Thoughts on Stats
after two PhyStat-nu Workshops

João Coelho

1st October 2016

Personal Motivation

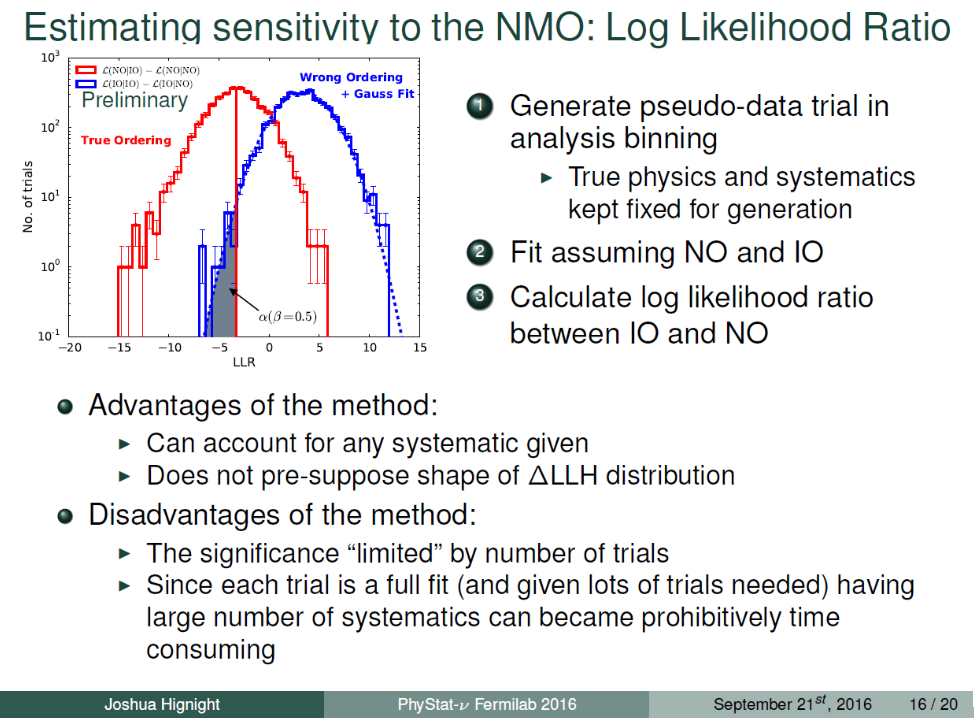


Table 14. Default parameter settings used for the LLR analysis. Where μ and σ are given, they refer to a Gaussian distribution.

Parameter	True value distr.	Initial value distr.	Treatment	Prior
θ_{23} (°)	{40, 42, ..., 50}	uniform over [35, 55] †	Fitted	No
θ_{13} (°)	8.42	$\mu = 8.42$, $\sigma = 0.26$	Fitted	Yes
θ_{12} (°)	34	$\mu = 34$, $\sigma = 1$	Nuisance	N/A
ΔM^2 (10^{-3} eV^2)	$\mu = 2.4$, $\sigma = 0.05$	$\mu = 2.4$, $\sigma = 0.05$	Fitted	No
Δm^2 (10^{-5} eV^2)	7.6	$\mu = 7.6$, $\sigma = 0.2$	Nuisance	N/A
δ_{CP} (°)	0	Uniform over [0, 360]	Fitted	No
Overall flux factor	1	$\mu = 1$, $\sigma = 0.1$	Fitted	Yes
NC scaling	1	$\mu = 1$, $\sigma = 0.05$	Fitted	Yes
$\nu/\bar{\nu}$ skew	0	$\mu = 0$, $\sigma = 0.03$	Fitted	Yes
μ/e skew	0	$\mu = 0$, $\sigma = 0.05$	Fitted	Yes
Energy slope	0	$\mu = 0$, $\sigma = 0.05$	Fitted	Yes

Note. The † indicates that the initial values for θ_{23} are generated in a special way: a total of seven initial values is tried. They are $x + i \times 5°$, where x is the randomly drawn value and $i \in [-3, -2, \ldots, 3]$.

- Both ORCA and PINGU use **fixed true values** for most parameters (except ΔM^2 and solar pars. for ORCA)
- This is ok, but in principle depends on what values are chosen from true parameter space

Backhouse (NOvA)

Coverage

- Frequentist coverage means: "if the true value of parameter x is A, 68% of experiments will include A in their confidence interval for x"
- FC procedure achieves this almost tautologously by throwing mock experiments at each A and finding the $\Delta\chi^2_{\text{crit}}$ that would have included that A in 68% of the experiments

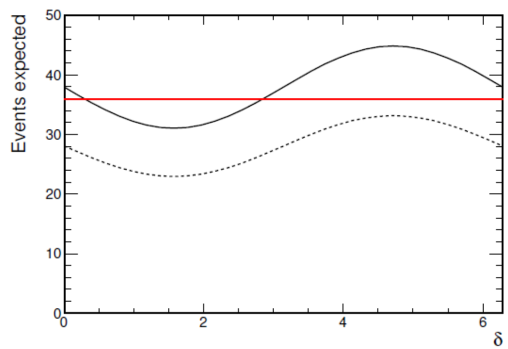
- In the presence of a parameter y not displayed on the plot (a "nuisance parameter")
- Want correct coverage *no matter the true value of that parameter*

- Obviously impossible in general, infinite array of possible values for y, all requiring different critical values in principle
- But *e.g.* for two gaussian variables profiling over y gives correct coverage, even without invoking FC corrections
- So how does it work out in practice for our experiment?

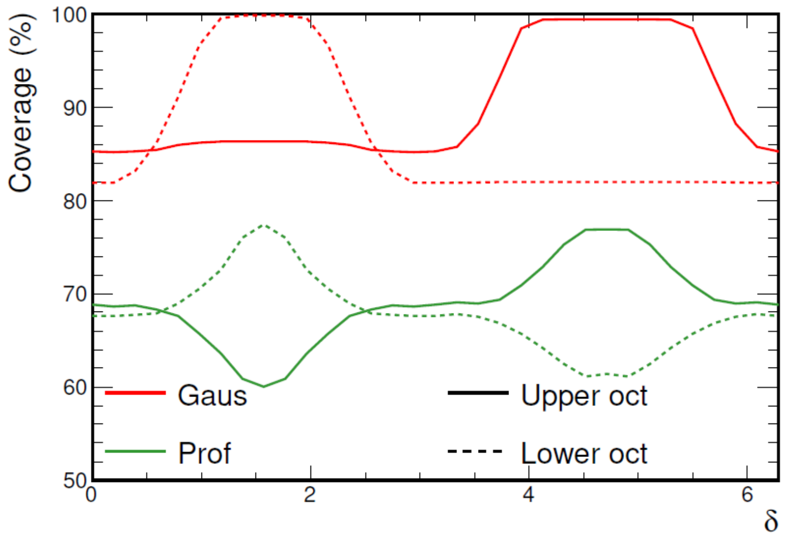
Backhouse (NOvA)

- **Could not find satisfactory** way to achieve proper **coverage** using a toy model inspired by nue appearance

