Systematic uncertainties in Upper Limits

J. Baudot, A. Di Canto, D. Greenwald, G. Inguglia, K. Kinoshita, T. Kuhr, F. Le Diberder, *D. Tonelli*, E. Prencipe, B. Yabsley

coll-statistics@belle2.org

Physics Week
Dec 3, 2020 (online)

What this talk will try to achieve

An (exclusion) limit is a measurement

Limits are just one-sided confidence (or Bayesian credibility) regions, that is, they are measurements, just as a standard central values +/- uncertainties (point estimates).

There's nothing exotic or special in the conceptual treatment of systematic uncertainties in limits that's not already in the treatment of systematic uncertainties in point estimates or two-sided confidence/credibility regions.

(In fact, because the small sample size typically dominates the precision in exclusion limits, systematic uncertainties are usually less relevant in limits than in two-sided confidence regions or point estimates — limits are never "precision measurements")

Systematic uncertainties isn't statistics

In fact, no statistics reference tells you which systematic sources to consider in a particular problem and how to determine their effects— this is up to the experimenter and her subjective judgement/understanding of the measurement process.

Statistical concepts just offer some conceptual means to guide their inclusion in the results and achieve a shared interpretation.

Rarely there is an unique, rigorously correct way to assess systematic uncertainties. Typically there are various reasonable ways.

Each has pros and cons that need to be balanced by the experimenter depending on the scientific goal at hand.

This is *not* to say that "anything goes": typically there are also (many more) wrong/unreasonable ways to assess systematic uncertainties.

I’d be happy if this talk could bring us closer to be able to (i) understand the implications and interpretations of some of the reasonable approaches (ii) identify (and discard) the wrong approaches

What this talk won't be

You won't get technical/empiric recipes, black boxes, or recommendations for tools

This is something many seek — I purposely choose not to provide it

- most ready-made recipes/tools/black boxes apply only to few simple cases — may not work as expected in your thesis problem. And — what is worse — you wouldn't know it.
- even if they apply to your problem, it's hardly a good idea to use casually something that impacts the results w/o appreciating limitations and implications.

There's a common prejudice that the statistical extraction of results (limit or else) is some sort of intellectual sophistry. Or, at best, a final technical appendix to the real scientific work, which is done elsewhere. I disagree.

Statistics is the language of science. It requires sufficient competence to ensure command of the techniques and understanding of implications.

You wouldn't trust black boxes while defining your analysis strategy, optimizing your selection, or defininig a fit model.

You shouldn't trust black boxes when dealing with statistical procedures.

Preliminaries

Preliminaries

Setting limits is part of statistical inference.

Central to any inference is the model **p(x|m)**, a mathematical construct that connects the physics parameter of interest m (i.e., cross section) with the observable quantity x (data, i.e. number of signal candidates).

The *assumption* of the model is the common assumption to all statistical approaches

When interpreted as a function of the data x (i.e., fixing $m = m_0$), $p(x|m_0)$ is the probability density function: expresses the probability for each data observation had the true value of the parameter been m_0.

When interpreted as a function of parameter m (fixing $x = x_0$) $p(x_0|m) = L(m)$ is the likelihood: expresses the likelihood to observe data x_0 for different choices of possibles m values.

Accelerated recap: Bayesian/frequentist inference

Measuring m consists in

(i) devising a model p(x|m) that conforms the phenomenon under study and connects the unobserved parameter m (i.e., cross section) to the observable quantity x (data, i.e. number of signal candidates in a distribution)

(ii) observing x

(iii) using the model and the data to infer information on m.

Bayesians — combine the model with prior probabilities for m to determine the posterior probability p(m|x), which expresses the probability for each value of parameter m given the data. ("Prior" == known or chosen before observing x)

Frequentists cannot define p(m|x)— they use the model and the probabilities for all other possible outcomes for x to determine for which values of m the model would produce the observed data x with highest probability

Don't want to spend time on the conceptual implications of each. We are free to choose either option provided that we do it in a conceptually consistent way and are aware of limitations and implications

Interpretation of results — coverage

An important aspect is to ensure proper communication of results. When a HEP paper reports an exclusion limit, we (consciously or unconsciously) tend to interpret the result as frequentist regardless of the approach used

PHYSICAL REVIEW LETTERS **125,** 161806 (2020)

(Received 26 July 2020; accepted 8 September 2020; published 14 October 2020)

We present a search for the direct production of a light pseudoscalar a decaying into two photons with the Belle II detector at the SuperKEKB collider. We search for the process $e^+e^- \to \gamma a$, $a \to \gamma\gamma$ in the mass range $0.2 < m_a < 9.7$ GeV/c^2 using data corresponding to an integrated luminosity of (445 ± 3) pb^{-1}. Light pseudoscalars interacting predominantly with standard model gauge bosons (so-called axionlike particles or ALPs) are frequently postulated in extensions of the standard model. We find no evidence for ALPs and set 95% confidence level upper limits on the coupling strength $g_{a\gamma\gamma}$ of ALPs to photons at the level of 10^{-3} GeV^{-1}. The limits are the most restrictive to date for $0.2 < m_a < 1$ GeV/c^2.

We expect that if ~100 experiments were to search for ALPS with our procedure, ~95 would report an exclusion range that does not contain the true value of the coupling strength (and ~5 an exclusion range that does contain it)

That is we implicitly assume “coverage”.

Hence, for proper communication of results it is generally believed (even by Bayesians HEP collegues) that coverage is a desirable property.

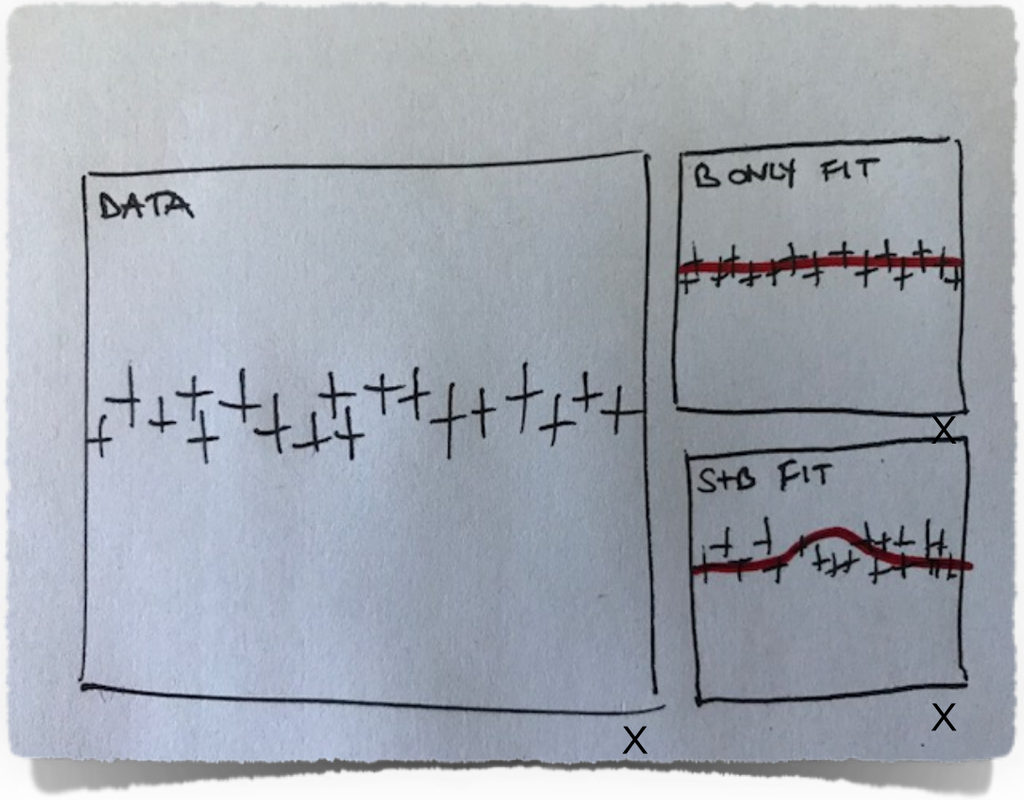
Toy limit example

Looking for a a signal over a background in a spectrum of data

Fit data with a model that allows for signal and background, “the (S+B) model”

$$p(x|N_s) = N_s[\text{Signal-bump}](x) + (1 - N_s)[\text{Flat-bckg}](x)$$

The estimate of the signal yield N̂s from data may be consistent or close to zero, which allows for setting an exclusion limit for the signal



(This is a toy example to illustrate the conceptual workflow — won’t spend time on peripheral details. Concepts would equally apply if it were a counting experiment instead of a fit, if data x and parameter m = Ns were multidimensional $\vec{x}$ and $\vec{m}$ etc.)

