- ❏ Better means with more meaningful significance estimates, and avoidance of bias.
- ❏ Do a “blind analysis”
  - – Blind means you design the experiment (analysis) before you look at the results.
- ❏ Goal is to know the sampling distribution.

BaBar routinely blinds it’s analyses, when there is a well-defined quatntity (e.g., CP asymmetry or branching fraction) being measured.

Blinding Methodologies

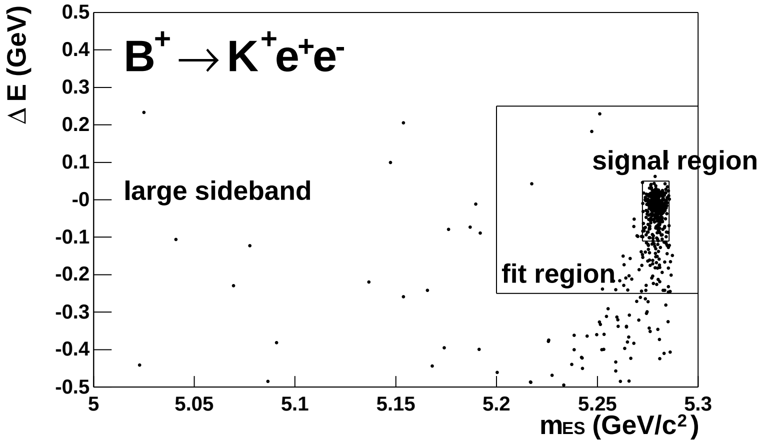
There are several basic approaches to blinding an analysis, which may be chosen according to the problem. Many variations on these themes!

- ❏ Don't look inside the box
- ❏ You can look, but keep the answer hidden
- ❏ Obscure the real data (e.g., adding simulated signal to data)
- ❏ Design on a dataset that will be thrown away
- ❏ Divide and conquer (e.g., BNL muon $g-2$ result, ω_p (magnetic field) and ω_a (muon precession) were analyzed independently, before combining to obtain $g-2$; arXiv:hep-0401008v3)

[Reference: Klein & Roodman, Ann.Rev.Nucl.Part.Sci. 55 (2005) 141.]

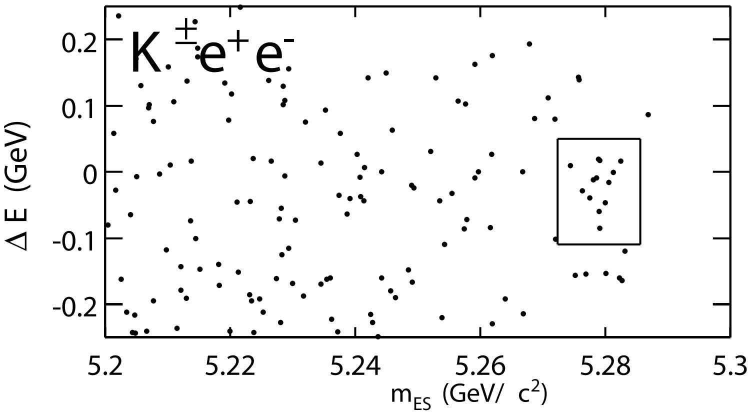
Blind Analysis – Don't look inside the box

- In this approach, the analysis is designed with the help of simulations, control samples, and sidebands.
- The data that will be fit for the result is kept invisible, until the analysis is deemed fixed.



Fit region & large sideband

(signal MC)

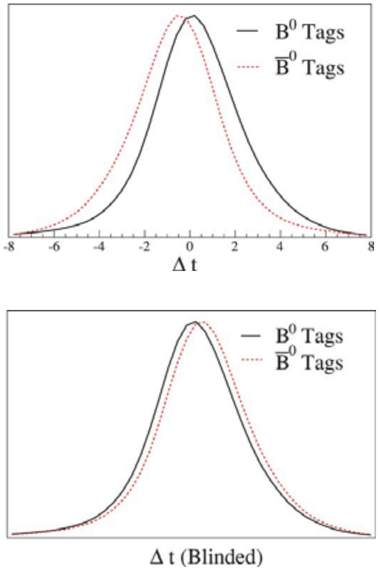


Unblinded fit region

(BaBar data)

Blind Analysis – Hiding the answer

- ❑ In this approach, no data is hidden, just the “answer” (typically, a number).
- ❑ For example, a hidden, possibly random, offset may be applied to the real answer to prevent it from being seen before the analysis is designed.



$$\Delta t(\text{Hidden}) = \begin{pmatrix} 1 \\ -1 \end{pmatrix} (\text{TAG})\Delta t + \text{Offset},$$

Top: Not hidden; Bottom: Hidden

BaBar, from Klein& Roodman, op. cit.

where TAG = ±1 according to the B flavor tag. This algorithm permits viewing the decay time distributions without revealing the asymmetry.