Upgraded toy



- Number expected $= (0.8 \text{ or } 1.2)(33 - 6\sin\delta)$

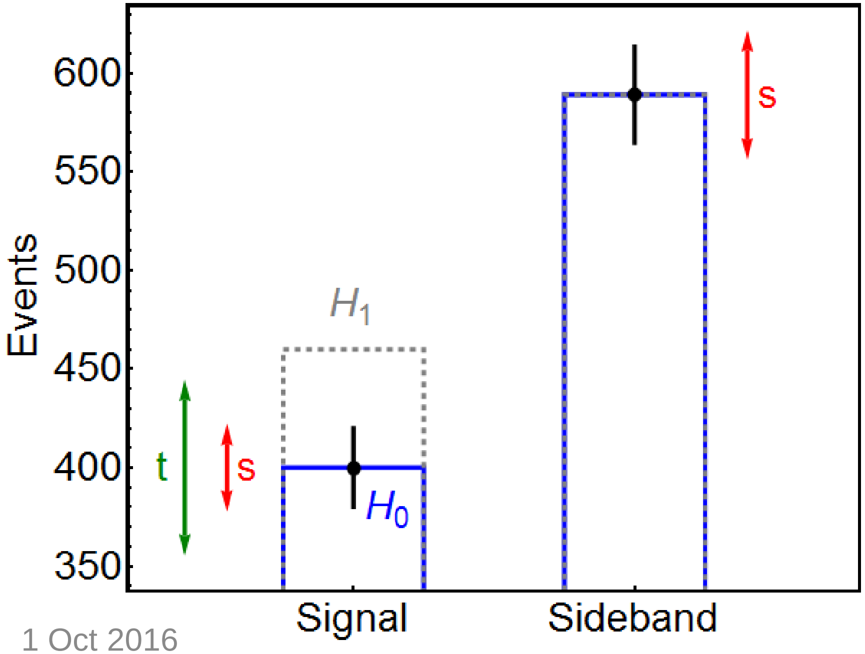


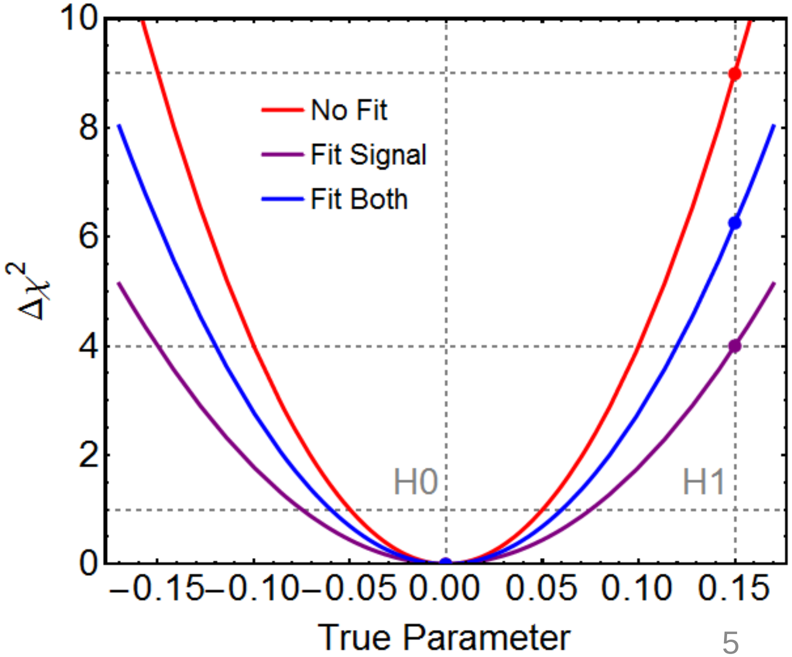
Pragmatism

- No satisfactory way to "integrate out" hierarchy or octant possible
- Continue to plot four curves

My Toy Model

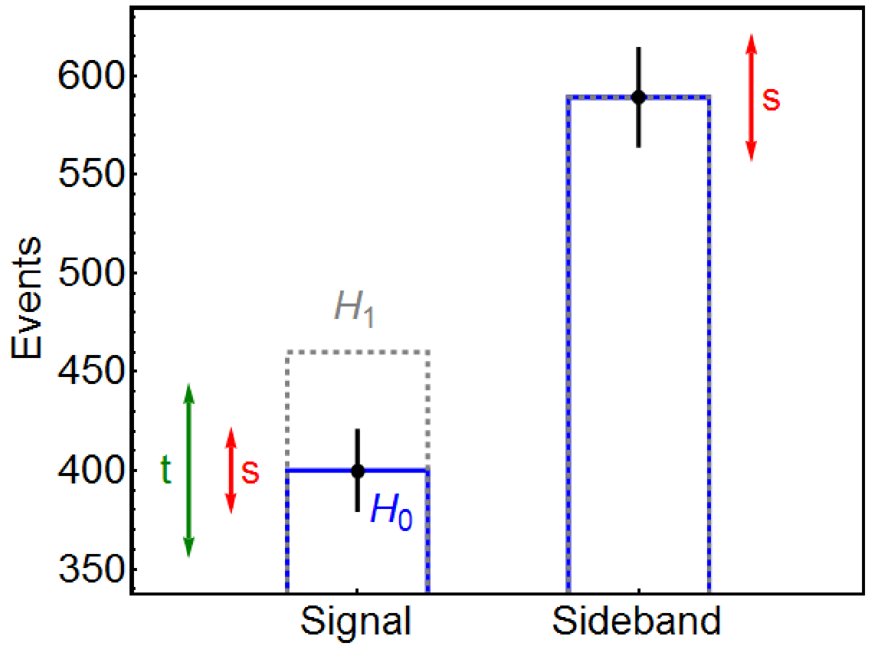
- Two bins:
  - Signal bin affected by par. of interest (t) and nuisance par. (s)
  - Sideband bin only affected by nuisance parameter (s)
- Three fitting approaches:
  - **No Fit**: Only look at signal bin and don't fit nuisance parameter
  - **Fit Signal**: Only look at signal bin and fit nuisance parameter
  - **Fit Both**: Look at both signal an sideband and fit nuisance par.

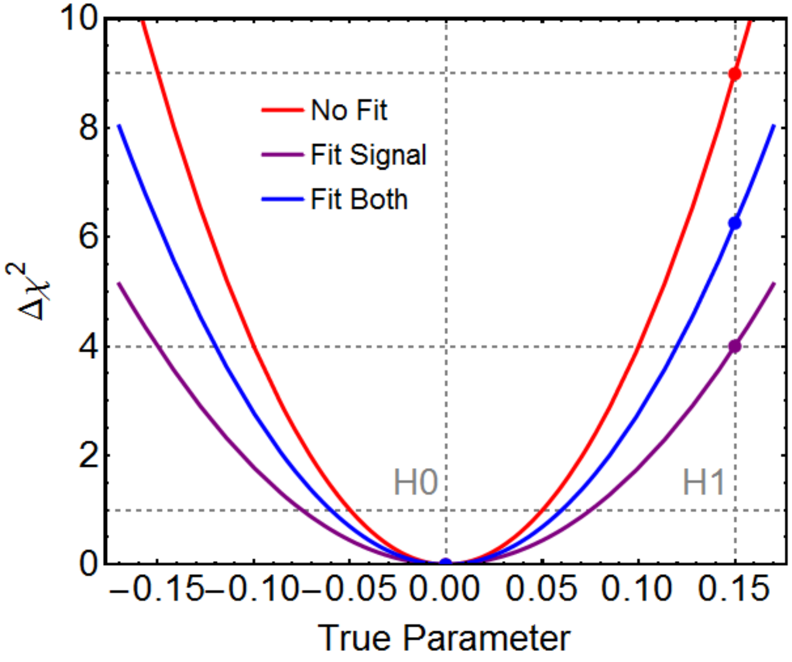




Hypothesis Testing

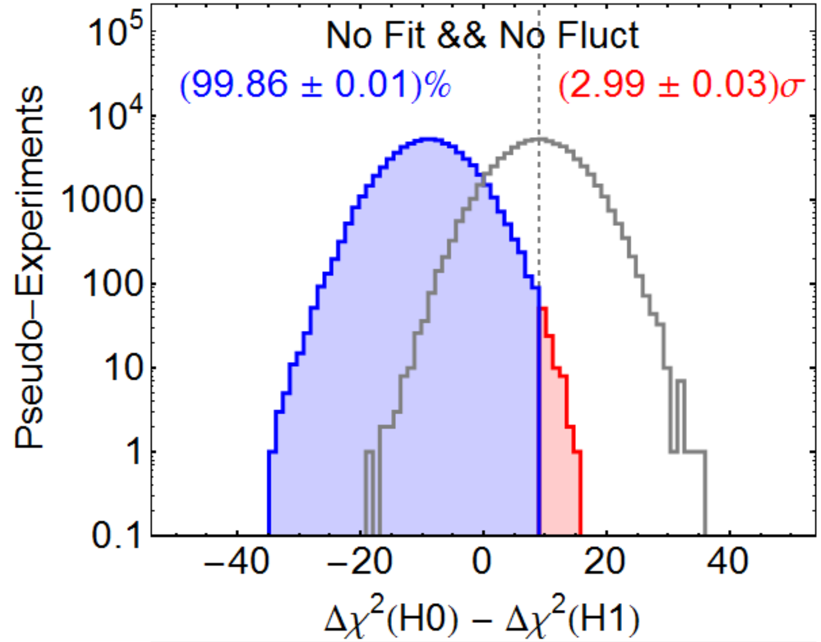
- Defined two hypothesis: H_0 ($t = 0$) and H_1 ($t = 15\%$)
- What should we expect?
  - **Stat. Significance: 3σ** (uncertainty is 5%)
  - **Stat. + Nuis. Significance: 2σ** (uncertainty is 7.5%)
  - **Stat. +Nuis w/ Sideband: 2.5σ** (uncertainty is 6%)

