

International Journal of Modern Physics A
© World Scientific Publishing Company

The Significance of HEP Observations

Frank C. Porter

*Physics Department 356-48, California Institute of Technology
Pasadena, CA 91125, USA
fcp@hep.caltech.edu*

Received Day Month Year

Revised Day Month Year

The subject of determining the significance of observations in particle physics is discussed. Several dangers are identified, all related to the problem of knowing the sampling probability distribution. Techniques for mitigating these pitfalls are presented.

Keywords: Statistics; significance.

PACS numbers: 02.50.-r, 02.50.Tt

1. Introduction

The discovery of a new phenomenon must be critically evaluated before acceptance. One criterion for acceptance is “statistical significance”: the observation is unlikely to be due to statistical fluctuation of known processes. It is desirable to quantify the measure of statistical significance in terms of probabilities, and there are standard approaches for this. However, there are several difficulties, both in principle and in execution. We review these difficulties, and suggest approaches to mitigate them.

A frequentist treatment is given. This is the appropriate context for descriptive statistics. For reporting significance measures, descriptive statistics seems called for, leaving the (Bayesian) interpretation up to the consumer of the information.

2. Significance as Hypothesis Test

It is conventional to use the language of hypothesis testing when discussing tests of significance. Physicists often avoid this terminology and even this perspective; however, it would likely prevent much confusion to simply adopt it. The “null hypothesis”, H_0 , is the hypothesis that there is no new effect. The “alternative hypothesis”, H_1 , is the hypothesis that there is a new effect. If the observation is sufficiently unlikely to occur in the null hypothesis, then we reject H_0 in favor of the alternative.

Statisticians define a “rejection region” corresponding to a given significance level, α : This is a region of sampling space which has probability α under the null hypothesis. Formally, α is known as the probability of making a “Type I error”,

2 Frank C. Porter

that is, the probability of rejecting the null hypothesis if the null hypothesis is true. In high energy physics 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. In this case, α is called the “ P -value”.

Consider a histogram drawn from a sampling distribution with a “flat background” with independent bin contents distributed according to $N(100, 10)$, plus a Gaussian “signal” of (exactly) 100 counts centered at $x = 0$ and standard deviation one. A particular sampling from this distribution is shown in Fig. 1a.

We do a simple cut-and-count fit to the data in Fig. 1a, using the sidebands ($|x| > 3$) to estimate the background. The background is subtracted from the observed counts in the signal region ($|x| < 3$) yielding a signal estimate of 194 ± 39 events. The significance (P -value) of this signal is given by the probability of obtaining a signal estimate at least as large (in absolute value for a two-tailed test) as that observed, where this probability is computed according to the null hypothesis that the data is sampled entirely from the background distribution:

$$H_0 : N_{\text{signal}} = 0; \quad (1)$$

$$H_1 : N_{\text{signal}} \neq 0. \quad (2)$$

In this example, this probability is $P = 5.7 \times 10^{-7}$, the probability of having a > 5 standard deviation fluctuation of a normal distribution.

Note that we have had an upward fluctuation of the estimated signal in Fig. 1a, due to a background fluctuation (since the signal is actually fixed). The expected signal is 100 events, corresponding to an expected significance of 2.7σ .

It should be stressed, since there is a common confusion, that 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 addressing how likely a signal of the estimated size would be to fluctuate to zero. It turns out to be a good approximation in this example only because the background-to-signal is large.

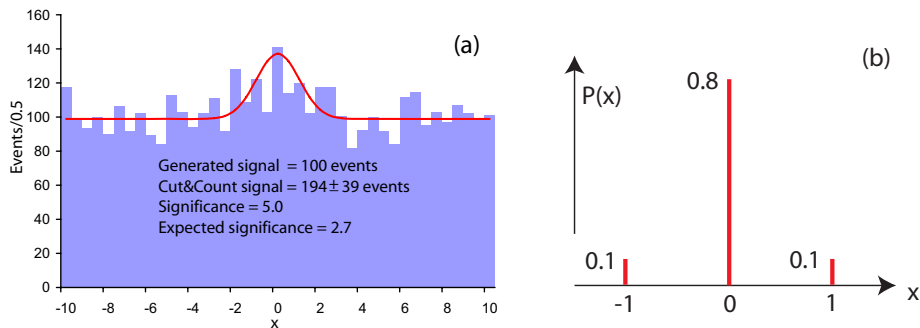


Fig. 1. (a) An example to demonstrate the computation of significance. (b) A sampling distribution with mean zero and standard deviation $\sigma = 0.2$, with a probability of 20% to encounter a 5σ fluctuation from the mean.

