Data Analysis and Statistical Methods in Experimental Particle Physics

Thomas R. Junk
Fermilab

UD0 Lecture Series 2012
November 8, 2012

Lecture Series Outline

Lecture 1: Introduction. Gaussian Approximations

Lecture 2: Data Analysis and Systematic Uncertainties

Lecture 3: Hypothesis Testing, Confidence Intervals

Lecture 4: Bayesian Inference, Miscellaneous Topics:
Binning, Smoothing, Unfolding

Lecture 1:

- Introduction
- Probability and Statistics
- Collider Experiments
- Common Conventions
- Likelihood Fits
- Goodness of Fit Tests

Useful 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.

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

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

(PHYSTAT2011)

I am also very impressed with the quality and thoroughness of Wikipedia articles on general statistical matters.

Figures of Merit

Our jobs as scientists are to

- **Measure quantities as precisely as we can**
  Figure of merit: the **uncertainty** on the measurement

- **Discover new particles and phenomena**
  Figure of merit: the **significance** of evidence or observation -- try to be first!
  Related: the **limit** on a new process

To be counterbalanced by:

- Integrity: All sources of systematic uncertainty must be included in the interpretation.
- Large collaborations and peer review help to identify and assess systematic uncertainty

Figures of Merit

Our jobs as scientists are to

- **Measure quantities as precisely as we can**
  Figure of merit: the **expected** uncertainty on the measurement

- **Discover new particles and phenomena**
  Figure of merit: the **expected** significance of evidence or observation -- try to be first!
  Related: the **expected** limit on a new process

To be counterbalanced by:

- Integrity: All sources of systematic uncertainty must be included in the interpretation.
- Large collaborations and peer review help to identify and assess systematic uncertainty

Expected Sensitivity is used in Experiment and Analysis Design

Probability and Statistics

Statistics is largely the inverse problem of Probability

Probability: Know parameters of the theory → Predict distributions of possible experiment outcomes

Statistics: Know the outcome of an experiment → Extract information about the parameters and/or the theory

Probability is the easier of the two -- solid mathematical arguments can be made.

Statistics is what we need as scientists. Much work done in the 20th century by statisticians.

Experimental particle physicists rediscovered much of that work in the last two decades.

In HEP we often have complex issues because we know so much about our data and need to incorporate all of what we know

Fermilab from the Air

Tevatron
ring radius=1 km

Protons on
antiprotons

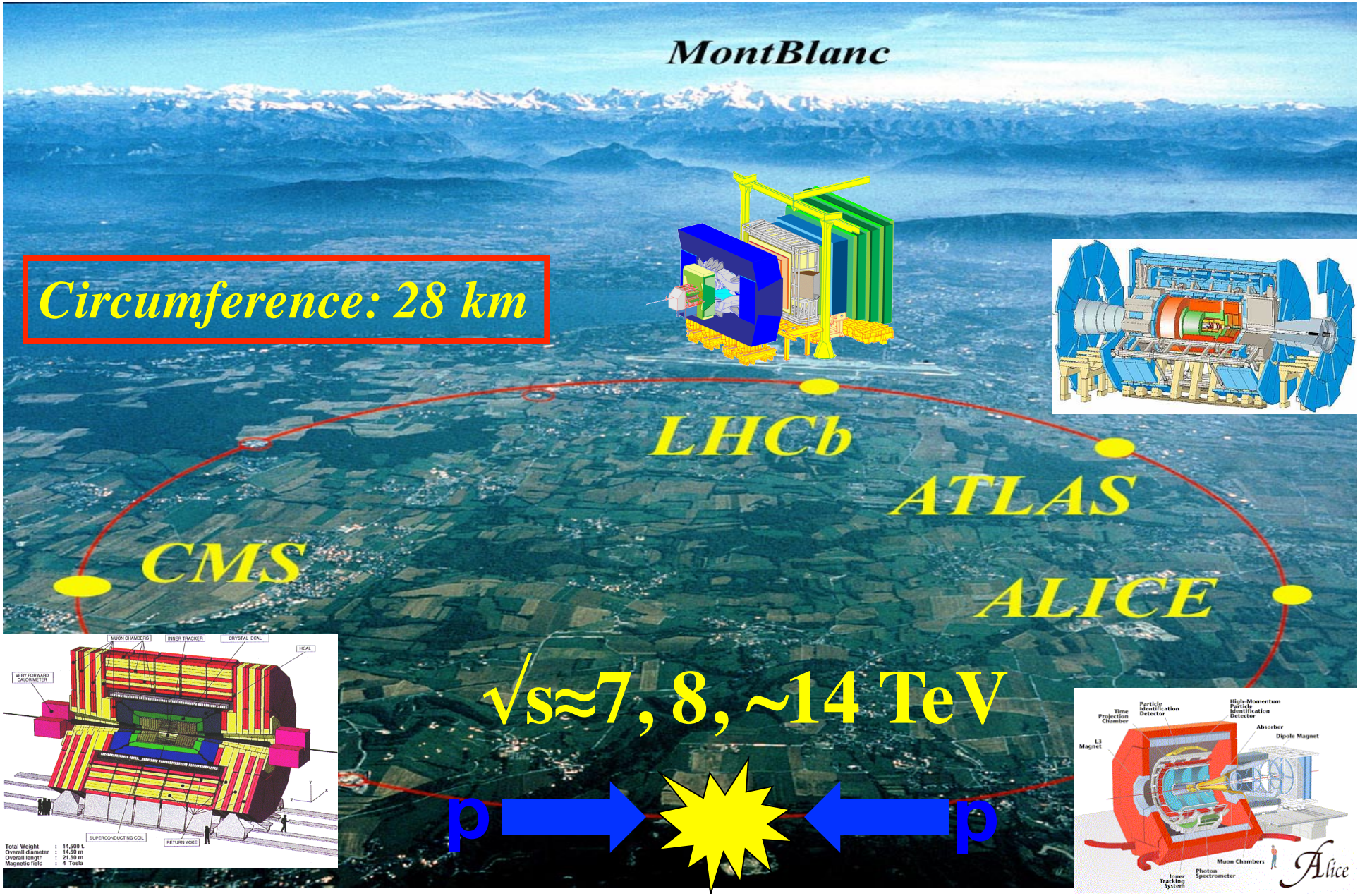
$$\sqrt{s_{p\bar{p}}} = 1.96 \text{ TeV}$$

Main Injector
commissioned in 2002

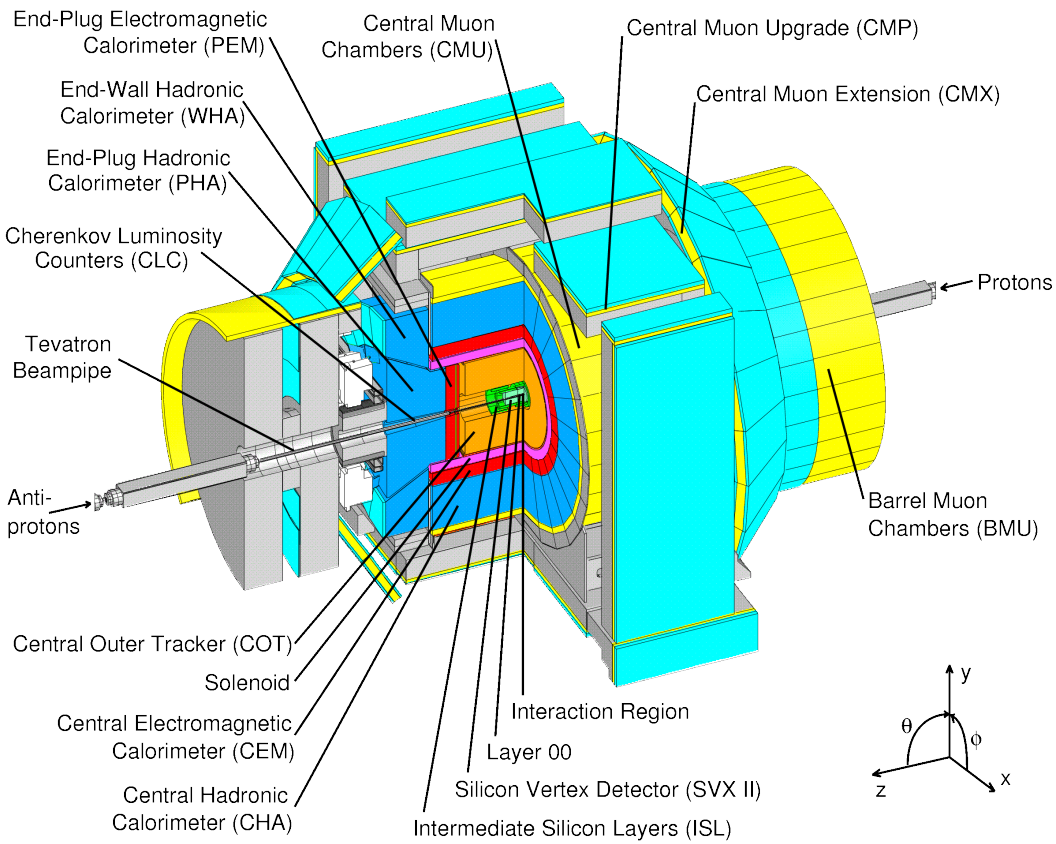
Recycler used
as another antiproton
accumulator

Start-of-store luminosities
exceeding 200×10^{30} now
are routine

The Large Hadron Collider



A Typical Collider Experimental Setup

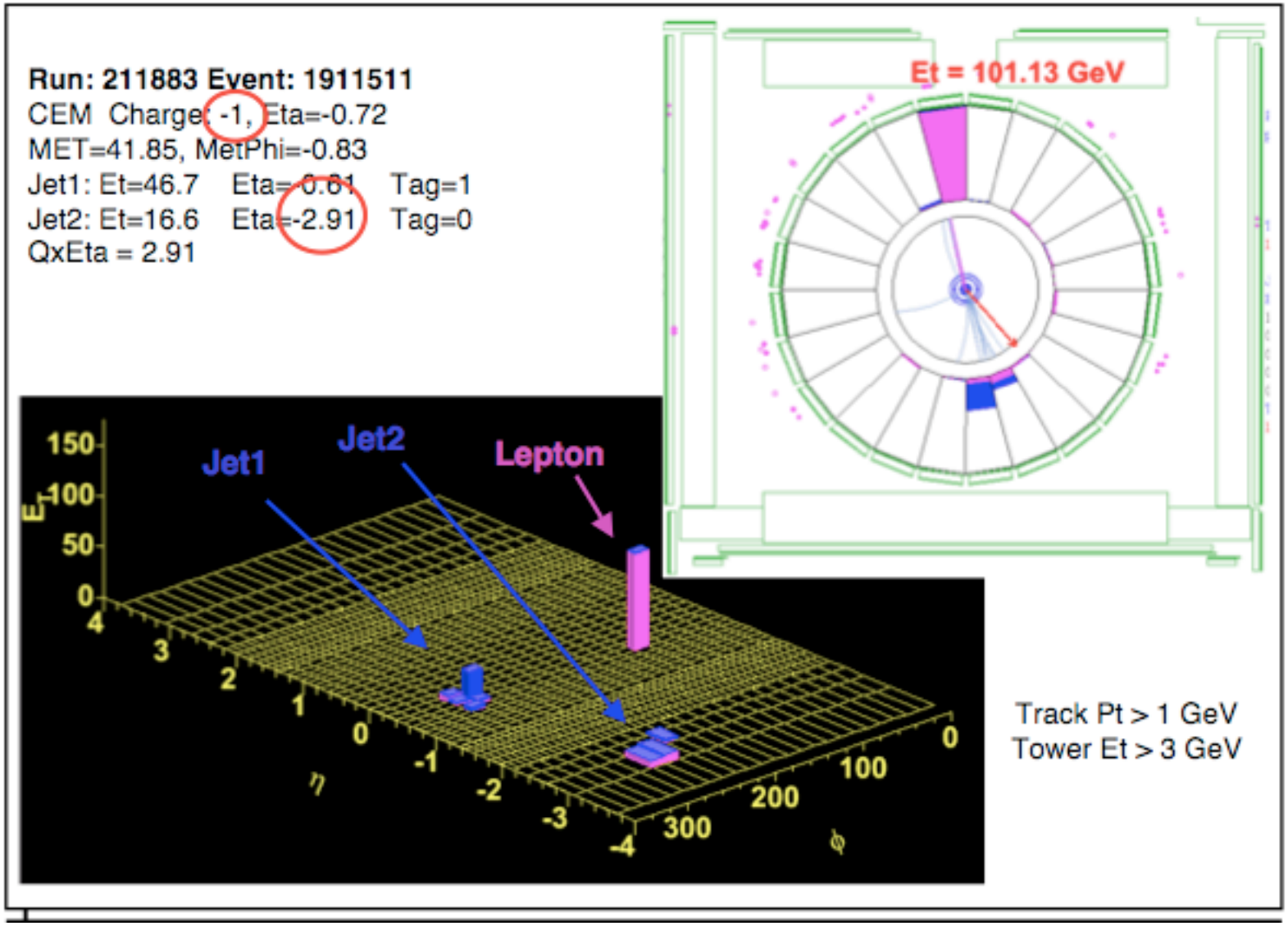


Counter-rotating beams of particles (protons and antiprotons at the Tevatron)
Bunches cross every 396 ns.

Detector consists of tracking, calorimetry, and muon-detection systems.

An online trigger selects a small fraction of the beam crossings for further storage.
Analysis cuts select a subset of triggered collisions.

An Example Event Collected by CDF (a single top candidate)



Some Probability Distributions useful in HEP

Binomial:

Given a repeated set of *N* trials, each of which has probability *p* of "success" and 1 - *p* of "failure", what is the distribution of the number of successes if the *N* trials are repeated over and over?

$$\mathrm{Binom}(k \mid N, p) = \binom{N}{k} p^k (1-p)^{N-k}, \quad \sigma(k) = \sqrt{\mathrm{Var}(k)} = \sqrt{Np(1-p)}$$

k is the number of "success" trials

Example: events passing a selection cut, with a fixed total *N*

The Gaussian approximate formula I remember:
Measuring p with observe k out of N (e.g. acceptance measurements)

$$p = k/N \qquad \sigma_p = \sqrt{\frac{p(1-p)}{N}}$$

Binomial Distributions in HEP

Formally, all distributions of event counts are really binomial distributions

→ The number of protons in a bunch (and antiprotons) is finite
→ The beam crossing count is finite

So whether an event is triggered and selected is a success or fail decision.

But – there are ~5×10^{13} bunch crossings if we run all year, and each bunch crossing has ~ 10^{10} protons that can collide. We trigger only 200 events/second, and usually select a tiny fraction of those events.

The limiting case of a binomial distribution with a small acceptance probability is Poisson.