Bayesian limits

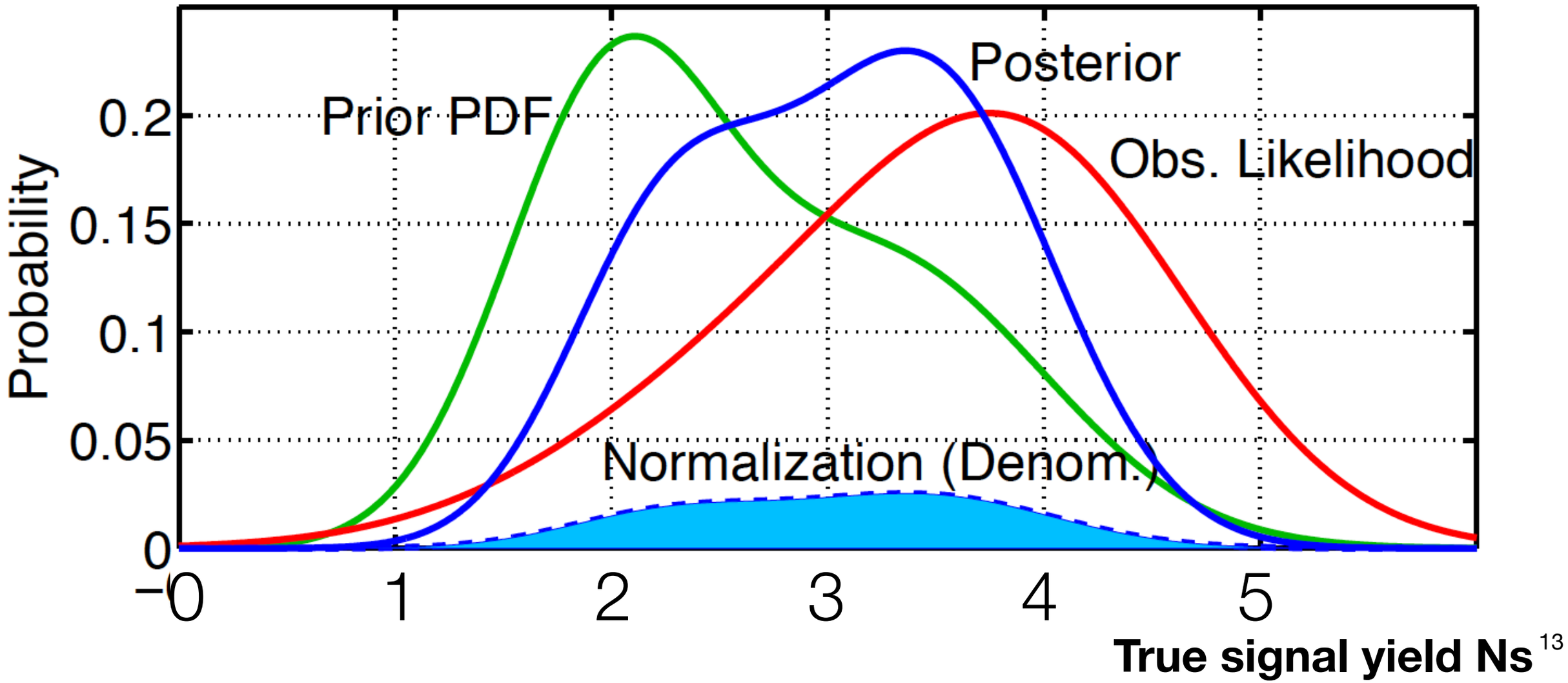
Probability for the parameter given the data

Likelihood of your data Prior probability (your assumption)

$$p(m|x) = \frac{p(x|m) \times p(m)}{p(x)}$$

Posterior probability

Normalization



Once you got the posterior, that's easy

Have likelihood L(Ns) = p(x|Ns) from data.
Assume prior p(Ns) [e.g., uniform]

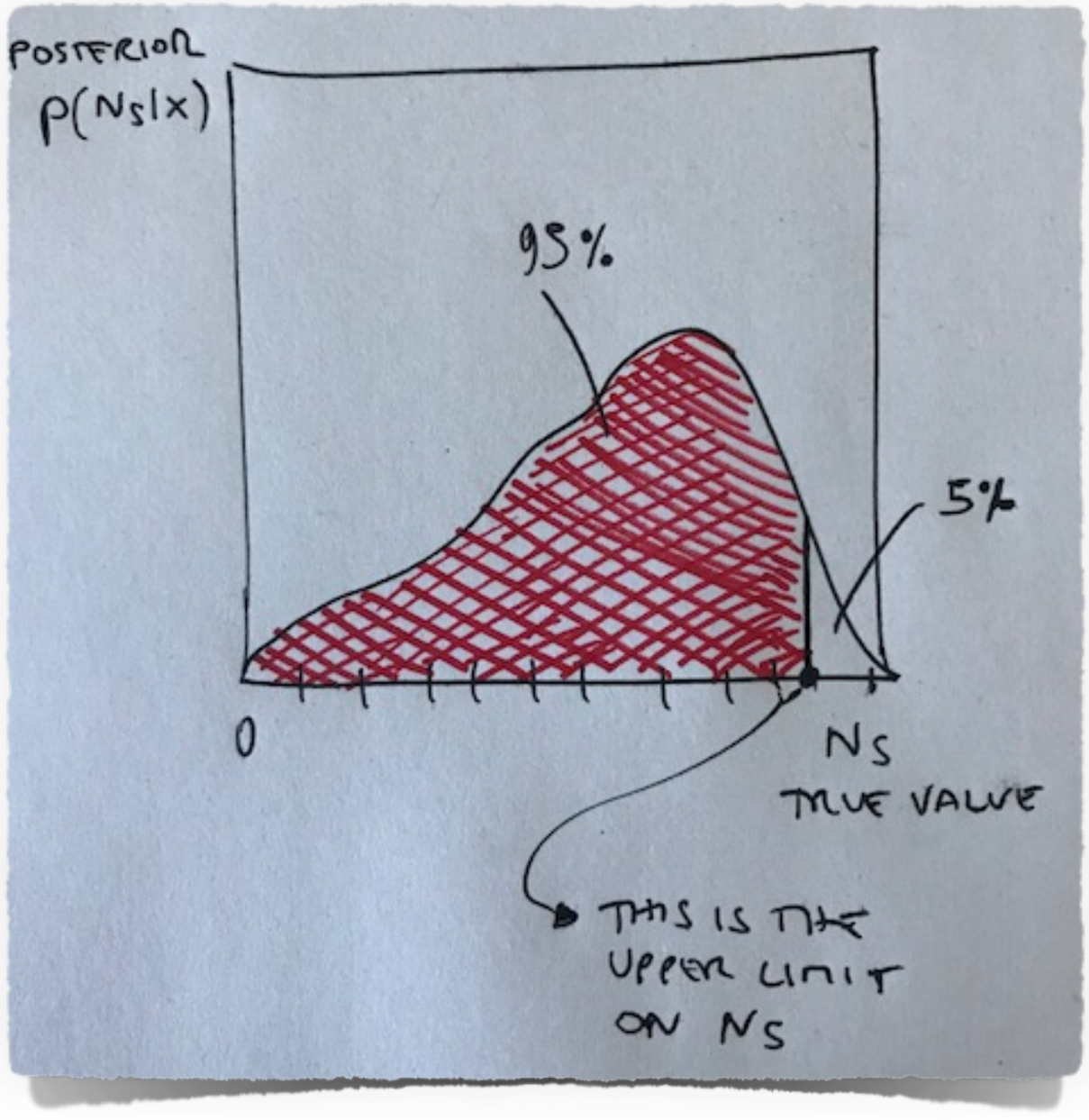
Integrate (marginalize) the posterior

Likelihood Prior

$$p(N_s|x) = \frac{p(x|N_s)p(N_s)}{\int_{N_s} p(x|N_s)p(N_s)dN_s}$$

until reaching a fractional area corresponding to the desired Bayesian credibility level, e.g. 95%.

(let's call the posterior's probability content "Bayesian credibility" — it keeps transparent the inherent subjectivity)



The corresponding signal yield defines the upper limit — may depend on prior.

