Statistical Methods for Experimental Particle Physics

Tom Junk

Fermilab

Pauli Lectures on Physics
ETH Zürich
30 January — 3 February 2012

Day 2:

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

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.

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

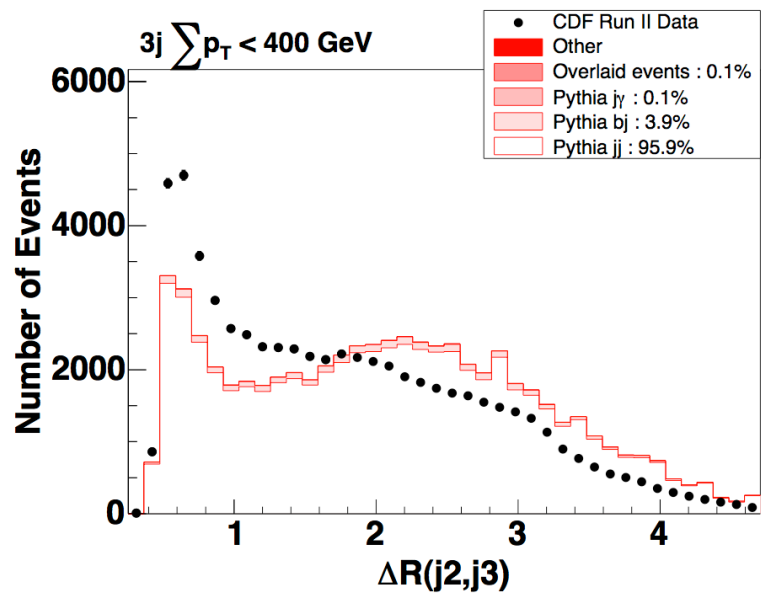
- Look in enough places and you'll eventually find a statistical fluctuation -- you may find some new physics, but probably also some statistical fluctuations along the way.

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

- What do you do when you see a discrepancy between data and prediction?
  1. Attribute it to a statistical fluctuation
  2. Attribute it to a systematic defect in the modeling of SM physics processes, the detector, or trigger and event selection effects
     - No matter how hard we work, there will always be some residual mismodeling.
     - Collect more and more data, and smaller and smaller defects in the modeling will become visible
  3. Attribute it to new physics

- Looking in many distributions will inevitably produce situations in which 1 and 2 are the right answer. Possibly 3, but if we only knew the truth! Trouble is, we'd always like to discover new physics as quickly as possible, so there is a reason to point out those discrepancies that are only marginal.

- In order to compute the look-elsewhere-effect, we need to have a prescription for how to respond to each possible discrepancy in any distribution.

  -- Run Monte Carlo simulations of possible statistical fluctuations and run each through the same interpretation machinery as used for the data to characterize its performance

Testing Just One Model – Difficulties in Interpretation

- Systematic effects in the modeling or new physics? ("old" physics vs. "new" physics)

- Use the data to constrain the "old" physics and improve the modeling
  - Tune Monte Carlo models to match data in samples known not to contain new physics.
  - Already a problem – how do we know this?
  - Examples: lower-energy colliders, e.g. LEP and LEP2, are great for tuning up simulations.

- Extrapolation of modeling from control samples to "interesting" signal samples – this step is fraught with assumptions which are guaranteed to be at least a little bit incorrect.

- But extrapolations with assumptions are useful! So we assign uncertainties, which we hope cover the differences between our assumptions and the truth

- But in a "global" search, it is less clear what's "signal" and what's "background". Which discrepancies can be used to "fix the Monte Carlo" and which are interesting enough to make discovery claims? It's a judgement call.

- Need to formalize judgement calls so that they can be simulated many times!

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.

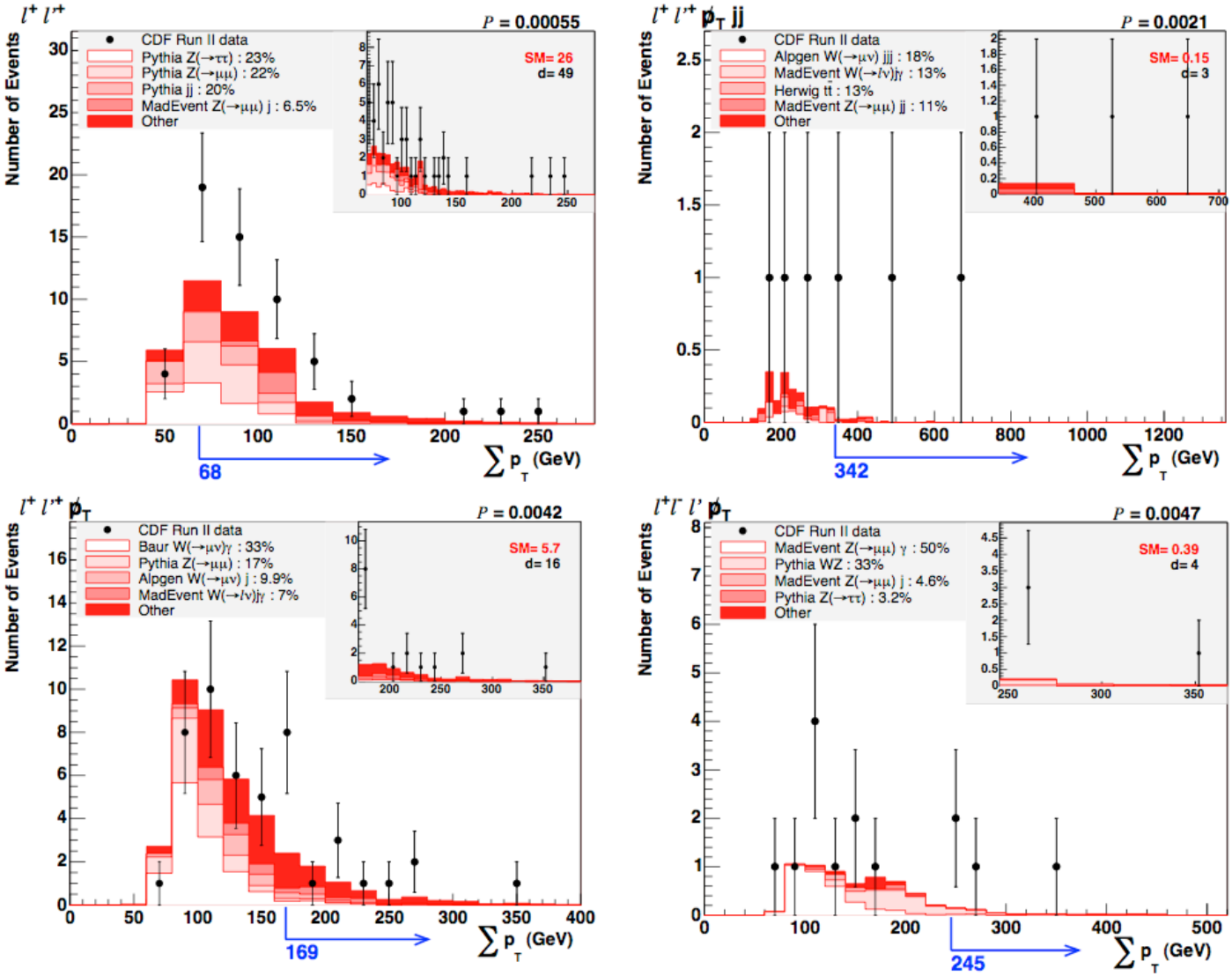
But we now have alternate hypotheses – no longer are we just testing the SM (really our clumsy Monte Carlo representation of it).

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.

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

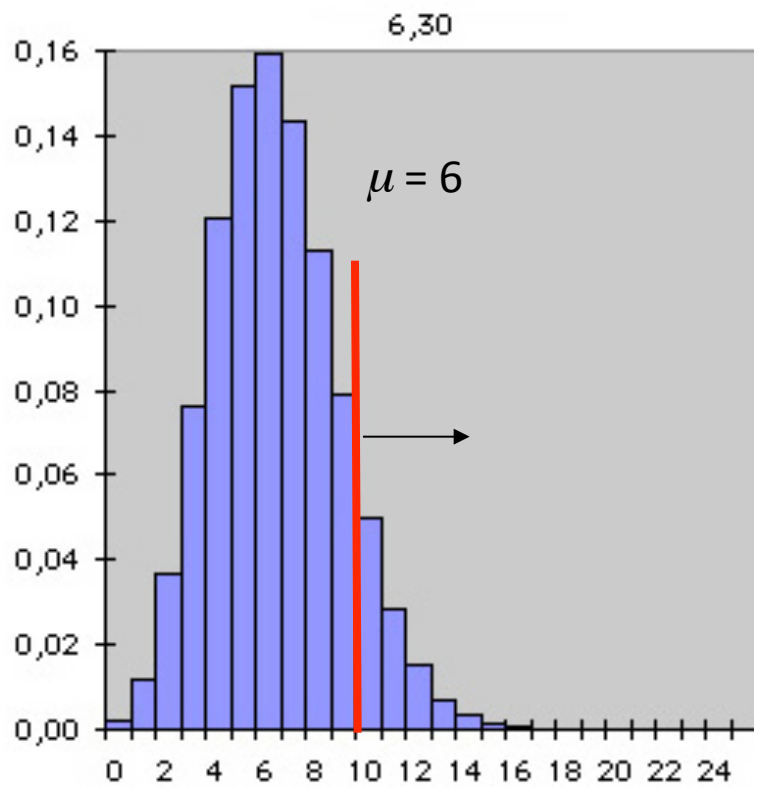
$p(n \geq n_{obs} | H_0)$

Example:

H_0: $b = \mu = 6$

$n_{obs} = 10$

p-value = 0.0839



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

But many often say that.
Especially the popular media!

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). 1 of 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, not very common in HEP, but people will understand)

- **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.

Common Standards of Evidence

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

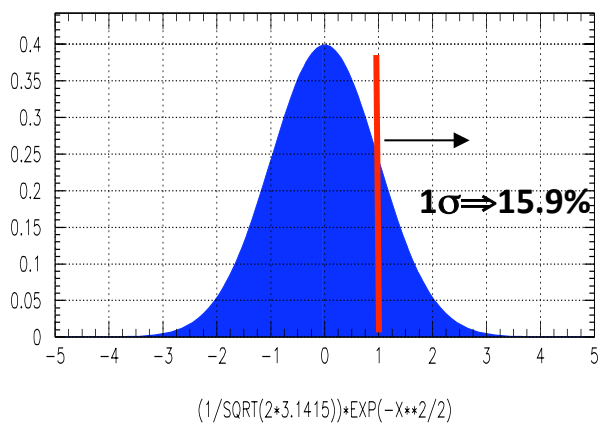
The number of "sigma" is called a "*z*-value" and is just a translation of a p-value using the integral of one tail of a Gaussian

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

Why 5 Sigma for Discovery?

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

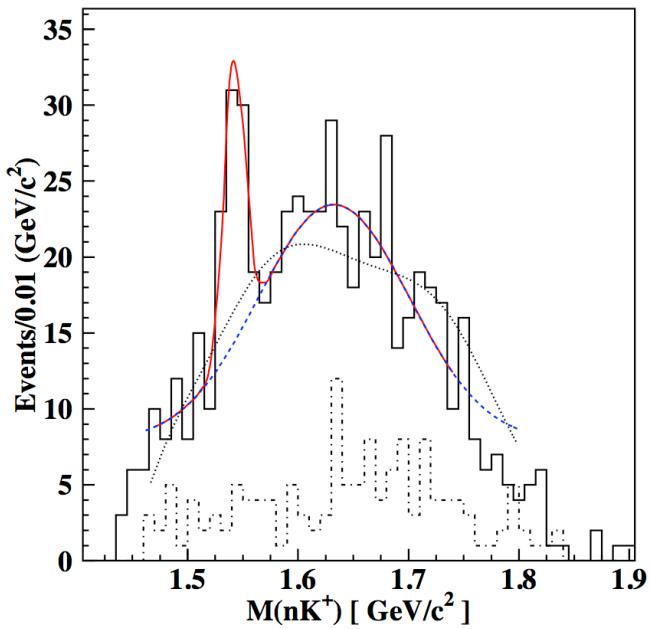
Meant to account for the Look Elsewhere Effect. A physicist estimated how many histograms would be looked at, and wanted to keep the error rate low.

Also too many 2σ and 3σ effects "go away" when more data are collected.

My personal opinion – not all estimations of systematic uncertainties are perfect – 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)).

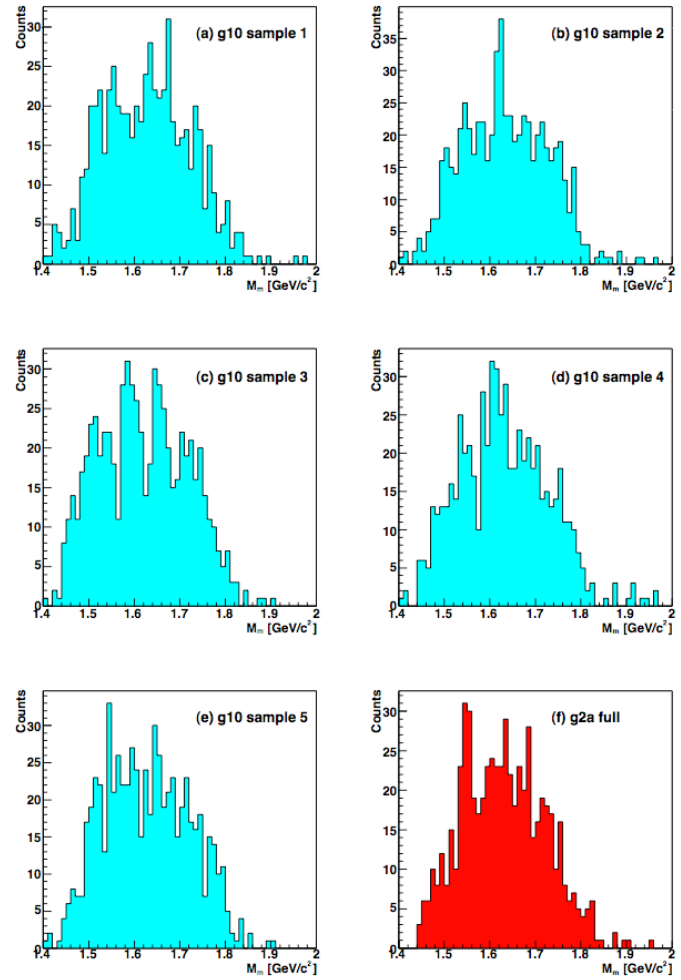
A Cautionary Tale – The Pentaquark "Discoveries"



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

Significance = 5.2 ± 0.6 σ

Watch out for the background function parameterization!