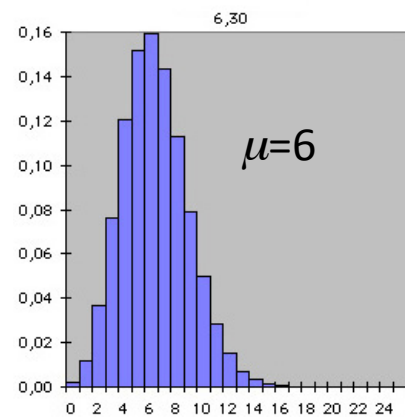
Useful for radioactive decay (large sample of atoms which can decay, small decay rate).

A case in which Poisson is not a good estimate for the underlying distribution of event counts: A saturated trigger (trigger on each beam crossing for example). – DAQ runs at its rate limit, producing a fixed number of events/second (if there is no beam).
To discuss later if we have time: "The Stopping Problem"

The Poisson Distribution

Limit of Binomial when $N \to \infty$ and $p \to 0$ with $Np = \mu$ finite

$$\mathrm{Poiss}(k \mid \mu) = \frac{e^{-\mu}\mu^k}{k!} \qquad \sigma(k) = \sqrt{\mu}$$



Normalized to unit area in two different senses

$$\sum_{k=0}^{\infty} \mathrm{Poiss}(k \mid \mu) = 1, \quad \forall \mu$$

$$\int_0^{\infty} \mathrm{Poiss}(k \mid \mu) d\mu = 1 \qquad \forall k$$

The Poisson distribution is assumed for all event counting results in HEP.

Composition of Poisson and Binomial Distributions

Example: Efficiency of a cut, say lepton p_T in leptonic W decay events at the Tevatron

Total number of W bosons: N -- Poisson distributed with mean μ

The number passing the lepton p_T cut: k

Repeat the experiment many times. ***Condition*** on N (that is, insist N is the same and discard all other trials with different N. Or just stop taking data).

$p(k) = \mathrm{Binom}(k|N,\varepsilon)$ where ε is the efficiency of the cut

Composition of Poisson and Binomial Distributions

If we no longer condition on N ($\mu = \sigma L$):

The number of W events passing the cut is just another counting experiment -- it must be Poisson distributed.

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

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

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

The sum of two Poisson-distributed numbers is Poisson-distributed with the sum of the means

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

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?

Application to a test of Poisson Ratios

Our composition formula from the previous page:

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

Say you have n_s in the "signal" region of a search, and n_c in a "control" region -- example: peak and sidebands

n_s is distributed as Poiss(s+b)
n_c is distributed as Poiss(τb)

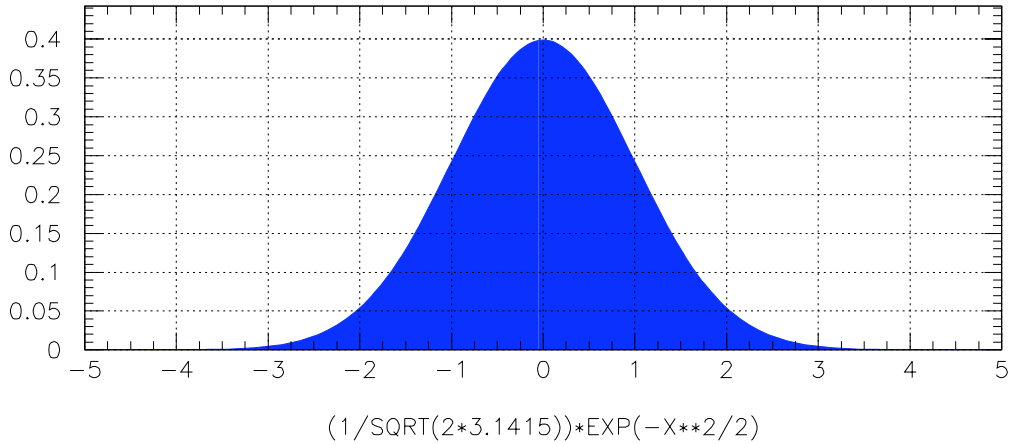
Suppose we want to test H_0: s=0. Then $n_s/(n_s+n_c)$ is a variable that measures $1/(1+\tau)$

The control region could be the weak link in the interpretation!

Another Probability Distribution useful in HEP

Gaussian:

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



It's a parabola on a log scale.

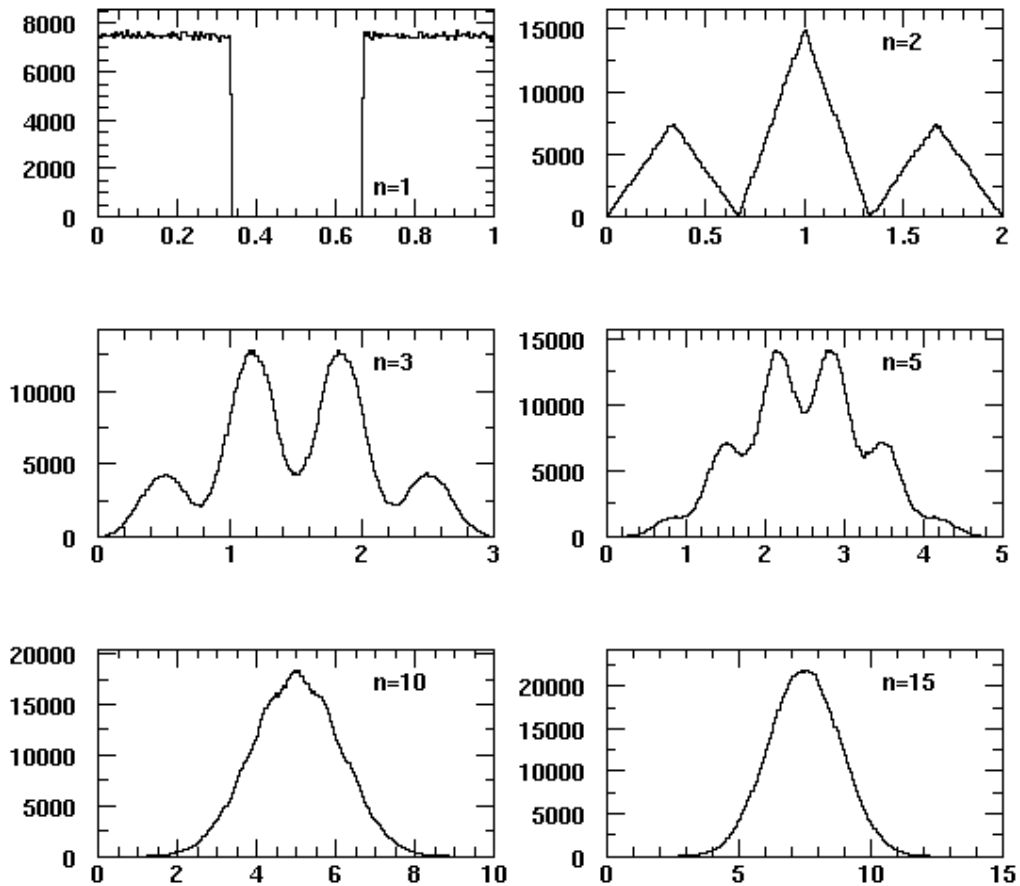
Sum of Two Independent Gaussian Distributed Numbers is Gaussian with the sum of the means and the sum in quadrature of the widths

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

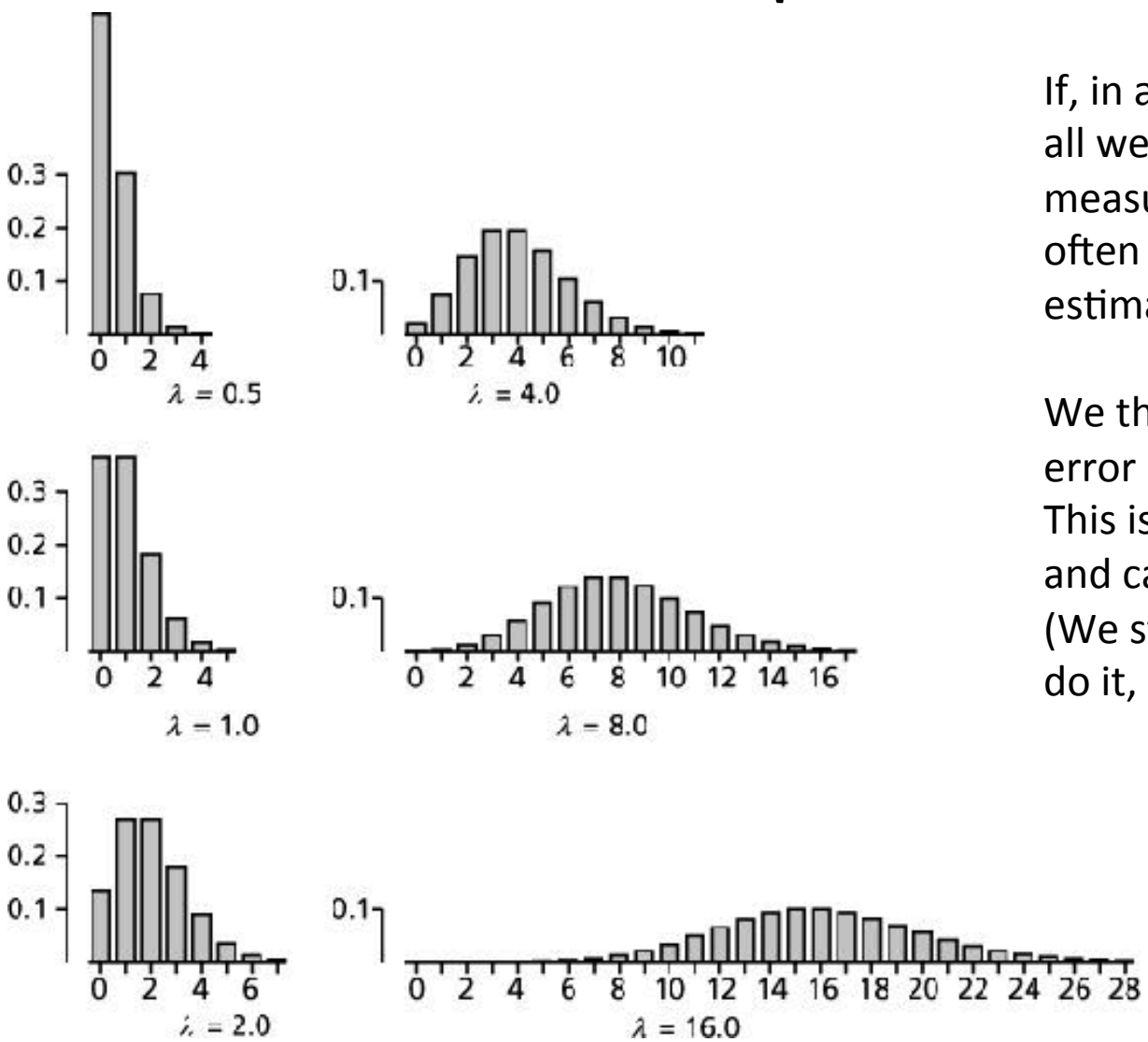
A difference of independent Gaussian-distributed numbers is also Gaussian distributed (widths still *add* in quadrature)

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 μ is Approximately Gaussian of width $\sigma = \sqrt{\mu}$



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

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

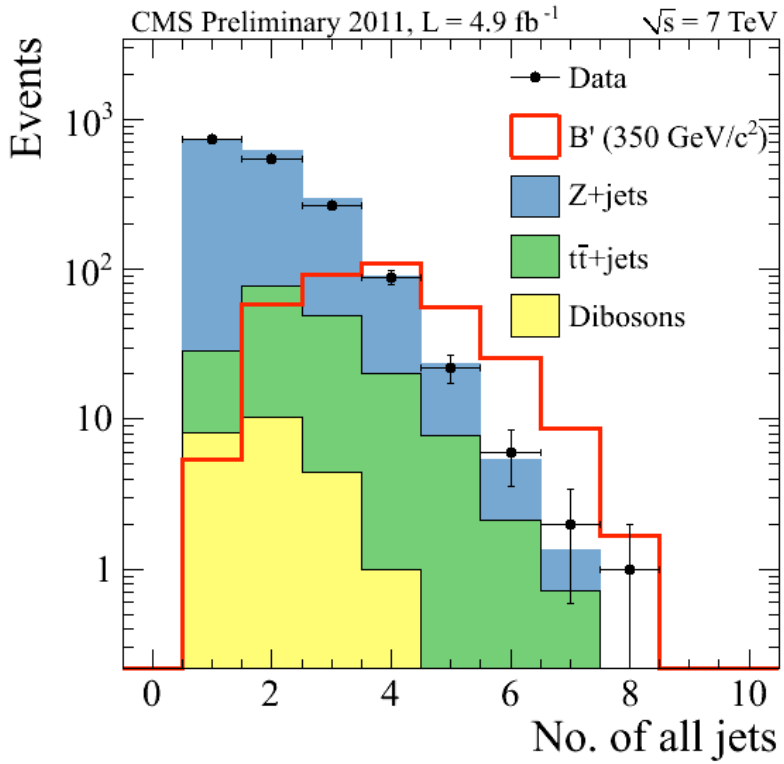
Why Put Error Bars on the Data?

- 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: $\sqrt{n} \neq \sqrt{\mu}$

The true value of μ 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.