Blinding a bump hunt

- No different in principle than our other blind analyses.
- The more you know (about the signal you are interested in), the better you can do.
- Yes, it is harder. It takes discipline, and it might even cost "sensitivity".
  - E.g., a simple approach is to tune analysis on data that will be thrown away. Ignore any "signals". Or tune on them...
- But it is at least worth thinking about.
- What if something goes wrong:
  - Fix known mistakes.
  - State deviation from planned protocol in publication.

Conclusions

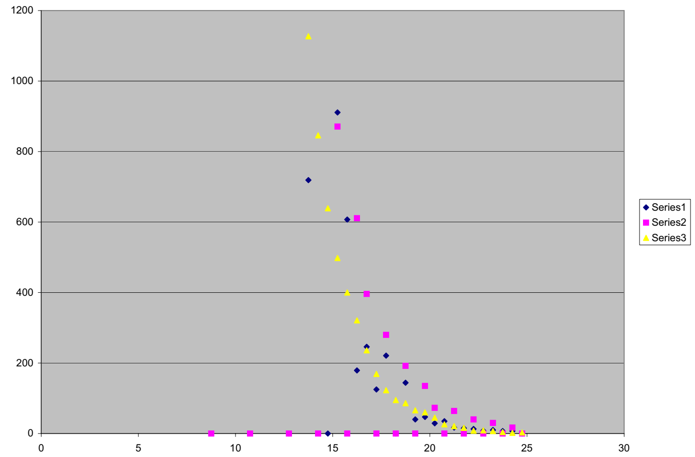
- Evaluation of significance is a “hypothesis test”. It is essentially the same problem as evaluating confidence intervals.
  - Except for the more obvious role played by improbable tails.
- Pitfalls amount to (not) knowing the sampling distribution.
- Techniques exist to avoid pitfalls:
  - Simulating the sampling distribution
  - Vary the nuisance/theory parameters
  - Blind your experiment
- Doing it properly requires patience and discipline; the benefit is a more meaningful, convincing result to yourself and to others.

Backup Slides

Toy Monte Carlo Computation of Tails

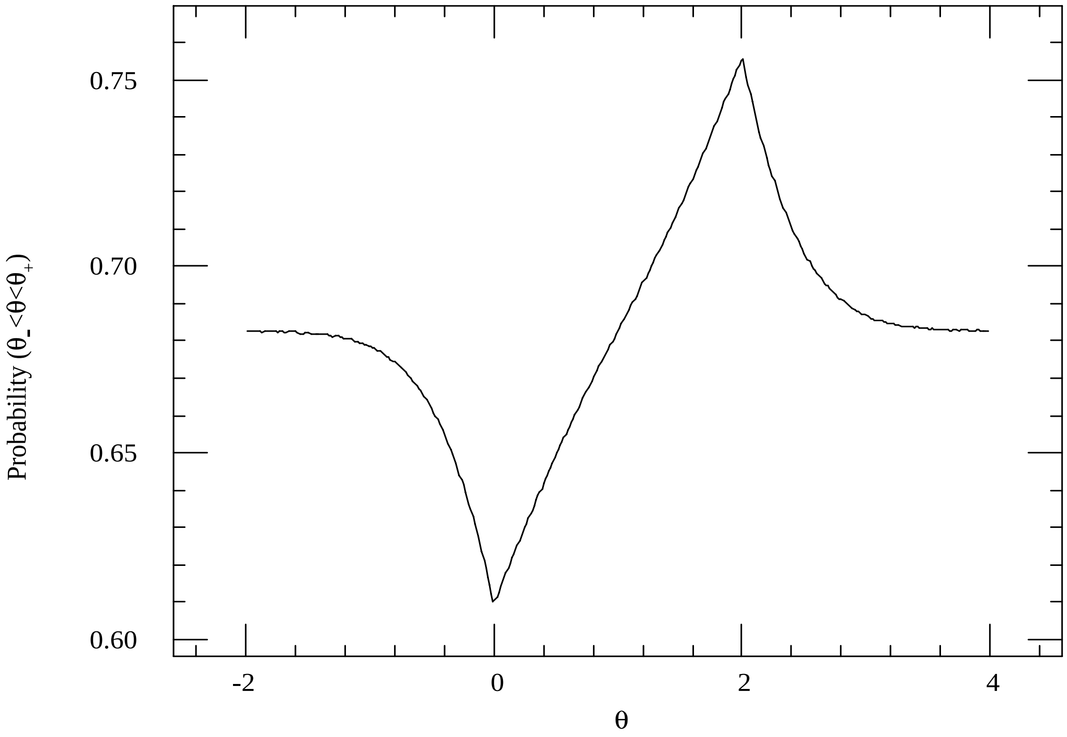
Example of counting register latches, where register is reset at fixed intervals (binomial sampling).

