Data Analysis and Statistical Methods in Experimental Particle Physics

Thomas R. Junk
Fermilab

TRISEP 2014
Sudbury, Ontario

A “Unicorn”

Lecture Outline

Lecture 1: Introduction, Gaussian Approximations

Lecture 2: Data Analysis and Systematic Uncertainties

Lecture 3: Frequentist and Bayesian Interpretation

Lecture 4: Higgs Boson Searches, Discovery, and Measurements

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

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

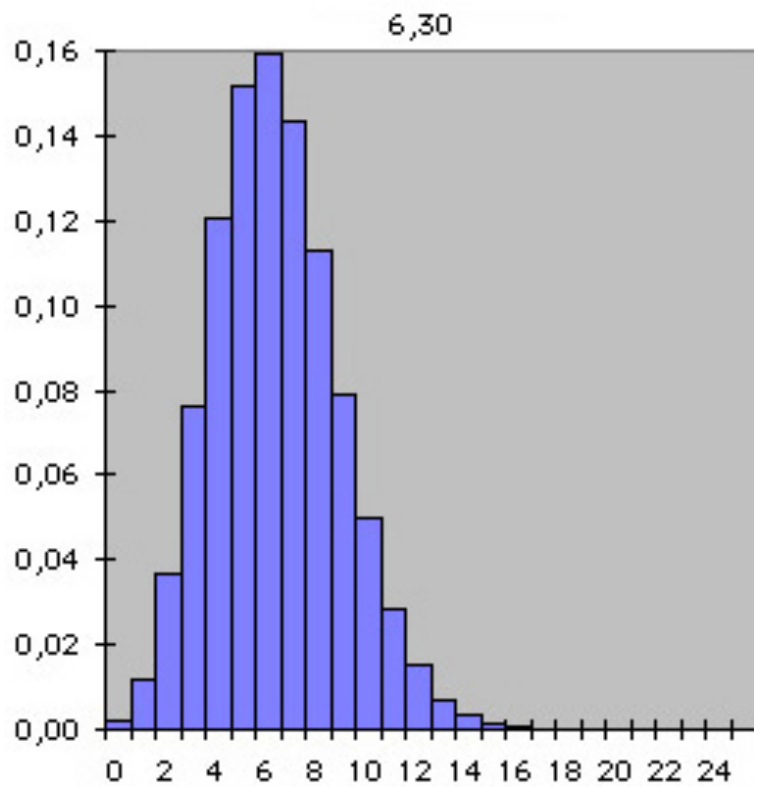
$p(n \geq n_{obs} | H_0)$

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.

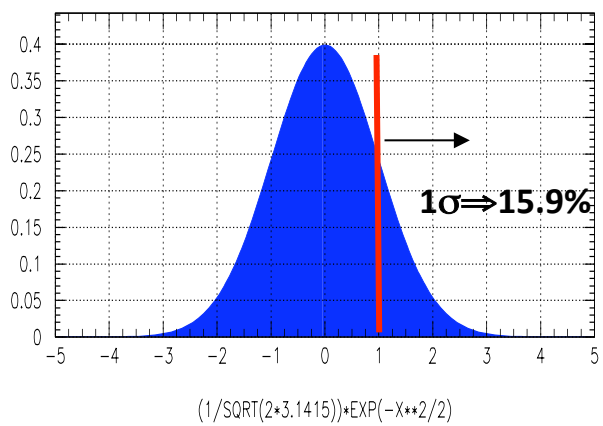
The number of "sigma" is called a "*z*-value" and is just a translation of a p-value using the integral of one tail of a Gaussian

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

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



Tip: most physicists talk about *p*-values now but hardly use the term *z*-value – we use the word "significance" instead

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.

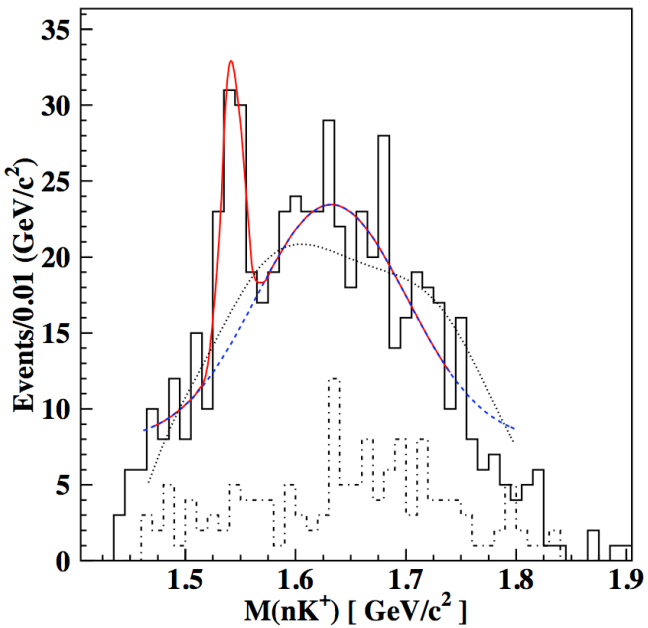
Some historical recollections:
http://www.huffingtonpost.com/victor-stenger/higgs-and-significiance_b_1649808.html

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)).

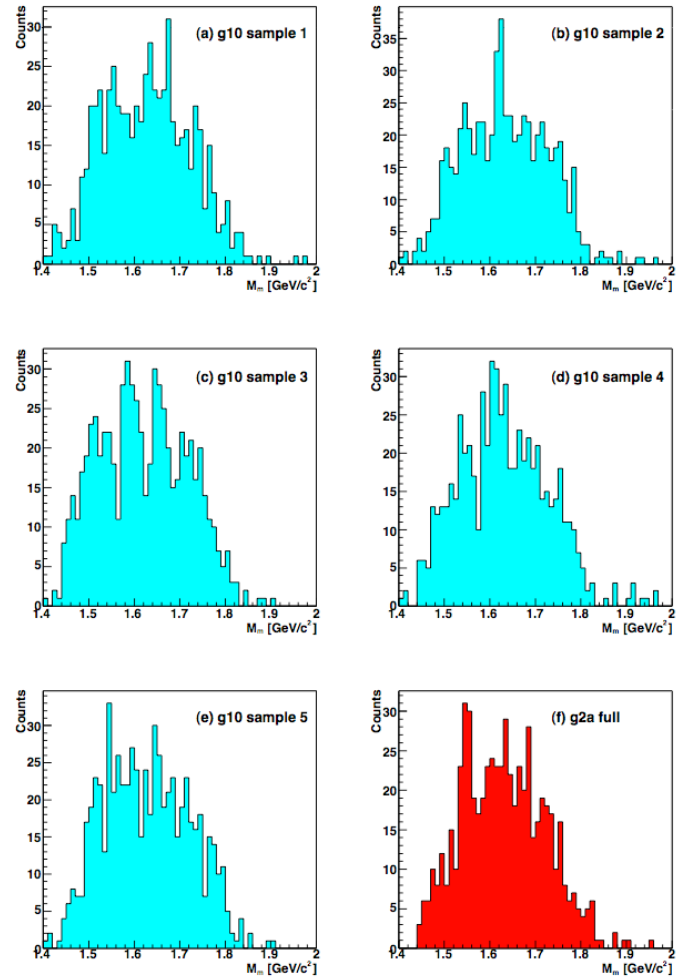
A Cautionary Tale – The Pentaquark "Discoveries"



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

Significance = 5.2 ± 0.6 σ

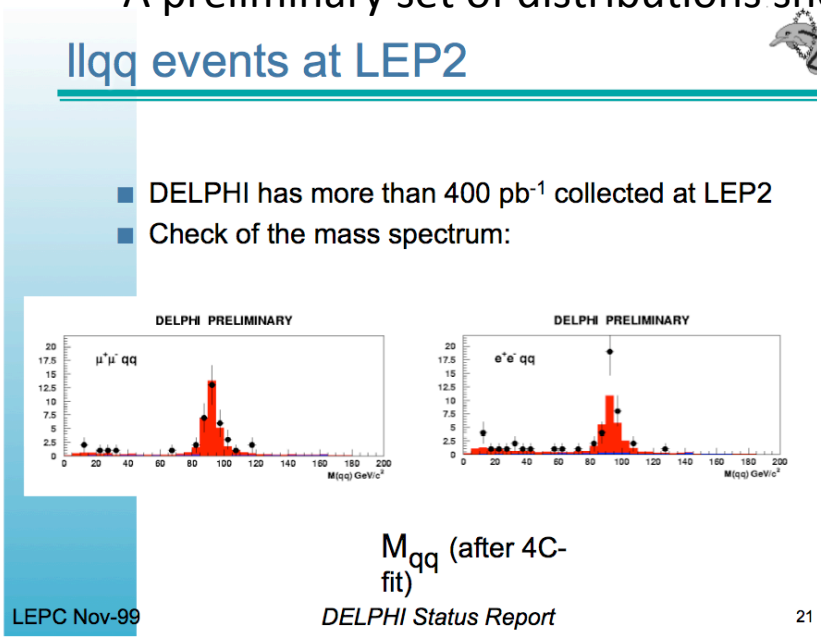
Watch out for the background function parameterization!

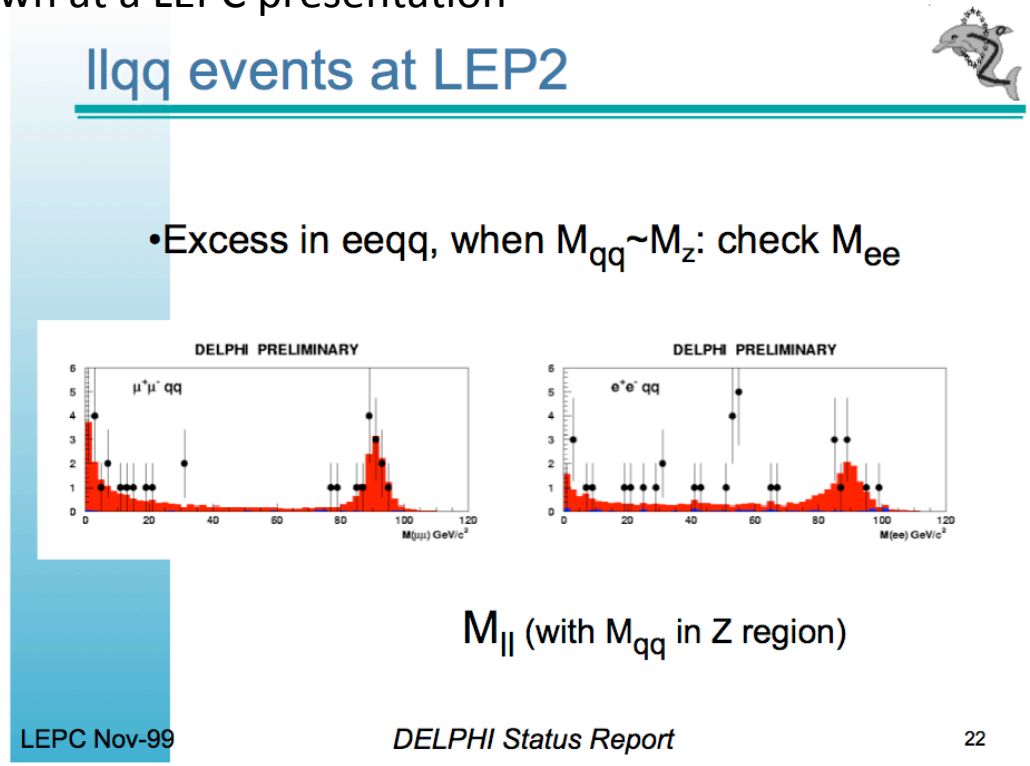


Five times the data sample
CLAS Collab., **Phys.Rev.Lett. 100 (2008) 052001**

Another Bump That Went Away

A preliminary set of distributions shown at a LEPC presentation





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”

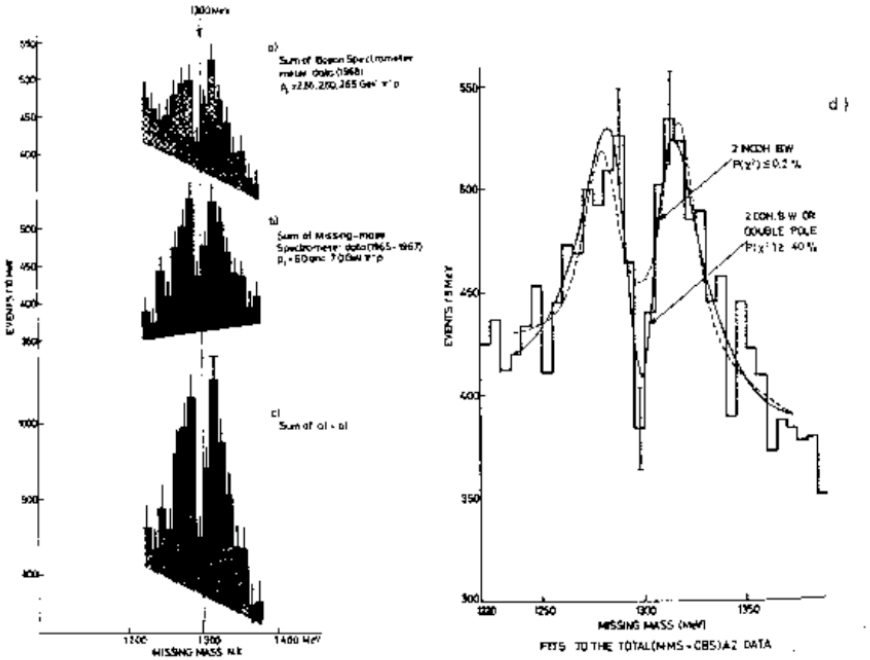


Figure 3: (a-c) Evidence for A_2 splitting in $\pi^- p \to pX^-$ collisions in the two CERN experiments, (d) same as (c) in 5 MeV bins fit to two hypotheses.

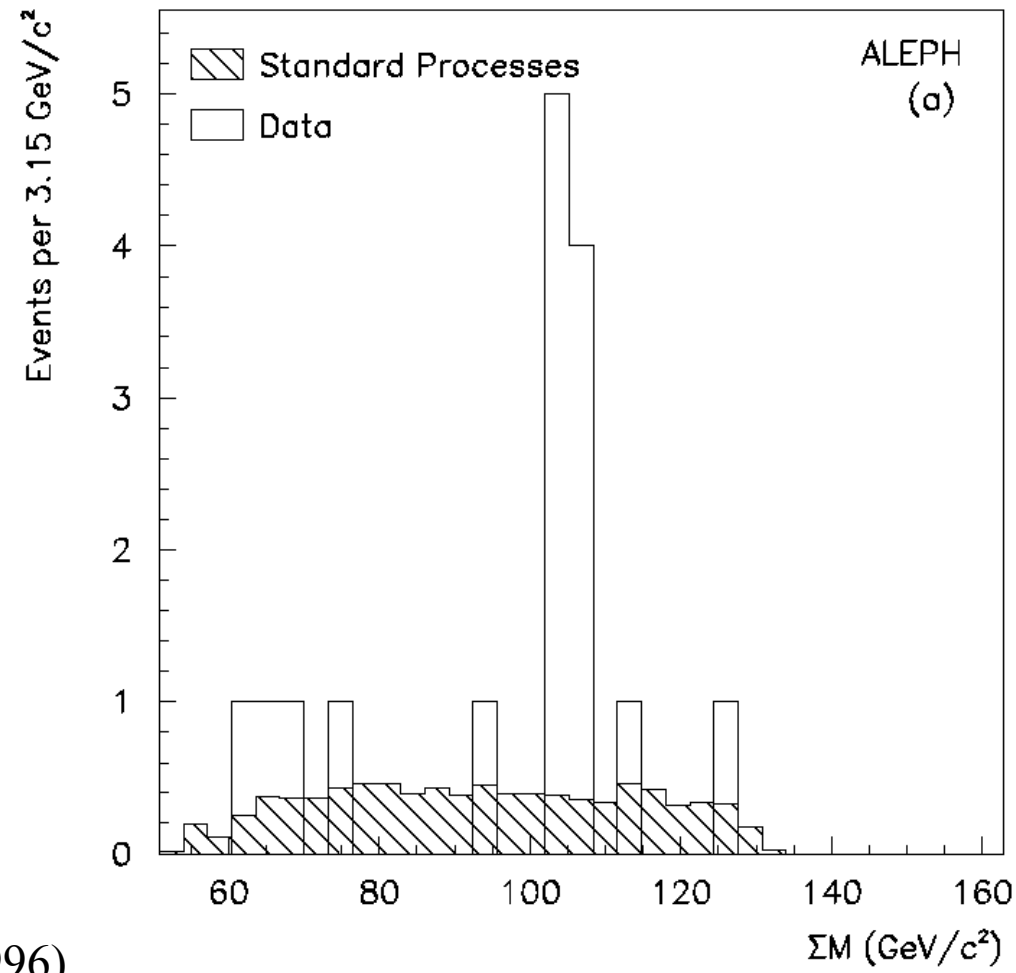
Text from Sheldon’s article:

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"



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

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

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.