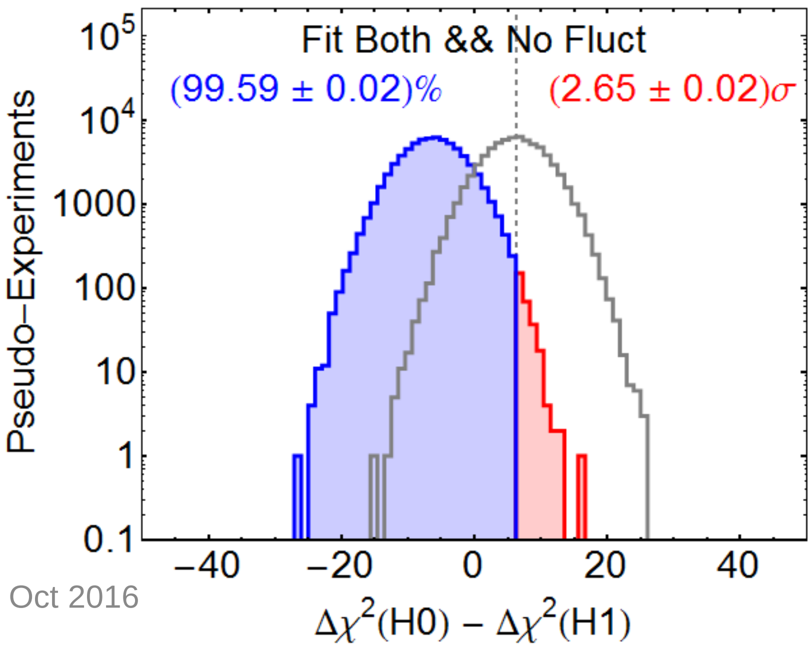


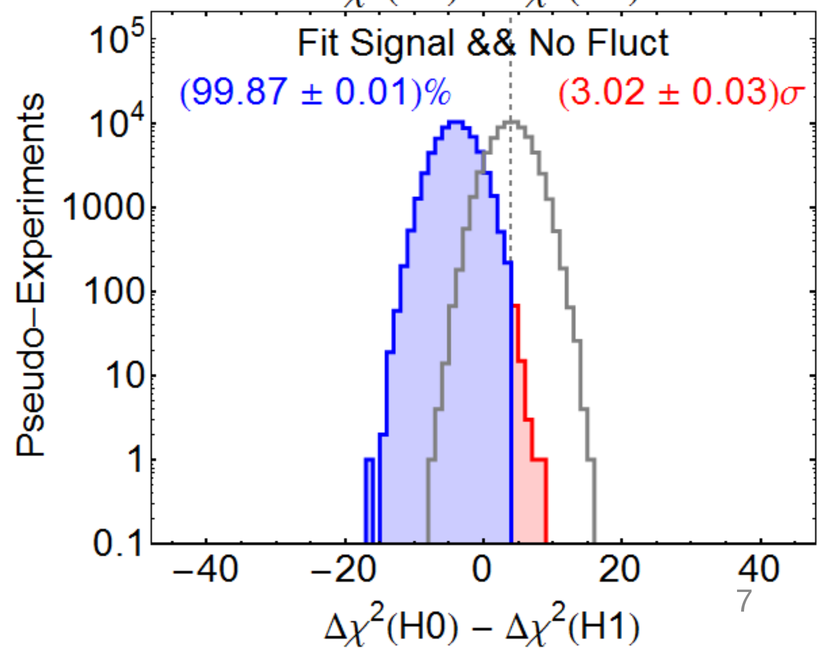


Test Statistic

- **Don't fluctuate** nuisance:
  - **3σ for signal bin only**
  - Independent of fitting nuisance
  - **Worse significance w/ sideband**
  - Not expected 2.5σ significance