Compute distribution of three statistics: (a) likelihood ratio; (b) chisq; (c) chisq for “equivalent” normal

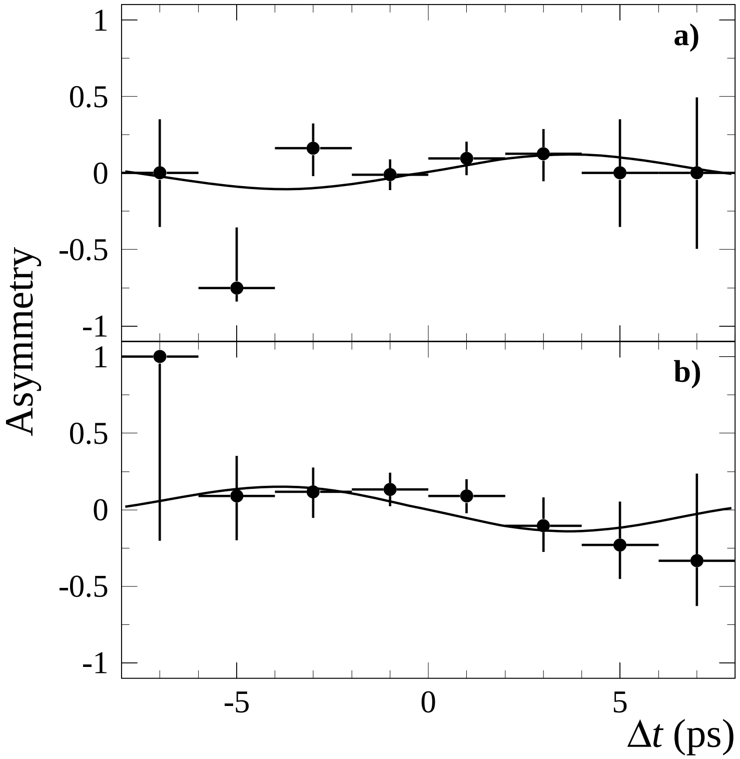


Stopping Problem: Normal likelihood function example

The likelihood function, as a function of θ, has the shape of a normal distribution, given any experimental result. The peak is at $\theta = m$, so m is the maximum likelihood estimator for θ. In spite of the normal form of the likelihood function, the sample mean is not sampled from a normal distribution. The interval defined by where the likelihood function falls by $e^{-1/2}$ does not correspond to a 68% CI:



Result of blind CP measurement



BaBar, Phys.Rev.Lett 86 (2001) 2515.

Blinding a bump hunt

❑ Decisions to make up front:

  - Where are we going to look? e.g., what invariant mass(es) are we interested in looking for something new?
  - What cuts do I make to select the data for the plot(s)?
  - Once I have the plot(s), how am I going to examine it for effects?
  - Plan for evaluating systematics.
  - The monkey wrench: What am I going to do if something goes awry?

Where are we going to look?

- ❑ Completely up to us what looks like it might be interesting.
  - – Could be "green-field" search, e.g., any mass, width regarded as interesting.
  - – Could be specific, e.g., another experiment saw something, or a theory predicts something.
- ❑ Just define the scope at the beginning.

What cuts do I make?

- ❏ Again, the more we know, the better off we are!
  - If we have good simulation of background processes, or background data, that's great! We'll use it.
  - Remember that the purpose of selection is to improve sensitivity to a possible signal.
  - The less we know, the bigger the risk that we won't be optimal. That's life. But being non-optimal doesn't mean wrong.
- ❏ One approach: Sacrifice a portion of data for cut optimization.
  - Ignore any "signals". Or tune on them. . .
  - If signal is real, it should show up in remaining blinded data. If it is a fluctuation, it will disappear.
  - Systematic effects will show up both places. But that's another matter.

Once I have the plot(s), how am I going to examine it for effects?

- Define in advance what structure will be interesting. Any mass? Constraints on width? Possibility of overlapping peaks?
  - In less constrained searches, requires automated search technology.
- How do we parameterize the null hypothesis (that is, the background)?
  - Again, the more you know, the easier this is.
  - If you don't know much, you can still specify general structure (e.g., order of polynomial) to cover smooth variations.