Statistical Methods in Neutrino Physics

Tom Junk

Fermilab

12th International Neutrino Summer School

August 7, 2019

Recommended Reading Material

- Particle Data Group reviews on Probability and Statistics.
  http://pdg.lbl.gov

- Frederick James, "Statistical Methods in Experimental Physics", 2nd edition, World Scientific, 2006

- Louis Lyons, "Statistics for Nuclear and Particle Physicists" Cambridge U. Press, 1989

- Glen Cowan, "Statistical Data Analysis" Oxford Science Publishing, 1998

- Roger Barlow, "Statistics, A guide to the Use of Statistical Methods in the Physical Sciences", (Manchester Physics Series) 2008.

- "Markov Chain Monte Carlo In Practice", W.R. Gilks, S. Richardson, and D. Spiegelhalter eds.

- Bob Cousins, "Why Isn't Every Physicist a Bayesian" Am. J. Phys **63**, 398 (1995).

Meetings on Statistics in HEP (with real Statisticians!)

The Phystat-Nu Series. Each has a link to suggested reading material for physicsists and for statisticians

https://indico.cern.ch/event/735431/ CERN, 2019
https://indico.fnal.gov/event/11906/ Fermilab, 2016
https://indico.ipmu.jp/indico/event/82/ At the IPMU Institute in Kashiwa, Japan, 2016

Phystat Series – tends to be collider-centric but still useful

http://indico.cern.ch/conferenceDisplay.py?confId=107747 Phystat 2011
http://www.physics.ox.ac.uk/phystat05/ Phystat 2005
http://www-conf.slac.stanford.edu/phystat2003/ Phystat 2013
http://conferences.fnal.gov/cl2k/

See Alex Himmel's talk at INSS 2017
https://indico.fnal.gov/event/13429/other-view?view=standard

And K. Cranmer's lectures at HCPSS 2013 http://indico.cern.ch/event/226365/
I am also very impressed with the quality and thoroughness of Wikipedia articles on general statistical matters.

The Scientific Method

"Classical Inference"

- Devise a hypothesis to test
  - Motivated by prior observations, or possibly not
  - Not already excluded
  - "Interesting" to the community or to everyone
    - Some hypotheses/measurements have technical value as inputs to subsequent high-profile measurements
  - Testable – it must predict something that is different from alternative hypotheses
  - "Null" vs. "Test" hypotheses (names not always applicable). You need at least two hypotheses to make a test
  - Precision measurements select from a continuous spectrum of hypotheses
  - A delicate balancing act – theorists work very hard to devise good hypotheses
- Design an experiment to test the hypothesis
  - Optimize the sensitivity at this stage
- Construct and operate the experiment
- Analysis: confront hypotheses with data.
  - Karl Popper: You can only rule out hypotheses, never prove one true.
- Estimate and include systematic uncertainties
- Report results!

Probability

- Testable hypotheses make predictions of observable data
- You need a full model of your experiment, including
  - The physics model being tested
  - Experimental apparatus
    - Beam flux and spectrum
    - Interaction cross sections (differential)
    - Detector response
    - Reconstruction and event selection
- Systematic uncertainties on all of the above

- Data are randomly drawn from true parent distributions which are not perfectly known.

Predictions of a model take the form of frequentist probabilities $p(\text{data}|\text{model})$. These are defined to be the fraction of experimental outcomes observing data in a large number of identical repeated experimental trials, assuming that the model is true.

The Binomial Distribution

Given *n* particles entering the detector, and each one has a probability *p* of interacting, then the distribution of the number of interactions *k* if the experiment is repeated many times is binomial.

The observed number of interactions is *k* (the "data").
n may also be observed (e.g. incoming charged particles in LArIAT or ProtoDUNE), but in neutrino experiments, it too is predicted from a flux and an exposure (running time and detector mass)

$$\mathrm{Binom}(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

where

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Some properties:

$$\langle k \rangle = pn$$

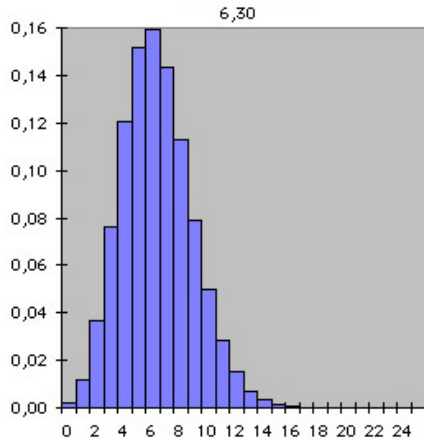
$$\sigma(k) = \sqrt{\mathrm{Var}(k)} = \sqrt{np(1-p)}$$

The sum of two binomally-distributed numbers with the same *p* is binomially distributed with that *p*. You can add your data together in a histogram.

The Poisson Distribution

In general, we start with *lots* of neutrinos, and very few interact with the detector. Binomial probabilities are difficult to work with – 10^{13}! is a big, big number.

$$\lim_{n\to\infty} \text{Binom}\left(k\middle|n, p=\frac{r}{n}\right) = \text{Poiss}(k|r) = \frac{r^k e^{-r}}{k!}$$



r is the rate.

$$\langle k \rangle = r$$

Var(*k*)=*r*

$$\sigma(k) = \sqrt{\text{Var}(k)} = \sqrt{r}$$

$$\sum_{k=0}^{\infty} \text{Poiss}(k|r) = 1 \quad \forall\, r$$

$$\int_{r=0}^{\infty} \text{Poiss}(k|r)dr = 1 \quad \forall\, k$$

Commonly used to model radioactive decay event distributions. They're really binomial, but the number of atoms is usually so big it is a great approximation

Composition of Poisson and Binomial Distributions

Say we have a rate of σL events, but our selection efficiency is ε

$$\text{Poiss}(k \mid \varepsilon\sigma L) = \sum_{N=0}^{\infty} \text{Binom}(k \mid N, \varepsilon)\text{Poiss}(N \mid \sigma L)$$

A more general rule: The law of conditional probability

P(A and B) = P(A|B)P(B) = P(B|A)P(A) more on this one later

And in general, $$P(A) = \sum_{B} P(A \mid B)P(B)$$

Joint Probability of Two Poisson Distributed Numbers

Example -- two bins of a histogram
Or -- Monday's data and Tuesday's data

$$\text{Poiss}(x \mid \mu) \times \text{Poiss}(y \mid \nu) = \text{Poiss}(x + y \mid \mu + \nu) \times \text{Binom}\left(x \mid x + y, \frac{\mu}{\mu + \nu}\right)$$

The sum of two Poisson-distributed numbers is Poisson-distributed with the sum of the means ("Raikov's Theorem")

$$\sum_{k=0}^{n} \text{Poiss}(k \mid \mu)\text{Poiss}(n - k \mid \nu) = \text{Poiss}(n \mid \mu + \nu)$$

Application: You can rebin a histogram and the contents of each bin will still be Poisson distributed (just with different means)

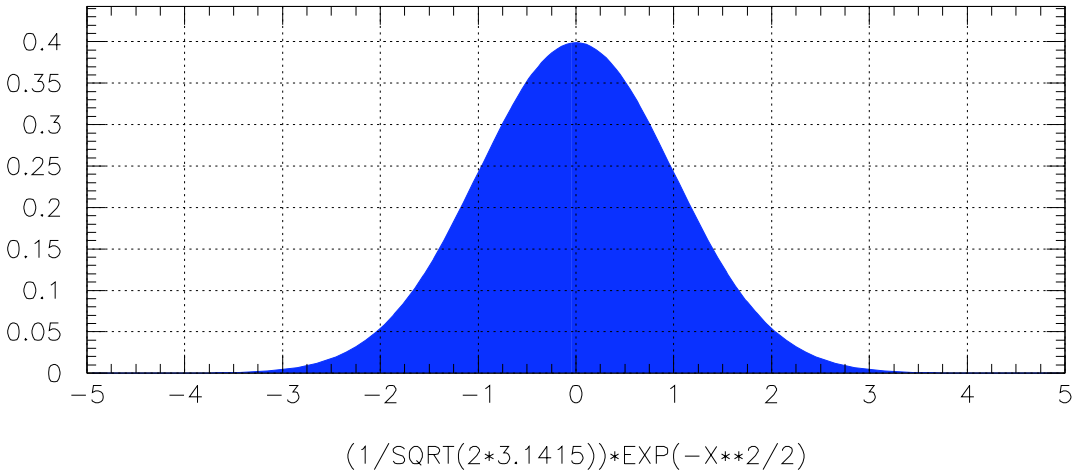
Question: How about the difference of Poisson-distributed variables?

The Gaussian (or "Normal") Distribution

$$\text{Gauss}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{\sigma^2}}$$

Mean: $\langle x \rangle = \mu$

Width: Var(x) = σ^2

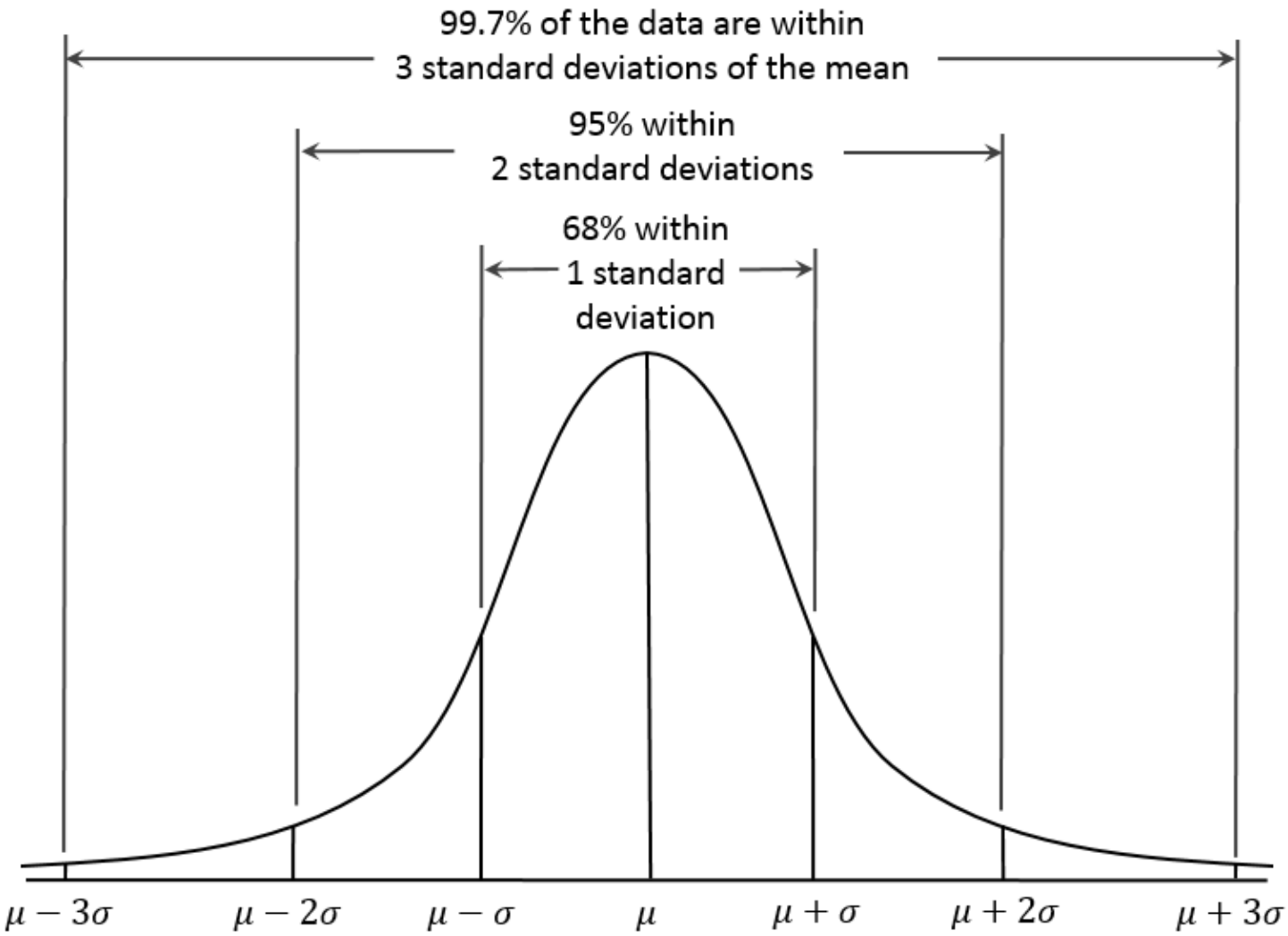


Sum of Two Independent Gaussian Distributed Numbers is Gaussian with the sum of the means and the sum in quadrature of the widths

$$\text{Gauss}\left(z,\mu+\nu,\sqrt{\sigma_x^2+\sigma_y^2}\right) = \int_{-\infty}^{\infty} \text{Gauss}(x,\mu,\sigma_x)\,\text{Gauss}(z-x,\nu,\sigma_y)\,dx$$

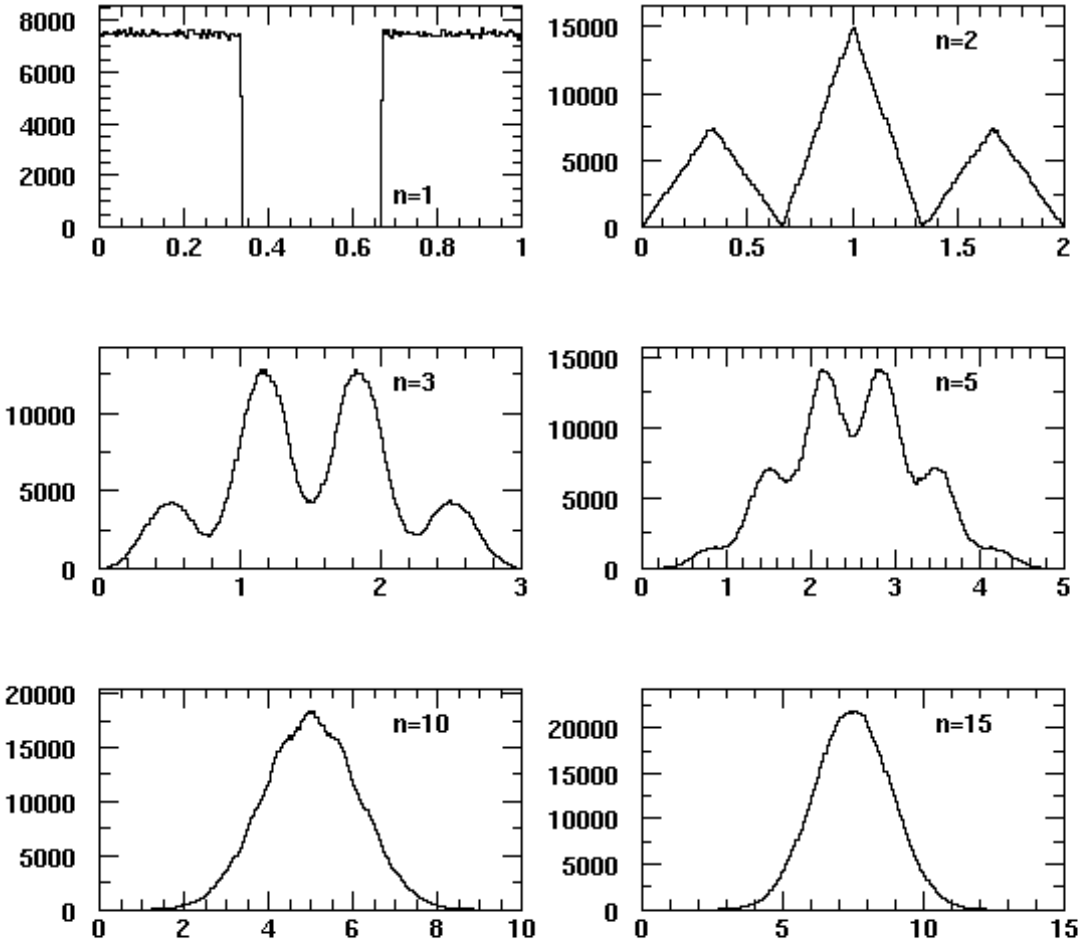
A difference of independent Gaussian-distributed numbers is also Gaussian distributed (widths still *add* in quadrature)

Areas under the Normal Distribution Curve



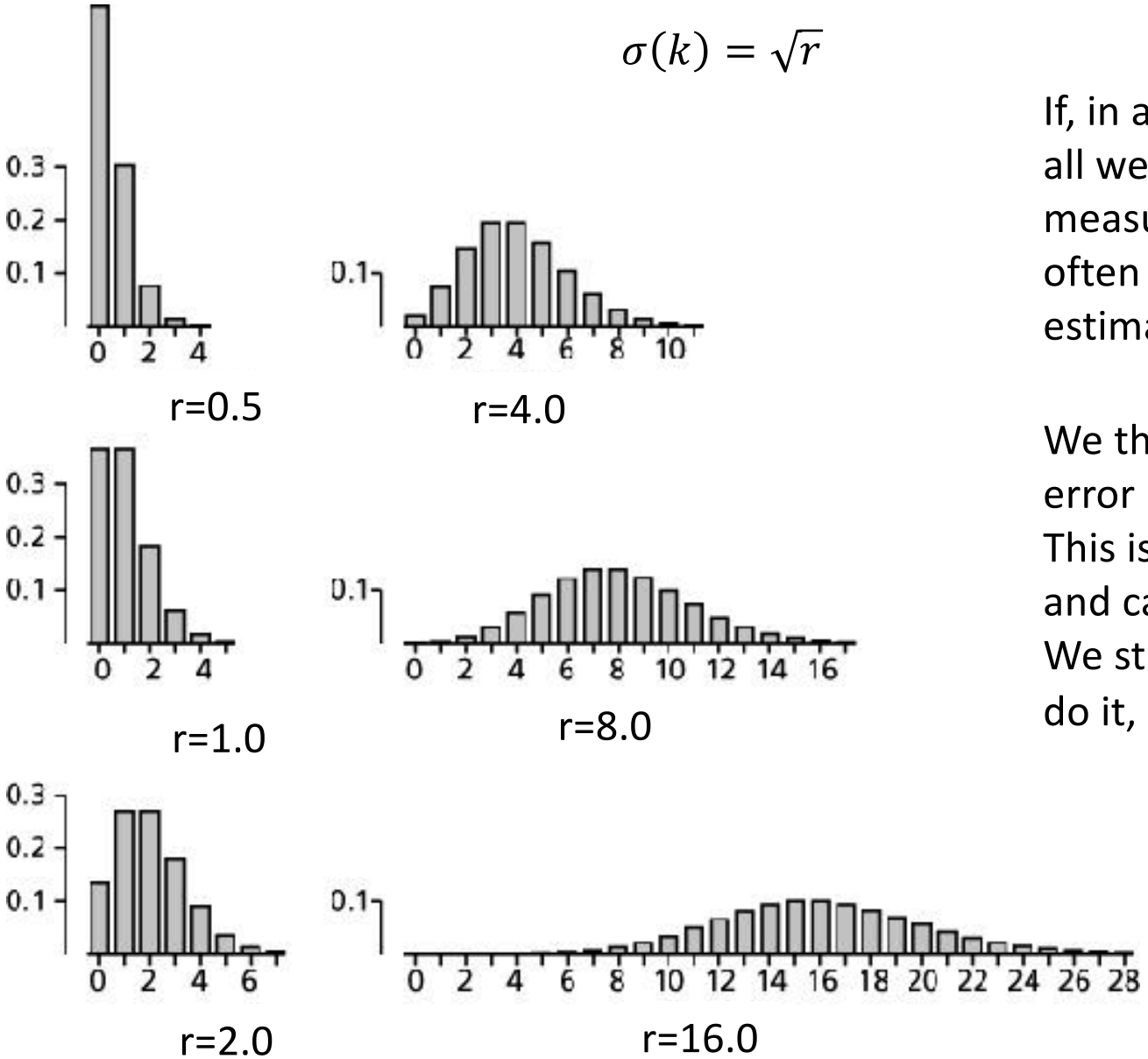
The Central Limit Theorem

The sum of many small, uncorrelated random numbers is asymptotically Gaussian distributed -- and gets more so as you add more random numbers in. *Independent of the distributions of the random numbers* (as long as they stay small).



Poisson for large *r* is Approximately Gaussian of width

$$\sigma(k) = \sqrt{r}$$



If, in an experiment all we have is a measurement n, we often use that to estimate *r*.

We then draw $\sqrt{k}$ ("root n") error bars on the data. This is just a *convention*, and can be misleading. We still recommend you do it, however.

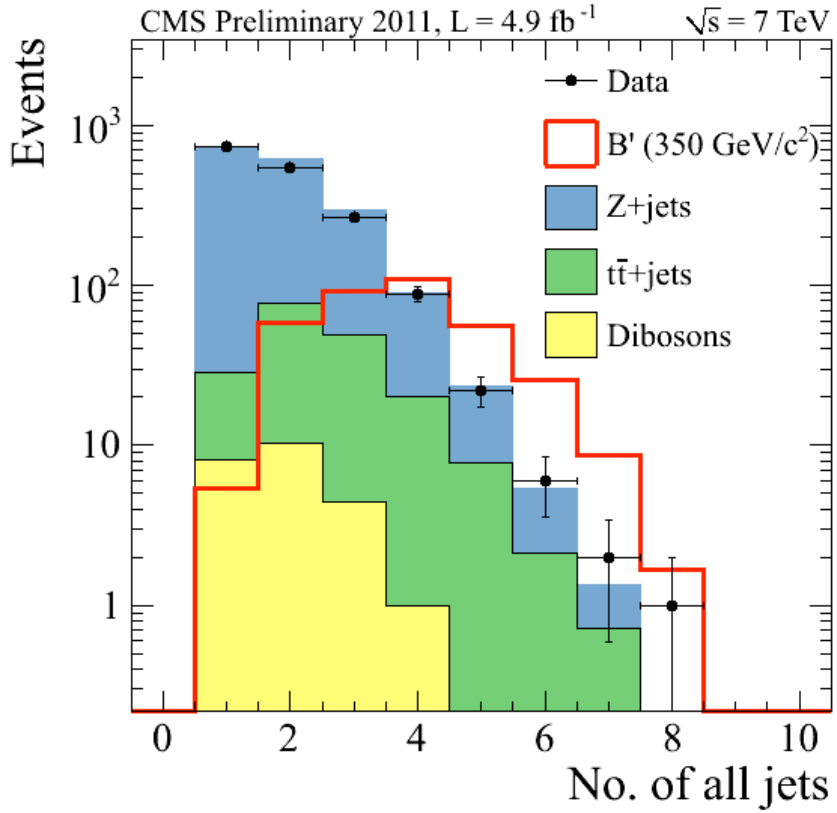
Why Put Error Bars on the Data?

- To identify the data to people who are used to seeing it this way
- To give people an idea of how many data counts are in a bin when they are scaled (esp. on a logarithmic plot).
- So you don't have to explain yourself when you do something different (better)

But: $\sqrt{k} \neq \sqrt{r}$

The true value of *r* is usually unknown

"Il n'est pas certain que tout soit incertain. (Translation: It is not certain that everything is uncertain.)" — Blaise Pascal, Pascal's Pensees



https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsEXO11066

Aside: Errors on the Data? (answer: no)

Standard to make MC histograms with no errors: Data points with error bars:

$$n_{\text{obs}} \pm \sqrt{n_{\text{obs}}}$$

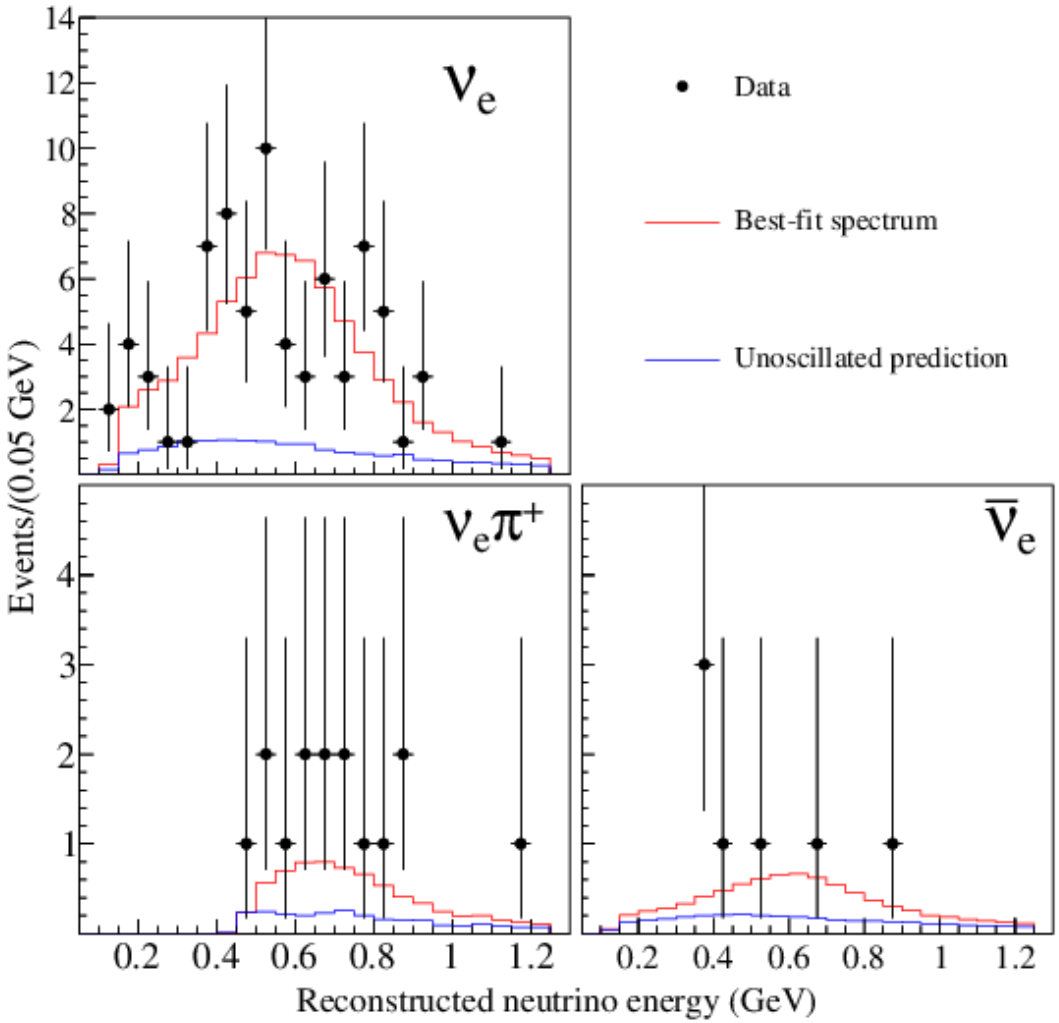
But we are not uncertain of n_{obs}! We are only uncertain about how to interpret our observations; we know how to count.

Collider example. Neutrino papers often put prediction uncertainties in separate tables.

	Signal	$ZZ^{(*)}$	Z + jets, $t\bar{t}$	Observed
4μ	2.09±0.30	1.12±0.05	0.13±0.04	6
$2e2\mu/2\mu2e$	2.29± 0.33	0.80±0.05	1.27±0.19	5
$4e$	0.90±0.14	0.44±0.04	1.09±0.20	2

ATLAS Collab., arXiv:1207.7214

Sometimes another convention is adopted for showing error bars on the data



There are several options.

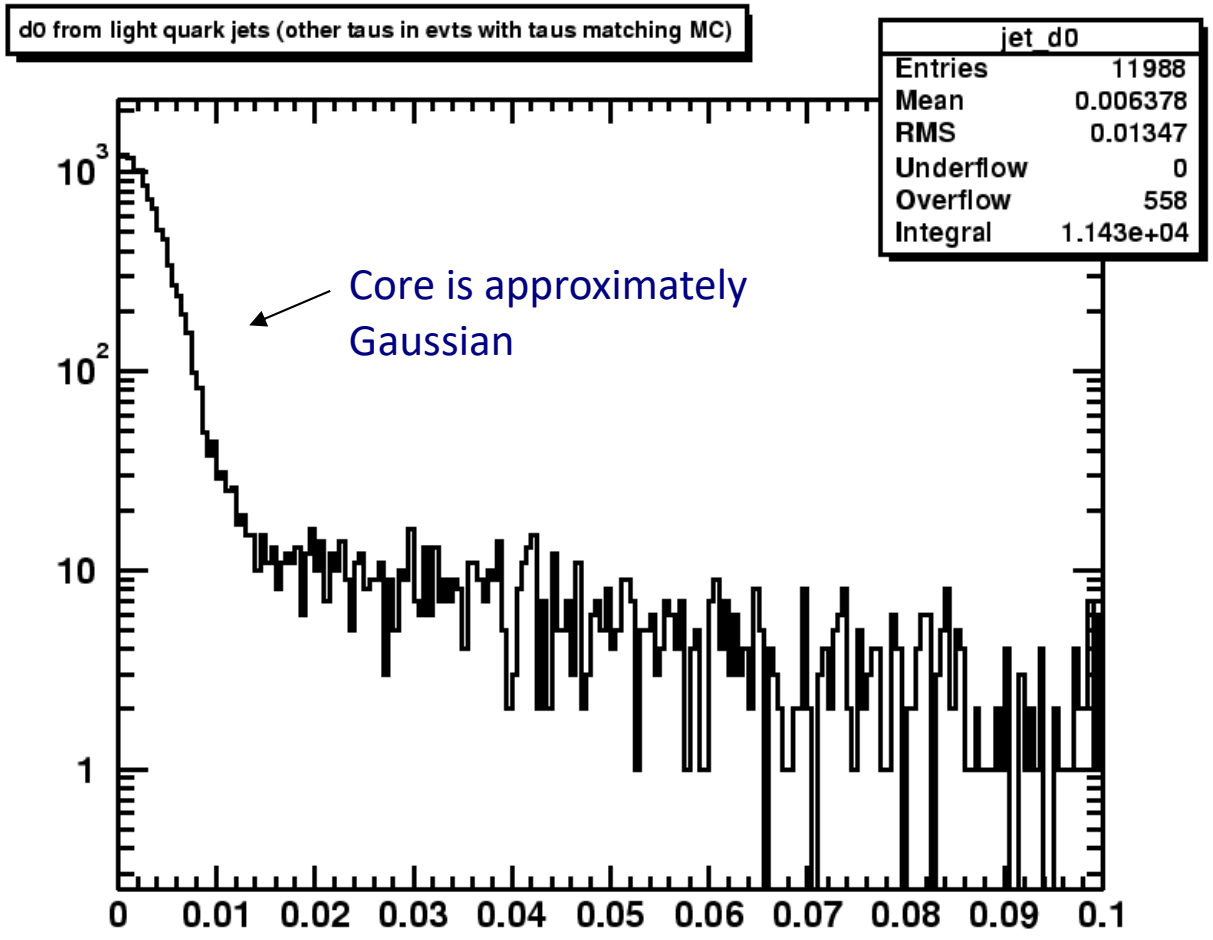
Need to explain which one is chosen

T2K Collaboration, Phys.Rev.Lett. 121 (2018) no.17, 171802

Not all Distributions are Gaussian

Track impact parameter distribution for example

Multiple scattering -- core: Gaussian; rare large scatters; heavy flavor, nuclear interactions, decays (taus in this example)



"All models are false. Some models are useful."

Statistics is the *inverse* of Probability

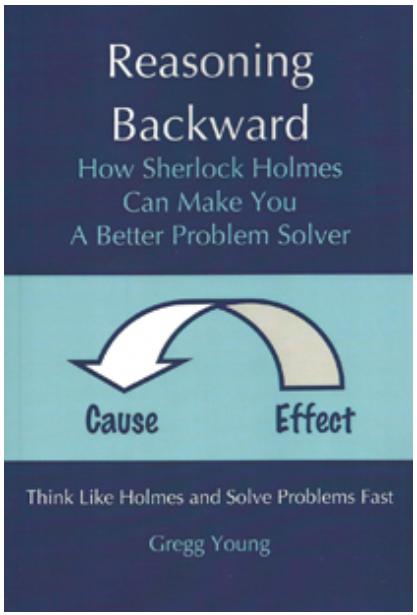
Probability:

Models with parameters → Predicted data distributions

Statistics:

Sample of data drawn from an unknown true distribution → Tests of models and parameter values

I have not read this book but I like the cover.



