Statistical methods for the Large Hadron Collider

Diego Tonelli — CERN and INFN Trieste

Seminario per il corso di statistica per fisici — Universita' di Trieste
Jan 10, 2017

The LHC

An accelerator that collides, 40 million times per second, protons against protons at center-of-momentum energies of 7 to13 TeV. Collisions are analyzed by 8000 scientists from 4 large collaborations to explore the fundamental structure of matter and its interactions.

Primary goal: settle conclusively the mechanism of spontaneous breaking of the electroweak symmetry that generates the masses of elementary particles.

Done

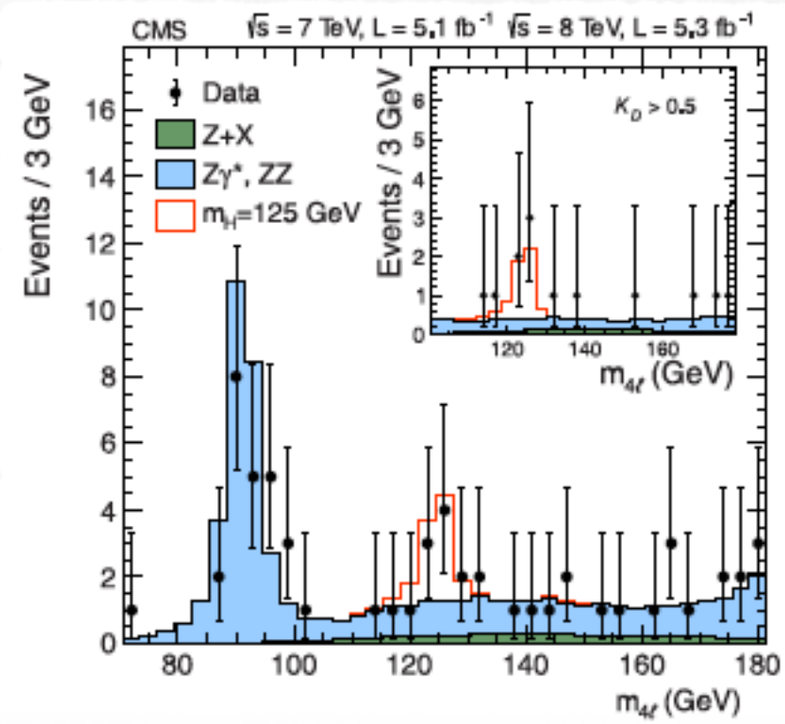
Contents lists available at SciVerse ScienceDirect

Physics Letters B

www.elsevier.com/locate/physletb

Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC ☆

CMS Collaboration *



Physics Letters B 716 (2012) 1–29

Contents lists available at SciVerse ScienceDirect

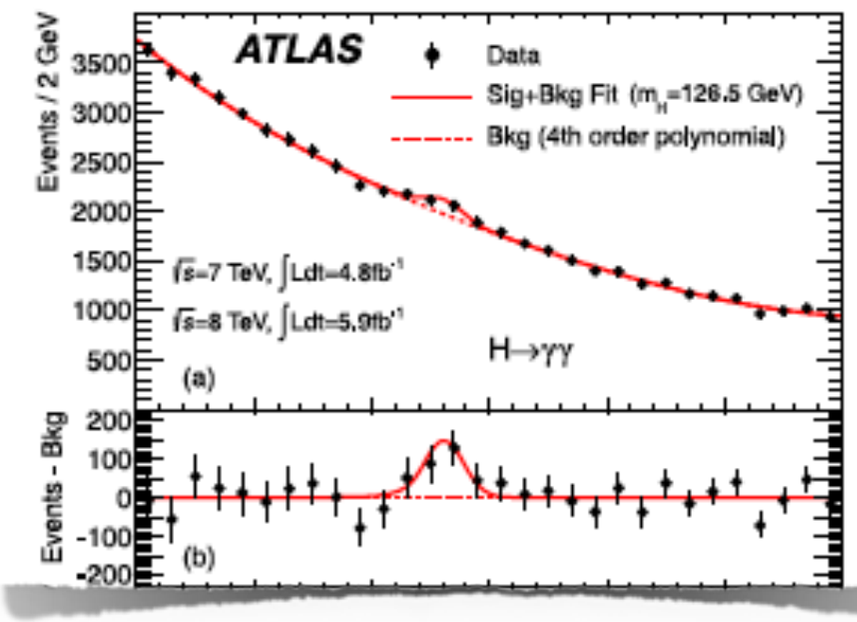
Physics Letters B

www.elsevier.com/locate/physletb

Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC ☆

ATLAS Collaboration *

This paper is dedicated to the memory of our ATLAS colleagues who did not live to see the full impact and significance of their contributions to the experiment.



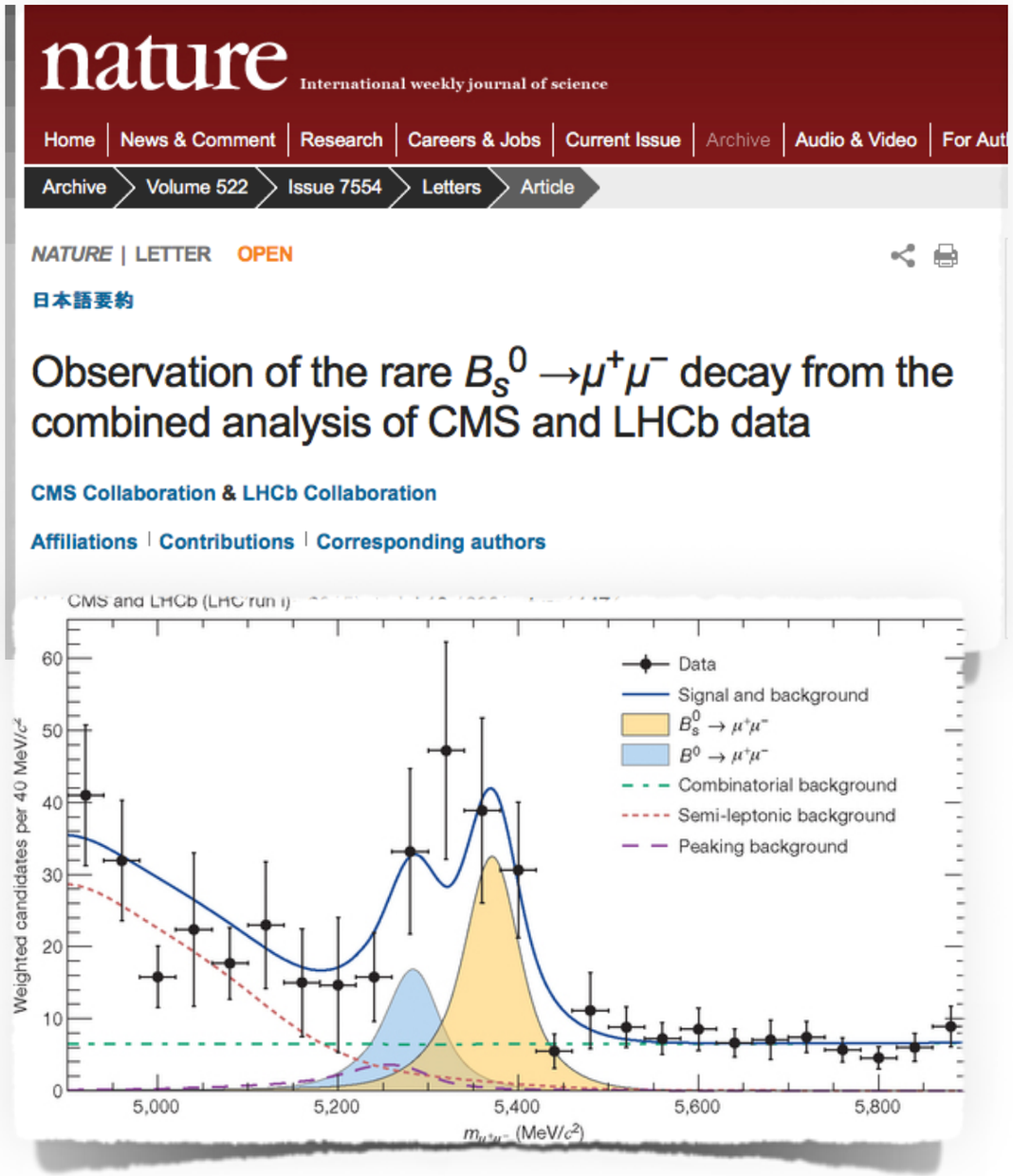
Not just Higgs

Hadron colliders at the energy frontier are machines with a broad discovery potential.

Most LHC physicists search for signs of the existence of new particles or interactions.

With luck, this effort may lead to discoveries. Otherwise, it will offer an improved understanding of known phenomena, useful to inform/guide future scientific decisions.

LHC experiments produce O(1000) physics measurements each year.