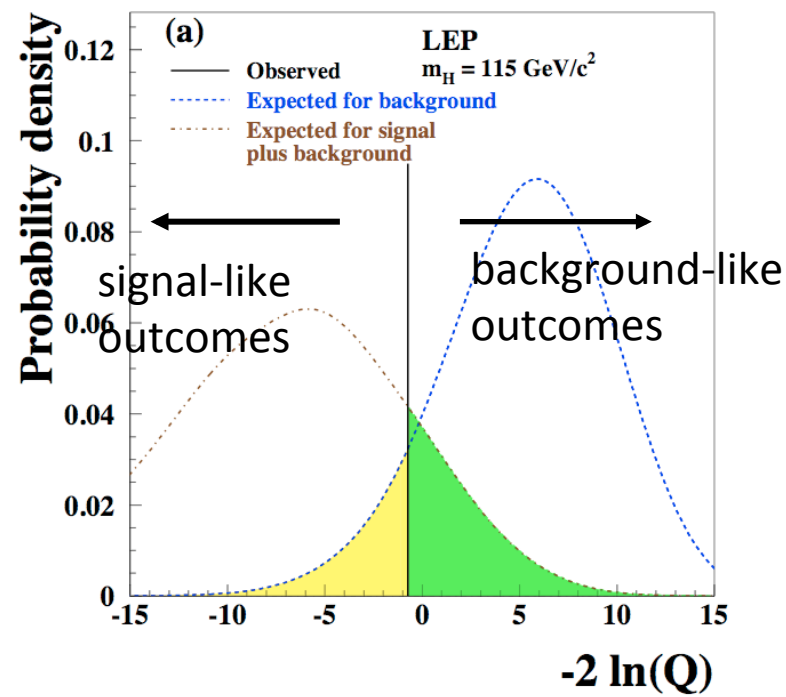
The word coverage comes from confidence intervals – are they big enough to contain the true value of a parameter being measured and what fraction of the time they do.

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 $\hat{\boldsymbol{\nu}}$ and $\hat{\hat{\boldsymbol{\nu}}}$?

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

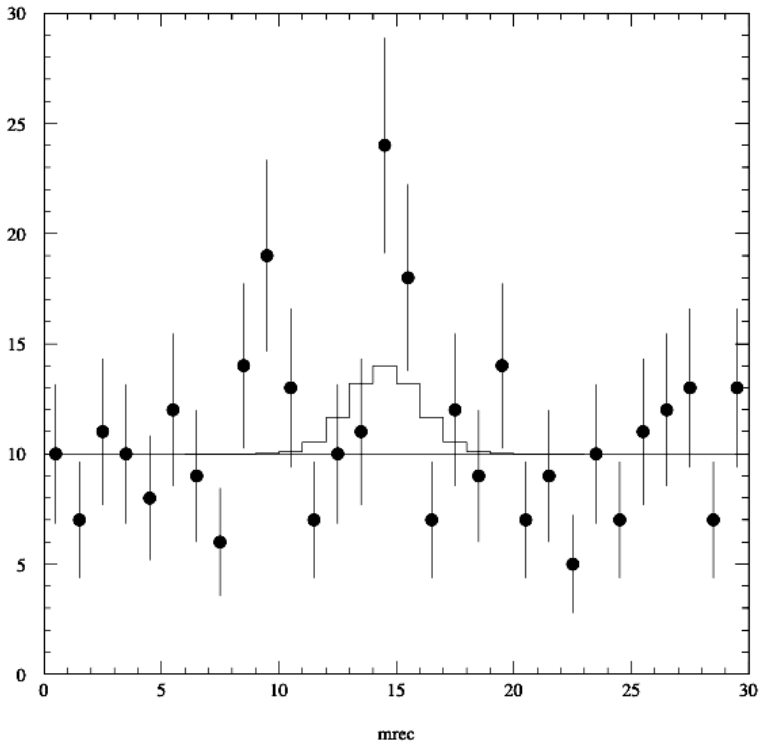
A *compound hypothesis* is less specific. It may have parameters whose values we are not particularly concerned about but which affects its predictions. These are called *nuisance parameters*, labeled ν.

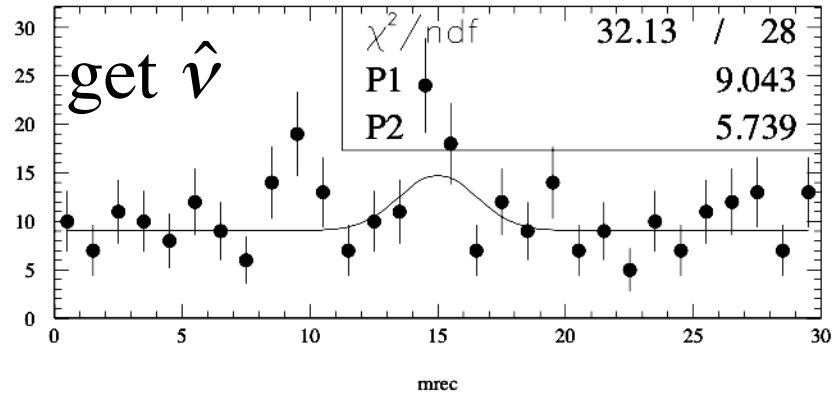
Example: H_0=SM. H_1=MSSM. Both make predictions about what may be seen in an experiment. A nuisance parameter would be, for example, the b-tagging efficiency. It affects the predictions but in the end of the day we are really concerned about H_0 and H_1.

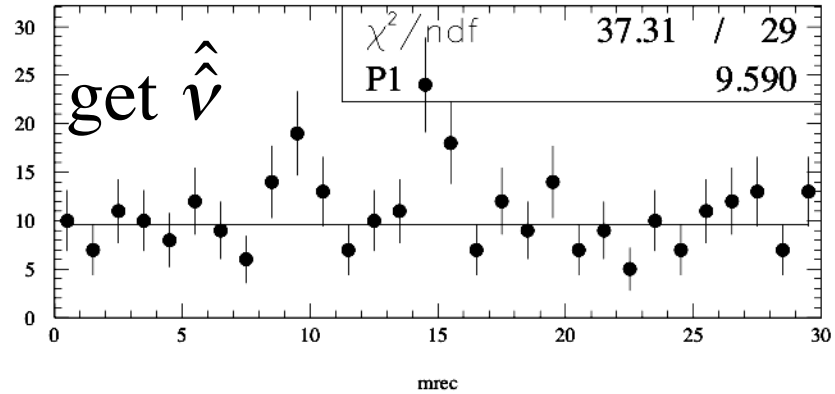
Fit twice! Once assuming H_0, once assuming H_1

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

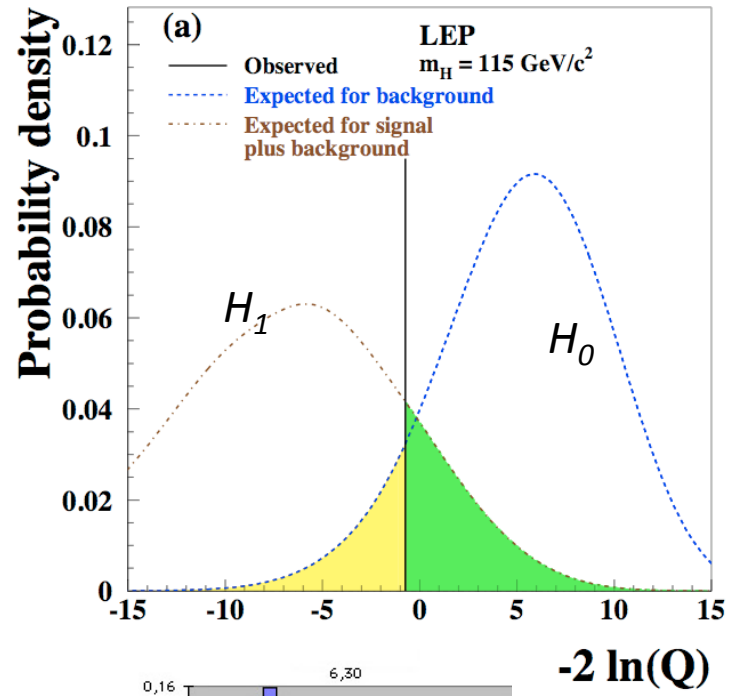
The "or-equal-to" is important here. For highly discrete distributions of possible outcomes – say an experiment with a background rate of 0.01 events (99% of the time you observe zero events, all the same outcome), then observing 0 events gives a p-value of 1 and not 0.01.

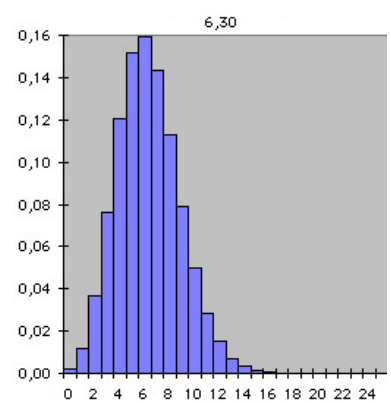
Shouldn't make a discovery with 0 observed events, no matter how small the background expectation! (or we would run the LHC with just one bunch crossing!).

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

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

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





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

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.

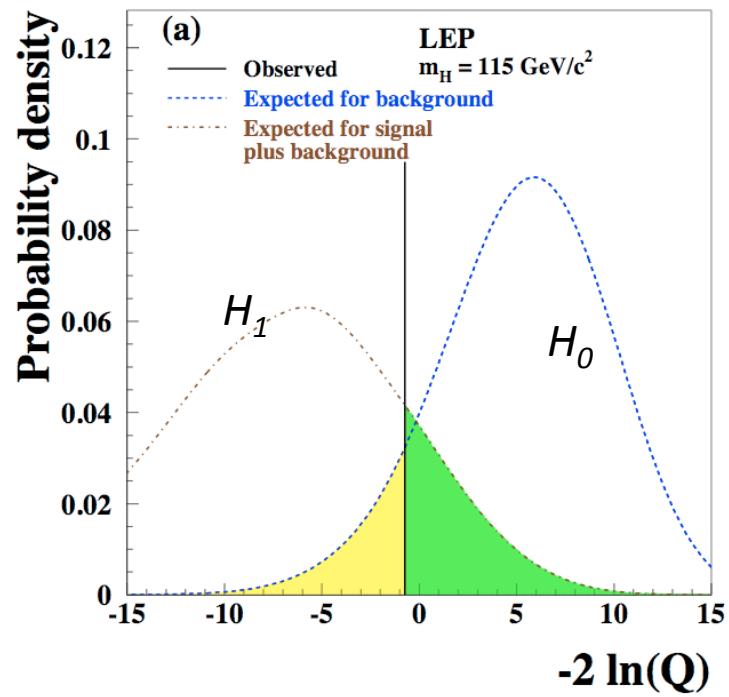
If it is small, reject H_1
The "or-equal-to" has similar effect here too.

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.

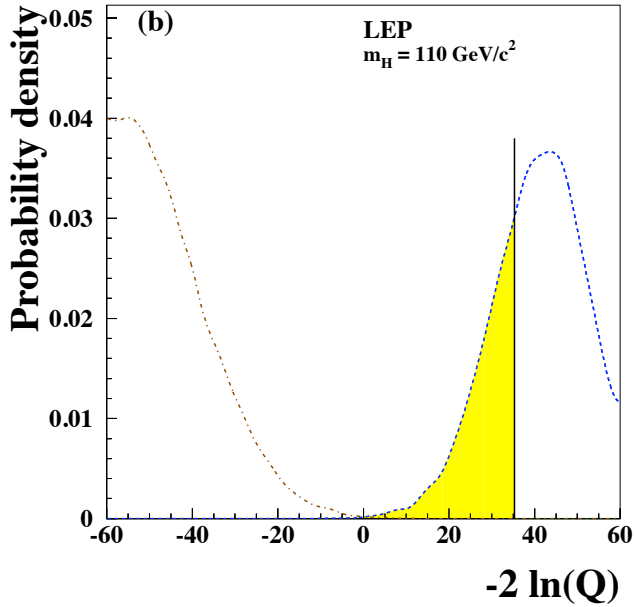
Lucky outcome: exclude at 97% CL
Do we exclude at the 50% CL?

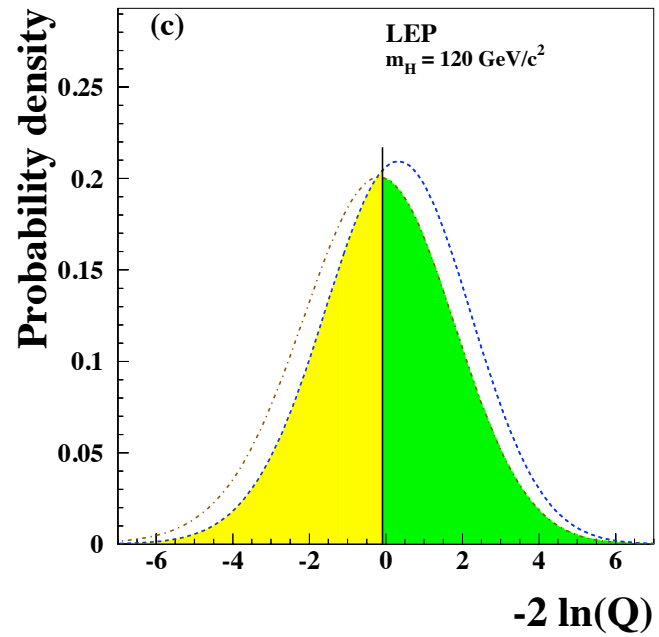
No! Set α once and for all (say 0.05). Then coverage is defined.



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

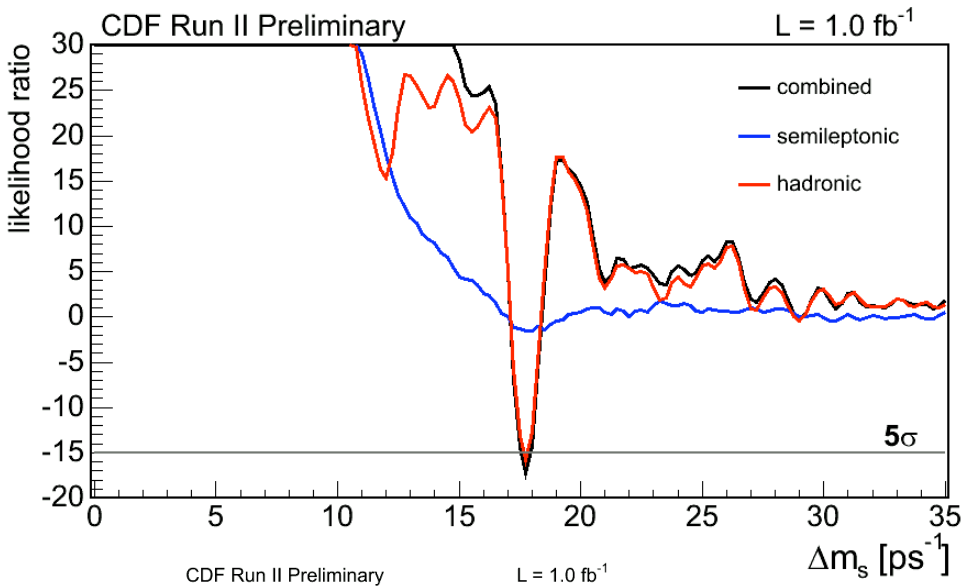
Can make no statement about the signal regardless of experimental outcome.

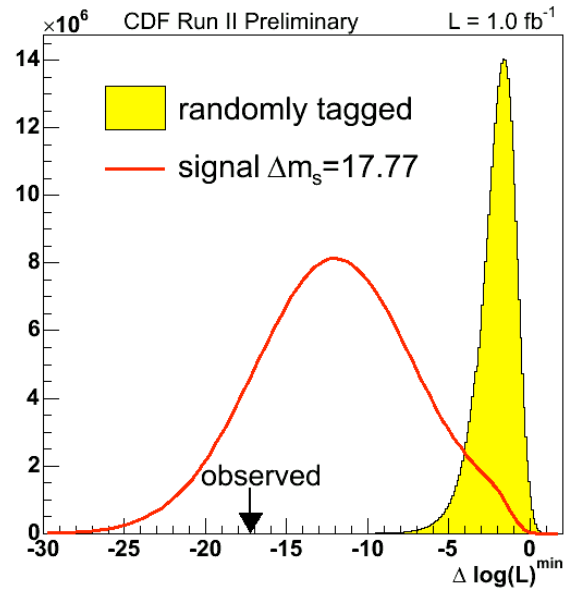
Unlikely (or implausible) outcomes are still possible of course!

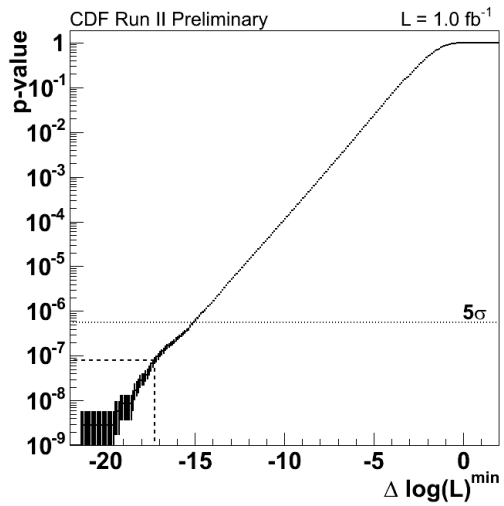
The usual sensitivity gauge: Median Expected p_0 if a signal is truly present.