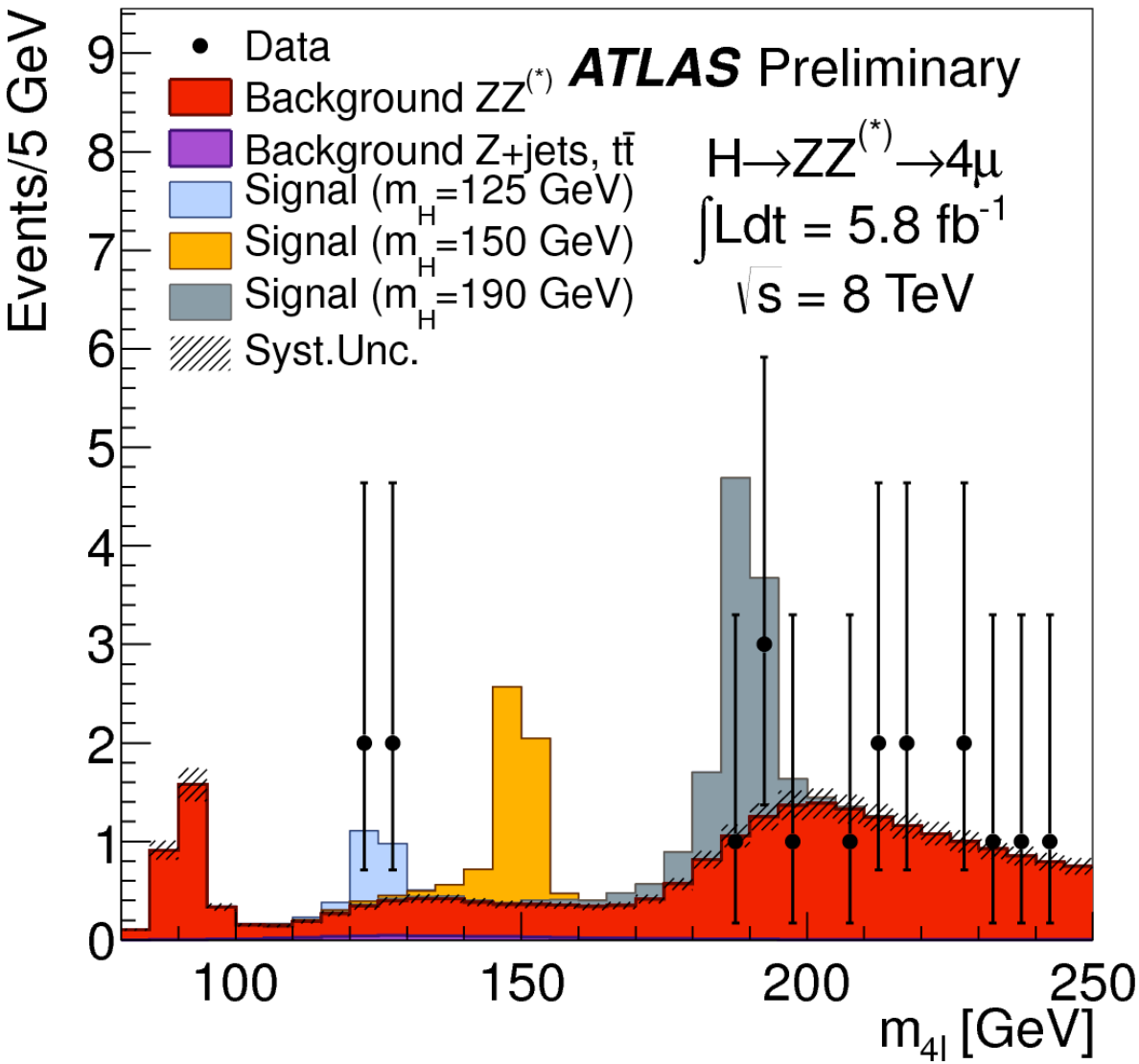
Correct presentation of data and predictions

Table 3: The numbers of expected signal ($m_H = 125$ GeV) and background events, together with the numbers of observed events in the data, in a window of size ±5 GeV around 125 GeV, for the combined $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV data.

	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



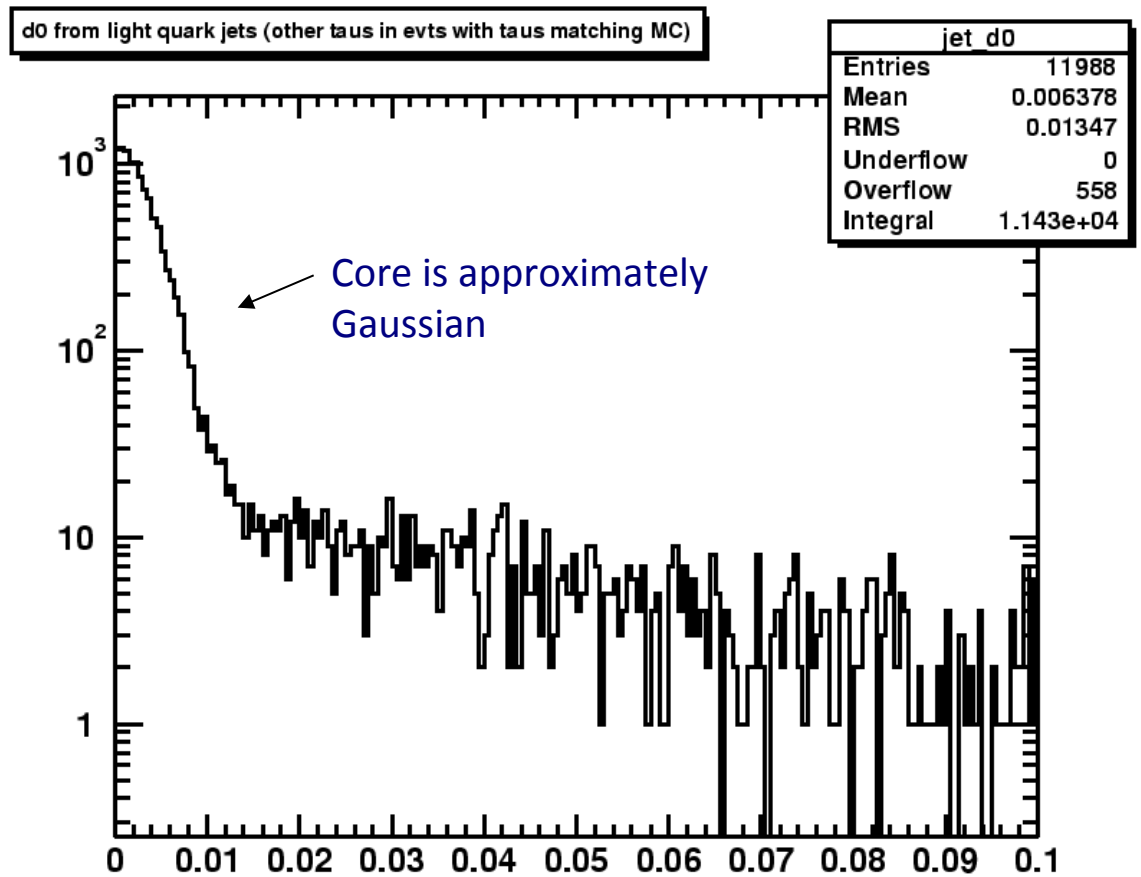
But there are several options.

Need to explain which one is chosen

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."

Different Meanings of the Idea "Statistical Uncertainty"

- Repeating the experiment, how much would we expect the answer to fluctuate?

  -- approximate, Gaussian

- What interval contains 68% of our belief in the parameter(s)?
  Bayesian credibility intervals

- What construction method yields intervals containing the true value 68% of the time?
  Frequentist confidence intervals

In the limit that all distributions are symmetric Gaussians, these look like each other. We will be more precise later.

Why Uncertainties Add in Quadrature

Common situation -- a prediction is a sum of uncertain components, or a measured parameter is a sum of data with a random error, and an uncertain prediction

"statistical" "systematic"

e.g., Cross-Section = (Data-Background)/(A*ε*Luminosity) where Background, Acceptance and Luminosity are obtained somehow from other measurements and models.

Probability distribution of a sum of Gaussian distributed random numbers is Gaussian with a sum of means and a sum of variances.

$$\int_{-\infty}^{\infty} g(x,\mu_1,\sigma_1)g(x'-x,\mu_2,\sigma_2)dx = g(x',\mu_1+\mu_2,\sqrt{\sigma_1^2+\sigma_2^2})$$

Convolution assumes variables are independent.

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

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

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

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

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

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

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:

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

The Variation of the Width Estimate

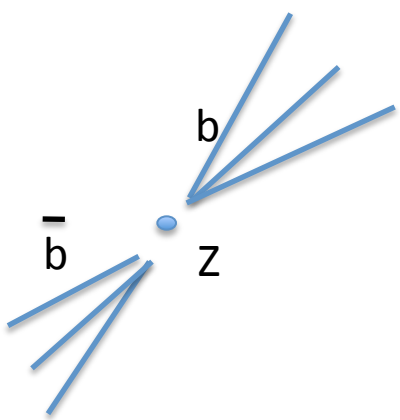
I got this formula from the Particle Data Group's Statistics Review
-- G. Cowan made the most recent revision

For Gaussian distributed numbers, the variance of σ^2_{est} is $2\sigma^4/(N-1)$.

The standard deviation of σ_{est} is therefore

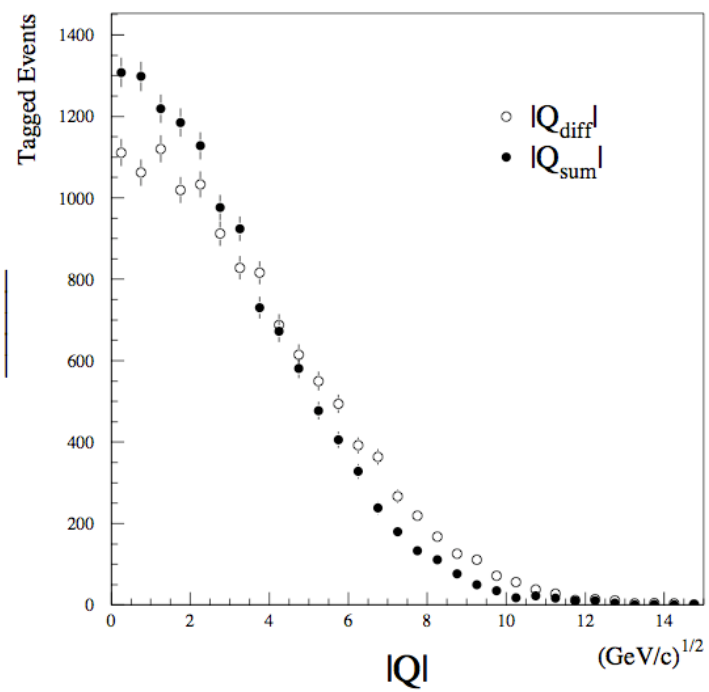
$$\sigma / \sqrt{2(N-1)}$$

I once had to use this formula for my thesis. Momentum-weighted charge determination of Z decay to bbbar in e^+e^- collisions



$$|Q_{diff}| = |Q| = \left| \sum_{tracks} q_i |\vec{p}_i \cdot \vec{t}|^{\kappa} \mathrm{sgn}(\vec{p}_i \cdot \vec{t}) \right|$$

$$Q_{sum} = \sum_{tracks} q_i |\vec{p}_i \cdot \vec{t}|^{\kappa}$$



Uncertainties That Don't Add in Quadrature

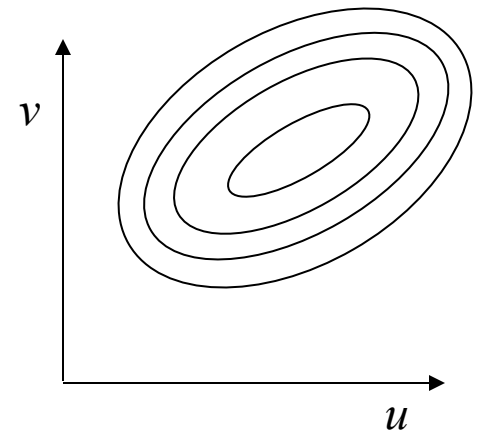
Some may be correlated! (or partially correlated). Doubling a random variable with a Gaussian distribution doubles its width instead of multiplying by $\sqrt{2}$

Example: The same luminosity uncertainty affects background prediction for many different background sources in a sum. The luminosity uncertainties all add linearly. Other uncertainties (like MC statistics) may add in quadrature or linearly.

Strategy: Make a list of independent sources of uncertainty -- these each may enter your analysis more than once. Treat each error source as independent, not each way they enter the analysis. Parameters describing the sources of uncertainty are called nuisance parameters (distinguish from parameter of interest)

Propagation of Uncertainties

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



If

$$x = au + bv$$

then

$$\sigma_x^2 = a^2\sigma_u^2 + b^2\sigma_v^2 + 2ab\sigma_{uv}^2$$

This can even vanish! (anticorrelation)

In general, if

$$x = f(u, v)$$

$$\sigma_x^2 = \left(\frac{\partial x}{\partial u}\right)^2 \sigma_u^2 + \left(\frac{\partial x}{\partial v}\right)^2 \sigma_v^2 + 2\left(\frac{\partial x}{\partial u}\right)\left(\frac{\partial x}{\partial v}\right)\sigma_{uv}^2$$

Relative and Absolute Uncertainties

If $x = auv$

then $\sigma_x^2 = a^2 v^2 \sigma_u^2 + a^2 u^2 \sigma_v^2 + 2a^2 uv \sigma_{uv}^2$

or, more easily, $$\frac{\sigma_x}{x} = \sqrt{\frac{\sigma_u^2}{u^2} + \frac{\sigma_v^2}{v^2} + 2\frac{\sigma_{uv}^2}{uv}}$$

"**relative uncertainties add in quadrature**" for multiplicative uncertainties (but watch out for correlations!)

The same formula holds for division (!) but with a minus sign in the correlation term.

Tip: I can never remember the correlation terms here! I always seek an uncorrelated basis to represent uncertainties, and derive what I have to.

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

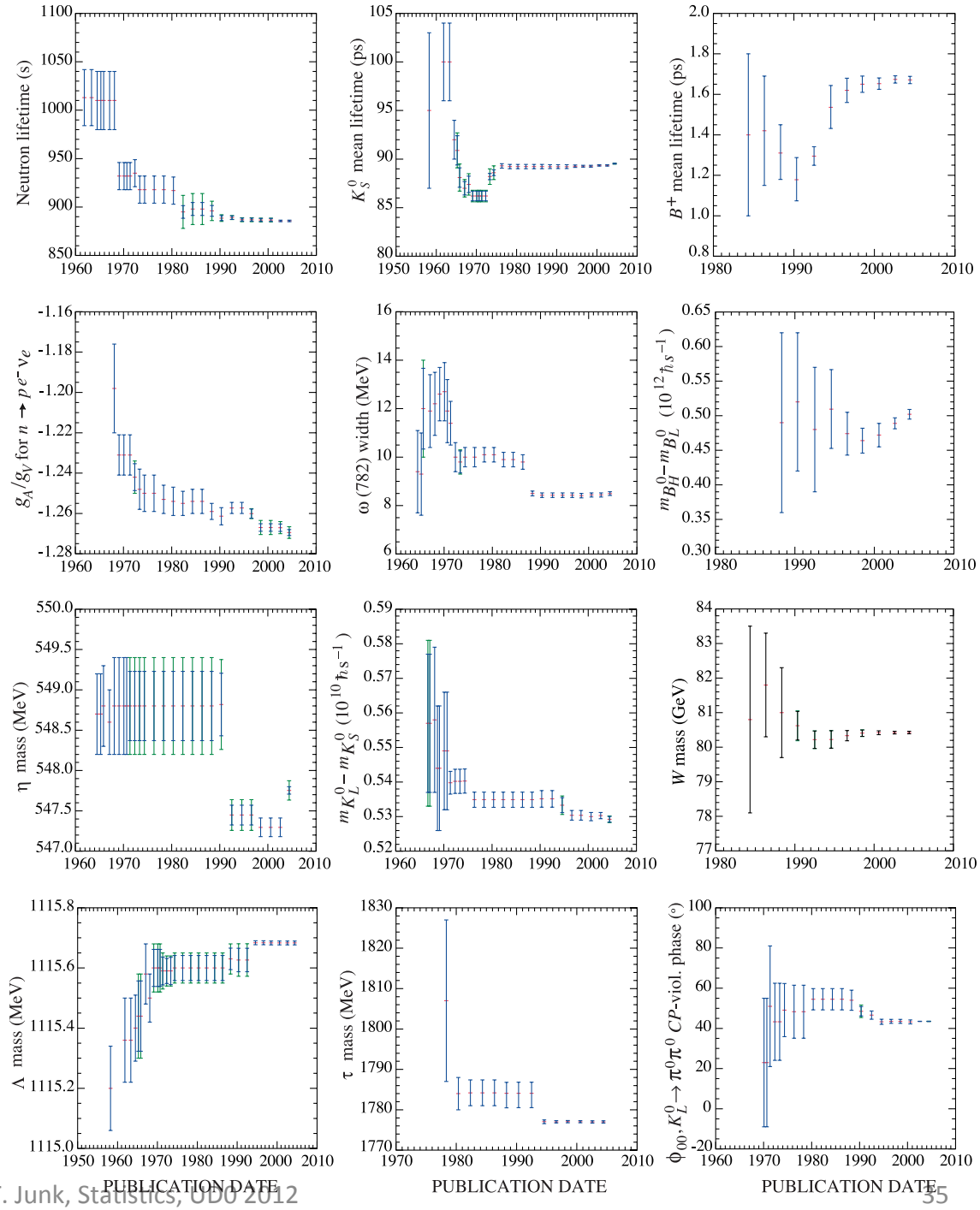
- Analyses need to be optimized -- shoot for the lowest uncertainty

- 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.