Guess which is easier! Inverse reasoning is usually *ill-posed*.
We seek to quantify the range of possible models consistent with data.

Back to probability: Characterization of performance of statistical methods.

Usually you need to calculate experiment sensitivity before you run the experiment.

Statistical Uncertainty on an Average of Independent Random Numbers Drawn from the Same Gaussian Distribution

Useful buzzword: "IID" = "Independent, identically distributed

N measurements, $x_i \pm \sigma$ are to be averaged

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

is an unbiased estimator of the mean μ

The square root of the variance of the sum is $\sqrt{N\sigma^2}$

so the standard deviation of the distribution of averages is

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{N}}$$

Worth Remembering this formula!

Estimating the Width of a Distribution

It's the square Root of the Mean Square (**RMS**) deviation from the true mean

n.b. Physics notation of RMS is RMS of the differences from the mean, not just raw RMS

$$\sigma_{est}(\mu_{true}\ \text{known}) = \sqrt{\frac{\sum_i (x_i - \mu_{true})^2}{N}}$$

BUT: The true mean is usually not known, and we use the same data to estimate the mean as to estimate the width. One **degree of freedom** is used up by the extraction of the mean.

This narrows the distribution of deviations from the average, as the average is closer to the data events than the true mean may be. An unbiased estimator of the width is:

$$\sigma_{est}(\mu_{true}\ \text{unknown}) = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{N-1}}$$

How Uncertainties get Used

- Measurements are inputs to other measurements -- to compute uncertainty on final answer need to know uncertainty on parts.
- Measurements are averaged or otherwise combined -- weights are given by uncertainties
- Analyses need to be optimized -- shoot for the lowest uncertainty
- Collaboration picks to publish one of several competing analyses -- decide based on sensitivity
- Laboratories/Funding agencies need to know how long to run an experiment or even whether to run. PINGU and P5.

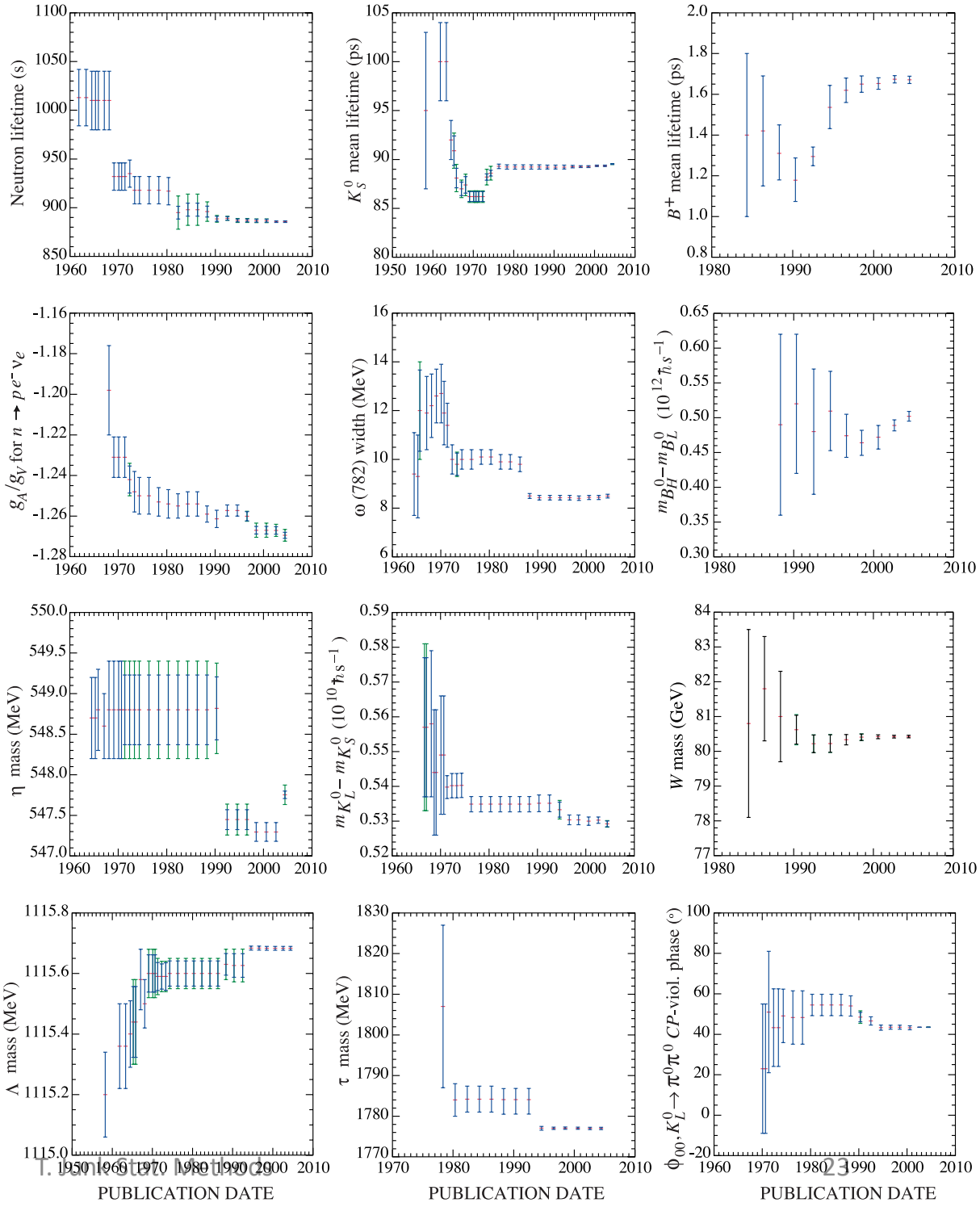
Statistical uncertainty: scales with data (1/sqrt(L). Systematic uncertanty often does too, but many components stay constant -- limits to sensitivity.

The Difference between "Error" and "Uncertainty"

Particle physicsists tend to use these words interchangeably, but they really mean different things.

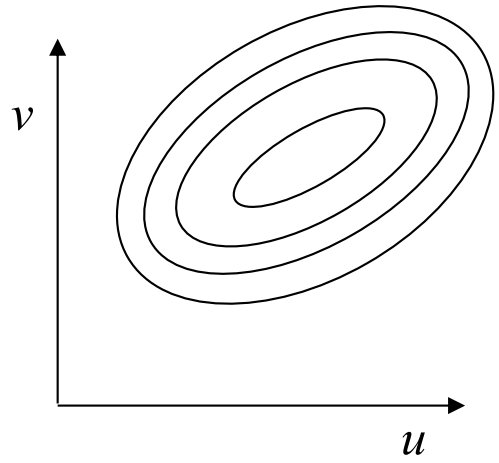
- Error = (measured – true): Usually the error is unknown
- Uncertainty: A prior or posterior distribution of the error, often represented as just one or two numbers.

Examples from the front of the PDG



Propagation of Uncertainties

Covariance: $\sigma^2_{uv} = \langle (u - \bar{u})(v - \bar{v}) \rangle$



If

$$x = au + bv$$

then

$$\sigma^2_x = a^2\sigma^2_u + b^2\sigma^2_v + 2ab\sigma^2_{uv}$$

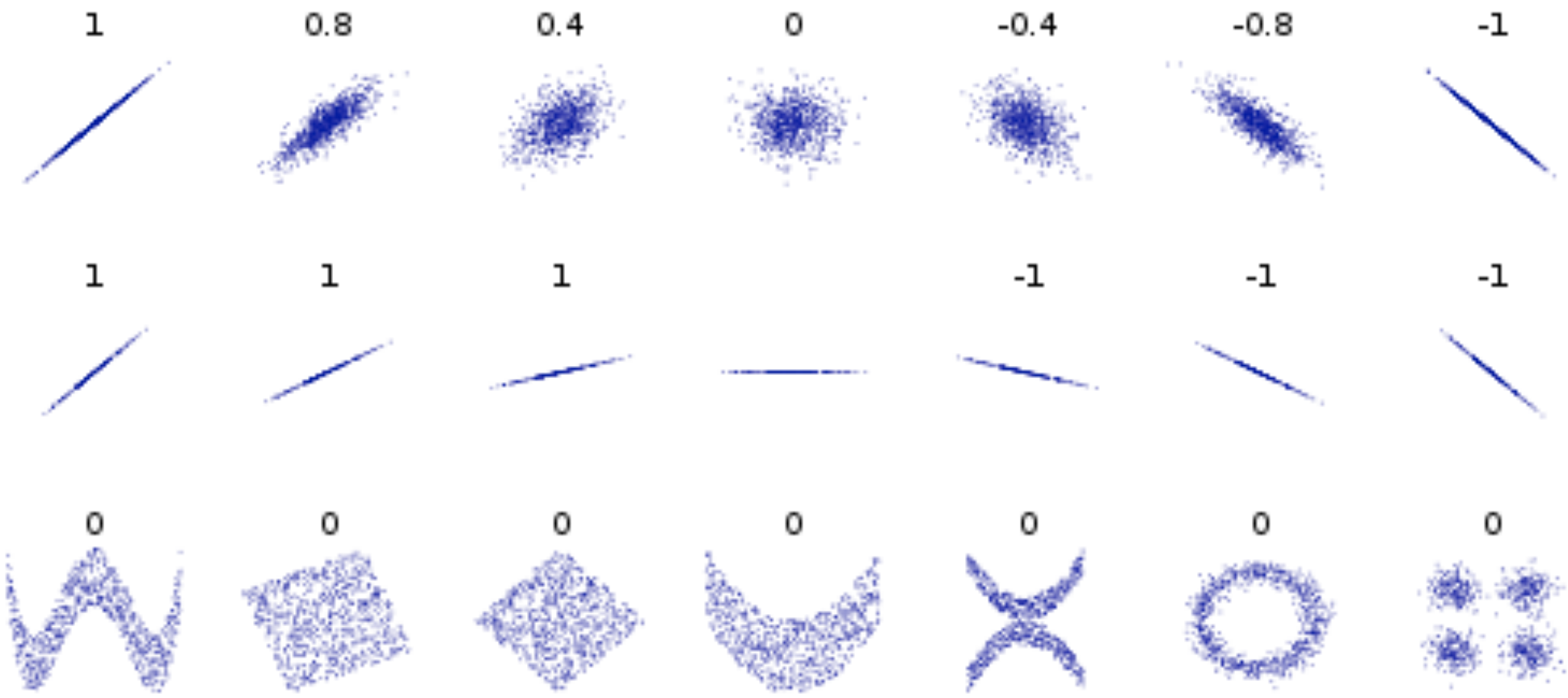
This can even vanish! (anticorrelation)

In general, if $x = f(u, v)$

$$\sigma^2_x = \left(\frac{\partial x}{\partial u}\right)^2 \sigma^2_u + \left(\frac{\partial x}{\partial v}\right)^2 \sigma^2_v + 2\left(\frac{\partial x}{\partial u}\right)\left(\frac{\partial x}{\partial v}\right)\sigma^2_{uv}$$

Zero Covariance Does NOT Imply Independent!

$$\rho_{xy} = \sigma_{xy}^2 \Big/ (\sigma_x \sigma_y)$$



http://en.wikipedia.org/wiki/Correlation_and_dependence

χ^2 Fitting and Goodness of Fit

For n independent Gaussian-distributed random numbers, the probability of an outcome (for known σ_i and μ_i) is given by

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} g(x_i, \mu_i, \sigma_i)$$

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-(x_i-\mu_i)^2/2\sigma_i^2}$$

If we are interested in fitting a distribution (we have a model for the μ_i in each bin with some fit parameters) we can maximize *p* or equivalently minimize

$$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} = -2 \ln p + c$$

σ_i includes stat. and syst. errors

For fixed μ_i this χ^2 has *n* degrees of freedom (DOF)

Counting Degrees of Freedom

$\chi^2 = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$ has n DOF for fixed μ_i and σ_i

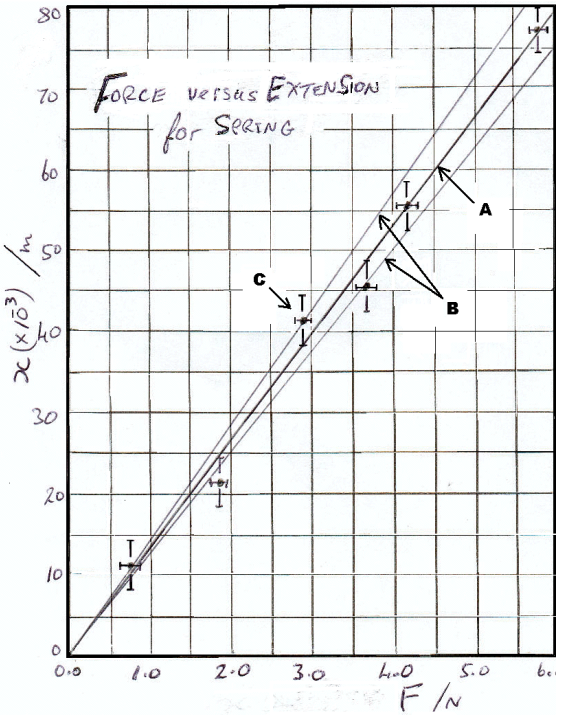
If the μ_i are predicted by a model with free parameters (e.g. a straight line), and χ^2 is minimized over all values of the free parameters, then

Approximate! Not always! (*cough*)

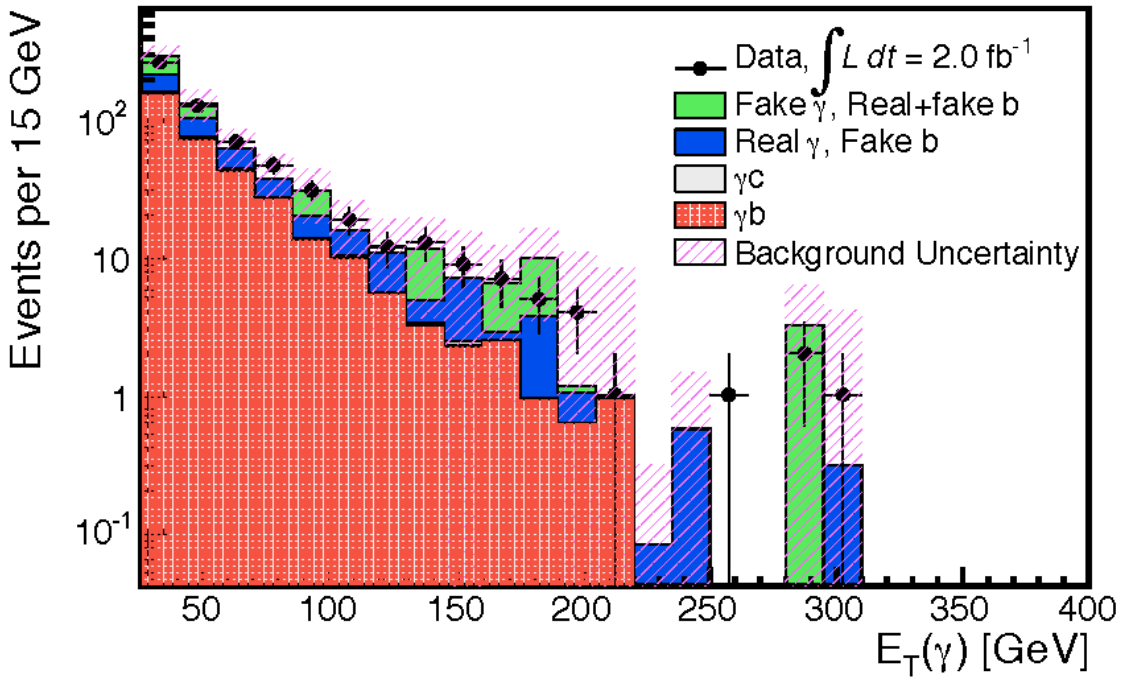
DOF = n - #free parameters in fit.

Example: Straight-line least-squares fit:
DOF = npoints - 2 (slope and intercept float)

With one constraint: intercept = 0,
6 data points, DOF = ?



MC Statistics and "Broken" Bins



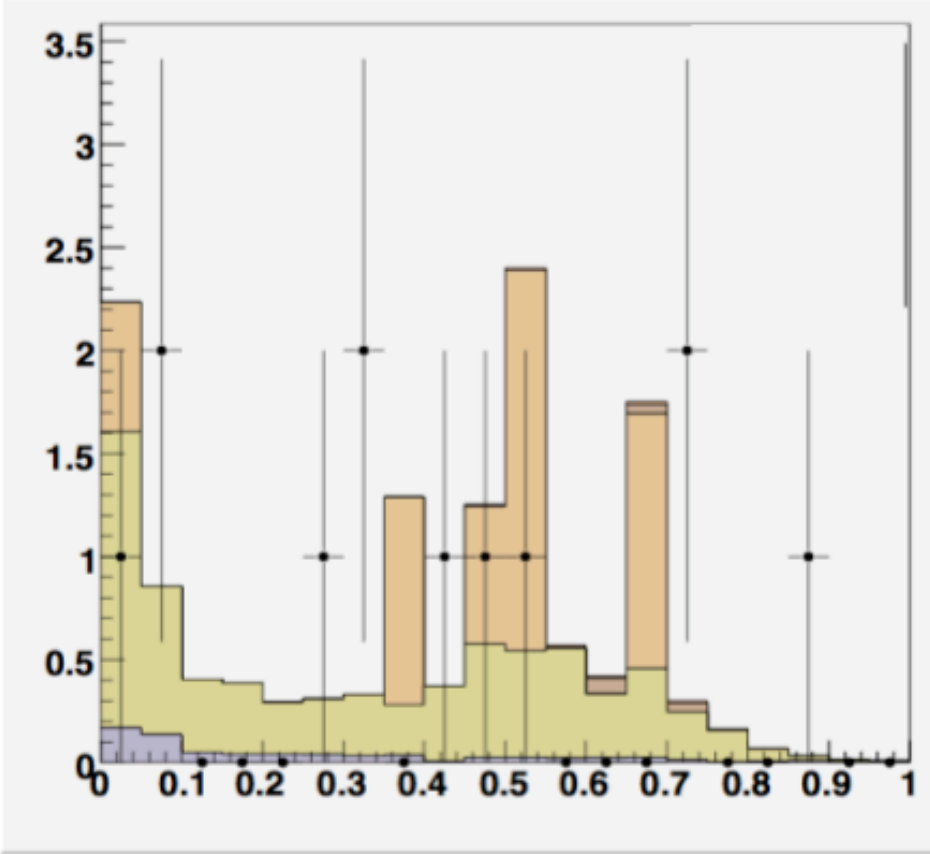
NDOF=?

- Automated tools cannot tell if the background expectation is really zero or just a downward MC fluctuation.
- Real background estimations are sums of predictions with very different weights in each MC event (or data event)
- Rebinning or just collecting the last few bins together often helps.

- Advice: Make your own visible underflow and overflow bins (do not rely on ROOT's underflow/overflow bins -- they are usually not plotted. Tools may ignore u/o bins

"Partially" Broken Bins? How Can we Tell the Bins are Broken?

An Extreme Example (names removed)



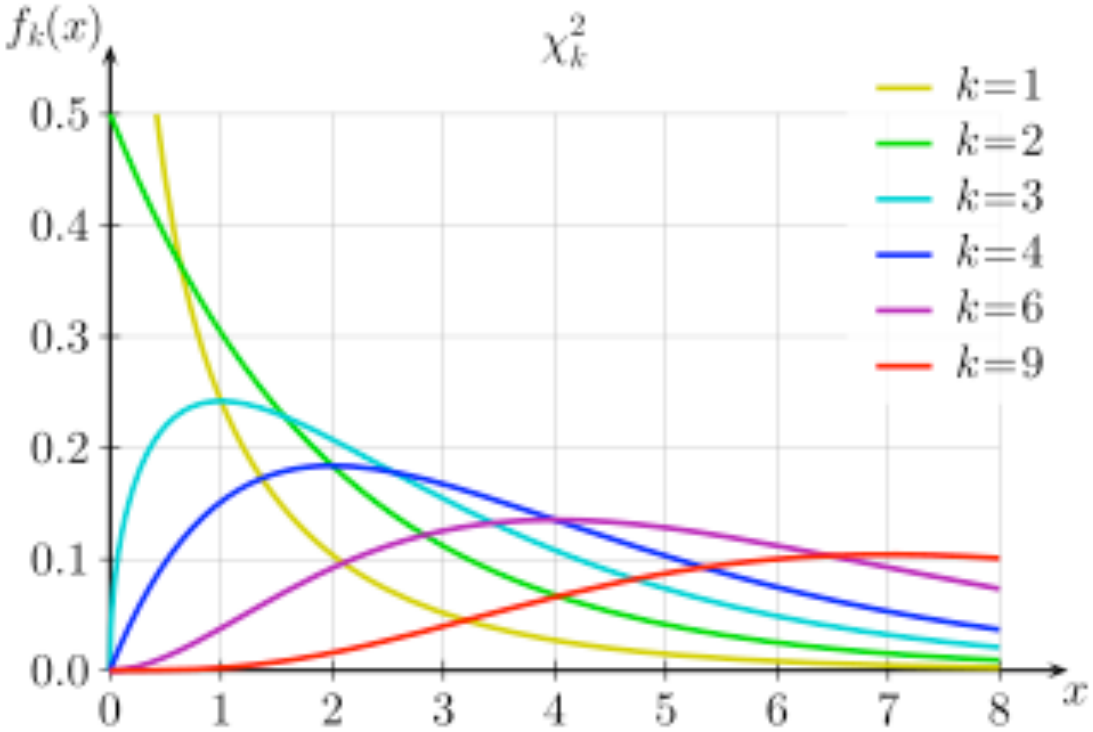
Questions: What's the shape we are trying to estimate?
What is the uncertainty on that shape?

There many not be enough information in this histogram to determine shape.

One bin may be right answer.

Orange contribution was estimated from a data sideband – hard just to run some more MC to fix the problem!

The χ^2 Distribution



Plot from Wikipedia:
"k" = number of degrees of Freedom

PDF $$\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}\, x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

Cumuliative Distribution $$\frac{1}{\Gamma\left(\frac{k}{2}\right)}\, \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$$

Mean: k

Assumes errors are Gaussian,
the model is true, and
and uncertainties are correct.

χ^2 and Goodness of Fit

- Gaussian-distributed random numbers cluster around μ_i ~68% within 1σ. 95% within 2σ. Very few outside 3 sigma.

TMath::Prob(Double_t Chisquare,Int_t NDOF)

Gives the chance of seeing the value of Chisquared or bigger given NDOF.

This is a **p-value** (more on these later)

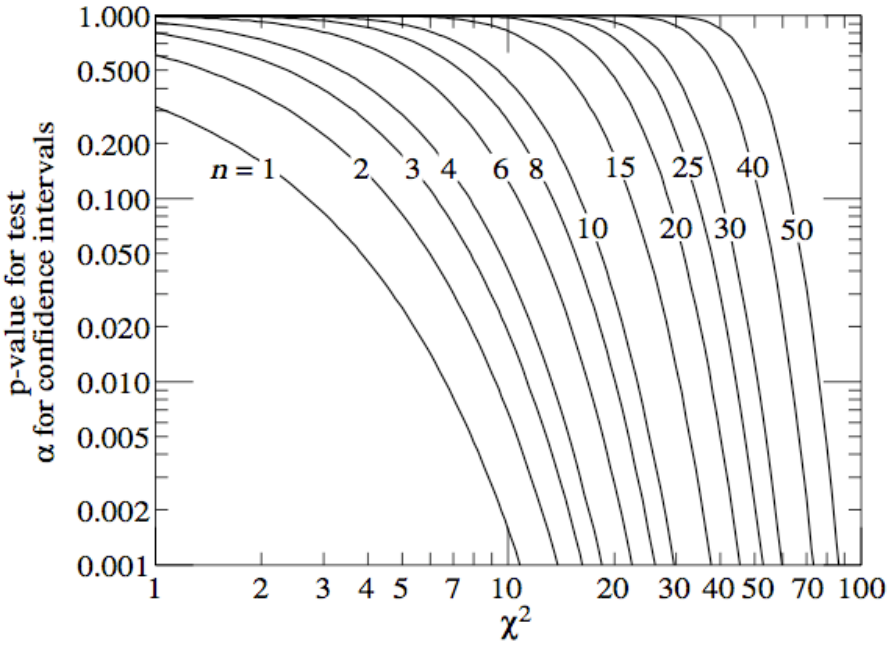
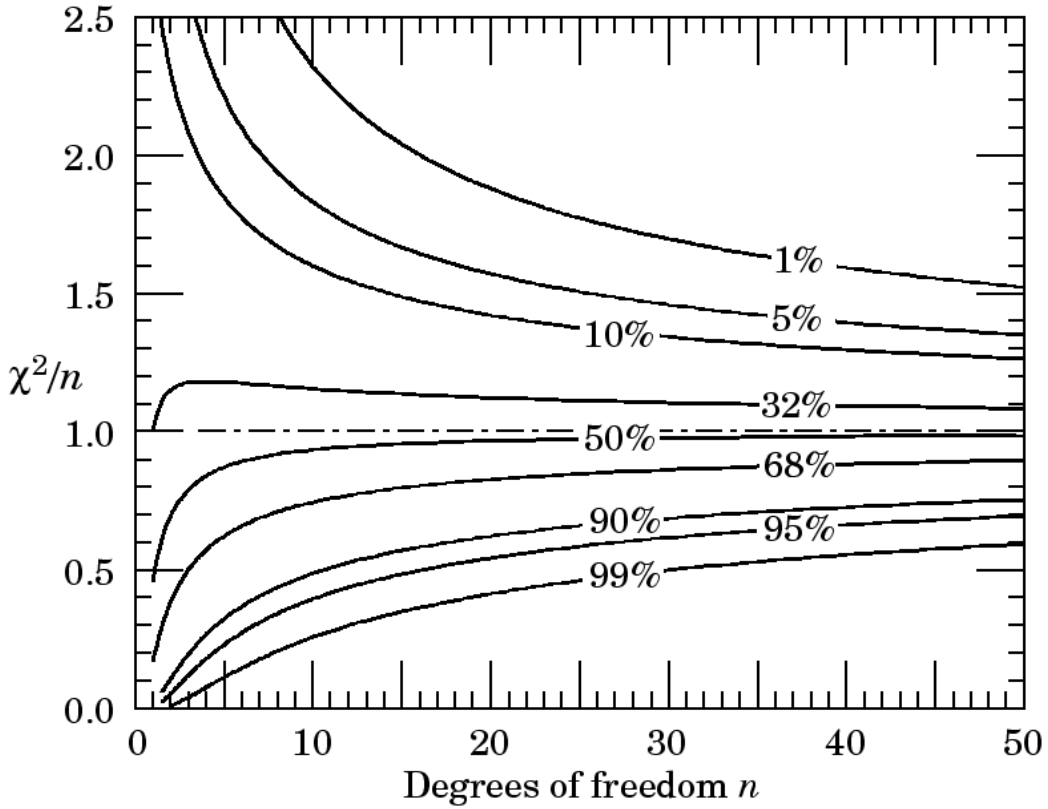


Figure 33.1: One minus the χ^2 cumulative distribution, $1 - F(\chi^2; n)$, for n degrees of freedom. This gives the p-value for the χ^2 goodness-of-fit test as well as one minus the coverage probability for confidence regions (see Sec. 33.3.2.4).

A Rule of Thumb Concerning χ^2

Average contribution to χ^2 per DOF is 1. χ^2/DOF converges to 1 for large n for distributions compared with the true model with correct uncertainties and Gaussian-distributed errors.



From the PDG Statistics Review

n.b. You can make χ^2/DOF as small as you like by overbinning Poisson data.

No such thing as an unbinned χ^2/DOF test (though some have tried)

An Example of a Dodgy Fit

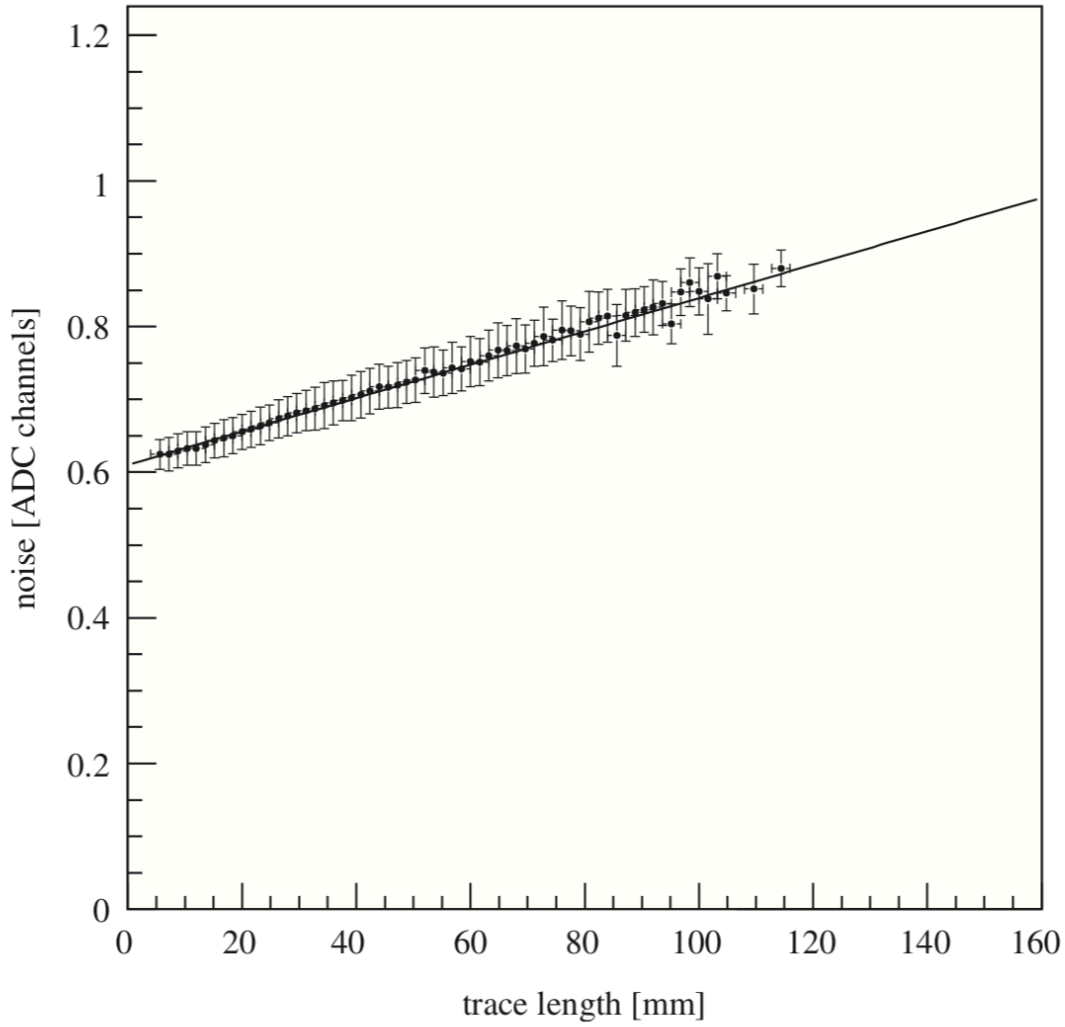


Fig. 65. Correlation of noise and the trace length on the pad-plane PCB board. A straight line was fit to the data.