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

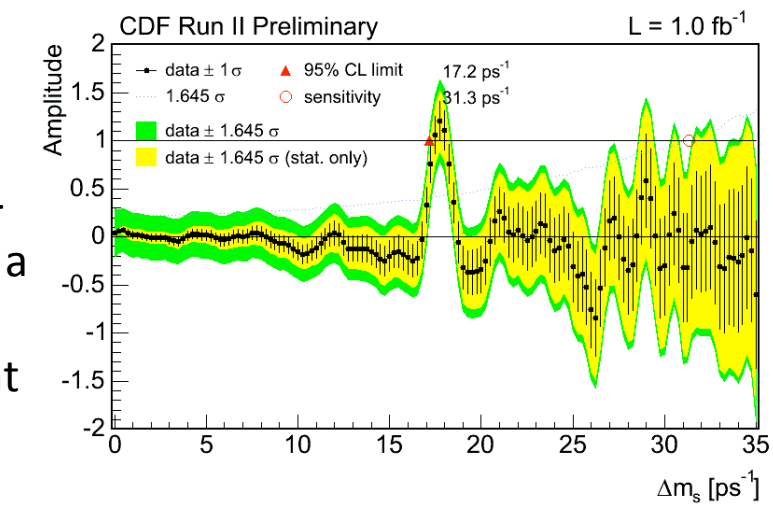
LLR Is not only used in Search Contexts – Precision Measurements too!







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

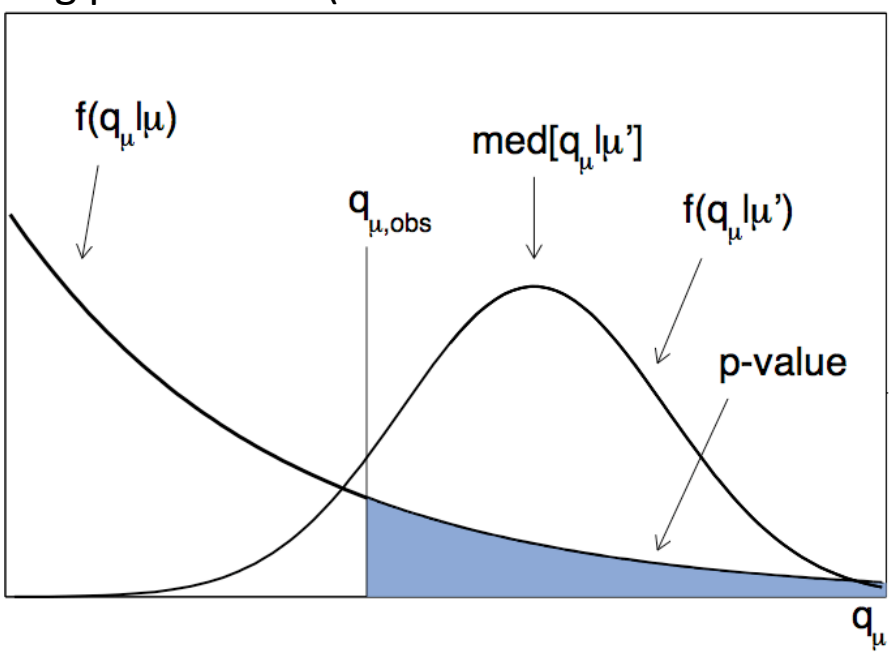
Distinguishing between μ=0 (zero signal, SM, Null Hypothesis) and μ>0 (the test hypothesis)

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

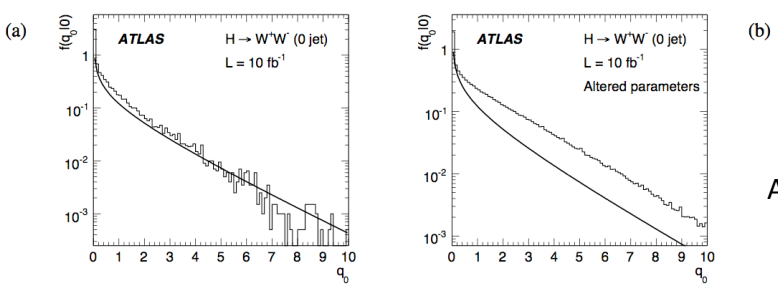
If the true value of the signal rate is given by μ, then q_μ is distributed according to a χ^2 distribution with one degree of freedom.

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*.

T. Junk TRISEP 2014 Lecture 3 27

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.

T. Junk TRISEP 2014 Lecture 3 30

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

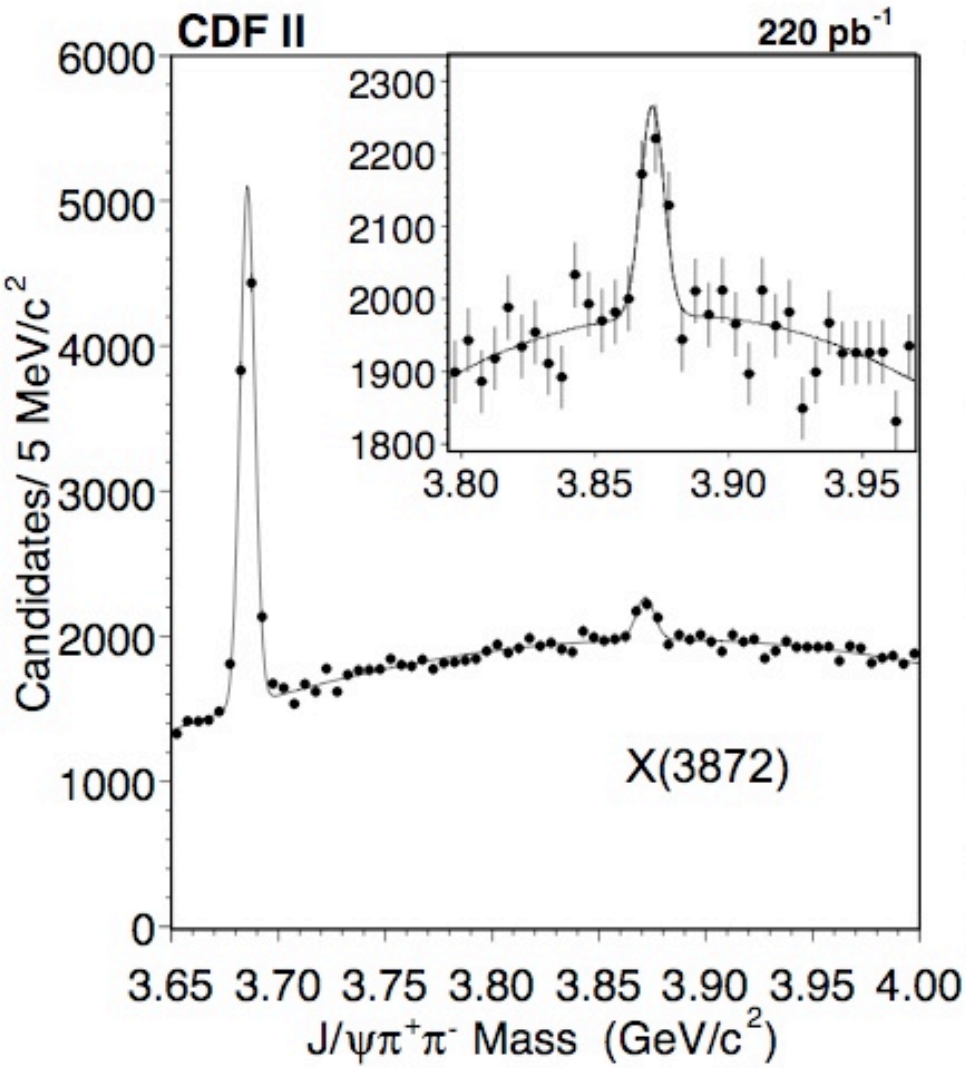
Probability that a future observation will be at least as extreme as the current observation assuming that the null hypothesis is true.

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.

The Traditional Solution to Large, Uncertain Backgrounds: Sideband Fits



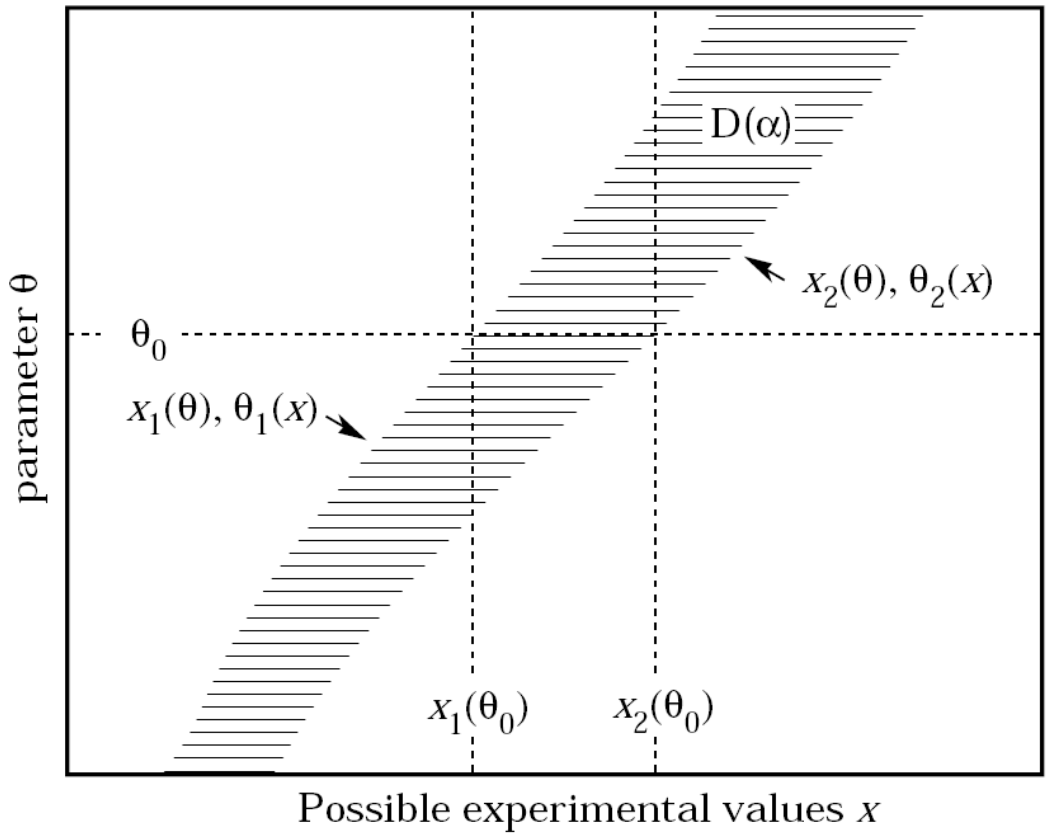
Guess a shape that fits the backgrounds, and fit it with a signal.

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.

Proper Coverage is Guaranteed!



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.

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!

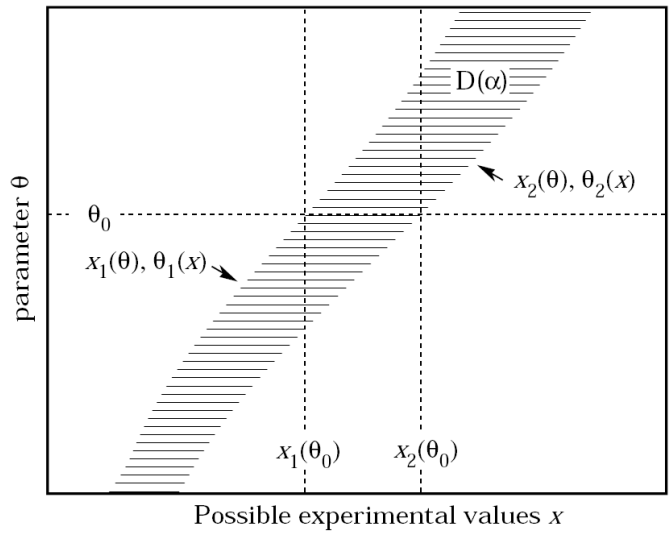
Take lowest x values: get lower limits.
Take highest x values: get upper limits.

Cousins and Feldman: Sort outcomes by the likelihood ratio.

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

R=1 for all x for some θ.

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

No empty intervals!

Some Properties of Frequentist Confidence Intervals

- Really just one: *coverage*. If the experiment is repeated many times, the intervals obtained will include the true value at the specified rate (say, 68% or 95%).

  Conversely, the rest of them (1-α) of them, must not contain the true value.

- But the interval obtained on a particular experiment may obviously be in the unlucky fraction. Intervals may lack credibility but still cover.

  Example: 68% of the intervals are from $-\infty$ to $+\infty$, and 32% of them are empty. Coverage is good, but power is terrible.

  FC solves some of these problems, but not all.
  Can get a 68% CL interval that spans the entire domain of θ.
  Imagine publishing that a branching ratio is between 0 and 1 at 68% CL.

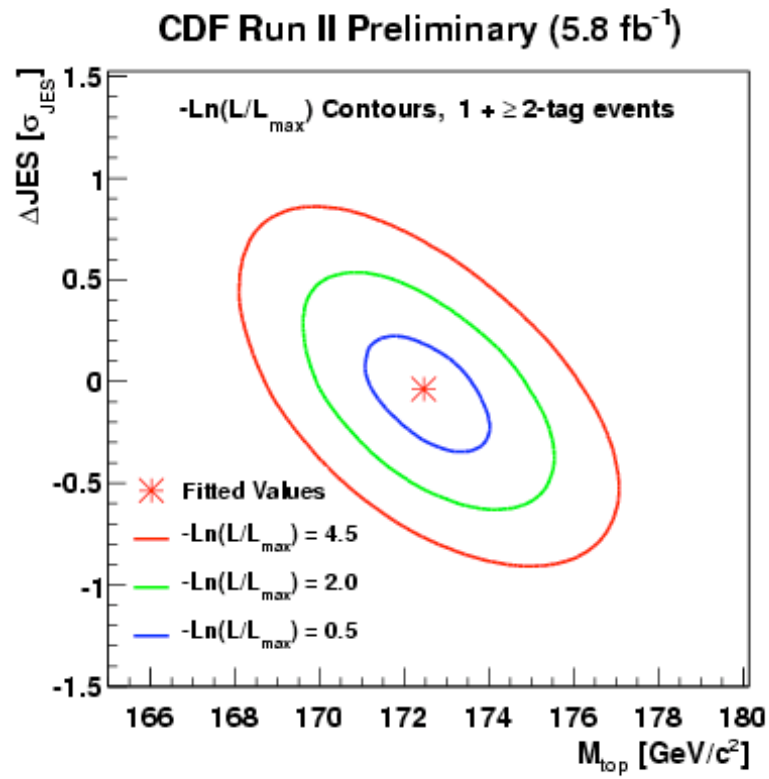
  Still possible to exclude models to which there is no sensitivity.

  FC assumes model parameter space is complete -- one of the models in there is the truth. If you find it, you can rule out others even if we cannot test them directly.

Treat Nuisance Parameters as Parameters of Interest!

- Somewhat arbitrary distinction, anyhow. Although you could argue this is what the Scientific Method is all about; separating nuisance parameters from parameters of interest.

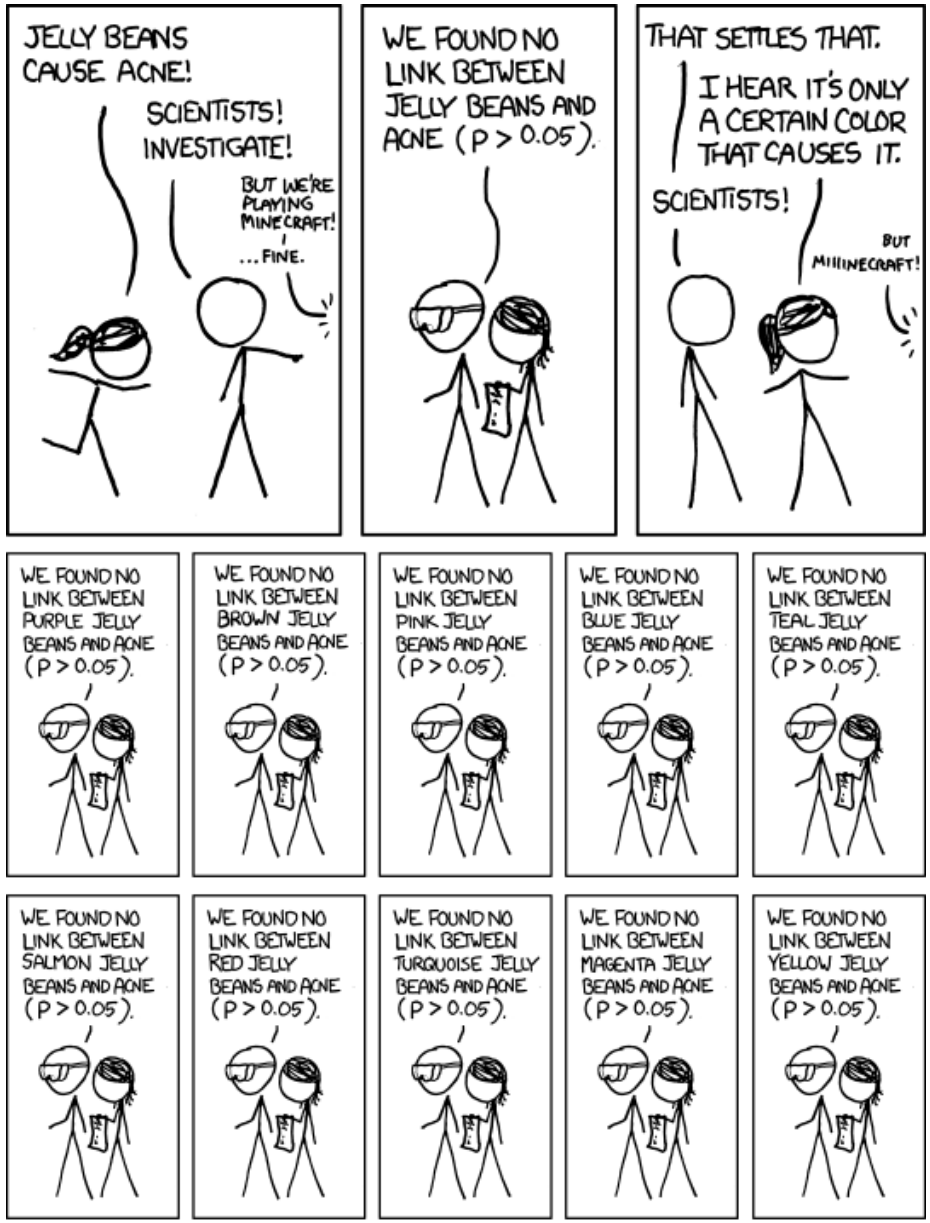
- Really only good if you have one dominant source of systematic uncertainty, and you want to show your joint measurement of the nuisance parameter and the parameter of interest.