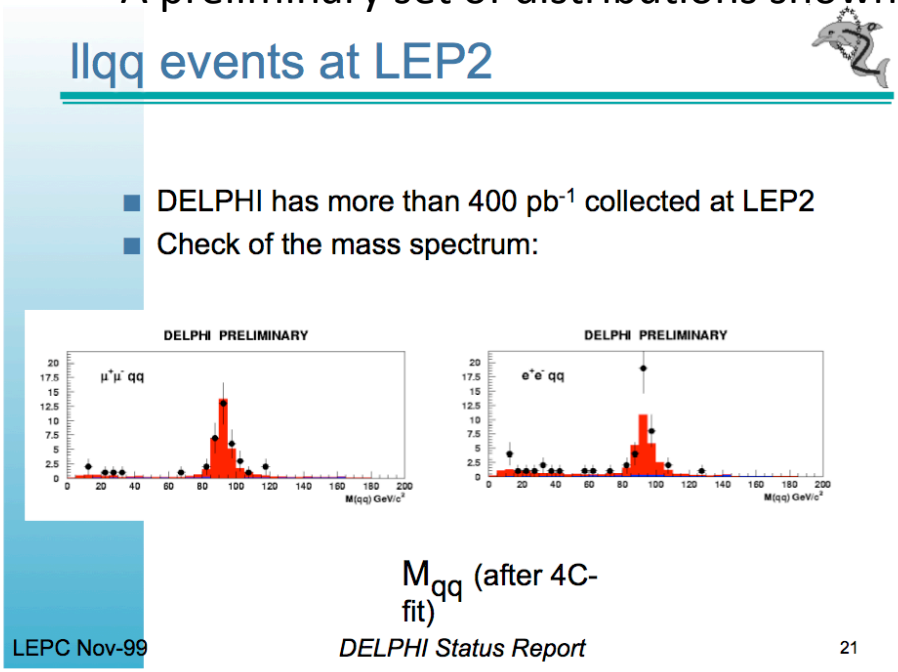


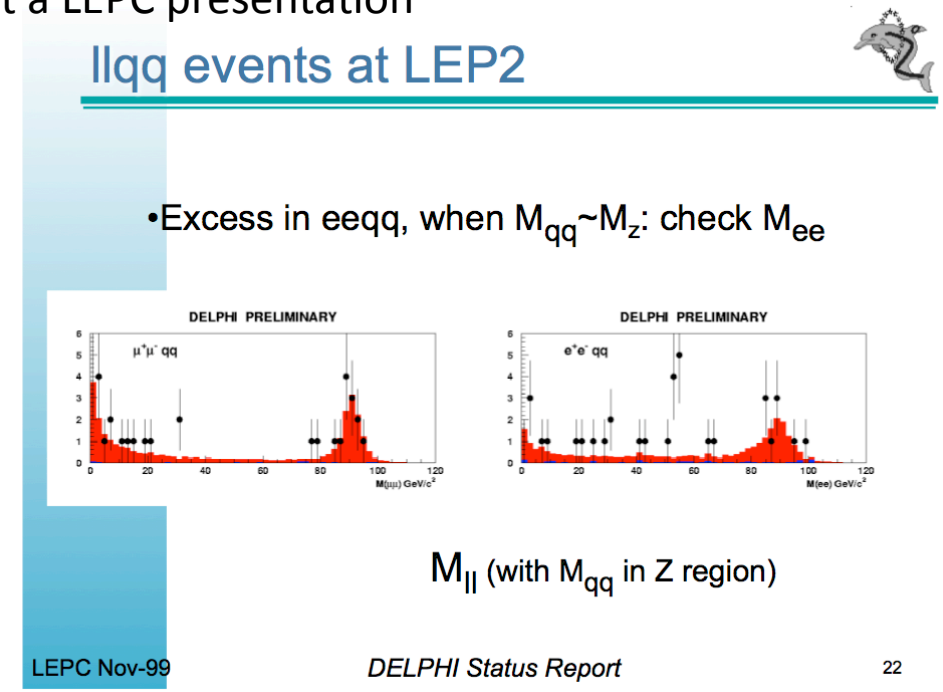
Five times the data sample
CLAS Collab., **Phys.Rev.Lett. 100 (2008) 052001**

n.b. the Bayesian analysis in this paper is flawed –
see the criticism by R. Cousins, **Phys.Rev.Lett. 101 (2008) 029101**

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”

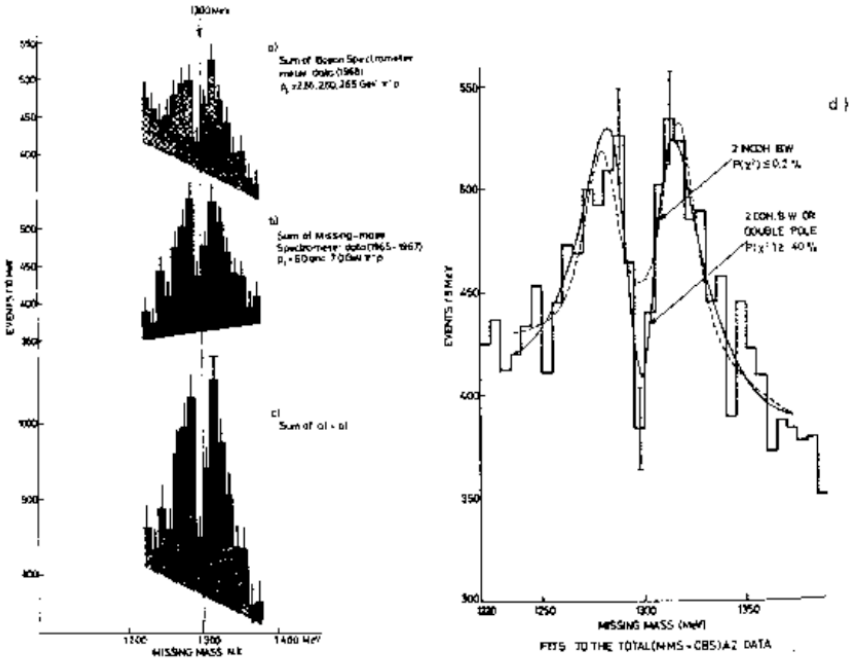


Figure 3: (a-c) Evidence for A_2 splitting in $\pi^- p \to pX^-$ collisions in the two CERN experiments, (d) same as (c) in 5 MeV bins fit to two hypotheses.

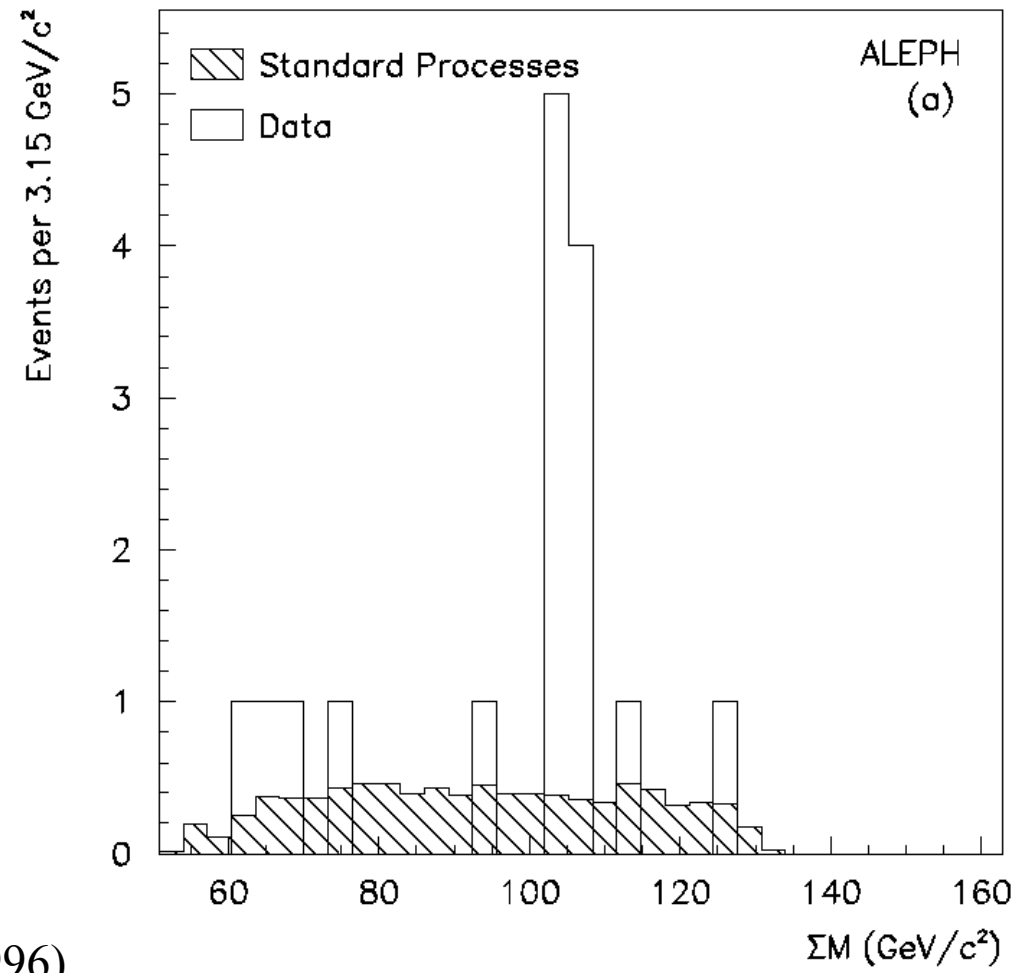
Text from Sheldon’s article:

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 $e^+e^- \rightarrow jjjj$

Sociological Issues

- Discovery is conventionally 5σ. In a Gaussian asymptotic case, that would correspond to a ±20% measurement.

- Less precise measurements are called "measurements" all the time

- We are used to measuring undiscovered particles and processes. In the case of a background-dominated search, it can take years to climb up the sensitivity curve and get an observation, while evidence, measurements, etc. proceed.

- Journal Referees can be confused.

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.

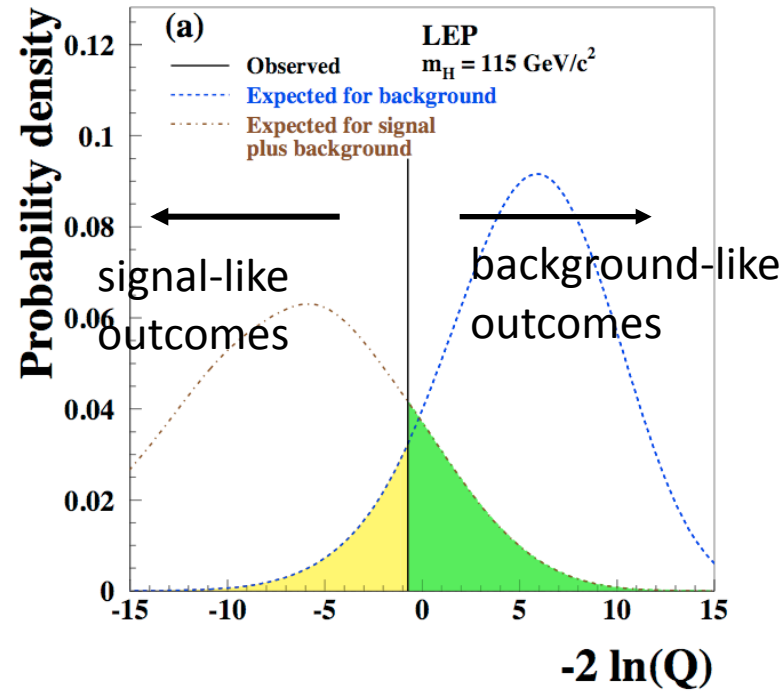
The word coverage comes from confidence intervals – are they big enough to contain the true value of a parameter being measured and what fraction of the time they do.

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



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

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

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

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.

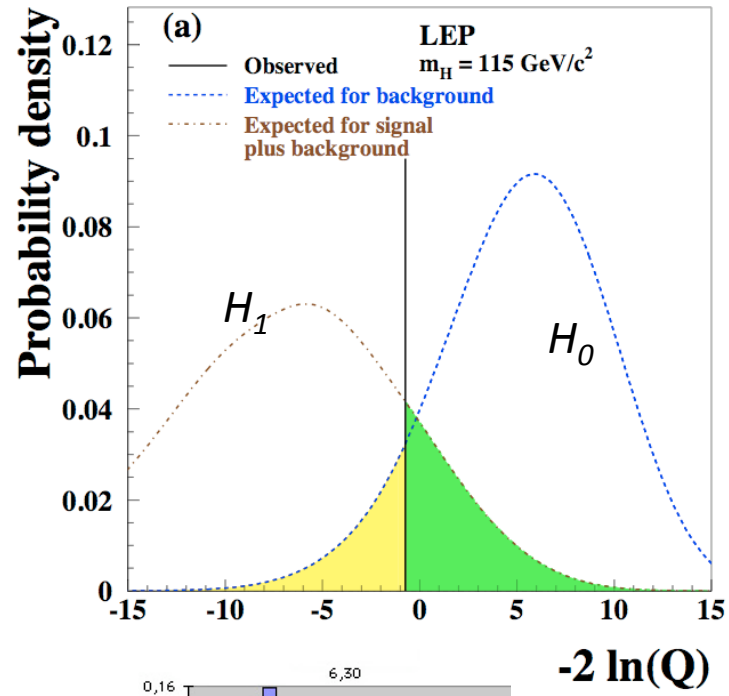
The "or-equal-to" is important here. For highly discrete distributions of possible outcomes – say an experiment with a background rate of 0.01 events (99% of the time you observe zero events, all the same outcome), then observing 0 events gives a p-value of 1 and not 0.01.

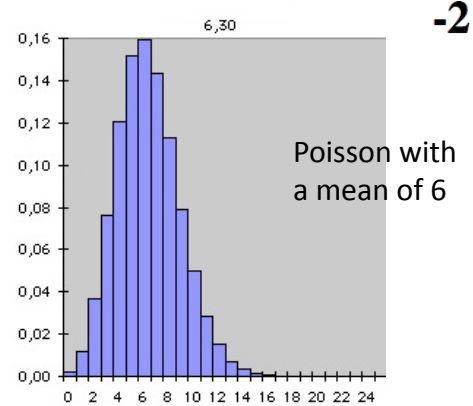
Shouldn't make a discovery with 0 observed events, no matter how small the background expectation! (or we would run the LHC with just one bunch crossing!).

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.

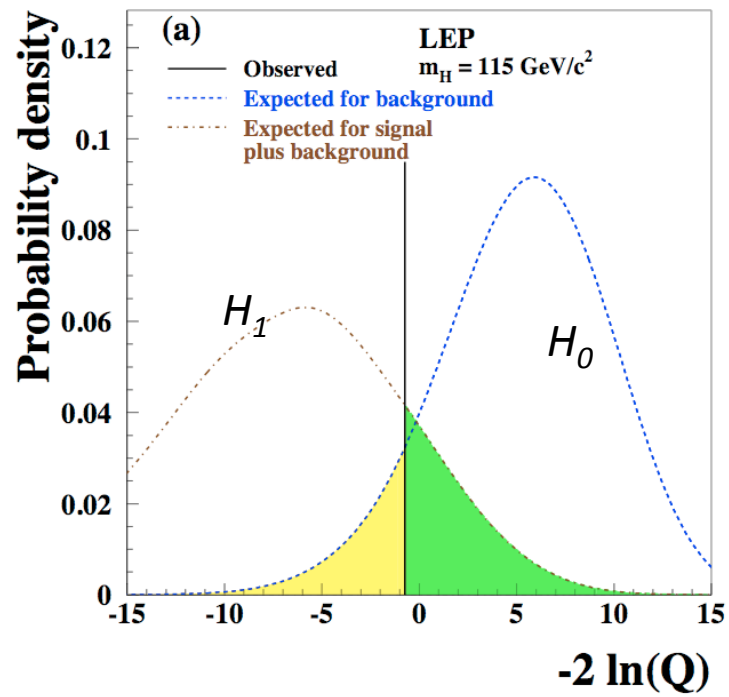
If it is small, reject H_1
The "or-equal-to" has similar effect here too.

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.

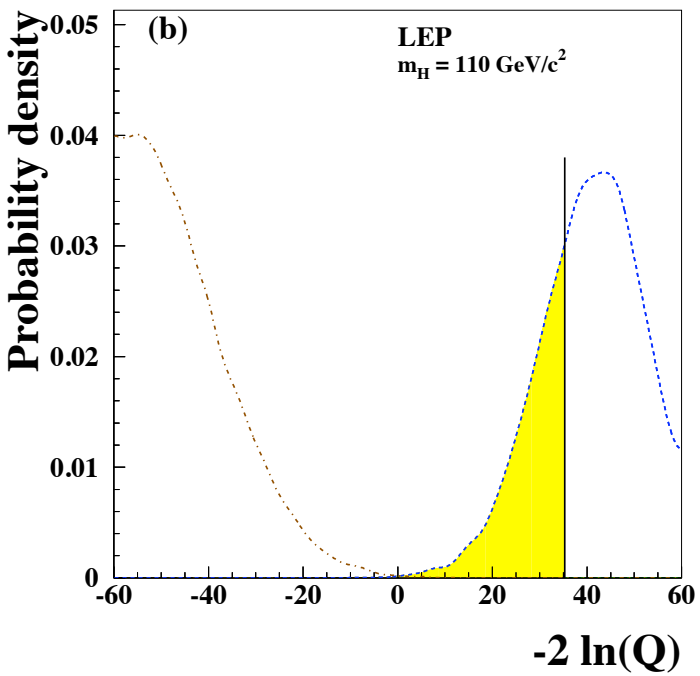
Lucky outcome: exclude at 97% CL
Do we exclude at the 50% CL?