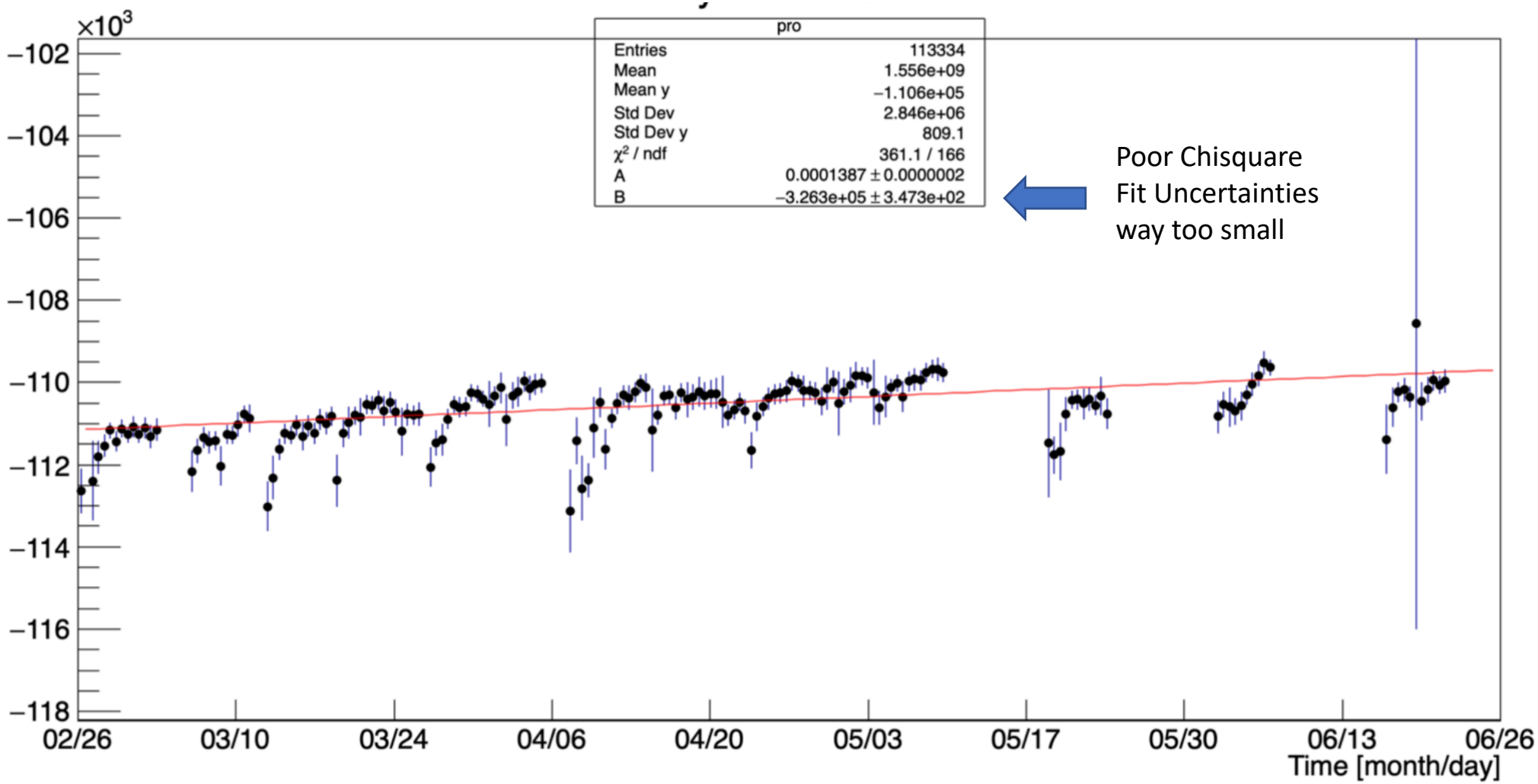
J. Alme et al. / Nuclear Instruments and Methods in Physics Research A 622 (2010) 316–367

Error bars are either correlated or overestimated

A naive least-squares fit will give uncertainties that are unreliable for the slope and intercept

Chisquared (not shown) per DOF is tiny

Another Questionable Fit



Labels removed. Model is naive (or error bars do not cover known reasons for deviation from the model). It was probably good enough for the purpose though.

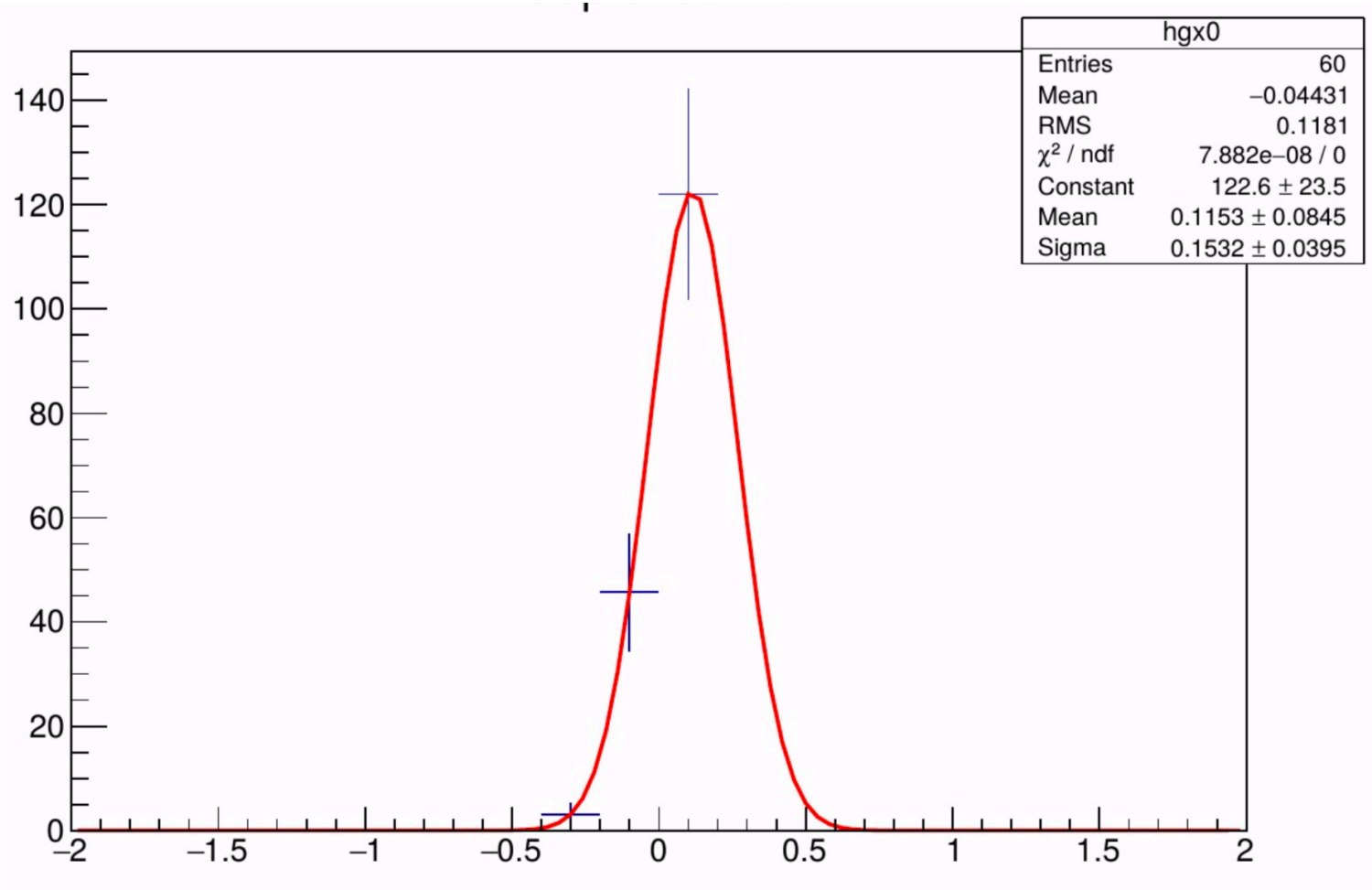
What ROOT Does By Default

Default is Neyman's chisquared

Error = sqrt(nobs) in each bin.

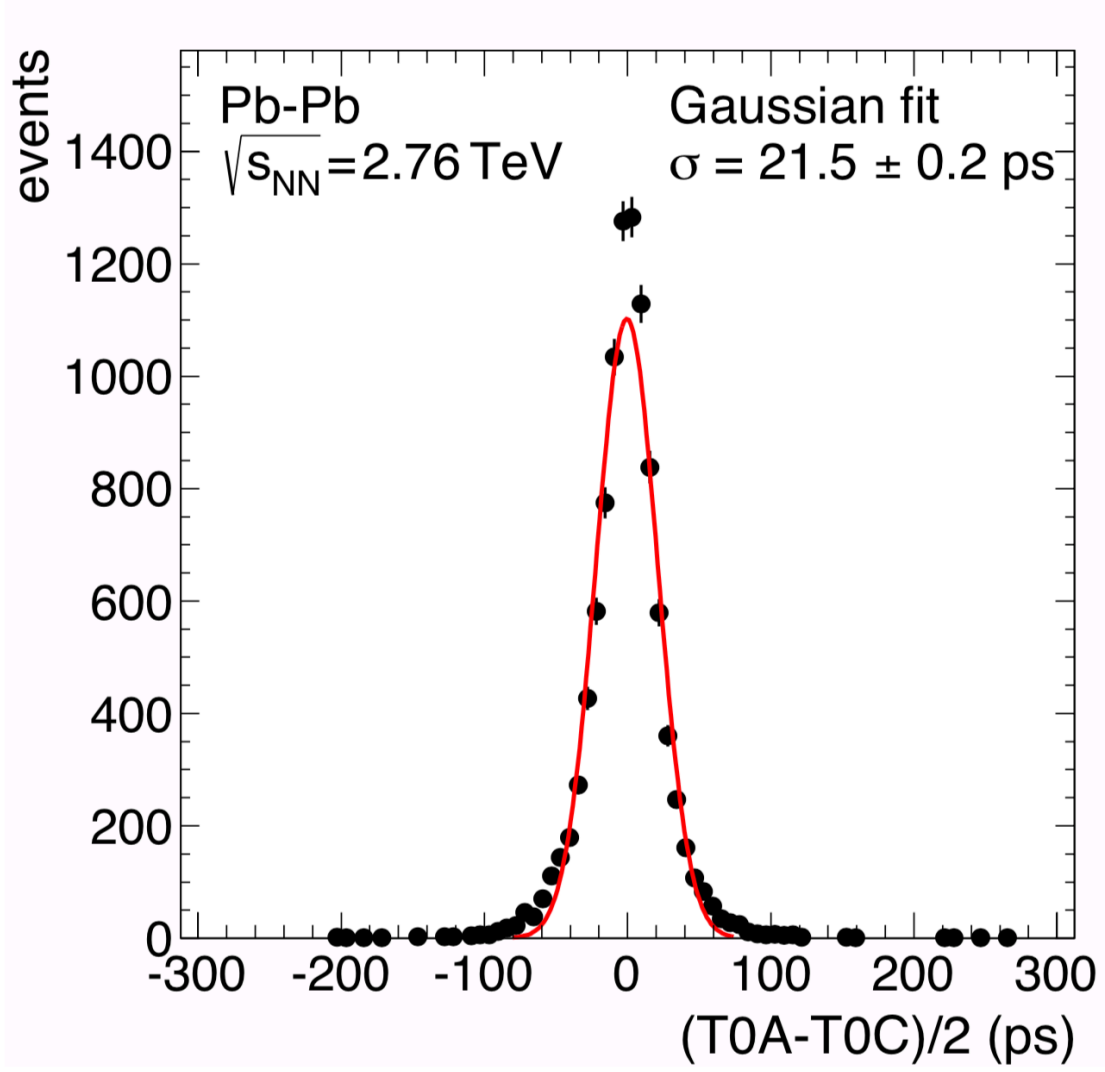
Ignores bins with nobs = 0.

Use the "L" option in TH1::Fit() to use the Poisson likelihood instead



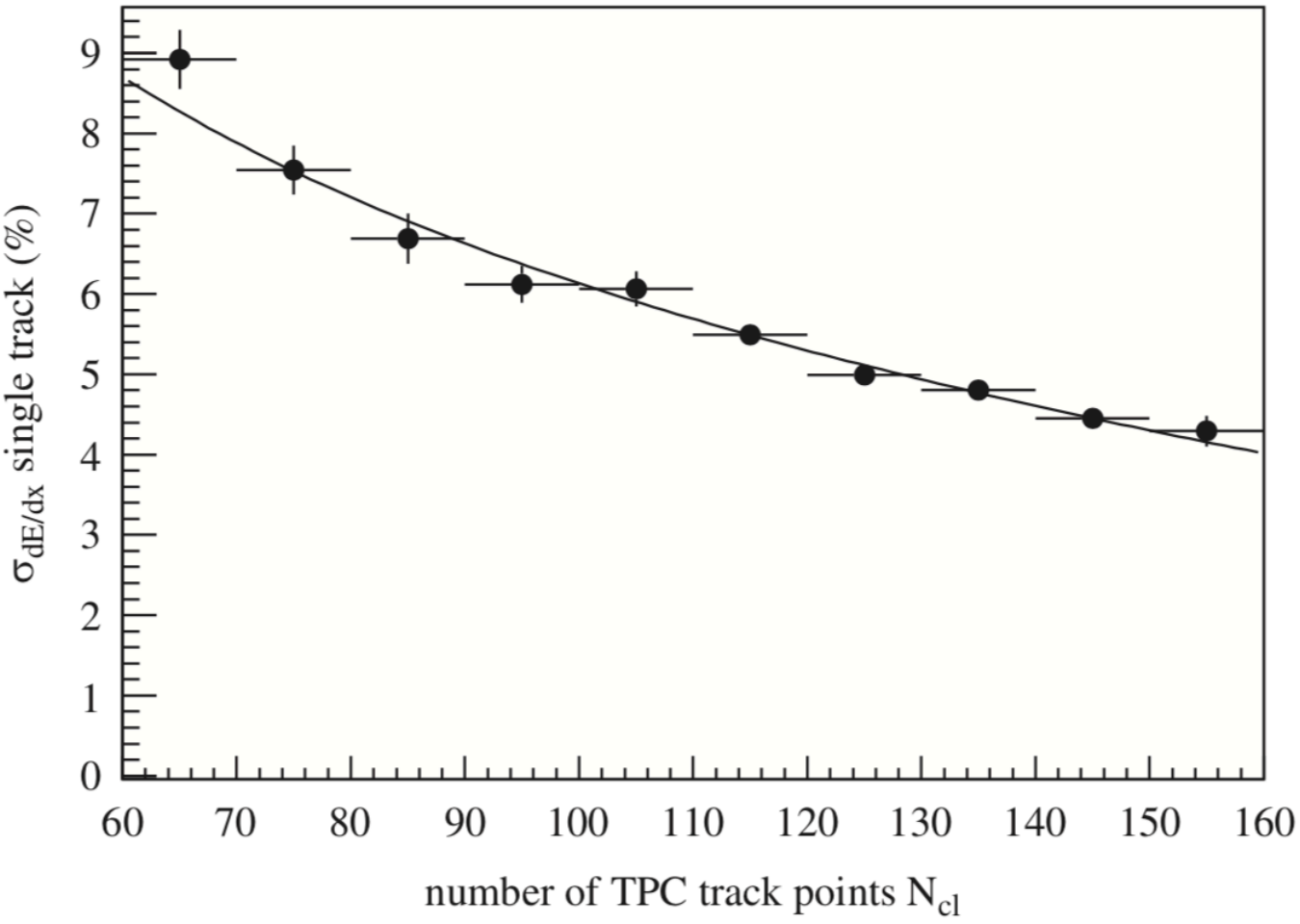
A Typical Situation: Data are both too wide *and* too narrow for the Gaussian model

Solution – try fitting a sum of two Gaussians. If the widths are similar, the uncertainties will be highly correlated.



ALICE Collab,
arXiv:1402.4476

A Reasonable Fit



The fact that the most discrepant point is the one on the end might raise a question but the fluctuations look real.

J. Alme et al. / Nuclear Instruments and Methods in Physics Research A 622 (2010) 316–367

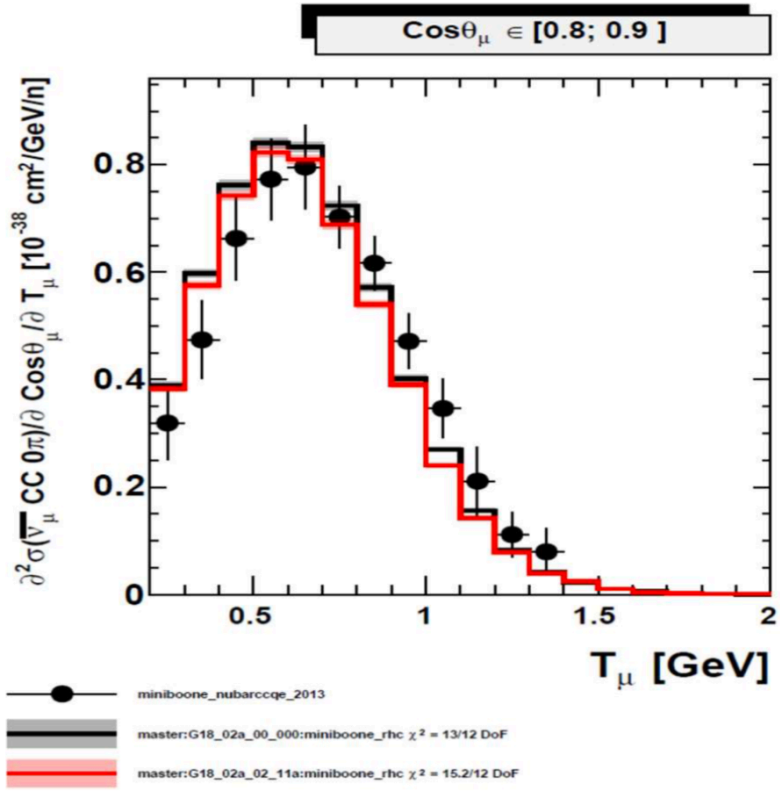
Other Goodness-of-Fit Tests

One you can do in your head (sort of)

The run test. Count how many consecutive data points are below the prediction, or how many are above the prediction in a row.

Residuals of mismodeled distributions often have a "wavy" structure to them.

Even if the χ^2 is good, the run test may show a problem



S. Dytman

http://npc.fnal.gov/wp-content/uploads/2018/11/fnal-dytman-nov18.pdf

There is no substitute for looking critically at your data! And your MC!

The Kolmogorov-Smirnov GOF Test

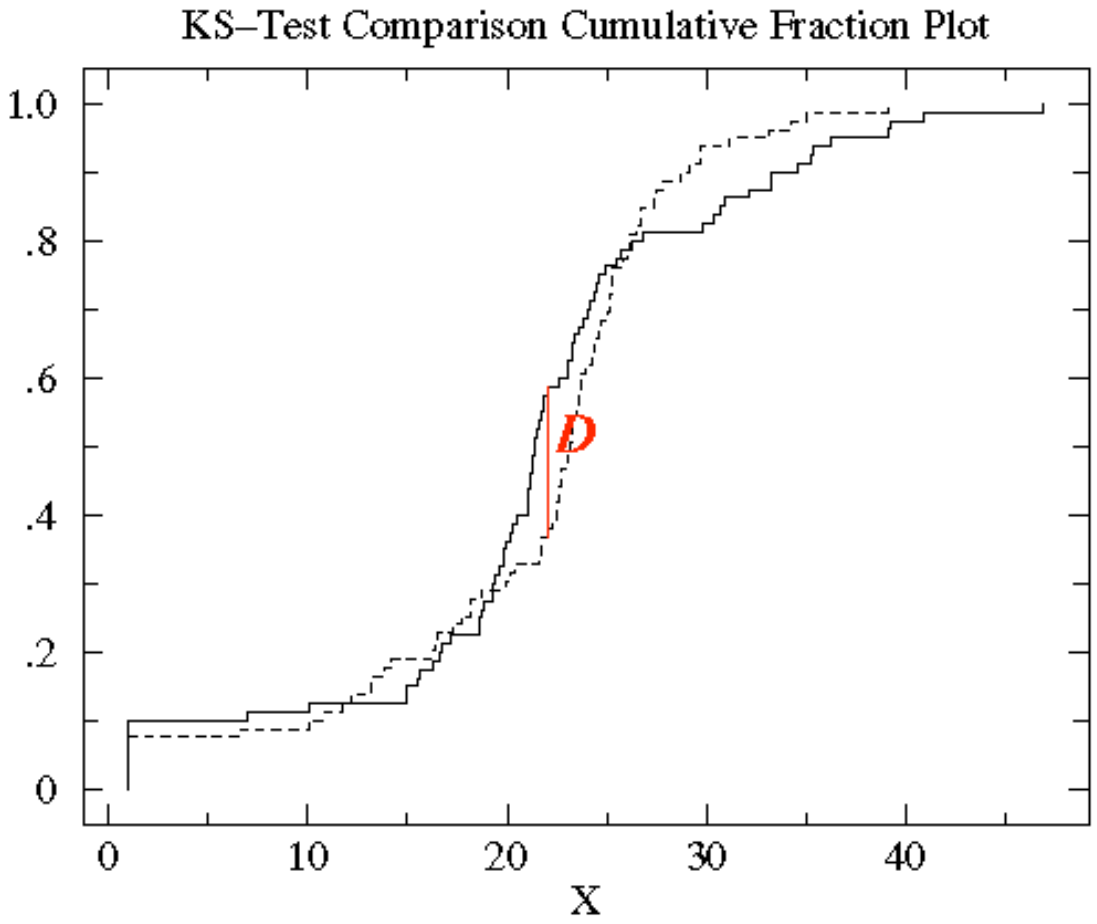
χ^2 Doesn't tell you everything you may want to know about distributions that have modeling problems.

Ideally, it is a test of two unbinned distributions to see if they come from the same parent distribution.

Procedure:

- Compute normalized, cumulative distributions of the two unbinned sets of events. Cumulative distributions are "stairstep" functions
- Find the maximum distance D between the two cumulative distributions

called the "KS Distance"



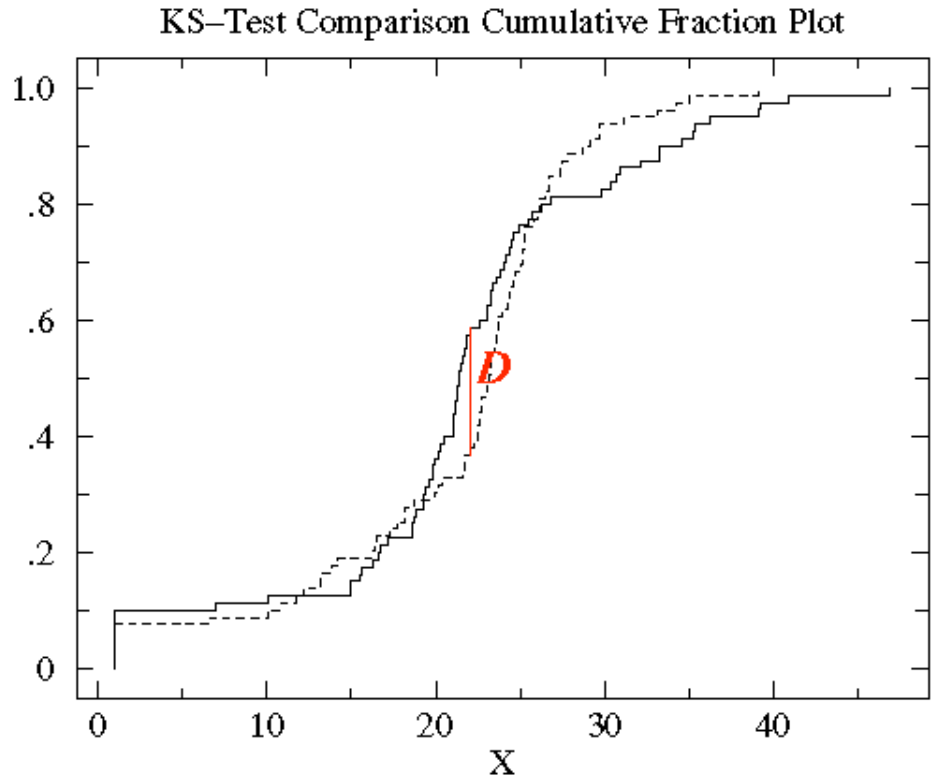
http://www.physics.csbsju.edu/stats/KS-test.html

The Kolmogorov-Smirnov GOF Test

- p-value is given by this pair of equations

$$z = D\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

$$p(z) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2z^2}$$



You can also compute the p-value by running pseudoexperiments and finding the distribution of the KS distance. Distributions are usually binned though – analytic formula no longer applies. Run pseudoexperiments instead.

See ROOT's TH1::KolmogorovTest() which computes both D and p. It is asymmetric in its treatment of histograms.

See also F. James, Statistical Methods in Elementary Particle Physics, 2nd Ed.

Including Correlated Uncertainties in χ^2

Example with

- Two measurements $a_1 \pm u_1 \pm c_1$ and $a_2 \pm u_2 \pm c_2$ of one parameter x
- Uncorrelated errors u_1 and u_2
- Correlated errors c_1 and c_2 (same source)

$$\chi^2(x) = \sum_{i,j=1,2} (x - a_i) C_{ij}^{-1} (x - a_j)$$

$$C = \begin{pmatrix} u_1^2 + c_1^2 & c_1 c_2 \\ c_1 c_2 & u_2^2 + c_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

If there are several sources of correlated error c_i^p then the off-diagonal terms become $\sum_p c_1^p c_2^p$

Combining Precision Measurements with **BLUE**

$$\chi^2(x) = \sum_{i,j=1,2} (x - a_i) C_{ij}^{-1} (x - a_j)$$

Procedure: Find the value of *x* which minimizes χ^2

This is a maximum likelihood fit with symmetric, Gaussian uncertainties.

Equivalent to a weighted average:

$$x_{best} = \sum_i w_i a_i \quad \text{with} \quad \sum_i w_i = 1$$

1 standard-deviation error from $\chi^2(x_{best} \pm \sigma_0) - \chi^2(x_{best}) = 1$

Can be extended to many measurements of the same parameter x.

More General Likelihood Fits

$$L = P\left(\text{data}|\vec{\theta},\vec{\nu}\right)$$

$\vec{\theta}$ "Parameters of Interest" oscillation parameters, cross-section, b.r.

$\vec{\nu}$ "Nuisance Parameters" Exposure, acceptance, detector resolution.

Strategy -- find the values of $\vec{\theta}$ and $\vec{\nu}$ which maximize *L*

Uncertainty on parameters: Find the contours in $\vec{\theta}$ such that

$\ln(L) = \ln(L_{max}) - s^2/2$,

to quote *s*-standard-deviantion intervals. Maximize L over $\vec{\nu}$ separately for each value of $\vec{\theta}$. Buzzword: "*Profiling*"

Be Sure to Normalize L

$$L = P\left(\text{data}|\vec{\theta},\vec{\nu}\right)$$

$\vec{\theta}$ "Parameters of Interest" oscillation parameters, cross-section, b.r.

$\vec{\nu}$ "Nuisance Parameters" Exposure, acceptance, detector resolution.

$$\sum_{\text{Possible\_data}} L\left(\text{data}|\vec{\theta},\vec{\nu}\right)=1$$

for all values of $\vec{\theta}$ and $\vec{\nu}$

Or your answers will be wrong.

More General Likelihood Fits

Advantages:

- "Approximately unbiased"
- Usually close to optimal
- Invariant under transformation of parameters. Fit for a mass or mass^2 doesn't matter.

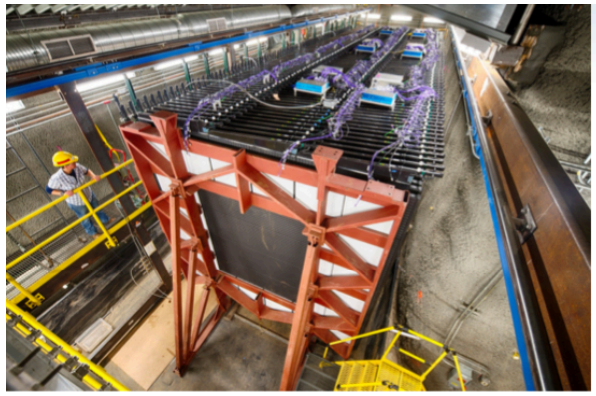
Unbinned likelihood fits are quite popular. Just need $L = P(\text{data}|\vec{\theta}, \vec{\nu})$

Warnings:

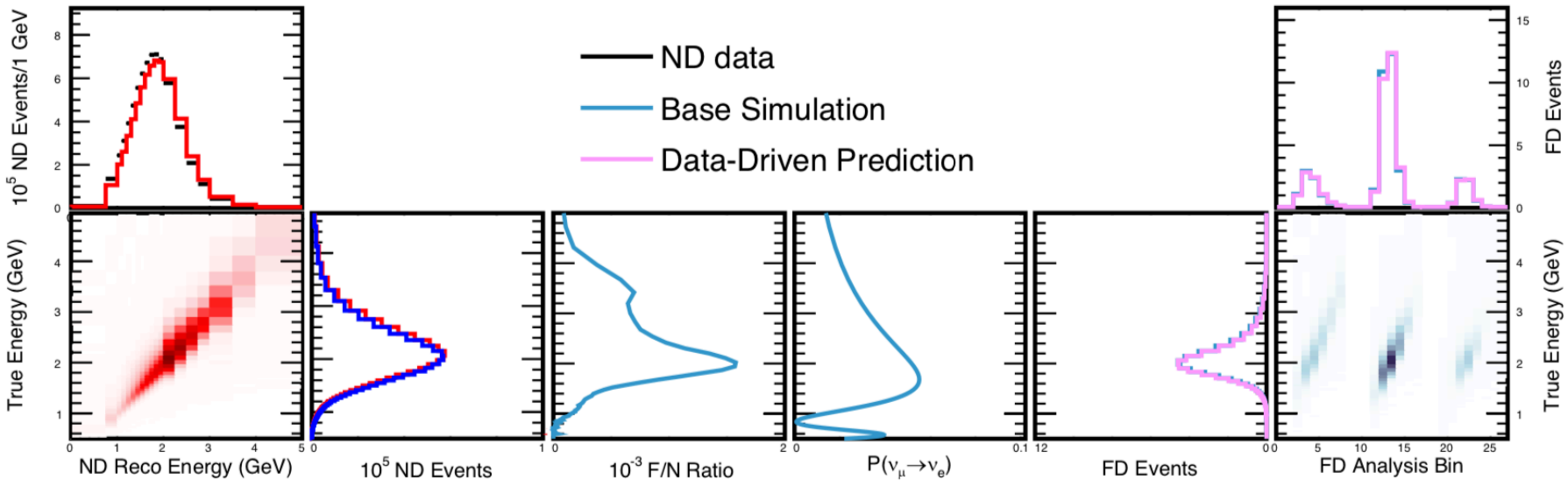
- Need to estimate what the bias is, if any.
- Monte Carlo Pseudoexperiment approach: generate lots of random fake data samples with known true values of the parameters sought, fit them, and see if the averages differ from the inputs.
- More subtle -- the uncertainties could be biased.

-- run pseudoexperiments and histogram the "pulls" (fit-input)/error -- should get a Gaussian centered on zero with unit width, or there's bias.

- Handling of systematic uncertainties on nuisance parameters by maximization can give misleadingly small uncertainties -- need to study L for other values than just the maximum (L can be bimodal)



Extrapolating from Near to Far

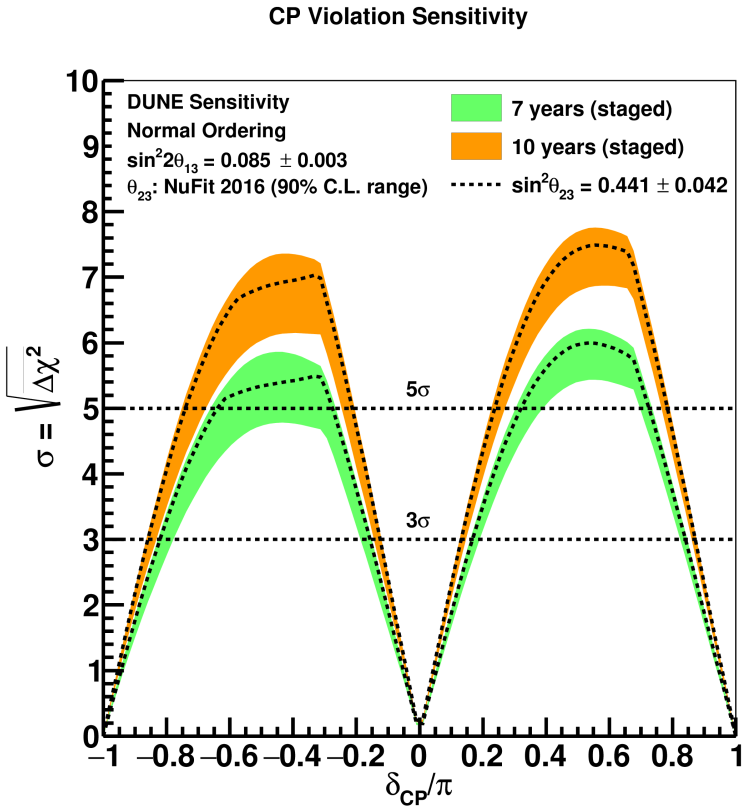


- Use the ND ν_μ sample to predict the FD ν_μ sample.
- Use the ND ν_μ sample to predict the FD ν_e signal.