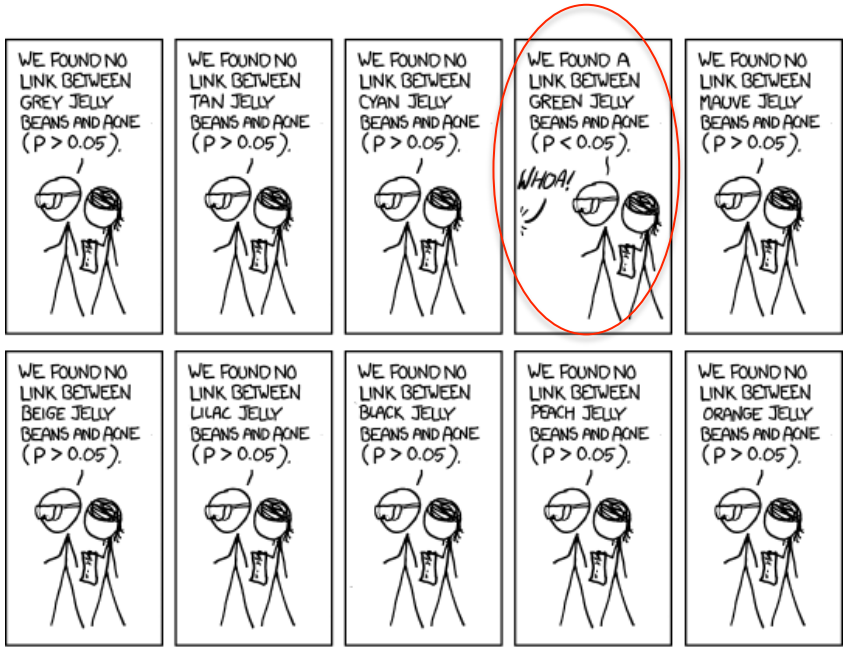


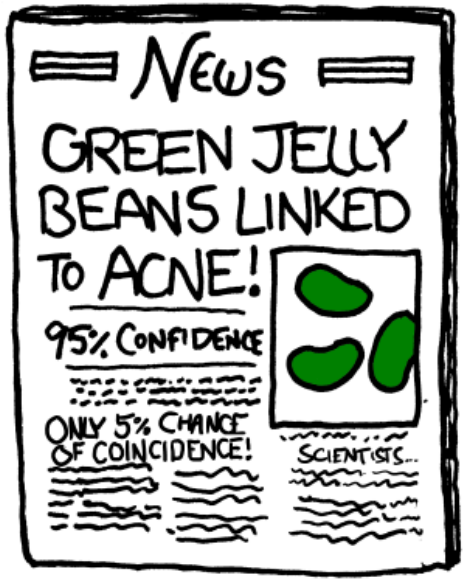
Doesn't generalize all that well to many parameters of interest.

Example: top quark mass (parameter of interest), vs. CDF's jet energy scale in all-hadronic ttbar events.

Multiple Testing: "Look Elsewhere"







xkcd.com

The Look-Elsewhere Effect

- Also called the "Trials Factor" or the effect of "Multiple Testing"
- Bump-hunters are familiar with it.

What is the probability of an upward fluctuation as big as the one I saw *anywhere* in my histogram?

-- Lots of bins → Lots of chances at a false discovery
-- Approximation (Bonferroni): Multiply smallest p-value by the number of **"independent" models** sought (not histogram bins!).
Bump hunters: roughly (histogram width)/(mass resolution)
Criticisms:
Adjusted p-value can now exceed unity!
What if histogram bins are empty?
What if we seek things that have been ruled out already?

Just as easy: The Dunn-Šidák correction, still assumes independence.
$p_{corrected} = 1 - (1-p_{min})^n$

The Look-Elsewhere Effect

More seriously, what to do if the p-value comes from a big combination of many channels each optimized at each m_H sought?

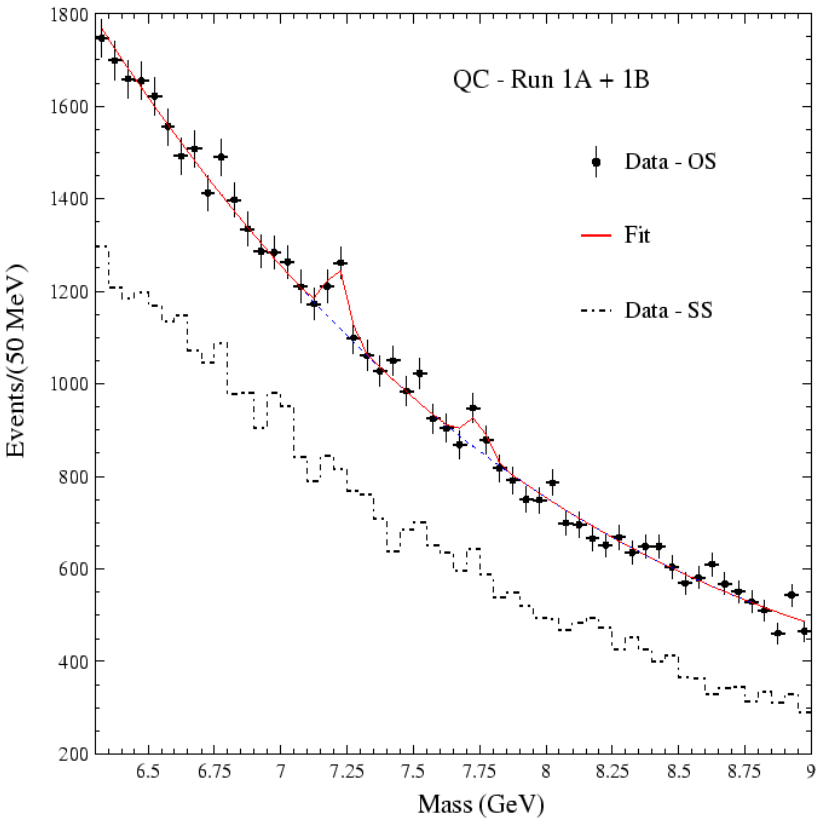
- Channels have different resolutions (or is resolution even the right word for a multivariate discriminant?
- Channels vary their weight in the combination as cross sections and branching ratios change with m_H

Proper treatment -- want a p-value of p-values!
(use the p-value as a test statistic)
Run pseudoexperiments and analyze each one at each m_H studied. Look for the distribution of smallest p-values.

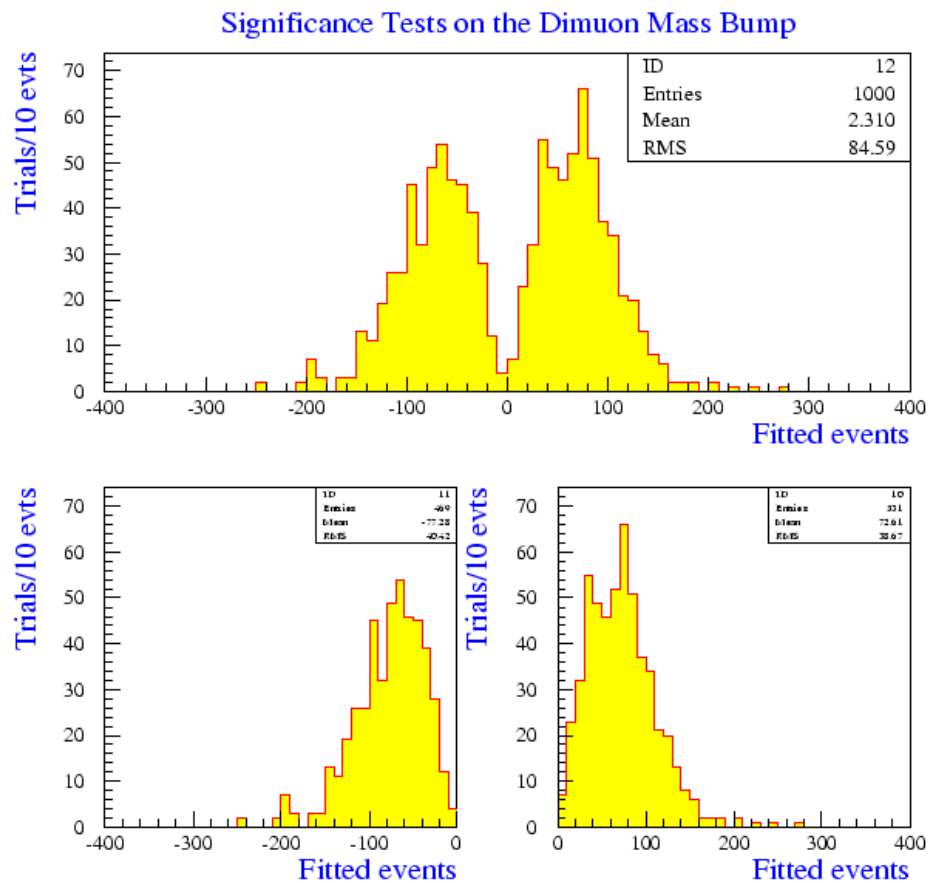
Next to impossible unless somehow analyzers supply how each pseudo-dataset looks at each test mass.

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!)

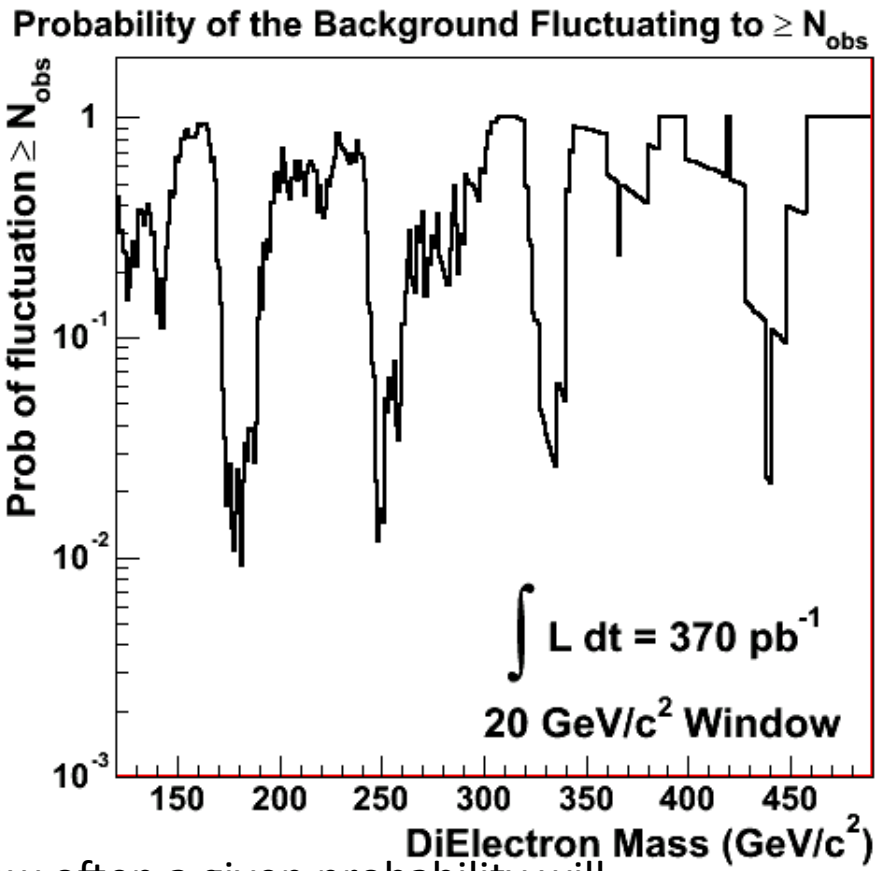


Null hypothesis pseudoexperiments with largest peak fit values

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

Looking Everywhere in a m_{ee} plot

- method:
  - scan along the mass spectrum in 1 GeV steps
  - at each point, work out prob for the bkg to fluctuate ≥ data in a window centred on that point
    - window size is 2 times the width of a Z' peak at that mass
  - sys. included by smearing with Gaussian with mean and sigma = bkg + bkg error



  - use pseudo experiements to determine how often a given probability will occur e.g. a prob ≤0.001 will occur somewhere 5-10% of the time

An Approximate LEE Correction for Peak Hunting

See E. Gross and O. Vitells, **Eur.Phys.J. C70 (2010) 525-530**.

Approximate formula applies to bump hunts on a smooth background.

Not all searches are like this – Multivariate Analyses are usually trained up at each mass separately, and there is not a single distribution we can look elsewhere in.

An interesting, very general feature:

As the expected significance goes up, so does the LEE correction

This makes lots of sense: LEE depends on the number of separate models that can be tested. As we collect more data, we can measure the position of the peak more precisely.

So we can tell more peaks apart from each other, even with the same reconstruction resolution.

Where is "Elsewhere?"

A collider collaboration is typically very large; >1000 Ph.D. students. ATLAS+CMS is another factor of two. (Four LEP collaborations, Two Tevatron collaborations).

Many ongoing analyses for new physics. The chance of seeing a fake bump somewhere is large. What is the LEE?

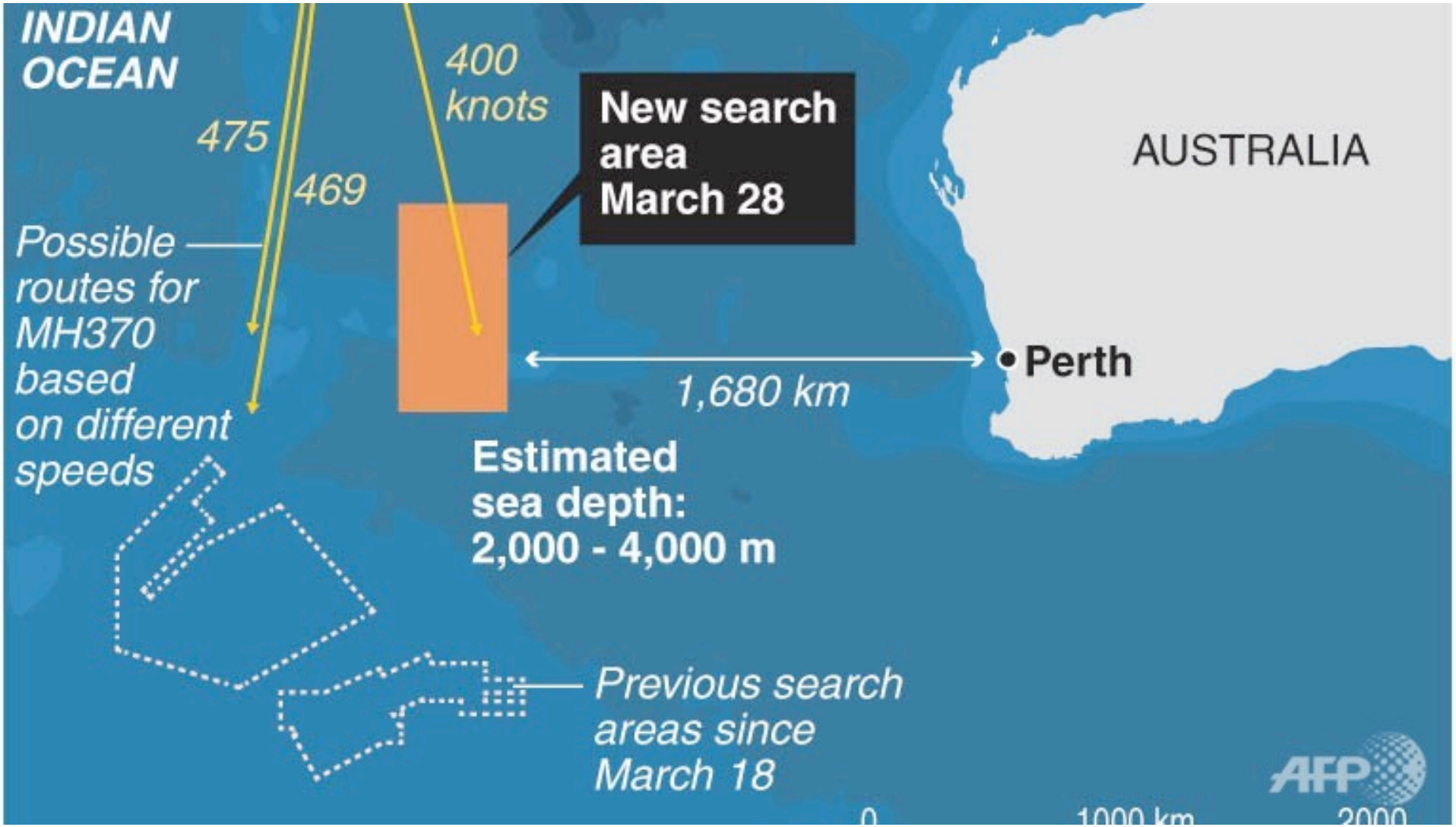
Do we have to correct our previously published p-values for a larger LEE when we add new analyses to our portfolio?