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

Examples from the front of the PDG



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

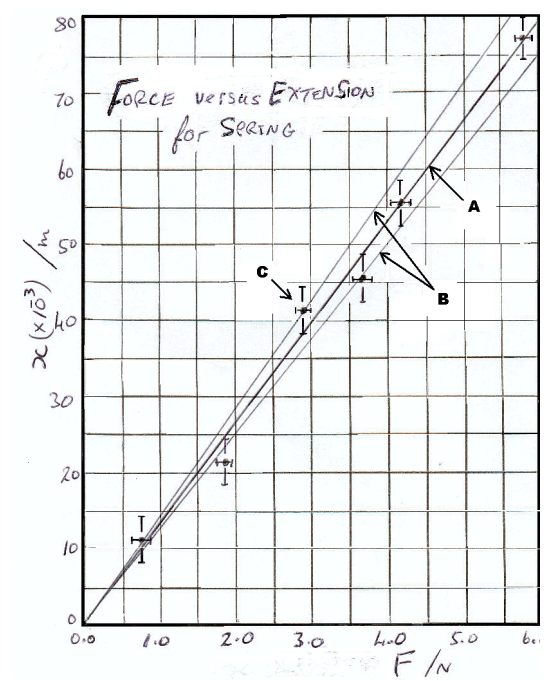
If the μ_i are predicted by a model with free parameters (e.g. a straight line), and χ^2 is minimized over all values of the free parameters, then

Approximate! Not always!
(*cough*)

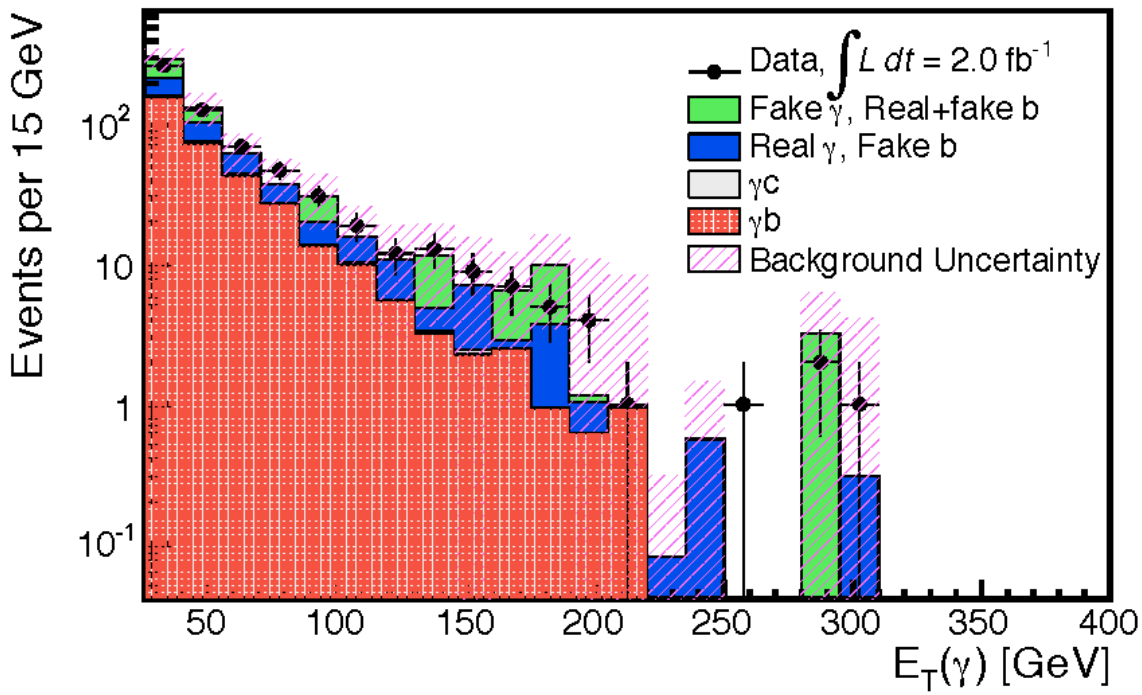
DOF = n - #free parameters in fit.

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

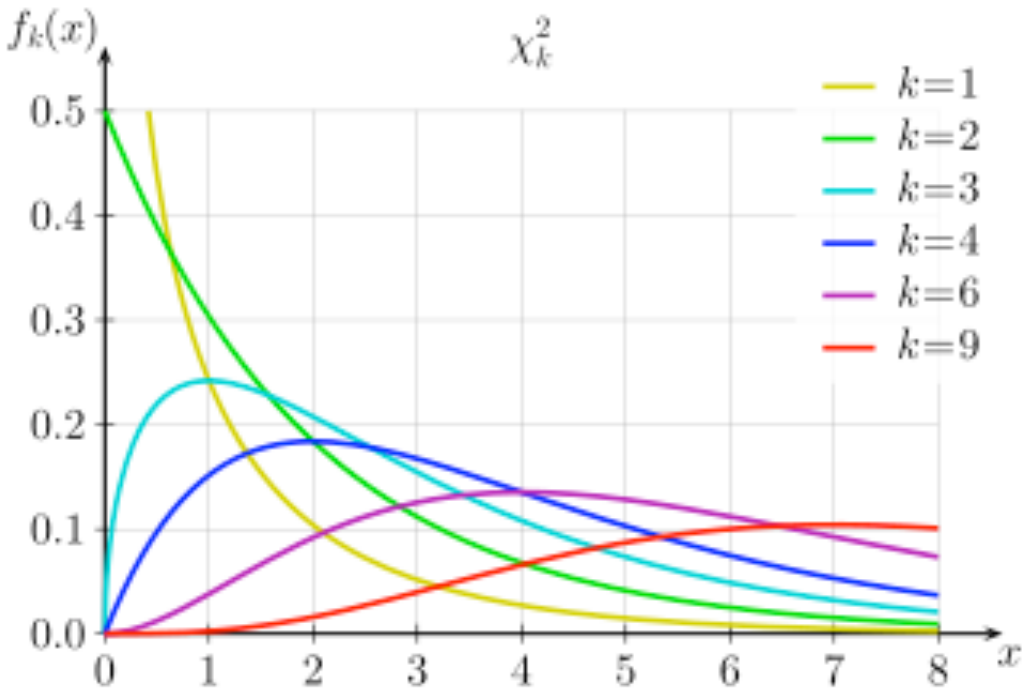


NDOF=?

- Limit calculators 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. Limit calculators should ignore ROOT's u/o bins).

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

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)

CERNLIB routine: PROB.

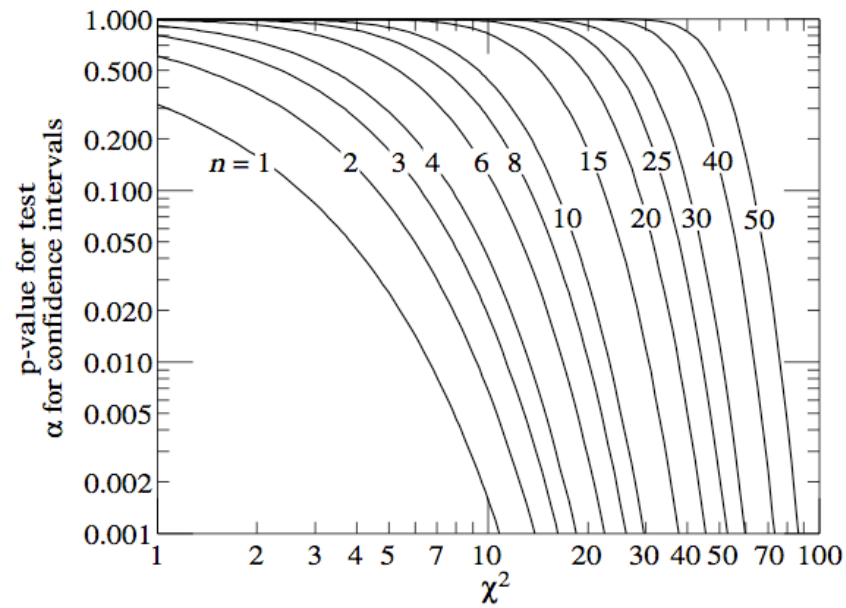
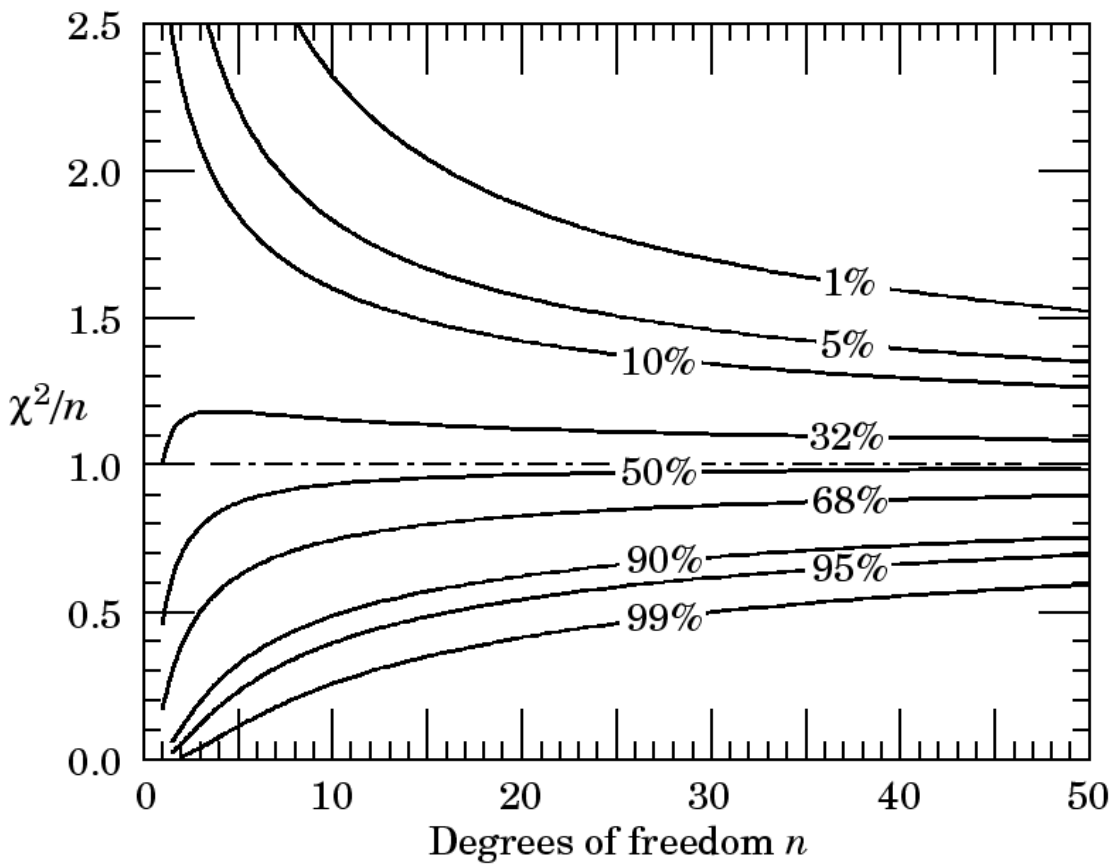


Figure 33.1: One minus the χ^2 cumulative distribution, $1 - F(\chi^2; n)$, for n degrees of freedom. This gives the p-value for the χ^2 goodness-of-fit test as well as one minus the coverage probability for confidence regions (see Sec. 33.3.2.4).