3. The Pitfalls

There are a variety of pitfalls that we encounter in trying to evaluate whether we have observed something new or not. All share a common theme: They involve a lack of understanding of the sampling distribution. We discuss several categories.

3.1. *The improbable tails*

Physicists like to report significance as “ $n\sigma$ ”, that is as a “number of standard deviations”; I have reflected this usage in the discussion so far. Unfortunately, usage is inconsistent, and it is suggested that more care be put into this reporting.

Quoting significance as “ $n\sigma$ ” implies that the observation is n standard deviations away from the value expected under the null hypothesis, where a standard deviation is computed according to $\sigma = \sqrt{\langle (x - \langle x \rangle)^2 \rangle}$. But we often don’t really mean this. Fig. 1b shows a distribution for which the probability of finding a 5σ effect is 20%. In spite of our intuition, 5σ fluctuations are not necessarily improbable.

When we quote a significance as $n\sigma$, we sometimes mean that the null probability (P -value) is given by the probability of a $n\sigma$ fluctuation of a normal distribution, i.e., $P = P(|x| > 5)$ for x from $N(0, 1)$, or $P = 5.7 \times 10^{-7}$ (two-tailed).

But often, we really do mean 5σ , usually presuming that the sampling distribution is approximately normal. It has also become popular in recent times to compute the change in log-likelihood, $\lambda \equiv -2\Delta \ln \mathcal{L}$, and call this “ $n\sigma$ ”. The two likelihoods compared are the value from the best fit assuming the alternative hypothesis H_1 and the maximum under the null hypothesis. The distribution of this statistic may be computed under the null hypothesis. But often it is quoted without investigating this probability, in the hope that it corresponds approximately to the normal case.

The example of Fig. 1a was known to be normal sampling. Often this is true approximately, by the central limit theorem. However, in computing significances we may be far into the tails of the distribution, and the assumption of normality may be unjustified even if the distribution is close to normal near the mean.

If there is any doubt, we need to compute the actual distribution of the significance statistic under the null hypothesis. This is usually done by simulation, with a “toy Monte Carlo”. To compute the tails may require a large amount of computing time. The possibility that even the simulation does not give an accurate representation of the tail distribution should be considered.

3.2. *Systematic unknowns and nuisance parameters*

Often a measurement of interest depends on the measurement of auxiliary quantities, such as backgrounds and efficiencies. These are called “nuisance parameters”; we don’t really care about them, but they are needed for the result of interest. The uncertainties due to the uncertain values of the nuisance parameters are treated as “systematic errors” and quoted separately. We might see a branching fraction measurement quoted as: $B(\text{new effect}) = 10 \pm 1 \pm 5$, where the ± 1 is the “statistical

4 *Frank C. Porter*

error”, and the ± 5 is the “systematic error”. This may represent a highly significant observation or not, depending on the source of the systematic uncertainty.

Perhaps the systematic uncertainty is from an uncertainty in the background subtraction. In this case, it is an additive uncertainty – the significance of the deviation from zero branching fraction is only “ 2σ ”. Alternatively, the systematic error could be due to uncertainty in the signal efficiency. In this case, it is a multiplicative uncertainty, and has little bearing on the significance of the deviation from zero, since 5 ± 0.5 is as significant as 10 ± 1 .

A fundamental difficulty is that even if the sampling distribution for the nuisance parameter estimators is known (except for the value of the nuisance parameter), it is not possible in general to derive exact P -values in a lower-dimensional parameter space. A notable exception is the multivariate normal distribution, where it is possible to compute probabilities in a lower dimension. However, not all of our measurements are normal to a good enough approximation, and the problem becomes more difficult. An approach to coping with this difficulty is to compute the distribution of the significance statistic assuming other true values of the nuisance parameter(s), besides the best estimate value.

Still worse is when the distribution of the nuisance estimators is not even known. An example here is “theoretical” uncertainties, for which no distribution exists (in the frequency sense), only some judgement. In this case, it may be important to repeat the significance computation for the range of plausible theoretical values. The worst-case value could then be used in quoting the significance, or alternatively the dependence of the significance on the uncertain parameter(s) may be described.

3.3. *The stopping problem*

There is a strong tendency to work on a measurement until we are convinced that we got it “right”, and then stop. We’ll investigate an example to illustrate how this can undermine estimates of significance. Our example is motivated by historical evidence that experimental measurements are sometimes biased by a preconception of what the answer “should” be. Such a preconception could be based on the result of another experiment or on theoretical prejudice.