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.

It's Good To have a Region of Interest

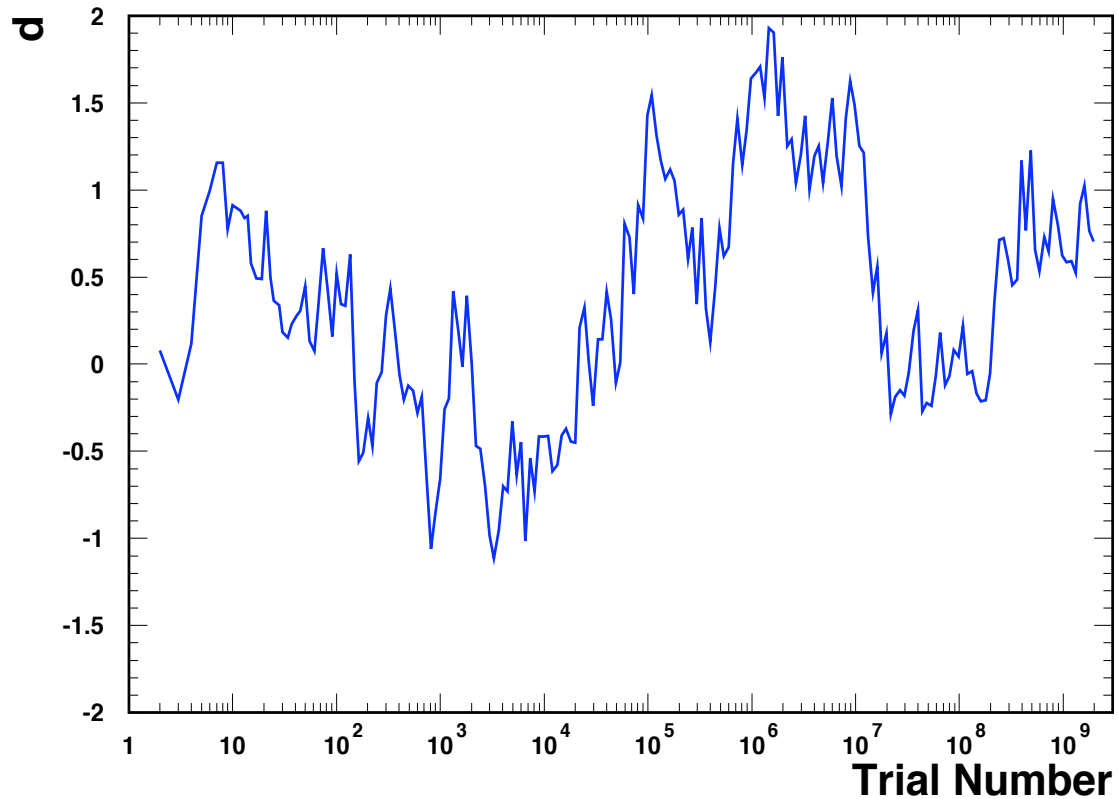


Look ElseWHEN

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"

No Discovery and No Measurement? No Problem!

- Often we are just not sensitive enough (yet) to discover a particular new particle we're looking for, even if it's truly there.

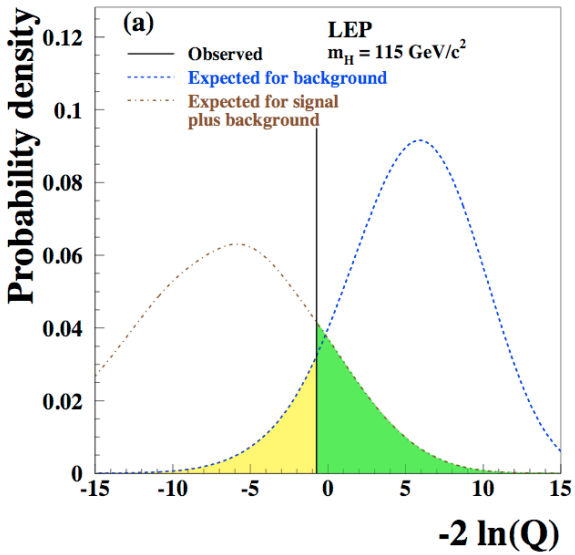
- Or we'd like to test a lot of models (each SUSY parameter choice is a model) and they can't all be true.

- It is our job as scientists to explain what we could have found had it been there. "How hard did you look?"

Strategy -- exclude models: set limits!

- Frequentist
- Semi-Frequentist
- Bayesian

CL_s Limits -- extension of the p-value argument



(apologies for the notation)

p-values:

$CL_b = P(-2\ln Q \geq -2\ln Q_{obs} \mid b \text{ only})$

Green area = $CL_{s+b} = P(-2\ln Q \geq -2\ln Q_{obs} \mid s+b)$

Yellow area = "$1-CL_b$" = $P(-2\ln Q \leq -2\ln Q_{obs} \mid b \text{ only})$

$CL_s \equiv CL_{s+b}/CL_b \geq CL_{s+b}$

Exclude at 95% CL if $CL_s<0.05$

Scale *r* until CL_s=0.05 to get r_{lim}

This step can take significant CPU

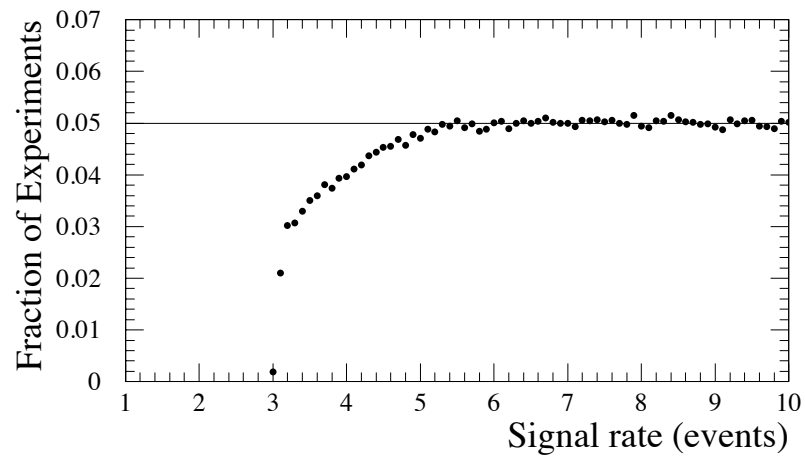
- Advantages:
  - Exclusion and Discovery p-values are consistent. Example -- a 2σ upward fluctuation of the data with respect to the background prediciton appears both in the limit and the p-value as such
  - Does not exclude where there is no sensitivity (big enough search region with small enough resolution and you get a 5% dusting of random exclusions with CL_{s+b})

Overcoverage on Exclusion

Coverage: The "false exclusion rate" should be no more than 1-Confidence Level

In this case, if a signal were truly there, we'd exclude it no more than 5% of the time. "Type-II Error rate" Excluding H_1 when it is true

Exact coverage: 5% error rate (at 95% CL)
Overcoverage: <5% error rate
Undercoverge: >5% error rate



T. Junk, NIM A434 (1999) 435.

Overcoverage introduced by the ratio $CL_s = CL_{s+b}/CL_b$
It's the price we pay for not excluding what we have no sensitivity to.

No similar penalty for the discovery p-value $1\text{-}CL_b$.

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

Rule of Three

From Wikipedia, the free encyclopedia

Rule of three may refer to:

- Rule of three (aviation), a rule of descent in aviation
- 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 (economics), a rule of thumb about major competitors in a free market
- Rule of three (mathematics), a computation method in mathematics
- Rule of three (medicine), for calculating a confidence limit when no events have been observed
- 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

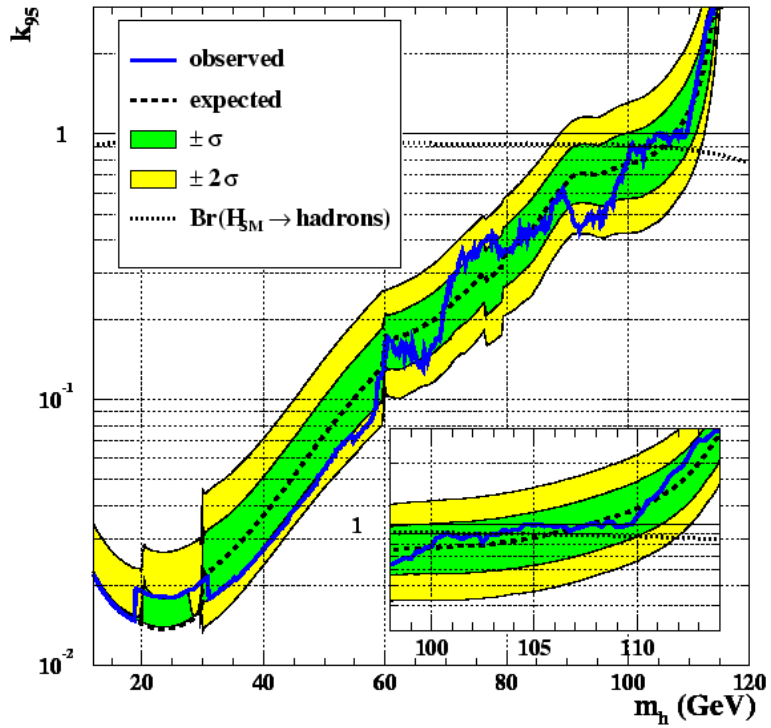
See also

- Rule of thirds, a compositional rule of thumb in photography
- Rule of thirds (diving), a rule of thumb for scuba divers

Exclusion Sensitivity

The usual figure of merit: median expected limit if a signal is truly absent.

again – average limits are pulled by tails in the distribution



OPAL's flavor-independent Higgs search. Some search channels valid at some m_H and not others (MC and ECM availability)

Median expected limit shown as dashed line. Distribution of expected limits assuming no signal shown in green and yellow bands.

An analysis that's optimized for exclusion need not be optimized for discovery.
Need 3 expected events to exclude, but you can discover with 1 event if the background expectation is low enough. MVA's separate events in categories of s/b, so they are the best way to optimize for both.

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"

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

$\{x\}$ = set of observations
$\{\theta\}$ = 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:

describes our belief in the different models parameterized by $\{\nu\}$

Bayes' Theorem

$p(\{\theta\} \mid data)$ 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 Limits

Including uncertainties on nuisance parameters ν

$$L'(data \mid r) = \int L(data \mid r, \nu)\pi(\nu)\,d\nu$$

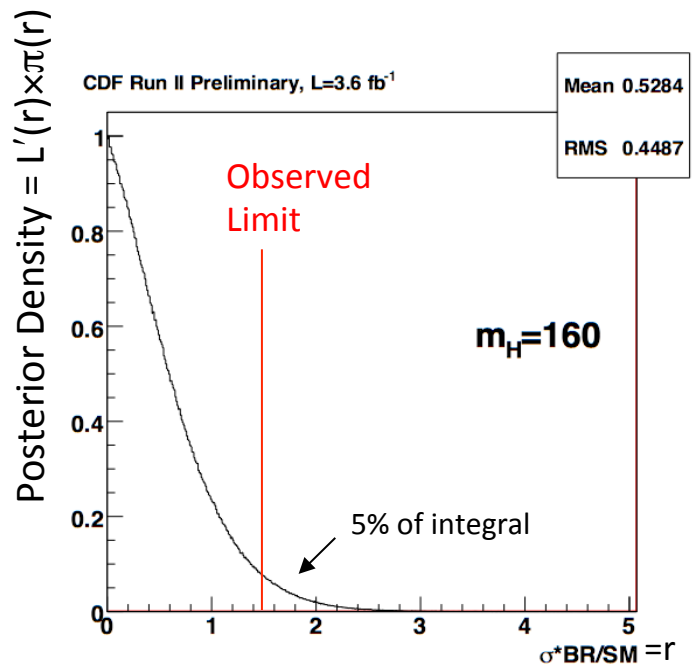
Typically $\pi(r)$ is constant
Other options possible.
Sensitivity to priors a concern.

where $\pi(\theta)$ 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! Look up "Metropolis-Hastings Algorithm" on Wikipedia

Useful for a variety of results:

Limits:

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



Bayesian Cross Section Extraction

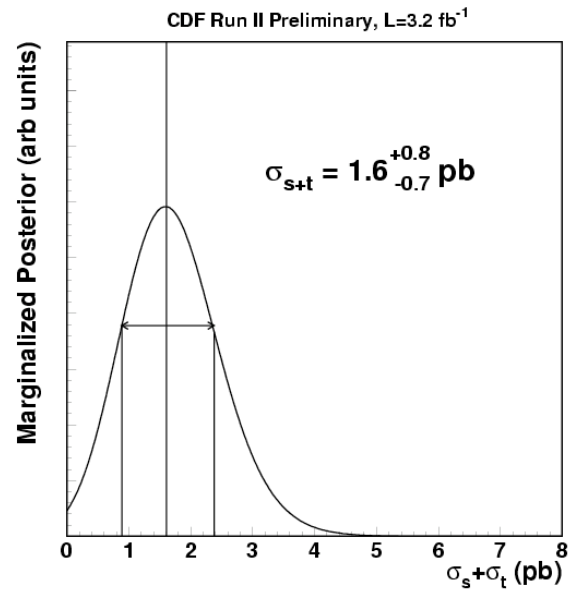
Same handling of nuisance parameters as for limits

$$L'(data \mid r) = \int L(data \mid r, \nu)\pi(\nu)\, d\nu$$

$$0.68 = \frac{\int_{r_{low}}^{r_{high}} L'(data \mid r)\pi(r)dr}{\int_{0}^{\infty} L'(data \mid r)\pi(r)dr}$$

The measured cross section and its uncertainty

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



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! (example: WW TGC@LEP2

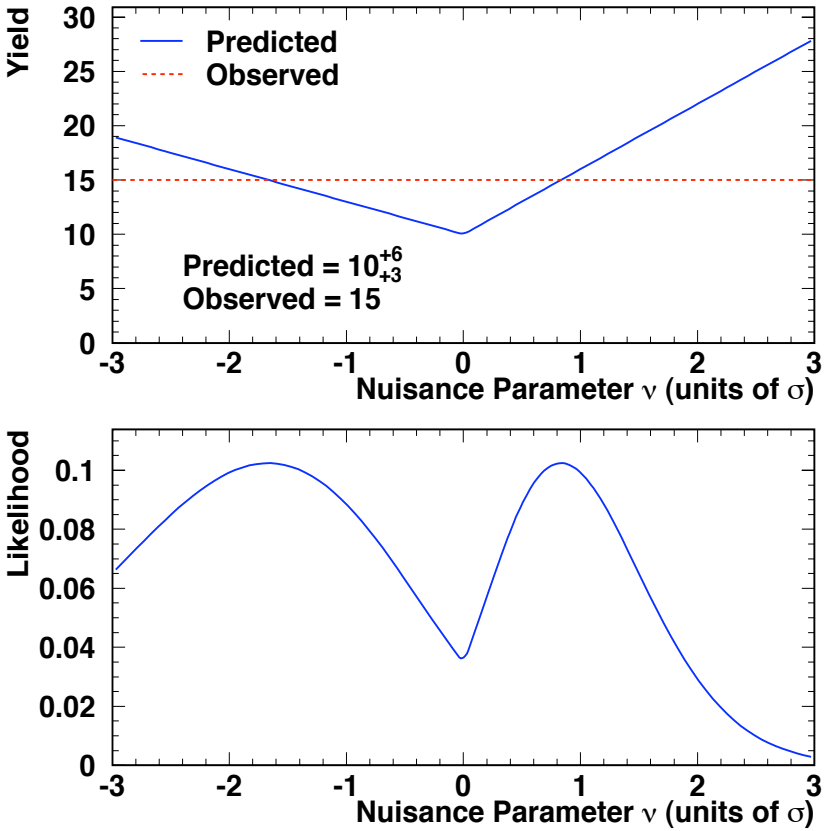
Coping with Systematic Uncertainty

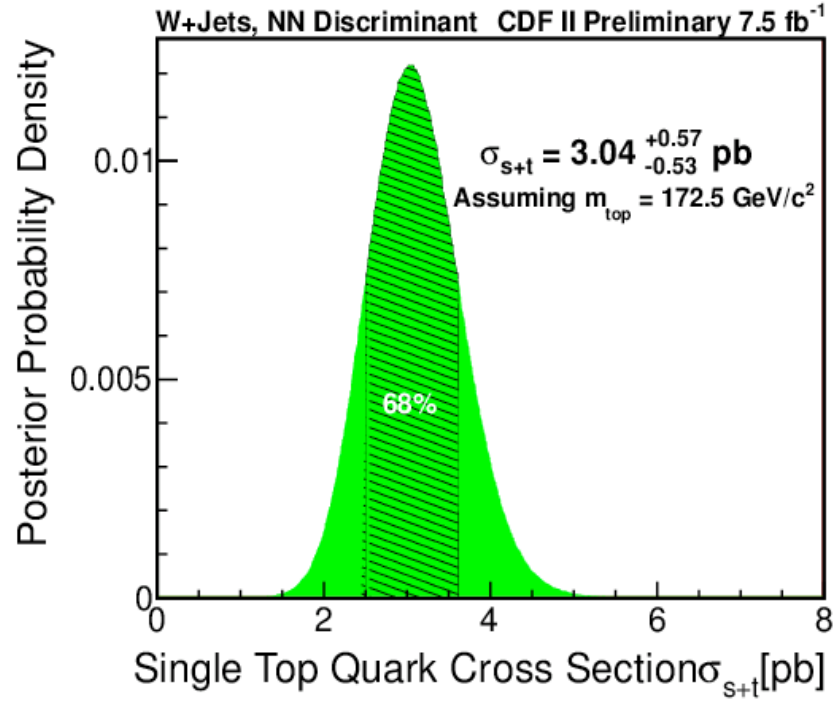
- "Profile:"
  - Maximize L over possible values of nuisance parameters include prior belief densities as part of the χ^2 function (usually Gaussian constraints)

- "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.