Alex Himmel at Phystat-Nu 2019
https://indico.cern.ch/event/735431/contributions/3137791/attachments/1783219/2902091/2019-01-23-lbl-stats.pdf

Sensitivity Projections

- Need to evaluate many experiment design choices quickly
- Usually involve fits to the Asimov dataset
  - Asimov data = median expected outcome, no Poisson fluctuation. From Isaac Asimov's short story *Franchise* in which one "typical" voter cast a ballot for everyone in the galaxy.



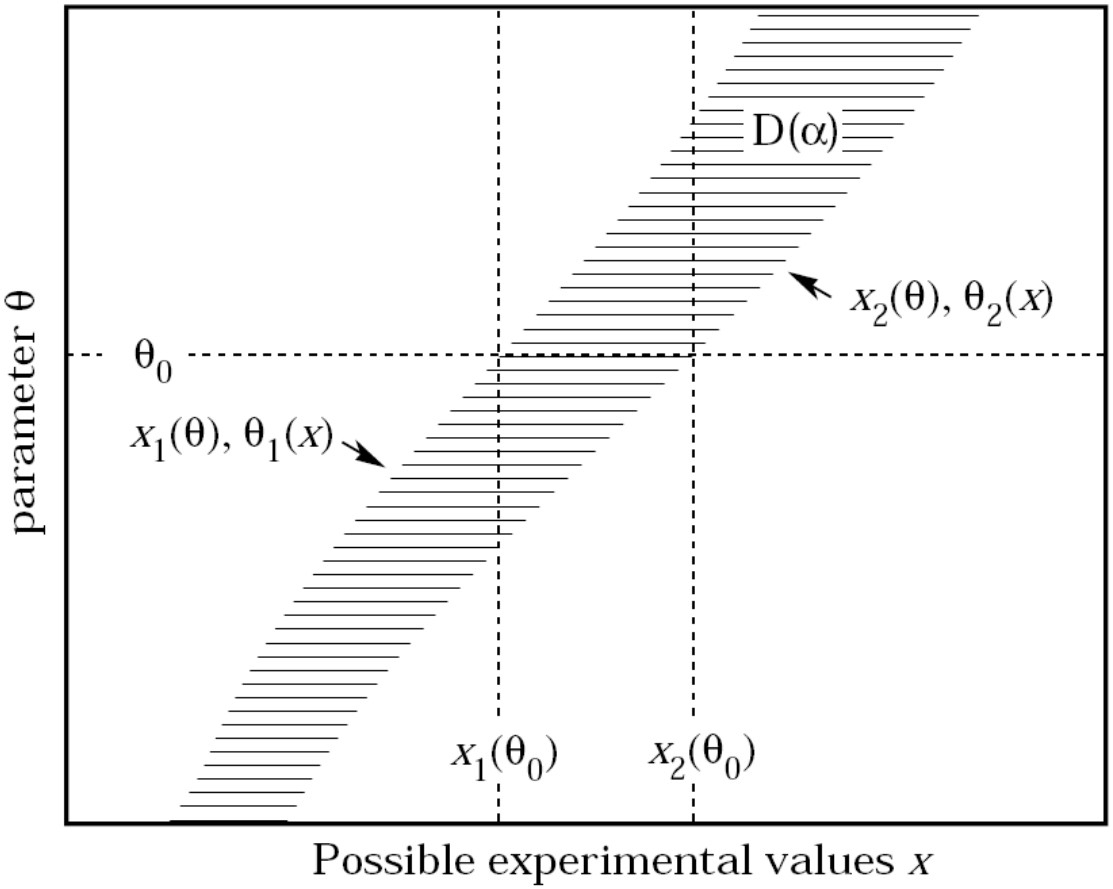
DUNE CDR sensitivity – TDR is similar

Units of sensitivity are usually $\sigma = \sqrt{\Delta\chi^2}$ using Gaussian approximations for the distribution of possible outcomes (68% are within $\pm 1\sigma$ of central)

The "Neyman Construction" of Frequentist Confidence Intervals

Essentially a
"calibration curve"

- Pick an observable *x* somehow related to the parameter θ you'd like to measure
- Figure out what distribution of observed x would be for each value of θ possible.
- Draw bands containing 68% (or 95% or whatever) of the outcomes
- Invert the relationship using the prescription on this page.



A pathology: can get an empty interval. But the error rate has to be the specified one. Imagine publishing that all branching ratios between 0 and 1 are excluded at 95% CL.

Proper Coverage is Guaranteed!

A Special Case of Frequentist Confidence Intervals: Feldman-Cousins

Each horizontal band contains 68% of the expected outcomes (for 68% CL intervals)

But Neyman doesn't prescribe which 68% of the outcomes you need to take!

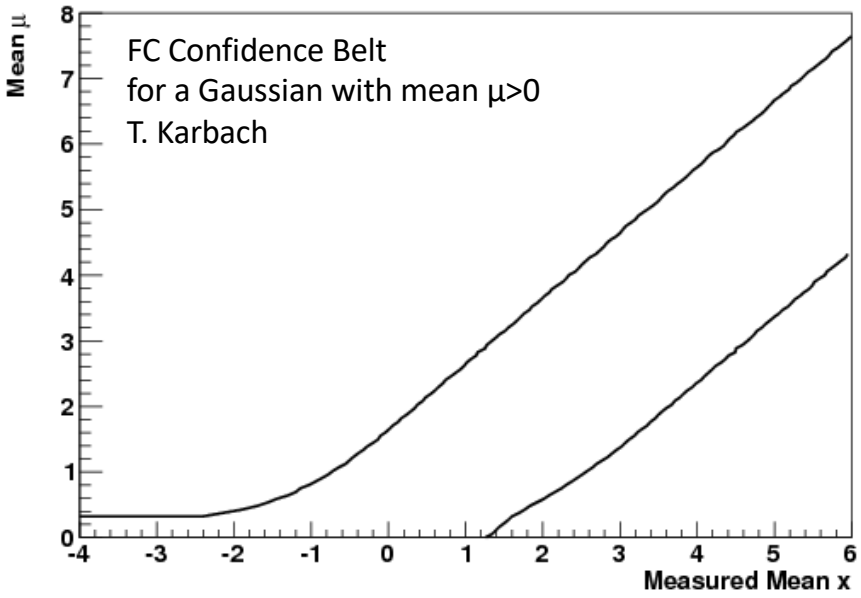
Take lowest x values: get lower limits.
Take highest x values: get upper limits.

Feldman & Cousins's Recommendation:
Sort outcomes by the likelihood ratio.

$R = L(x|\theta)/L(x|\theta_{best})$

For all x, R=1 for some θ. And R≤1 always.

Picks 1-sided or 2-sided intervals --
no flip-flopping between limits and 2-sided intervals.

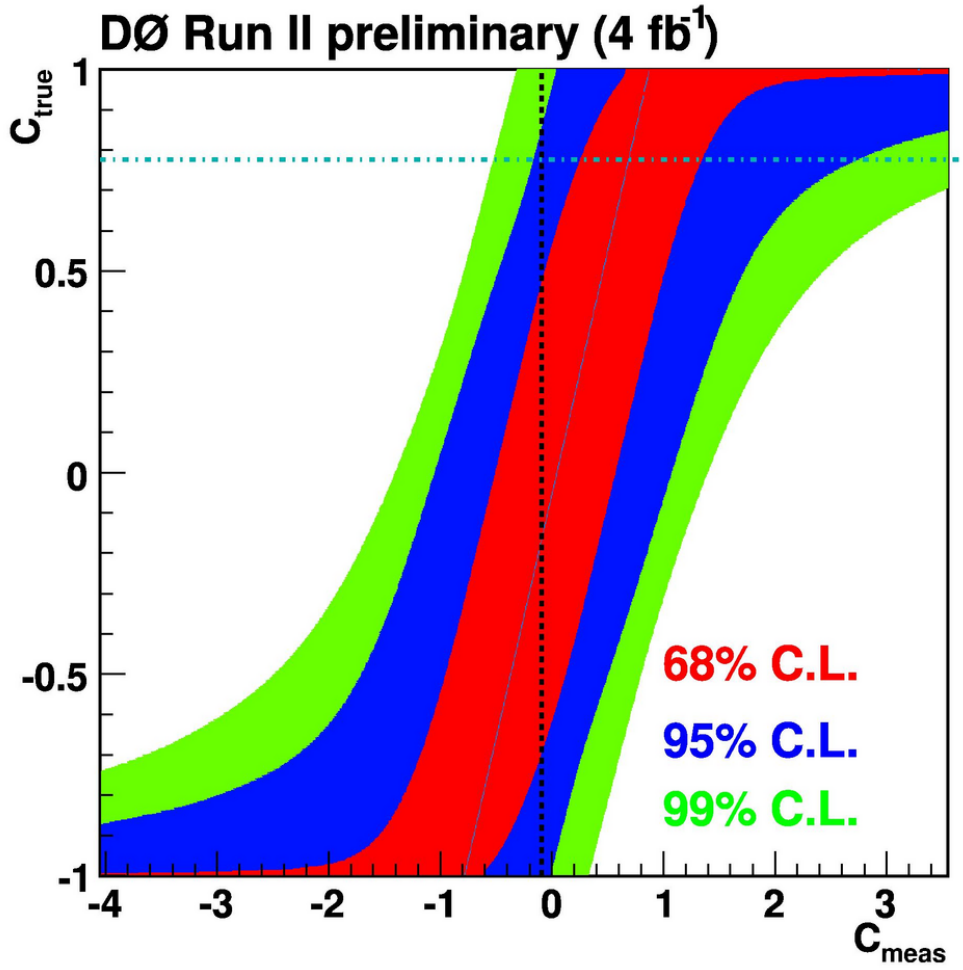


G. Feldman and R. Cousins,
"A Unified approach to the classical statistical analysis of small signals"
Phys.Rev.D57:3873-3889,1998.
arXiv:physics/9711021

Also explained in Kendall & Stuart in the 1940's.

No empty intervals!

A 1D FC Belt Example with Bounds on Both Sides



A top-quark polarization correlation measurement.
Branching ratios, $\sin^2\theta$, and other variables have similar constraints

Two-Dimensional FC Constructions

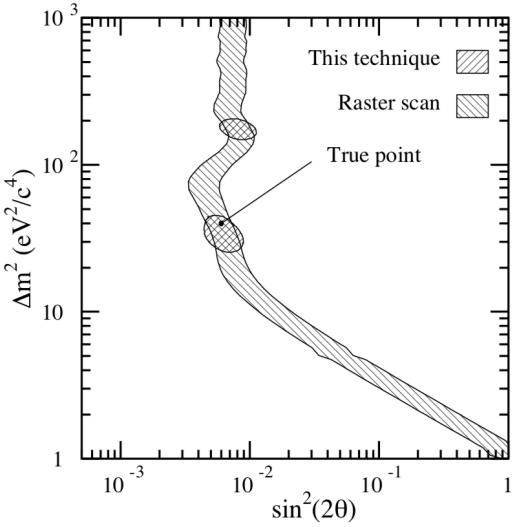
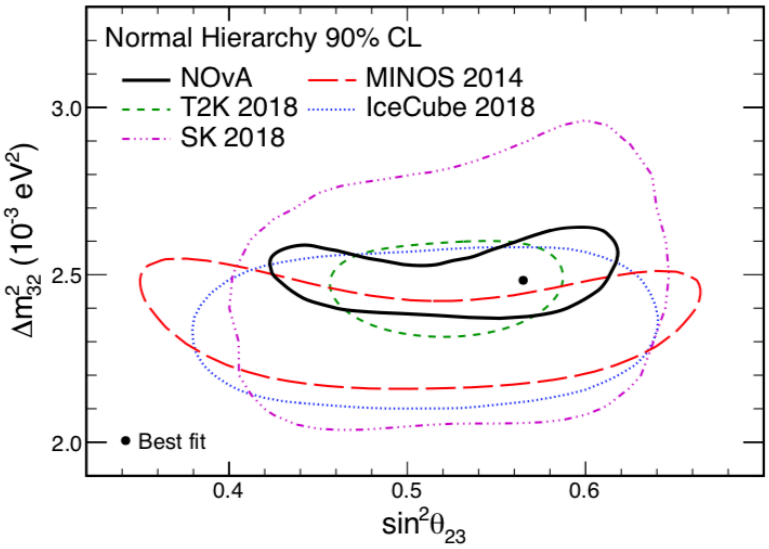
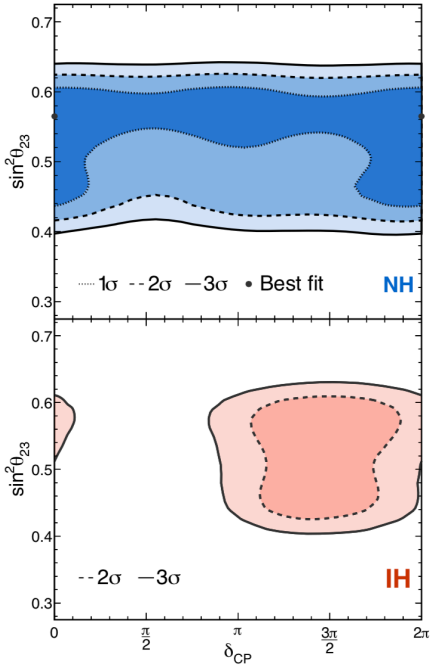


FIG. 12. Calculation of the confidence regions for an example of the toy model in which $\Delta m^2 = 40$ (eV/c^2)^2 and $\sin^2(2\theta) = 0.006$, as evaluated by the proposed technique and the Raster Scan.





F&C, Phys.Rev. **D57** (1998) 3873-3889.

NOvA Collab. https://arxiv.org/abs/1906.04907

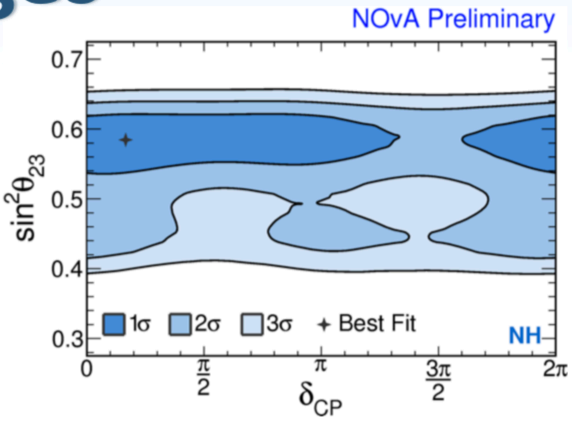
See A. Sousa CHEP 2018 for modern computational details
https://indico.cern.ch/event/587955/contributions/2938131/attachments/1685595/2710354/Sousa_SciDac4_NOvA_HPC_CHEP2018.pdf

Three-dimensional space of parameters of interest is much harder.

A. Sousa, CHEP 2018

Computational Challenges

- Need $\Delta\chi^2$ distributions for each point in sampled parameter space. Minimal set requires:
  - **1,200 points** total for ten 1D Profiles, 60 points each for 2 octants of θ_{23}
  - **471 points** total for four 2D Contours, after optimizing for regions of interest in parameter space



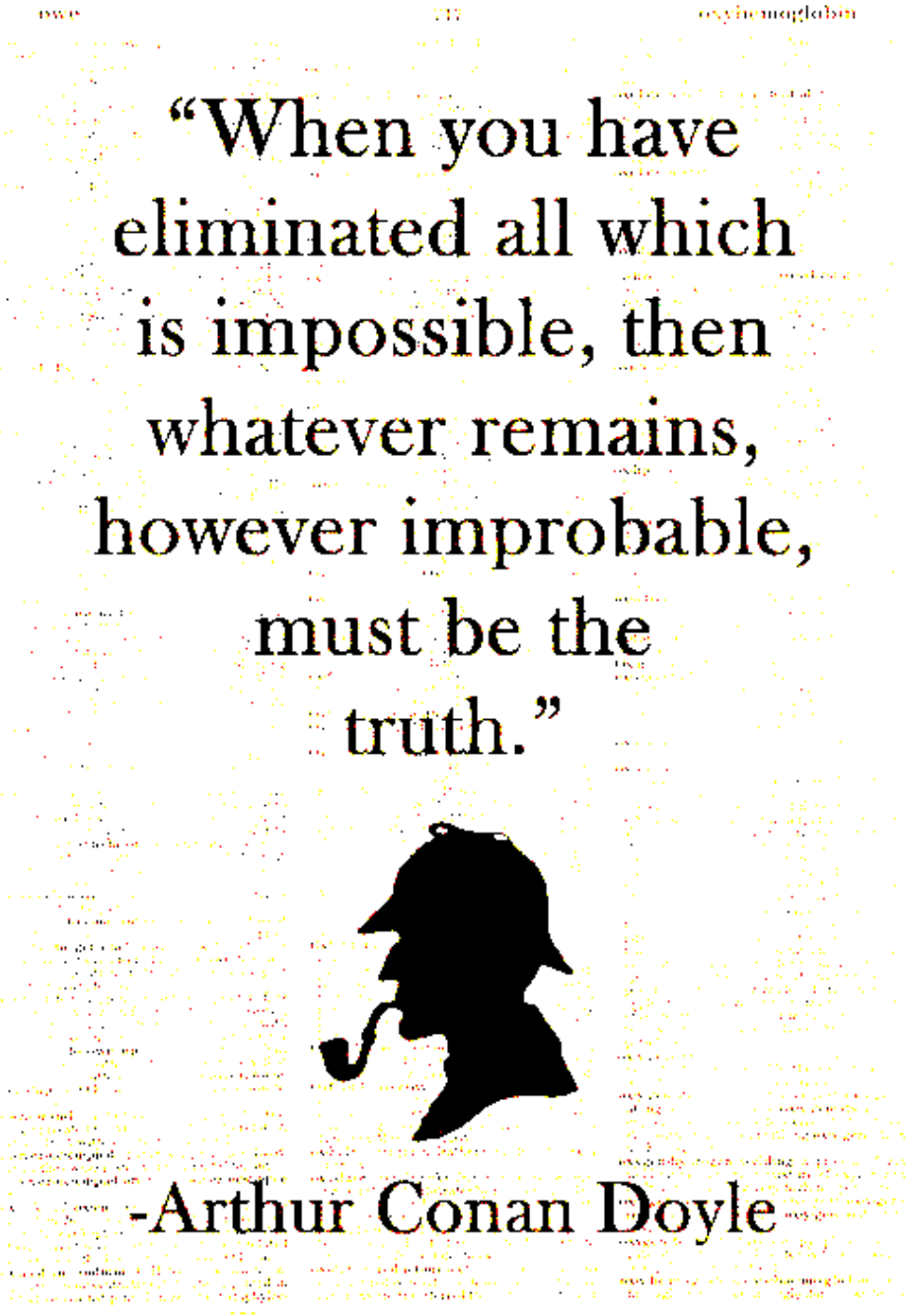
- For each point, need at least **4,000 pseudoexperiments** to generate accurate empirical distribution
  - Depends on how large the critical value corresponding to desired confidence level is (up to 3σ for NOvA)
  - Depends on number of systematic uncertainties included
  - Computing $\Delta\chi^2$ for each pseudoexperiment takes between $\mathcal{O}$(10 min) to $\mathcal{O}$(1 hour) for fits with high-level of degeneracy

Required No. of Points	Minimum No. of Pseudoexperiment
1,671	6,684,000

- Previously done with FermiGrid + OSG resources - results obtained in **~4 weeks**
  - FermiGrid provides a total of ~200M CPU-hours/year (50% CMS, 7% NOvA). Use of OSG opportunistic resources by NOvA doubles FermiGrid allocation (NOvA total of **~30M CPU-hours/year**)

- 2018 analysis includes new antineutrino dataset + longer list of systematics ⇒ FermiGrid + OSG not enough to get to results in timely fashion

2018 analysis: 10 histograms to fit, 2 runs of ~20M CPU hours each at NERSC



Two complaints

1) Elimination may be erroneous (Type-II error) which can create a false discovery (Type-I error)

2) How do we know all possibilities have been considered?

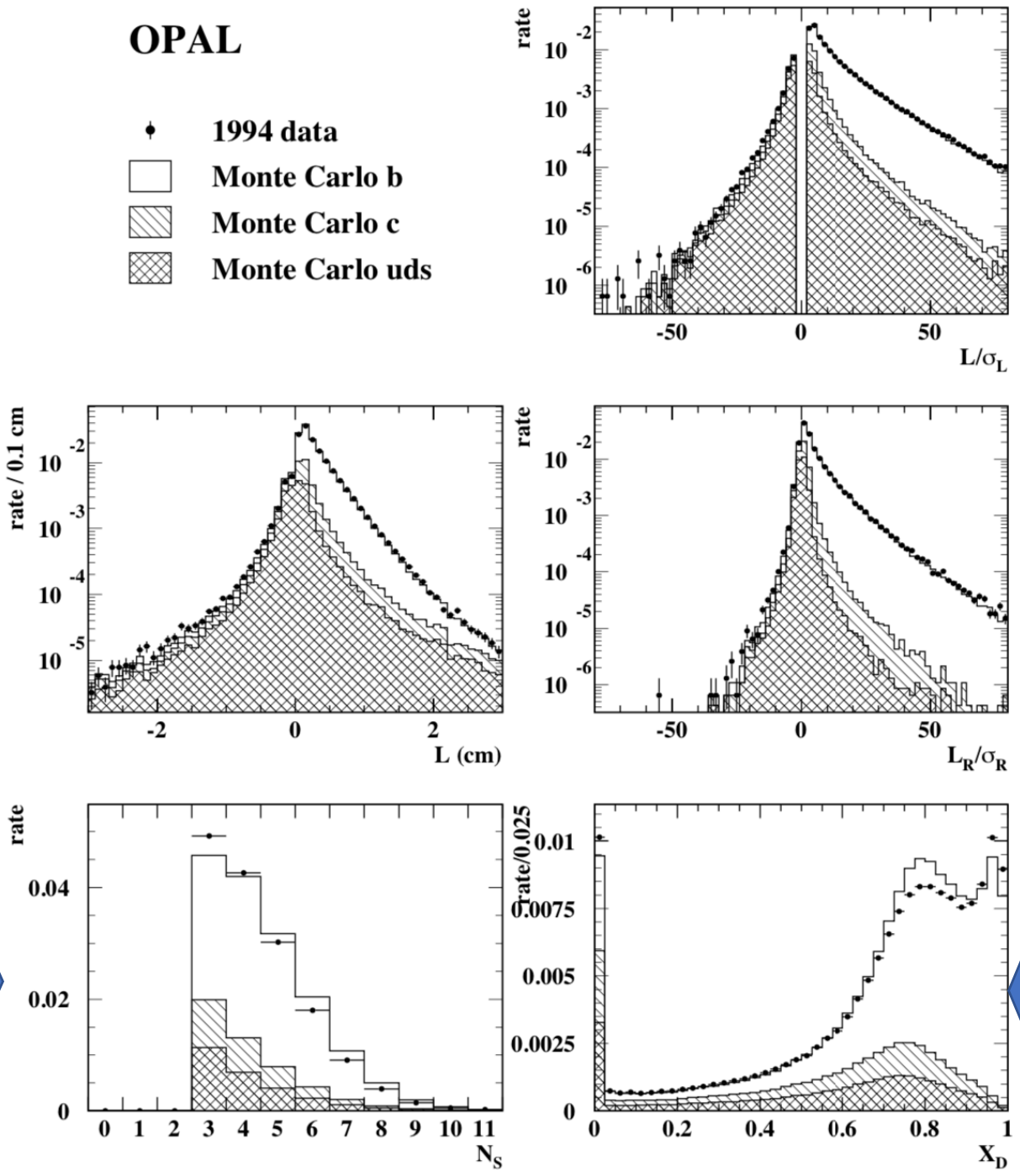
Nature may be *outside* the chosen model space entirely.

If we know *a priori* somehow that the truth is in our model space then this is okay.

Some model spaces, such as properties of exotic particles, may consist entirely of untrue models and we would like to rule out the entire space if we can. Otherwise we may just "discover" the part of parameter space we cannot test.

Putting the null hypothesis in the model space does not ensure completeness.

The incompleteness in the model space is usually in the nuisance parameter portion.



A Measurement of R(b) using a double-tagging method
Eur.Phys.J.C8:217-239,1999

We had two distributions we couldn't model perfectly.

We thought for *sure* it was our imperfect detector modeling.

Tweaking parameters one way fixed one distribution and made the other one worse, and vice versa.

We ended up taking a systematic uncertainty on detector modeling due to this.

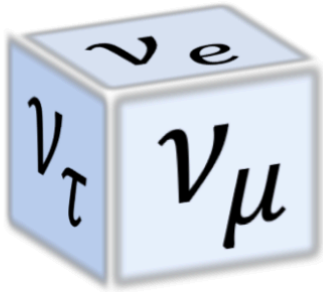
We later discovered that the decay multiplicity of B hadrons measured by ARGUS and put in Pythia was ~1 track off the world average.

It took LEP-2 data with WW decays (no B's) to discover this.

K. McFarland, Phystat-nu 2019 summary talk

https://indico.cern.ch/event/735431/contributions/3137831/attachments/1785482/2906728/Neutrino_Summary_PHYSTATnu_2019_final.pdf

Canonical cautionary tale: MiniBooNE

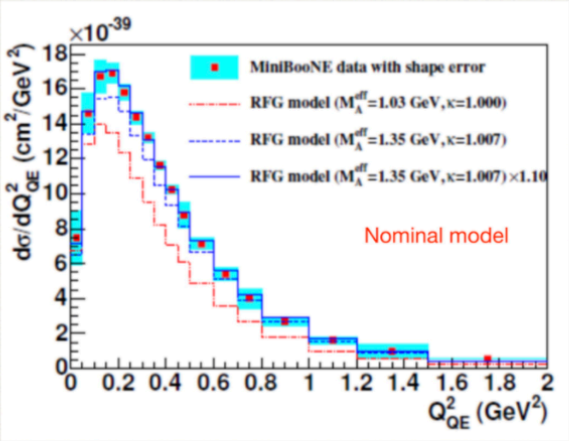


Experiments confronting data/MC discrepancies

Experiments need a model that describes their data

However, often, **data/MC agreements are handled in a non-satisfactory way**

- Overemphasising own data - breaking consistency with other neutrino data
- Largely ignoring complementary constraints from charged-lepton and hadron scattering



A typical (and conveniently old and non-controversial) example comes from the MiniBooNE experiment:

Tweaking axial form factor parameter

- Axial mass 1.03 -> 1.35 GeV
- Not consistent with bubble chamber results

Tweaking Pauli blocking

- Not consistent with textbook physics

Good description of own data.
But wrong physics!

22

25 January 2019 K. McFarland, Neutrino Summary

- MiniBooNE observed a discrepancy in its "CCQE" (charged current quasielastic) events vs Q^2.
  - Attributed to axial form factor and Pauli blocking, just an event distortion in Q^2.
  - We understand now this is, at least in part, due to multinucleon production with a different energy-momentum transfer relationship.
- Burying the difference in form factor means misreconstructing E_ν.

The neutrino experiment experience [combining data]
Speaker: Prof. Constantinos Andreopoulos (Liverpool, STFC/RAL) 30

Systematic Uncertainties

Roger Barlow's fine advice: https://arxiv.org/abs/hep-ex/0207026

Penalty for diligence:

- Someone who has one watch always knows what time it is. Someone with two watches is never quite sure.

Paraphrasing Kyle Cranmer and Costas Andreopoulos, systematic uncertainties come in three categories:

- The GOOD
  - Those that are constrained with auxiliary measurements
- The BAD
  - (educated) Guesses
- The UGLY
  - Forgotten, omitted, dismissed, or unknown

Sometimes a GOOD systematic can have BAD or UGLY components, e.g. extrapolating from a control sample into a signal sample requires some knowledge or guesswork.

Reasons for Another Kind of Probability

- So far, we've been (mostly) using the notion that probability is the limit of a fraction of trials that pass a certain criterion to total trials.

- Systematic uncertainties involve many harder issues. Experimentalists spend much of their time evaluating and reducing the effects of systematic uncertainty.

- We also want more from our interpretations -- we want to be able to make decisions about what to do next.
  - Which HEP project to fund next?
  - Which theories to work on?
  - Which analysis topics within an experiment are likely to be fruitful?

  These are all different kinds of bets that we are forced to make as scientists. They are fraught with uncertainty, subjectivity, and prejudice.

  Non-scientists confront uncertainty and the need to make decisions too!

Bayes' Theorem

Law of Joint Probability:

$$p(A \text{ and } B) = p(A|B)p(B) = p(B|A)p(A)$$

Events A and B interpreted to mean "data" and "hypothesis"

$$p(\theta|\text{data}) = \frac{L(\text{data}|\theta)\pi(\theta)}{\int L(\text{data}|\theta')\pi(\theta')d\theta'}$$

θ = set of model parameters

A frequentist would say: Models have no "probability". One model's true, others are false. We just can't tell which ones (maybe the space of considered models does not contain a true one).

Better language: $p(\theta|\text{data})$ describes our belief in the different models parameterized by θ

Bayes' Theorem

$p(\theta|\text{data})$ is called the "posterior probability" of the model parameters

$\pi(\theta)$ is called the "prior density" of the model parameters

The Bayesian approach tells us how our existing knowledge before we do the experiment is "updated" by having run the experiment.

This is a natural way to aggregate knowledge -- each experiment updates what we know from prior experiments (or subjective prejudice or some things which are obviously true, like physical region bounds).

Be sure not to aggregate the same information multiple times! (groupthink)

We make decisions and bets based on all of our knowledge and prejudices

"Every animal, even a frequentist statistician, is an informal Bayesian." See R. Cousins, "Why Isn't Every Physicist a Bayesian", Am. J. P., Volume 63, Issue 5, pp. 398-410

How I remember Bayes's Theorem

$$p(\text{hypothesis}|\text{data}) = \frac{p(\text{data}|\text{hypothesis}) \times p(\text{hypothesis})}{p(\text{data})}$$

Posterior "PDF"
("Credibility")

"Likelihood Function"
("Bayesian Update")

"Prior belief
distribution"

Normalize this so that

$$\int p(\text{hypothesis}|\text{data}) d(\text{hypothesis}) = 1$$

for the observed data

Bayesian Upper Limits

Including uncertainties on nuisance parameters $\boldsymbol{\nu}$

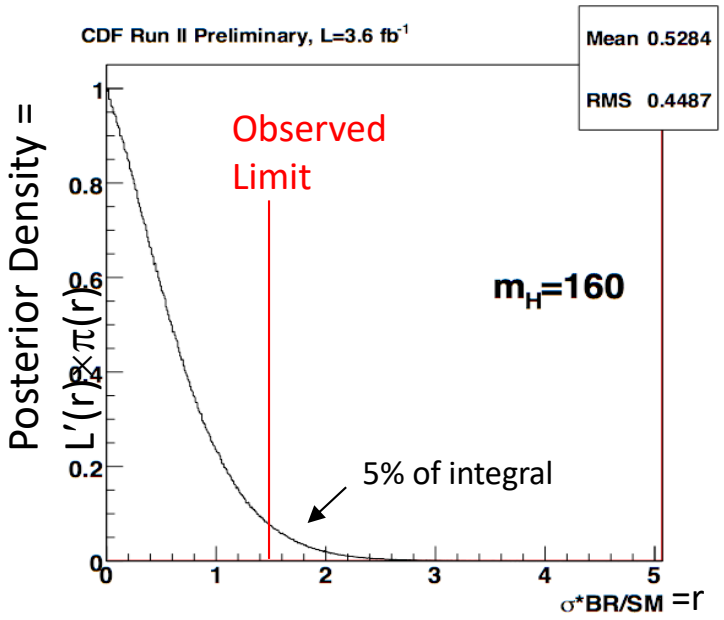
$$L'(\text{data}|r) = \int L(\text{data}|r,\nu)\pi(\nu)d\nu$$

Typically $\pi(r)$ is constant
Other options possible.
See the PDG stats reveiw
Sensitivity to priors a concern.

where $\pi(\boldsymbol{\nu})$ encodes our prior belief in the values of the uncertain parameters. Usually Gaussian centered on the best estimate and with a width given by the systematic. The integral is high-dimensional. Markov Chain MC integration is quite useful! The Metropolis-Hastings Algorithm and variants are very useful.