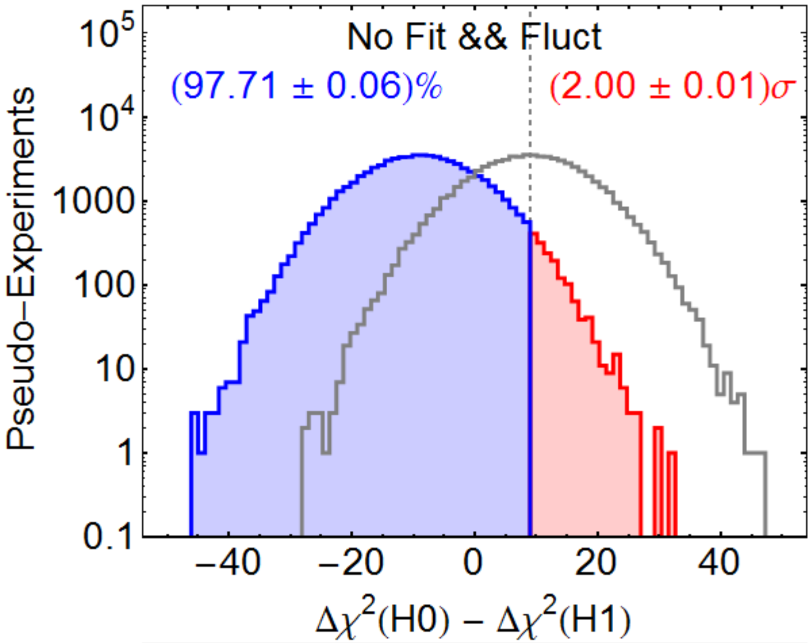


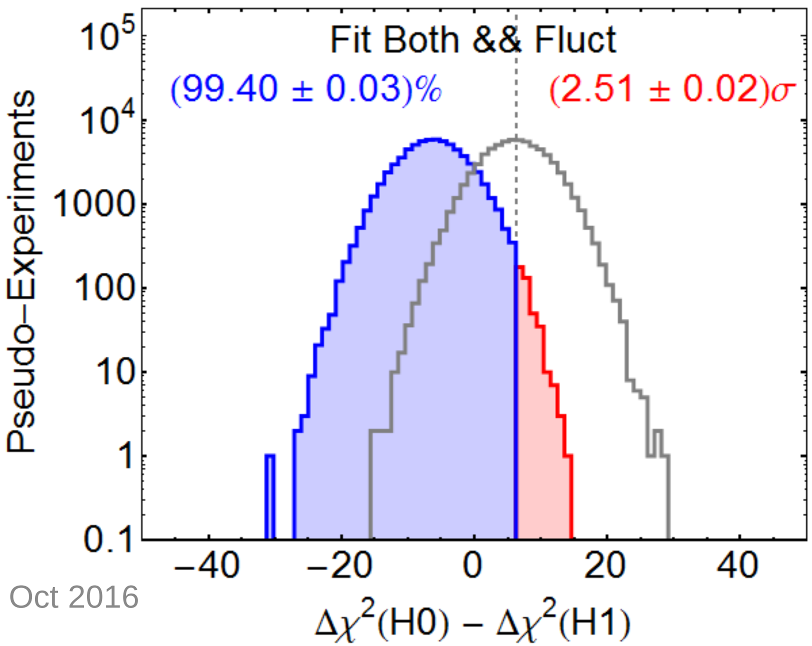


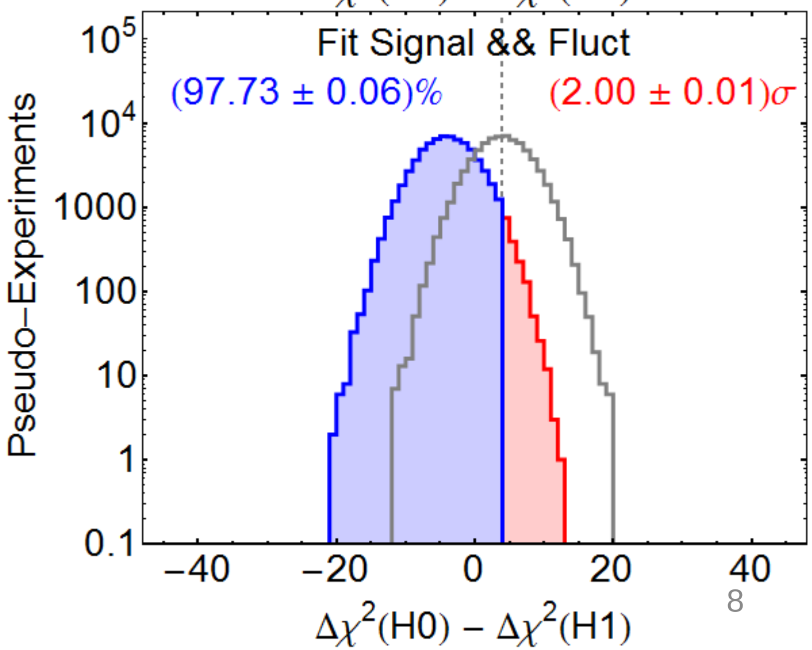


Test Statistic

- **Fluctuate** nuisance:
  - **2σ for signal bin only**
  - Independent of fitting nuisance
  - **Better significance w/ sideband**
  - Expected 2.5σ significance
  - Same median values

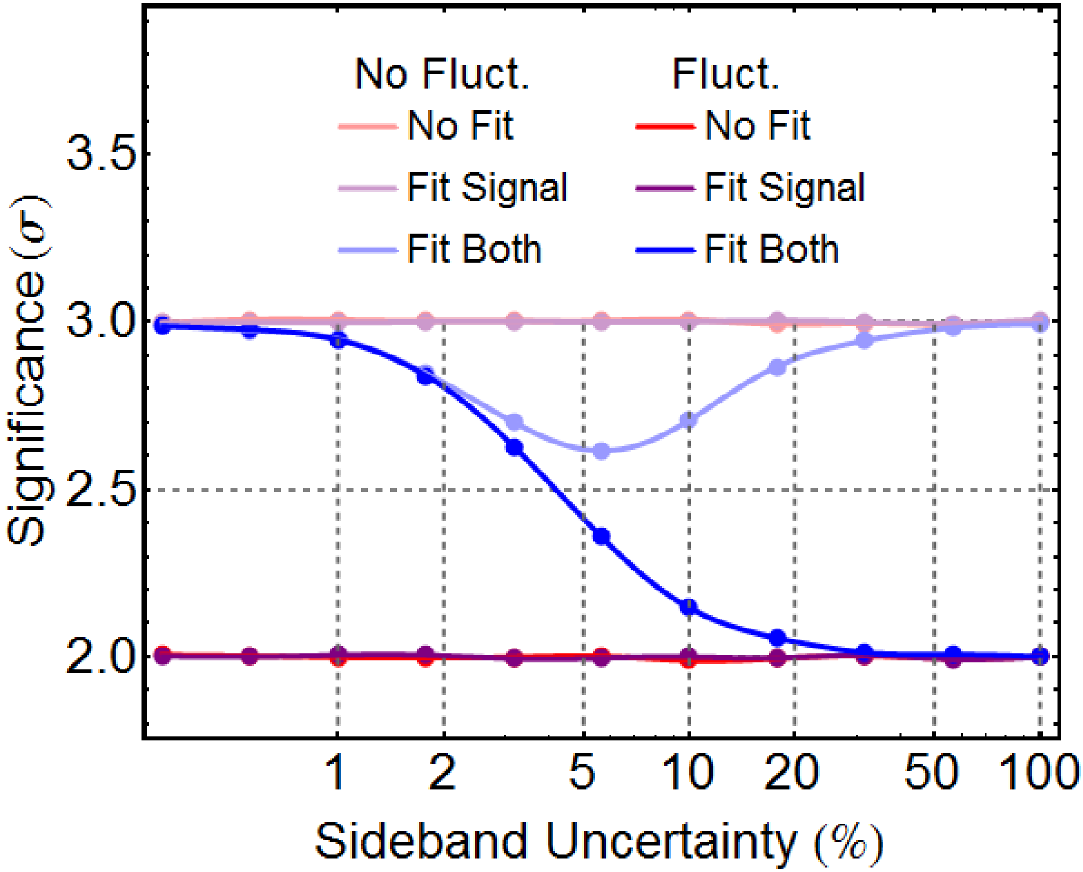






When does it matter?

- Sideband is relevant if it reduces nuisance uncertainty
- If sideband fits nuisance very well, not fluctuating may be ok
- However, no reason to not fluctuate since TS distribution should be identical for significance to match



Except…

Bob Cousins

Conditioning (cont.)

- The 1958 thought expt of David R. Cox focused the issue:
  - Your procedure for weighing an object consists of flipping a coin to decide whether to use a weighing machine with a 10% error or one with a 1% error; and then measuring the weight. (Coin flip result is ancillary stat.)
  - Then "surely" the error you quote for your measurement should reflect which weighing machine you actually used, and not the average error of the "whole space" of all measurements!
  - But classical most powerful Neyman-Pearson hypothesis test uses the whole space!
- In more complicated situations, ancillary statistics do not exist, and it is not at all clear how to restrict the "whole space" to the relevant part for frequentist coverage.
- In methods obeying the likelihood principle, in effect one conditions on the exact data obtained, giving up the frequentist coverage criterion for the guarantee of relevance.

Summary

- Lots of very interesting material from both the Tokyo and Fermilab organised PhysStat-nu workshops
  - https://indico.fnal.gov/conferenceDisplay.py?confId=11906
  - http://indico.ipmu.jp/indico/internalPage.py?pageId=1&confId=82 (Broken?)
- Tokyo workshop has a live summary document:
  - http://www.hep.ph.ic.ac.uk/~yoshiu/PhyStat-nu-IPMU-2016-Summary-Draft/

- General consensus:
  - p-value (sigma) is not good enough to inform us
  - Experiments should report both Frequentist and Bayesian results
  - When using Bayesian method, must explore sensitivity to priors

Summary

- MH, CPV, θ_{23} oct., all introduce violations of Wilk's theorem.

- No clear answer on best practices for treating as nuis. pars.

- My toy model says **we should sample random true values** in order to obtain correct sensitivities

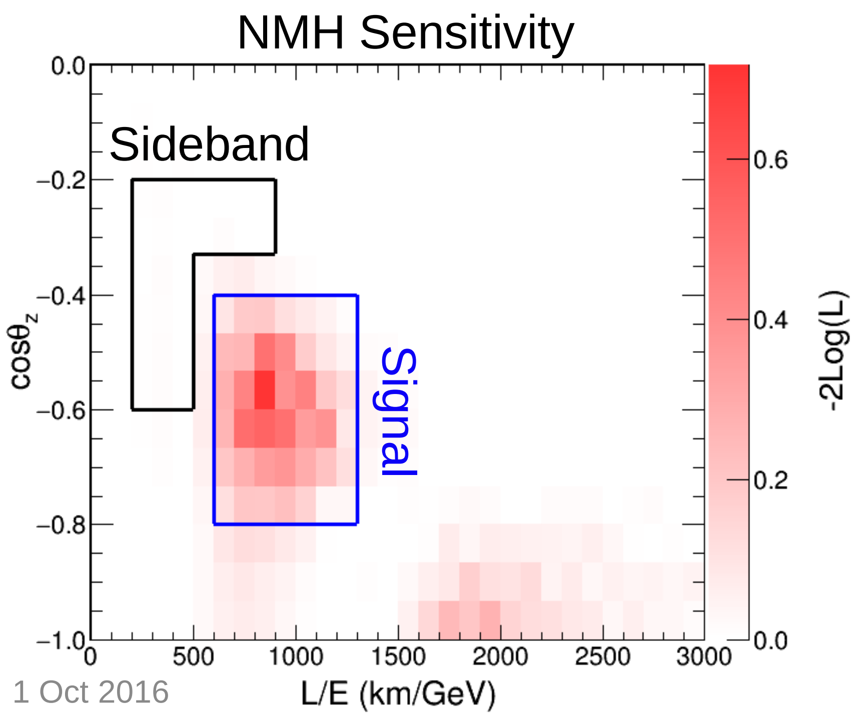
- Also some discussion on conditioning frequentist method