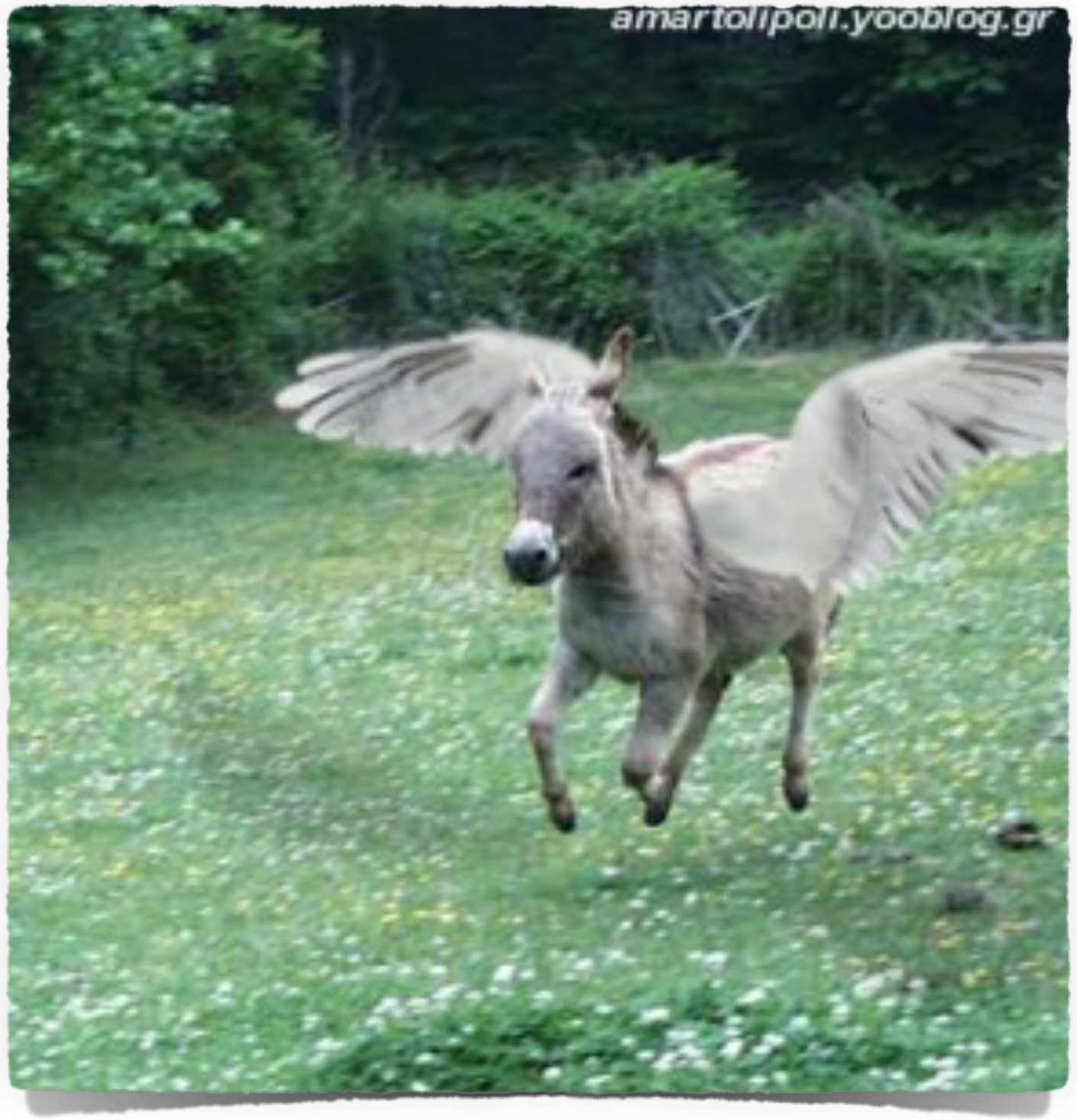


A proper statistical treatment of data is a key aspect of many of these measurements: minimize the risk of drawing wrong conclusions and maximize the amount and quality of extracted information.

Like for all hadron colliders at the energy frontier, the premier goal of the LHC is to observe unexpected physics phenomena (if they exists within its reach)

Hence, the chief LHC statistical challenge is to devise techniques to test efficiently whether the data support the solid observation of an unexpected physics phenomenon or not.

Why do I need statistics at all to discover anything?

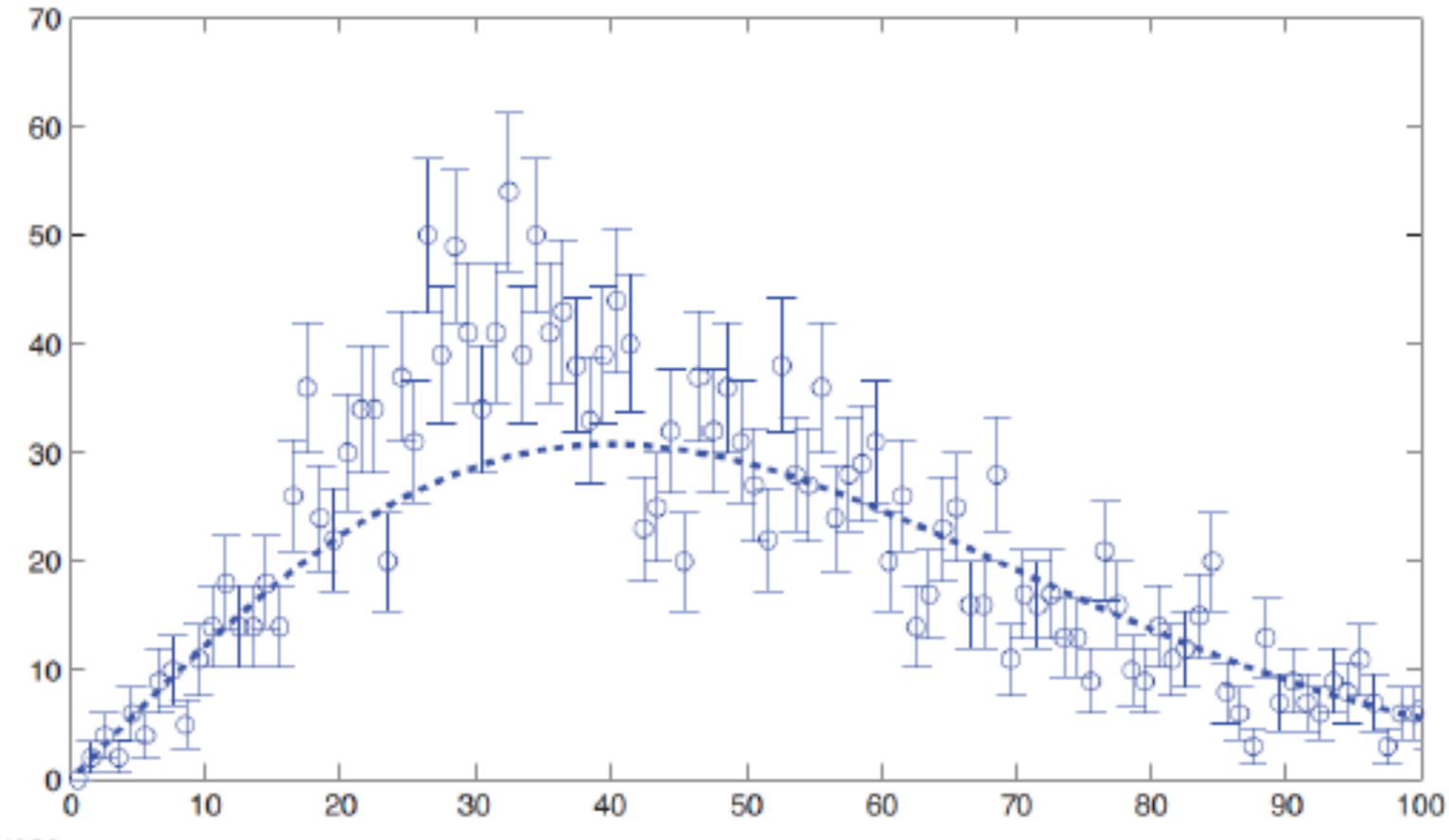


Not all sciences need statistics

An enthomologist has little doubt when he/she stumbles upon a previously unobserved insect. No need for histograms, or sophisticated data analysis. One “signal event” suffices when background is known to be zero certainly.