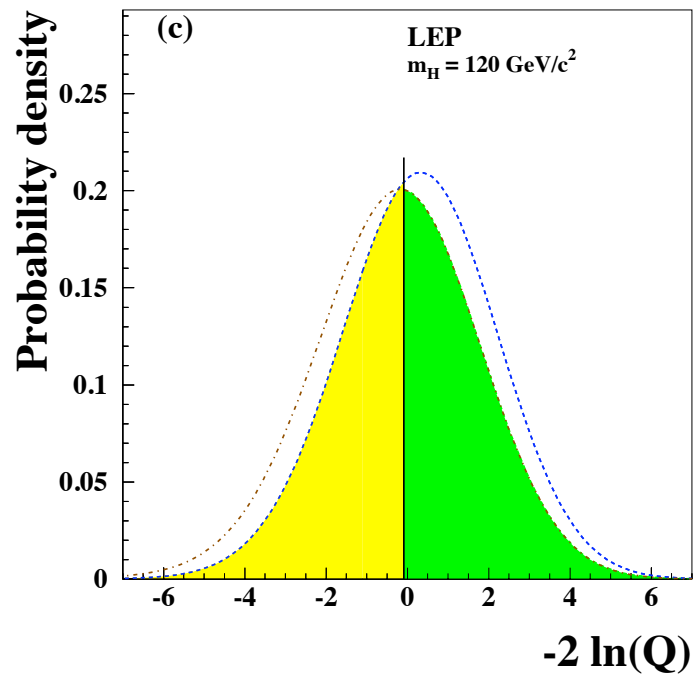
No! Set α once and for all (say 0.05). Then coverage is defined.



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



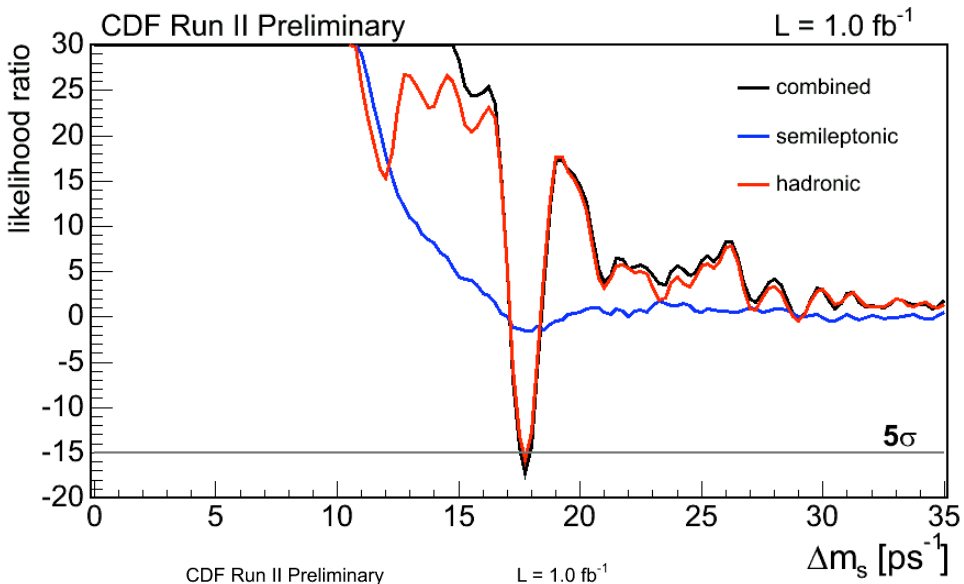


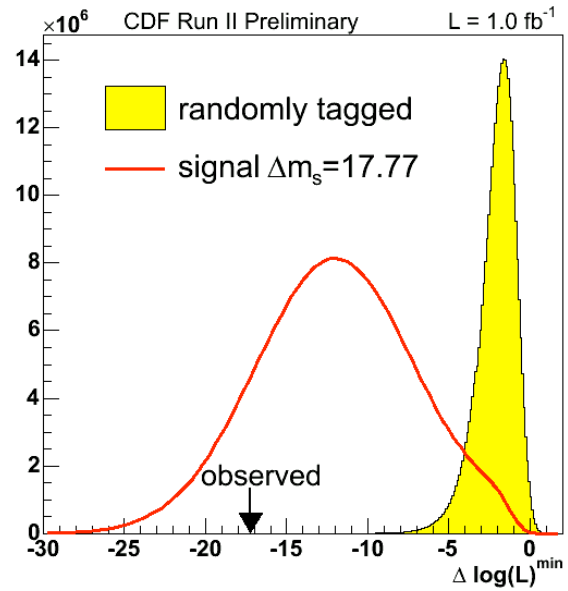
signal p-value very small.
Signal ruled out.
Possible to exclude both H_0 and H_1 (-2lnQ=0).
Possible to get outcomes that make you pause to reconsider the modeling. Say -2lnQ<-100 or -2lnQ>+100

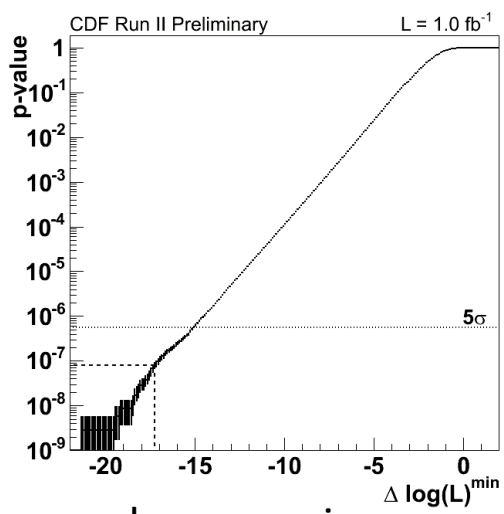
Can make no statement about the signal regardless of experimental outcome.

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

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

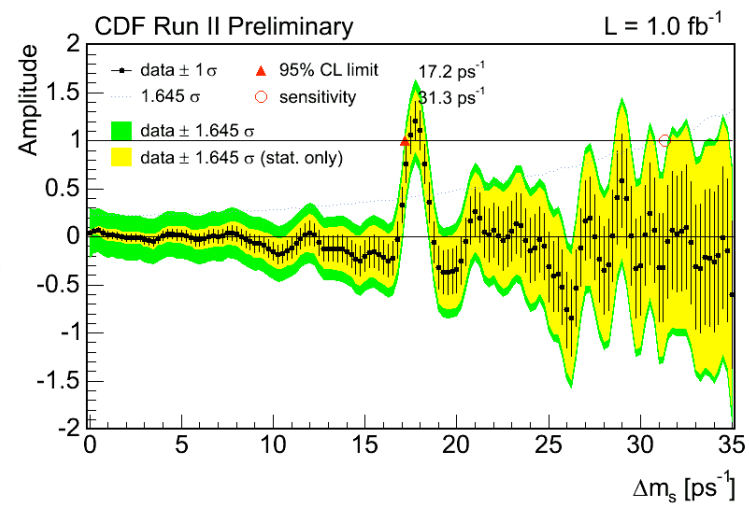






Log-scale comparison of observation and no-signal outcome distribution

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!! More on this later.)

The Classical Two-Hypothesis Likelihood Ratio

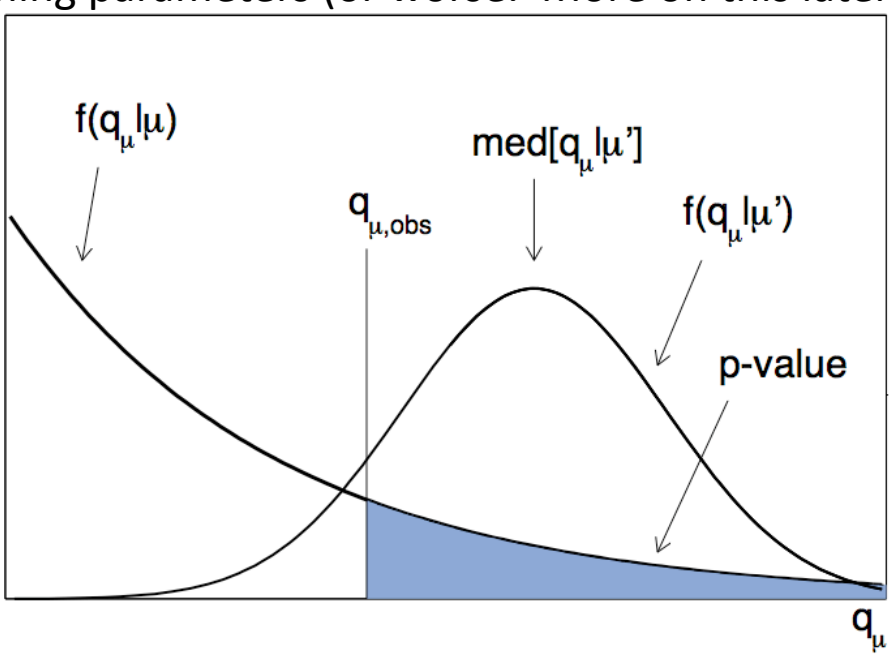
Distinguishing between μ=0 (zero signal, SM, Null Hypothesis) and μ>0 (the test hypothesis)

Assumption Warning! Signal rates scale with a single parameter μ

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

$\hat{\mu}$ 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

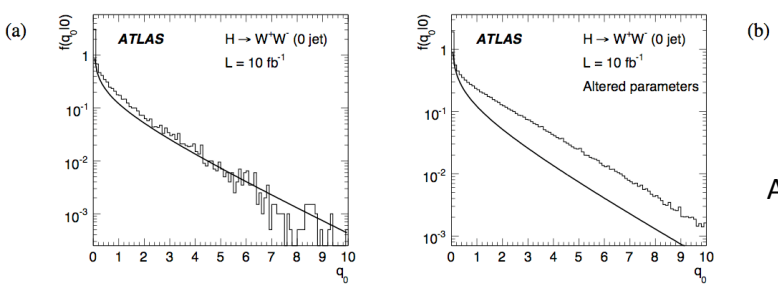
If the true value of the signal rate is given by μ, then q_μ is distributed according to a χ^2 distribution with one degree of freedom.

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.



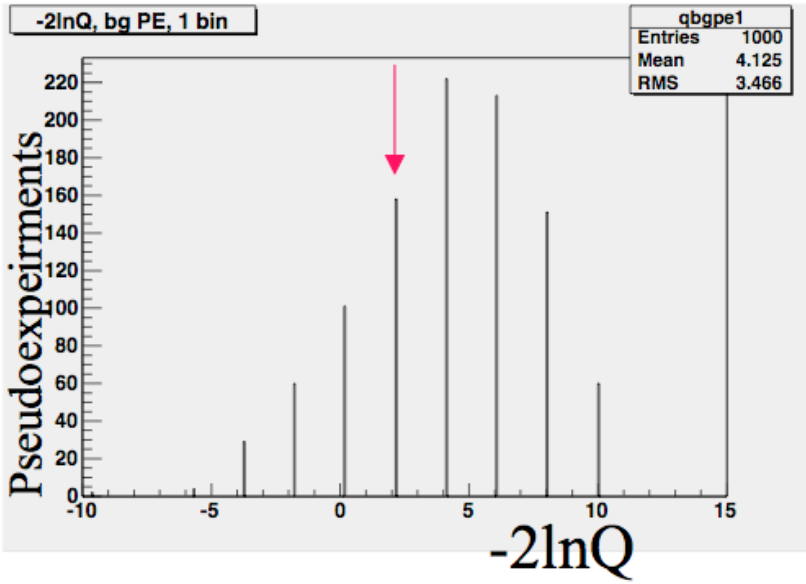
ATLAS performance projections, CERN-OPEN-2008-020

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.

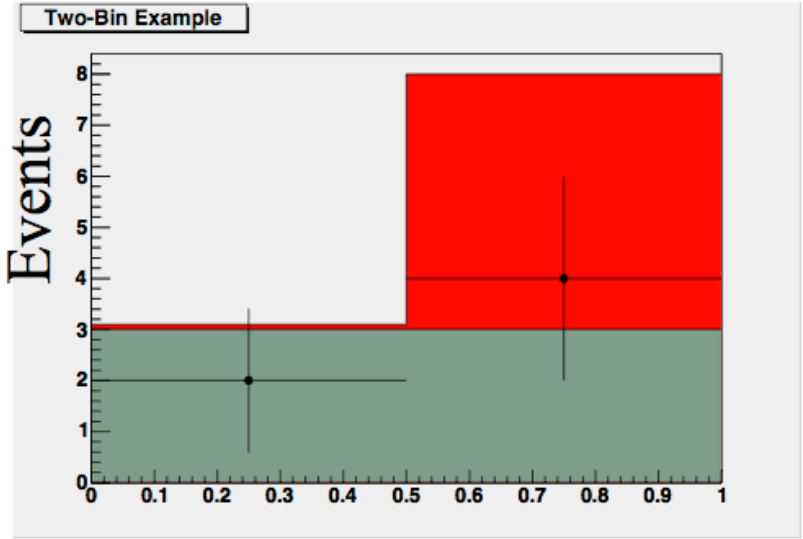
Multibin Searches and Discreteness

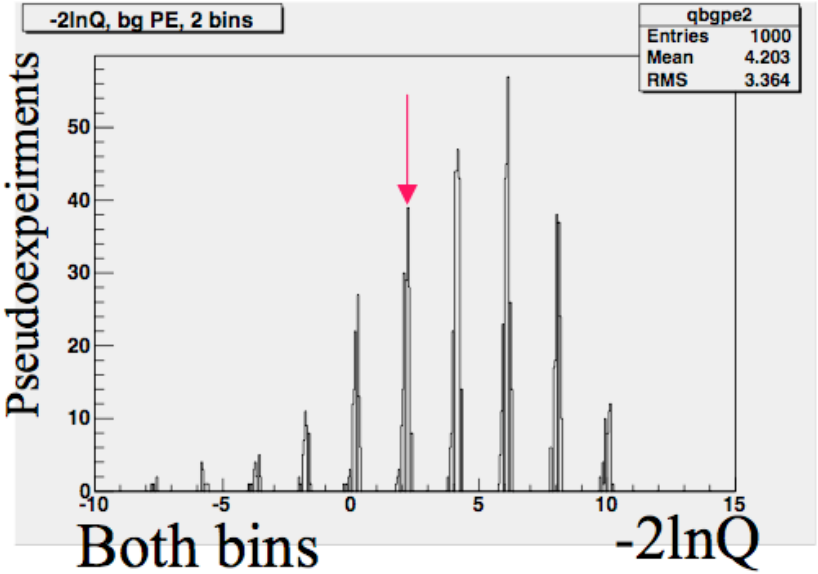
A discrete example with two bins, no systematic errors, very different s/b:

Bin 1: s=0.1, b=3, data=2
Bin 2: s=5, b=3, data=4



Background pseudoexperiments, second bin only





Multibin Searches and Discreteness

- The p-value you get for a 1-bin search can be modified by as much as the probability of a particular outcome by adding in a second bin with almost no power.

- Consequence of discreteness and the definition of p-value "Split Outcomes"

$1\text{-}CL_b = \text{p-value(bg)} = p(-2\ln Q \leq -2\ln Q_{obs} | \text{bg hyp})$

- An indication that your analysis can be optimized further
- Distribution of p-value(bg) is not uniform but discrete in this case, and really is a convolution of discrete outcomes

- Many bins of different s/b cure this apparent problem.