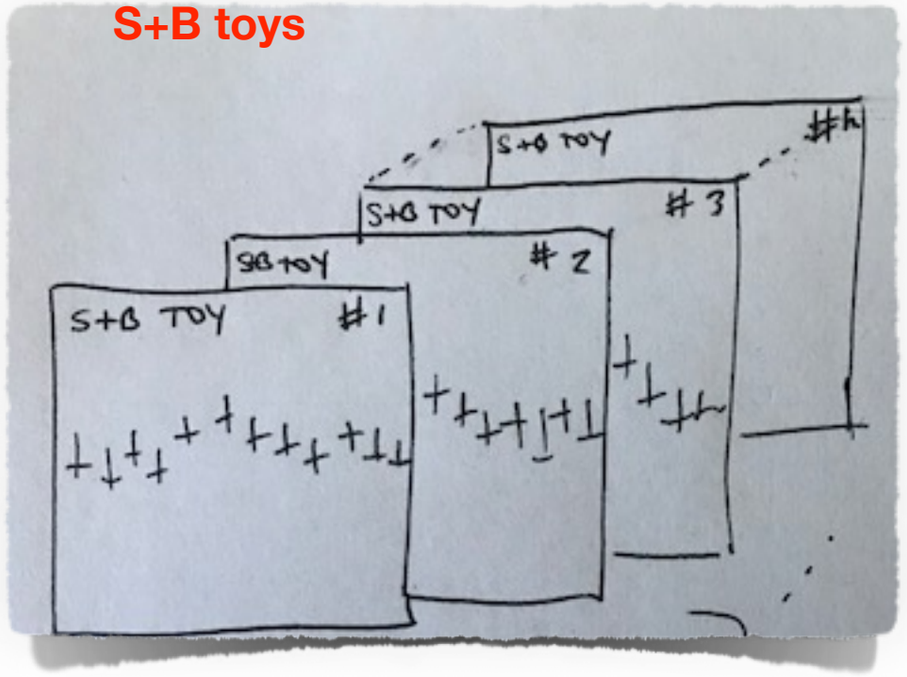
Frequentist limits

Back to our example — toys

Assume a signal strength N_s^j (a choice for the true value of signal yield Ns) and generate toy simulations by drawing pseudodata drawn from the S+B model, each of same luminosity as the experimental data.

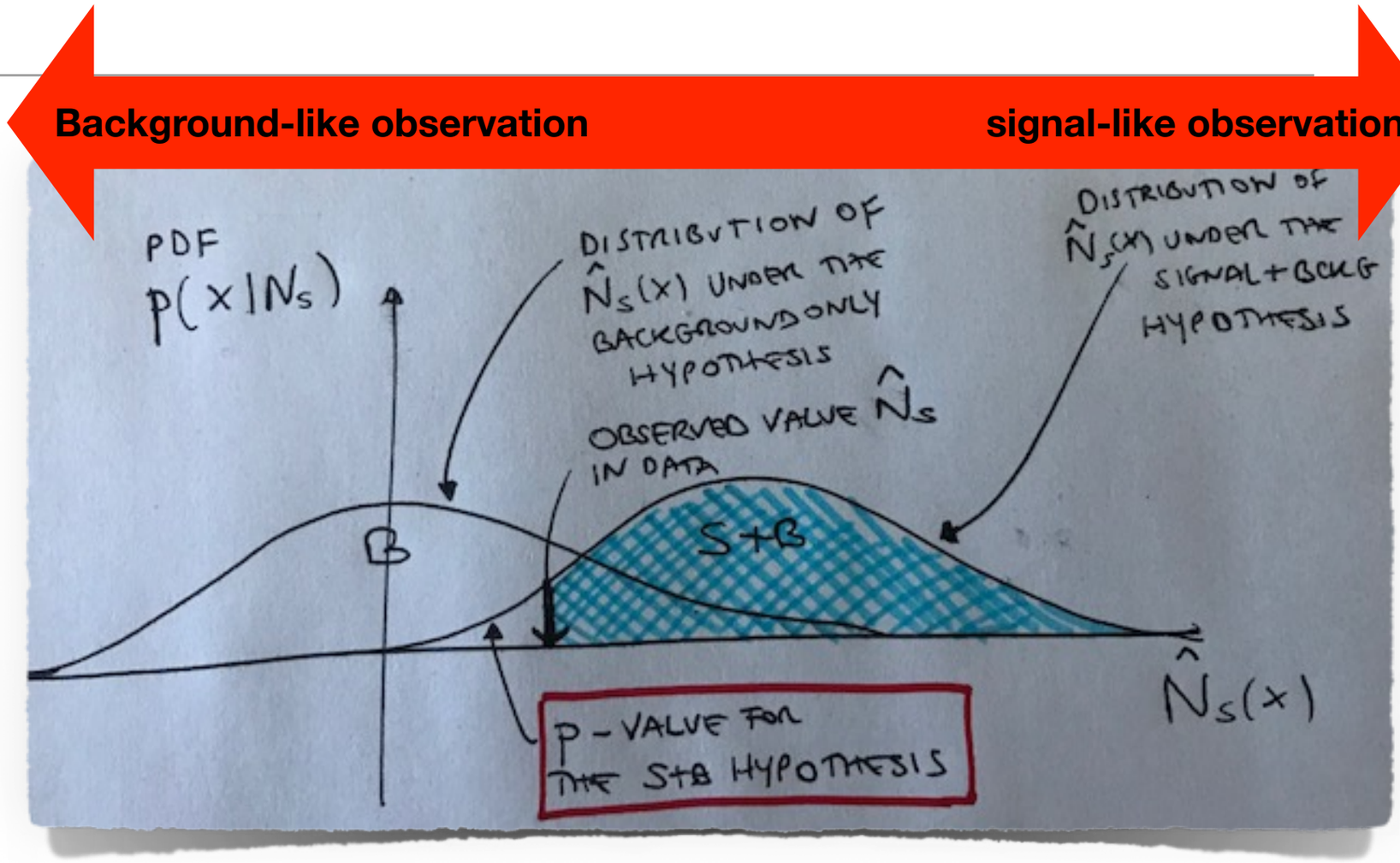
Fit each toy of each ensemble with the S+B model and get a result for $\hat{N}_s$

Then plot the distributions of the $\hat{N}_s$ results



p-value = 1- CL

The location of the data observation relative to the curves corresponding to the two hypotheses ("B-only" or "S+B") offers a measure of compatibility of the data with either.



The fractional integral of the S+B curve over values as bcgk-like as the one we observed, or more, is the p-value for the "S+B hypothesis".

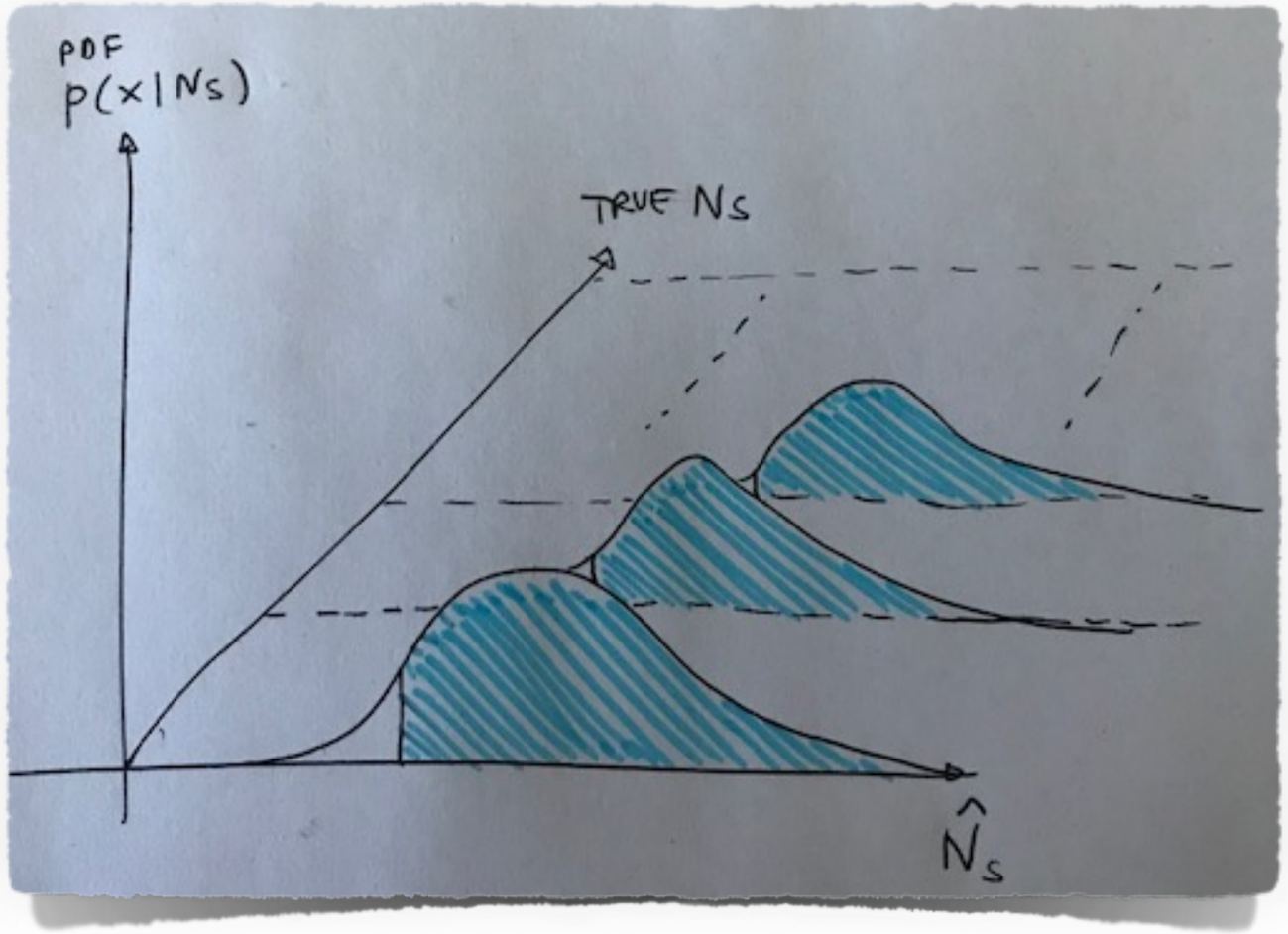
The smallest such p-value, the highest the incompatibility of our data with the S+B hypothesis: it would be unlikely to observe our data if model S+B was the one realized in nature. That is, our data disfavor the S+B model, or "exclude the S+B model at a confidence level CL = 1- p" (e.g., 95% CL if p-value = 5%)