Limits:

$$0.95 = \frac{\int_0^{r_{\lim}} L'(\text{data}|r)\pi(r)dr}{\int_0^{\infty} L'(\text{data}|r)\pi(r)dr}$$



Bayesian Cross Section (or rate) Extraction

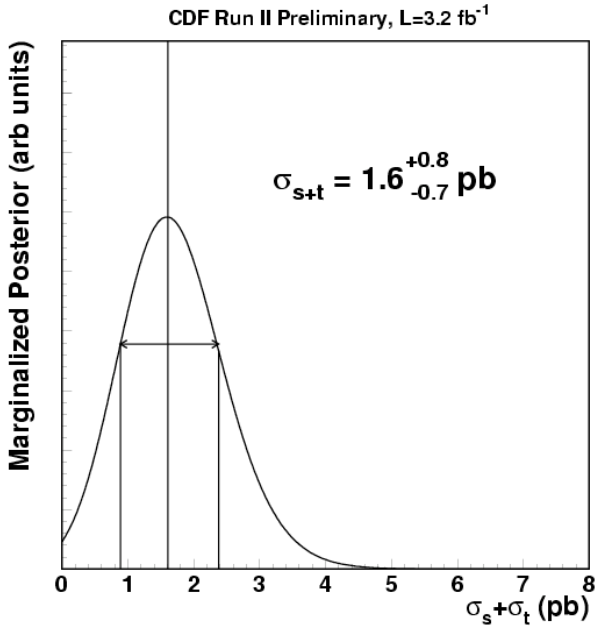
Same handling of nuisance parameters as for limits

$$L'(\text{data}|r) = \int L(\text{data}|r,\nu)\pi(\nu)d\nu$$

$$0.68 = \frac{\int_{r_{\text{low}}}^{r_{\text{high}}} L'(\text{data}|r)\pi(r)dr}{\int_0^{\infty} L'(\text{data}|r)\pi(r)dr}$$

The measured cross section and its uncertainty

$$r = r_{\text{max}}{}^{+(r_{high}-r_{\text{max}})}_{-(r_{\text{max}}-r_{low})}$$



Usually: shortest interval containing 68% of the posterior (other choices possible). Use the word "credibility" in place of "confidence"

If the 68% CL interval does not contain zero, then the posterior at the top and bottom are equal in magnitude.
The interval can also break up into smaller pieces!

Also easily generalizable to many parameters of interest.

Systematic Uncertainties

Encoded as priors on the nuisance parameters $\pi(\boldsymbol{\nu})$.

Can be quite contentious -- injection of theory uncertainties and results from other experiments -- how much do we trust them?

Do not inject the same information twice.

Some uncertainties have statistical interpretations -- can be included in L as additional data. Others are purely about belief. Theory errors often do not have statistical interpretations.

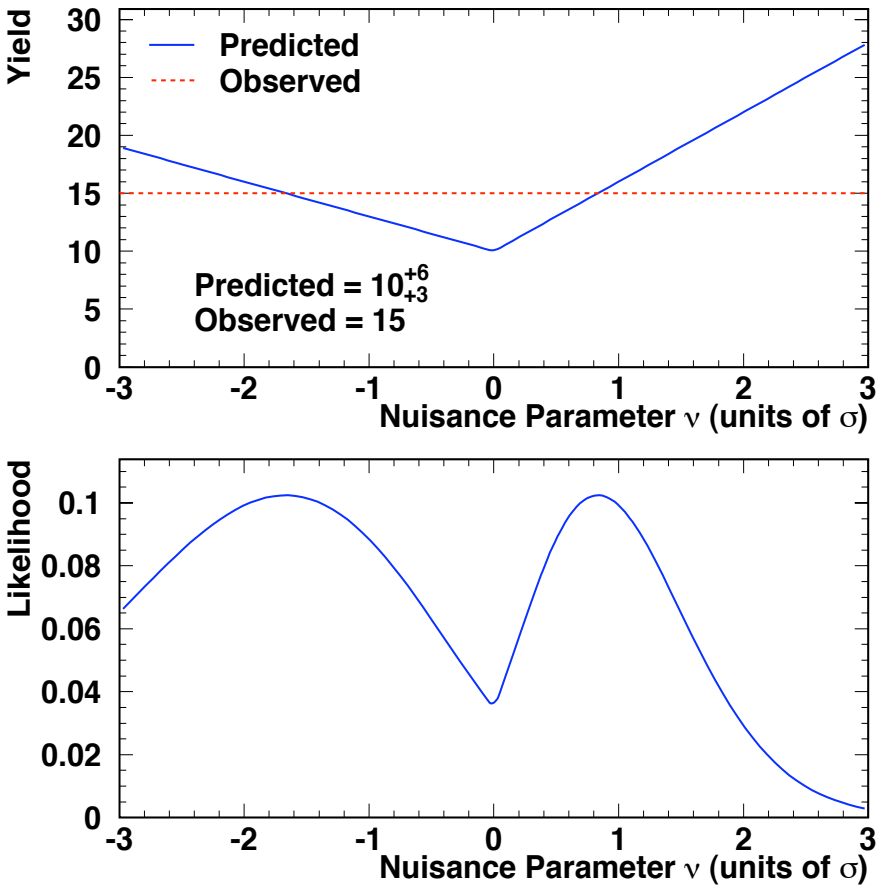
Coping with Systematic Uncertainty

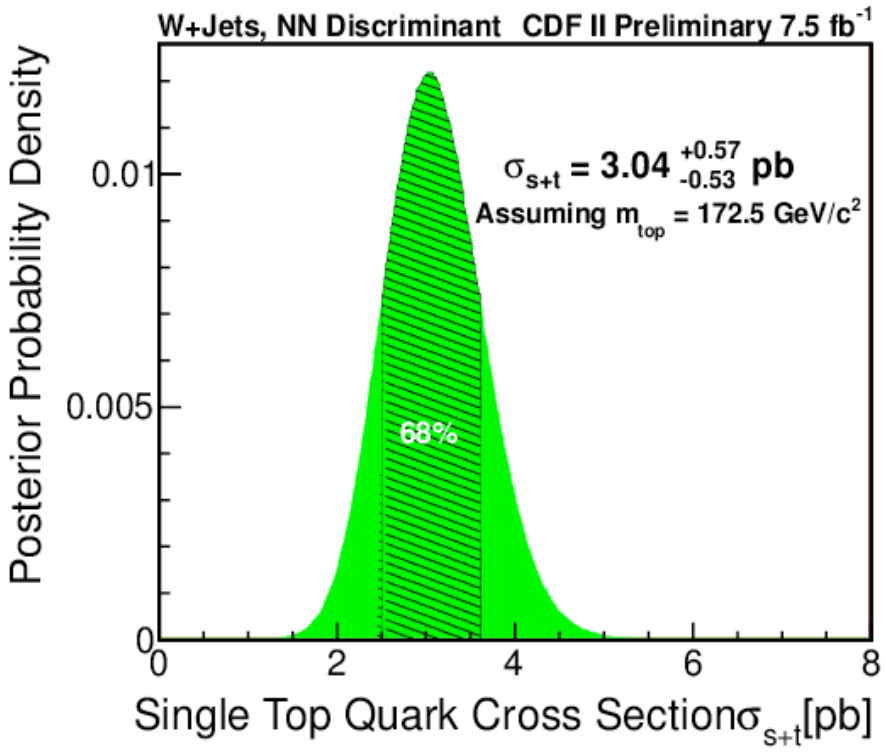
- "Profile:"
  - Maximize *L* over possible values of nuisance parameters include prior belief densities as part of the χ^2 function (usually Gaussian constraints)

- "Marginalize:"
  - Integrate L over possible values of nuisance parameters (weighted by their prior belief functions -- Gaussian, gamma, others...)
  - Consistent Bayesian interpretation of uncertainty on nuisance parameters

- Aside: MC "statistical" uncertainties are systematic uncertainties

Parameter Estimation – Marginalize or Profile?





If Pred = 10^{-6}_{-3}, and obs=15, then the likelihood would have one maximum, but it would have a corner. MINUIT may quote inappropriate uncertainties as the second derivative isn't well defined.

The corner can be smoothed out – See
R. Barlow, http://arxiv.org/abs/physics/0406120,
http://arxiv.org/abs/physics/0401042
http://arxiv.org/abs/physics/0306138

But I know of no way
to get rid of the double-peak
Nor should there be a way --
it can be a real effect. See the LEP2 TGC measurements

Even Bayesians have to be a little Frequentist

- A hard-core Bayesian would say that the results of an experiment should depend only on the data that are observed, and not on other possible data that were not observed.

- But we still want the sensitivity estimated! An experiment can get a strong upper limit not because it was well designed, but because it was lucky.

How to optimize an analysis before data are observed?

So -- run Monte Carlo simulated experiments and compute a Frequentist distribution of possible limits. Take the median-- metric independent and less pulled by tails.

Cross-Checks in Data Subsamples

Do not improve discovery sensitivity (see Barlow's advice)

Give confidence that the model(s) are predicting the data adequately.

See an excess of events in a specific energy range? Look in all the corners of the detector!

See something weird in the data – does it persist in other kinematic bins, run periods, etc.

Be careful of trigger bias – you will only see the events you seek!

A warning – splitting data up into small pieces enhances the look-elsewhere effect.
You will see discrepancies just due to randomness if you look in enough places.

"All data are infinitely unlikely" -- R. McPherson

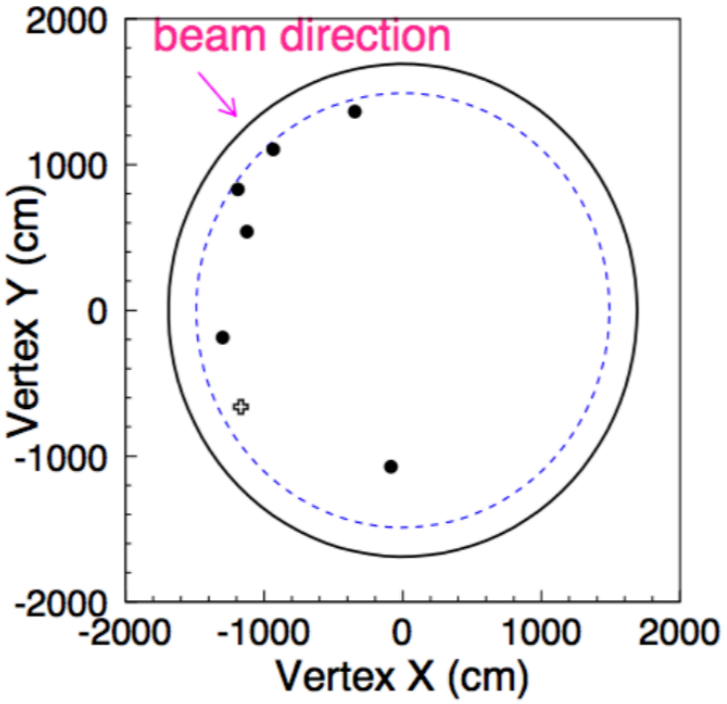
A hazard: post-hoc analysis placing cuts around special events. The probability of seeing one is smaller and smaller the tighter the cuts around the observed event get.

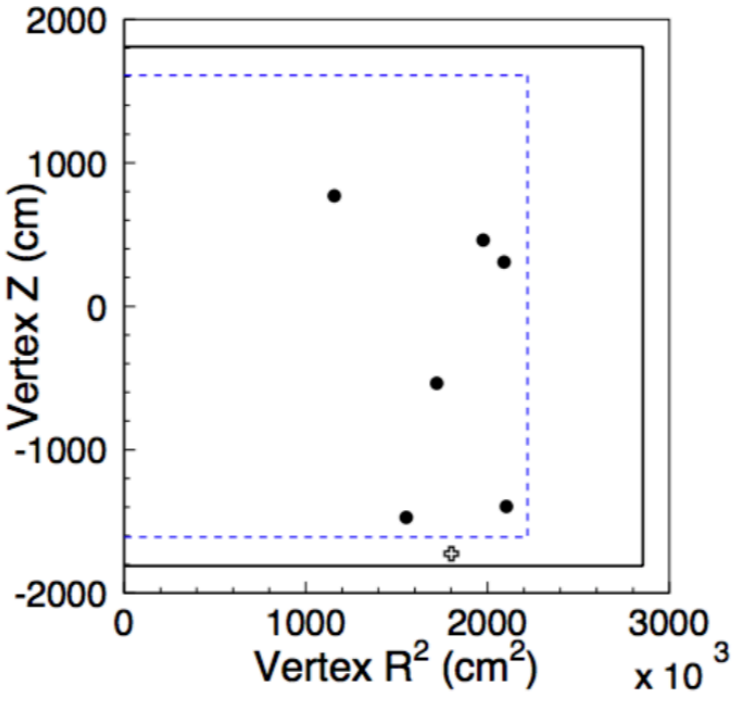
T. Kuffner: Using the data to select the model to test invalidates classical inference.

Real Life Examples

T2K 2011

Vertex distribution of ν_e candidate events



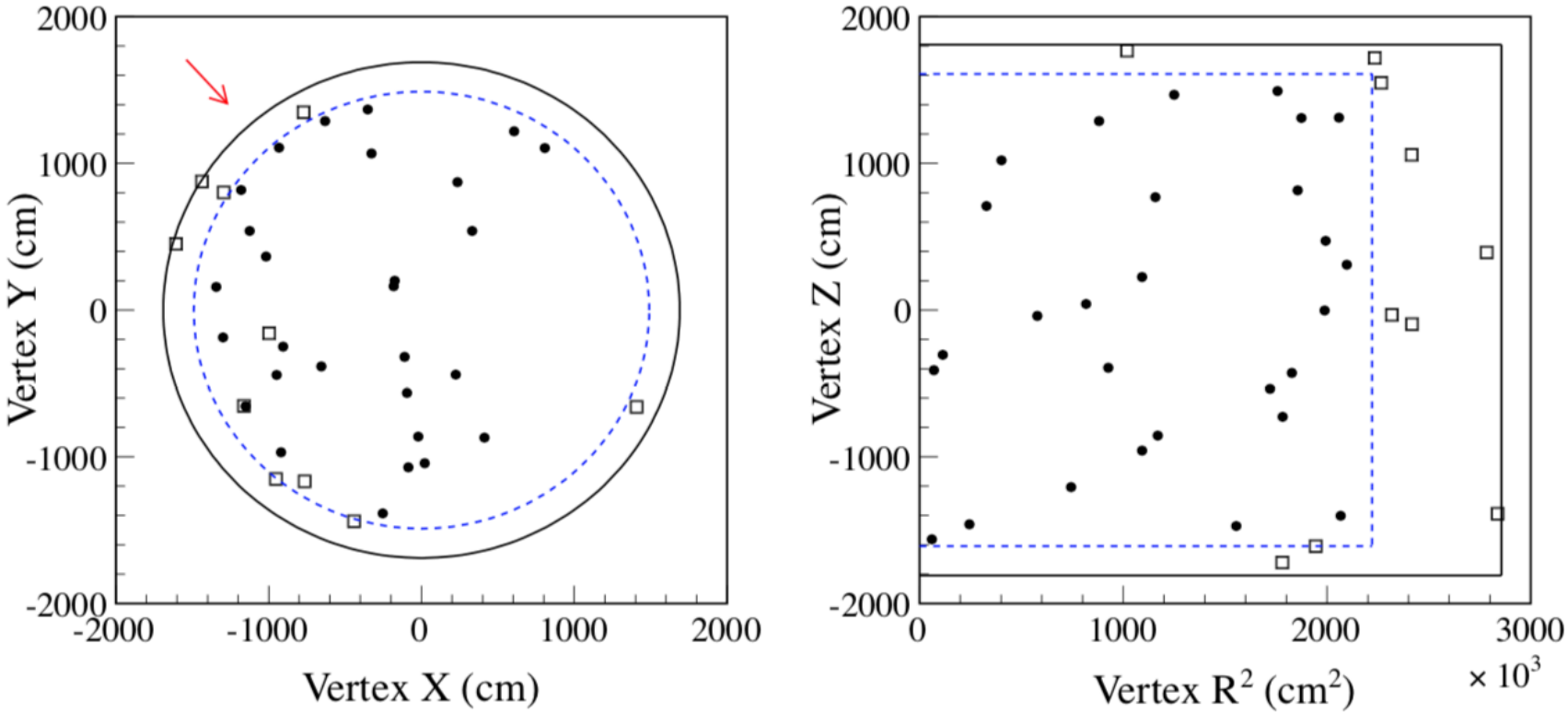


Event outside FV

These events are clustered at large R
→ Perform several checks. for example

* Check distribution of events outside FV → no indication of BG contamination
* Check distribution of OD events → no indication of BG contamination
* K.S. test on the R^2 distribution yields a p-value of 0.03

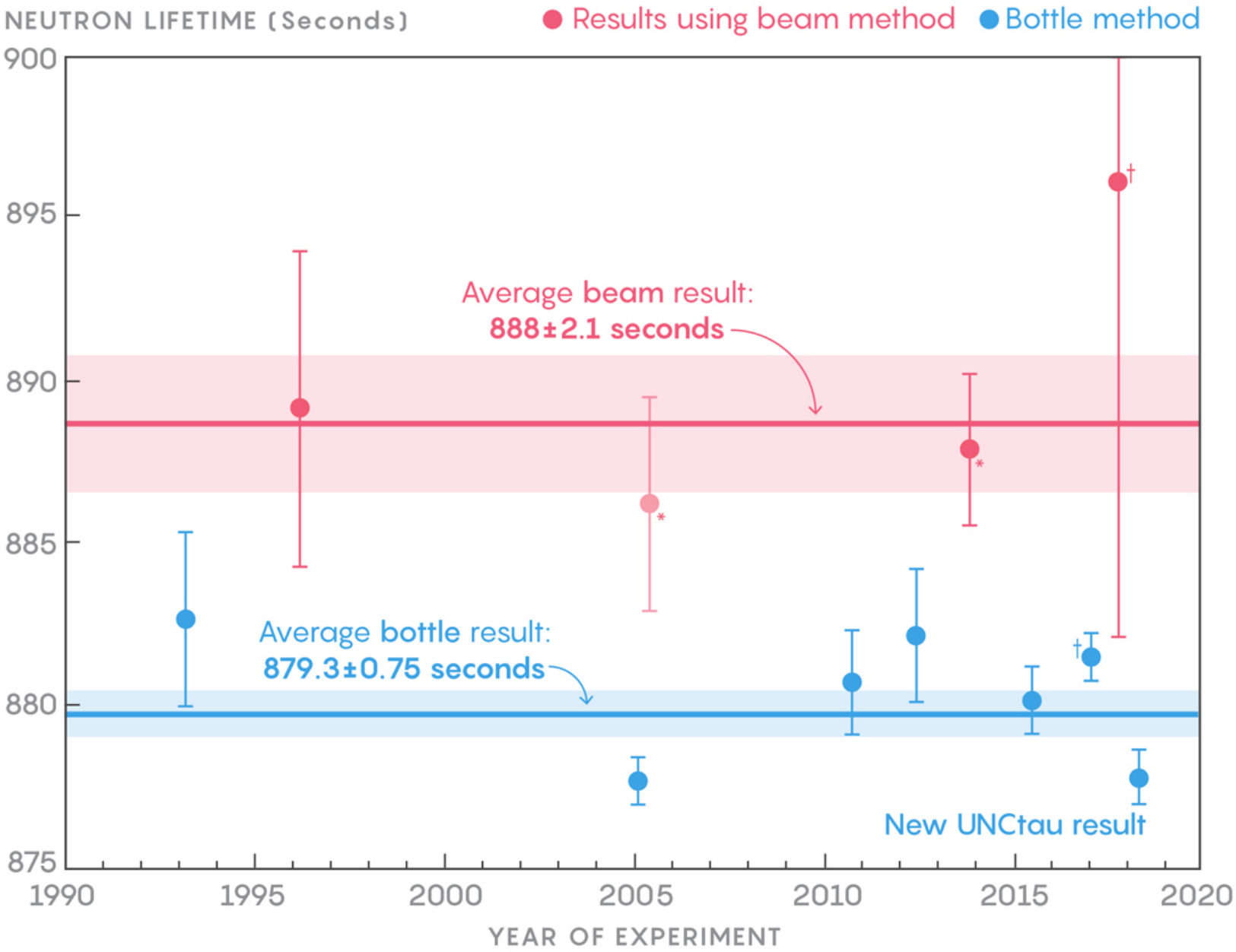
A Solution: Take More Data if You Can



Open square points fail fiducial volume cut.

"We observe no unexpected clustering and combined KS tests for uniformity in r^2 and z yields a p-value of 0.6"

T2K Collaboration, Phys.Rev. D91 (2015) no.7, 072010



*Nico result (2005) was superseded by an updated and improved result, Yue (2013);

†Preliminary results

https://www.quantamagazine.org/neutron-lifetime-puzzle-deepens-but-no-dark-matter-seen-20180213/

Hypothesis Testing

- Simplest case: Deciding between two hypotheses. Typically called the *null* hypothesis H_0 and the *test* hypothesis H_1
- Can't we be even simpler and just test one hypothesis H_0?
  - Data are random -- if we don't have another explanation of the data, we'd be forced to call it a random fluctuation. Is this enough?
  - H_0 may be broadly right but the predictions slightly flawed
  - Look at enough distributions and for sure you'll spot one that's mismodeled. A second hypothesis provides guidance of where to look.
- Popper: You can only prove models wrong, never prove one right.

All models are wrong; some are useful.

- Proving one hypothesis wrong doesn't mean the proposed alternative must be right.

Frequentist Hypothesis Testing: Test Statistics and p-values

Step 1: Devise a quantity that depends on the observed data that ranks outcomes as being more signal-like or more background-like.

Called a test statistic. Simplest case: Searching for a new particle by counting events passing a selection requirement.

Expect b events in H_0, $s+b$ in H_1.

The event count n_{obs} is a good test statistic.

Step 2: Predict the distributions of the test statistic separately assuming:

H_0 is true

H_1 is true

(Two distributions. More on this later)

Frequentist Hypothesis Testing: Test Statistics and p-values

Step 3: Run the experiment, get observed value of test statistic.

Step 4: Compute p-value

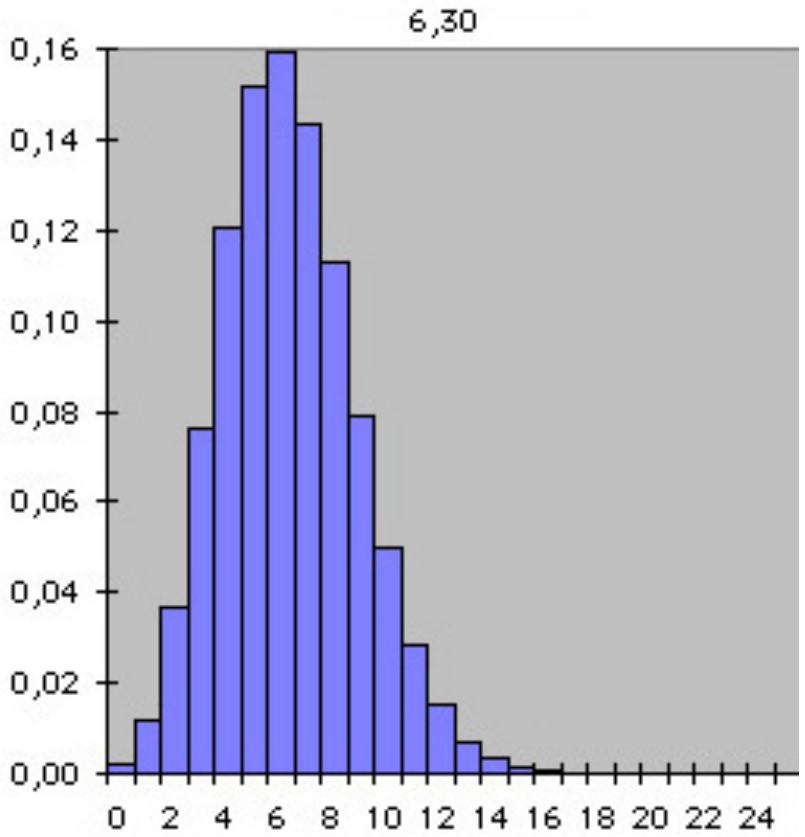
$p(n \geq n_{obs} | H_0)$

Example:

H_0: $b = \mu = 6$

$n_{obs} = 10$

p-value = 0.0839



But many often say that.

Especially the popular media!

A p-value is **not** the "probability H_0 is true"

So what *is* the p-Value?

A *p*-value is **not** the "probability H_0 is true" -- this isn't even a Frequentist thing to say anyway. If we have a large ensemble of repeated experiments, it is not true that H_0 is true in some fraction of them!

p-values are uniformly distributed assuming that the hypothesis they are testing is true (and outcomes are not too discretized).

Why not ask the question – what's the chance $N=N_{obs}$ (no inequality). Each outcome may be vanishingly improbable. What's the chance of getting exactly 10,000 events when a mean of 10,000 are expected? (it's small). How about 1 if 1 is expected?

If $p < p_{crit}$ then we can make a statement. Say $p_{crit}=0.05$. If we find $p < p_{crit}$, then we can exclude the hypothesis under test at the 95% CL.

What does the 95% CL mean? It's a statement of the *error rate*.

In no more than 5% of repeated experiments, a false exclusion of a hypothesis is expected to happen if exclusions are quoted at the 95% CL.

Type I and Type II Error Rates

(statistics jargon, getting more common in HEP)

- **Type I Error rate**: The probability of excluding the Null Hypothesis H_0 when H_0 is true. Also known as the **False Discovery Rate.**

- **Type II Error rate**: The probability of excluding the Test Hypothesis H_1 when H_1 is true. The **False Exclusion Rate.**

Typically a desired false discovery rate is chosen – this is the value of p_{crit}, also known as α. Then if $p < \alpha$, we can claim evidence or discovery, at the significance level given by α.

We discover new phenomena by ruling out the SM explanation of the data!
-- the Popperian way to do it – we can only prove hypotheses to be false.

In some cases neither H_0 nor H_1 has any *a priori* prejudice for it, like the neutrino mass hierarchy. I'm not sure which gets called Type I and Type II in that case; arbitrary.

Common Standards of Evidence

Physicists like to talk about how many “sigma” a result corresponds to and generally have less feel for *p*-values.

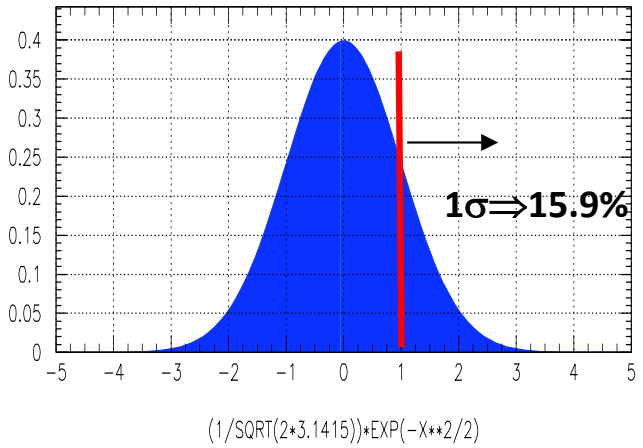
The number of “sigma” is called a “*z*-value” or "z-score" and is just a translation of a p-value using the integral of one tail of a Gaussian

Double_t zvalue = - TMath::NormQuantile(Double_t pvalue)

Folklore:
95% CL -- good for exclusion
3σ: “evidence”
5σ: “observation”
Some argue for a more subjective scale.

z-value (σ)	p-value
1.0	0.159
2.0	0.0228
3.0	0.00135
4.0	3.17E-5
5.0	2.87E-7

$$pvalue = \frac{\left(1 - erf\left(zvalue/\sqrt{2}\right)\right)}{2}$$



One-Sided

Tip: most physicists talk about *p*-values now but hardly use the term *z*-value – we use the word “significance” instead (try not to conflate with "sensitivity" which is expected significance)

Why 5 Sigma for Discovery?

From what I hear: It was proposed in the 1970's when the technology of the day was bubble chambers.

Meant to account for the Look Elsewhere Effect. A physicist estimated how many histograms would be looked at, and wanted to keep the error rate low.

Also too many 2σ and 3σ effects "go away" when more data are collected. Psychology literature with $p<0.05$ criteria – lots of irreproducible results.

Some historical recollections:
http://www.huffingtonpost.com/victor-stenger/higgs-and-significiance_b_1649808.html

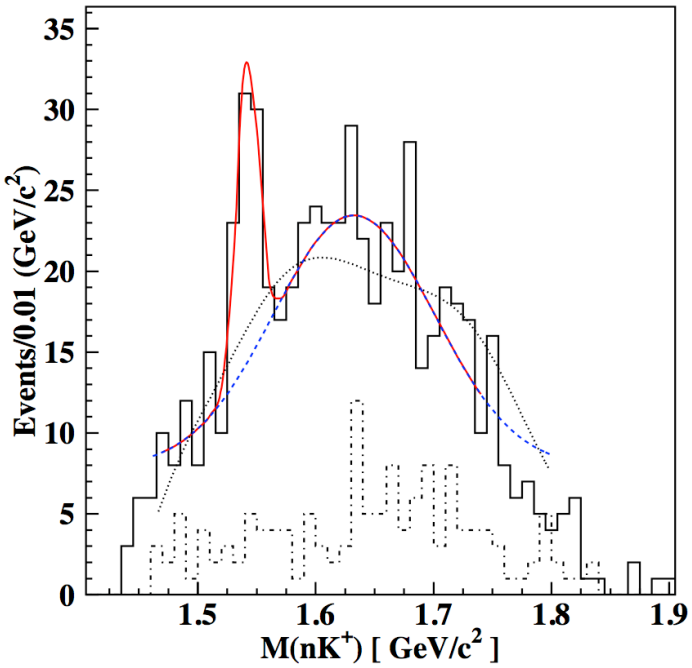
And a modern take on the matter from L. Lyons: https://arxiv.org/abs/1310.1284

Not all estimations of systematic uncertainties are perfect, and extrapolations from typical 1σ variations performed by analyzers out to 5σ leave room for doubt.

Some effects go away when additional uncertainties are considered. Example – CDF Run I High-E_T jets. Not quark compositeness, but the effect could be folded into the PDFs.

If a signal is truly present, and data keep coming in, the expected significance quickly grows (s/sqrt(b) grows as sqrt(integerated luminosity)).

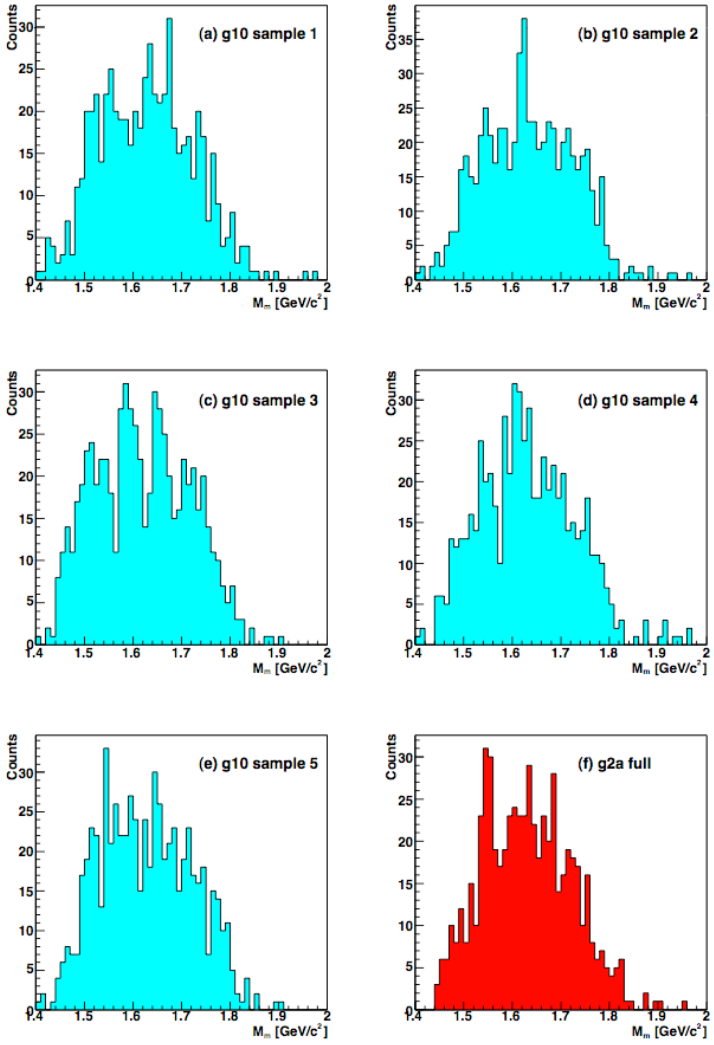
A Cautionary Tale – The Pentaquark "Discoveries"



CLAS Collab., **Phys.Rev.Lett. 91 (2003) 252001**

Significance = 5.2 ± 0.6 σ

Watch out for the background function parameterization!