Suppose we do an experiment to measure a parameter θ corresponding to the mean of a Gaussian distribution of standard deviation one. Suppose further that we have a prejudice that $\theta > 1$. Subconsciously, we make measurements until the sample mean, $m = \frac{1}{n} \sum_{i=1}^n x_i$, is greater than one, or until becoming convinced after N measurements. We then use the sample mean as an estimate of θ .

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

$$f(m, n; \theta) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(m-\theta)^2}, & n = 1, m > 1 \\ 0, & n = 1, m < 1 \\ \frac{1}{\pi} e^{-(m-\theta)^2} \int_{-\infty}^1 e^{-(x-m)^2} dx, & n = 2. \end{cases} \quad (3)$$

This distribution is shown in Fig. 2a.

The likelihood function, as a function of θ , has the shape of a normal distribution, given any experimental result. The peak is at $\theta = m$, hence the sample mean is the maximum likelihood estimator for θ . However, in spite of the form of the likelihood, the sample mean is not sampled from a normal distribution. This is a not widely-appreciated distinction: The form of the likelihood function does not imply the form of the sampling distribution. In frequency statistics, it is the sampling distribution that is crucial to computing probabilities. The oft-touted suggestion that showing the likelihood function provides a complete picture is not generally valid.

The implication for computing significance is that treating the sample mean as sampled from a normal distribution will give an incorrect result. This is illustrated in Fig. 2b, in which the probability of an apparent 4σ fluctuation is plotted as a function of the value of the parameter θ . Depending on θ , the actual probability may be as much as twice as likely as the experimenter thinks.

In our scenario we think we are taking n samples from a normal distribution, and make probability statements (about significance) according to a normal distribution for the sample mean. We get an erroneous result because of the mistake in the distribution. If we realize that the sampling (of the sample mean) was actually from a non-normal distribution, we can do an analysis to obtain more valid results. However, the context of the example is that we do not, in fact, realize this.

There is a familiar, related phenomenon: First “observations” of a new process tend to be biased high. Especially in exploratory investigations, null results tend not to be reported quantitatively. The first claim of a new effect will preferentially occur with a positive fluctuation. This suggests the importance, in summarizing knowledge, of averaging in earlier null results to get best estimates. It further calls for reporting of results permitting averaging, rather than only quoting upper limits.

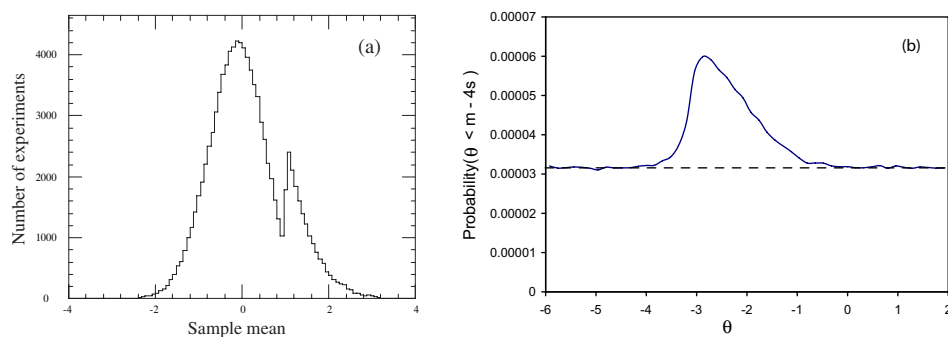


Fig. 2. (a) Histogram of the sampling distribution for the sample mean, according to Eqn. 3, for $\theta = 0$.¹ (b) The probability to observe a “ 4σ ” fluctuation in the stopping problem example as a function of the distribution parameter θ . The dashed line shows the probability for a 4σ fluctuation of a normal distribution.

6 *Frank C. Porter*

3.4. The bump hunt conundrum

In a typical HEP exploratory search for unexpected phenomena, a histogram is made and examined for unexplained structure. We call this sort of search “bump hunting”. When we see an interesting structure (e.g., Fig. 3a), the first question is usually “Is it significant?”.

Quite often, the analysis in a bump hunt is developed concurrently with looking at the data. In this case, it is impossible to compute the significance of an effect. The problem, once again, is that we don’t know the sampling distribution under the null hypothesis, and thus we cannot compute probabilities.

Recall the example of Fig. 1. We assume that we know the (flat) sampling distribution under the null hypothesis, and that we are looking for an effect described by a Gaussian of mean zero and standard deviation one. As long as these assumptions are correct, we make correct inferences of significance.

Now consider the example in Fig. 3b. It shows a “mass distribution” plotted in bins that are larger than the resolution. The bump hunting here consists of looking for one-bin peaks, as signs of narrow structure. The observed histogram looks flat except for a bump near threshold. It is desired to know the significance of this peak.