The Classical Two-Hypothesis Likelihood Ratio

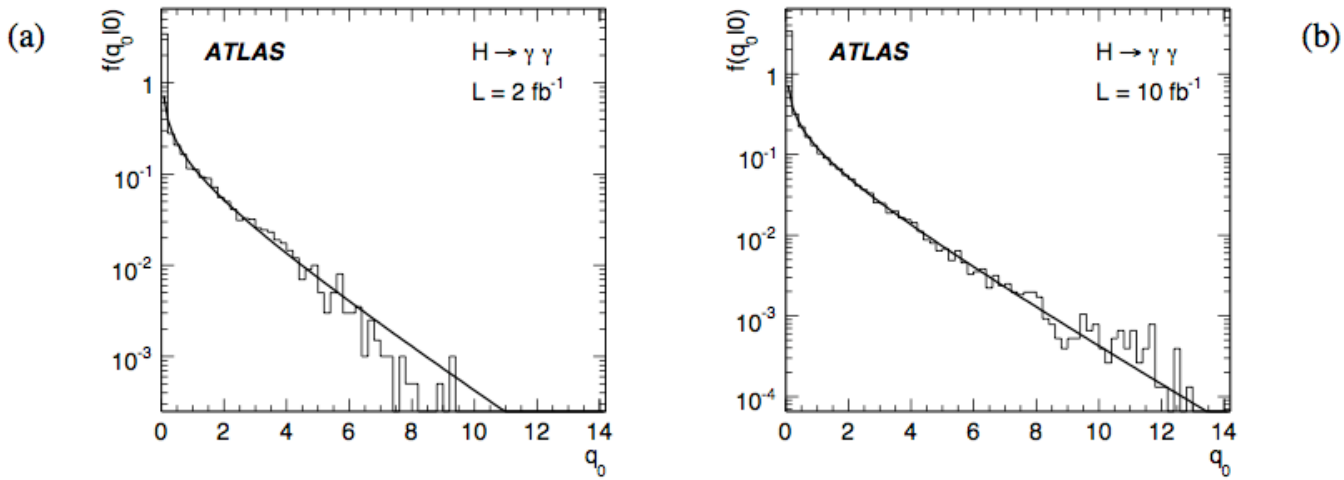


Figure 6: The distribution of the test statistic q_0 (for $H \to \gamma\gamma$), under the null background only hypothesis, for $m_H = 120$ GeV with an integrated luminosity of 2 (a) and 10 (b) fb^{-1}. A $\frac{1}{2}\chi^2_1$ distribution is superimposed.

ATLAS performance projections, CERN-OPEN-2008-020

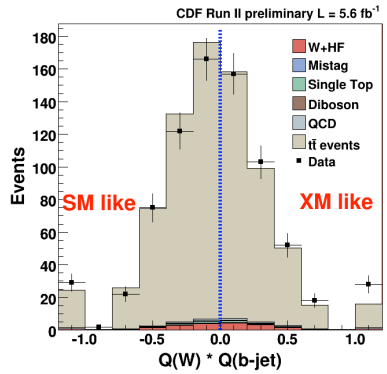
The big δ-function at q_0=0 is for those outcomes for which the best signal fit is zero or negative. The null hypothesis is exactly as good a description as the test hypothesis. If the null is really true, this should happen ½ of the time.

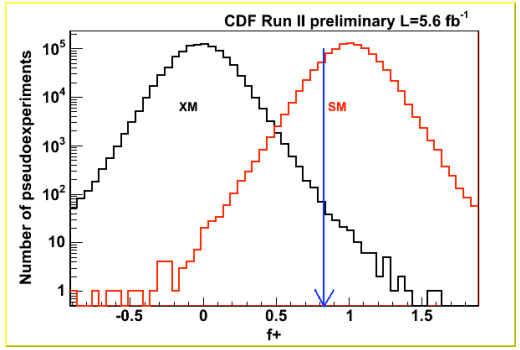
The Classical Two-Hypothesis Test

In the case that one or the other, H_0 or H_1 must be true, then β is a function of α for a single requirement on q.

This is useful when testing very discrete possibilities – for example, the charge of the top quark:

H_0: q(top) = 2*e*/3
H_1: q(top) = 4*e*/3.





These are the only allowed possiblities assuming t→Wb (Wbbar) proceeds.

See: CDF Collab., **Phys.Rev.Lett. 105 (2010) 101801.** Even here we introduced no-decision regions to keep with our 95% CL exclusion and 3σ, 5σ conventions

For problems with a more continuous test hypothesis,
LEP and the Tevatron choose not to fit for the signal rate μ in the test statistic as it makes for a more symmetrical presentation, and one can read off CL_{s+b}

Conditioning the Ensemble and the Stopping Rule

- Something the analyzers did a few years ago (smaller data sample) which wasn't optimal (it wasn't wrong, just not optimal):

All pseudoexperiments to compute the p-values were generated with the same total number of events.

Test statistic – counting same charge vs. opposite charge events and form an asymmetry:
$A = (N_{SM}-N_{XM})/(N_{SM}+N_{XM})$
– very discrete!

Each possible experimental outcome had a high probability of occurring. An asymmetry of zero in particular was highly likely! The "or-equal-to" part of the p-value rule was a large piece of the expected p-value (and we want small p-values for making significant tests)

Jargon: The ensemble was Conditioned on $N_{total}=N_{data}$

This is a "Slippery Slope"! How much like the observed data must the simulated outcomes be? At least in this case there's a clear answer.

The Stopping Rule

- Statisticians always ask: "When do you stop taking data and analyze and publish the results?"
- Important in order to define the sample space from which an experimental outcome has been drawn (cannot compute p-values without an answer to this).

Some options:

- Run until you get a desired N_{total} events (hardly anyone does this – although one year, SLC ran until 10,000 Z's were detected offline by SLD. Ambiguous, because extra ones can be found by changing cuts or unearthing unanalyzed tapes).
- Run until you get a desired $N_{selected}$ events passing some analysis requirement. If you are looking for a rare process with little or no background, you could be running for a long time! Also, the distribution of -2lnQ looks odd in this case. Worries of bias (the last event is always a selected one!).
- Stop when you get a small p-value. Possible, since p-values fluctuate between 0 and 1. As more data arrive, newer data overwhelm older data and the p-value is effectively re-sampled from [0,1] (takes exponentially more data to do this). See the "Law of the Iterated Logarithm". Called "Sampling to a foregone conclusion." Avoid at all costs even the perception of this.
- The most common case: HEP experiments run until the money runs out. Analyze all the data (if possible), and assume the experimental outcome was drawn from a large sample of experiments with the same total integrated luminosity.
- Variation: It can be an individual's money (or time, or patience) that limits a specific analysis
- Variation: Analyze a subset of the data that can be processed in time for a major conference.

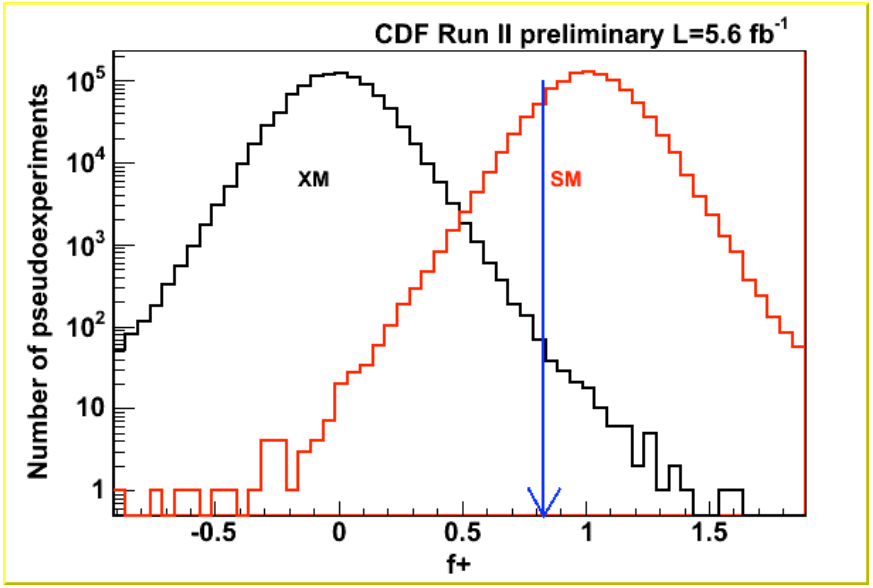
Back to the Q_{top} Example

- Sampling outcomes from a Poisson-distributed N_{total} based on a predicted event rate provided more distinct values of

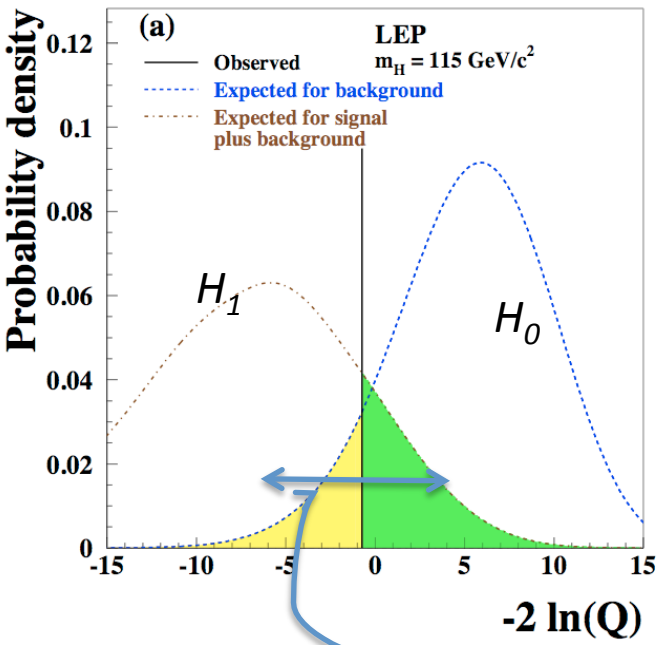
$A = (N_{SM}-N_{XM})/(N_{SM}+N_{XM})$; $f+ = N_{SM}/N_{total}$

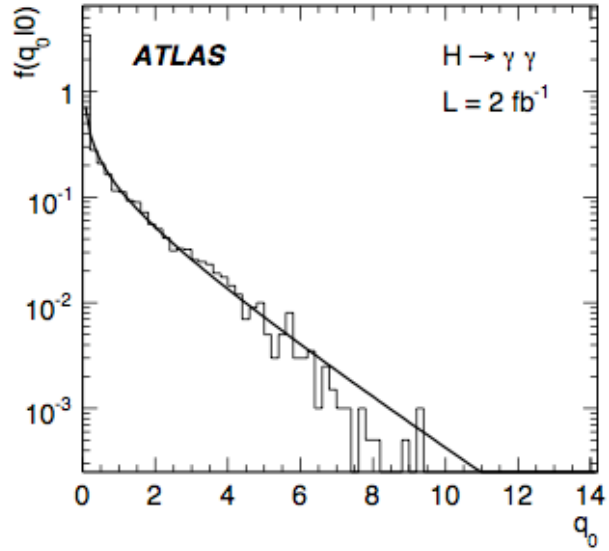
Example: for an odd N_{total}, A=0 is impossible. For even N_{total}, A=0 is likely!

The median expected p-value for a signal was smaller. And believable since the data are drawn from a Poisson distribution and not fixed *a priori*.



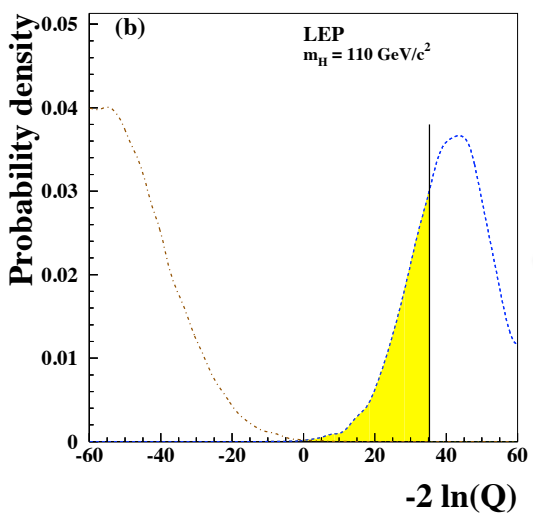
No-Decision Regions



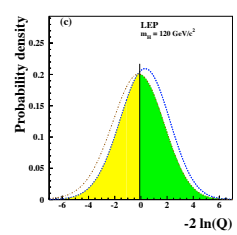


Can only test H_0 with this distribution q_{SM} is good for testing H_1 but it too has a delta-function at zero.

No Decision region
Outcomes in here are not sufficient to rule out either H_0 at 3σ or H_1 at 95% CL.
Need more data (or a better accelerator)



This example has no no-decision region. All outcomes either exclude H_0 or H_1, and some outcomes exclude both!



very small signal. no-decision region consists of all outcomes.

Gauging the Sensitivity of an Analysis

We need this for many, many reasons!

- Decide which experiments to fund and build
- Decide how long to run them
- Decide how to trigger on events
- Decide how to optimize the analysis
- Compare competing efforts. Some experiment may get "lucky" – that does not mean that they were necessarily better at what they were doing.

The classical β is not used (much, if at all) in HEP. Mostly because we allow for no-decision regions, and outcomes that do not look like either hypothesis, and because we stick to the 95% for exclusion, and 3σ and 5σ evidence and discovery error rates.

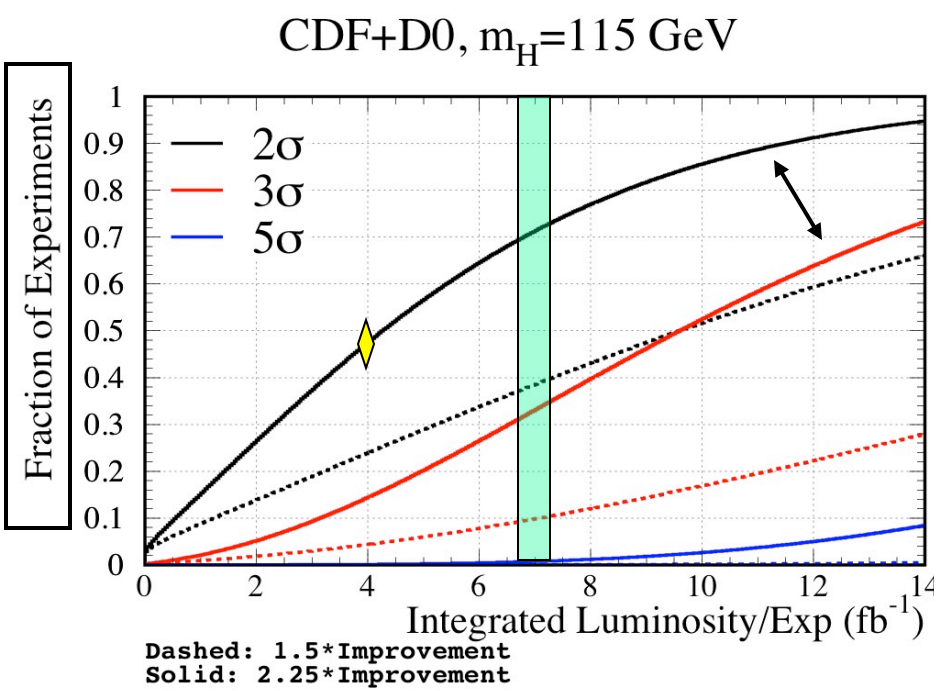
Today's currancy:

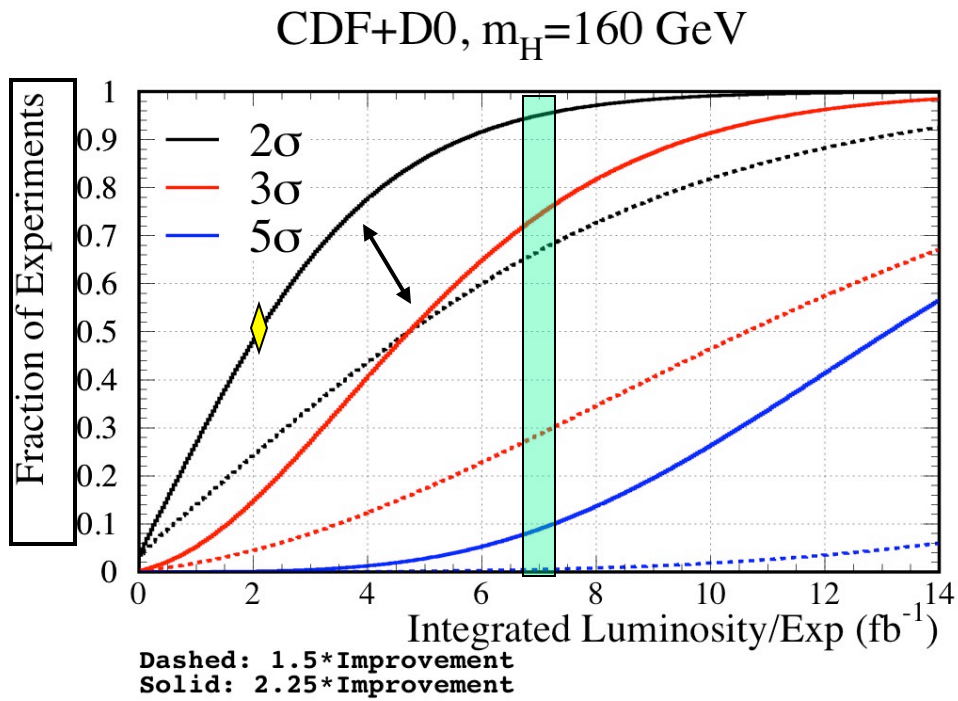
1) Median Expected p-value assuming a signal is present

2) Median Expected limit on cross section assuming a signal is absent

3) Median Expected width of the measurement interval for measured parameters

On any given roll of the dice





"further" @ 115 GeV

7 fb^{-1} => 70% experiments w/2σ
30% experiments w/3σ

"further" @ 160 GeV

7 fb^{-1} => 95% experiments w/2σ
75% experiments w/ 3σ

Tevatron experiments have achieved the 1.5 factor improvement already.