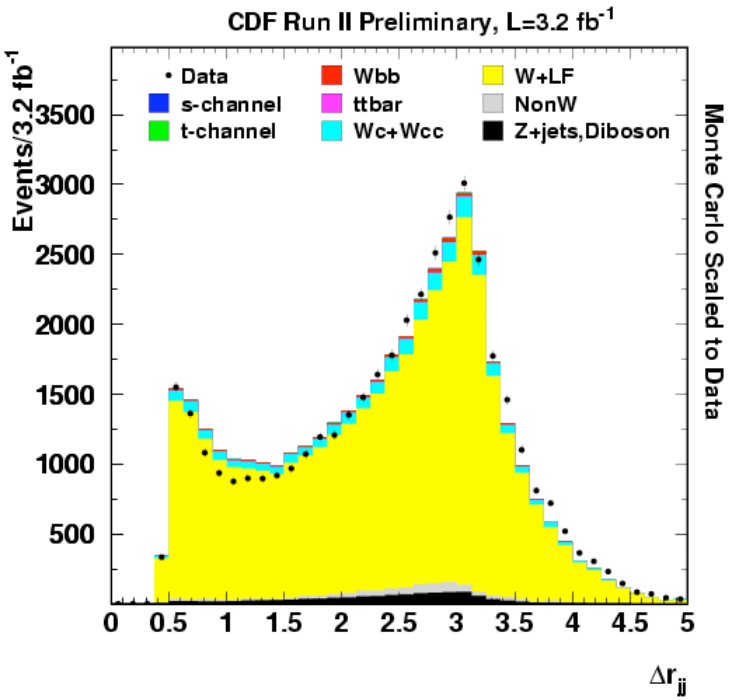
A Rule of Thumb Concerning χ^2

Average contribution to χ^2 per DOF is 1. χ^2/DOF converges to 1 for large n



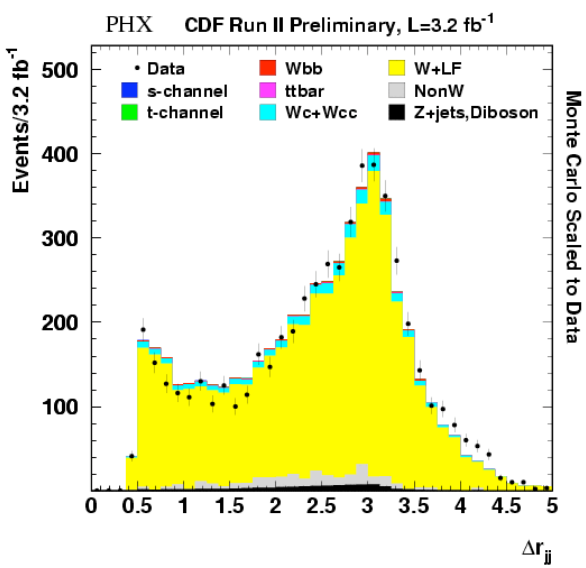
From the PDG Statistics Review

Chisquared Tests for Large Data Samples

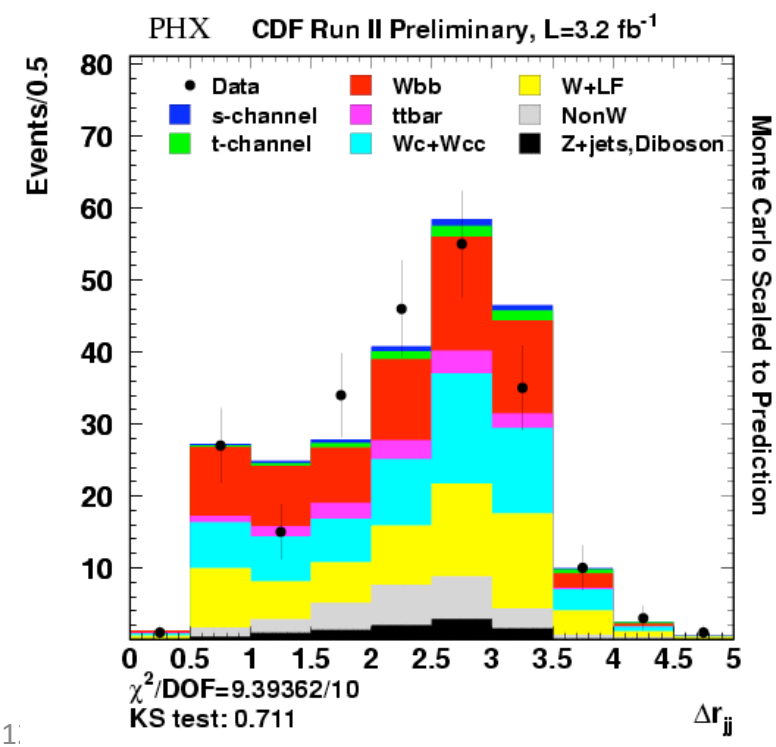


A large value of χ^2/DOF -- p-value is microscopic. We are very very sure that our model is slightly wrong. With a smaller data sample, this model would look fine (even though it is still wrong.

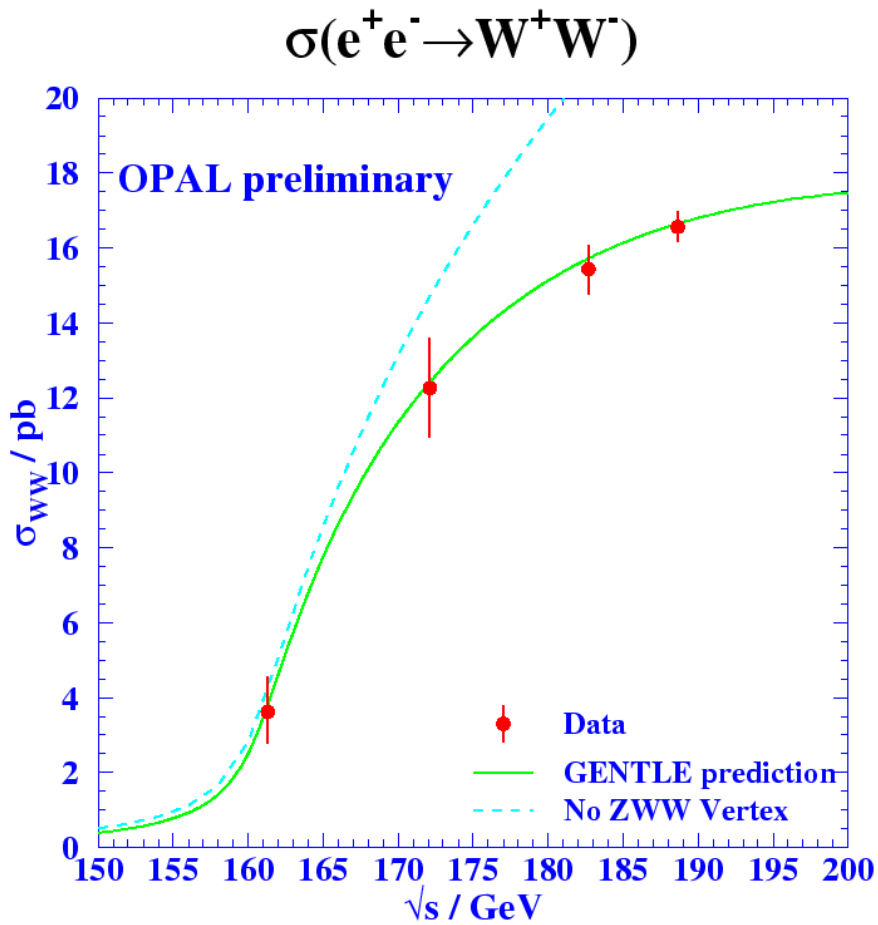
χ^2 depends on choice of binning.



Smaller data samples: harder to discern mismodeling.



χ^2 Can Sometimes be so Good as to be Suspicious

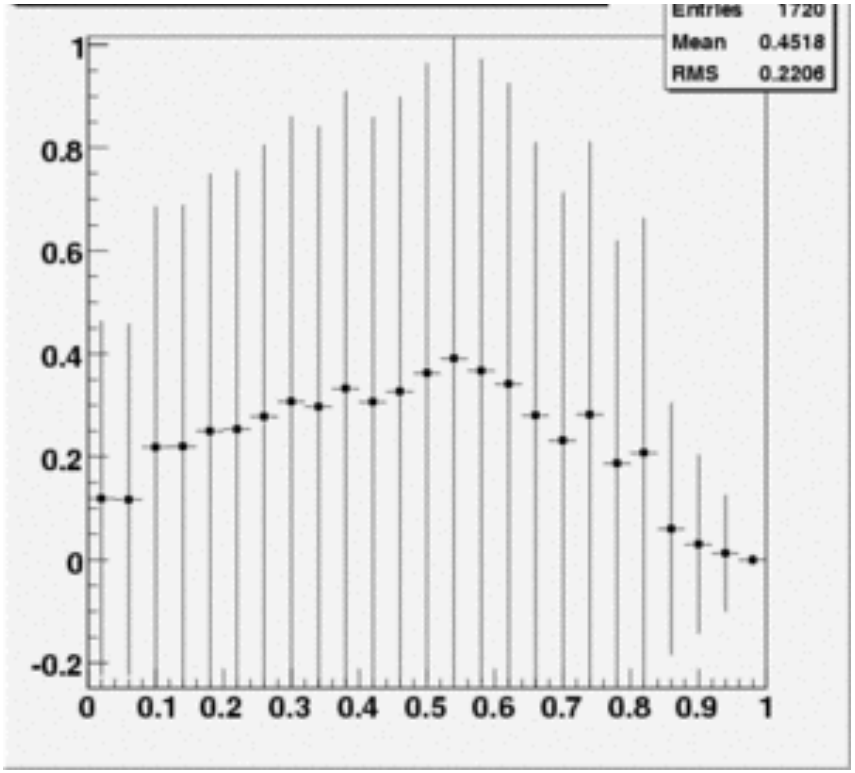


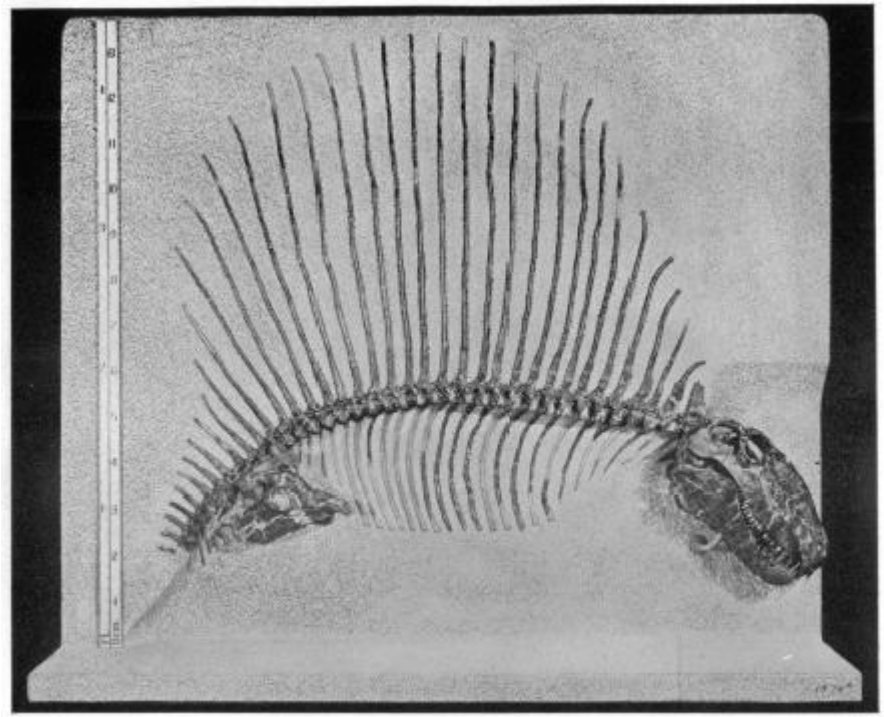
It should happen sometimes! But it is a red flag to go searching for correlated or overestimated uncertainties

no free parameters in model

(happy ending: further data points increased χ^2)

Or Really Suspicious

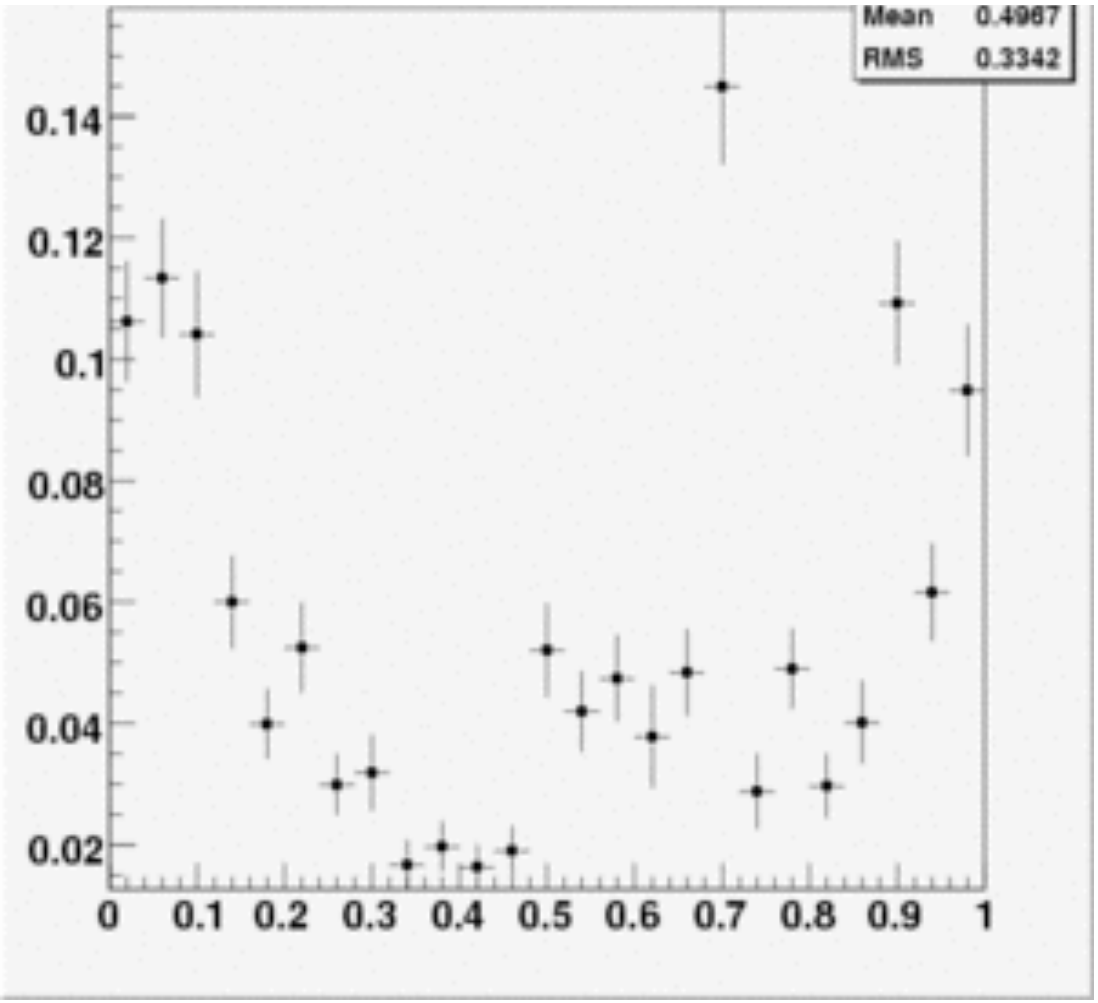




Dimetrodon incisivus, mounted skeleton, No. 4636 A. M. N. H. Mounted by A. Hermann. Greatly reduced. Scale in feet and decimetres.

Cause of the problem: Not calling TH1::Sumw2()
Uncertainties are overestimated in this case.

It can go the other way – Is this an estimate of a smooth function?



χ^2 is really bad for any smooth function

Uncertainties probably underestimated.

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)

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

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

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**

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

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

$$L = P(\text{data} \mid \vec{\nu}, \vec{\theta})$$

$\vec{\nu}$: "Parameters of Interest" mass, cross-section, b.r.
$\vec{\theta}$: "Nuisance Parameters" Luminosity, acceptance, detector resolution.