  Also known as the "likelihood principle"

- 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.

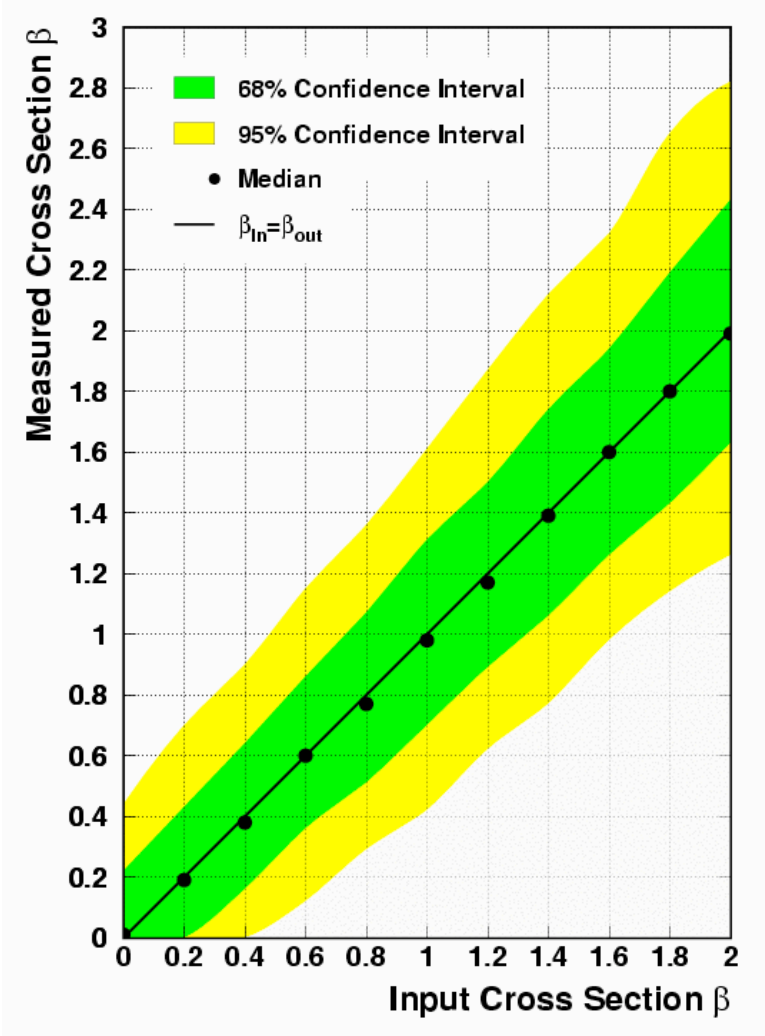
  But even Bayesian/Frequentists have to be Bayesian: use the Prior-Predictive method -- vary the systematics on eachc pseudoexperiment in calculating expected limits. To omit this step ignores an important part of their effects.

Even Bayesians have to be a little Frequentist

We would like to know how the cross section calculations behave in an ensemble of possible experimental outcomes.

Procedure:

- Inject a signal.
- Vary systematics on each pseudoexperiment (which integrates over them in the ensemble)
- Calculate Bayesian cross section for each outcome and plot distribution.
- Black line is the median, not the mean
- Check the width of the distribution against the quoted uncertainties. Specifically, the distribution of (meas-inject)/uncertainty Should be a Unit-width Gaussian (when not up against zero).



This is in fact a Neyman construction! Can do Feldman-Cousins with this (correct for fit biases, if any).

Advantages and Disadvantages of Bayesian Inference

- Advantages:
  - Allows input of *a priori* knowledge:
    - positive cross-sections
    - positive masses
  - Gives you "reasonable" confidence intervals which don't conflict with *a priori* knowledge
  - Easy to produce cross-section limits
  - Depends only on observed data and not other possible data
  - No other way to treat uncertainty in model-derived parameters
- Disadvantages:
  - Allows input of *a priori* knowledge (AKA "prejudice") (be sure not to put it in twice...)
  - Results are metric-dependent (limit on cross section or coupling constant? -- square it to get cross section).
  - Coverage not guaranteed
  - Arbitrary edges of credibility interval (see freq. explanation)

Extras

A Dilemma – Can't we test just *one* model?

Something experimentalists come up with from time to time:

- Make distributions of every conceivable reconstructed quantity
- Compare data with Standard Model Predictions
- Use to test whether the Standard Model can be excluded
- Example: CDF's Global Search for New Physics Phys.Rev. D **79** (2009) 011101

The case *for* doing this:

- We might miss something big and obvious in the data if we didn't
- Searches that are motivated by specific new physics models may point us away from actual new physics.

More potential for discovery if you look in more places.

Example: Discovery of Pluto. Calculations from Uranus's orbit perturbations were flawed, but if you look in the sky long enough and hard enough you'll find stuff. Even without calculations it's still a good idea to look in the sky for planetoids.

Testing Just One Model – Difficulties in Interpretation

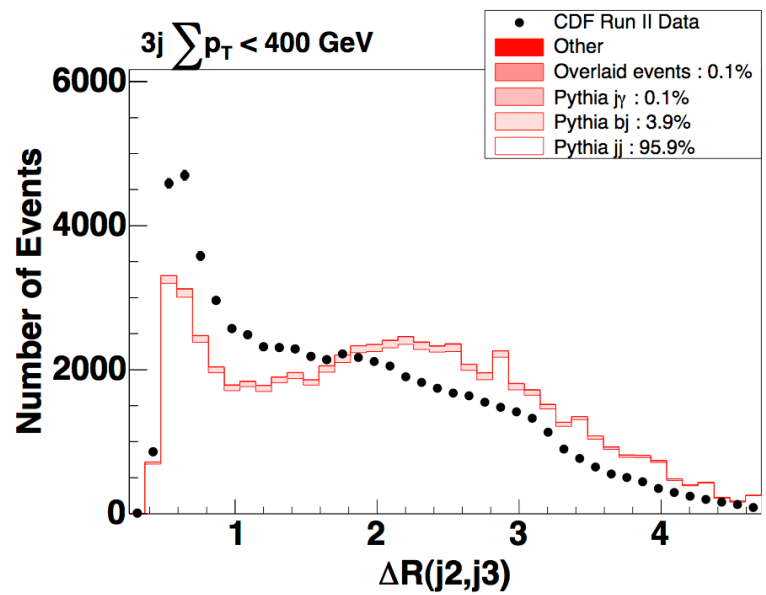
- Look in enough places and you'll eventually find a statistical fluctuation -- you may find some new physics, but probably also some statistical fluctuations along the way.

This is straightforward to correct for – called the "Trials Factor" or the "Look Elsewhere Effect", or the effect of multiple testing. To be discussed later.

- More worrisome is what to do when systematic flaws in the modeling are discovered.

Example: angular separation between the two least energetic jets in three-jet events.

Not taken as a sign of new physics, but rather as an indication of either generator (Pythia) or detector simulation (CDF's GEANT simulation) mismodeling. Or an issue with modeling trigger biases. Each of these is a responsibility of a different group of people.



Phys.Rev. D79 (2009) 011101

Testing Just One Model – Difficulties in Interpretation

- Need a definition of what counts as "interesting" and what's not. Already, using triggered events at a high-energy collider is a motivation for seeking highly-energetic processes, or signatures of massive new particles previously inaccessible.

- Analyzers chose to make ΣP_T distributions for all topologies and investigate the high ends, seeking discrepancies.

We just lost some generality! Some new physics may now escape detection.

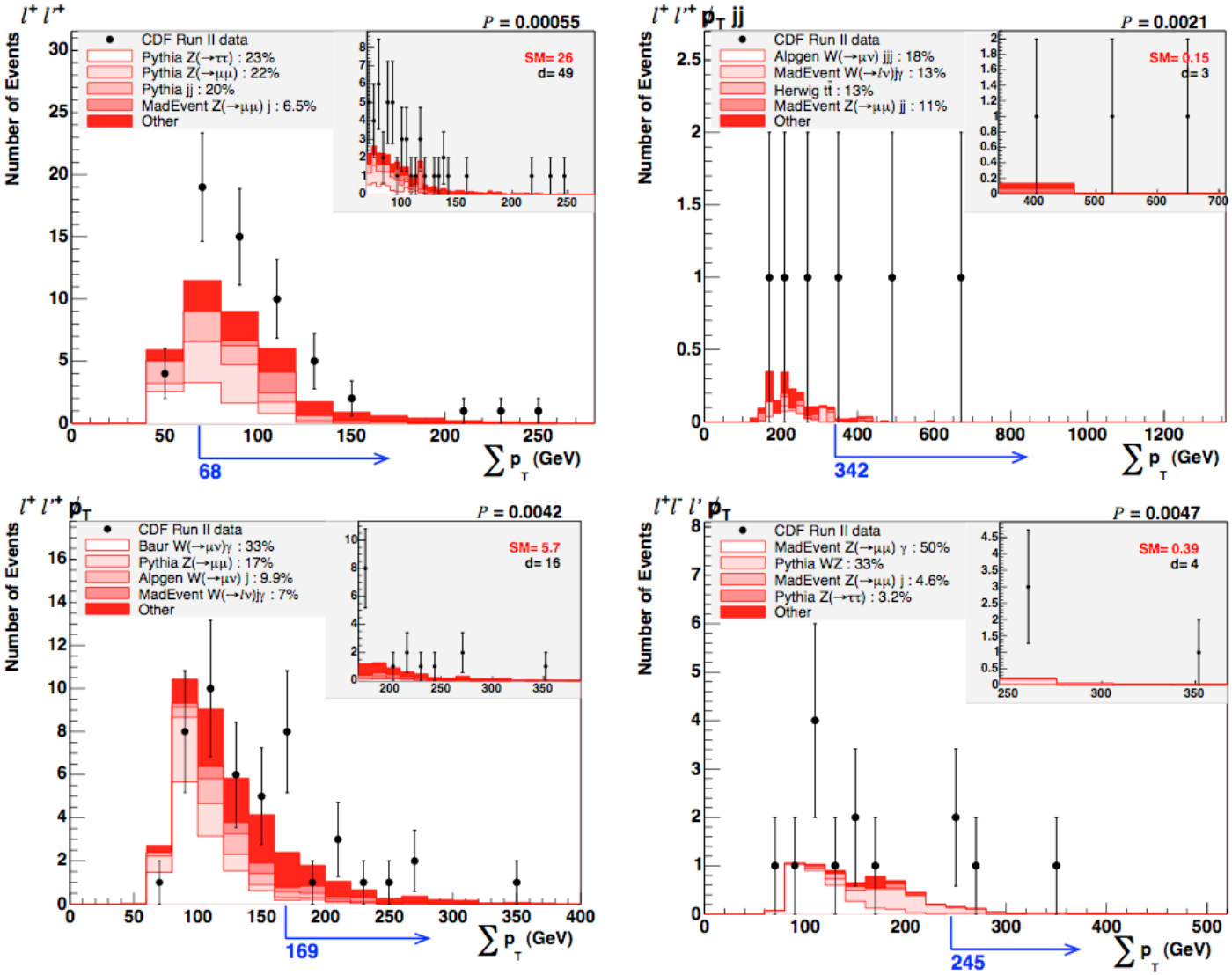
But we now have alternate hypotheses – no longer are we just testing the SM (really our clumsy Monte Carlo representation of it).

Boxed into a corner trying to test just one model

- Of course our MC is wrong (that's what systematic uncertainty is for)
- Of course the SM is incomplete (but is it enough to describe our data?)

But without specifying an alternative hypothesis, we cannot exclude the null hypothesis ("maybe it's a fluctuation. Maybe it's mismodeling.")

The Most Discrepant ΣP_T distributions



like-sign dileptons, missing p_T – modeling of fakes and mismeasurement is always a question.

Phys.Rev. D79 (2009) 011101

Searching for Everything All at Once

- A global search also is less optimal than a targeted search
  - Targeted searches can take advantage of more features of the signal (and background) processes than just particle content and ΣP_T.
  - The Global search suffers from a much larger Look-Elsewhere Effect
  - The Global search may not benefit as much from sideband constraints of backgrounds, although CDF's did adjust some non-new-physics nuisance parameters to fit the data the best.

- Global Search distributions must be hidden from blind analyzers – they unblind everything.

In practice, this isn't much of a problem due to different event selection criteria

In spite of all of the difficulties, it is *still* a good idea to do this. We absolutely do not want to miss anything.

But a signal of new physics would have to be pretty big for us to stumble on it. It's hard to manufacture serendipity.

Consequences of Not Fitting

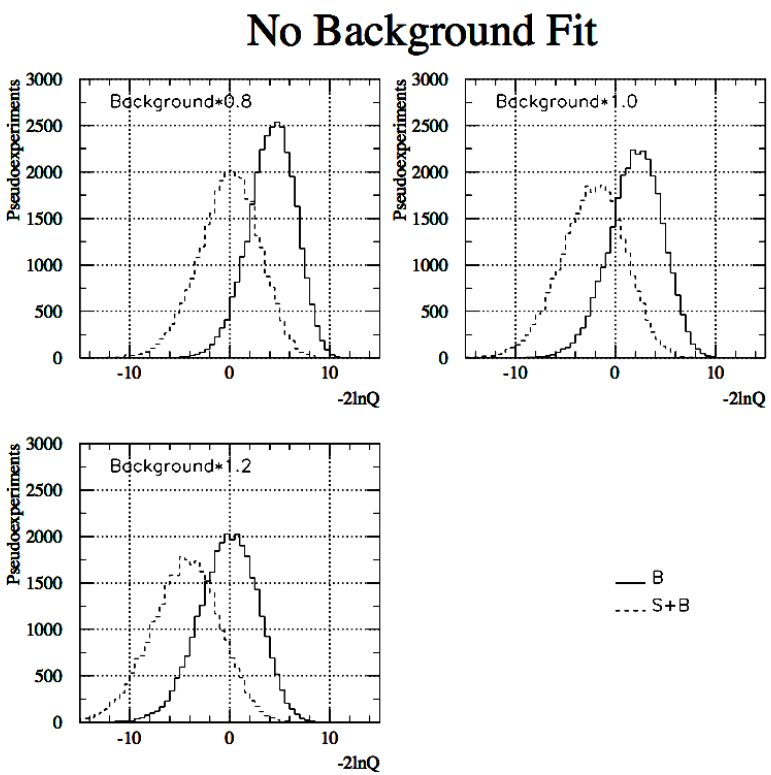
See Favara and Pieri, **hep-ex/9706016**

They found that channels, or bins within channels are better off being neglected in the interpretation of an analysis in order to optimize its sensitivity.

If the systematic uncertainty on the background b exceeds the expected signal s, then that search isn't of much use. Fitting backgrounds helps constrain them however, and sidebands with little or no signal still provide useful information, but you have to fit to get it.