Testing multiple signal strengths

We only tested one hypothesis of signal strength (i.e. our limit only excludes a specific BF value)

Typically one is interested to test a whole range of signal strengths (e.g, BF < 10^{-9}).

Repeat the previous procedure on multiple ensembles of toys, each ensemble generated assuming a different signal strength

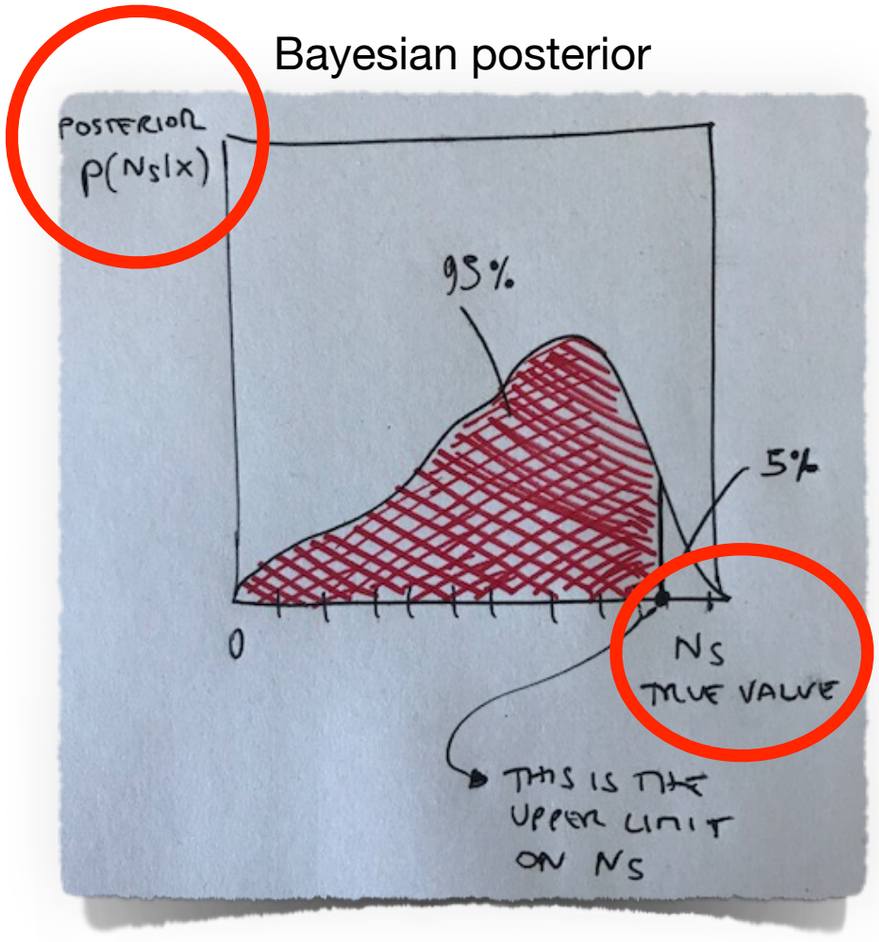


Your measurement excludes at 90% CL the signal strength yielding a p-value of 5% in data, and all higher strengths.

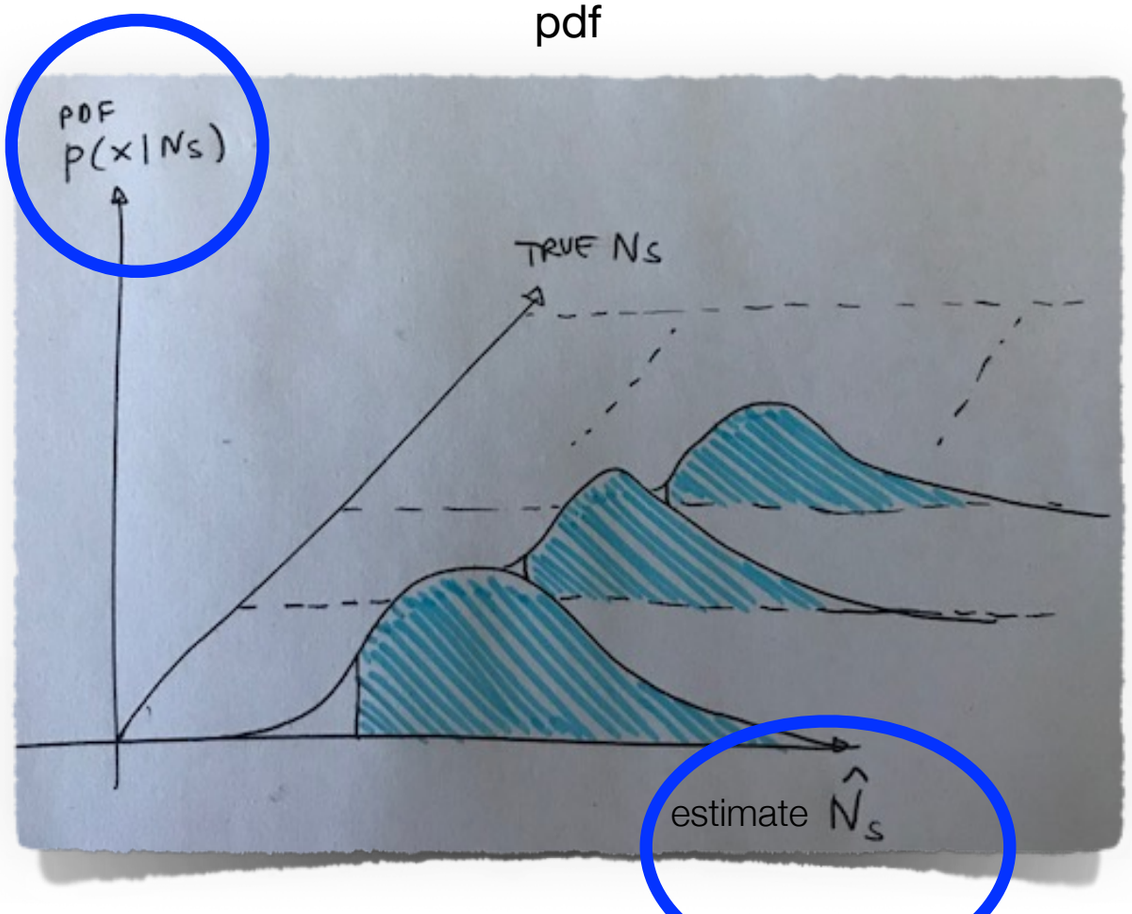
(This toy example illustrates the general, first-principle approach to construct a frequentist limit that is, to study the properties of your estimator in toys. Massive toy generation is often avoided by exploiting the asymptotic properties of the likelihoods — but those are all special cases of this general one)

Remember p(x|m) ≠ p(m|x)

Bayesian posterior



pdf



Curves above have different meanings

p(m|x) is the posterior probability for the parameter m: it expresses the probability that each value of m is true given the data. p(x|m) is the probability density function: it expresses the probability to observe data given a value of m.

This is important

P(A|B) is NOT equal to P(B|A).

Variable A: “pregnant”, “not pregnant”

Variable B: “male”, “female”.

P(pregnant | female) ~ 3% but

P(female | pregnant) >>> 3% !

Systematic uncertainties

The model

The model p(x|m) **is assumed**. One's best approximation/idealization of the actual relation between m and x relevant for the problem at hand.