The usual approach to computing the significance is to estimate the background level under the peak, and then compute, under the null hypothesis of no signal, the probability of a fluctuation to the observed level. Here, the background level is estimated from the sidebands to be 100 events, with negligible statistical uncertainty. The excess in the peak (“signal”) is 40 ± 12 events. With a background standard deviation of 10 events (the counts in each bin were generated from a Gaussian distribution), the probability of a ≥ 40 event upward fluctuation is $P = 3.2 \times 10^{-5}$.

So, our result looks very significant. But did we really model the sampling distribution properly? No. First, we admit that we would have accepted a fluctuation as interesting in any bin. Thus, we should divide our probability estimate by 100. Our probability is no longer so spectacular, although it is still small. However, our second admission is that this was not the only histogram we looked at. Perhaps we

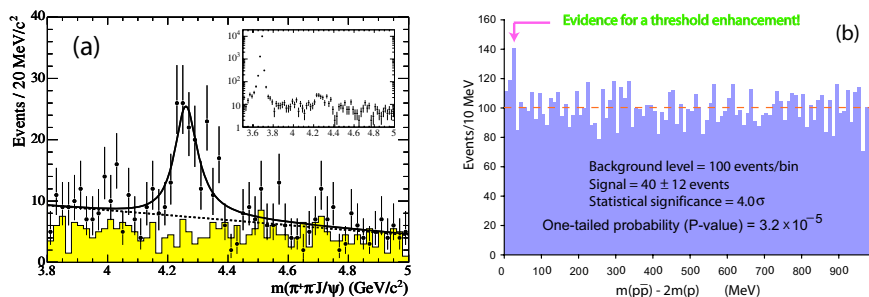


Fig. 3. Bump hunting: (a) The $\pi^+\pi^-J/\psi$ mass spectrum in the initial state radiation process $e^+e^- \rightarrow \gamma\pi^+\pi^-J/\psi$ from the BaBar experiment.² (b) Spectrum of the $p\bar{p}$ mass from threshold to 1 GeV. The dotted line indicates the estimated background level.

varied the cuts, or looked at other masses, until we found something interesting.^a

The significances we quote for bumps in exploratory analyses usually aren't P -values. We still give these numbers, generally as number of "sigma"s. What we really mean is: "If I had done a controlled analysis, and had been interested in the observed values 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." With this understanding, it is perhaps not utterly useless, as we can interpret it in the context of experience. However, we are occasionally reminded that experience is that we may be fooled. The pentaquark story is, I think, an example. We have also made great discoveries with this approach. Just remember that the quoted significances are highly misleading as quantitative statements.

One way we mitigate this pitfall is "conservatism", applied to the interpretation rather than to the computation. That is, we make up for the unreliability of our probability calculation by requiring very great computed significance before claiming a new effect. Thus, " 3σ " isn't regarded as especially unlikely, even though the implied probability under the null hypothesis is only 0.13% for a one-tail test.

4. Blinding the Analysis

An important technique to avoid the pitfalls is to "blind" the analysis. The goal of blinding is to ensure a known sampling distribution under the null hypothesis. Several approaches to blinding exist, which may be chosen as appropriate to the problem (for a review, see Ref. 3). We list several used in high energy physics:

- (1) Hide the answer in a box, don't look inside until ready.
- (2) You can look, but keep the answer hidden via an unknown transformation.
- (3) Obscure the real data. For example, let the data be visible, but add simulated signal to it, to be removed only when the analysis methodology is final.
- (4) Design the analysis on a dataset that will be discarded.
- (5) "Divide and conquer": The idea is to separate the analysis into pieces that will be combined to get the quantity of interest only when the pieces are final.

A nice example of this methodology is the muon anomalous magnetic moment measurement⁴. In this experiment, teams independently measured the magnetic field and the muon precession; neither alone gives a clue to the value of $g - 2$.

We'll elaborate on a couple of these to give the flavor.

4.1. Blinding 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. An illustration of is the measurement of the

^aIn this example, I re-generated the simulated experiment until I got a " 4σ " effect somewhere, and then stopped. The sampling distribution was $N(100, 10)$ for each bin.

8 *Frank C. Porter*

process $B^\pm \rightarrow K^\pm e^+ e^-$ in BaBar.⁵ Figure 4a shows a scatterplot of simulated signal data in the relevant two kinematic variables. The “large sideband” region is the only visible region of the data prior to fixing the analysis. Note that all of the data that will be used in the fit for the results is kept blind, including the (smaller) sidebands around the signal region. The final unblinded data is shown in Fig. 4b.