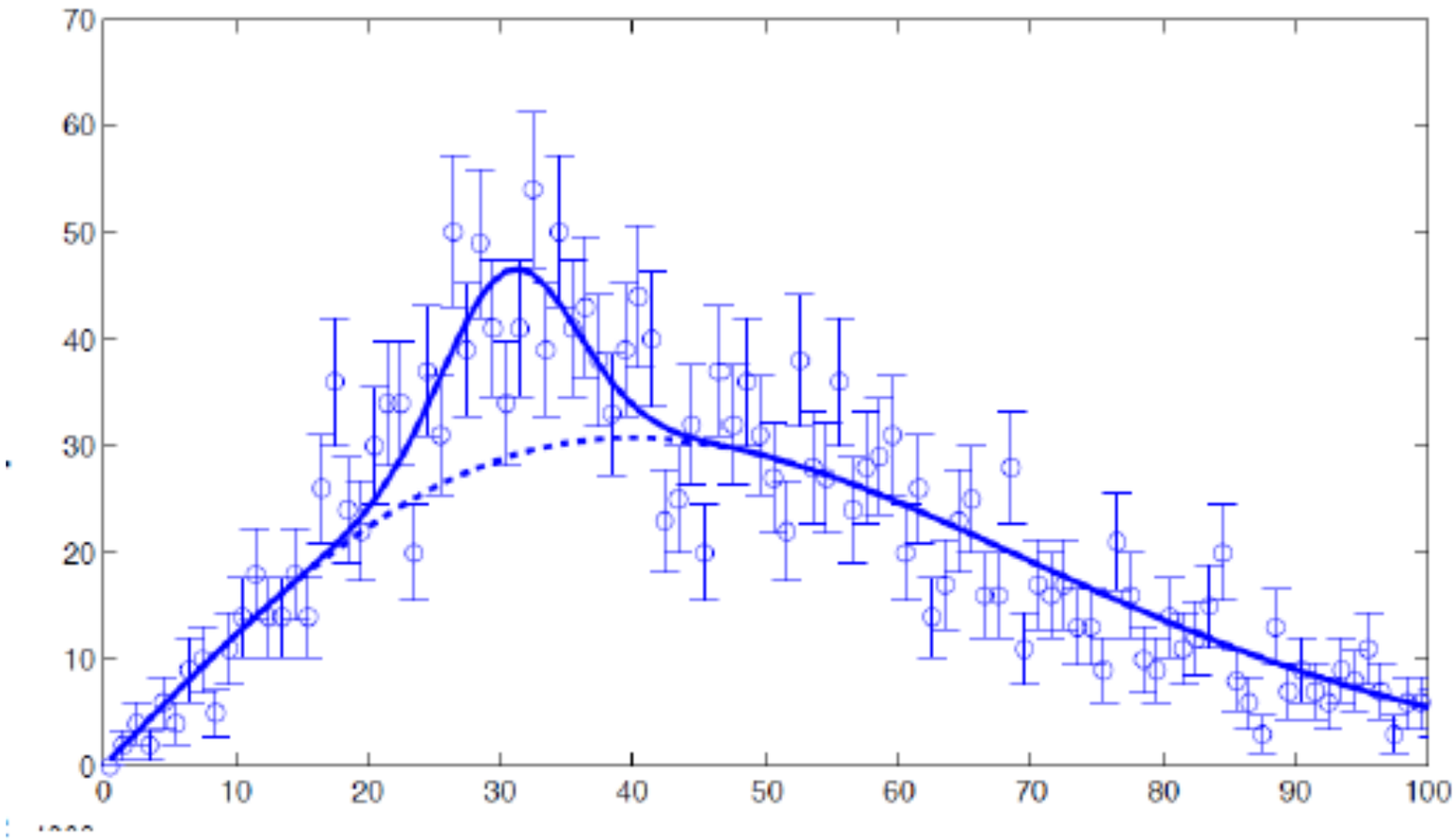
Background only?

Zillions of collisions, each recorded through millions of electronic channels, and reconstructed using complex kinematic/dynamical constraints. A lot of information to process and digest.



However, at the end of the day, it boil down to studying whether a small number of data distribution are compatible with expectations from known processes only ("background") or if they indicate contributions of new phenomena as well ("signal").

Or is there signal as well?



The challenge: how compatible data are with expectations from background? Is there a signal lurking? If so, what would be the statistical significance? And what is the most powerful way of telling the background apart from the signal+background ?