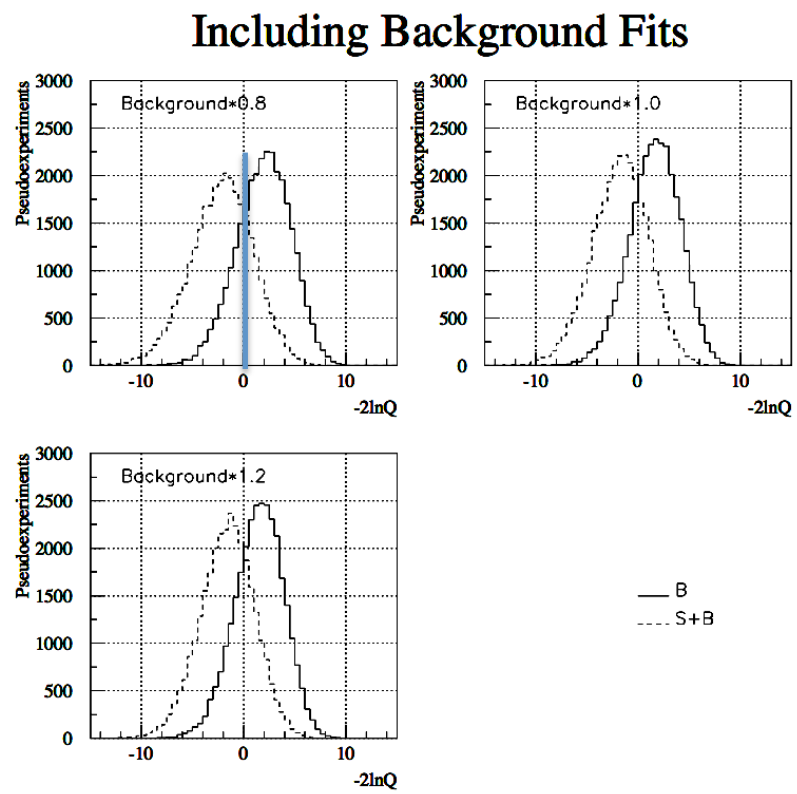
We also initially tried running LEP-style CL_s programs on the Tevatron Higgs searches, and got limits that were a factor of two worse than with fitting. The limits with fitting matched older ones done by a Bayesian prescription (more on that later)

Fitting Nuisance Parameters to Reduce Sensitivity to Mismodeling

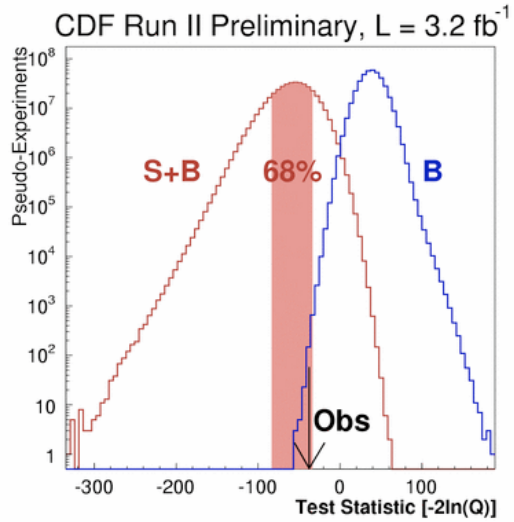


Means of PDF's of -2lnQ very sensitive to background rate estimation.

Still some sensitivity in PDF's residual due to prob. of each outcome varies with bg estimate.



Fitting and Fluctuating



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

- Monte Carlo simulations are used to get p-values.
- Test statistic -2lnQ is not uncertain for the data.
- Distribution from which -2lnQ is drawn is uncertain!

- Nuisance parameter fits in numerator and denominator of -2lnQ **do not incorporate systematics into the result**.
  Example -- 1-bin search; all test statistics are equivalent to the event count, fit or no fit.

- Instead, we fluctuate the probabilities of getting each outcome since those are what we do not know. Each pseudoexperiment gets random values of nuisance parameters.

- Why fit at all? It's an optimization. Fitting reduces sensitivity to the uncertain true values and the fluctuated values. For stability and speed, you can choose to fit a subset of nuisance parameters (the ones that are constrained by the data). Or do constrained or unconstrained fits, it's your choice.

- If not using pseudoexperiments but using Wilk's theorem, then the fits are important for correctness, not just optimality.

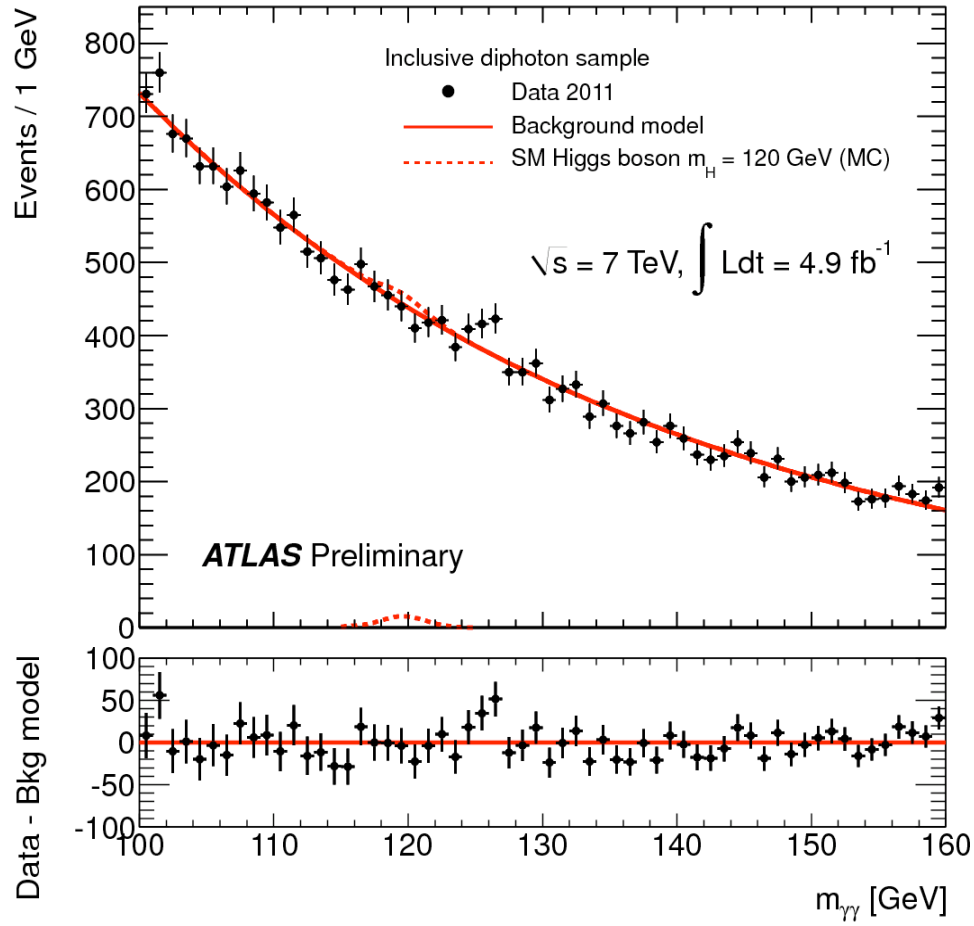
An interesting Bias Bill Murray Showed at The Next Stretch of the Higgs Magnificent Mile Conference

Seek a bump on a smooth background
Example: LHC (or Tevatron) H→γγ search.

Allow m_H to float and pick the m_H that maximizes the fitted cross section.

The fitted cross section will be biased upwards and the position resolution of "lucky" outcomes will be worse than unlucky ones even if a signal is truly present.

Why? A true bump can coalesce with a fluctuation either to the left or to the right of the bump (two chances to fluctuate upwards).



Effect can be substantial! Calibrate with simulated experimental outcomes (FC).

Choosing a Region of Interest

- I do not have a foolproof prescription for this, just some thoughts.

- Analyses are designed to optimize sensitivity, but LEE dilutes sensitivity. There is a penalty for looking for many independently testable models. Can we optimize this?

- But you should always do a search anyway! If you expect to be able to test a model, you should.

- Testing previously excluded models? We do this anyway, just in case some new physics shows up in a way that evaded the previous test.

- There is no such thing as a model-independent search. Merely building the LHC or the Tevatron means we had something in mind. And the SM (or just our implementation of it) is wrong, but possibly not in a way that is both interesting and testable.

Blind Analysis

- Fear of intentional or even unintentional biasing of results by experimenters modifying the analysis procedure after the data have been collected.

- Problem is bigger when event counts are small -- cuts can be designed around individual observed events.

- Ideal case -- construct and optimize experiment before the experiment is run. Almost as good -- just don't look at the data

- Hadron collider environment requires data calibration of backgrounds and efficiencies

- Often necessary to look at "control regions" ("sidebands") to do calibrations. Be careful not to look "inside the box" until analysis is finalized. Systematic uncertainties must be finalized, too!

Non-Blind Analyses

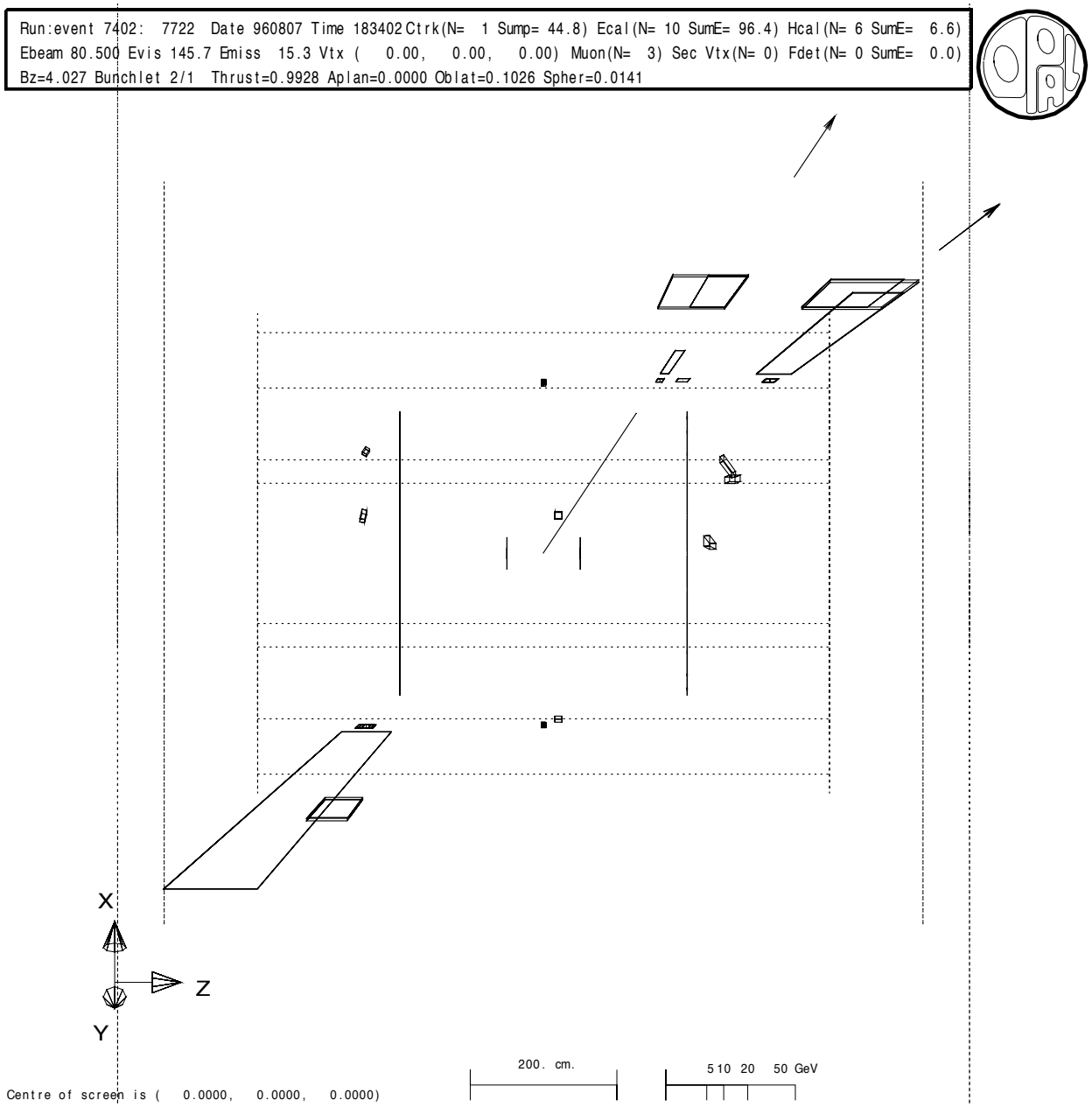
- More of a concern, but many factors keep analyzers from selecting (or excluding) only their favorite events

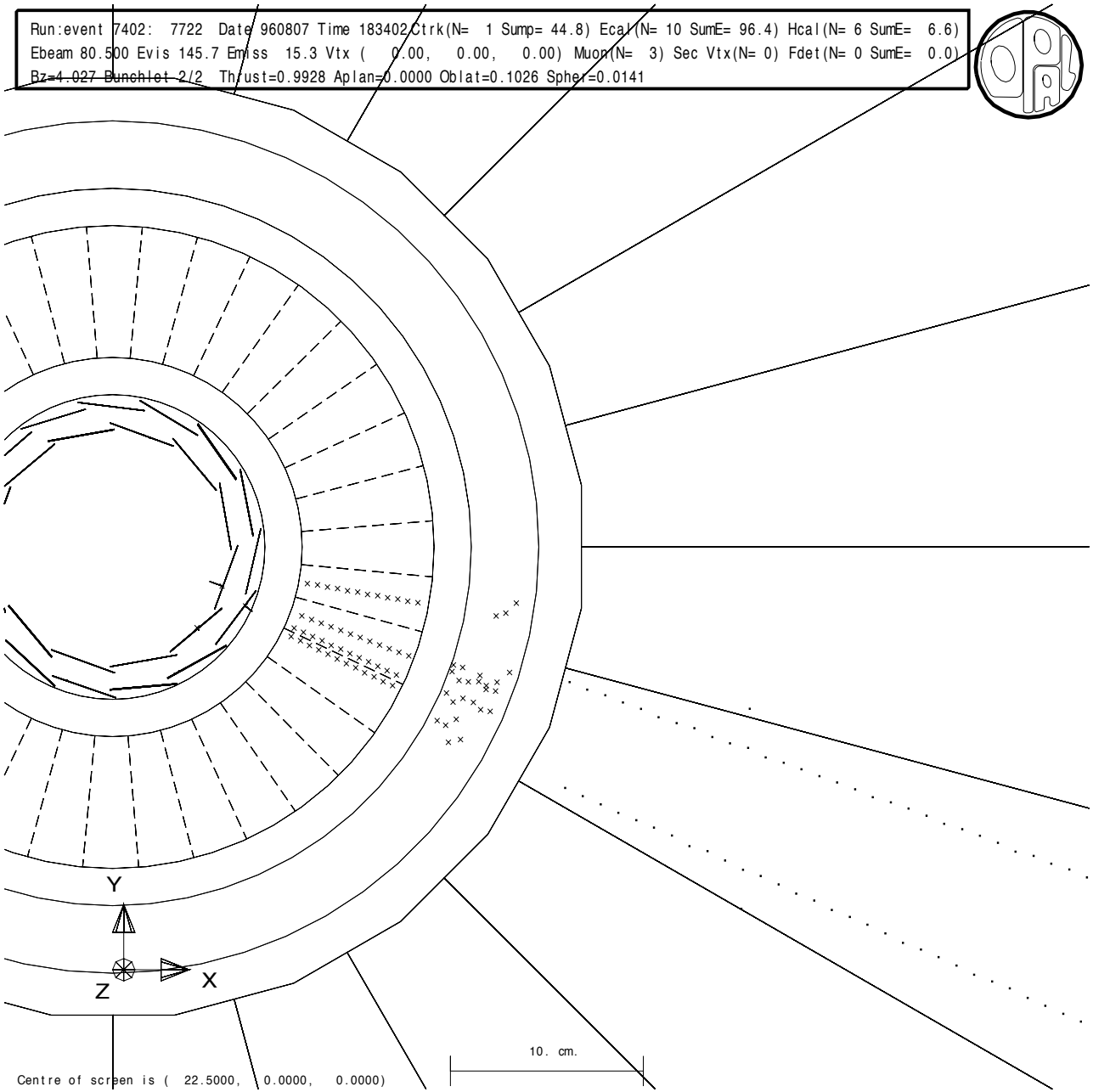
  - Standardized jet definition. Jet energy scale, resolution, modeling is typically approved for a small number of jet algorithms and parameter choices
  - Jet and lepton E_T and η requirements are typically standardized so previous signal efficiency and background estimate tools can be re-used.
  - Changes to an analysis – new selection requirements, or new MVA's must be justified in terms of improved sensitivity (better discovery chances, lower expected limits, or smaller cross section uncertainties)

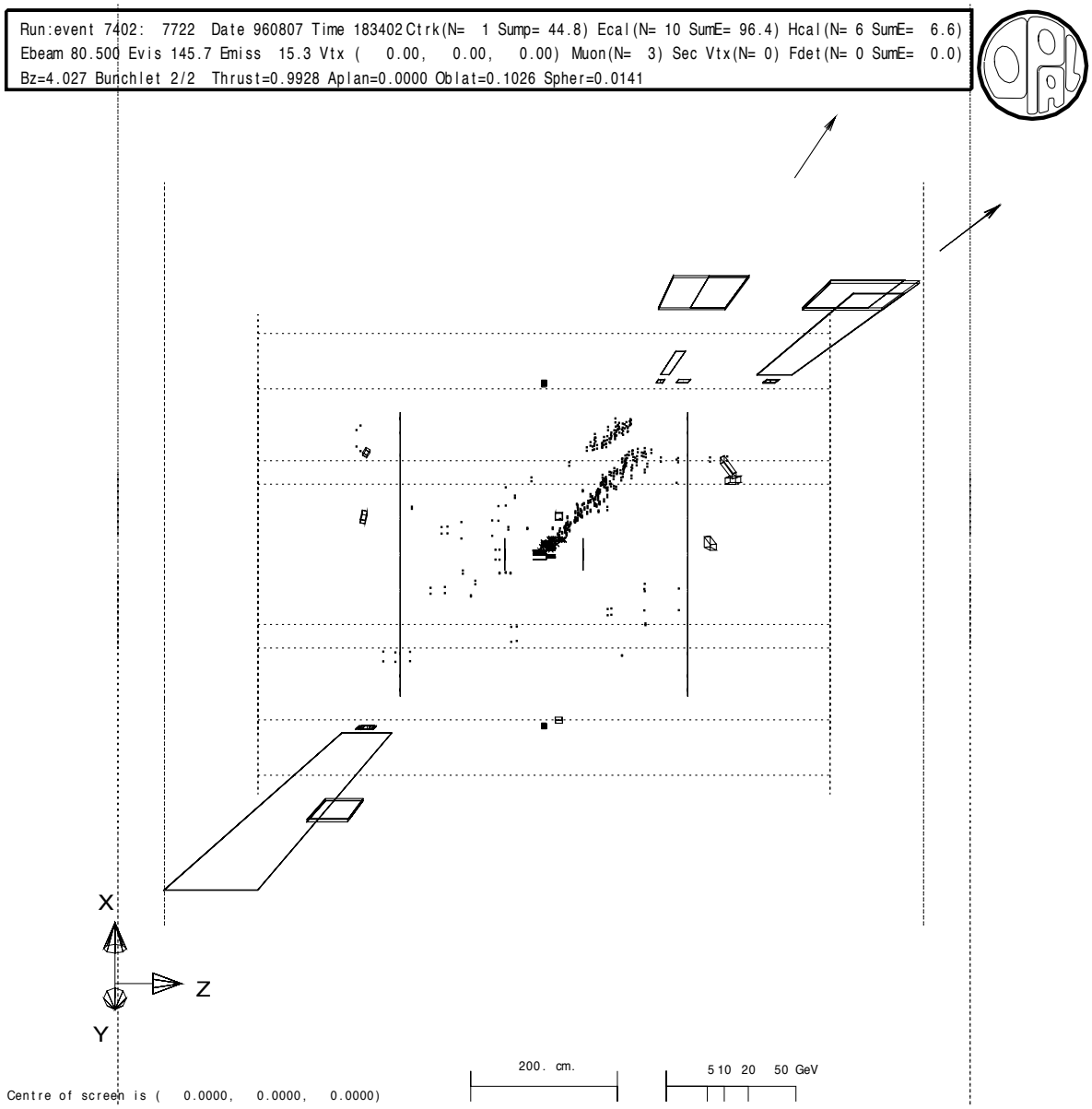
  -- Still possible to devise many improvements to an analysis, all of which improve the sensitivity, but only those that push the observed result in a desired direction are chosen. We frequently discuss all kinds of improvements so it is not that frequent that we throw a good one away for an unjustifiable reason.