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

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

$\ln(L) = \ln(L_{max}) - s^2/2$, to quote *s*-standard-deviantion intervals. Maximize L over $\vec{\theta}$ separately for each value of $\vec{\nu}$. Buzzword: "*Profiling*"

More General Maximum Likelihood Fits

Advantages:

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

$$L = \prod_{i=1}^{Nevents} P(\vec{x}_i \mid \vec{\theta}, \vec{\nu})$$

Unbinned likelihood fits are quite popular. Just need Product over events of probability of observing each event. Overcomes information loss introduced by grouping events into bins. Unbinned likelihoods are usually poor estimates of goodness of fit though.

General Warnings about all Maximum Likelihood Fits

- 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)

Example of a problem:

Using Observed Uncertainties in Combinations Instead of Expected Uncertainties

Simple case: 100% efficiency. Count events in several subsets of the data. Measure *K* times the same integrated luminosity. each with $n_i \pm \sqrt{n_i}$

Total: $$N_{tot} = \sum_{i=1}^{K} n_i$$

Weighted average: (from **BLUE**)

Best average: $n_{avg} = N_{tot}/K$

$$n_{avg} = \frac{\sum_{i=1}^{K} n_i/\sigma_i^2}{\sum_{i=1}^{K} 1/\sigma_i^2} = \frac{\sum_{i=1}^{K} n_i/n_i}{\sum_{i=1}^{K} 1/n_i} = \frac{K}{\sum_{i=1}^{K} 1/n_i}$$

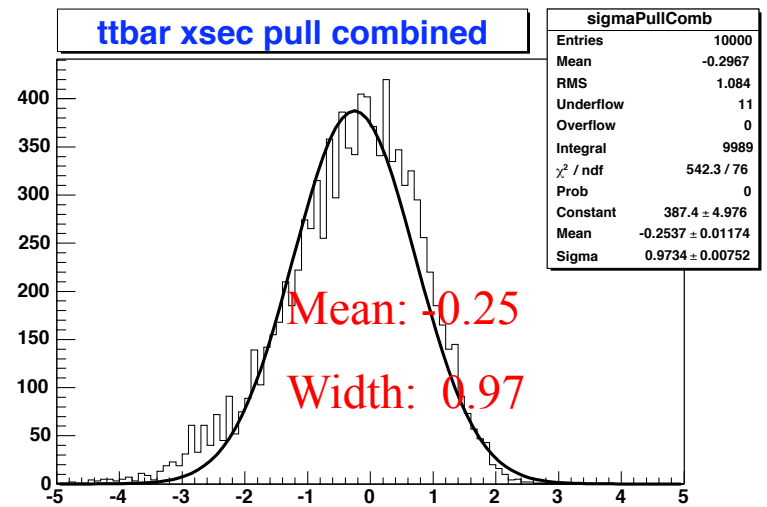
crazy behavior (especially if one of the n_i=0)

What Went Wrong?

-- low measurements have smaller uncertainties than larger measurements.

True uncertainty is the scatter in the measurements for a fixed set of true parameters

Solution: Use the expected error $\sqrt{\mu}$ for the true value of the parameter after averaging -- need to iterate!



But: Sometimes the "observed" uncertainty carries some real information! Statisticians prefer reporting "observed" uncertainties as lucky data can be more informative than unlucky data.

Example: Measuring M_Z from one event -- leptonic decay is better than hadronic decay.

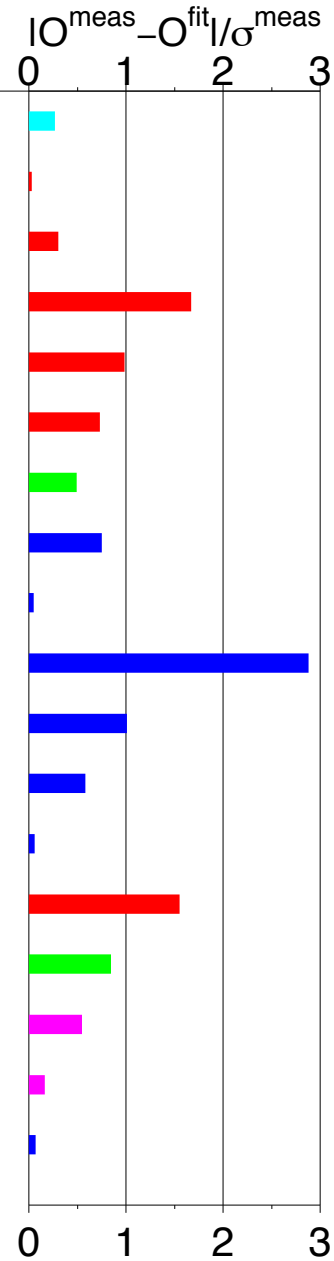
A Prominent Example of Pulls -- Global Electroweak Fit

χ^2/DOF = 18.5/13

probability = 13.8%

Didn't expect a 3σ result in 18 measurements, but then again, the total χ^2 is okay

	Measurement	Fit
$\Delta\alpha^{(5)}_{had}(m_Z)$	0.02750 ± 0.00033	0.02759
m_Z [GeV]	91.1875 ± 0.0021	91.1874
Γ_Z [GeV]	2.4952 ± 0.0023	2.4959
σ^0_{had} [nb]	41.540 ± 0.037	41.478
R_l	20.767 ± 0.025	20.742
$A^{0,l}_{fb}$	0.01714 ± 0.00095	0.01645
$A_l(P_\tau)$	0.1465 ± 0.0032	0.1481
R_b	0.21629 ± 0.00066	0.21579
R_c	0.1721 ± 0.0030	0.1723
$A^{0,b}_{fb}$	0.0992 ± 0.0016	0.1038
$A^{0,c}_{fb}$	0.0707 ± 0.0035	0.0742
A_b	0.923 ± 0.020	0.935
A_c	0.670 ± 0.027	0.668
A_l(SLD)	0.1513 ± 0.0021	0.1481
$\sin^2\theta^{lept}_{eff}(Q_{fb})$	0.2324 ± 0.0012	0.2314
m_W [GeV]	80.385 ± 0.015	80.377
Γ_W [GeV]	2.085 ± 0.042	2.092
m_t [GeV]	173.20 ± 0.90	173.26



March 2012

Bounded Physical Region

What happens if you get a best-fit value you know can't possibly be the true?

Examples:

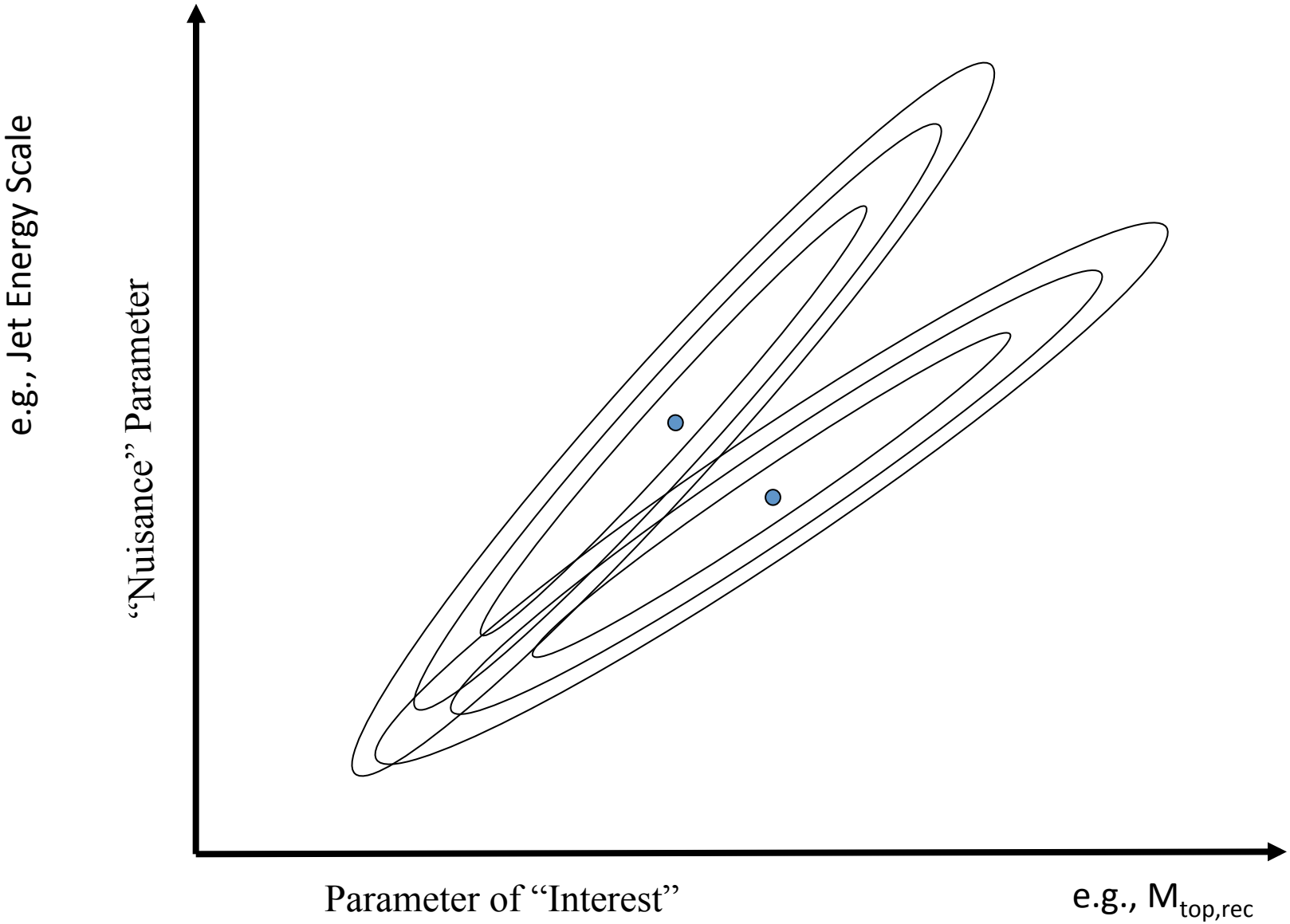
Cross Section for a signal < 0
m^2(new particle) < 0
$\sin\theta < -1$ or $> +1$

These measurements are important! You should report them without adjustment.
(but also some other things too)

An average of many measurements without these would be biased.
Example: Suppose the true cross section for a new process is zero.
Averaging in only positive or zero measurements will give a positive answer.

Later discussion: confidence intervals and limits -- take bounded physical regions into account. But they aren't good for averages, or any other kinds of combinations.

Odd Situation: BLUE Average of Two Measurements not Between the Measured Values



An Exercise: What is the Expected Difference in a Measured Value when a Cut is Tightened or Loosened?

Assume no systematic modeling problems with the variable that is being cut on.

Usually this is what we'd like to test. The result of a measurement will depend on the event selection, but it will have statistical and systematic components.

Let's estimate the statistical component.

Total measurement: $x_1 \pm \sigma_1$. (stat uncertainty only)
Tighten cuts: get: $x_2 \pm \sigma_2$.
Make a measurement in the exclusive sample (what was cut out): $x_3 \pm \sigma_3$.
Weighted averages: x_2 and x_3 are independent.

$$x_1 = \frac{\frac{x_2}{\sigma_2^2} + \frac{x_3}{\sigma_3^2}}{\frac{1}{\sigma_2^2} + \frac{1}{\sigma_3^2}} \qquad \sigma_1 = \frac{1}{\sqrt{\frac{1}{\sigma_2^2} + \frac{1}{\sigma_3^2}}}$$

An Exercise: What is the Expected Difference in a Measured Value when a Cut is Tightened or Loosened?

Would like to know what the width is of the distribution x_1-x_2 (total minus the new version with the tighter cut).

Strategy: Solve for x_1-x_2 in terms of x_2 and x_3, which are the independent variables, with independent uncertainties. Propagate the uncertainties in x_2 and x_3 to x_1-x_2.

$$x_1 - x_2 = x_2\left(\frac{\sigma_1^2}{\sigma_2^2} - 1\right) + x_3\left(\frac{\sigma_1^2}{\sigma_3^2}\right)$$

$$\sigma_{x_1 - x_2} = \sqrt{\sigma_2^2\left(\frac{\sigma_1^2}{\sigma_2^2} - 1\right)^2 + \sigma_3^2\left(\frac{\sigma_1^2}{\sigma_3^2}\right)^2}$$

And after a small amount of work, $\sigma_{x_1 - x_2} = \sqrt{\sigma_2^2 - \sigma_1^2}$

check: If the new cut is the same as the old cut, no difference in measurements!
Assumes: Gaussian, uncorrelated measurement pieces.

The Kolmogorov-Smirnov Test

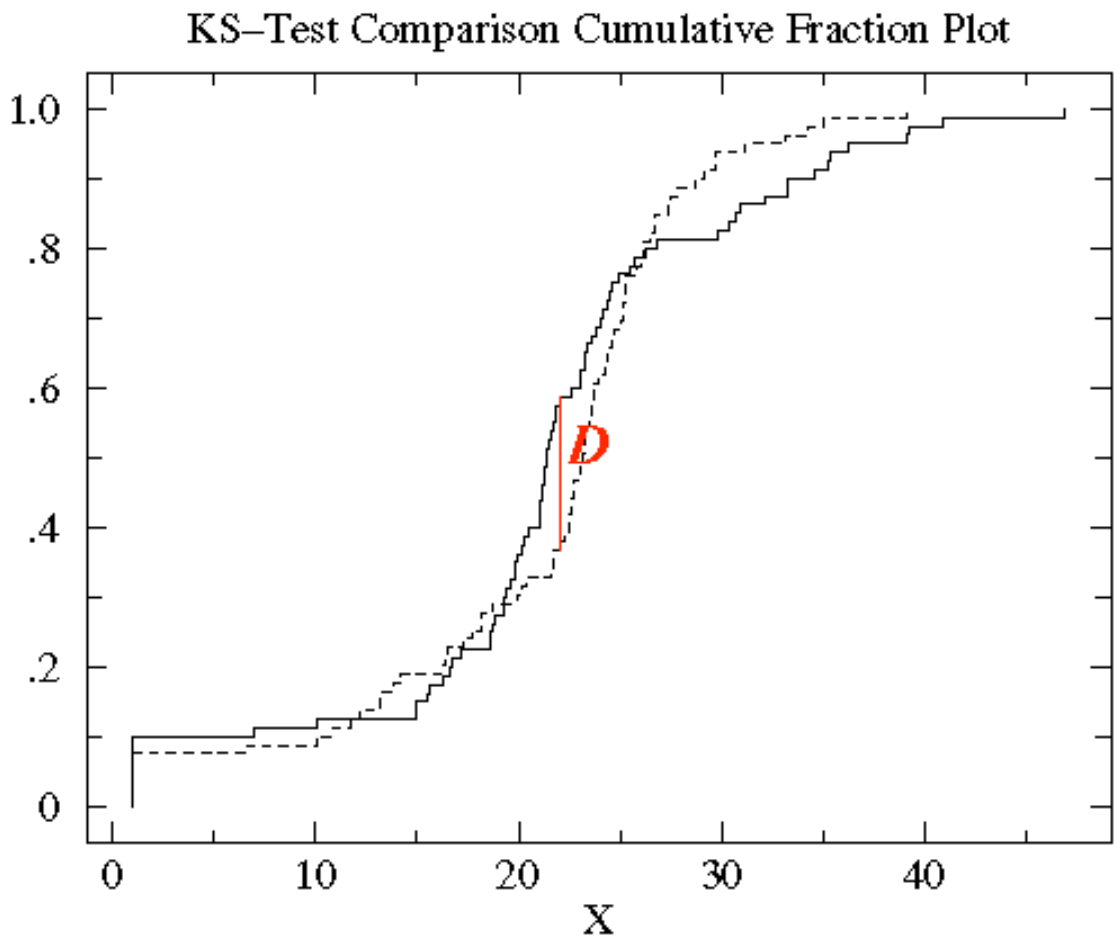
χ^2 Doesn't tell you everything you may want to know about distributions that have modeling problems.