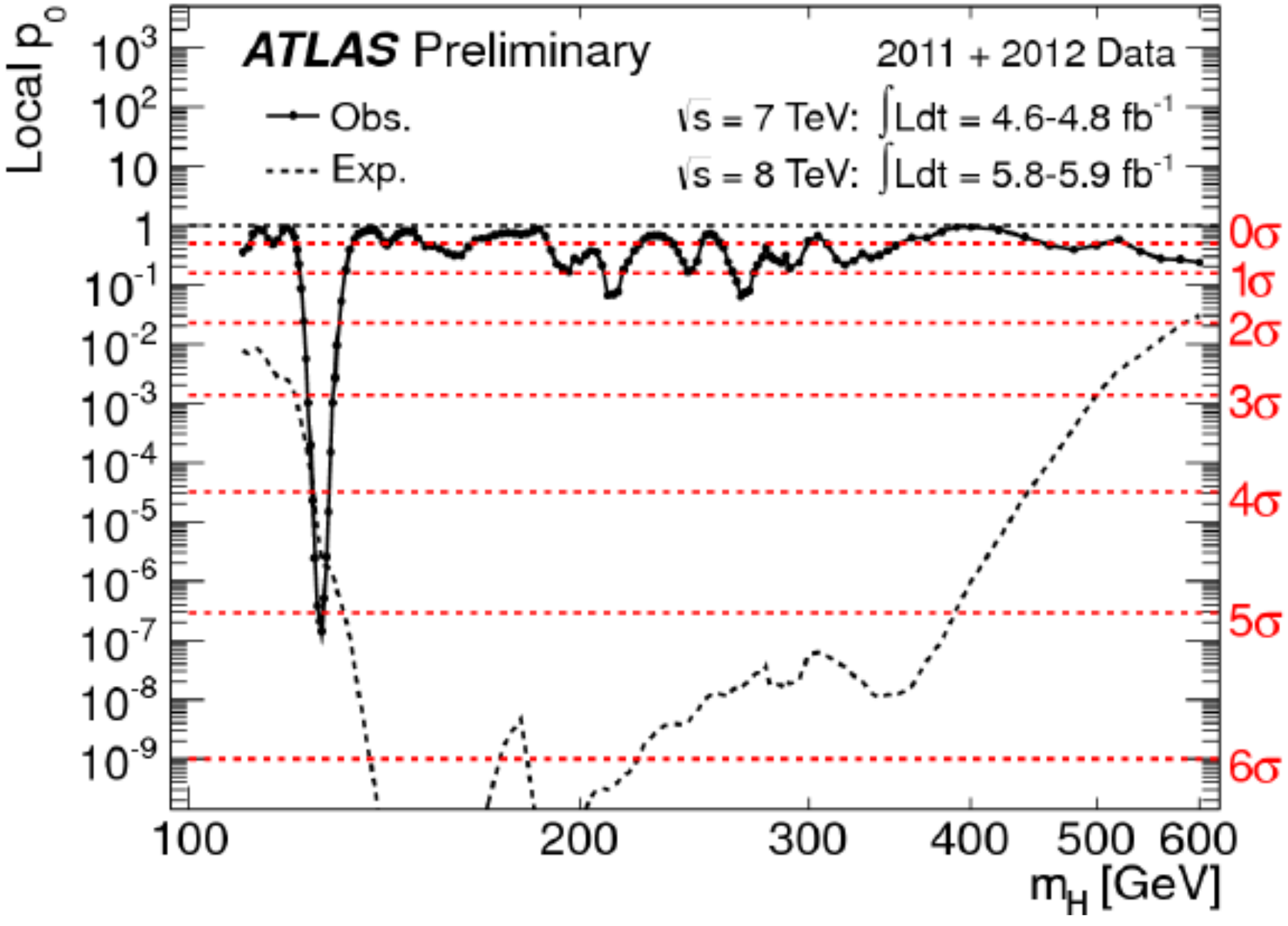
Today

- p-values, look-elsewhere-effect, 5-sigma folklore and all that

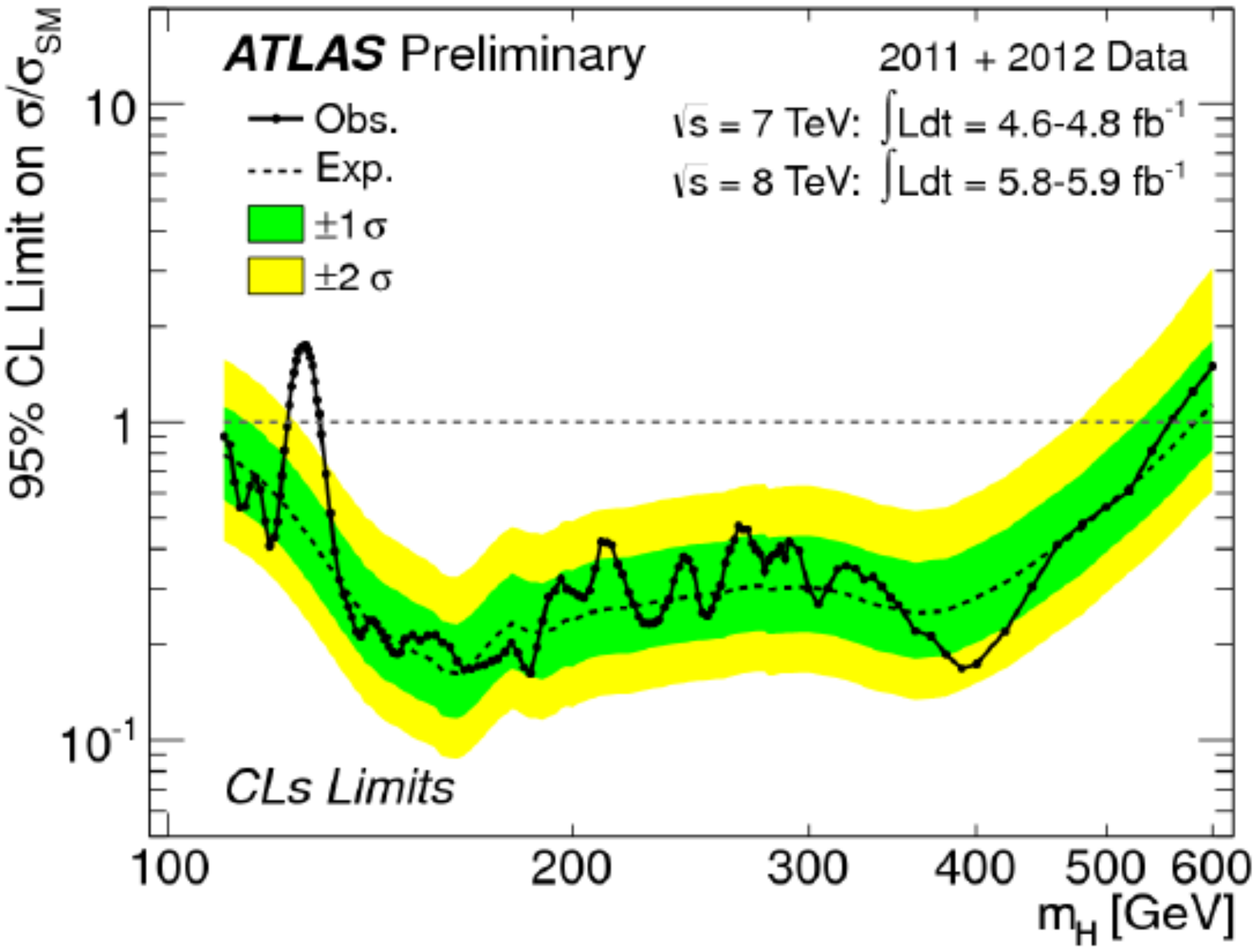
In god we trust, all others bring data

W. E. Deming

What is the p-value plot? What is the local p-value? What is the look-elsewhere-effect?



What does the “Brazil plot” mean? What is CLs?



Caveats

I am not a statistician nor did I give any original contribution to statistics. Just an enthusiastic practitioner, somewhat educated through 10+ years of data analysis in collider experiments.

Please, please, please: do interrupt me to ask questions. This is essential to keep us awake. Also, feel free to follow-up at diego.tonelli@cern.ch

Will make my slides available to prof. E. Milotti soon so that he can share them.

Is there a deviation? Is it significant?

Experimentalists often need to judge if an apparent anomaly in the observed data qualifies as a significant departure from the expectations of known phenomena or, rather, if it's likely to arise from statistical fluctuations of known phenomena.

This is the first thing you do if you suspect you may have a discovery (and in many other less exciting cases)

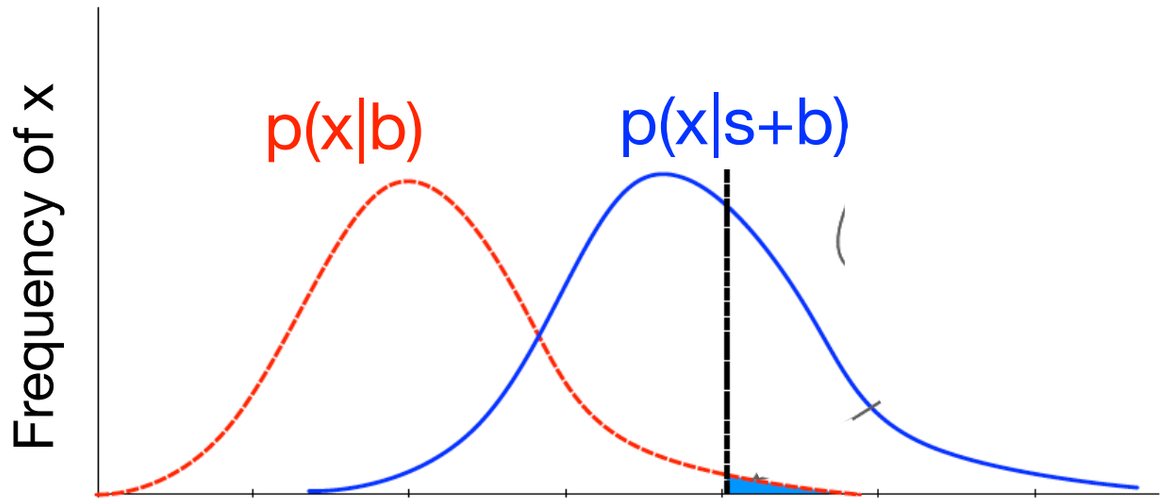
At LHC (and in particle physics at large) this is mostly addressed using "p-values", a (strongly debated) concept of frequentist statistics.

A p-value is a random variable that provides a quantitative evaluation of the probabilities to be observing a genuine anomaly or a fluke.

(Check this out for an entertaining piece on the birth of the p-value notion http://priceonomics.com/the-guinness-brewer-who-revolutionized-statistics/)

Step1: p-value ingredients

Two hypotheses. 1. only known phenomena contribute ("bckg-only" — or "null" — hypothesis). 2. new phenomena contribute as well ("signal" — or "alternate" — hypothesis)



function x of the data that is distributed differently btw the two hypotheses

Choose function x of the data (e.g., signal-event count), whose distribution under the null p(x|b) differs from that under the signal hypothesis p(x|s+b). Construct both, p(x|b) and p(x|s+b) (most labor-intensive step, typically done using simulation).

Set, prior to looking at data, the acceptable false-positive rate: how signal-like the observed x should be to exclude the null? Or how background-like x should be to exclude the signal.

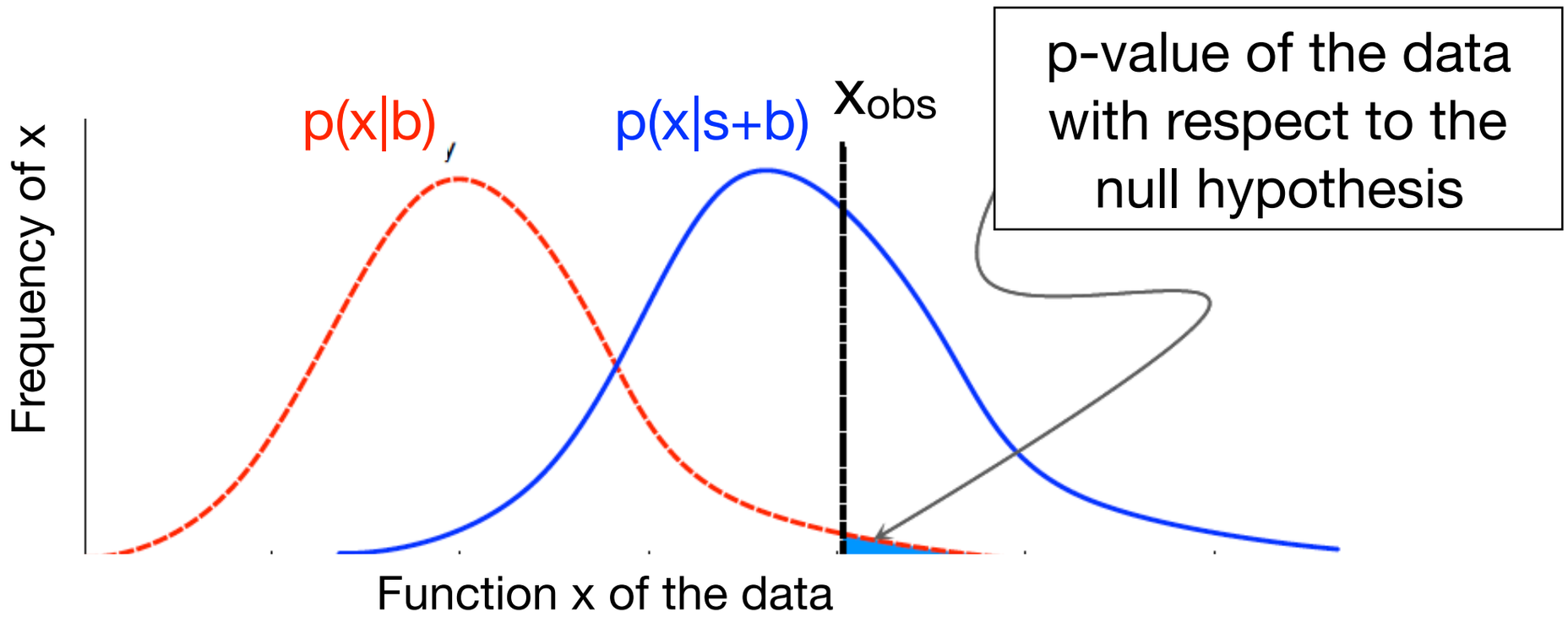
Step 2: look at the data

That is, look at what particular value x_{obs} the quantity x takes up in your data

Discovering a new effect

The location of x_{obs} relative to the two shapes may offer a quantitative measure of the probability that one is observing a fluctuation or a new phenomenon.

p-value is the relative fraction of the integral of the bckg-only model over values of x as signal-like as that observed and more. The smaller the p-value, the stronger the evidence against the null. If p-value < false-positive rate, the bckg-only is excluded at a confidence level CL = 1- (p-value).