  -- Always a concern – Analyzers keep working and fixing bugs until they get the answer they like, and then stop. We would like review to be exhaustive!

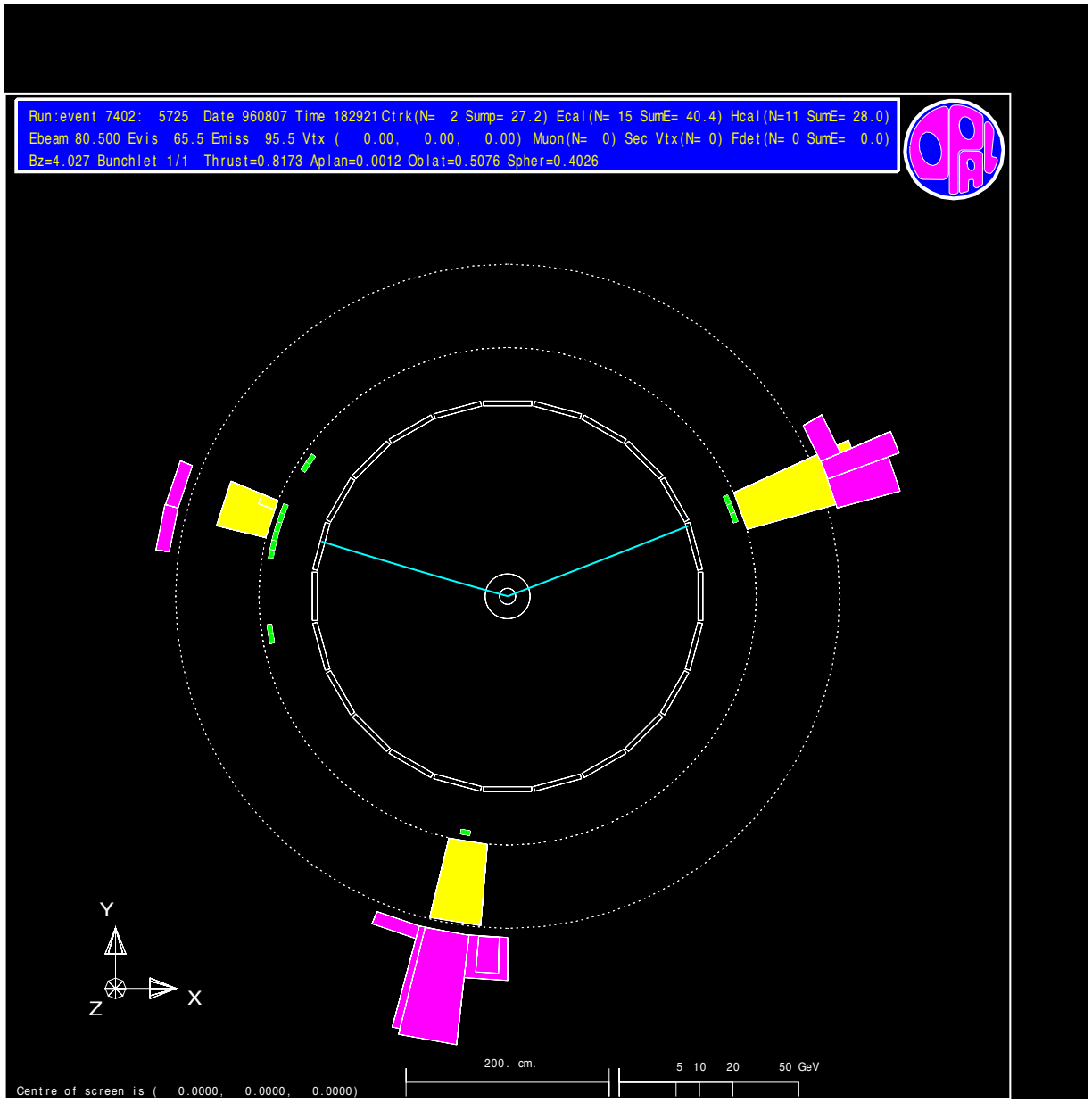
  A special case – re-doing an analysis with a slightly larger data set.
  Good practice for future work. If a flaw was found in the previous work, all the better!


Run:event 7402: 7722 Date 960807 Time 183402 Ctrk(N= 1 Sump= 44.8) Ecal(N= 10 SumE= 96.4) Hcal(N= 6 SumE= 6.6)
Ebeam 80.500 Evis 145.7 Emiss 15.3 Vtx (0.00, 0.00, 0.00) Muon(N= 3) Sec Vtx(N= 0) Fdet(N= 0 SumE= 0.0)
Bz=4.027 Bunchlet 2/1 Thrust=0.9928 Aplan=0.0000 Oblat=0.1026 Spher=0.0141
X
Z
Y
200. cm.
5 10 20 50 GeV
Centre of screen is (0.0000, 0.0000, 0.0000)


Run:event 7402: 7722 Date 960807 Time 183402 Ctrk(N= 1 Sump= 44.8) Ecal(N= 10 SumE= 96.4) Hcal(N= 6 SumE= 6.6)
Ebeam 80.500 Evis 145.7 Emiss 15.3 Vtx (0.00, 0.00, 0.00) Muon(N= 3) Sec Vtx(N= 0) Fdet(N= 0 SumE= 0.0)
Bz=4.027 Bunchlet 2/2 Thrust=0.9928 Aplan=0.0000 Oblat=0.1026 Spher=0.0141
Y
X
Z
10. cm.
Centre of screen is (22.5000, 0.0000, 0.0000)


Run:event 7402: 7722 Date 960807 Time 183402 Ctrk(N= 1 Sump= 44.8) Ecal(N= 10 SumE= 96.4) Hcal(N= 6 SumE= 6.6)
Ebeam 80.500 Evis 145.7 Emiss 15.3 Vtx (0.00, 0.00, 0.00) Muon(N= 3) Sec Vtx(N= 0) Fdet(N= 0 SumE= 0.0)
Bz=4.027 Bunchlet 2/2 Thrust=0.9928 Aplan=0.0000 Oblat=0.1026 Spher=0.0141
X
Z
Y
200. cm.
5 10 20 50 GeV
Centre of screen is (0.0000, 0.0000, 0.0000)

G. Wilson's event from OPAL

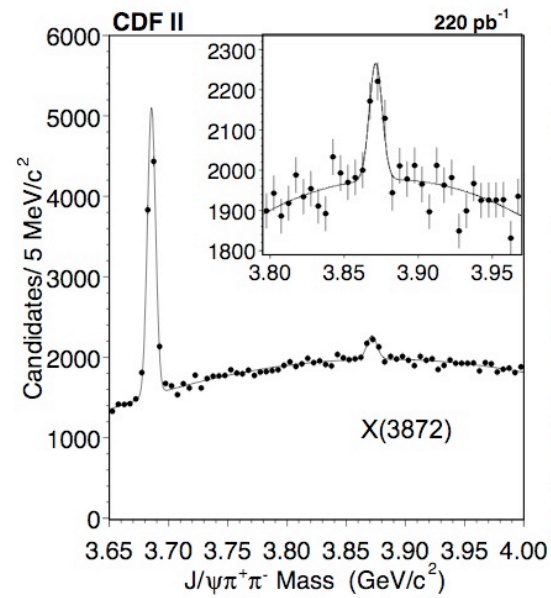


Asymmetric Uncertainties and Priors

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

Treatment of Asymmetric Uncertainties

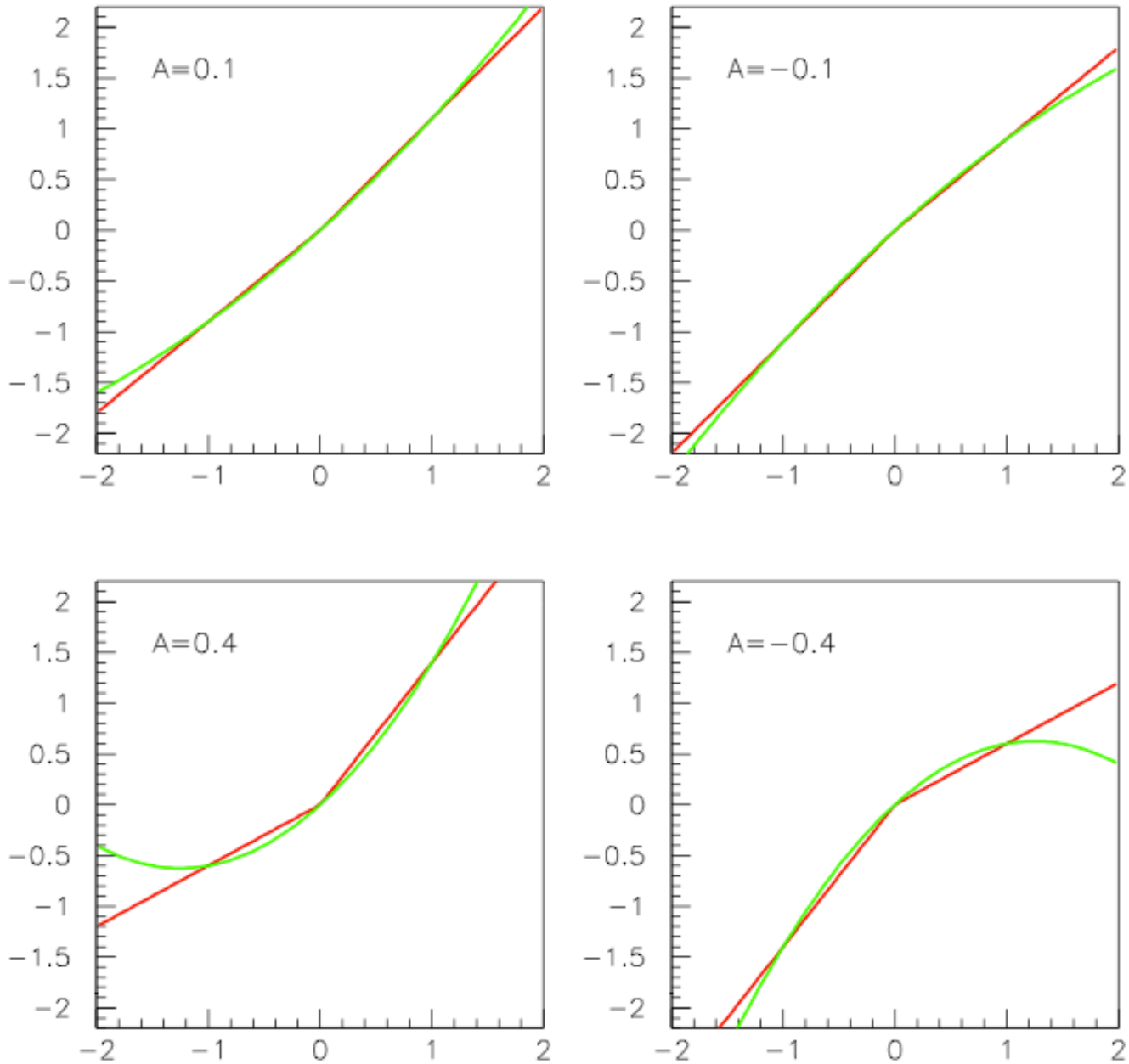
These cases are pretty clear – the underlying parameter, the energy scale, has a (Gaussian? Your choice) distribution, while it has a nonlinear, possibly non-monotonic *impact* on the model prediction.

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.

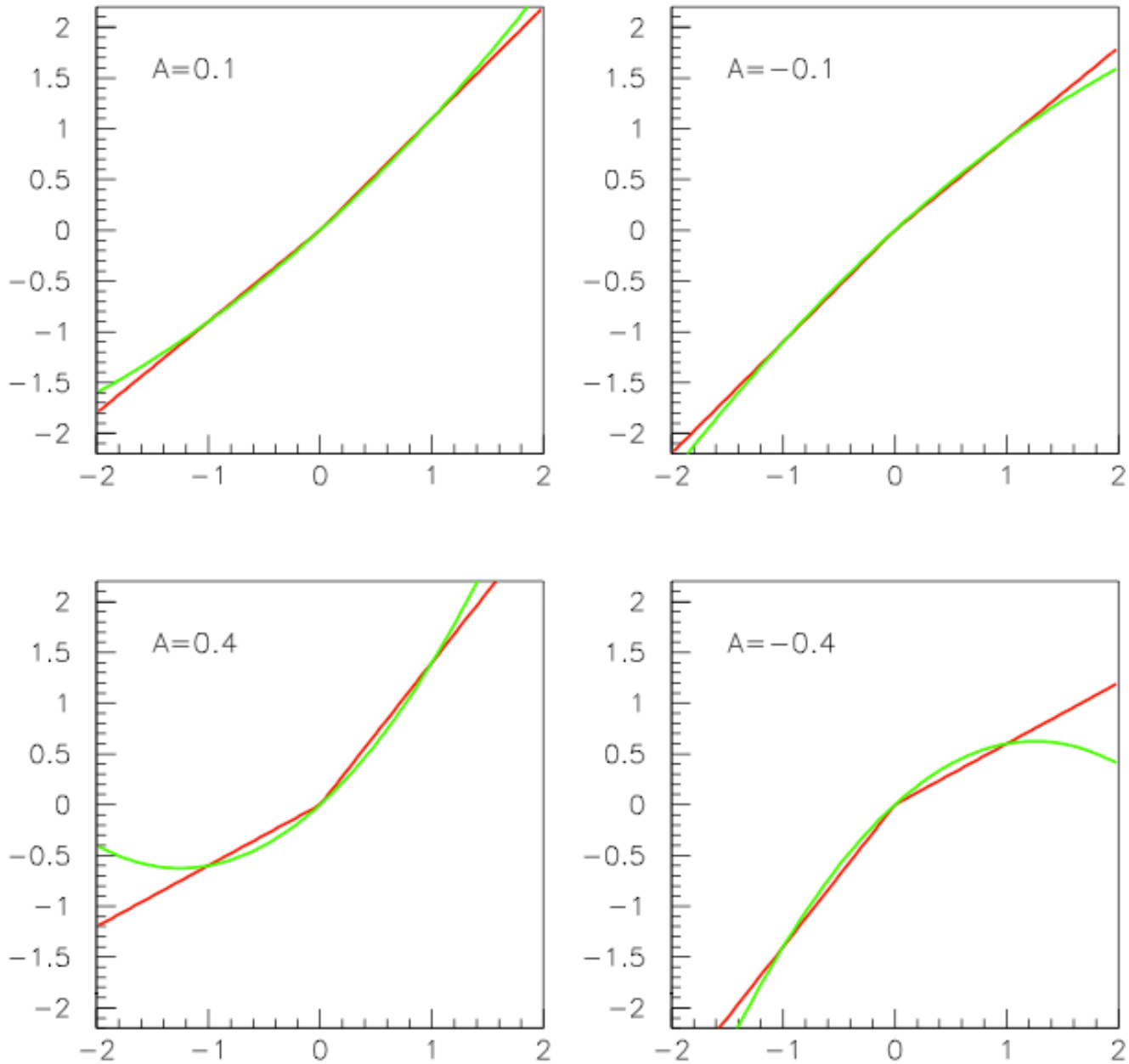
See R. Barlow,
"Asymmetric Systematic Errors", arXiv:physics/0306138
"Asymmetric Statistical Errors", arXiv:physics/0406120

Quadratic Impacts of Asymmetric Uncertainties



R. Barlow

Quadratic Impacts of Asymmetric Uncertainties



R. Barlow