Five times the data sample
CLAS Collab., **Phys.Rev.Lett. 100 (2008) 052001**

Coverage

A statistical method is said to ***cover*** if the Type-I error rate is no more than the claimed error rate α. Exclusions of test hypotheses (Type-II errors) also must cover – the error rate cannot be larger than stated.

95% CL limits should not be wrong more than 5% of the time if a true signal is present.

If the results are wrong a smaller fraction of the time, the method ***overcovers***.
If the results are wrong a larger fraction of the time, the method ***undercovers***.

Undercoverage is a serious accusation – it has a similar impact as saying that the quoted uncertainties on a result are too small (overselling the ability of an experiment to distinguish hypotheses).

Note: Coverage is a property of a method, not of an individual result. In some cases we may even know that a result is in the unlucky 5% of outcomes, but that individual outcome does not have a coverage property – only the set of possible outcomes.

The word coverage comes from confidence intervals – are they big enough to contain the true value of a parameter being measured and what fraction of the time they do.

p-values and -2ln*Q*

p-value for testing H_1 = $p(-2\ln Q \geq -2\ln Q_{obs} | H_1)$
The green-shaded area to the right.

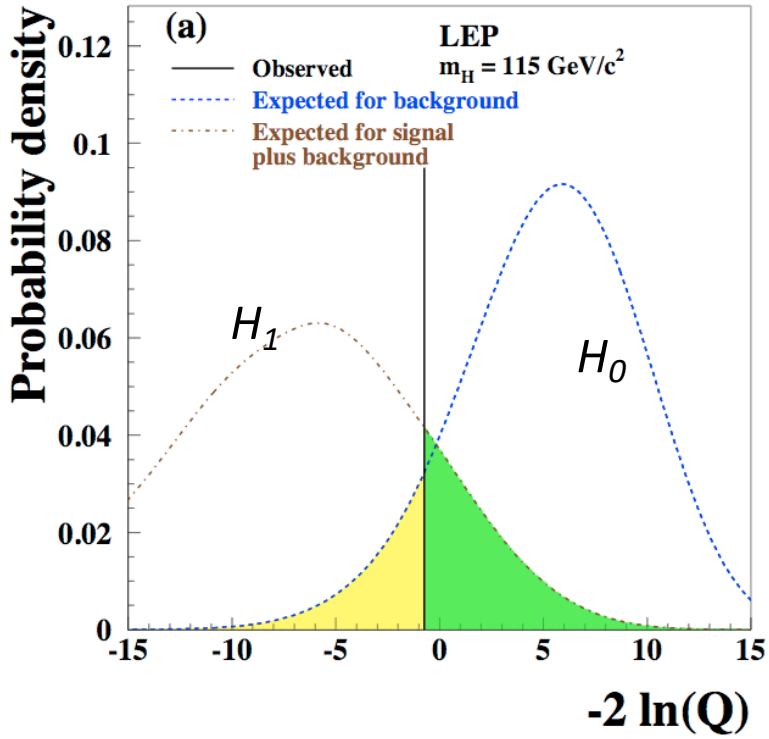
If it is small, reject H_1
The "or-equal-to" has similar effect here too.

This one is called CL_{s+b} (again, not my choice of words). p-values are not confidence levels.

Note: If we quote the CL as the p-value, we will always exclude H_1, just at different CL's each time.

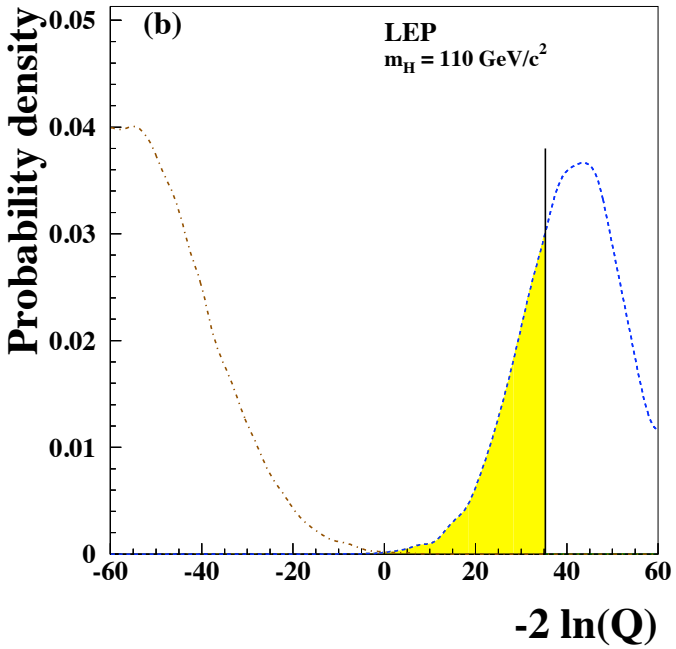
Lucky outcome: exclude at 97% CL
Do we exclude at the 50% CL?

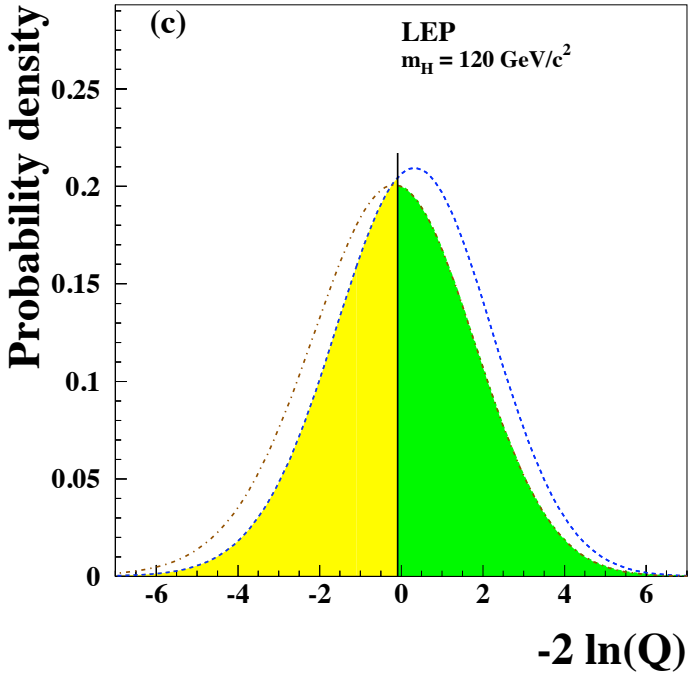
No! Set α once and for all (say 0.05). Then coverage is defined.



From which distribution was the data drawn? We know what the data are; we don't know what the distribution is!

More Sensitivity or Less Sensitivity





signal p-value very small.
Signal ruled out.
Possible to exclude both H_0 and H_1 (-2lnQ=0).
Possible to get outcomes that make you pause to reconsider the modeling. Say -2lnQ<-100 or -2lnQ>+100

Can make no statement about the signal regardless of experimental outcome.

Unlikely (or implausible) outcomes are still possible of course!

The usual sensitivity gauge: Median Expected p_0 if the alternate hypothesis is true

n.b. average p-values tend to be pulled by long tails in the p-value distrib.

Sensitivity is important!

Null results are valid scientific results, but the test must be sensitive for them to be meaningful.

Quote an upper limit!

Some fields of study (not HEP) have addressed this by publishing only positive results with $p<0.05$. Many of these results are not reproducible.

Searching for Anything and Everything

An example from astronomy:

J. Rafael Martinez-Galarza, Harvard & Smithsonian
Finding Needles in the Haystack: Outlier Detection in Astronomical Datasets
Video at: https://events.fnal.gov/colloquium/events/event/martinez-galarza-colloq-2019/

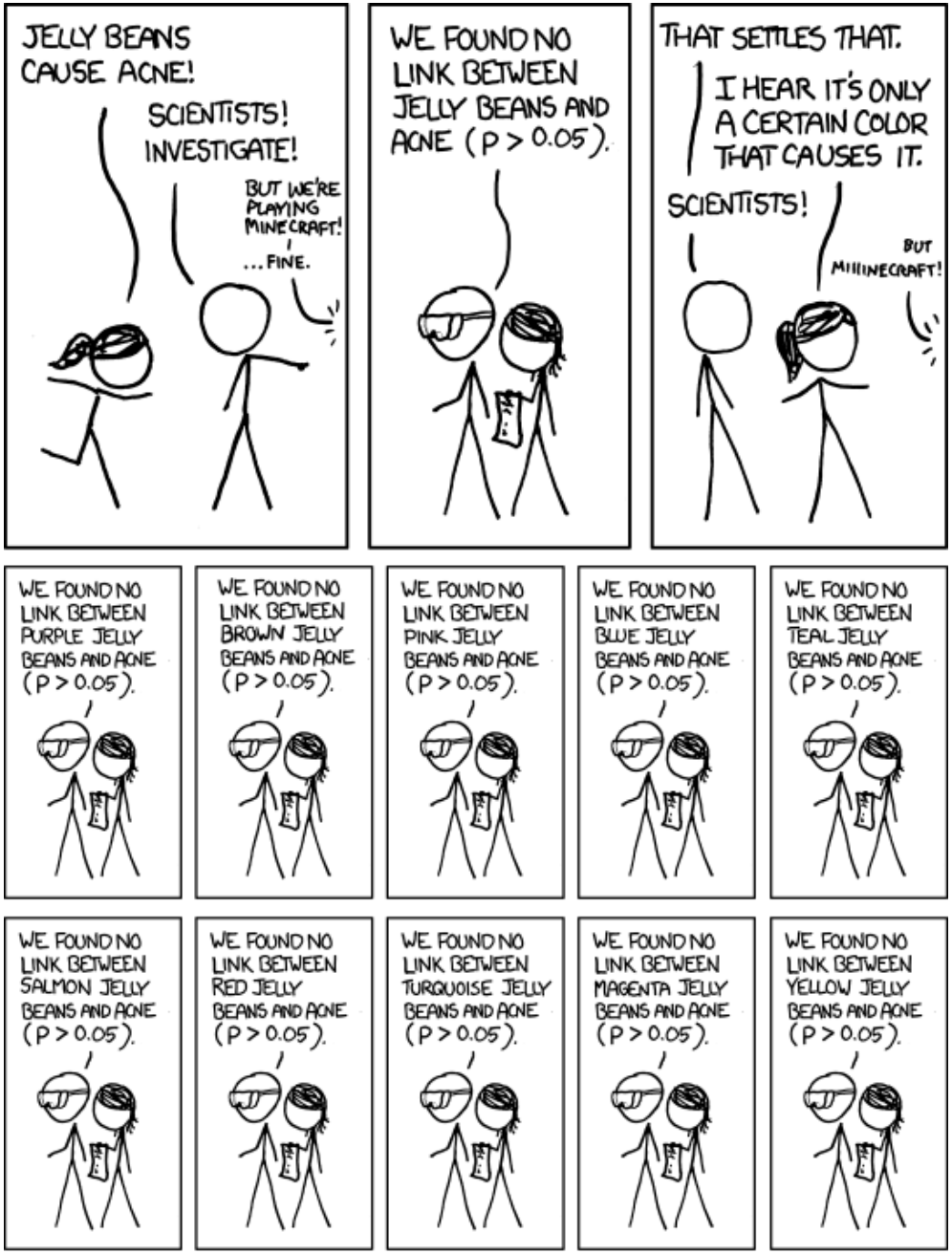
A collider example:

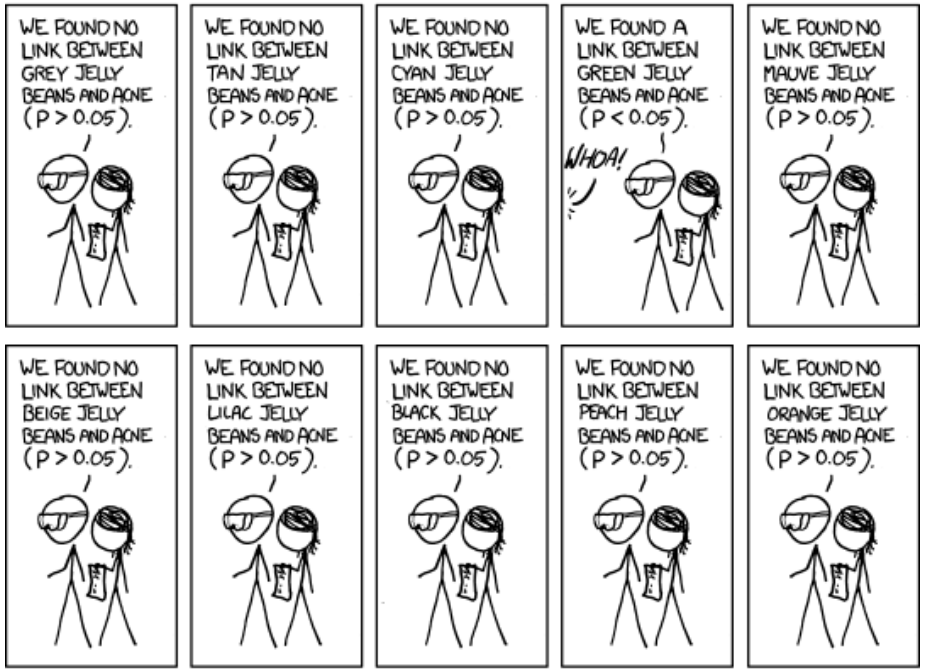
CDF Collab., "Global Search for New Physics" https://arxiv.org/abs/0809.3781
Phys.Rev.D79:011101,2009

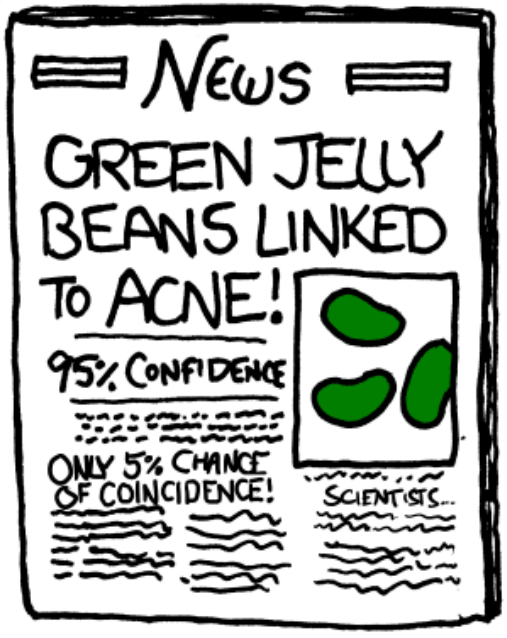
Search for outliers and you will find them! With statistically limited samples of data, you will find them even if there is no new physics to be discovered.

Also called "Cherry Picking" and "p-Hacking"

https://xkcd.com/882/







Look-Elsewhere

Multiple Independent Tests, each with a specified Type-1 Error Rate (false signal), ought to produce Errors at the specified rate (or they aren't powerful enough, or the error rate can be claimed to be lower).

A single analysis can involve many multiple tests. Classic example: A bump-hunt on a histogram.

Old-fashioned way to handle it (Bonferroni) – multiply p-value by the number of independent tests.

Better and just as easy: Dunn-Šidák Correction

- Given *m* different null hypotheses and a familywise alpha level of α, each null hypotheses is rejected that has a p-value lower than

$$\alpha_{SID} = 1 - (1 - \alpha)^{\frac{1}{m}}.$$

From Wikipeida. Sometimes it isn't exactly clear what *m* is.

Where is "Elsewhere?"

A collider collaboration is typically very large; >1000 Ph.D. students.
DUNE has more than 1000 physicists, and there are many neutrino experiments.

Many ongoing analyses for new physics. The chance of seeing a statistical fluctuation that looks like new physics somewhere is large. What is the LEE?

Do we have to correct our previously published p-values for a larger LEE when we add new analyses to our portfolio?

How about the physicist who goes to the library and hand-picks all the largest excesses? What is LEE then?

"Consensus" at the Banff 2010 Statistics Workshop: LEE should correct only for those models that are tested within a single published analysis. Usually one paper covers one analysis, but review papers summarizing many analyses do not have to put in additional correction factors.

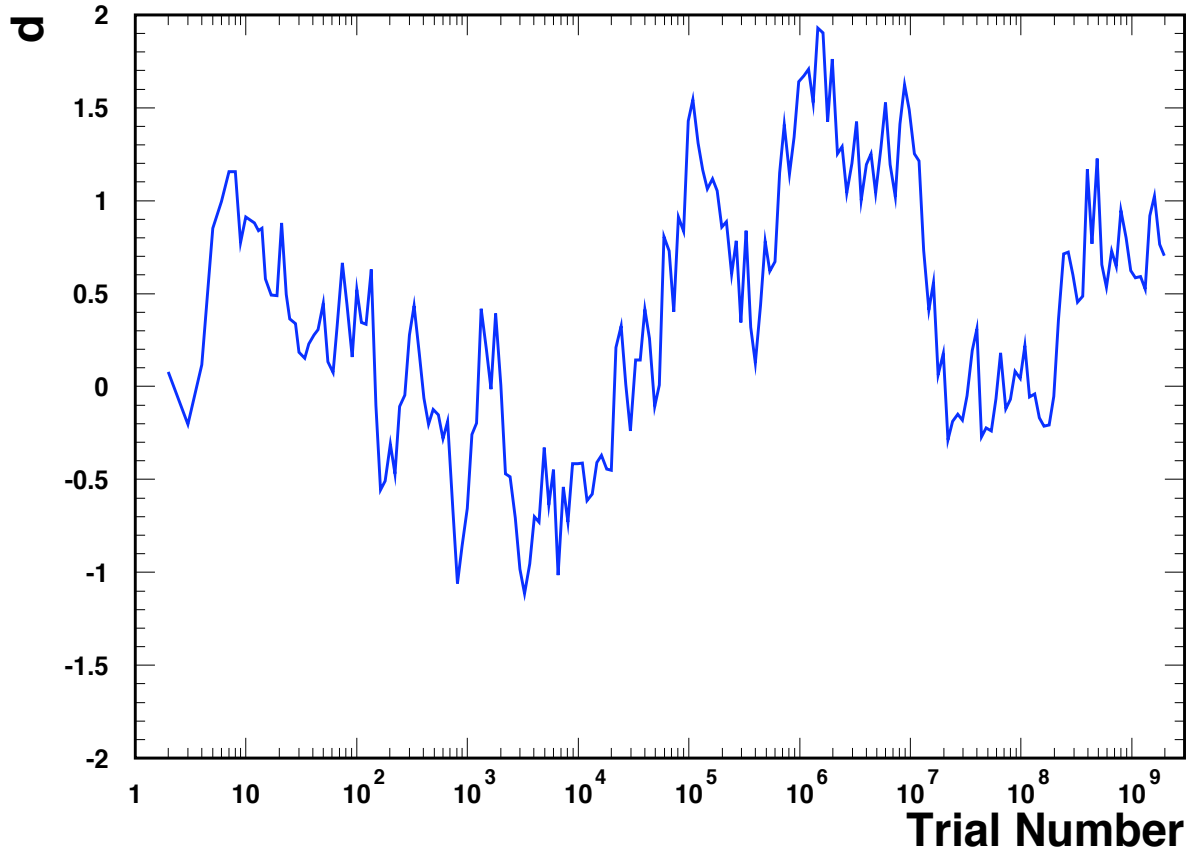
Caveat lector.

Look ElseWHEN

Running averages converge on correct answer, but the deviations in units of the expected uncertainty have a random walk in the logarithm of the number of trials

$$d_n = \frac{\sum_{k=1}^{n} r_k / n}{1/\sqrt{n}}$$

The r_k are IID numbers drawn from a unit Gaussian.



It's possible to cherry-pick a dataset with a maximum deviation. "Sampling to a foregone conclusion"

Stopping Rule: In HEP, we (almost always!) take data until our money is gone. We produce results for the major conferences along the way. Some will coincidentally stop when the fluctuations are biggest. We take the most recent/largest data sample result and ignore (or should!) results performed on smaller data sets. p-values still distributed uniformly from 0 to 1. A recipe for generating "effects that go away"

Packaging Results for Future Use

- Particle physicists have been quite bad at this. Mostly because our criteria for success are so stringent.
- We spend huge amounts of money and effort building and running experiments, and do a great job extracting results.
- But future physicists may discover a problem in the modeling (detector, physics in MC models, parameter values assumed, etc.)
- Re-performing an experiment can be prohibitive.
- Data presented in papers may be corrected using erroneous models.
  - Sometimes one has to uncorrect and recorrect the old results
  - happens all the time in mining old publications for inputs to parton distribution functions
  - Also happens in experimental neutrino data
- Uncorrected data may lack enough explanation of experimental details

Packaging Results for Future Use

Goals of documentation for future readers

- Should be able to reproduce your result
- Should be able to combine your result with other experiment's data
  - The systematic uncertainties and correlations are always the hard part!
- Should be able to change model assumptions and arrive at new results.
  - A systematic uncertainty that used to be BAD or UGLY may now be GOOD

Newer technologies help a lot:

http://hepdata.net

(hepdata.com is something else entirely)

Shared tools: ROOFIT, ROOSTATS, RECAST

APS Journals allow submission of supplemental data.

DOE now requires a "data management plan" which includes standards for preserving data in machine-readable format. At the very least, data shown in published plots must now be made machine readable.

Packaging Results for Future Use

This is all great, but the real issue with preserving data is preserving analysis techniques and software, and handling of systematic uncertainties and correlations.

Calibrations must be applied to data (and MC), and sometimes *ad hoc* corrections are needed to get the most out of a data set.

Future non-collaborators may get their hands on your data and make false discoveries, or just discoveries that a particular part of the detector wasn't working part of the time.

In collider physics, each grad student may need and develop a tiny tweak to a calibration that only he or she needs. In neutrino physics, there are fewer analyses, and thus all tweaks need to be mainstreamed.

Blind Analysis Procedures

Validate analysis as much as possible with simulation and control sample data

Collaboration sign-off on the analysis without looking at signal-region data

Data are “unblinded” (or hidden offsets revealed)

A hard-line approach: Collaboration must approve the unblinded result and submit for publication, even if it contains mistakes that are obvious only when the signal region data are investigated.

“Blind”, not “deaf and dumb”: Allow review of possible mistakes. But then we’re not really blind, are we?

A practical concern: One analysis group’s calibration sample is another’s signal sample!
They can accidentally unblind each other!
Do we need to keep people out of each others’ meetings?
Collaboration by-laws usually prohibit denying access to data or to analysis meetings.
Usually a “good-faith effort”

Summary

- Statistical treatment in HEP is getting better each year
- Powerful tools exist that cover many use cases.
- Documentation is also quite good

- But: Some calculations are prohibitively expensive when done "by the book" e.g. 5σ sensitivities.
- Many methods you might come across are *approximations* designed to get close to a result that is otherwise difficult to compute
- Knowing *a priori* the distribution of a test statistic helps avoid the need to throw lots of toy experiments.
- Often these approximations break down, and often in the cases we are most interested in for neutrino physics.
- If something seems fishy, it probably is.
- Talk to experts if you need help! Especially when starting your analysis, or designing your experiment. Not right before your paper or thesis deadline!

Extras

xkcd.com

Stephen Hawking was famous for making bets he was happy to lose.

Treatment of Asymmetric Uncertainties

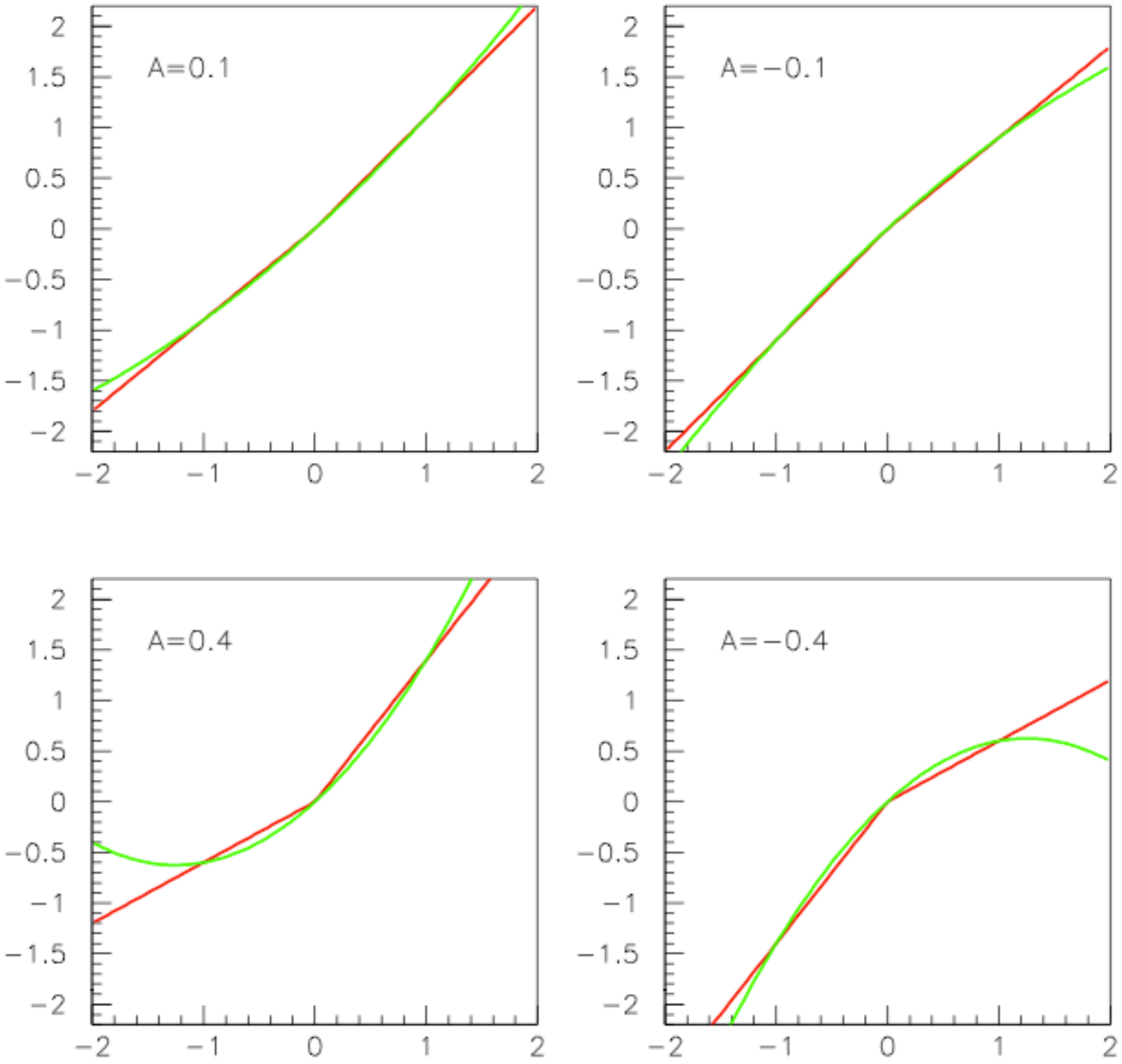
These cases are pretty clear – the underlying parameter, the energy scale, has a (Gaussian? Your choice) distribution, while it has a nonlinear, possibly non-monotonic *impact* on the model prediction.

The same parameter may have a linear, symmetrical impact on another model prediction, and we will have to treat them as correlated in statistical analysis tools.

Treatment is ambiguous when little is known why the uncertainties are asymmetric, or it is not clear how to extrapolate/interpolate them.

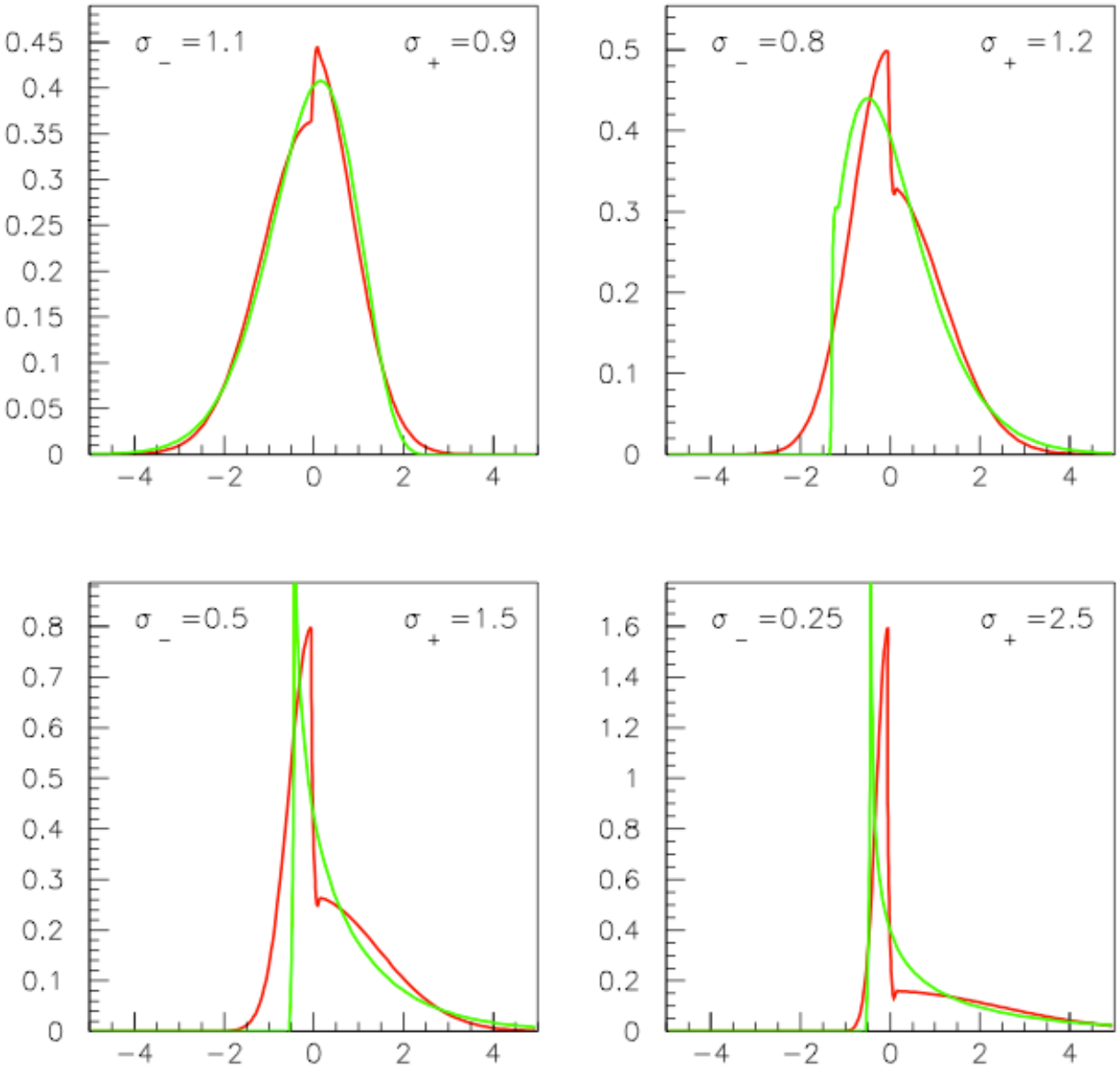
See R. Barlow,
"Asymmetric Systematic Errors", arXiv:physics/0306138
"Asymmetric Statistical Errors", arXiv:physics/0406120

Quadratic Impacts of Asymmetric Uncertainties



R. Barlow

Resulting Prior Distributions for alternative handling of Asymmetric Impacts

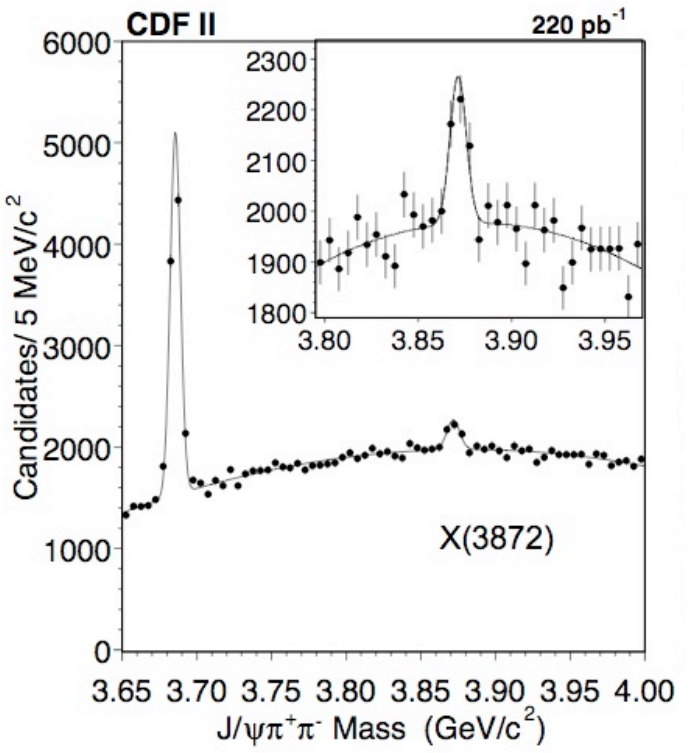


R. Barlow

Integrating over Systematic Uncertainties Helps Constrain their Values with Data

Nuisance parameters: $\boldsymbol{\nu}$
Parameter of Interest: r

Example: suppose we have a background rate prediction that's 50% (fractionally) uncertain -- goes into $\pi(\boldsymbol{\nu})$. But only a narrow range of background rates contributes significantly to the integral. The kernel falls to zero rapidly outside of that range.