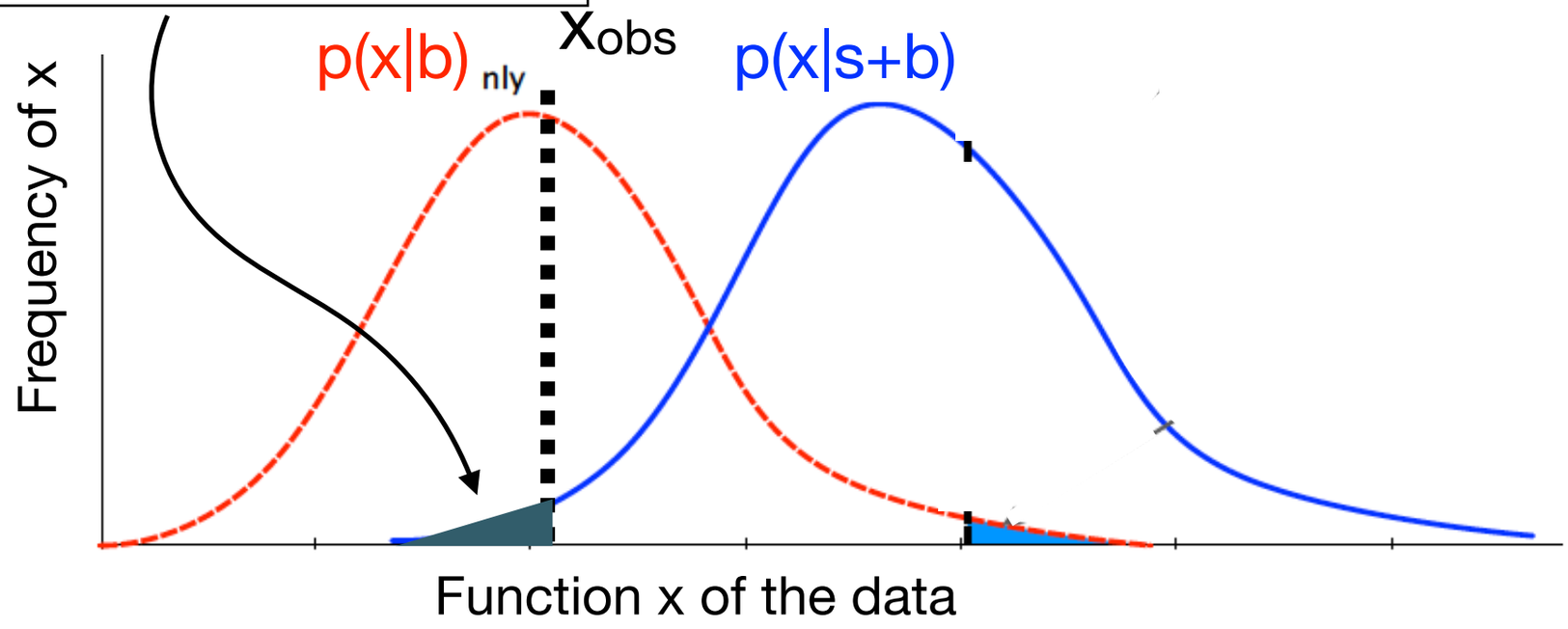


Excluding a new effect

If the purpose is to exclude a new effect, one tests the signal hypothesis, and quotes the p-value with respect to that.

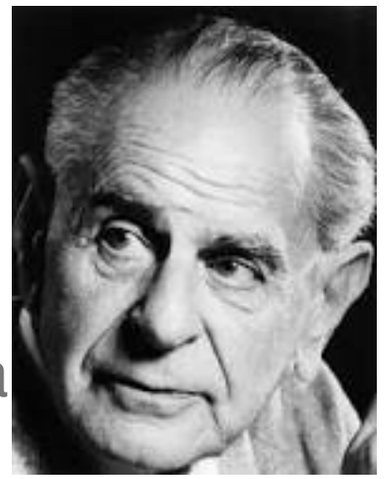
Is the relative fraction of the integral of the signal model over values of x as background-like as that observed and more. The smaller the p-value, the stronger the evidence against the signal hypothesis.

p-value of the data with respect to the signal hypothesis



Testing nature the Popperian way

Cannot prove that an hypothesis is true, only that it is false.



Karl Popper (1902-1994)

"Discover" a signal by excluding its absence with a high level of significance (i.e., by observing that the probability to observe our data if background only contributes is tiny)

"Exclude" a signal by excluding its presence with a high level of significance.

A p-value is not a probability! It is a random variable (function of the data) that is distributed uniformly between 0 and 1 if the tested hypothesis is true.

It does not express the probability that an hypothesis is true or false!

Wrong claim "The measurement shows that the probability for hypothesis blah is .."

P-values connect to the probability to observe x_{obs} or a more extreme value *if an hypothesis were true*. Proper claim: "Assuming that the hypothesis blah holds, the probability to observe a fluctuation as extreme as that observed in our data or more is…"

HEP lingo and folklore

Physicists have less feel for p-values than for “sigmas…” .HEP lingo goes like “at how many sigma such and such result is significant”

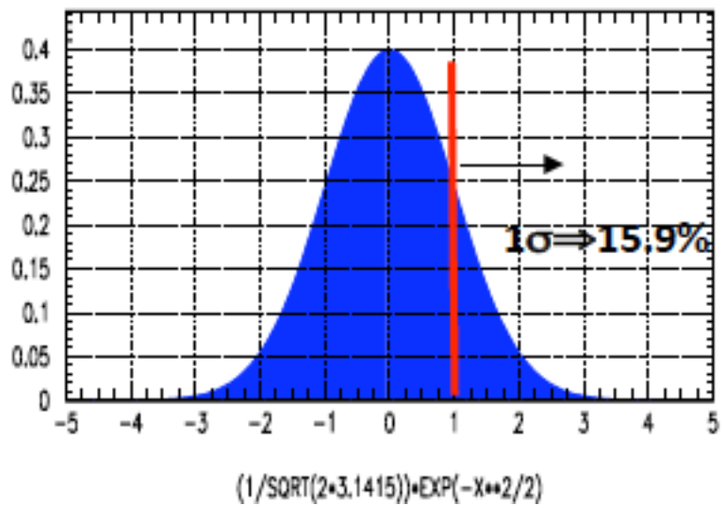
The “number of sigma” (or z-value) is just a remapping of p-values into integrals of one tail of a Gaussian. It expresses by how many sigma from the mean my observation would be if the test statistic x would be distributed as Gaussian

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

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

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



Examples: p-values in coin tossing

Check if a coin is fair. The probability to observe j heads in n trials is binomial

$$f(j;n,p) = \binom{n}{j} p^j (1-p)^{n-j} = \frac{n!}{(n-j)!j!} p^j (1-p)^{n-j}$$