The "Asimov" Approximation for Computing Median Expected Sensitivity

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

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

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

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

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

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

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

A Case in which the Asimov Approximation Breaks Down

Usually it's a very good approximation.

Poisson discreteness can make it break down, however.

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

The median outcome is 0 events, not 0.2 events.

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

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

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

Some Comments on Fitting

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

  More on systematics later

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

- 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"

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

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

Other Possible ways to Incorporate Systematic Uncertainties in P-Values

For a nice (lengthy) review, see

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

Confidence interval method

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

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

Plug-in p-value

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

Double use of the data; ignores uncertainty in best-fit values of uncertain parameters.

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

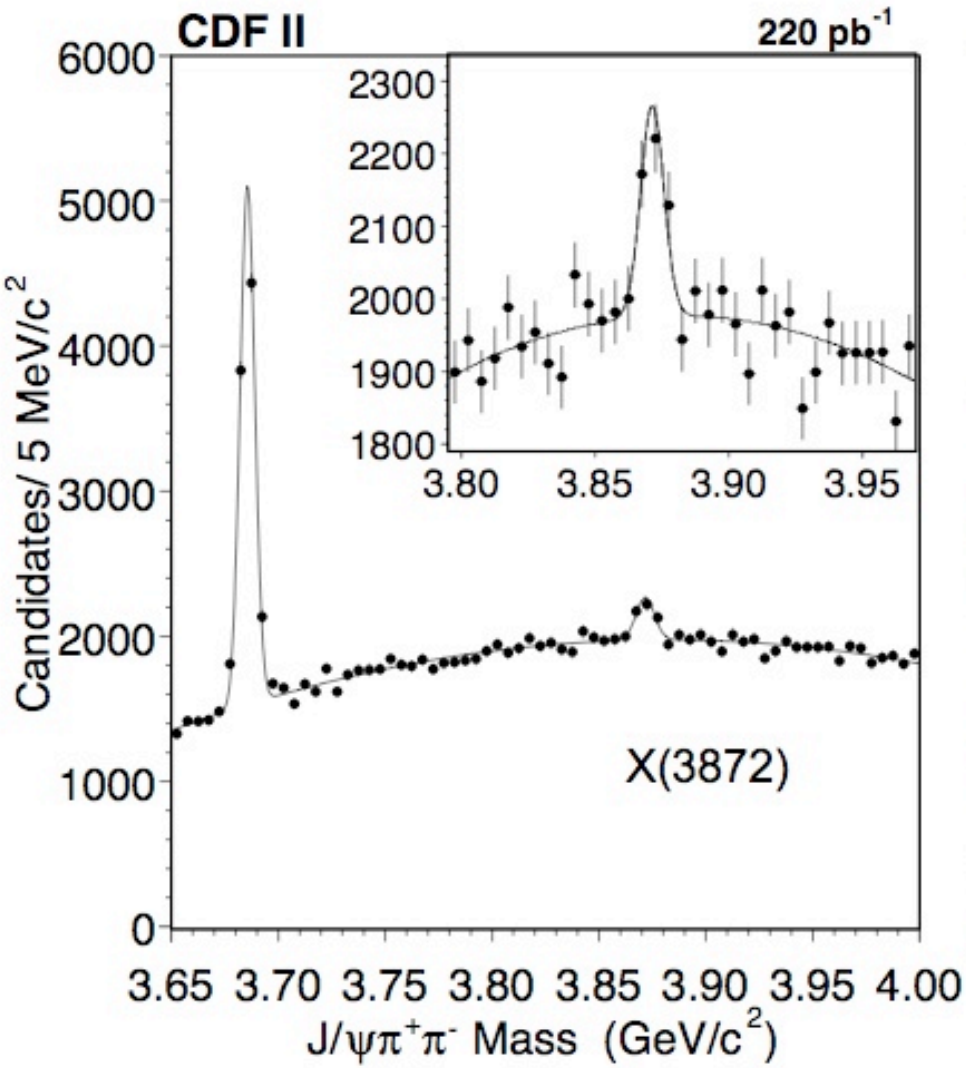
Probability that a future observation will be at least as extreme as the current observation assuming that the null hypothesis is true.

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.

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

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

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

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

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

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

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

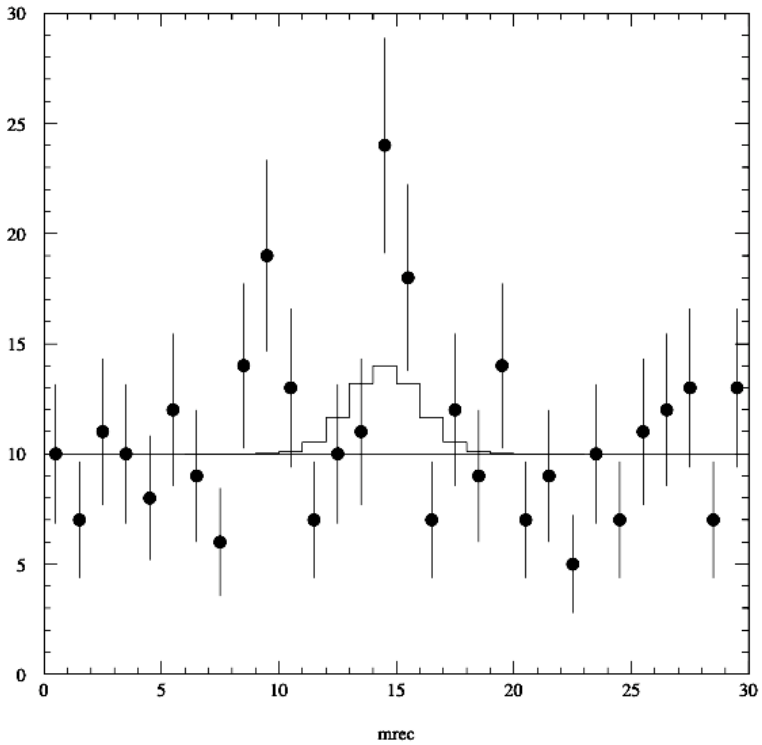
A *compound hypothesis* is less specific. It may have parameters whose values we are not particularly concerned about but which affects its predictions. These are called *nuisance parameters*, labeled θ.

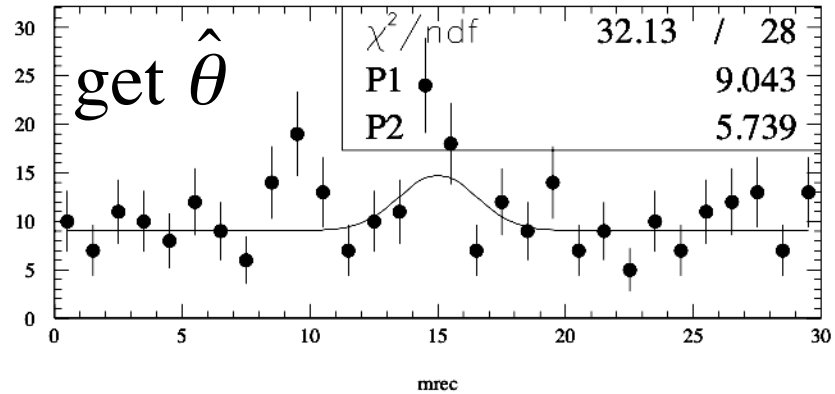
Example: H_0=SM. H_1=MSSM. Both make predictions about what may be seen in an experiment. A nuisance parameter would be, for example, the b-tagging efficiency. It affects the predictions but in the end of the day we are really concerned about H_0 and H_1.

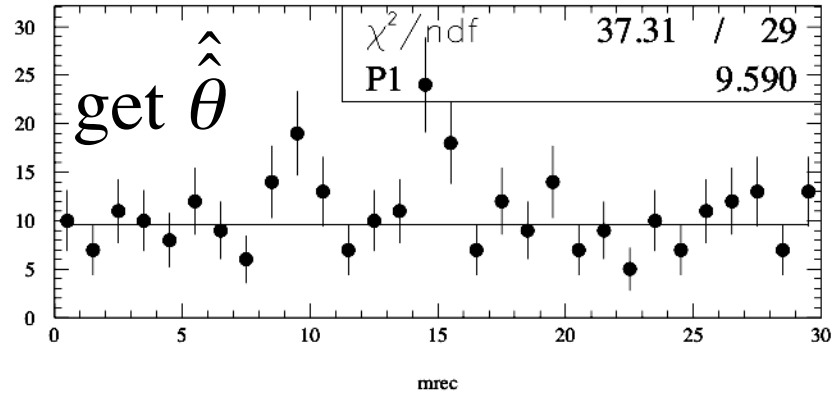
Fit twice! Once assuming H_0, once assuming H_1

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

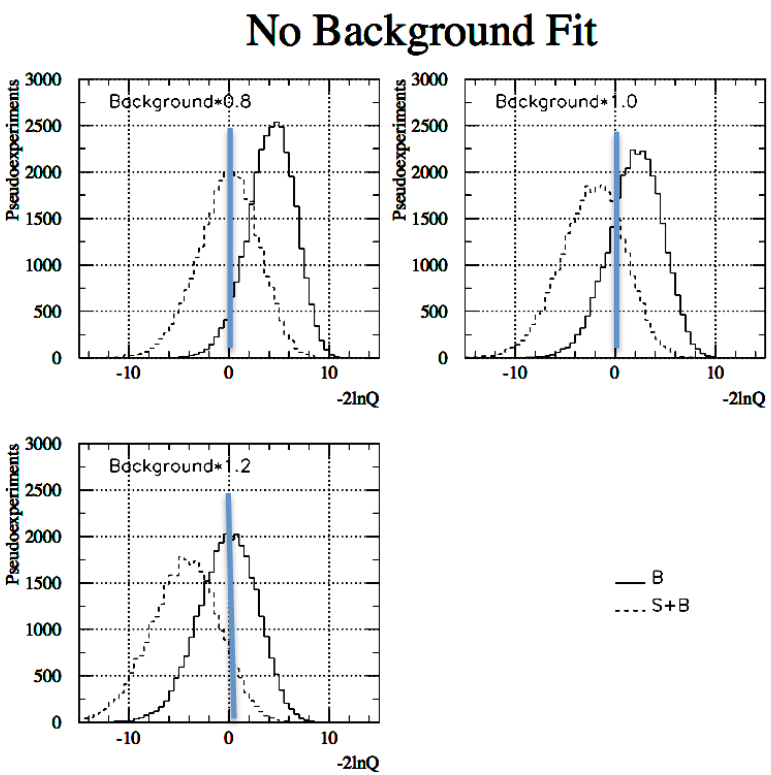
Fit to flat bg, Separate fit to flat bg + known signal shape.
The background rate is a nuisance parameter $\theta = b$
Use fit signal and bg rates to calculate Q.
Fitting the signal is a separate option.





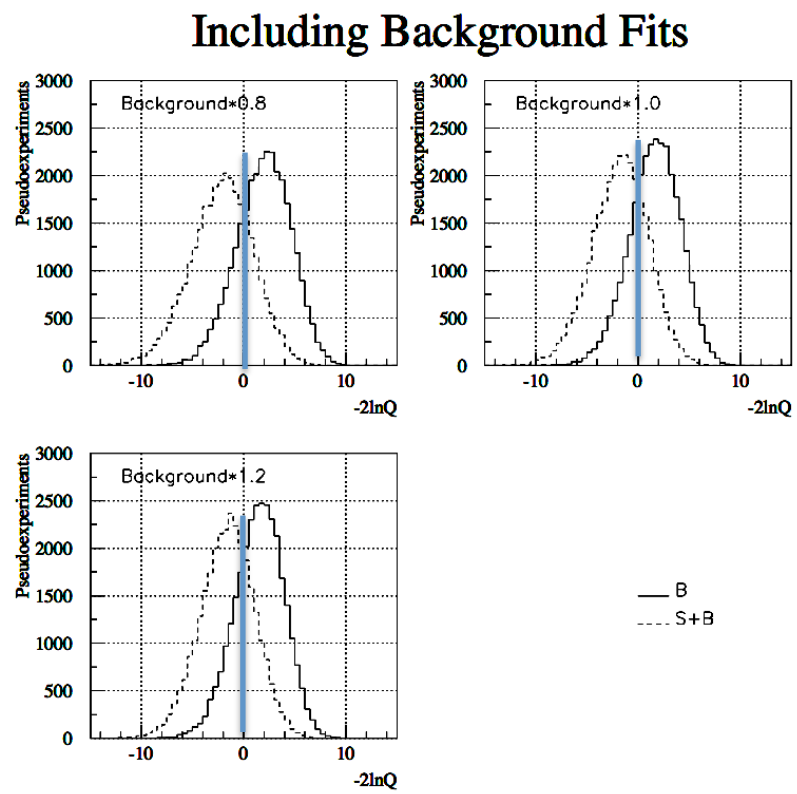


Fitting Nuisance Parameters to Reduce Sensitivity to Mismodeling

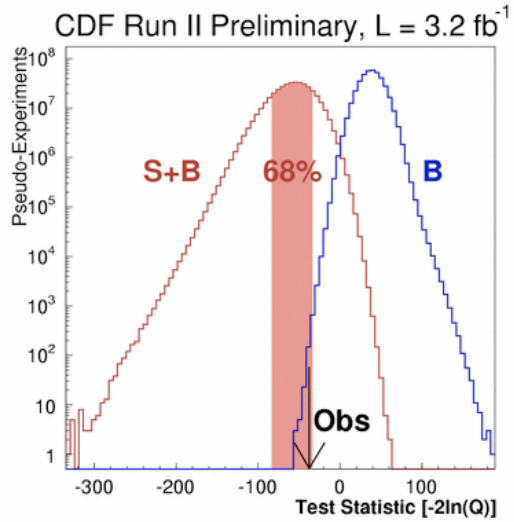


Means of PDF's of -2lnQ very sensitive to background rate estimation.

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



Fitting and Fluctuating



$$-2\ln Q \equiv LLR \equiv -2\ln\left(\frac{L(\text{data}\,|\,s+b,\hat{\theta})}{L(\text{data}\,|\,b,\hat{\hat{\theta}})}\right)$$

- 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.

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)

The "On-Off Problem"

Banff Challenge I

http://newton.hep.upenn.edu/~heinrich/birs/

Single counting experiment – select events in a "signal region" n_{on}.

Don't need a signal model, other than that more events is more signal-like. All test statistics are equivalent, ranking outcomes by n_{on}.