4.2. *Hiding the answer*

Sometimes we can let the data be visible in graphical form, but obscure the numerical result of interest. For example, a hidden, perhaps random, offset may be applied to the real answer to prevent it from being seen during the analysis design. An example is the BaBar CP -violation measurement.³ Fig. 5 shows the Δt distributions, in which CP violation appears as a difference between the B^0 and \bar{B}^0 “tags” and as an asymmetry about 0. These asymmetries are obscured from view while the analysis is being tuned, by applying a transformation to the two Δt distributions.

4.3. *Blinding a Bump Hunt*

In section 3.4 we discussed the difficulty of evaluating significance in a “bump hunting” exploratory analysis. For example, consider the search for structure in some invariant mass spectrum. Can the methods of blind analysis be applied to this situation so that meaningful measures of significance may be computed?

There is no difference in principle in this case from our other blinding examples. The difficulty arises from the broad range of phenomena one may be interested in discovering. We might not know the location, width, kinematic regime, or even the particle content of interest. The analysis must encompass all interesting possibilities.

One relatively straightforward approach to this problem is to divide the data up into two samples, one used to design the analysis and one to use for the results. The dataset used for the analysis design is discarded once the analysis is designed. This costs sensitivity, but can be an effective method. Note that the details of how

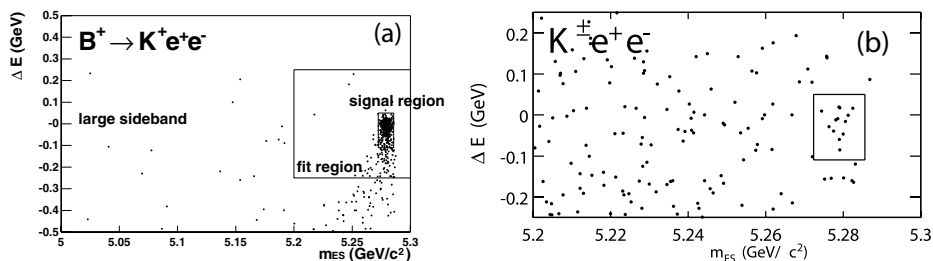


Fig. 4. Illustration of the blinding methodology in BaBar’s $B \rightarrow K e e$ analysis. (a) Signal Monte Carlo in the two relevant kinematic variables, showing the large sideband region, the fit region, and the region where most of the signal is concentrated. (b) Data after unblinding, in the fit region. The small box is the signal region as defined in the left plot.

one uses the design sample don't much matter; one can even tune cuts in ways that enhance apparent peaks. If there is actually no effect, then the most probable result will be that nothing significant is observed when the blinded sample is observed.

5. Conclusions

The problem of computing significance in frequentist statistics is really the same as computing other probabilities, such as confidence intervals or goodness-of-fit. However, the tails of the distribution often play a larger role in evaluating significance, and extra care is required to ensure that the sampling distribution is both known and properly computed.

Acknowledgments

I am indebted to my BaBar colleagues for several examples and stimulating discussions. This work is supported in part by DOE grant DE-FG02-92-ER40701.

References

1. F. C. Porter, *Nucl. Inst. Meth.* **A368**, 793 (1996).
2. BaBar Collab. (B. Aubert *et al.*), *Phys. Rev. Lett.* **95**, 142001 (2005).
3. J. Klein and A. Roodman, *Ann. Rev. Nucl. Part. Sci.* **55**, 141 (2005).
4. Muon g-2 Collab. (G. W. Bennett *et al.*), *Phys. Rev. D* **73**, 072003 (2006).
5. BaBar Collab. (B. Aubert *et al.*), *Phys. Rev. Lett.* **91**, 221802 (2003).

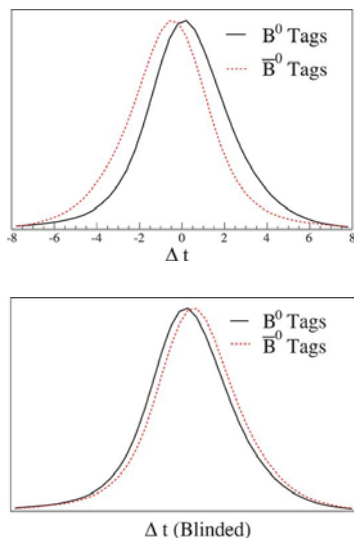


Fig. 5. Example of blinding an analysis by hiding the answer via a transformation on the data.³ Top: unblinded result; Bottom: blinded data.