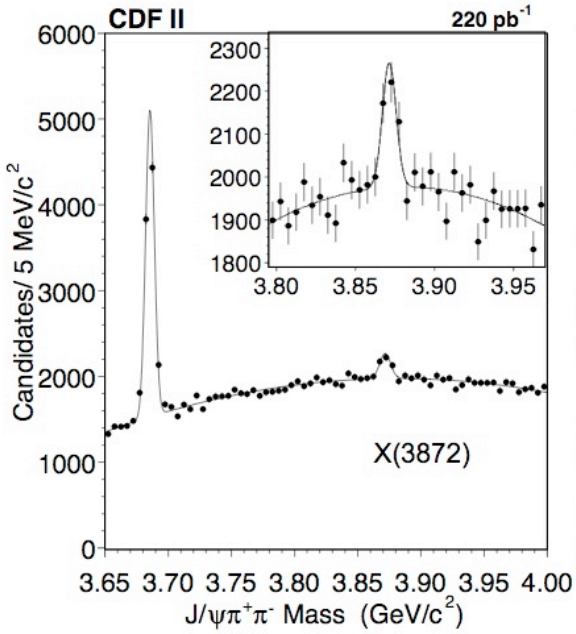
Can make a posterior probability distribution for the background too -- narrow belief distribution.

Asymmetric Uncertainties and Priors

Measurements, and even theoretical calculations, frequently are assigned asymmetric uncertainties:

Value = 10^{+2}_{-1}, or more extremely, 10^{+2}_{+2} (ouch). When the uncertainties have the same sign on both sides, it is worthwhile to check and see why this is the case.

Example – we seek a bump in a mass distribution by counting events in a small window around where the bump is sought.



The detector calibration has an energy uncertainty (magnetic field or chamber alignment for tracks, or much larger effect, calorimeter energy scales for jets).

Shift the calibration scale up – predicted peak shifts out of the window → downward shift in expected signal prediction.

Shift the calibration down – predicted peak shifts out of the other side of the window → downward shift in expected signal prediction

Hypothesis Testing

- *p*-values
- Coverage and Power
- Test Statistics and Optimization
- Incorporating Systematic Uncertainties
- Multiple Testing ("Look Elsewhere Effect")

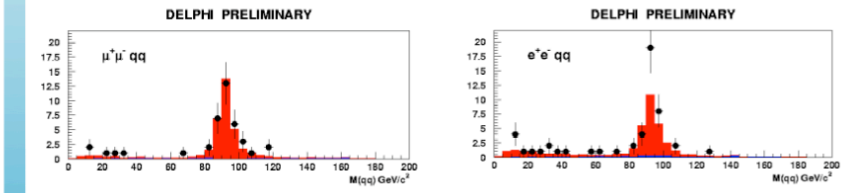
Thus the unfacts, did we possess them, are too imprecisely few to warrant our certitude...
J. Joyce, Finnegan's Wake

Another Bump That Went Away

A preliminary set of distributions shown at a LEPC presentation

llqq events at LEP2

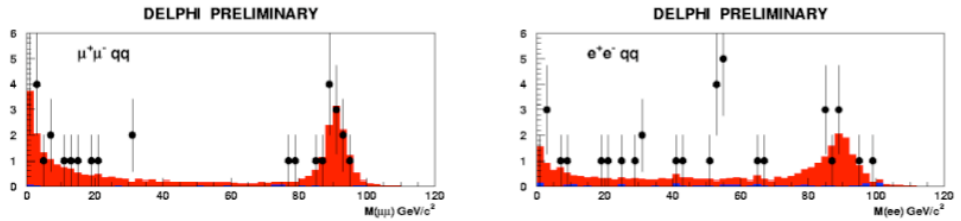
- DELPHI has more than 400 pb^{-1} collected at LEP2
- Check of the mass spectrum:



M_{qq} (after 4C-fit)

llqq events at LEP2

•Excess in eeqq, when $M_{qq} \sim M_z$: check M_{ee}



M_{ll} (with M_{qq} in Z region)

Benefit of having four LEP experiments – at the very least, there's more data.
This one was handled very well – cross checked carefully.

But, they shared models – Monte Carlo programs, and theoretical calculations.

The Literature is Full of Bumps that Went Away

See Sheldon Stone, "Pathological Science", **hep-ph/0010295**

My personal favorite is the "Split A_2 resonance"

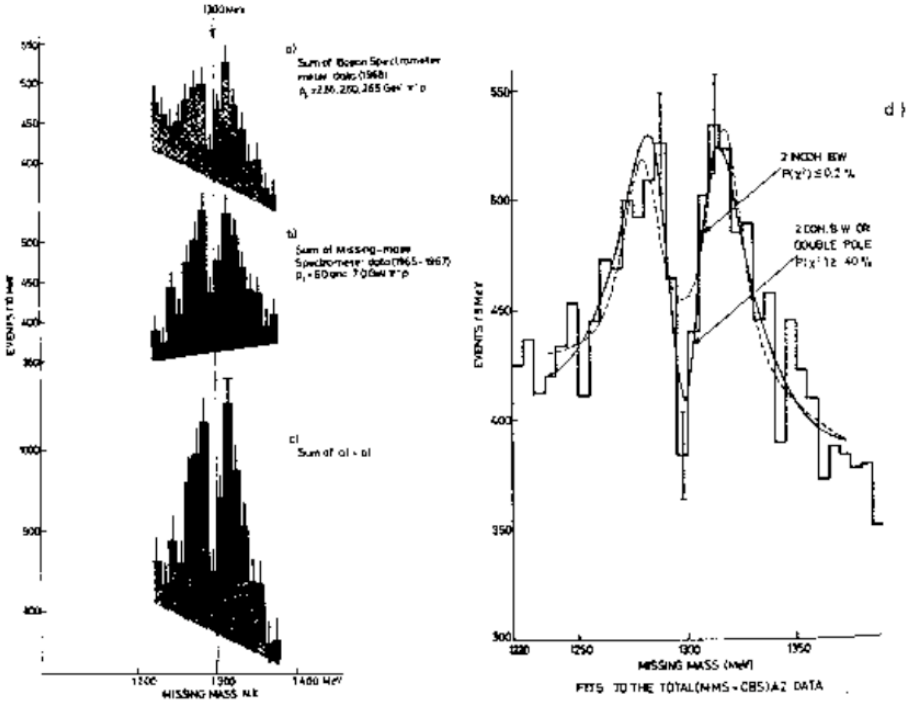


Figure 3: (a-c) Evidence for A_2 splitting in $\pi^- p \to pX^-$ collisions in the two CERN experiments, (d) same as (c) in 5 MeV bins fit to two hypotheses.

Text from Sheldon's article:

How did this happen? I have heard several possible explanations. In the MMS experiment, I was told that they adjusted the beam energy so the dip always lined up! Another possibility was revealed in a conversation I had with Schübelin, one of the CBS physicists. He said: "The dip was a clear feature. Whenever we didn't see the dip during a run we checked the apparatus and always found something wrong." I then asked him if they checked the apparatus when they did see the dip, and he didn't answer.

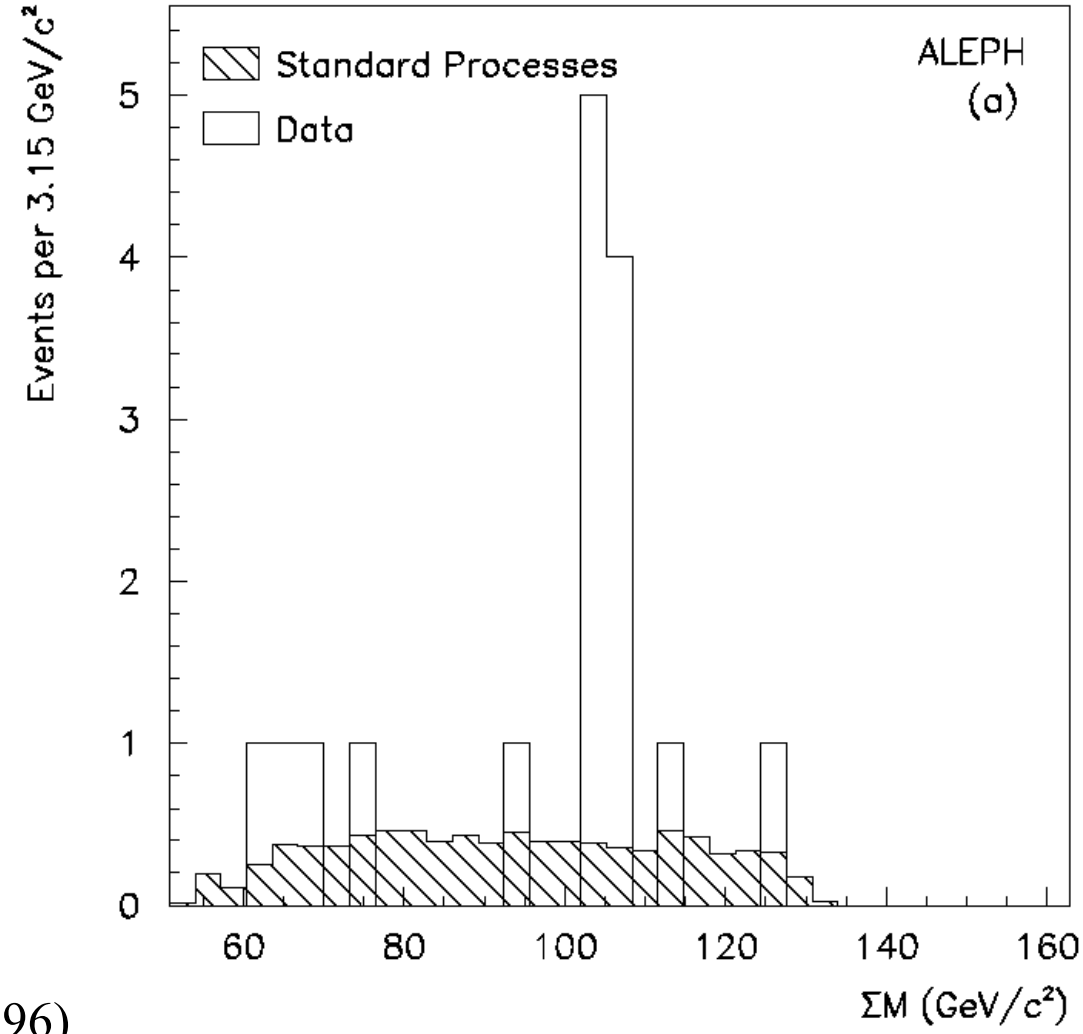
What about the other experiments that did see the dip? Well there were several experiments that didn't see it. Most people who didn't see it had less statistics or poorer resolution than the CERN experiments, so they just kept quiet. Those that had a small fluctuation toward a dip worked on it until it was publishable; they looked at different decay modes or t intervals, etc. (This is my guess.)

At Least ALEPH Explained What They Did

"the width of the bins is designed to correspond to twice the expected resolution ... and their origin is deliberately chosen to maximize the number of events found in any two consecutive bins"

LEP ended up running a little extra at 130 GeV to collect more data to test this hypothesis.

ALEPH Collaboration, Z. Phys. C71, 179 (1996)



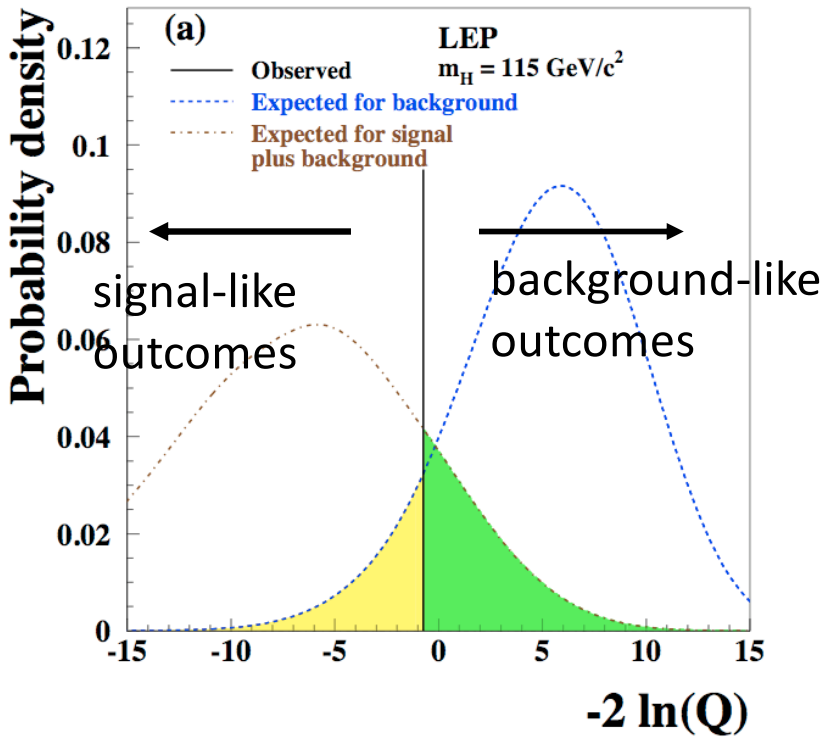
Dijet mass sum in $e^+e^- \rightarrow jjjj$

A More Sophisticated Test Statistic

What if you have two or more bins in your histogram? Not just a single counting experiment any more.

Still want to rank outcomes as more signal-like or less signal-like

Neyman-Pearson Lemma (1933): The likelihood ratio is the "uniformly most powerful" test statistic



$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}|H_1,\hat{\nu})}{L(\text{data}\,|\,H_0,\hat{\hat{\nu}})}\right)$$

yellow=p-value for ruling out H_0. Green= p-value for ruling out H_1

Acts like a difference of Chisquareds in the Gaussian limit

$$-2\ln Q \rightarrow \Delta\chi^2 = \chi^2(data\,|\,H_1) - \chi^2(data\,|\,H_0)$$

What's with $\hat{\nu}$ and $\hat{\hat{\nu}}$?

$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}|H_1,\hat{\nu})}{L(\text{data}\,|\,H_0,\hat{\hat{\nu}})}\right)$$

We parameterize our ignorance of the model predictions with nuisance parameters.

A model with a lot of uncertainty is hard to rule out!

-- either many nuisance parameters, or one parameter that has a big effect on its predictions and whose value cannot be determined in other ways

$\hat{\nu}$ maximizes *L* under H_1

$\hat{\hat{\nu}}$ maximizes *L* under H_0

What's with $\hat{\nu}$ and $\hat{\hat{\nu}}$?

A *simple hypothesis* is one for which the only free parameters are parameters of interest.

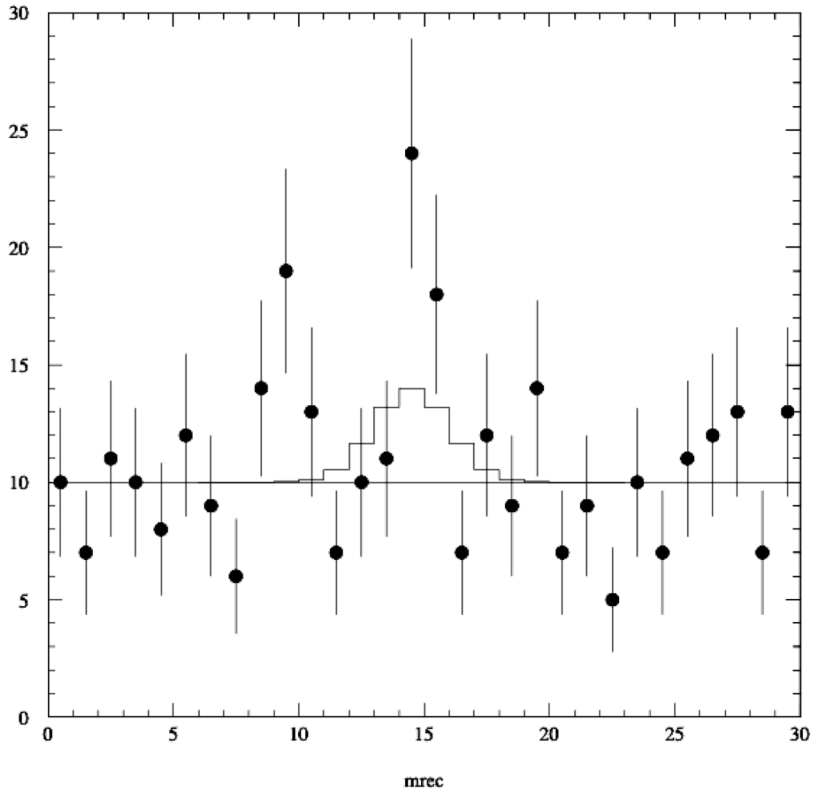
A *compound hypothesis* is less specific. It may have parameters whose values we are not particularly concerned about but which affects its predictions. These are called *nuisance parameters*, labeled ν.

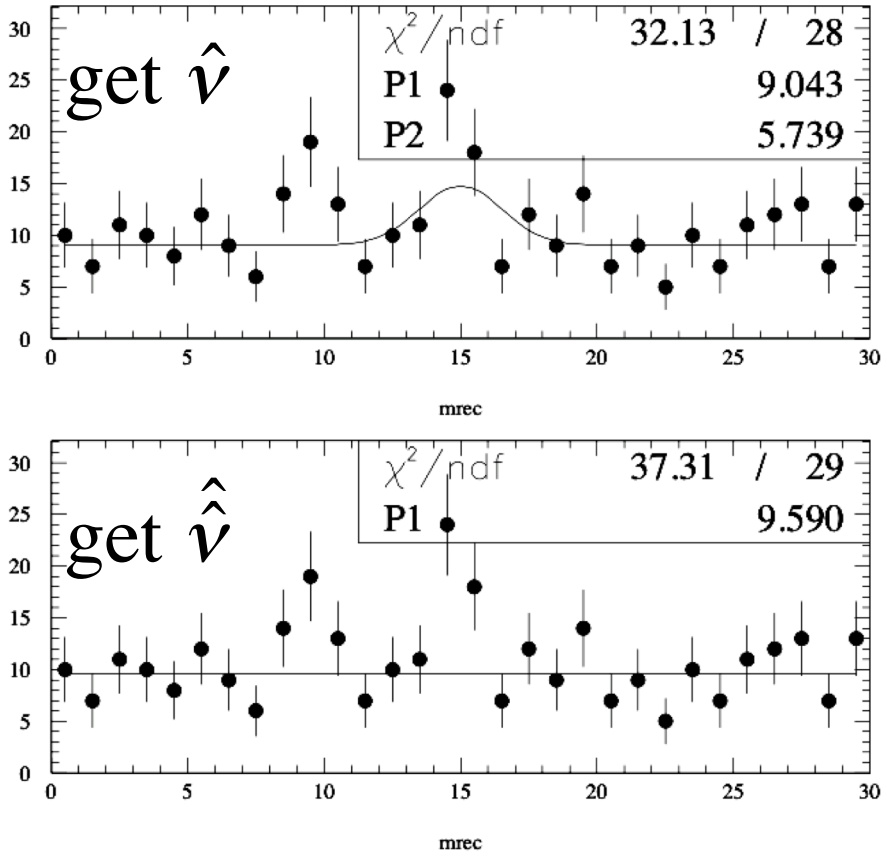
Example: H_0= Normal Mass Ordering. H_1= Inverted Mass Ordering. Both make predictions about what may be seen in an experiment. A nuisance parameter would be, for example, the beam flux. It affects the predictions but in the end of the day we are really concerned about H_0 and H_1.

Fit twice! Once assuming H_0, once assuming H_1

Example: flat background, 30 bins, 10 bg/bin, Gaussian signal.
Run a pseudoexperiment (assuming s+b).

Fit to flat bg, Separate fit to flat bg + known signal shape.
The background rate is a nuisance parameter ν = b
Use fit signal and bg rates to calculate Q.
Fitting the signal is a separate option.





p-values and -2ln*Q*

p-value for testing H_0 = $p(-2\ln Q \le -2\ln Q_{obs} | H_0)$
The yellow-shaded area to the right.

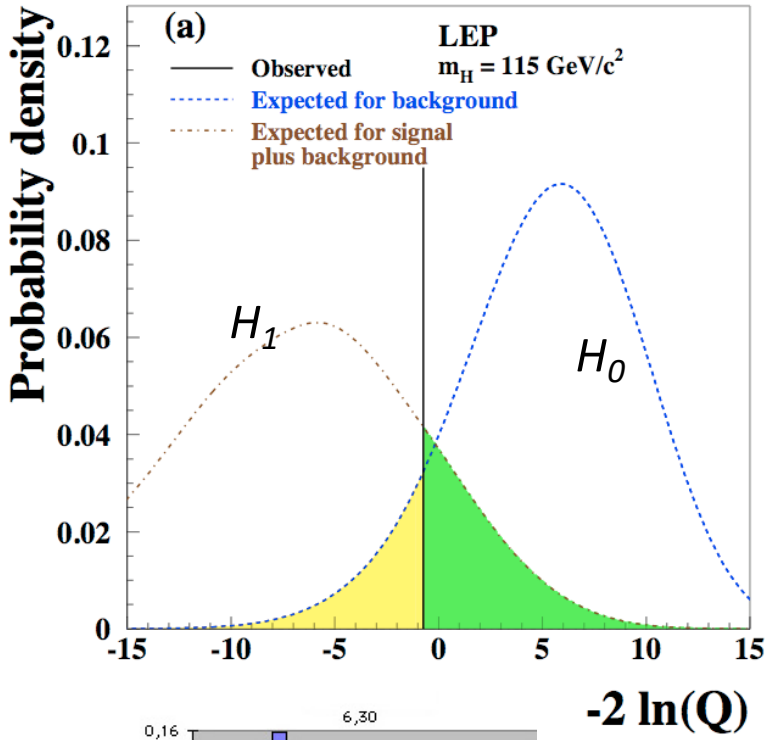
The "or-equal-to" is important here. For highly discrete distributions of possible outcomes – say an experiment with a background rate of 0.01 events (99% of the time you observe zero events, all the same outcome), then observing 0 events gives a p-value of 1 and not 0.01.

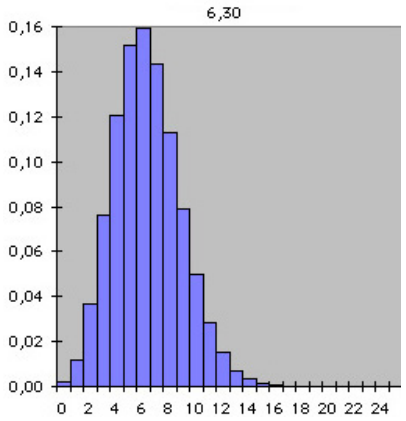
Shouldn't make a discovery with 0 observed events, no matter how small the background expectation! (or we would run the LHC with just one bunch crossing!).

This *p*-value is often called "1-CL_b" in HEP. (apologies for the notation! It's historical)

$CL_b = p(-2\ln Q \ge -2\ln Q_{obs} | H_0)$

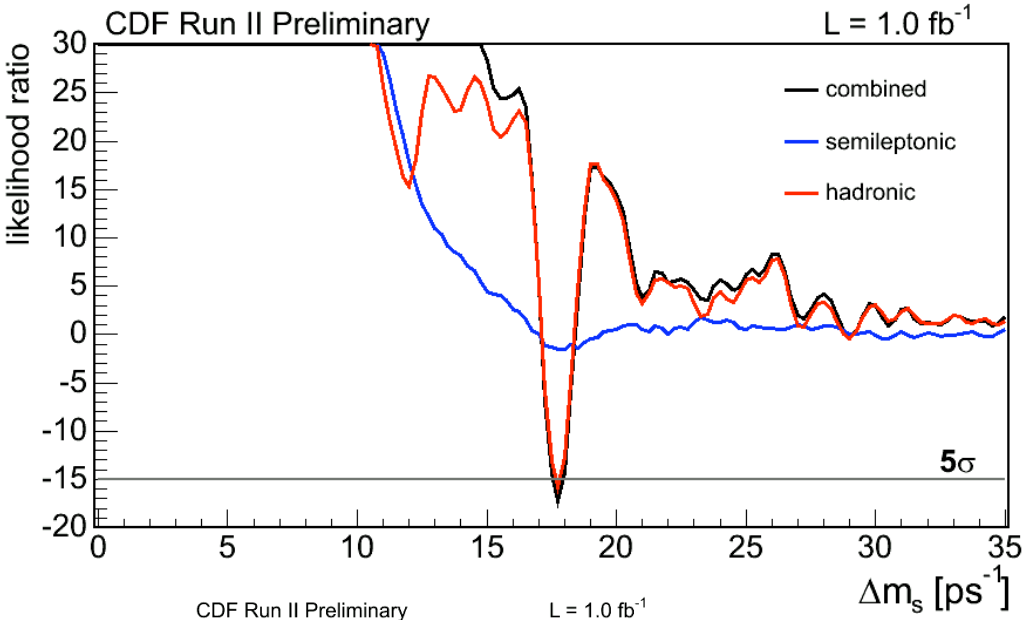
Due to the "or equal to"'s $(1-CL_b) + CL_b \ne 1$

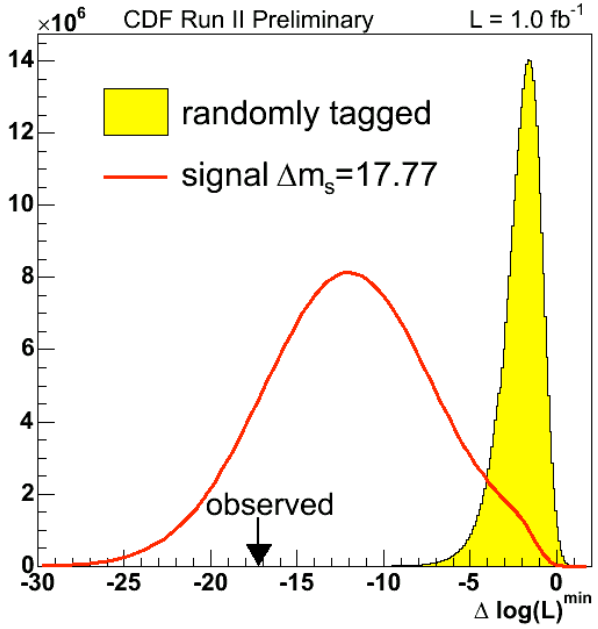


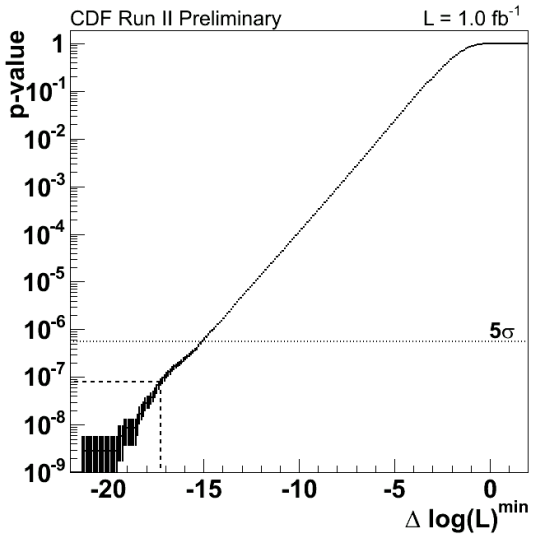


For an experiment producing a single count of events all choices of test statistic are equivalent. *Usually* more events = more signal-like.

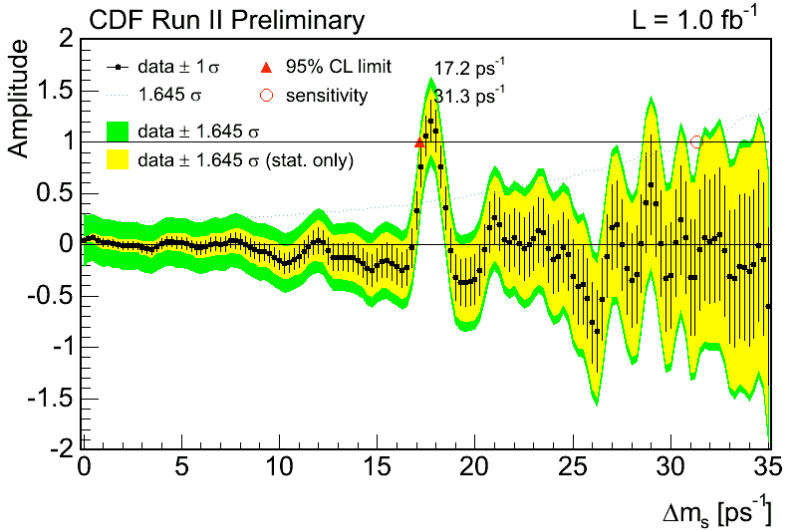
LLR Is not only used in Search Contexts – Precision Measurements too!







Mixing rate – more akin to a cross section measurement



Phys. Rev. Lett 97, 242003 (2006)

Power

The Type-I Error Rate is α or less for a method that covers. But I can cover with an analysis that just gives a random outcome – in α of the cases, reject H_0, and in 1-α of the cases, do not reject H_0.

But we would like to reject H_1 when it is false.

The quoted Type-II error rate is usually given the symbol β (but some use 1-β).

For excluding models of new physics, we typically choose β=0.05, but sometimes 0.1 is used (90% CL limits are quoted sometimes but not usually in HEP).

Classical two-hypothesis testing (not used much in HEP, but the LHC may lean towards it).

H_0 is the null hypothesis, and H_1 is its "negative". We know *a priori* either H_0 or H_1 is true. Rejecting H_0 means accepting H_1 and vice versa (n.b. not used much in HEP)

Example: H_0: The data are described by SM backgrounds
H_1: There is a signal present of strength μ>0. Can also be μ≠0 but most models of new physics add events. (Some subtract events! Or add in some places and subtract in others!!)