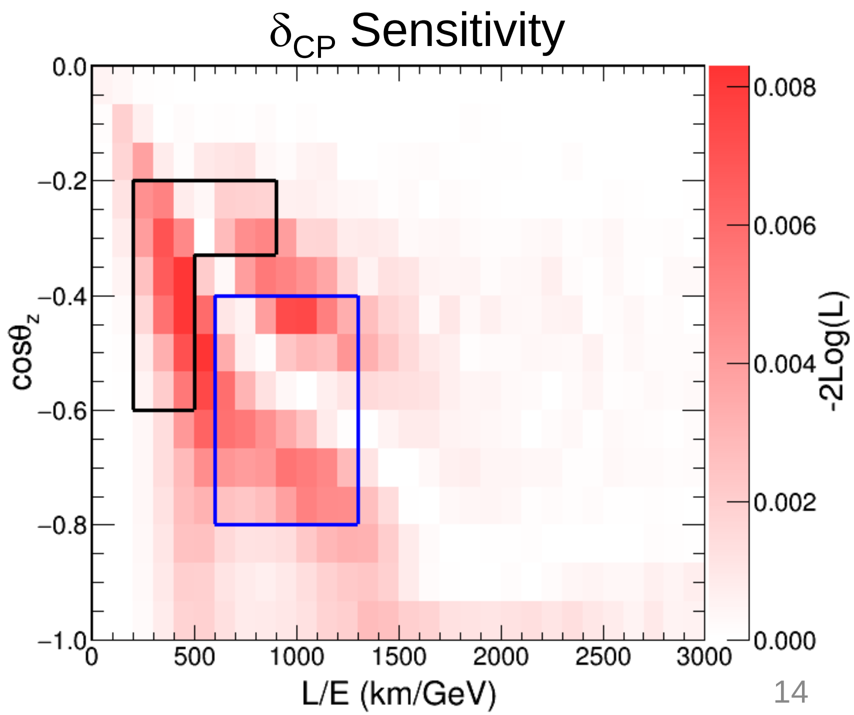
- **No guaranteed coverage**, but not all statisticians care (Bayesians)

Backup Slides

Toy Model Concept

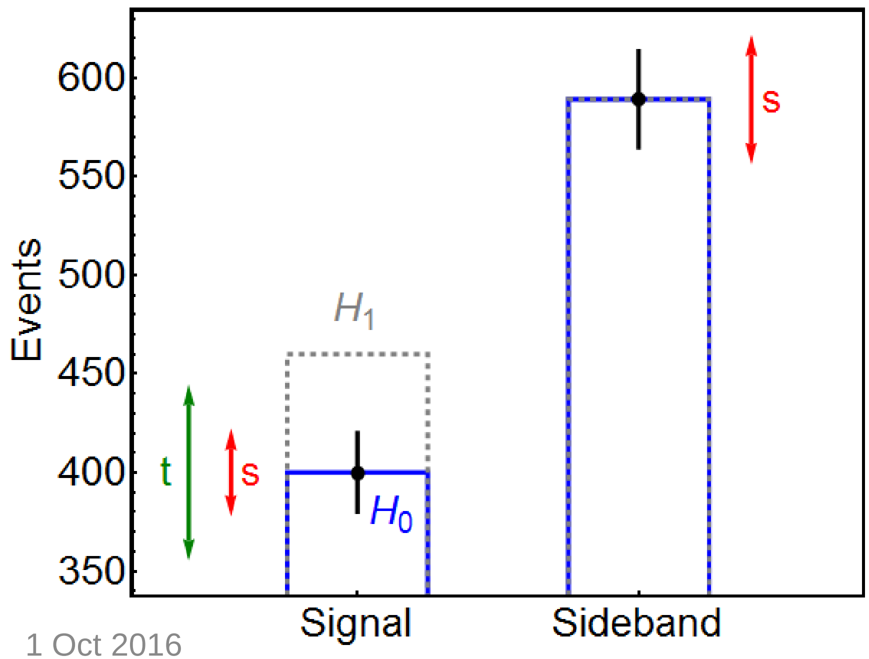
- MH sensitivity is limited to small regions of θ_z and E
- Other parameters, e.g. δ_{CP}, affect different regions
- Two bins:
  - **Signal bin** affected by par. of interest (MH) and nuisance par. (δ_{CP})
  - **Sideband bin** only affected by nuisance parameter (δ_{CP})
  - Sideband can be used to reduce impact of δ_{CP} (in principle)

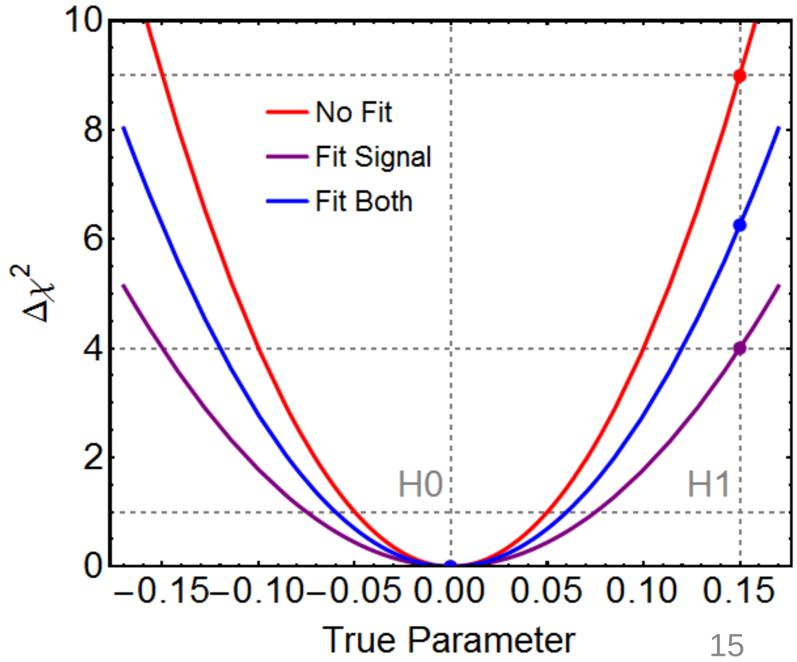




Some Nice Numbers

- Choose some nice properties:
  - Set signal bin to 400 events (**5% Stat.** Uncertainty)
  - Set nuisance uncertainty to 5.6% (**7.5% Stat. + Nuis.** Uncertainty)
  - Sideband size controls precision to measure nuisance par.
  - This example has sideband uncertainty at 4.1%
  - **Reduces to 6% Stat. + Nuis.** Uncertainty