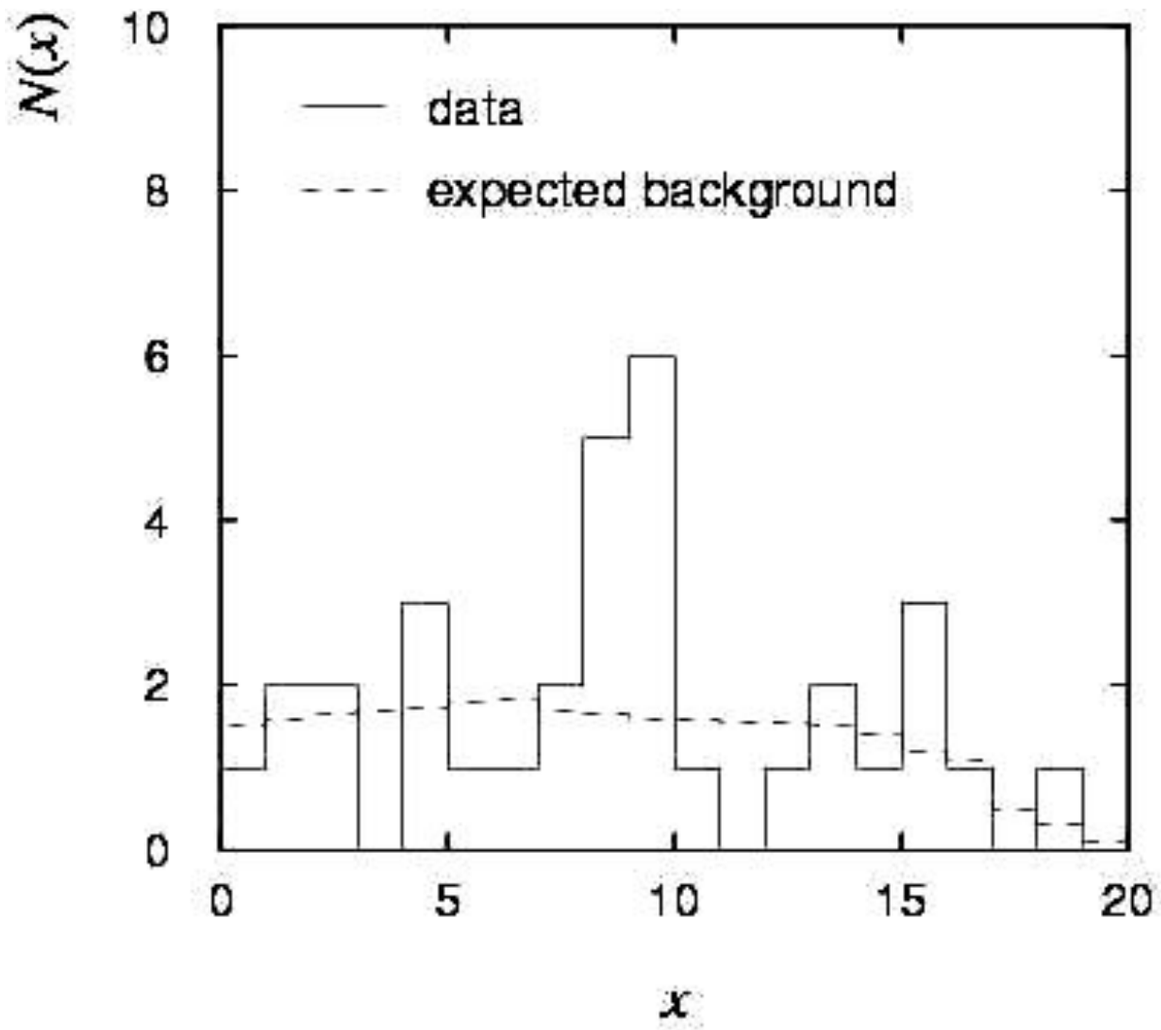
Null hypothesis: the coin is fair (p=0.5). Get 17 heads out of 20 trials. Regions of data space with equal or lesser compatibility with null, relative to j=17 include n=17, 18, 19, 20, *0, 1, 2, 3.*

P(n=0,1,2,3,17,18,19,or 20) = 0.26%.

Hence, if the null were true (coin is fair) and we would repeat the experiment many times, only 0.26% of the times we would obtain a result as extreme or more than that observed.

p-values in mass peak

Suppose you measure a value x for each event and bin the resulting distribution.



The count in each bin is a Poisson random variable, whose mean in the bckg-only hypothesis is given by the dashed line

$$P(n;s,b) = \frac{(s+b)^n}{n!}e^{-(s+b)}$$

Observe a peak of 11 events in two central bins, with expected background 3.2 events

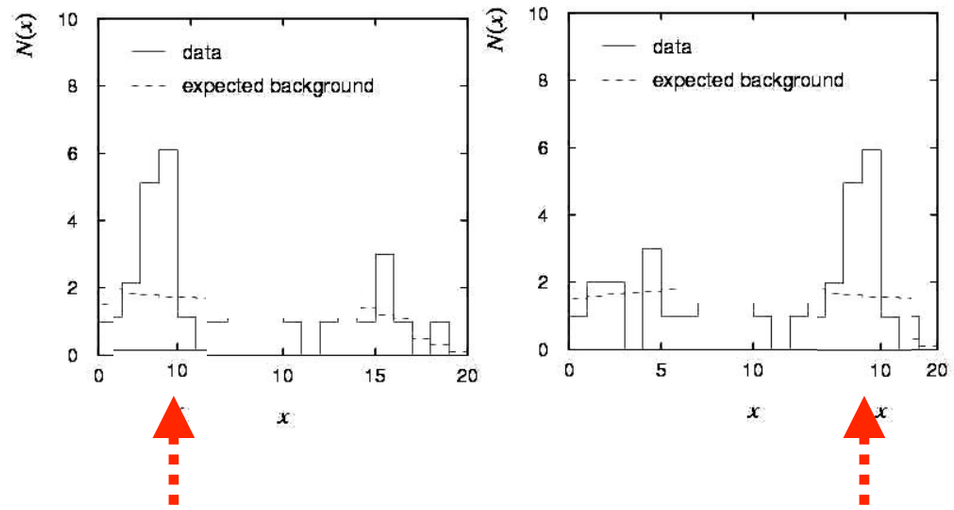
P-value for the background-only hypothesis is P(n>=11, b=3.2, s=0) = $5*10^{-4}$

Is this evaluation fair or biased?

Look-elsewhere

It's biased because it only accounts for the chances of a upward fluctuation in that very position at x~9.

That's the "local p-value".



Peak could have been observed here

..or here

To get the "p-value" need to account for the chances that such excess could show up in any pair of adjacent bins. With 20 bins (10 pairs of adjacent bins) the local p-value gets multiplied by ≈10 to get the global p-value.

Lots of bins (that is, search channels) implies lots of chances at fluctuations.

Correcting for multiple testing

When quoting p-values, need to include the effect of multiple testing.

That is, properly accounting that we have also been “looking elsewhere” from the region where the anomaly is observed in our very data sample, but have dismissed all the search channels boringly showing uneventful, background-like behavior.

The larger the size of the test space, the higher the probabilities to observe rare fluctuations.

That is why in HEP the standard conventional threshold for credibly claiming a solid observation of an unexpected effects is kept very high (5σ)

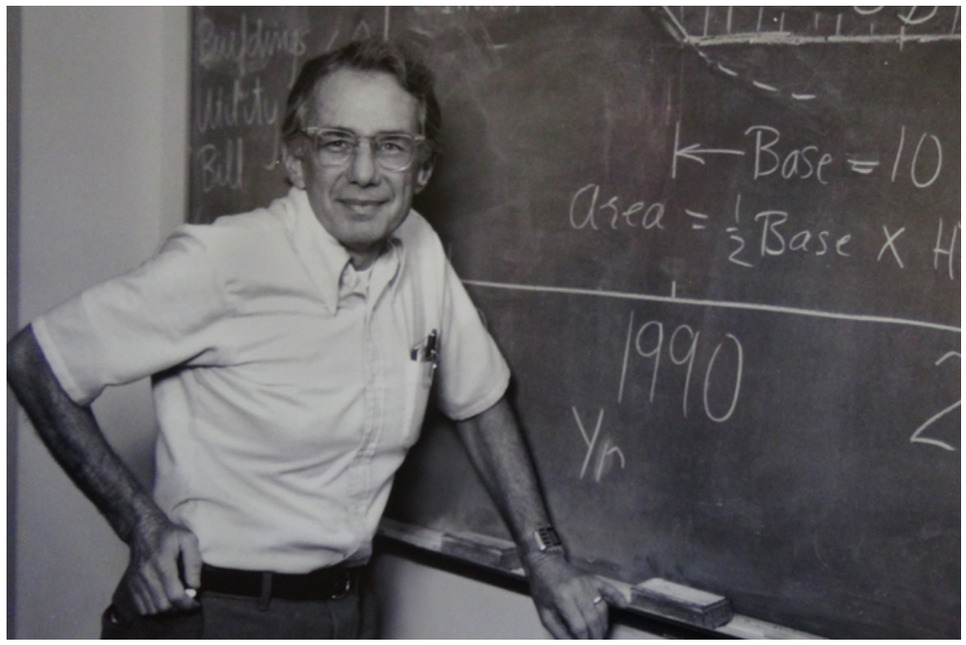
Claim a new effect only when the chances for it to be resulting from a fluctuation due to known phenomena are 0.000029% or less…

Short aside (not LHC stuff)

The birth of the 5σ criterion

Far-out hadrons

In 1968, Art H. Rosenfeld at UC Berkeley surveyed the searches for exotic hadrons that did not fit the then-new static quark model.



He noted that the number of discovery claims quite matched with the number of statistical fluctuations expected in the data sets analyzed.

Rosenfeld blamed the large mutliple testing corrections needed to account for the massive use of combination of observed particles to construct mass spectra containing potential exotic excesses.

"[...] This reasoning on multiplicities, extended to all combinations of all outgoing particles and to all countries, leads to an estimate of 35 million mass combinations calculated per year. How many histograms are plotted from these 35 million combinations? A glance through the journals shows that a typical mass histogram has about 2,500 entries, so the number we were looking for, h is then 15,000 histograms per year. [...] Our typical 2,500 entry histogram seems to average 40 bins. This means that therein a physicist could observe 40 different fluctuations one bin wide, 39 two bins wide, 38 three bins wide... This arithmetic is made worse by the fact that when a physicist sees 'something', he then tries to enhance it by making cuts..."