The Classical Two-Hypothesis Likelihood Ratio

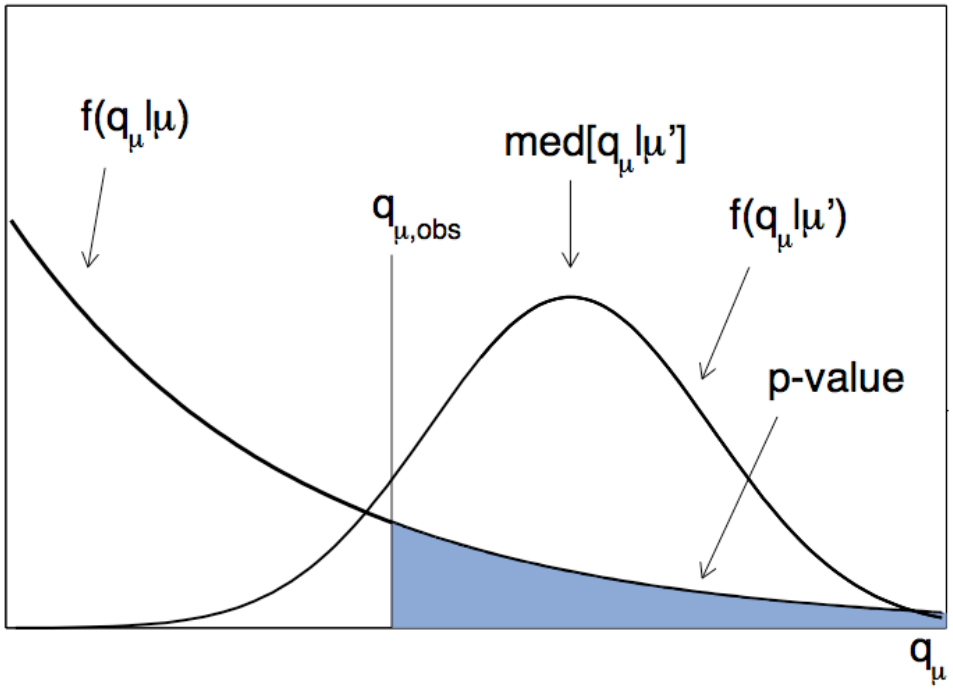
Distinguishing between μ=0 (zero signal, SM, Null Hypothesis) and μ>0 (the test hypothesis)

Assumption Warning! Signal rates scale with a single parameter μ

$$q_\mu \equiv 2\ln\left(\frac{L(\text{data}|\hat{\mu},\hat{\nu})}{L(\text{data}\,|\,\mu,\hat{\hat{\nu}})}\right)$$

$\hat{\mu}$ is the best-fit value of the signal rate. Can be zero. Your choice to allow it to go negative.

μ is quadratically dependent on coupling parameters (or worse. More on this later).



Larger q_0 is more signal-like

q_μ>0 always because H_1 is a superset of H_0 and therefore always fits at least as well.

ATLAS performance projections, CERN-OPEN-2008-020

Wilks's theorem

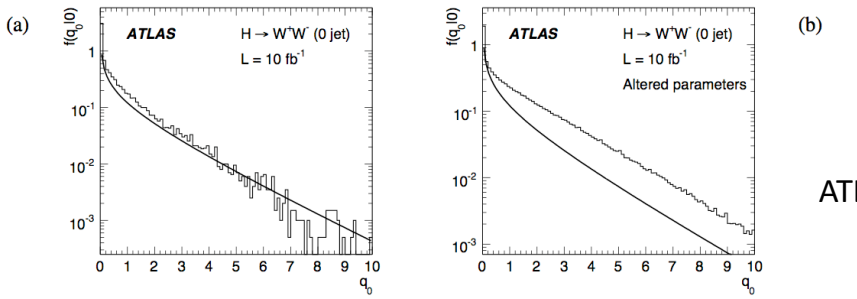
If the true value of the signal rate is given by μ, then q_μ is distributed according to a χ^2 distribution with one degree of freedom.

Assumptions: Underlying PDFs are Gaussian (this is never the case)

Systematic uncertainties also complicate matters. If a systematic uncertainty which has no a priori constraint can fake a signal, then there is no sensitivity in the analysis.

Example: data = signal + background, single counting experiment.
If the background is completely unknown a priori, there is no way to make any statement about the possibility of a signal. So q_μ=0 for all outcomes for all μ.

Poisson Discreteness also makes Wilkes's theorem only approximate.



ATLAS performance projections, CERN-OPEN-2008-020

Figure 8: The distribution of the test statistic q_0 for $H+0j \to WW+0j$, under the background-only hypothesis, with the same fixed QCD WW shape parameters used at both the generator and the fit level, for $m_H = 150$ GeV and for an integrated luminosity of 10 fb^{-1} (a) with the same shape parameters for event generation and fitting; (b) with altered shape parameters. A $\frac{1}{2}\chi^2_1$ distribution is superimposed.

The "Asimov" Approximation for Computing Median Expected Sensitivity

We seek the median of some distribution, say a p-value or a limit (more on limits later).

- CPU constraints computing p-values, limits, and cross sections
- Need quite a few samples to get a reliable median. Usually many thousands.
- I use the uncertainty on the mean to guess the uncertainty on the median (not true for very discrete or non-Gaussian distributions

$$\sigma_{avg} = RMS / \sqrt{n-1}$$

- Often have to compute median expectations many times when optimizing an analysis

But: The median of a distribution is the entry in the middle.

Let's consider a simulated outcome where data = signal(pred)+background(pred), and compute only one limit, p-value, or cross section, and call that the median expectation.

Named after Isaac Asimov's idea of holding elections by having just one voter, the "most typical one" cast a single vote, in the short story *Franchise*.

A Case in which the Asimov Approximation Breaks Down

Usually it's a very good approximation.

Poisson discreteness can make it break down, however.

Example: signal(pred)=0.1 events, background(pred)=0.1 events.

The median outcome is 0 events, not 0.2 events.

In fact, 0.2 events is not a possible outcome of the experiment at all!

For an observed data count that's not an integer, the Poisson probability must be generalized a bit (seems to work okay):

$$p_{Poiss}(n,r) = \frac{r^n e^{-r}}{\Gamma(n+1)}$$

Some Comments on Fitting

- Fitting is an optimization step and is not needed for correctly handling systematic uncertainties on nuisance parameters.

  More on systematics later

- Some advocate just using -2lnQ with fits as the final step in quoting significance (Fisher, Rolke, Conrad, Lopez)

- But we do not know the distribution from which the data fit is drawn – could have gotten "lucky" or not.

- Fits can "fail" -- MINUIT can give strange answers (often not MINUIT's fault). Good to explore distributions of possible fits, not just the one found in the data.

Incorporating Systematic Uncertainties into the p-Value

Two plausible options:

"Supremum p-value"

Choose ranges of nuisance parameters for which the p-value is to be valid

Scan over space of nuisance parameters and calculate the p-value for each point in this space.

Take the largest (i.e., least significant, most "conservative") p-value.
"Frequentist" -- at least it's not Bayesian. Although the choice of the range of nuisance parameter values to consider has the same pitfalls as the arbitrary choice of prior in a Bayesian calculation.

"Prior Predictive p-value"

When evaluating the distribution of the test statistic, vary the nuisance parameters within their prior distributions. "Cousins and Highland"

$$p(x) = \int p(x \mid \nu) p(\nu) d\nu$$

Resulting p-values are no longer fully frequentist but are a mixture of Bayesian and Frequentist reasoning. In fact, adding statistical errors and systematic errors in quadrature is a mixture of Bayesian and Frequentist reasoning. But very popular. Used in ttbar discovery, single top discovery.

Other Possible ways to Incorporate Systematic Uncertainties in P-Values

For a nice (lengthy) review, see

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

Confidence interval method

Use the data twice – once to calculate an interval for a nuisance parameter, and a second time to compute supremum p-values in that interval, and correct for the chance that the nuisance parameter is outside the interval.

Hard to extend to cases with many (hundreds!) of nuisance parameters

Plug-in p-value

Find the best-fit values of the uncertain parameters and calculate the tail probability assuming those values.

Double use of the data; ignores uncertainty in best-fit values of uncertain parameters. Works best when the data strongly constrain the important uncertainties.

Other Possible ways to Incorporate Systematic Uncertainties in P-Values

Fiducial method – See Luc's note. I do not know of a use of this in a publication

Posterior Predictive p-value

Probability that a future observation will be at least as extreme as the current observation assuming that the null hypothesis is true.

Advantages: Uses measured constraints on nuisance parameters
Disadvantages: Cannot use it to compute the sensitivity of an experiment you have yet to run.

In fact, all methods that use the data to bound the nuisance parameters in the pseudoexperiment ensemble generation cannot be used to compute the *a priori* sensitivity of an experiment with systematic uncertainties.

Of course the sensitivity of an experiment is a function of the true values of the nuisance parameters.

Look-Elsewhere Effect Comments

Generally, no LEE for limits.

BUT: Taking the union of excluded parameter spaces is not well defined (breaks coverage. Points would have multiple chances to be excluded).

We overlay exclusion contours all the time.

LEE for p-values for sure. Bayes Factor? Question for Jim Berger

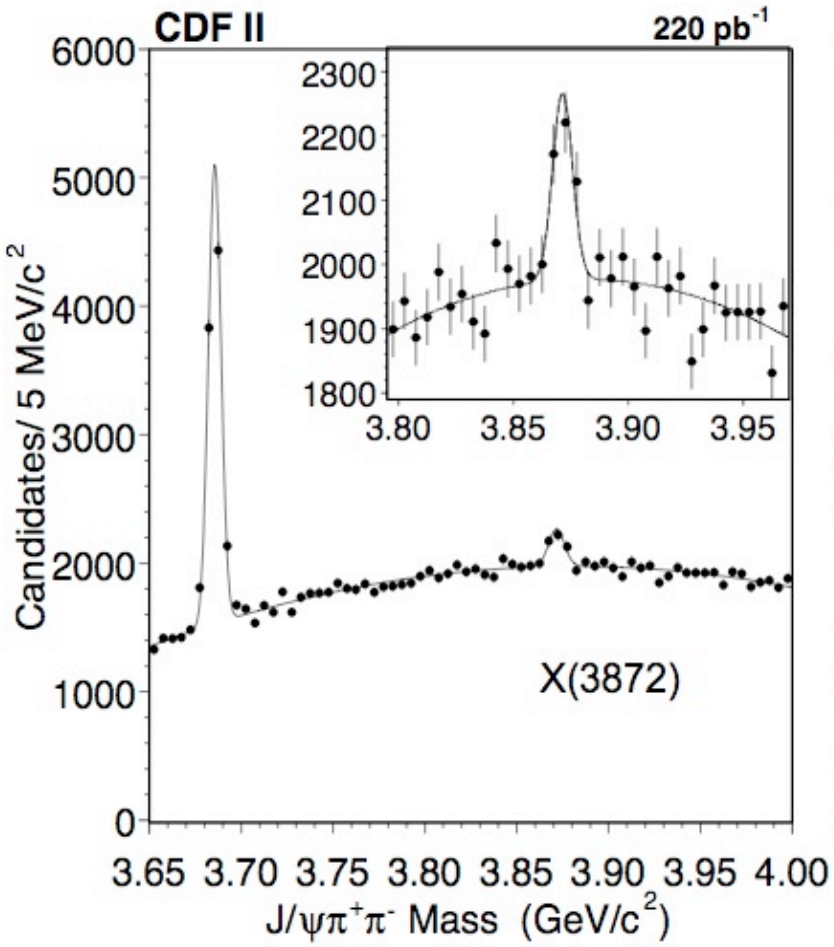
LEE depends on how many independent testable models. More dimensionality of parameters of interest not present on the null, the larger the LEE
Example: LEP MSSM Higgs search – multiple small excesses. Theorists found models to predict them all simultaneously, did not compute LEE.

LEE when most of the model space has been excluded already?

Not all dimensions are created equally. (delta msquared, sin^2 theta)

"On-Off" Example

Select events with J/ψ(→ll) $\pi^+\pi^-$ candidates. Lots of nonresonant background whichis poorly understood *a priori*, but there's a *lot* of it.



Typical strategy:
Fit the background outside of the signal peak, and interpolate the background under the signal to subtract it off.

The ratio of events in the sidebands to the background prediction under the signal is called **τ**

Guess a shape that fits the backgrounds, and fit it with a signal.

"Weak" Sideband Constraints

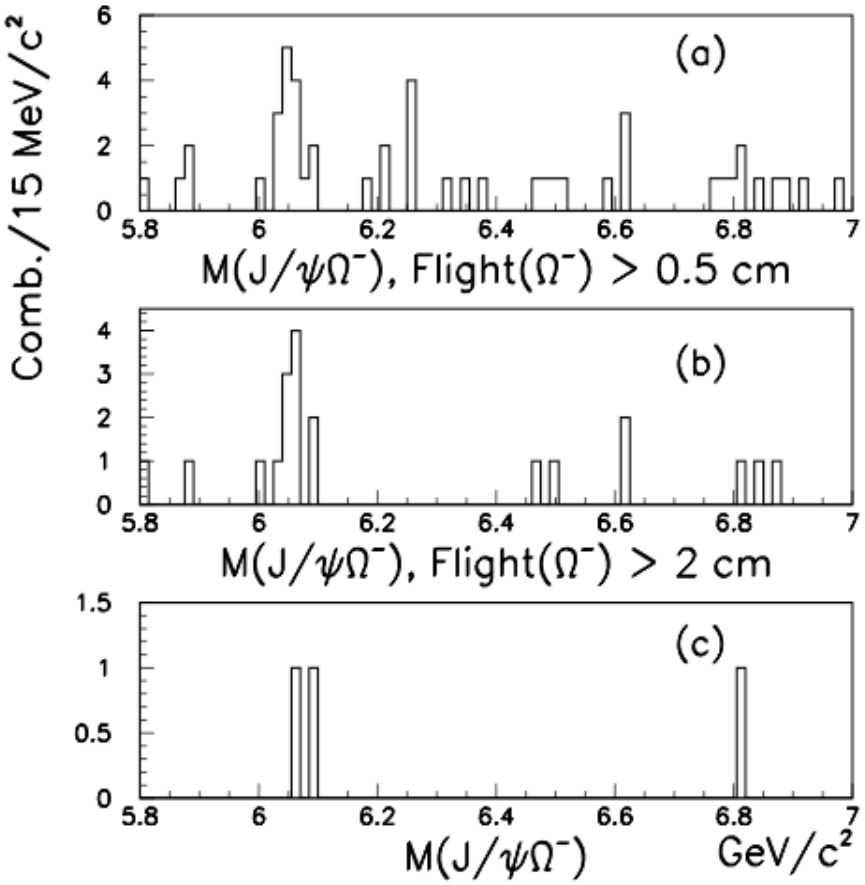


FIG. 8: (a,b) The invariant mass distribution of $J/\psi\,\Omega^-$ combinations for candidates where the transverse flight requirement of the Ω^- is greater than 0.5 cm and 2.0 cm. (c) The invariant mass distribution of $J/\psi\,\Omega^-$ combinations for candidates with at least one SVXII measurement on the Ω^- track. All other selection requirements are as in Fig. 5(c).

CDF's Ω_b observation paper:

Phys.Rev. D80 (2009) 072003

No Sideband Constraints?

Example: Counting experiment, only have a priori predictions of expected signal and background

All test statistics are equivalent to the event count – they serve to order outcomes as more signal-like and less signal-like. More events == more signal-like.

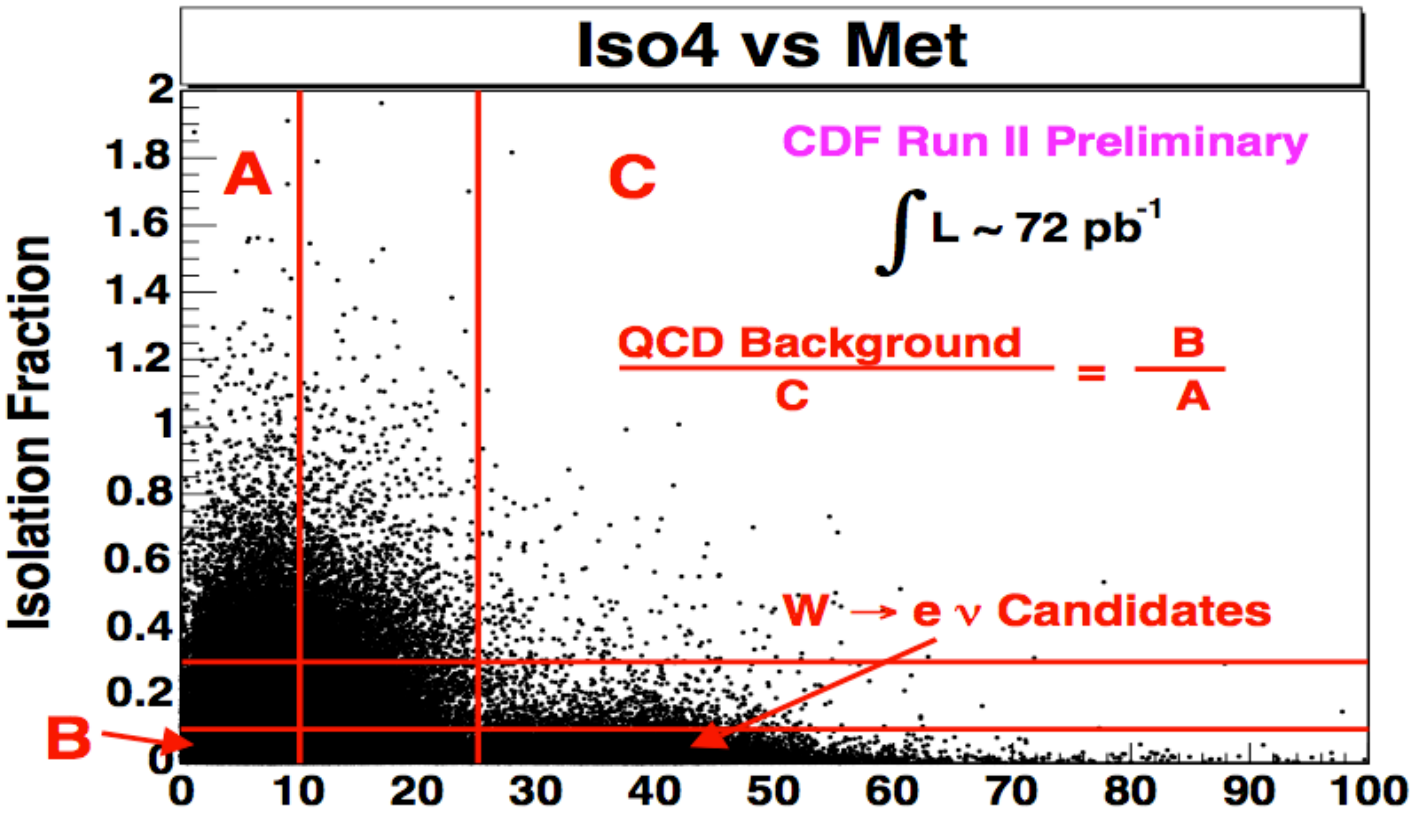
Classical example: Ray Davis's Solar Neutrino Deficit observation. Comparing data (neutrino interactions on a Chlorine detector at the Homestake mine) with a model (John Bahcall's Standard Solar Model). Calibrations of detection system were exquisite. But it lacked a standard candle.

How to incorporate systematic uncertainties? Fewer options left.

Another example: Before you run the experiment, you have to estimate the sensitivity. No sideband constraints yet (except from other experiments).

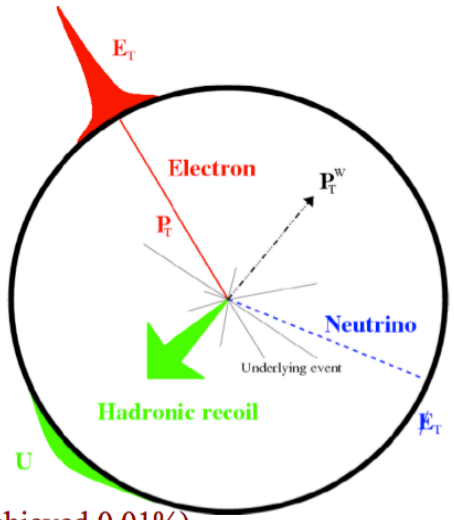
"ABCD" Methods

CDF's W Cross Section Measurement



Want QCD contribution to the "D" region where signal is selected.

Assumes: MET and ISO are uncorrelated sample by sample
Signal contribution to A,B, and C are small and subtractable



Isolation fraction=

Energy in a cone of radius 0.4 around lepton candidate not including the lepton candidate / Energy of lepton candidate

ABCD methods are really just on-off methods where **τ is measured using data samples**

"ABCD" Methods

Advantages

- Purely data based, good if you don't trust the simulation
- Model assumptions are injected by hand and not in a complicated Monte Carlo program (mostly)
- Model assumptions are intuitive

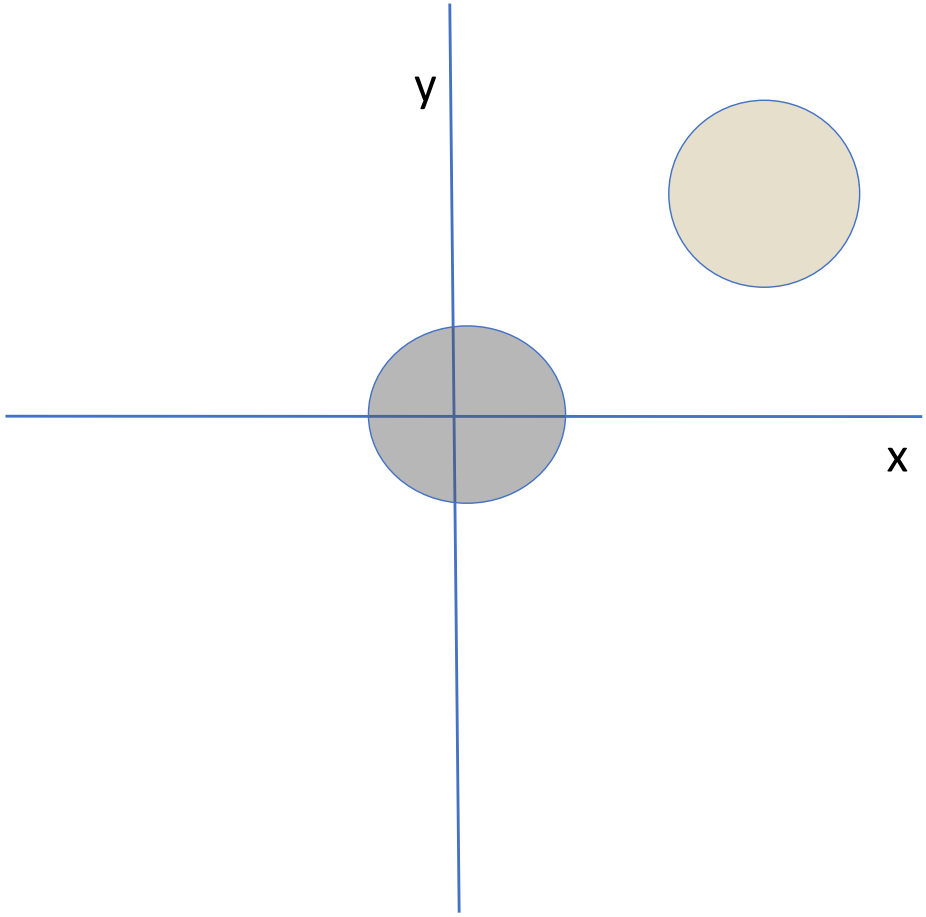
Disadvantages

- The lack of correlation between MET and ISO assumption may be false. e.g., semileptonic B decays produce unisolated leptons and MET from the neutrinos.
- Even a two-component background can be correlated when the contributions aren't by themselves.
- Another way of saying that extrapolations are to be checked/assigned sufficient uncertainty
- Works best when there are many events in regions A,B, and C. Otherwise all the problems of low stats in the "Off" sample in the On/Off problem reappear here. Large numbers of events → Gaussian approximation to uncertainty in background in D
- Requires subtraction of signal from data in regions A, B, and C → introduces model dependence
- Worse, the signal subtraction from the sidebands depends on the signal rate being measured/tested.

→ A small effect if s/b in the sidebands is small

→ You can iterate the measurement and it will converge quickly

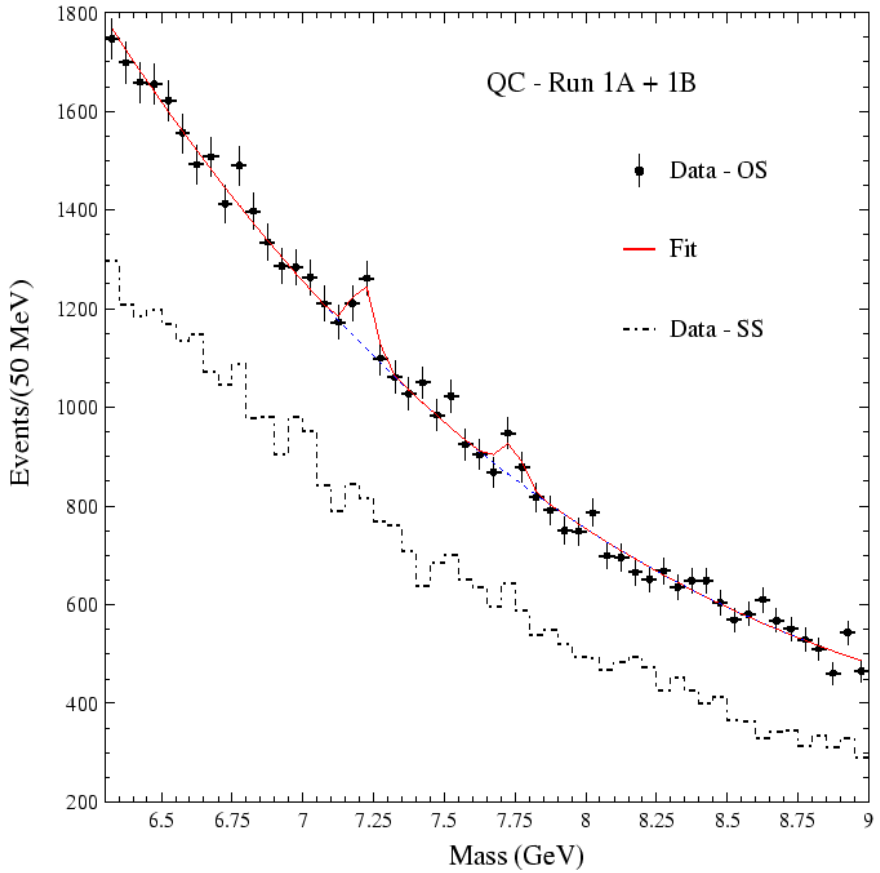
The Sum of Uncorrelated 2D Distributions may be Correlated



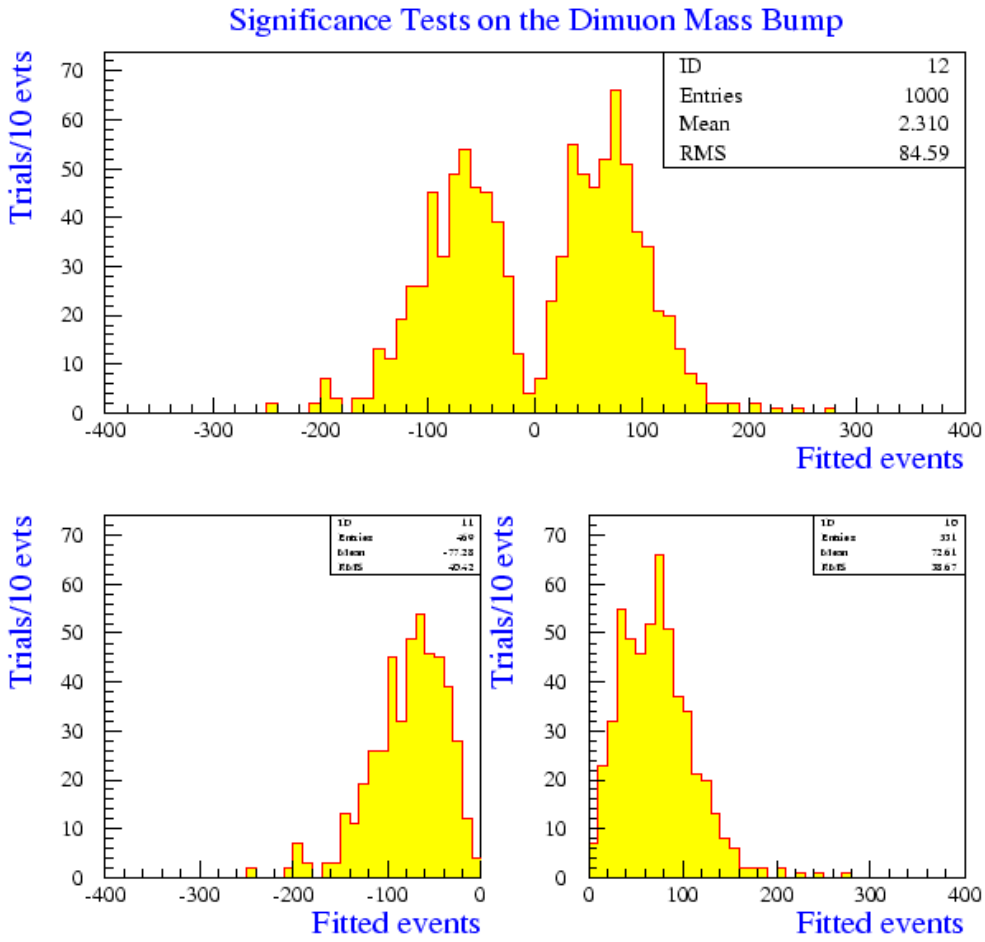
Knowledge of one variable helps identify which sample the event came from and thus helps predict the other variable's value even if the individual samples have no covariance.

An internal CDF study that didn't make it to prime time – dimuon mass spectrum with signal fit

(not enough PE's)



249.7±60.9 events fit in bigger signal peak (4σ? No!)



Null hypothesis pseudoexperiments with largest peak fit values

Looks like a lot of spectra in S. Stone's article

A Useful Tip about Limits

It takes almost exactly 3 expected signal events to exclude a model.

If you have zero events observed, zero expected background, then the limit will be 3 signal events.

$$p_{Poiss}(n=0,r) = \frac{r^0 e^{-r}}{0!} = e^{-r}$$

If p=0.05, then r=-ln(0.05)=2.99573

You can discover with just one event and very low background, however!
Example: The Ω^- discovery with a single bubble-chamber picture.

Cut and count analysis optimization usually cannot be done simultaneously for limits and discovery.

But MVA's take advantage of all categories of s/b and remain optimal in both cases; but you have to use the entire MVA distribution

Extending Our Useful Tip About Limits

It takes almost exactly 3 expected signal events to exclude a model.

If you have zero events observed, zero expected background, and no systematic uncertainties, then the limit will be 3 signal events.

Call s=expected signal, b=expected background. r=s+b is the total prediction.

$$L(n=0,r)=\frac{r^0 e^{-r}}{0!}=e^{-r}=e^{-(s+b)}$$

$$0.95=\frac{\int_0^{r_{\lim}} L'(data\,|\,r)\pi(r)dr}{\int_0^{\infty} L'(data\,|\,r)\pi(r)dr}=\frac{-e^{-(s+b)}\Big|_0^{r_{\lim}}}{-e^{-(s+b)}\Big|_0^{\infty}}=e^{-r_{\lim}}$$

The background rate cancels! For 0 observed events, the signal limit does not depend on the predicted background (or its uncertainty). This is also true for CL_s limits, but not PCL limits (which get stronger with more background)

If p=0.05, then r=-ln(0.05)=2.99573

Rule of three

From Wikipedia, the free encyclopedia

Rule of three may refer to:

Science and technology [edit]

- Rule of three (C++ programming), a rule of thumb about class method definitions
- Rule of three (computer programming), a rule of thumb about code refactoring
- Rule of three (mathematics), a method in arithmetic
- Rule of three (medicinal chemistry), a rule of thumb for lead-like compounds
- Rule of three (statistics), for calculating a confidence limit when no events have been observed

Other [edit]

- Rule of three (aviation), a rule of descent in aviation
- Rule of three (economics), a rule of thumb about major competitors in a free market
- Rule of threes (survival), a quick reference for how long one can survive in an emergency situation
- Rule of Three (Wicca), a tenet of Wicca
- Rule of three (writing), a principle of writing
- *Rule of Three*, a series of one-act plays by Agatha Christie
- The Bellman's Rule of Three in *The Hunting of the Snark*, a poem by Lewis Carroll

See also [edit]

- Three-sigma rule, for a normal distribution in statistics
- Triumvirate, a political regime dominated by three powerful individuals
- Rule of thirds, a compositional rule of thumb in photography
- Rule of thirds (diving), a rule of thumb for scuba divers
- Rule of thirds (military), a rule of thumb regarding the distribution of available manpower

Ambiguous or Missing Data

A “Unicorn”