Background is uncertain: rate μ_b is unknown. Constrain μ_b with an auxiliary data sample (events failing the signal region requirements but passing other requirements – usually a subset of the signal region requirements).

Measure n_{off} events in the auxiliary sample.

Assume there is no signal in the "off sample"

Suppose further we know the value of $\tau=\mu_{off}/\mu_b$.
(note: this is almost never true – we have some uncertainty on τ, but more on uncertainties later).

See also Cousins, Linnemann, and Tucker, NIM A 595 (2008)

The "On-Off Problem"

- Still an oversimplification of a real HEP search.
- Cousins, Linnemann, and Tucker explored methods that are always conservative for computing discovery p-values.
- Joint binomial probablity for on vs. off counts works well.
- Bayesian analysis taking a uniform prior for μ_{off} (proportional to μ_b) ended up being numerically coincident with joint binomial probability. More on Bayesian techniques later.

But It's biased! The average of the posteriors for repeated outcomes in the off sample is $\mu_{off}+1$. Bin more finely, can make the total background estimate on average as big as you like.

Conservative for p-values – overestimates the background on average. Aggressive for limits!

Other techniques are biased as well – observe $n_{off}=0$ → if you infer $\mu_{off}=0\pm0$ gives p-values too small some of the time. See later talk on smoothing and density estimation.

Making Cousins, Linnemann, and Tucker – make sure $\tau > 5$ (my thesis advisor said run 5x as much MC as you have data). Sometimes you cannot, however.

The Trials Factor

- Also called the "Look Elsewhere Effect"
- 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 Šidák correction, still assumes independence.
$p_{corrected} = 1 - (1-p_{min})^n$

The Trials Factor

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

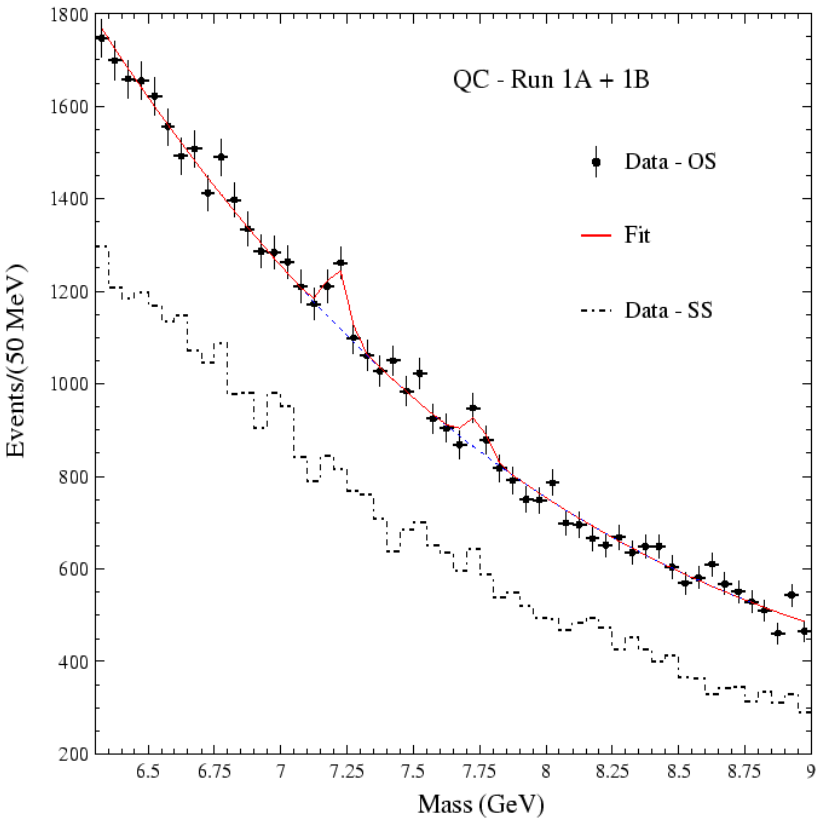
- Channels have different resolutions (or is resolution even the right word for a multivariate discriminant?
- Channels vary their weight in the combination as cross sections and branching ratios change with m_H

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.

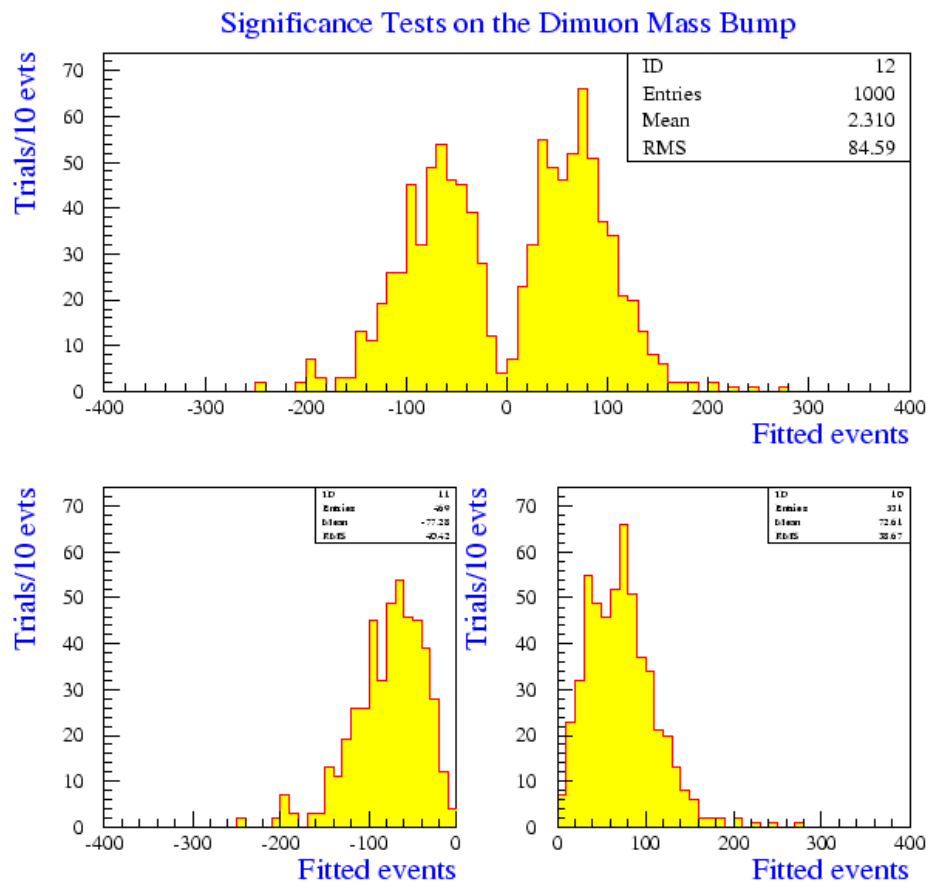
Next to impossible unless somehow analyzers supply how each pseudo-dataset looks at each test mass.

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!)



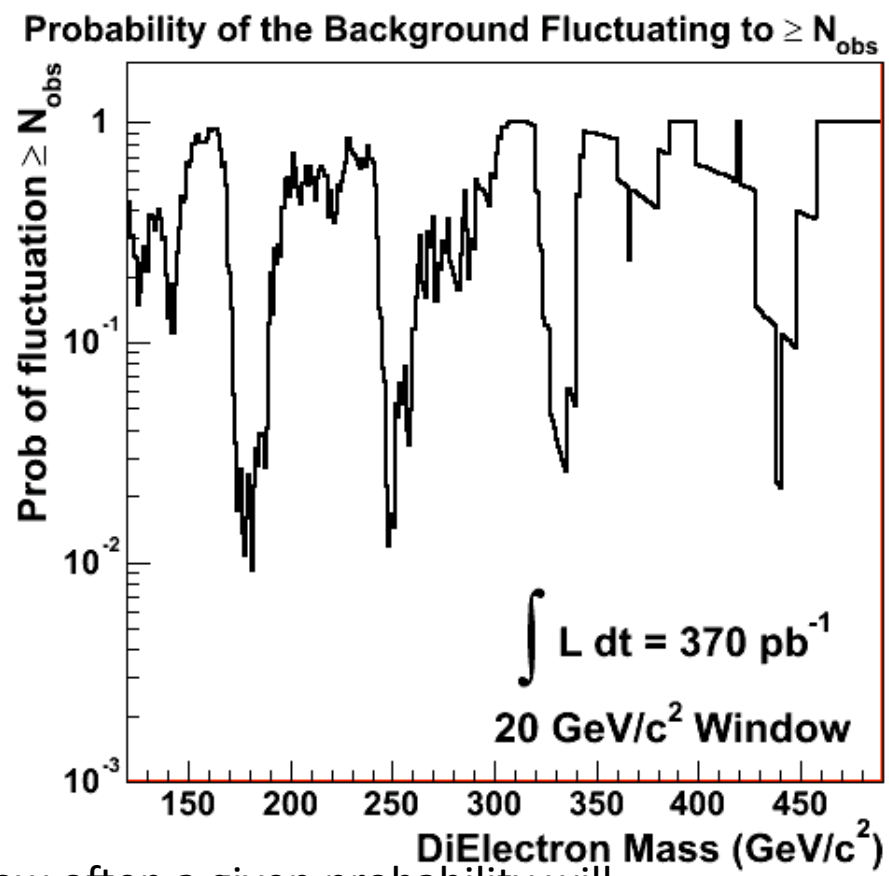
Null hypothesis pseudoexperiments with largest peak fit values

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

Single Pseudoexperiment



  - 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.

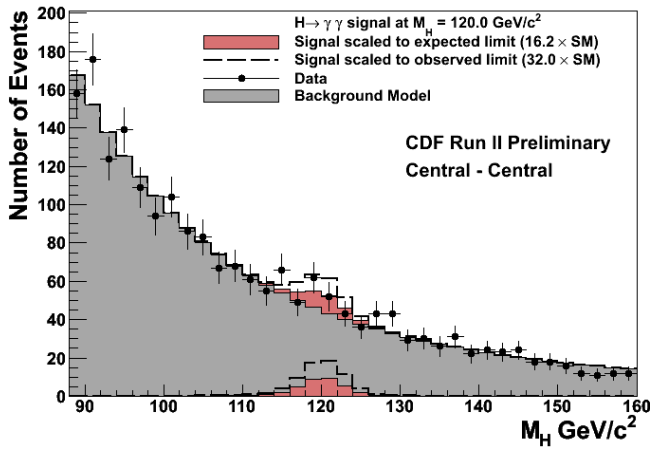
An interesting, very general feature:

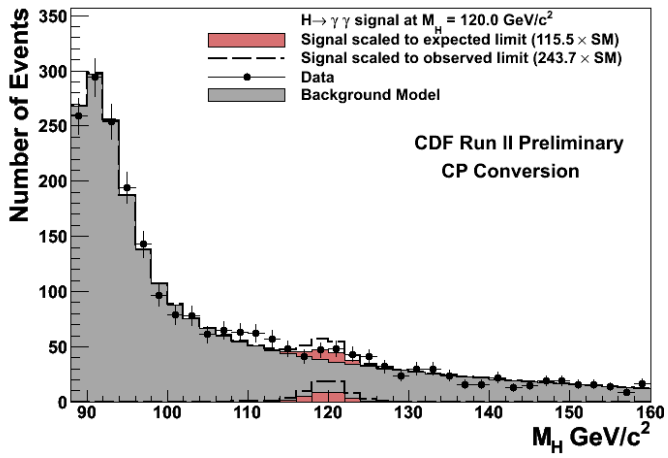
As the expected significance goes up, so does the LEE correction

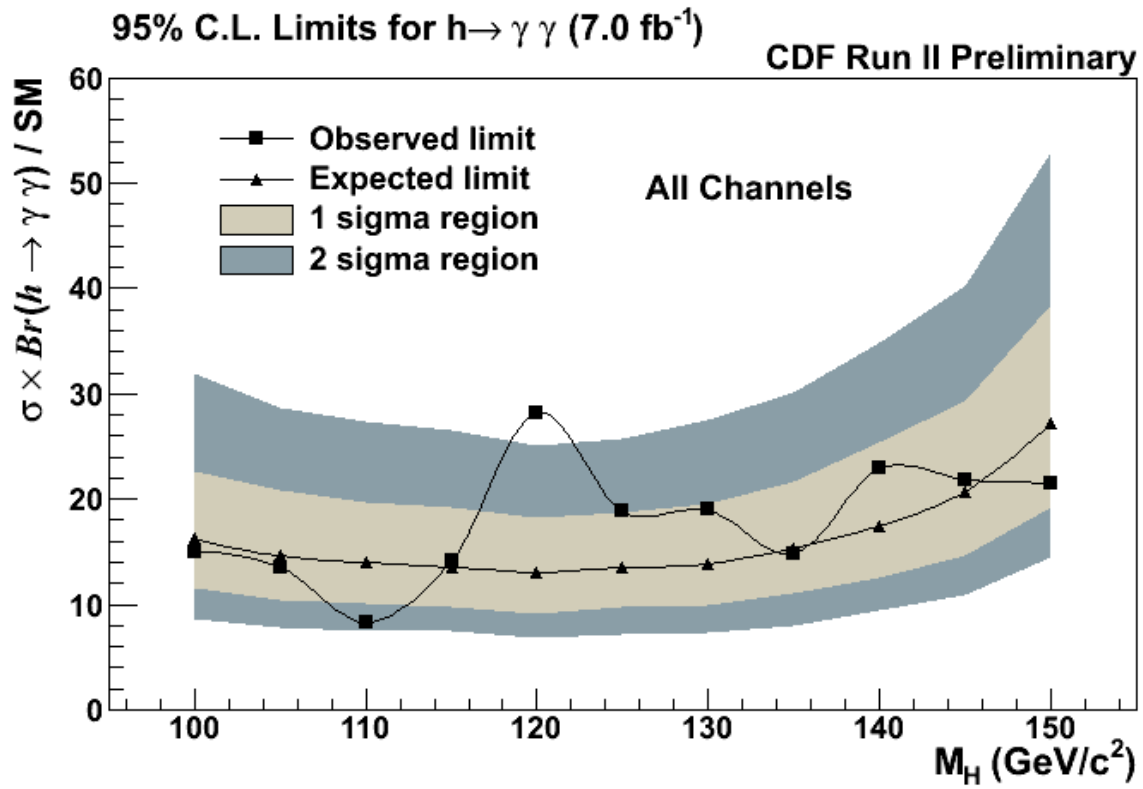
In hindsight, this makes lots of sense: LEE depends on the number of separate models that can be tested. As we collect more data, we can measure the position of the peak more precisely.

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

CDF's 2011 H→γγ Search







+2 other channels with smaller excesses

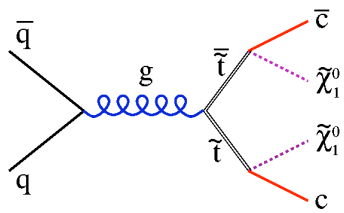
Insufficient sensitivity to a SM Higgs boson. Rate ruled out by other searches (gg→H→WW for example). So we know the bump is a stat fluctuation.

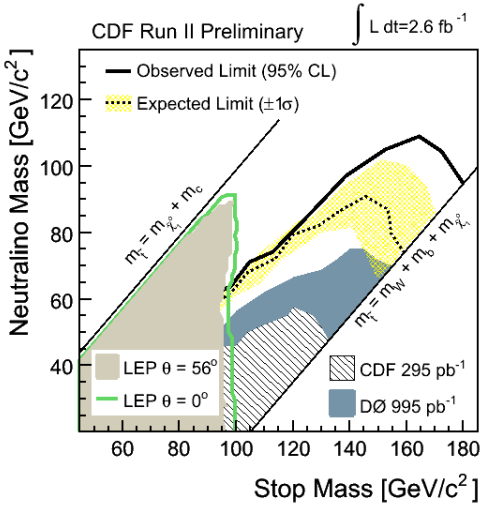
Where is "Elsewhere?"