"

Far-out hadrons

*"In summary of all the discussion above, I conclude that each of our 150,000 annual histograms is capable of generating somewhere between 10 and 100 deceptive upward fluctuations [...] To the theorist or phenomenologist the moral is simple: **wait for nearly 5σ effects**. For the experimental group who has spent a year of their time and perhaps a million dollars, the problem is harder... go ahead and publish... but they should realize that any bump less than about 5σ **calls for a repeat of the experiment.**"*

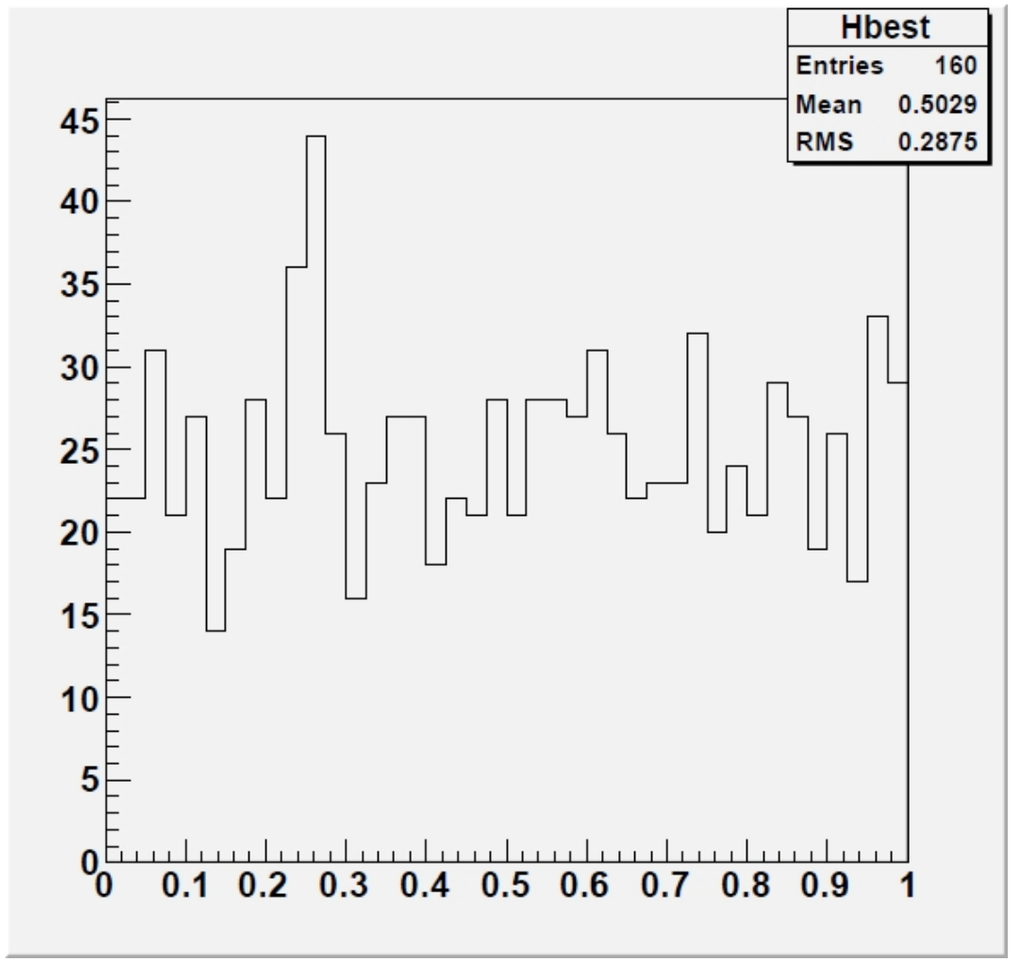
Rosenfeld also mentions the semiserious GAME test by his colleague, Gerry Lynch

"My colleague Gerry Lynch has instead tried to study this problem 'experimentally' using a 'Las Vegas' computer program called Game. Game is played as follows. You wait until a unsuspecting friend comes to show you his latest 4-sigma peak. You draw a smooth curve through his data (based on the hypothesis that the peak is just a fluctuation), and punch this smooth curve as one of the inputs for Game. The other input is his actual data. If you then call for 100 Las Vegas histograms, Game will generate them, with the actual data reproduced for comparison at some random page. You and your friend then go around the halls, asking physicists to pick out the most surprising histogram in the printout. Often it is one of the 100 phoneys, rather than the real '4-sigma' peak."

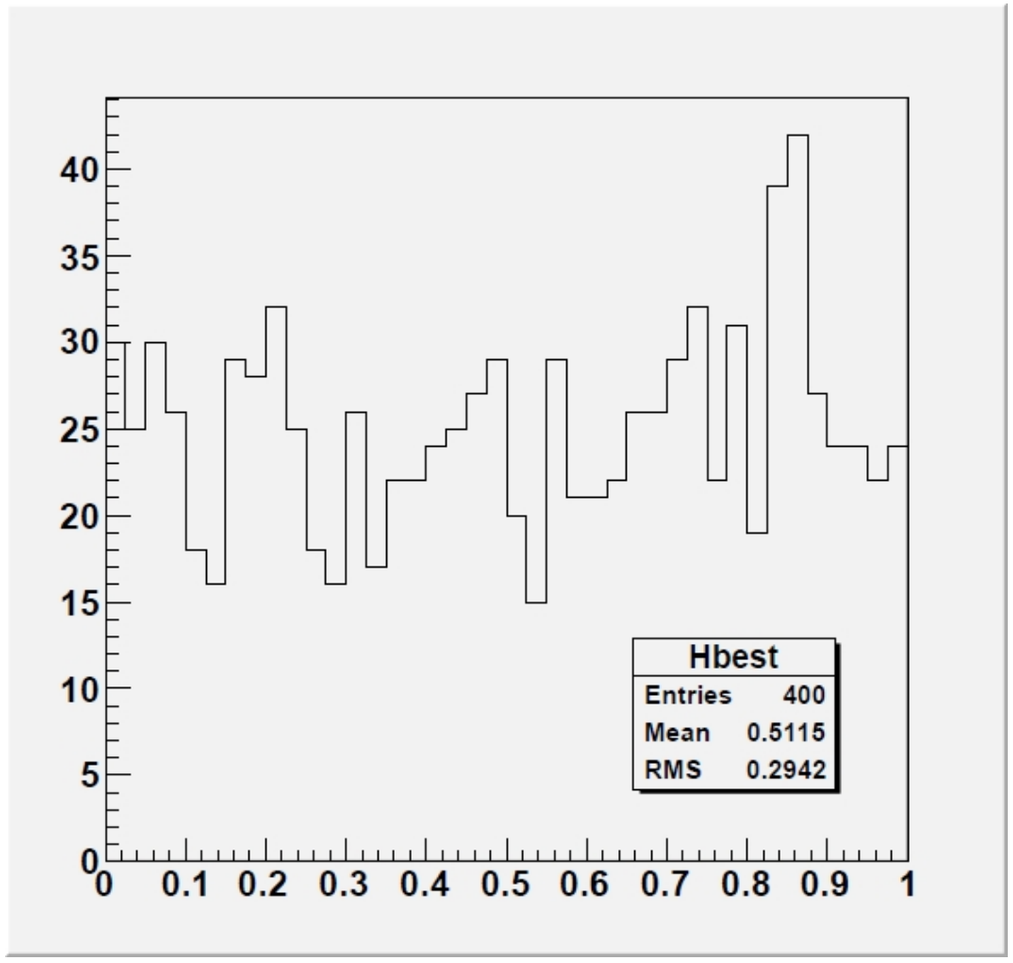
Let’s play GAME

PS: Each histogram selected as the one with the most striking pair of adjacent bins from a set of 100 histograms generated according to a uniform distribution