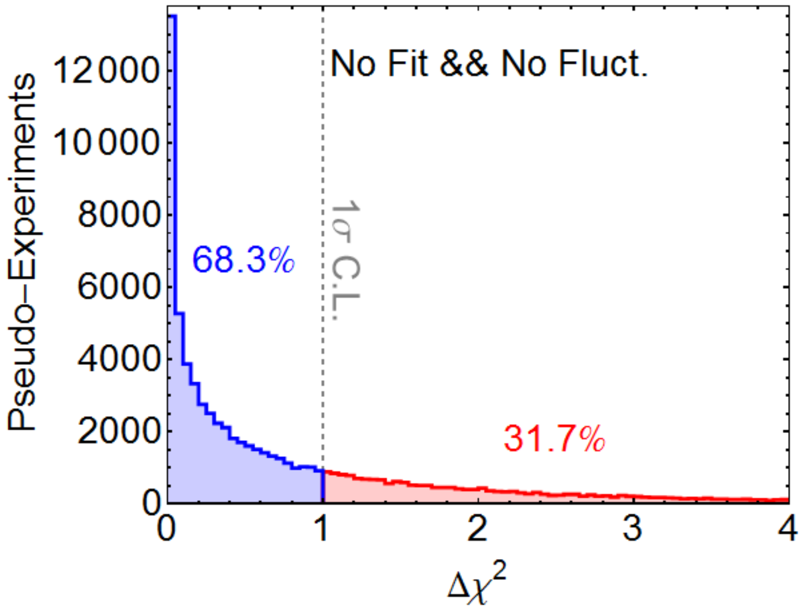


Feldman-Cousins

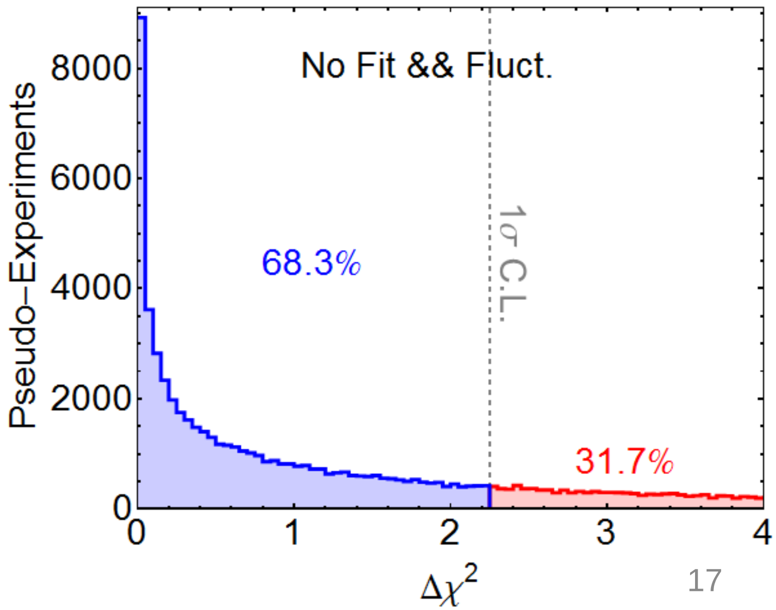
- Need to interpret value of $\Delta\chi^2$ at each value of t
  1. Define a procedure to use in data
     - Count number of events
     - Choose a fitting method, e.g. fit nuisance in signal and sideband bins
     - Compute $\Delta\chi^2$ at a particular point in par. of interest space
  2. Simulate N experiments of possible results you might get
     - **Gaussian or Poisson statistics**
     - **Different experimental setups (some systematics)**
     - **Different possible worlds (vary physics parameters)**
  3. Count experiments that correspond to certain results
     - **Use same procedure as defined for data**
     - How many experiments have $\Delta\chi^2 < Y$?
     - What value of Y contains 90% of the experiments? (or 68%, 95% ...)
  4. Interpret likelihood of getting the observed data

Fluctuate or Not?

- E.g. : No nuisance fit
- Fix nuisance value:
  - **No impact from nuisance**
  - Expect statistics only result
  - 1σ C.L. at familiar $\Delta\chi^2 = 1$

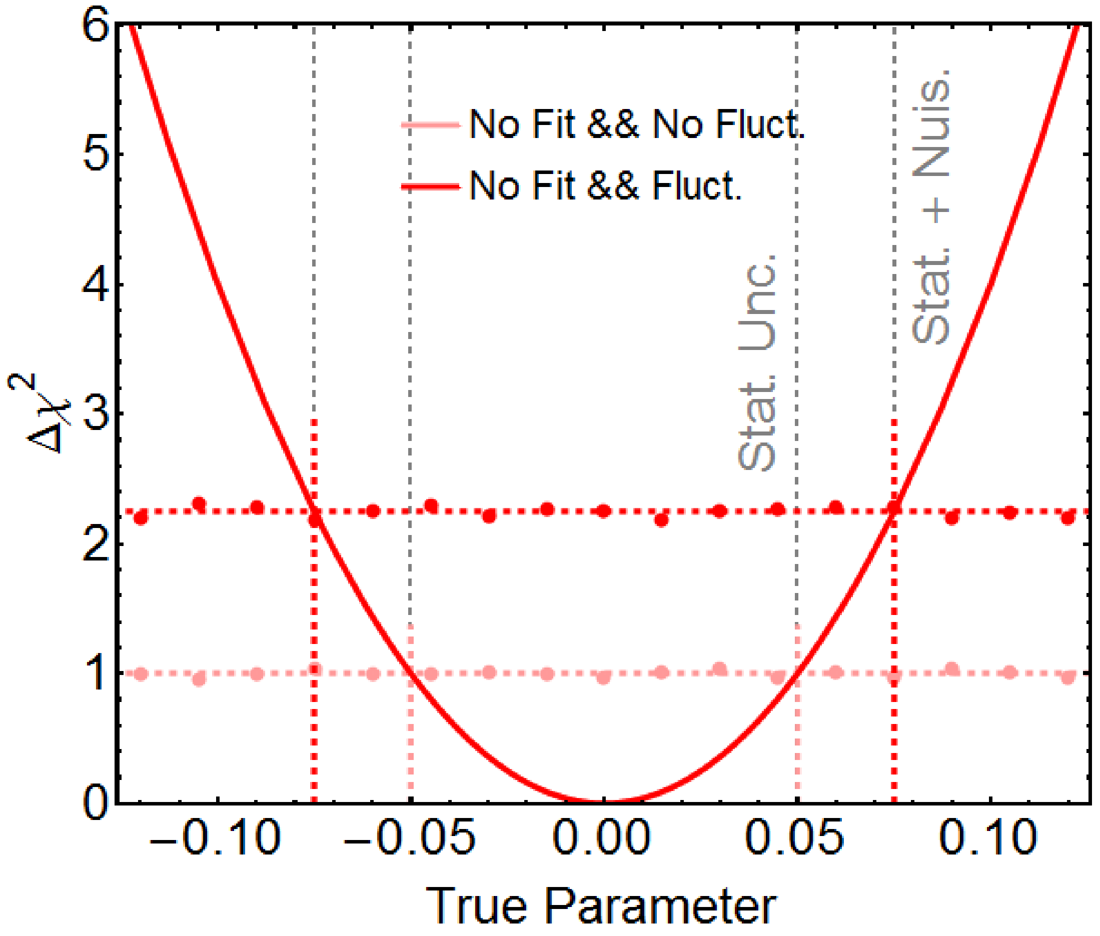


- Fluctuate nuisance value:
  - Assume gaussian prior
  - Effect is to increase typical value of $\Delta\chi^2$
  - **1σ C.L. moves to $\Delta\chi^2$ ~ 2.25**