- Most searches for new physics have a "region of interest"
  - Definition is a choice of the analyzer/collaboration
  - Often bounded below by previous searches, bounded above by kinematic reach of the accelerator/detector
  - Limits the amount of work involved in preparing an analysis. Sometimes a 2D search involves lots of training of MVA's and checking sidebands and validation of inputs and outputs

Example: A search for pair-produced stop quarks which decay to c+Neutralino

If $M_{stop}>m_W+m_b+m_{neutralino}$ then another analysis takes over.





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 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?

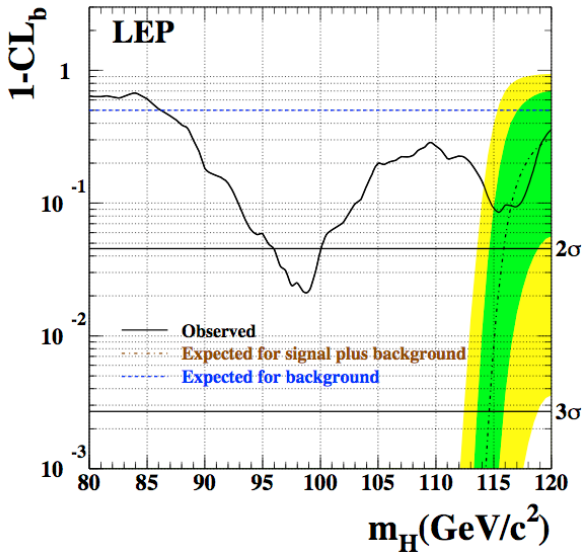
"Consensus" at the Banff 2010 Statistics Workshop: LEE should correct only for those models that are tested within a single published analysis. Usually one paper covers one analysis, but review papers summarizing many analyses do not have to put in additional correction factors.

Caveat lector.

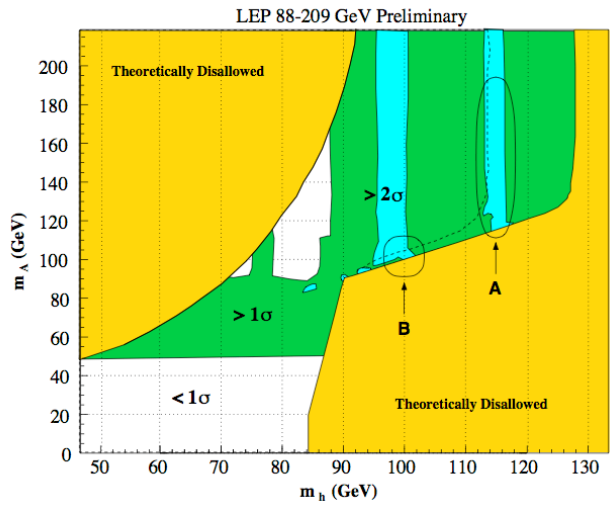
Where is "Elsewhere?"

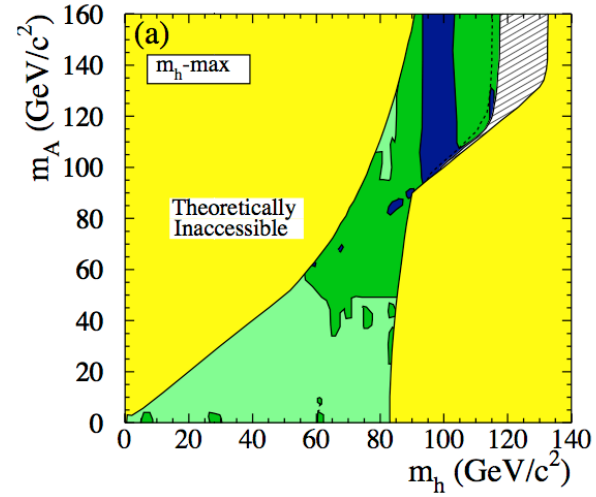
LEE is often hard enough to evaluate. Right way to do it – compute p-value of p-values simulate experiment assuming zero signal many times and for each simulated outcome find the model with the smallest p-value.
Multidimensional models are harder, and LEE is worse.

Kane, Wang, Nelson, Wang, Phys. Rev. D **71**, 035006 (2005)



ALEPH, DELPHI, L3, OPAL, and the LHWG
Phys.Lett. B565 (2003) 61-75





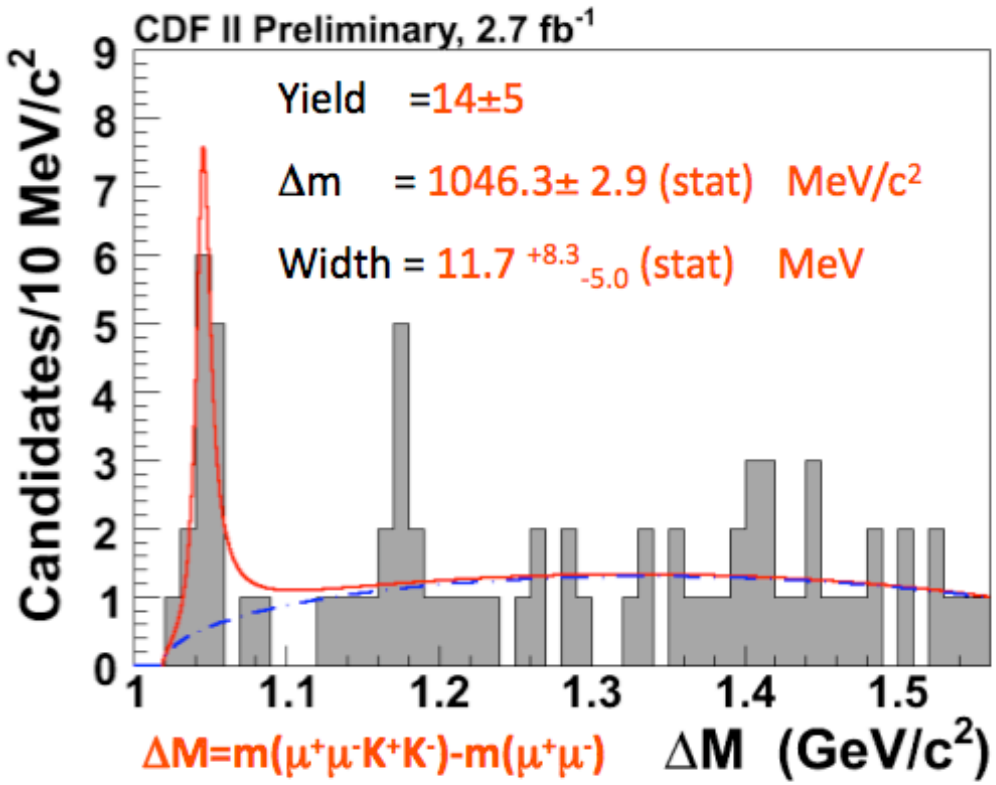
ALELPH, DELPHI, L3, OPAL, and the LHWG
Eur.Phys.J. C47 (2006) 547-587

Two excesses seen; proposed models explain both with two Higgs bosons. Combined local significance is greater, but LEE now is much larger (and unevaluated). Published plot grays out region beyond experimental sensitivity.

Search for structures in $J/\psi\phi$ mass--Data

- We model the Signal (S) and Background (B) as:

S: S-wave relativistic Breit-Wigner B: Three-body decay Phase Space



Convoluted with resolution (1.7 MeV)

Slide from K. Yi, Fermilab Joint Experimental/Theoretical Physics Seminar, March 17, 2009

How many bumps do you see?

$\sqrt{(-2\log(L_{max}/L_0))}$=5.3, need Toy MC to determine significance for low statistics

What if we don't have a signal model, and we're just on a hunting expedition? What's LEE now?

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!

LEP2's Energy Strategy, Blindness, and LEE

Every month brought a new beam energy. Sometimes new energies would be introduced at the end of a fill ("mini-ramps").

Experimenters did not have time to re-optimize analyses for the new energies – same cuts applied to new data; effectively blind.

But lots of new MC had to be generated, and lots of validation work for the new data.

Any experiment that rapidly doubles its dataset is in a luxurious position! Bumps in the data (even non-blind ones) can quickly be confirmed or refuted with new data.

Similarly, the untested window of m_H that was left from the previous year that was tested with the new data was small at the end – very little LEE!

LHC is now in its best phase! New energies, and rapid doubling of the data sample make most questions much cleaner!

Conversely, slowly-increasing data samples, or analyzing the data of a completed experiment favors blinding analyses.

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!

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.

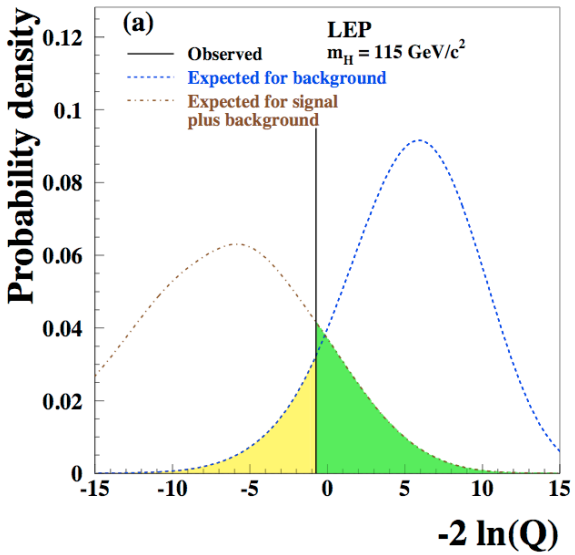
- Or we'd like to test a lot of models (each SUSY parameter choice is a model) and they can't all be true.

- 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

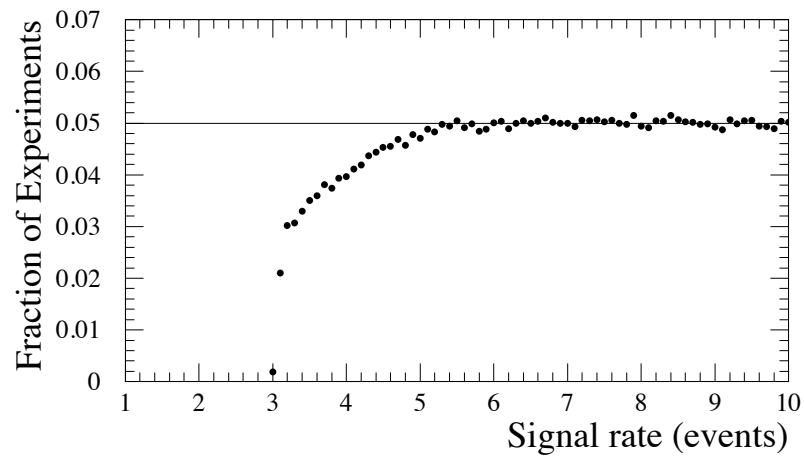
- Advantages:
  - Exclusion and Discovery p-values are consistent. Example -- a 2σ upward fluctuation of the data with respect to the background prediciton appears both in the limit and the p-value as such
  - Does not exclude where there is no sensitivity (big enough search region with small enough 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.

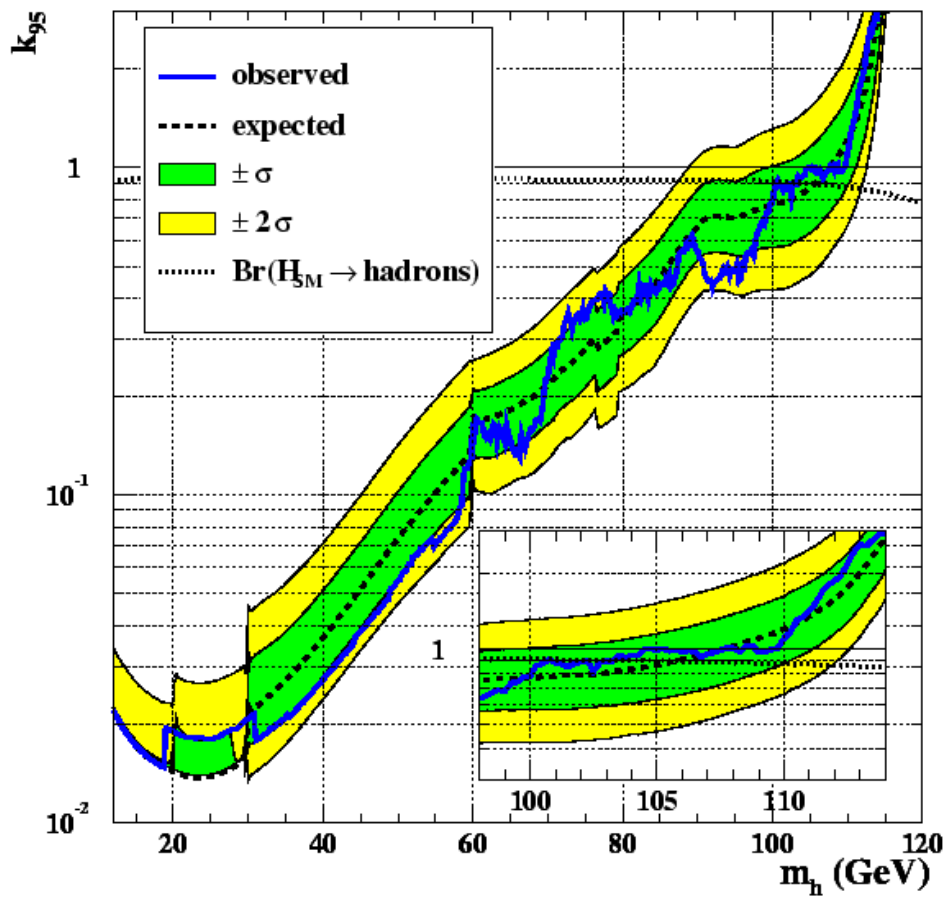
But MVA's take advantage of all categories of s/b and remain optimal in both cases; but you have to use the entire MVA distribution

Different kinds of analyses switching on and off

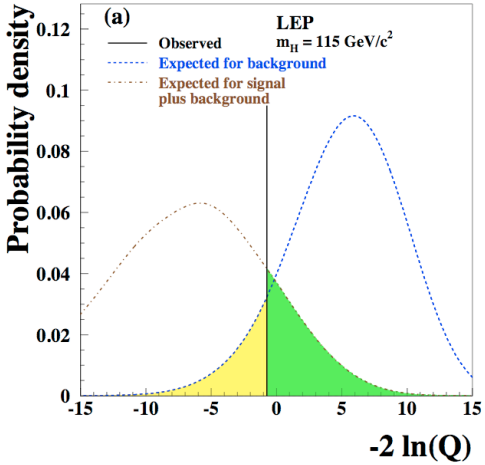
OPAL's flavor-independent hadronically-decaying Higgs boson search.

Two overlapping analyses: Can pick the one with the smallest median CL_s, or separate them into mutually exclusive sets.

Important for SUSY Higgs searches.

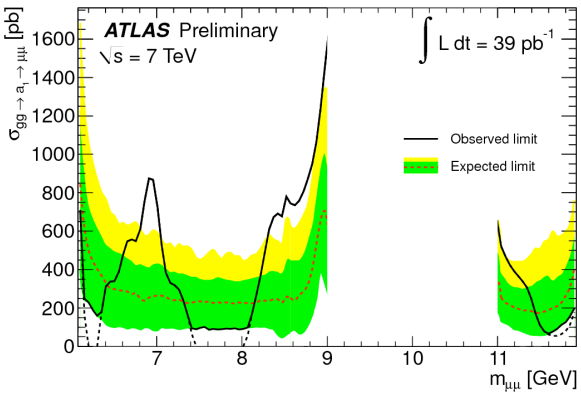


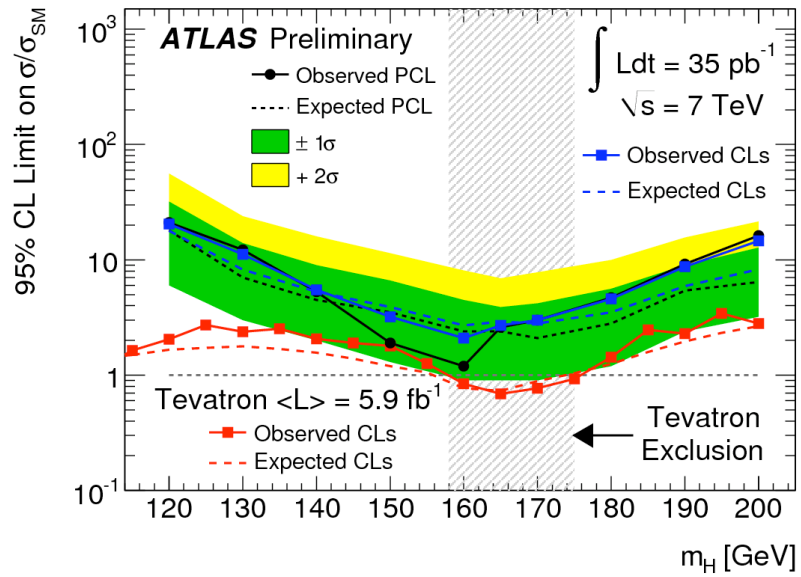
Power Constrained Limits (PCL)



Just use CL_{s+b}<0.05 to determine exclusion.