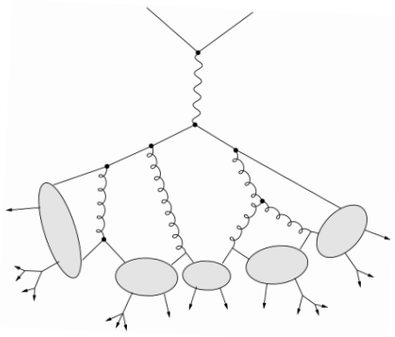
Monte Carlo modeling

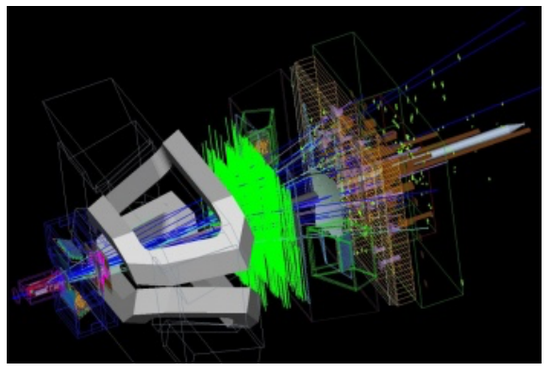
$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

$$+ \underbrace{\bar{L}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})L + \bar{R}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g'YB_{\mu})R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

$$+ \underbrace{\frac{1}{2}\left|(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})\,\phi\right|^{2} - V(\phi)}_{W^{\pm},Z,\gamma,\text{and Higgs masses and couplings}}$$

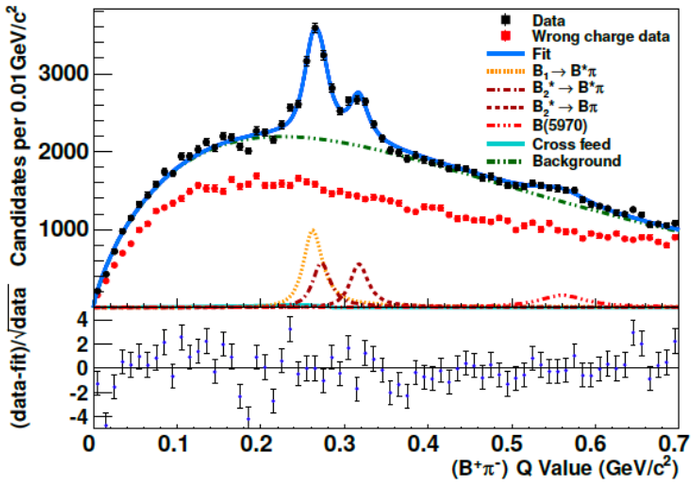
$$+ \underbrace{g''(\bar{q}\gamma^{\mu}T_{a}q)\,G^{a}_{\mu}}_{\text{interactions between quarks and gluons}} + \underbrace{(G_{1}\bar{L}\phi R + G_{2}\bar{R}\phi_{c}L + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$





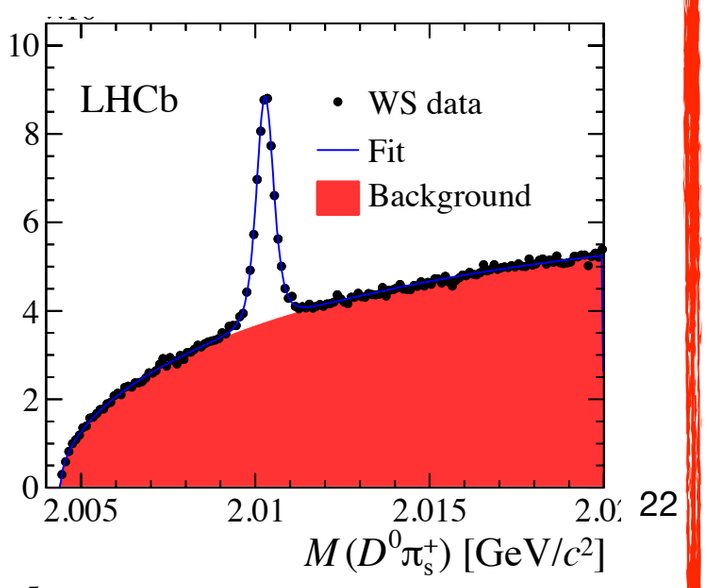
Data driven modeling

- Sideband subtraction
- Same-charge candidates
- Mixed-event candidates
- ABCD methods
- …



Effective modeling

Empirical modeling



Systematics = the model is imperfect

To account for the uncertainty associated with the model approximation/ idealization, allow for additional dependencies on unknown nuisance parameters in the model— p(x|m,s).

The values s are unknown and uninteresting but do influence the results.

Lack of knowledge of s introduces an uncertainty in the p(x|m,s) shape.

Not only you don't know exactly what value of x would be observed if m had a definite value, you don't even know exactly how probable each possible x value is.

The uncertainty *in the shape* of p(x|m) reflects into the systematic uncertainty of the inference.

So what?

Systematic uncertainties increase the dimensionality of our inference: they add unknown parameters.

“Well, I’ve got an additional unknown parameter. What’s the big deal?”

Do nothing and treat s as just an additional parameter to be inferred from the data (“let’s fit it”). E.g, determine a two-dimensional confidence region in the (m, s) space that then gets projected onto the space of the parameter of interes m to set the desired limit.

In practice this is rarely a good idea.

- Using the *finite* statistical information of our data to determine s in addition to m is a waste of statistical power that degrades the statistical precision on m.
- The technical and computational complexity of constructing confidence regions in high number of dimensions gets quickly intractable.

Hence, both frequentist and Bayesian devise methods to reduce the number of parameters to be inferred.

Bayesian treatment of nuisance parameters

Marginalization over the nuisance parameters

Bayesians reduce dimensionality by “averaging” over the space of the unknown nuisance parameters.

A straightforward generalization of the standard Bayesian treatment.

Assume a prior for any of the nuisance parameters s and integrate (oft-called “marginalize”) over the nuisance parameters to obtain a posterior that no longer depend on the nuisance parameters.

$$p(m|x) = \frac{\int_s p(x|m,s)p(m,s)ds}{\int_m \int_s p(x|m,s)p(m,s)dsdm} = \frac{P(x|m)p(m)}{\int_m P(x|m)P(m)dm} = \frac{P(x|m)p(m)}{P(x)}$$

The price to pay is an enhanced dependence on the subjective assumptions on the priors, which may get critical in high dimensions and may spoil the requirement of coverage that we all want for our results.

The point is that “averaging” the effects of an unknown, over an arbitrary/ subjective metric even, offers no guarantee that the final result will bracket the true value with the desired probability content.

Frequentist treatment nuisance parameters

Profiling the likelihood

Reduce the dimensionality of the problem by derivation: replace the likelihood with a lower-dimensional function obtained by maximizing the likelihood wrt the nuisance parameters (“profiling”).

For each point m_0 in the space of true m values, the likelihood $L(m_0,s)$ gets replaced by its profile-likelihood $PL(m_0,\hat{s}(m_0))$, where $\hat{s}(m_0)$ is the value that maximizes L(m,s) in that point.

The profile-likelihood is not a likelihood in that it does not meet all the nice mathematical properties of the likelihood but is demonstrated to approximate it very well in many cases.

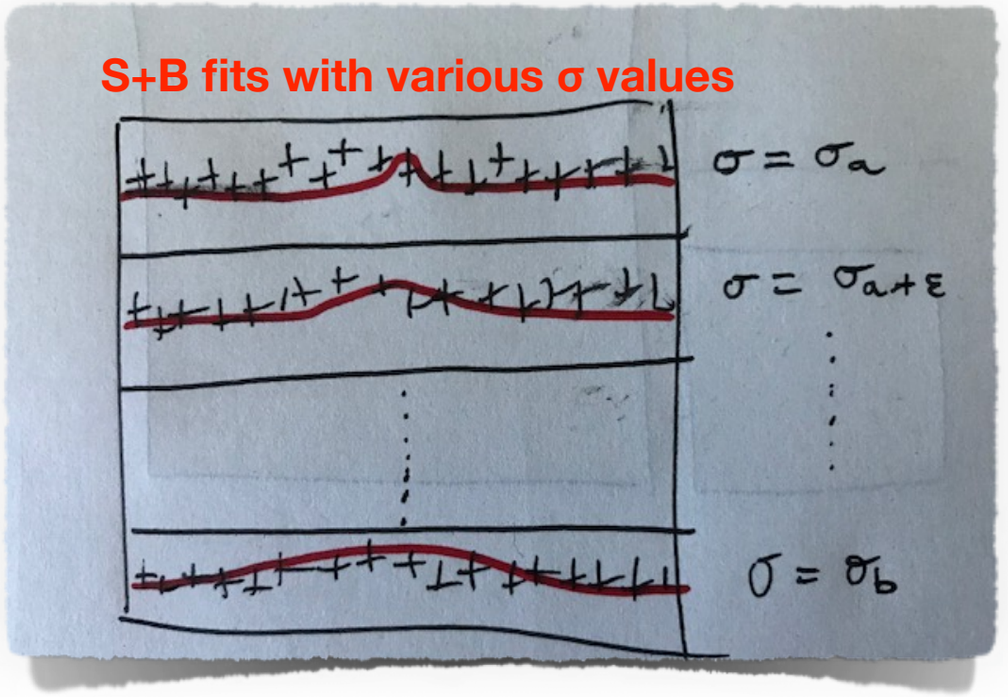
In practice

Coming back to our example

The S+B model now depends on an additional (nuisance) parameter: the detector resolution σ

Assume to know that σ ranges between the extremes σ_a and σ_b from simulation or control sample studies.