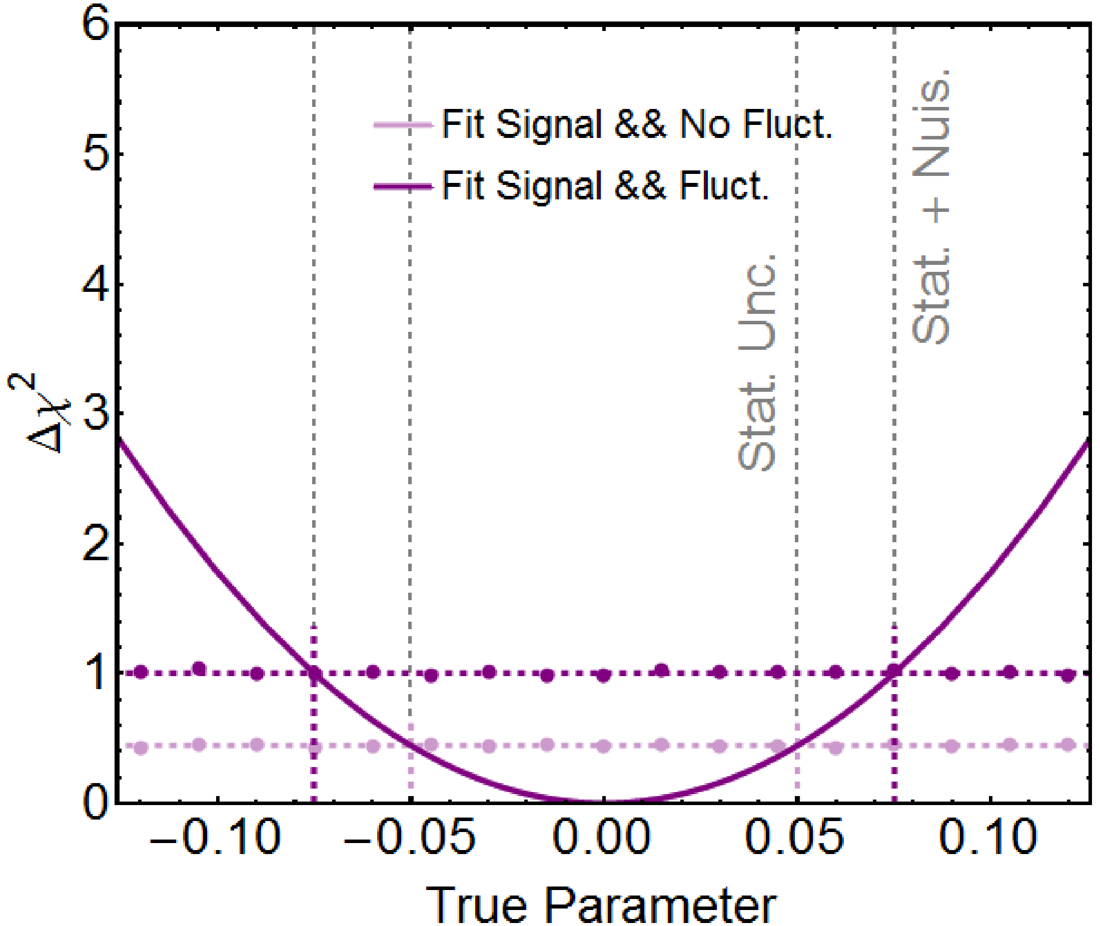
Fluctuate or Not?

- As expected, not fluctuating nuisance leads to parameter of interest being constrained to 5% (Stat. Uncertainty)
- When fluctuating, we get the expected 7.5% (Stat. + Nuis.)



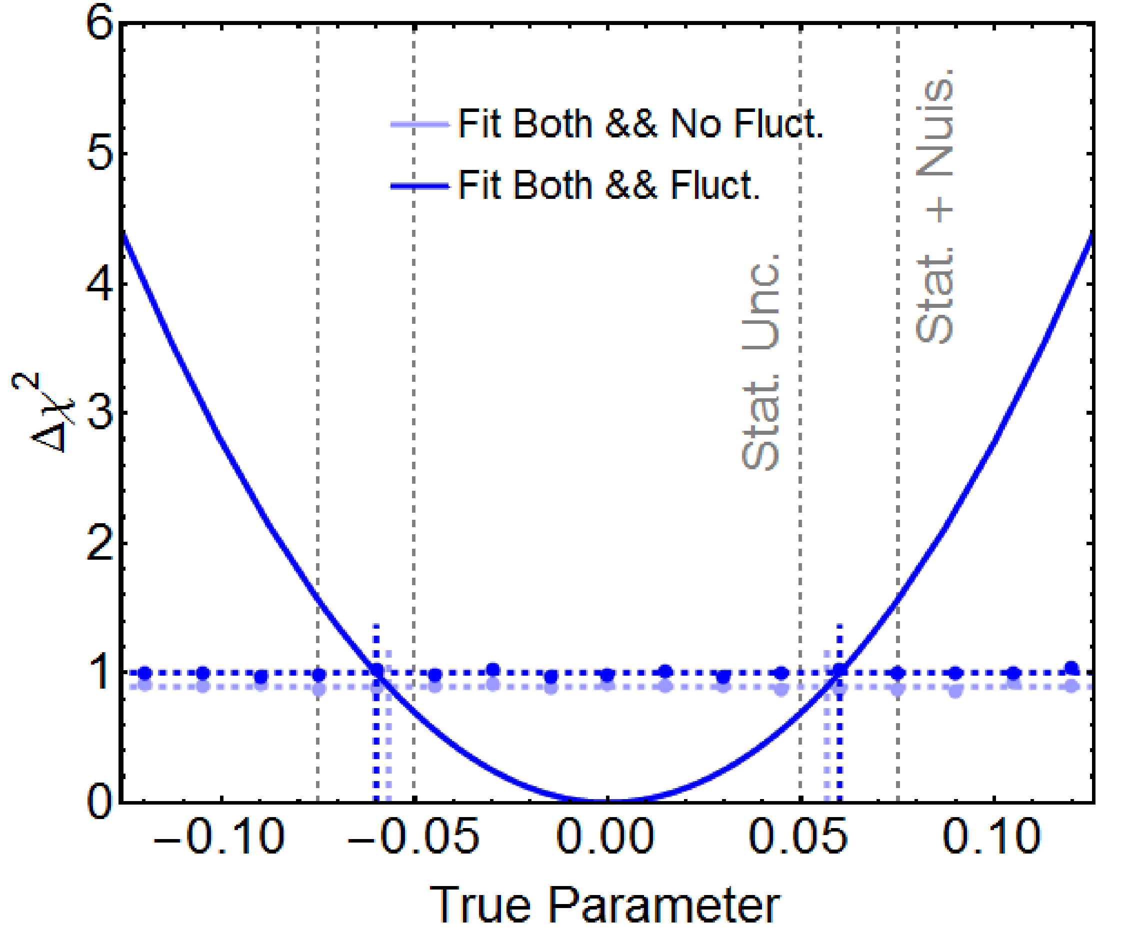
Fluctuate or Not?

- Fitting for the nuisance parameter in signal bin only doesn't change the situation
- Single bin can't distinguish it from the parameter of interest



Fluctuate or Not?

- Fitting in both bins does improve the precision
- Whether to fluctuate has smaller impact, but not zero
- Should depend on how well we measure sideband



David van Dyk

Frequentist or Bayesian?

Do you have to choose??

- Bayes prescribes methodology.
- Frequentists evaluate methods.
- Frequency evaluation of Bayesian methods.
- Model fitting: often little difference in fits and errors.
- Why not control rate of false detection
  and assess probability of new physics?
- Why throw away half of your tool box?

I'm impressed with the openness of neutrino researchers to both Bayesian and Frequency based methods.

- Lots of Bayesian and Frequentist proposals at PhyStat-ν.
- My experience with cosmologists and particle physicists.

David van Dyk

Strategies

What is a physicists to do?

- Controlling false discovery is critical in physical sciences.
- Comparing p-values with a predetermined significant level can control false discovery.... *if used with care, e.g., no cherry picking!*
- When confronted with small p-values researchers *...even statisticians!!...* may believe H_0 is unlikely.
- Bayesian solutions can better quantify likelihood of H_0 / H_1.
- **Solution:** Compute both *global* p-value *and* Bayes Factor.

But be Careful...

1. *quantification of p-values in non-standard problems*
2. *choice and validation of prior distributions*

remain challenging!

Xiao-Li Meng

But what is *Statistical/Probabilistic* Inference?

- An ultimate intellectual game: **"to guess wisely and to guess meaningfully the errors in our guesses."** (*XL-Files*, Oct 2015)
- Impossible to access exact errors, but **a full spectrum** of possibilities for accessing **probabilistic errors**.
- Balancing the **degree of inexactness (Relevance)** & the **reliance on assumptions (Robustness).**

Pure Frequentist (Fully unconditional)

Most Robust but Least Relevant

Pure Bayesian (Fully conditional)

Most Relevant but Least Robust

But life is about *compromise*:

Conditional frequentist, Objective Bayesian, Fiducial ...