But if the resulting limit is more than 1σ more stringent than the median expectation, quote the 1σ limit instead





Advantages:

- More powerful than CL_s or Bayesian limits while still covering
- Does not exclude where there is no sensitivity

Disadvantage:

- 1σ constraint is arbitrary – balance desire for a more powerful method with acceptability of limits. A 2σ constraint defeats the purpose entirely for example.

An Interesting Feature of Power Constrained Limits

As with CL_s (and Bayesian limits, see later), if we observe 0 events and expect b=0 events, then the limit on the signal rate is r=-ln(0.05)=2.99573 ~ 3 events at 95% CL

The median expected limit is also 3 events since the median observation assuming the null hypothesis is 0 events (can rank outcomes easily with just one bin).

What if we expect some background b?

Observe 0 events, and CL_{s+b}<0.05 means s+b<3. So the limit will be 3-b events.
You get a better observed limit with more background expectation.
Not in itself a problem – a feature of most limit procedures.

But the median expectation is still 0 events for b<-ln(0.5) = 0.69 events. So the median expected limit decreases as the background rate increases. We get rewarded for designing a worse analysis!

(exercise: show why the median outcome is 0 events for a rate of 0.69)

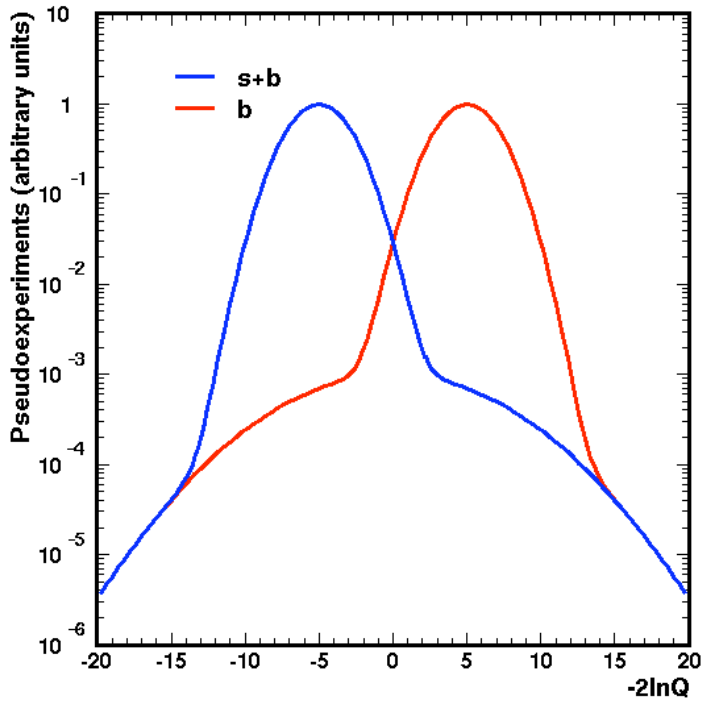
Interesting Behavior of CL_s

CL_s may not be a monotonic function of -2lnQ

Tails in the -2lnQ distribution shared in the s+b and b-only hypothesis (fit failures)

Distributions are sums of two Gaussians each. The wide Gaussian is centered on zero.

Practical reason this could happen – every thousandth experimental outcome, the fit program "fails" and gives a random answer.



CL_s=1 for
-2lnQ < -15 or
-2lnQ > +15

Not really a pathology of the method, but rather a reflection that the test statistic isn't always doing its job of separating s+b-like outcomes from b-like outcomes in some fraction of the cases.

Interesting Behavior of CL_s

Poisson Discreteness and ordering of outcomes can make the result "jump" when the model parameters tested vary by small amounts.

This is a hint of non-optimality – add more bins with different s/b usually fixes this problem. But there's another effect going on here.

-2lnQ = LLR is, without fits is given by the log of a ratio of Poisson probabilites, and serves as an Ordering Principle to sort outcomes as more signal-like or less.

$$Q = \frac{\prod_{i=1}^{n_{bins}} \frac{e^{-(s_i+b_i)}(s_i+b_i)^{n_i}}{n_i!}}{\prod_{i=1}^{n_{bins}} \frac{e^{-b_i} b_i^{n_i}}{n_i!}}$$

$$-2\ln Q = LLR = 2\sum_{i=1}^{n_{bins}} s_i - 2\sum_{i=1}^{n_{bins}} n_i \ln\left(1+\frac{s_i}{b_i}\right)$$

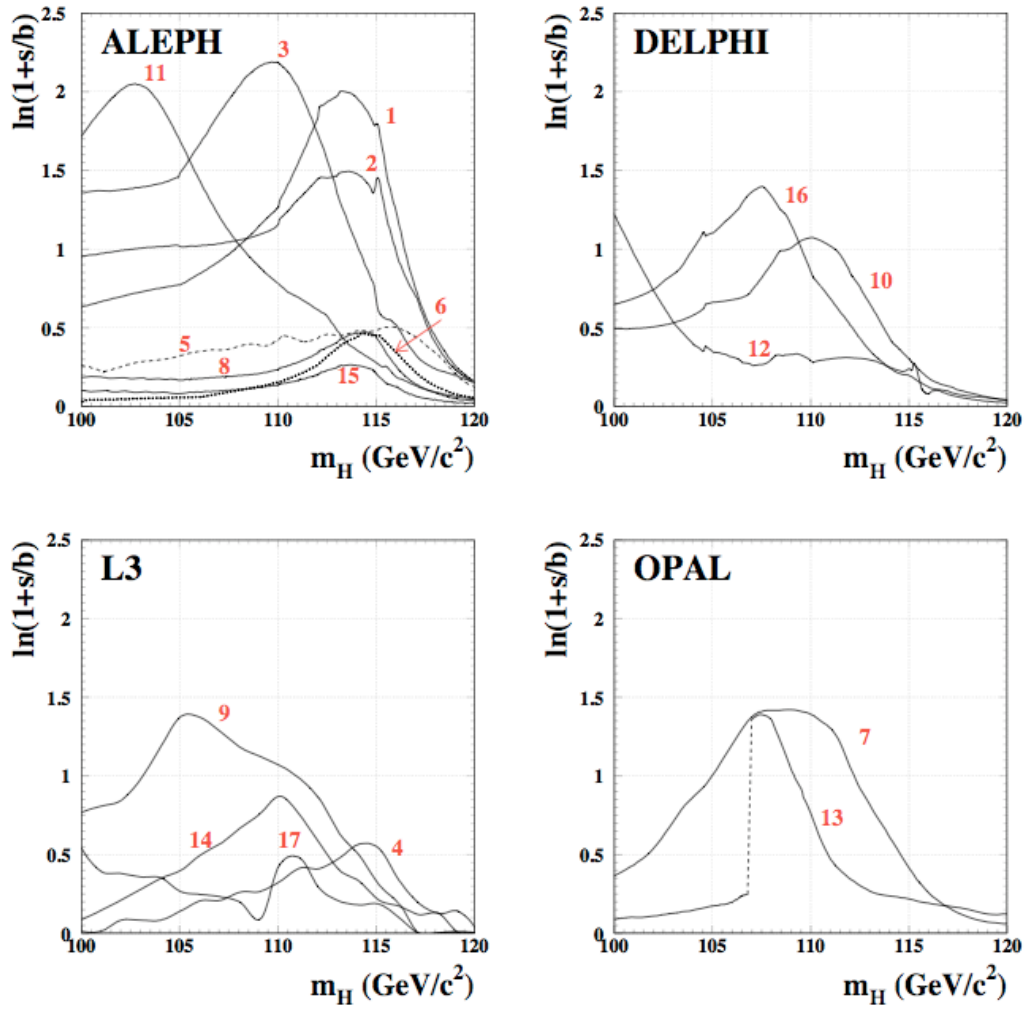
Aside from constant offsets and scales that do not affect the ordering of outcomes, it is a weighted sum of events where the weight is $\ln(1+s_i/b_i)$ where s_i/b_i is the local signal to background ratio. Each event can be assigned an s/b value.

Individual Candidates Can Make a Big Difference

At LEP -- can follow individual candidates' interpretations as functions of test mass

if s/b is high enough near each one.

Fine mass grid -- smooth interpolation of predictions -- some analysis switchovers at different m_H for optimization purposes



Interesting Behavior of CL_s

In a calculation of -2lnQ without fits, events are weighted by their local s/b with the function ln(1+s/b).

So which outcome is more signal-like in a two-bin example:

Bin	Predicted s/b	Outcome 1	Outccome 2
1	1.0	20	16
2	5.0	20	21
Total n*ln(1+s/b)		**49.7**	**48.7**

Outcome 1 is more signal-like

Let's now scale the s/b's down by a factor of 10 (looking for a smaller signal). If the events were weighted with s/b, this wouldn't matter. But ln(1+s/b) is a nonlinear function (which is approximately s/b only for small s/b)

Bin	Predicted s/b	Outcome 1	Outccome 2
1	0.1	20	16
2	0.5	20	21
Total n*ln(1+s/b)		**10.01**	**10.04**

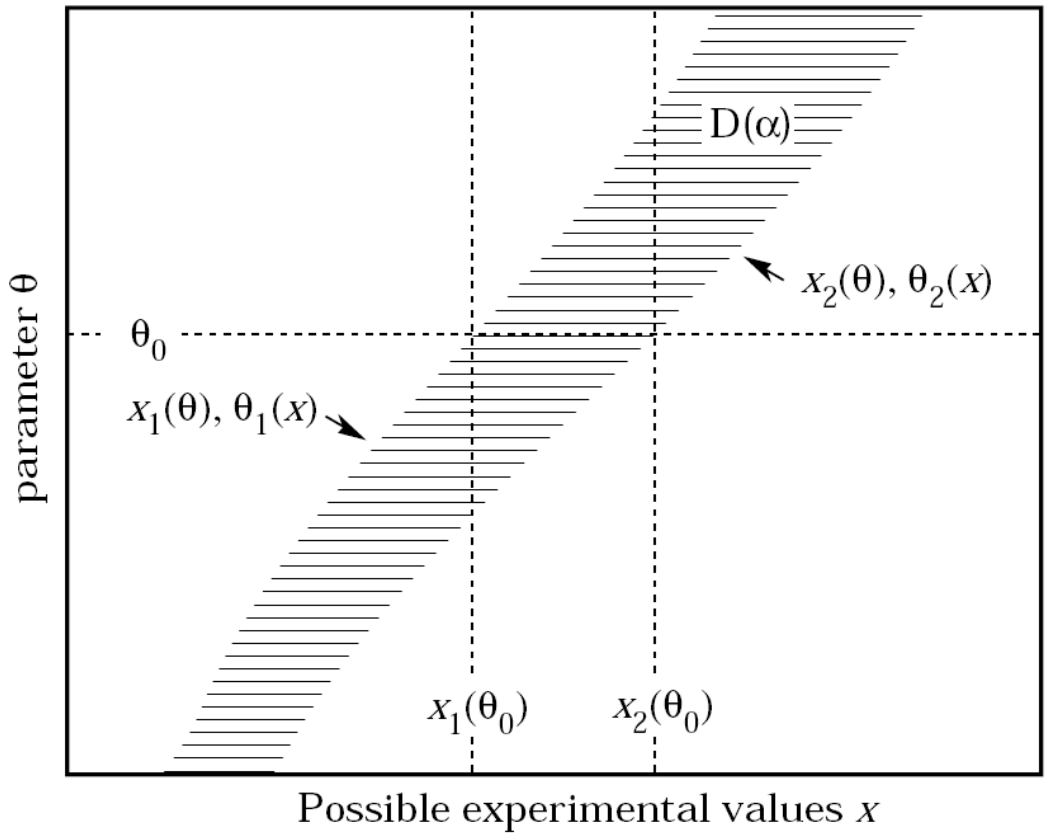
Outcome 2 is more signal-like

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.

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.

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!

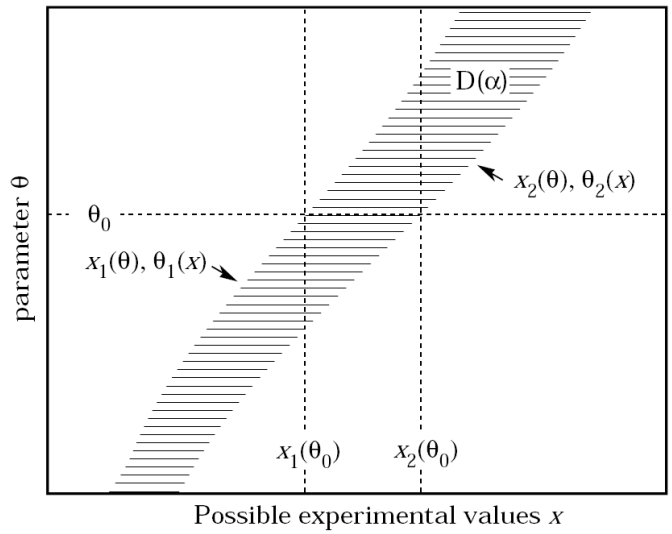
Take lowest x values: get lower limits.
Take highest x values: get upper limits.

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.



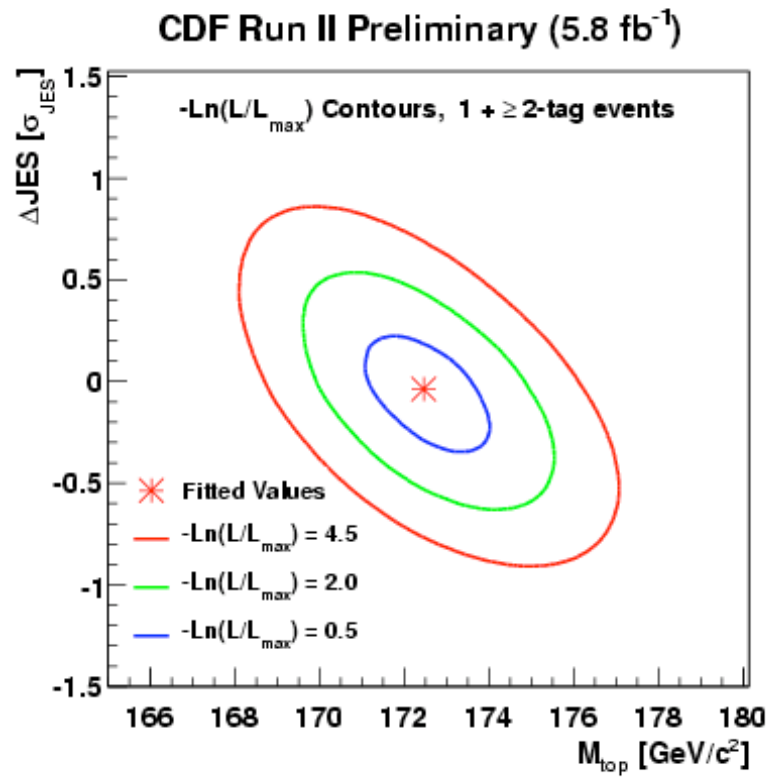
G. Feldman and R. Cousins,
"A Unified approach to the classical statistical analysis of small signals"
Phys.Rev.D57:3873-3889,1998.
arXiv:physics/9711021

No empty intervals!

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.

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