This is a common condition in many cases: even if we don't know the value of a nuisance parameter and its model, we are typically able to reliably bracket it within a range



$$p(x|N_s, \sigma) = N_s[\text{Signal-bump}](x, \sigma) + (1 - N_s)[\text{Flat-bckg}](x)$$

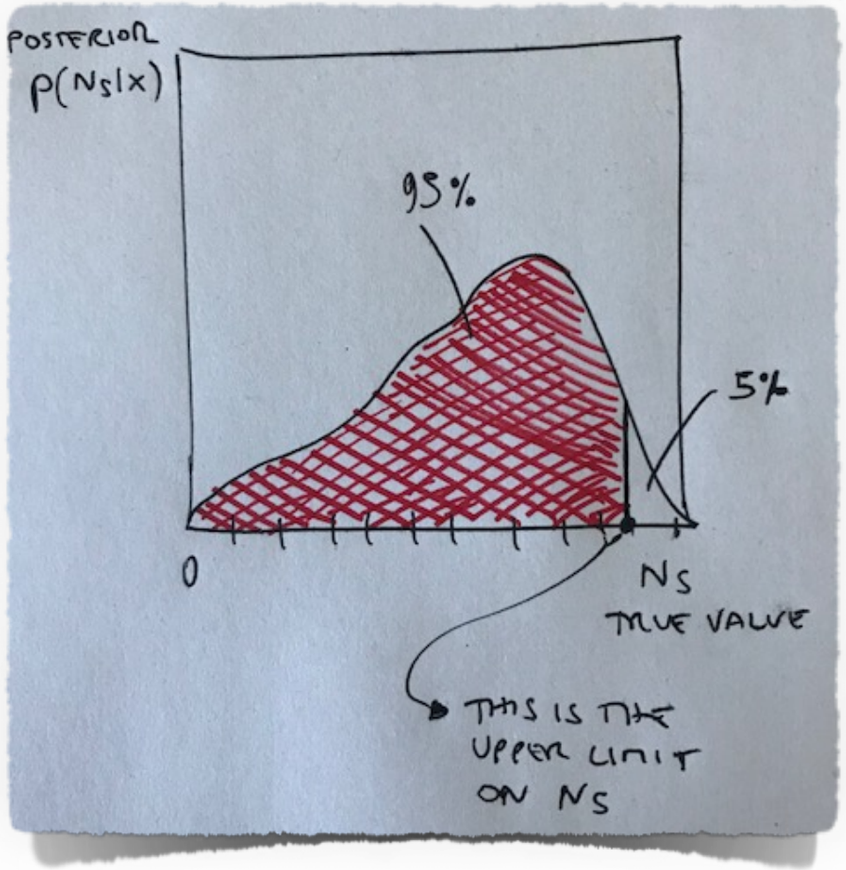
Bayesian case

Assume a prior for the distribution of σ between σ_a and σ_b and marginalize

$$p(N_s|x) = \frac{\int_{\sigma_a}^{\sigma_b} p(x|N_s,\sigma)p(N_s,\sigma)d\sigma}{\int_{N_s}\int_{\sigma_a}^{\sigma_b} p(x|N_s,\sigma)p(N_s,\sigma)d\sigma dN_s} = \frac{P(x|N_s)p(N_s)}{\int_{N_s} P(x|N_s)P(N_s)dN_s} = \frac{P(x|N_s)p(N_{N_s})}{P(x)} \quad (1)$$

Then continue as in the case of no systematics: integrate the posterior p(Ns|x) from zero until the value of the integral is 95%: the corresponding value of Ns is an upper limit on the signal yield at 95% CL.

Note that now the posterior p(Ns|x) differs from the no-systematics case, since it has been "marginalized" (think about a fancy 'average') not only over the values of Ns but also over the values of σ.



Frequentist case

Here there is a further complication.

In the case without nuisance parameters, limit setting requires repeating the measurement over ensembles of toys generated at different true values of m.

We just learned that the “repeating the measurement” part is straightforward: just replace likelihood with a profile likelihood.

But what about the toys? Which true values of σ should one use in their generation?

This is a very important point where decisions have important implications on results, but that gets unfortunately often overlooked (or voluntarily swept under the carpet).

People typically focus lots of attention on how to treat nuisance parameters in *fitting* (that is when repeating the measurement). But how to treat them in *generation* has usually significant impact on the variance of the final results.

How to treat nuisance parameters in generation

"plugin method" — only generate toys using the very σ value $\hat{\sigma}$ estimated on data. Equivalent to assume that the true values of the nuisance parameters are *exactly* those measured in data. This is by far the most used (..requires less work…)

"supremum p-value method" — generate multiple ensembles of toys, each at a different true value of σ chosen over a grid that scans the full allowed range. For each tested signal strength, use the true value σ yielding the worse p-value (the one yielding the weakest limit). I've only seen this used by the CKMFitter group in the early 2000s (most work)

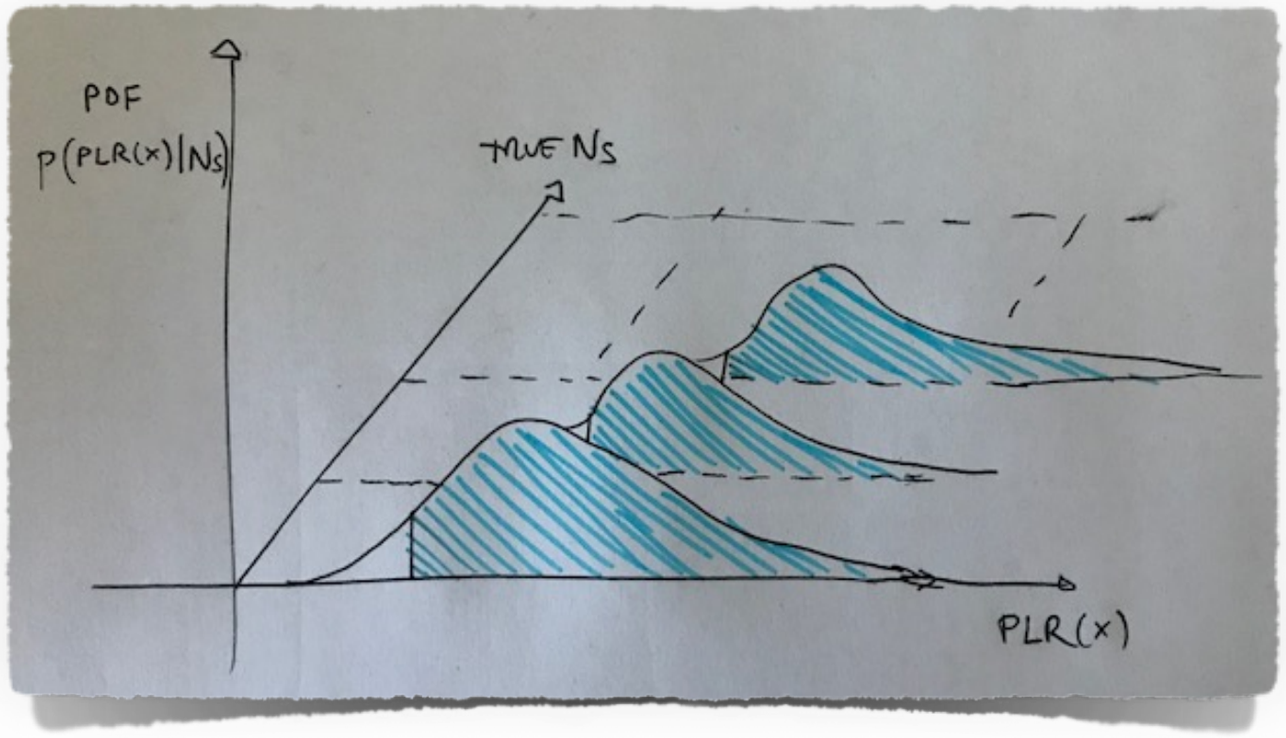
In medio stat virtus?