Xiao-Li Meng

A Unified Picture of BFF (and Inference)?

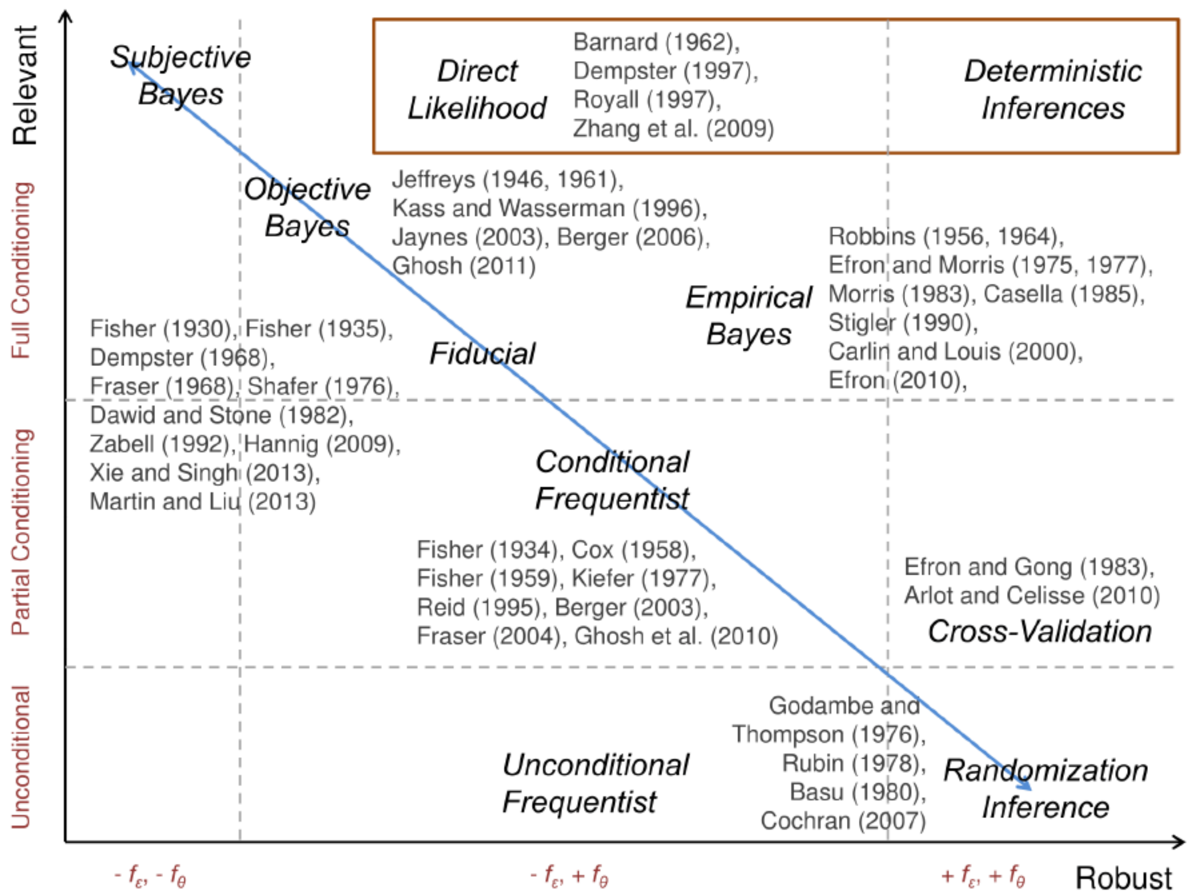
BFF 20/21

Xiao-Li Meng

Choose Your Replication!

Basu Ex

Summary



Steve Biller

This gets you nowhere!!

"A Frequentist uses impeccable logic to answer the wrong question, while a Bayesian answers the right question by making assumptions that nobody can fully believe in."

P.G. Hamer

ALWAYS ask the right question, even if the answer isn't necessarily straight-forward!!
(this is what being a scientist is all about)

Steve Biller

90% CL/CI upper bounds on a possible average signal level from a simple counting experiment

	Initial Test: B=5, n=2	Improved Cuts: B=0.5, n=0
Standard Frequentist	0.32	1.8 **(worse)**
Feldman-Cousins	1.73	1.94 **(worse)**
Bayesian (prior uniform in rate)	3.13	2.3 **(better)**

Can appear to be overly strict bounds on the average signal strength

New analysis technique: suppresses backgrounds by a factor of 10 with no loss in signal efficiency!

Steve Biller

What's the way out??

Pragmatism!

There is no "correct" choice of prior!

- Where possible, use informative priors or follow standard conventions (if they exist);
- Otherwise, choose simple prior forms that are easy to understand and visualise (*e.g.* uniform);
- Use common (*i.e.* standardised) parameter choices that "make sense" for these priors;
- If there's an ambiguity that leads to a non-conservative bound, show the sensitivity to the choice of prior

Bob Cousins

Continuous Mass Hierarchy variable?

The +1 and -1 for MH appear in the equations as simply that: arithmetic signs.

Various authors (e.g., Capozzi, Lisi, and Marrone, PRD 89 013001) have suggested replacing ± 1 with (unbounded) continuous variable α.

Reminiscent of continuous "number of light neutrino species" (which recall had BSM physics interpretation).

In frequentist treatment, I think it is mostly a matter of presentation, since results from discrete way map to continuous way, and vice versa (particularly if F-C construction is used for confidence interval for α, with relevant set of C.L.'s).

I encourage continuous α approach as part of toolkit.

But…Eligio Lisi has explained to me that α is highly correlated with Δm^2, and contributes to increase its overall uncertainty. This leads to the undesired result that power is lost due to consideration of unphysical (or at least non-SM) values of MH. Ugh.

NOTE added after talk: I mis-stated Eligio's point above at the time of the talk; I believe that it is now repaired. -BC

Bob Cousins

Addition of Nuisance Parameter δ to MH Test

Small variation of nuisance parameters seems not to upset the formalism, and some relevant examples with toys still give nicely Gaussian distribution of LR test statistic. However the situation can become harder – see talk by Sara Algeri at Tokyo.

If the CP phase δ is treated as a nuisance parameter in the MH determination, then great care is needed.

Providing the MH results as a function of δ (same δ in numerator and denominator of LR) would seem to be mandatory, before attempting to "eliminate" δ by profiling or marginalizing..

Bob Cousins

But something about "eliminating" δ_{CP} reminds me of the quote by "likelihoodist" A.W.F. Edwards:

"Let me say at once that I can see no reason why it should always be possible to eliminate nuisance parameters. Indeed, one of the many objections to Bayesian inference is that it always permits this elimination."

(commenting on J.D. Kalbfleisch and J.D. Sprott, J. Roy. Stat. Soc. Series B 32, 175 (1970). See my paper Oxford05.)

For further reading:

For PhyStat 2005, I wrote, "Treatment of nuisance parameters in high energy physics, and possible justifications and improvements in the statistics literature". Small compared to:

Luc Demortier, "P Values: What They Are and How to Use Them" http://www-cdf.fnal.gov/~luc/statistics/cdf8662.pdf (174 pages!)

Louis Lyons

Wilks' Theorem, contd

Examples: Does Wilks' Th apply?

1) H0 = polynomial of degree 3
   H1 = polynomial of degree 5

YES: ΔS distributed as χ^2 with ndf = (d-4) – (d-6) = 2

2) H0 = background only
   H1 = bgd + peak with free M_0 and cross-section

NO: H0 and H1 nested, but M_0 undefined when H1→ H0. $\Delta S \neq \chi^2$
(but not too serious for fixed M)

3) H0 = normal neutrino hierarchy *********
   H1 = inverted hierarchy *********

NO: Not nested. $\Delta S \neq \chi^2$ (e.g. can have $\Delta\chi^2$ negative)

N.B. 1: Even when W. Th. does not apply, it does not mean that ΔS is irrelevant, but you cannot use W. Th. for its expected distribution.

N.B. 2: For large ndf, better to use ΔS, rather than S_1 and S_0 separately