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"



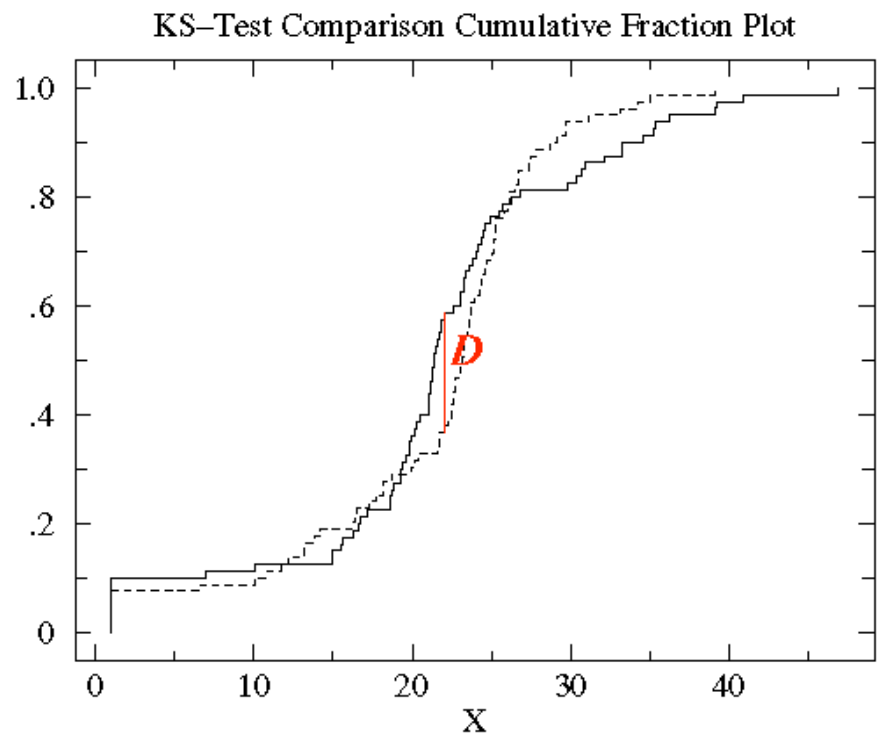
http://www.physics.csbsju.edu/stats/KS-test.html

The Kolmogorov-Smirnov 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.

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

Cautions with the Binned Kolmogorov-Smirnov Test

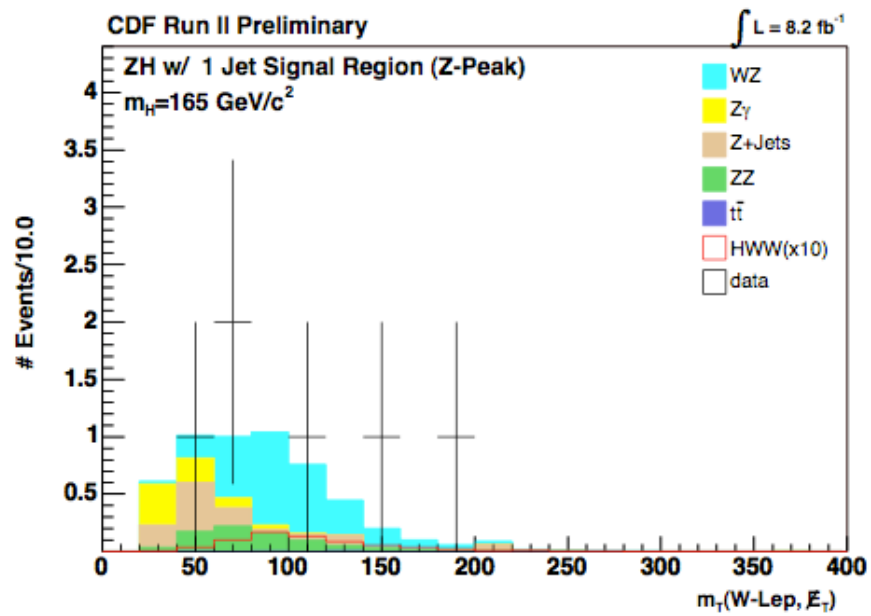
// Statistical test of compatibility in shape between
// THIS histogram and h2, using Kolmogorov test. /

Double_t TH1::KolmogorovTest(const TH1 *h2, Option_t *option)

The pseudoexperiment option "X" only varies the the "this" histogram (h1) and not h2 but it draws pseudoevents with the histogram normalization of h2.

This procedure makes sense if the "this" histogram is a smooth model, and h2 has statistically limited data in it. Exchanging h1 and h2 gives you different KS p-values (although the same D!)

Putting in the histograms in reverse order can make for some very large KS p-values – I've seen talks in which all the KS p-values are 0.99 or higher.

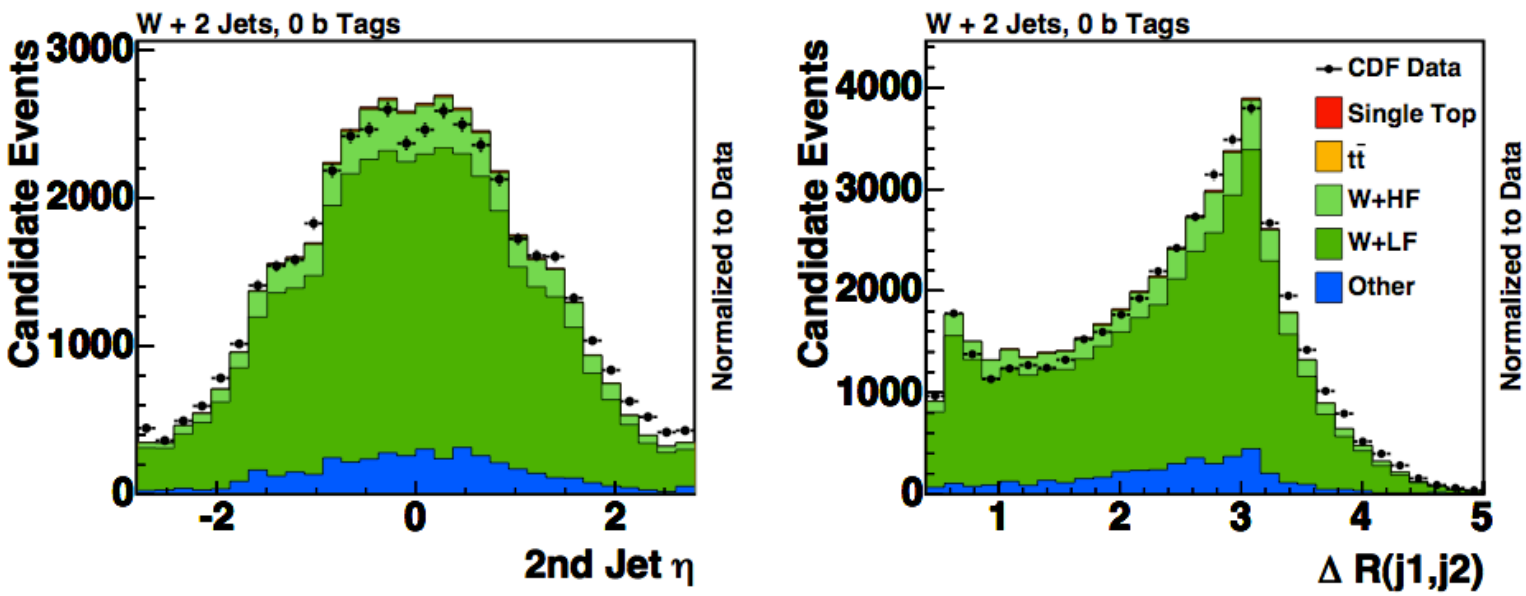


The Run Test

Count the maximum number of neighboring positive deviations from the data and the prediction, and also negative deviations. If there are many deviations of the same sign in a row, even if the χ^2 looks okay, it is a sign of mismodeling.

Typically we don't go to the trouble of computing p-values for the run test. But it's a handy thing to remember when reviewing the modeling of distributions in the process of approving analyses. What's the chance of getting 10 fluctuations of the same sign in a row? (2^{-9}, but watch the Look Elsewhere Effect, to be described later.

Only works in 1D. Can be sensitive to the overall normalization (which we may care less about than shape mismodeling)



Two (or more) Parameters of Interest

For quoting Gaussian uncertainties on **single** parameters. Ellipse is a contour of 2ΔlnL=1

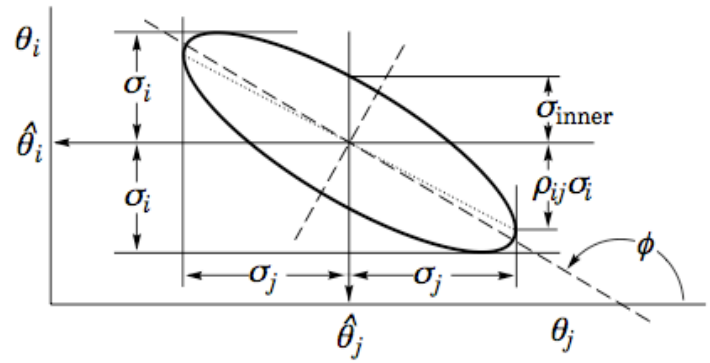


Figure 33.5: Standard error ellipse for the estimators $\widehat{\theta}_i$ and $\widehat{\theta}_j$. In this case the correlation is negative.

For displaying joint estimation of several parameters

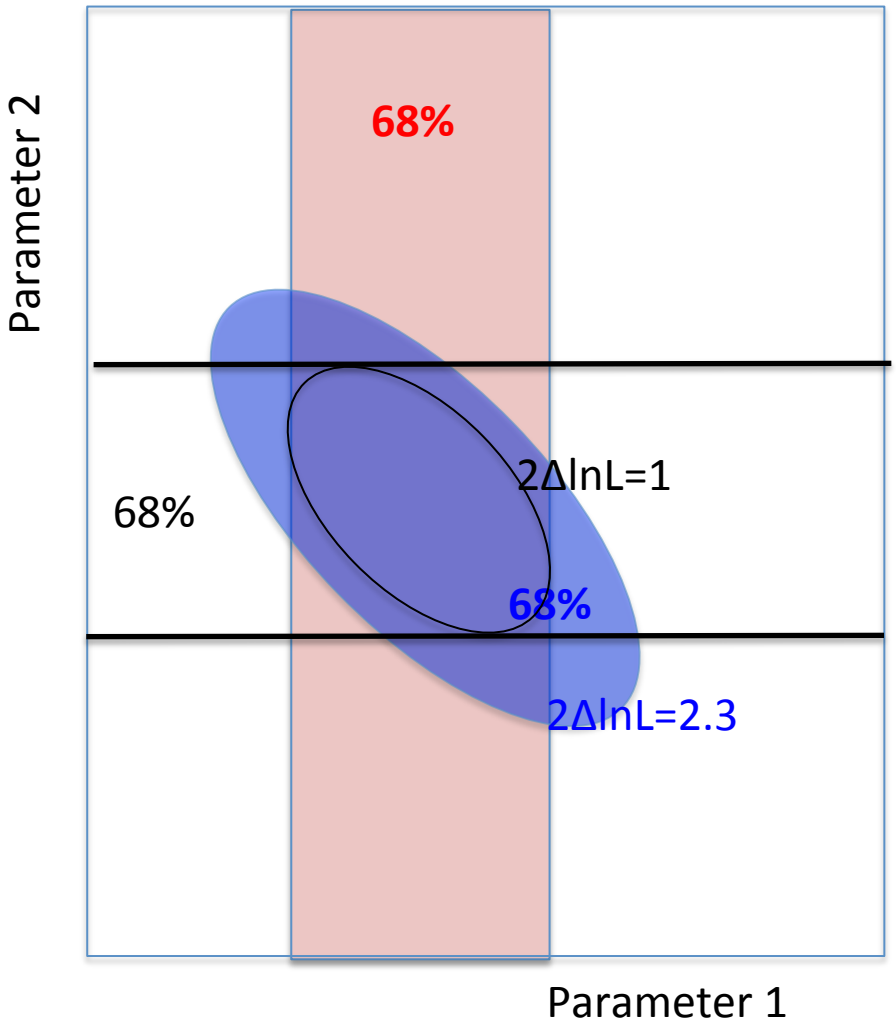
Table 33.2: $\Delta\chi^2$ or $2\Delta \ln L$ corresponding to a coverage probability $1-\alpha$ in the large data sample limit, for joint estimation of m parameters.