"Berger and Boos:" generate multiple ensambles of toys at σ values sampled in a plausible range centered on their estimates $\hat{\sigma}$ in data. JASA, 89, 427 (1994)

https://arxiv.org/pdf/0810.3229.pdf Phys. Rev. Lett 100 161802.

Plug-in at work

Produce an ensemble of S+B toys assuming in generation a true value N_s^j for the signal yield and a true value $\hat{\sigma}$ for the resolution — $\hat{\sigma}$ has been previously estimated in simulation/control samples.



For each set of toys generated at true value N_s^j , construct the profile likelihood L(Ns, $\hat{\sigma}$') by maximizing L(Ns, σ) with respect to σ and then maximize over Ns to get the distribution of $\hat{N}s$ and construct a curve

Repeat for all sensible choices of true values N_s^k

The resulting limit covers the true value of N_s only if the true value of σ in reality is the value $\hat{\sigma}$ assumed in generation. Undercoverage is possible otherwise.

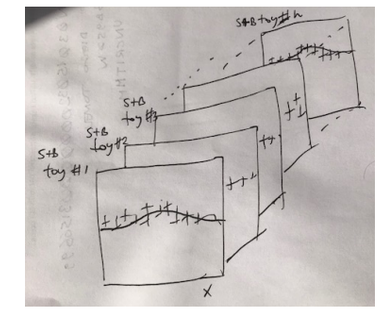
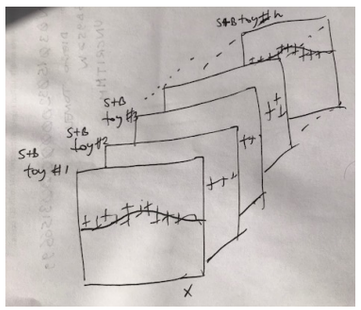
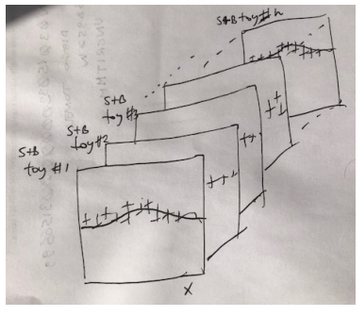
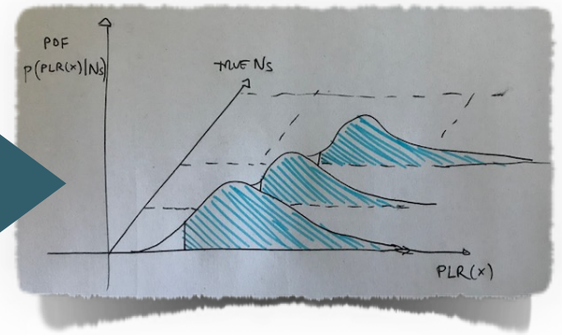
Supremum at work

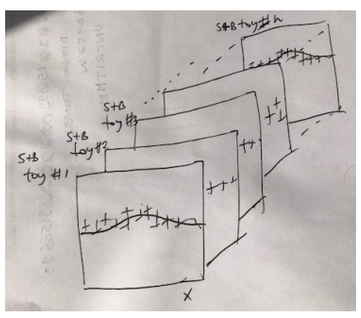
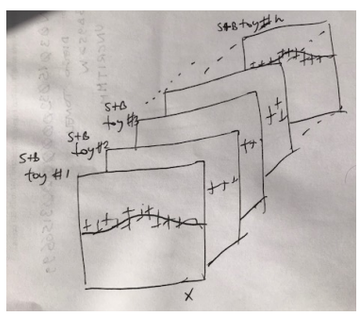
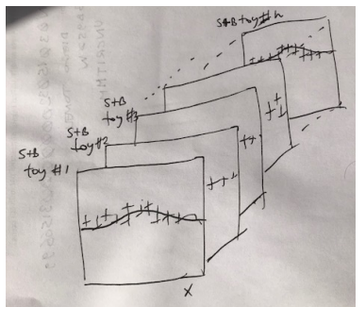
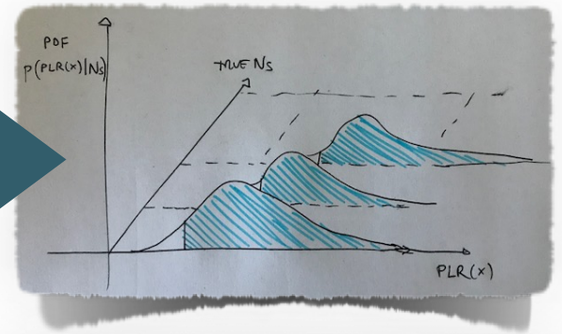
Generate with $N_s = N_s^1$

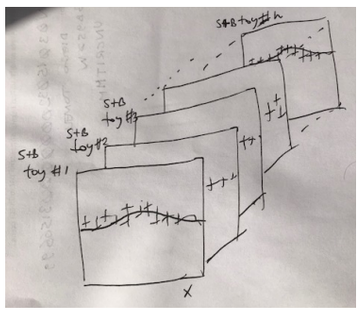
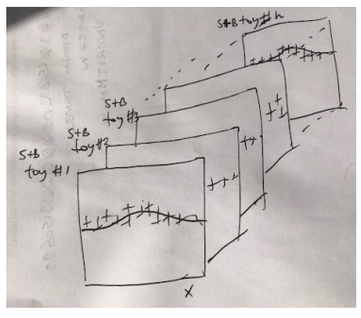
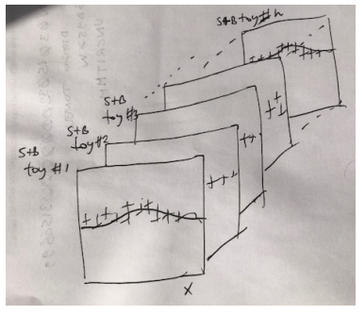
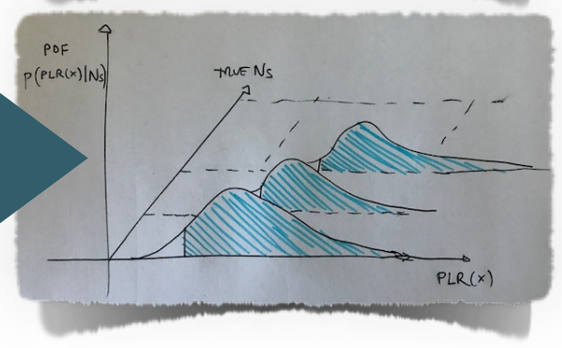
$N_s = N_s^2$

$N_s = N_s^n$

Generate with $\sigma = \sigma_a$ $\sigma = \sigma_a + \varepsilon$

For each tested value N_s, construct the confidence band by using the curve corresponding to the true σ value that makes our limit weakest. The resulting limit covers the true value regardless of the unknown value of the resolution

An hybrid approach — Cousins-Highlands

Marginalize the likelihood until it is free from any nuisance parameter dependence and then proceed switching to with a frequentist procedure

Very much used in the late 90's,

Nuclear Instruments and Methods in Physics Research A320 (1992) 331–335
North-Holland

NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A

Incorporating systematic uncertainties into an upper limit

Robert D. Cousins
Physics Department, University of California, Los Angeles, CA 90024, USA

Virgil L. Highland
Physics Department, Temple University, Philadelphia, PA 19122, USA

Received 27 March 1991 and in revised form 19 February 1992

We discuss the problem of incorporating the uncertainty in the experimental sensitivity into the calculation of an upper confidence limit on a branching ratio or similar quantity. If the number of events is small or zero but without background, the correction to the usual result is given by a simple, easily applied formula. The case of an accurately known background also has a simple solution.

The main limitation of this hybrid method is that it does not have fully Bayesian properties nor it has fully frequentists properties, which makes it hard to qualify its performance and determine a straightforward interpretation

What can go wrong

(in addition to technical bugs, snafus, etc.)

Coverage

The single most serious failure of any limit setting procedure is to screw the coverage (exclude something that exists). Remember: coverage is a frequentist concept but it is generally seen as desirable in HEP also by Bayesian experimenters too.

Experimental Limits on the Decays $K_L^0 \to \mu^+\mu^-$, e^+e^-, and $\mu^\pm e^\mp$†

Alan R. Clark, T. Elioff,* R. C. Field, H. J. Frisch, Rolland P. Johnson, Leroy T. Kerth, and W. A. Wenzel
Lawrence Radiation Laboratory, University of California, Berkeley, California
(Received 9 April 1971)

We have performed a search at the Bevatron for the decays $K_L^0 \to \mu^+\mu^-$, e^+e^-, and $\mu^\pm e^\mp$ with a double magnetic spectrometer using wire spark chambers. Over 10^6 observed $K_L^0 \to \pi^+\pi^-$ decays determine the normalization for the di-lepton decay modes. No e^+e^- or $\mu^\pm e^\mp$ events were observed. For each of these decays the upper limit on the branching ratio relative to all modes is 1.57×10^{-9} (90% confidence level). For the decay $K_L^0 \to \mu^+\mu^-$, the limit is 1.82×10^{-9} (90% confidence level).