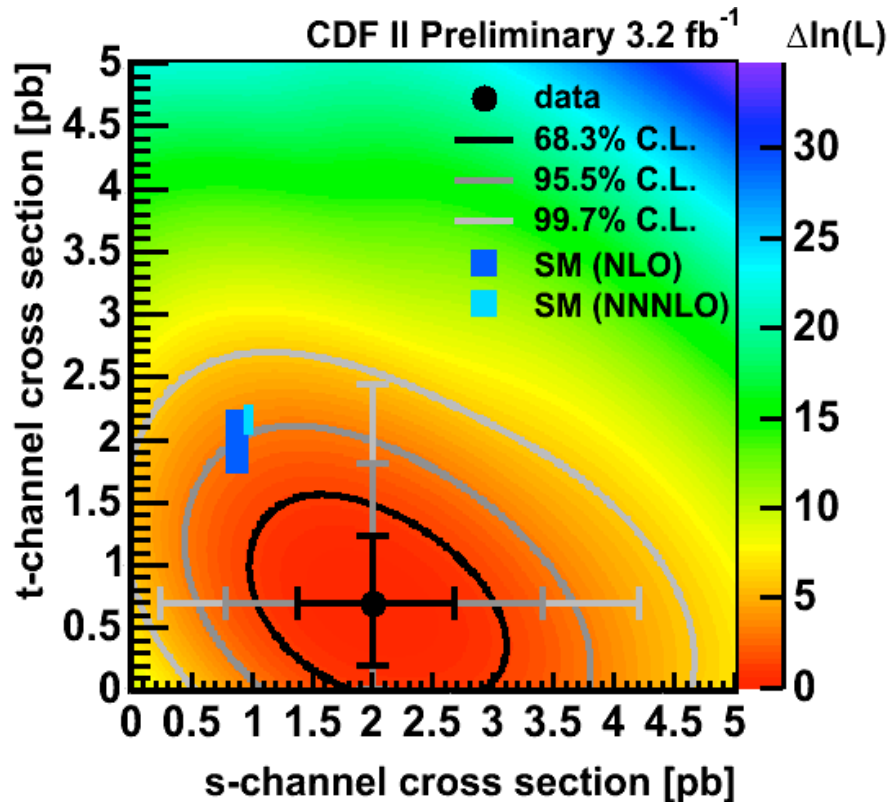
$(1-\alpha)$ (%)	$m=1$	$m=2$	$m=3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

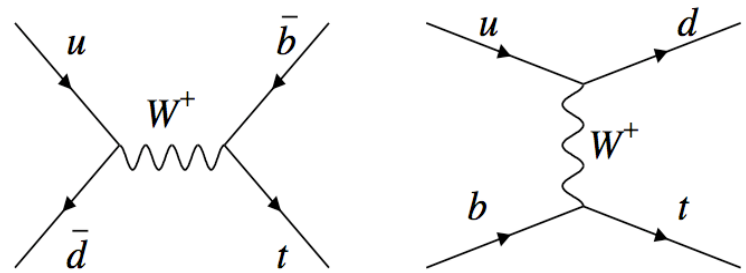
From the 2011 PDG Statistics Review

http://pdg.lbl.gov/2011/reviews/rpp2011-rev-statistics.pdf

1D or 2D Presentation

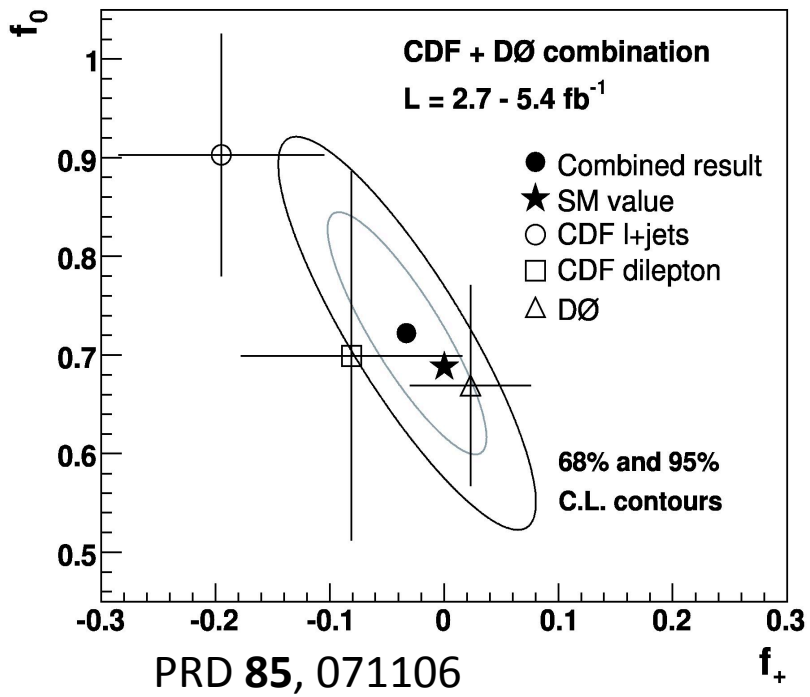




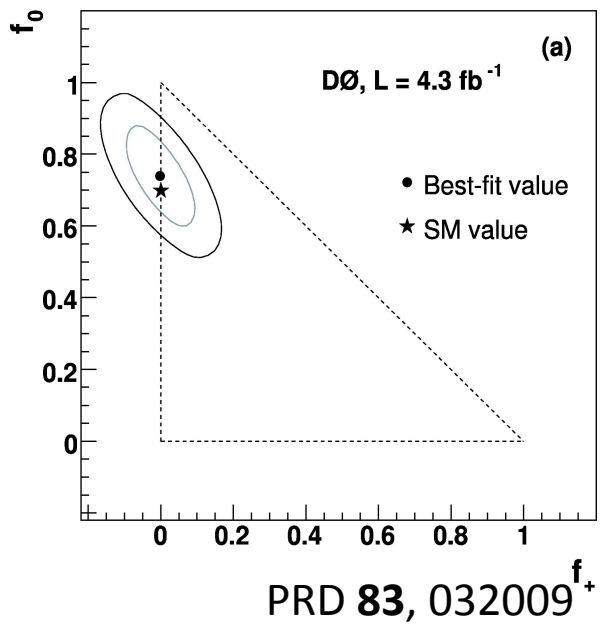


I prefer when showing a 2D plot, showing the contours which cover in 2D. The 2ΔlnL=1 contour only covers for the 1D parameters, one at a time.

A Variety of ways to show 2D Fit results

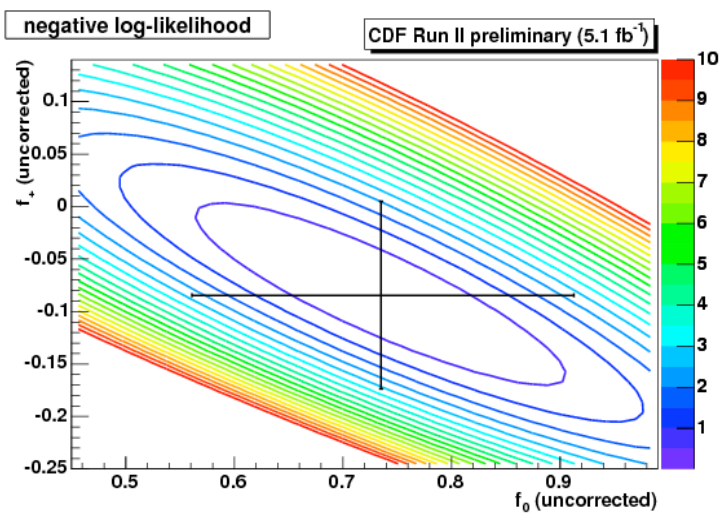


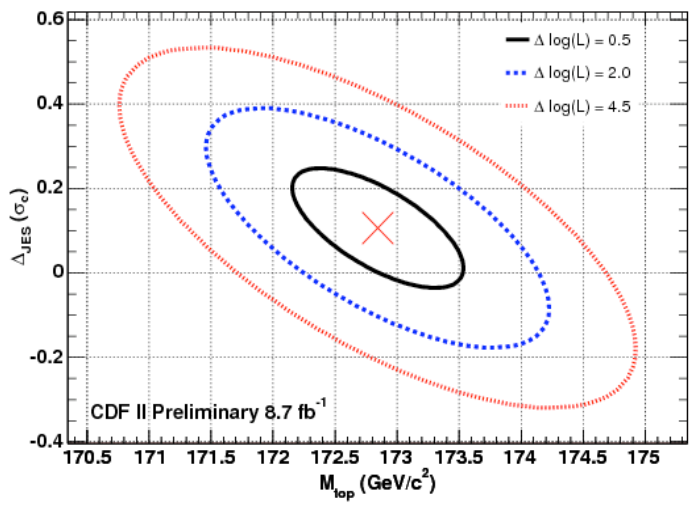
PRD **85**, 071106



PRD **83**, 032009

68% and 95% contours

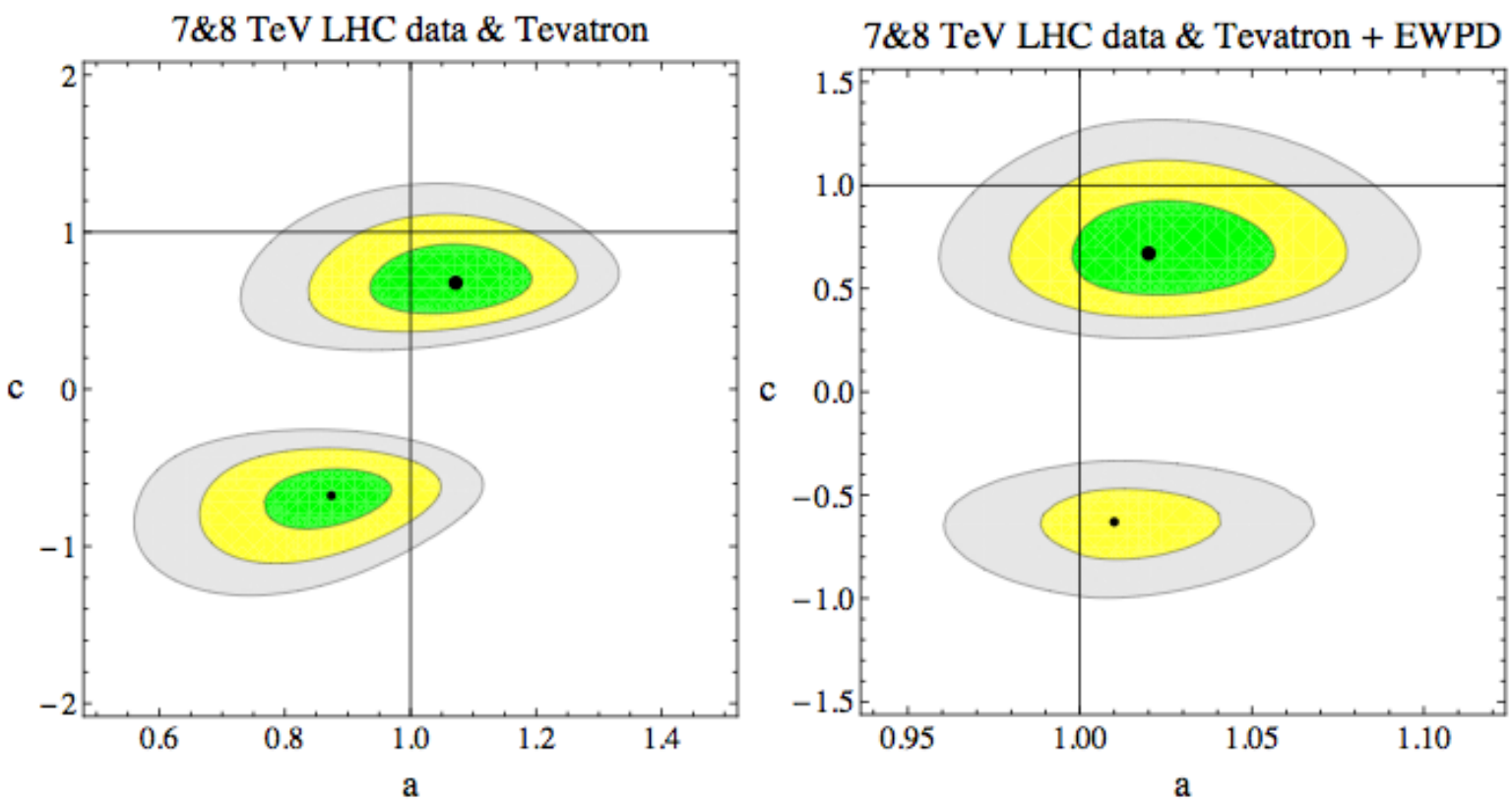




http://www-cdf.fnal.gov/physics/new/top/2011/WhelDil/index.html

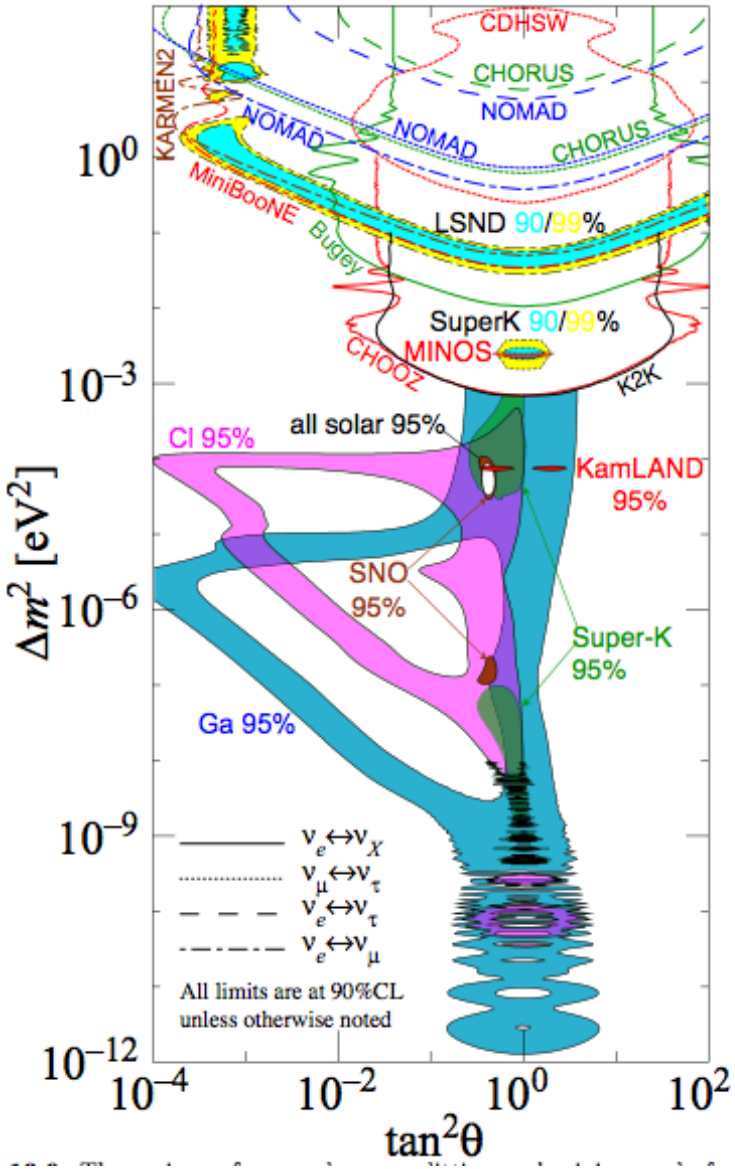
Sometimes you have multimodal Likelihoods

These come up frequently because predicted yields are at least quadratic functions of coupling parameters



Espinosa, Grojean, Mühlleitner, Trott http://arxiv.org/abs/arXiv:1207.1717

Or Even More Interesting Dependence on Model Parameters



PDG Neutrino Oscillation Review, 2012