K_L^0 $I(J^P) = 1/2(0^-)$

See related reviews:

V_{ud}, V_{us} the Cabibbo Angle, and CKM Unitarity
CP Violation in K_L^0 Decays
$\Delta S = \Delta Q$ in K^0 Decays

▼ **Charge conjugation × Parity (CP) or Lepton Family number (LF) violating modes, or $\Delta S = 1$ weak neutral current ($S1$) modes**

Γ_{23}	$\mu^+\mu^-$	$(6.84 \pm 0.11) \times 10^{-9}$	225

Coverage — Bayesian case

In the Bayesian approach, coverage is more fragile as it's not built-in by construction in the procedures.

The conceptual `average' associated with the marginalization over various nuisance-parameter scenarios, with an arbitrary metric, exposes to potentially large violations of coverage

The measurement lives in one and only nuisance-parameter scenario (though unknown): averaging over scenarios might compensate/cancel the effects associated with the actual realized scenario, leading to unrealistic optimistic results.

Health checks

- Prior sensitivity: repeat the measurement with different choices of prior densities and study the dependence of the results. Large sensitivity to prior choice suggest that results are driven by the subjective assumptions than data
- Coverage: generate ensembles of toys for various values of the physics and nuisance parameters, repeat the procedure on them and check that the resulting limits exclude the true values with the desired Bayesian credibility

Coverage - frequentist case

In the frequentist approach, coverage is built-in in the procedure.

A limit based on likelihood-ratio ordering (aka Feldman-Cousins) where the confidence band is constructed in the full dimensionality of all physics and nuisance parameters and then projected into the physics parameters has rigorous coverage.

However, the necessary simplifications associated with realistic numbers of dimensions may and do jeopardize the coverage properties.

Using profile-likelihoods, use only a subset (if not one value only as in plugin) of possible true values of nuisance parameters in toy generation, etc, are all subject to cause undercoverage.

- Coverage: generate ensembles of toys for various values of the physics and nuisance parameters, repeat the procedure on them and check that the resulting limits exclude the true values with the desired CL.

What you shouldn’t be doing

(but probably have done, are doing, and may do…)

Ingenious systematic embedding

It is still common to encounter “creative” ways of accounting for systematic effects in limits. Most are based on somehow embedding systematics with the statistical portion of the inference and then use result in standard limit-setting procedure.

Typical example: likelihood of the data is convolved with a “likelihood” for the nuisance parameter (typically assumed Gaussians) to construct a pseudolikelihood that is then used in a standard limit procedure (Bayesian or frequentist).

Stuff like that has no statistical support — you are on your own.

- the “likelihood” assumed for the nuisance parameter is arbitrary. If you were to know a model for your nuisance parameters, then they would no longer be nuisance parameters and would naturally be included in your p(x|m,s) model.
- the resulting pseudolikelihood has unknown properties

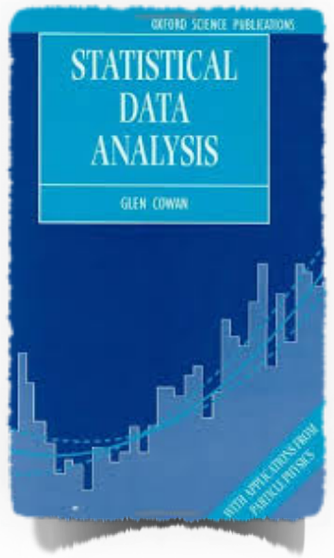
In many case results are not strongly wrong, despite the conceptual anarchy, especially when systematics are a small perturbation of the statistical fluctuations. But these procedures should be deprecated.

Summary

- Limits are not exotic animals, they are *measurements* just as your standard point estimates or two-sided confidence regions
- Treatment of systematic uncertainties in limits is subjected to the same assumptions, limitations, implications of other forms of measurements. Stay away from black boxes, ready recipes etc. make an effort to understand the concepts: it's just as fun and creative as other parts of your analysis.
- The HEP community expects coverage from your result.
- Systematic uncertainties are a parametrization of imperfect knowledge of the model
- Bayesians treat them just as any physics parameters: "average" over the arbitrary metric determined by priors.
- Frequentist replace the likelihood with a lower-dimensional function that has nice properties but is not a likelihood. In addition, they should ensure to treat correctly the nuisance parameter unknowns in generation too.
- Whatever is your inference philosophy (Bayesian or frequentist or else): (i) use a statistical procedure consistently (eg, don't mix Bayesian and frequentist) (ii) document it in detail including assumptions/simplifications etc (iii) make sure results have coverage or say it clearly if they don't.

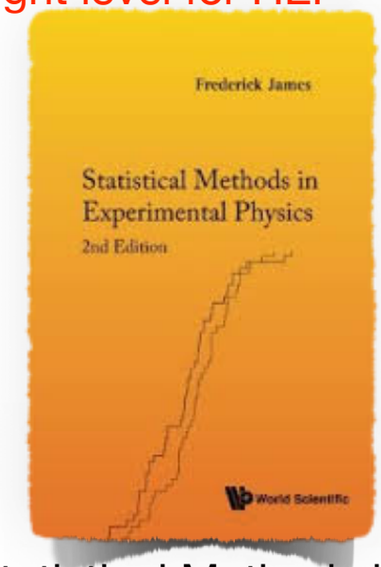
Basics of confidence-band construction

- Good starting point



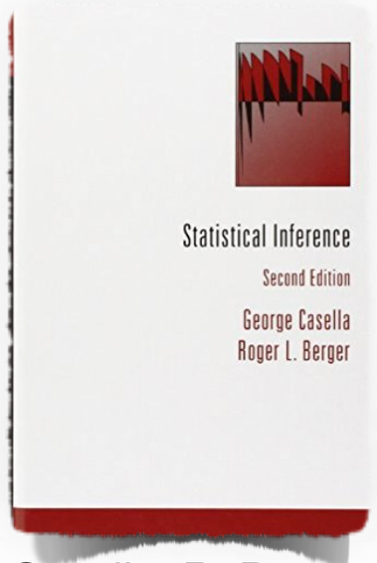
G. Cowan,“Statistical data analysis”

- Very good book at the right level for HEP



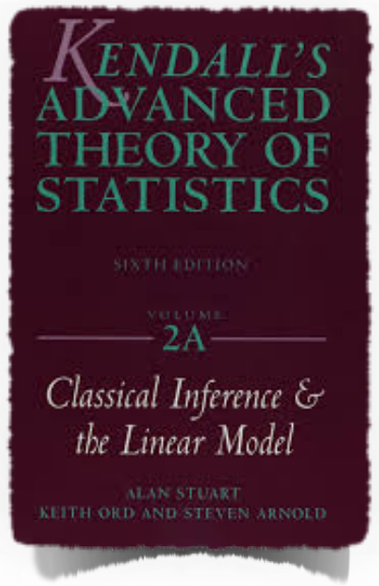
F. James, “Statistical Methods in Experimental Physics, data analysis”

- Advanced book



G. Casella, R. Berger, “Statistical Inference

- Ultimate bible



A. Stuart, et al “Kendall’s Advanced Theory of Statistics Vol 2A”

For the bravest of all..(sec 4)

CDF/MEMO/STATISTICS/PUBLIC/8662
Version 4.00
June 13, 2007

P Values: What They Are and How to Use Them

Luc Demortier[1]
Laboratory of Experimental High-Energy Physics
The Rockefeller University

"Far too many scientists have only a shaky grasp of the statistical techniques they are using. They employ them as an amateur chef employs a cook book, believing the recipes will work without understanding why. A more cordon bleu *attitude to the maths involved might lead to fewer statistical soufflés failing to rise."*
in "Sloppy stats shame science," The Economist, Vol. 371, No. 8378, pg. 74 (June 5th 2004).

Abstract

This note reviews the definition, calculation, and interpretation of p values with an eye on problems typically encountered in high energy physics. Special emphasis is placed on the treatment of systematic uncertainties, for which several methods, both frequentist and Bayesian, are described and evaluated. After a brief look at some topics in the area of multiple testing, we examine significance calculations in spectrum fits, focusing on a situation whose subtlety is often not recognized, namely when one or more signal parameters are undefined under the background-only hypothesis. Finally, we discuss a common search procedure in high energy physics, where the effect of testing on subsequent inference is incorrectly ignored.

https://www-cdf.fnal.gov/~luc/statistics/cdf8662.pdf