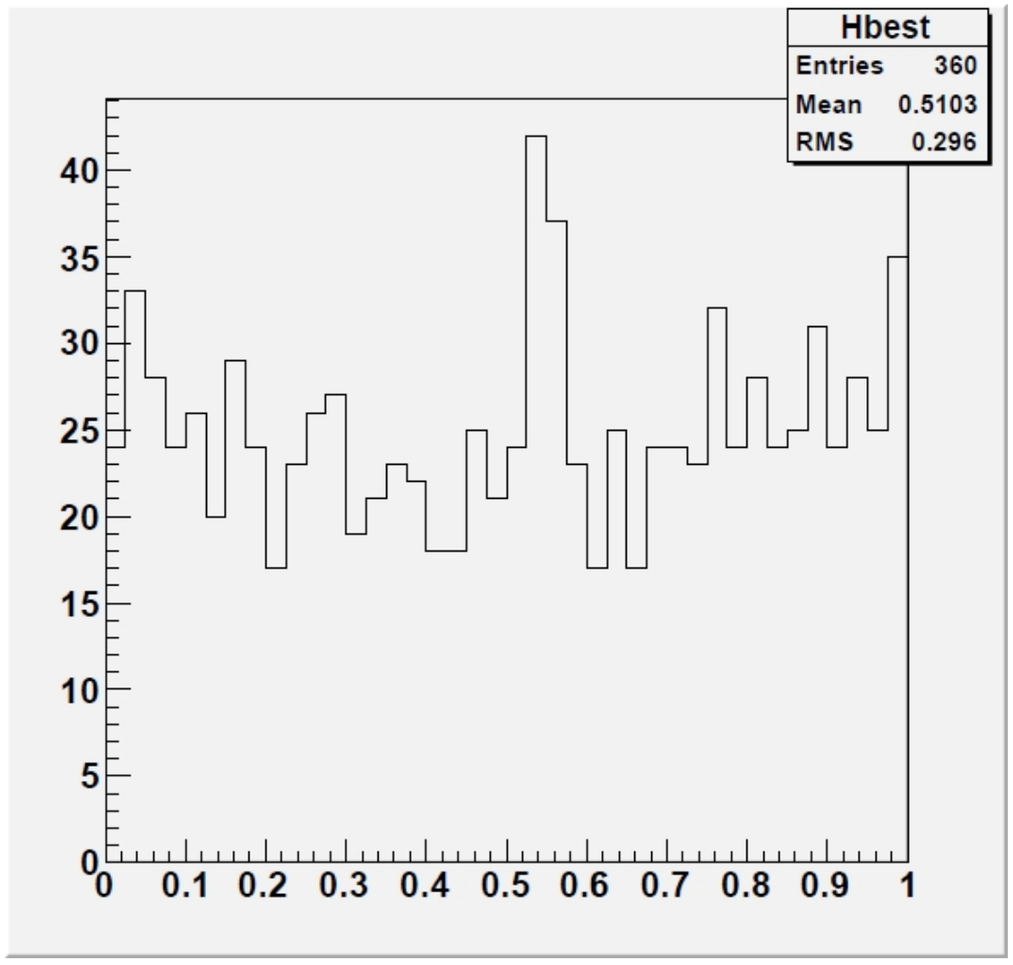
Two-bin bumps



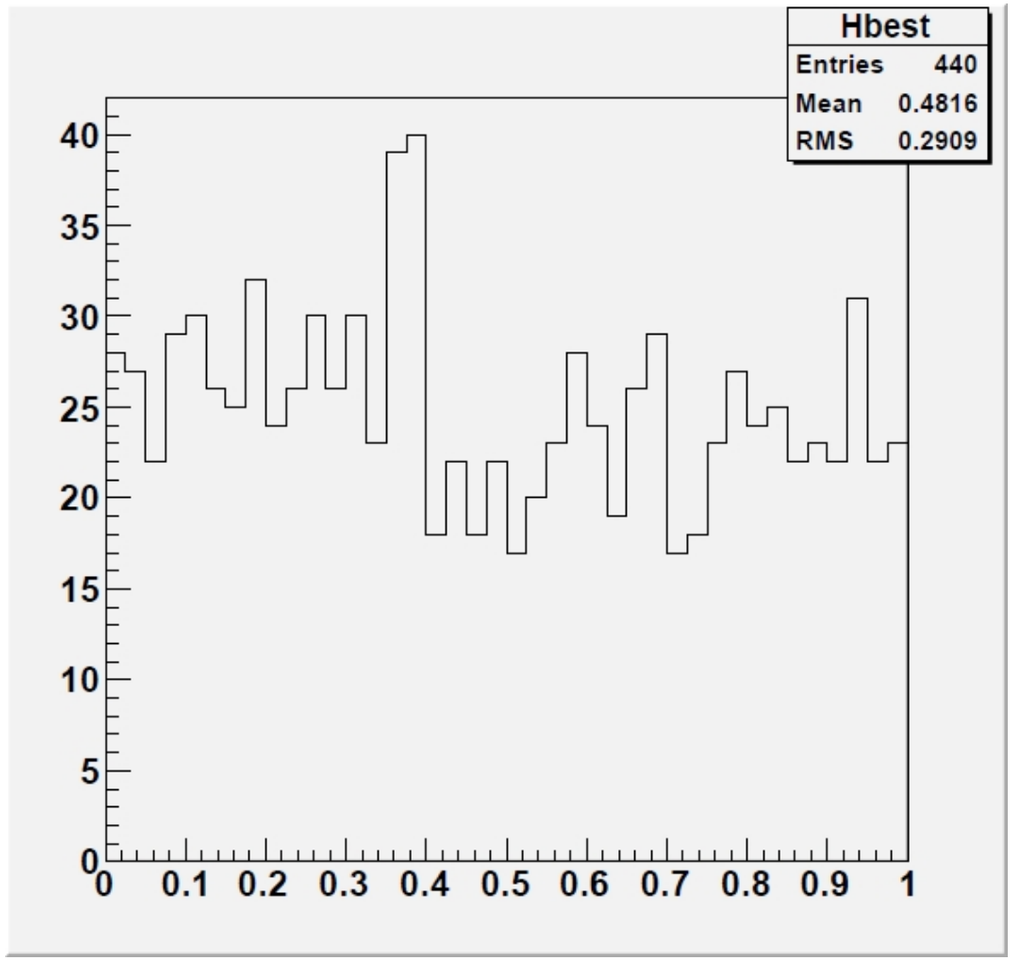
Two-bin bumps



Two-bin bumps



Two-bin bumps



End aside

Dealing with the effect of multiple testing

Various semiempiric recipes to determine a LEE-correction for local p-values.

Rough bump-hunting: multiply the local p-value by the range of the inspected histogram divided by the typical resolution on the inspected parameter.

Bonferroni-Dunn: multiply the local p-value by the number of independent models (not bins!) sought [C.E. Bonferroni, Teoria statistica delle classi e calcolo delle probabilità, Istit. Sup. di Scienze Econ. e Comm. di Firenze (1936); J.O. Dunn, Ann. of Mathematical Stat., 30 (1), 192 (1959) and J. of the American Stat. Assoc. 56 (293) 52, (1961)]

Some issues: adjusted p-value can exceed unity (!); unclear how to account for empty histogram bins or for regions where new phenomena have already been excluded by previous experiments.

Dealing with the effect of multiple testing (cont'd)

Dunn-Sidak: global p-value = 1- (1- local p-value)^n assuming n independent tests
Z.K. Sidak, J. of the American Stat. Assoc. 62 (318) 626, (1967)

Gross-Vitells for bump hunt over smooth background: more involved but precise estimation of correction E. Gross and O. Vitells, Eur. Phys. J. C70, 525 (2010), 525

Sufficient for a semiqualitative feel of the effect. Harder in analyses like Higgs searches, where p-value results from combining many channels, each contributing a different weight and entering with different experimental sensitivities.

Ideal would be a p-value of p-values. Take p-value as test statistic and look at the distribution of smallest p-values. Hard and laborious.

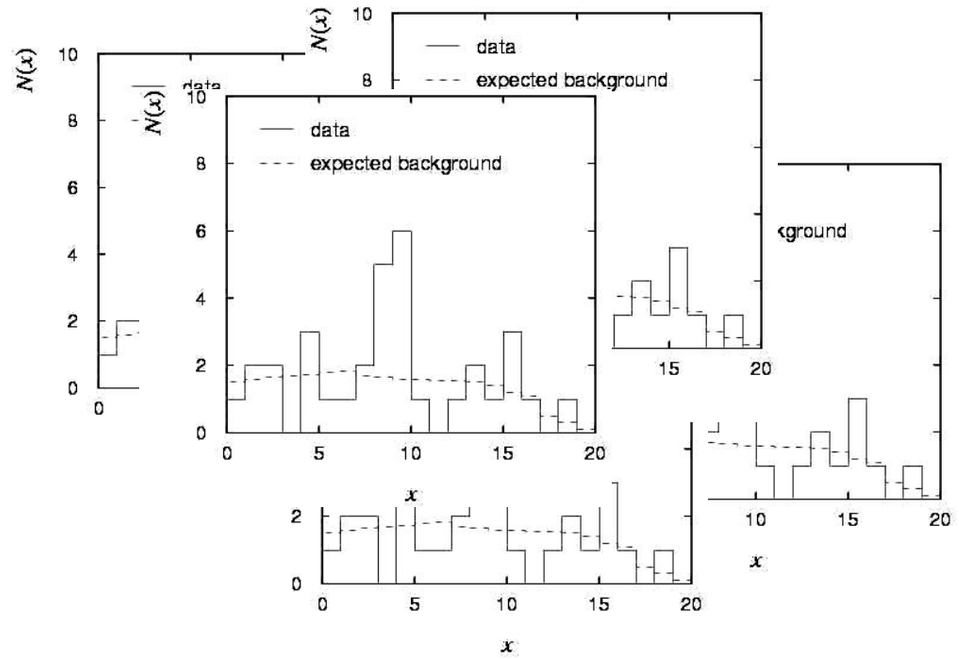
Where is “elsewhere”?

Tenths, or hundreds, or thousands of distributions may have been inspected, in the same analysis or in other analyses.

Should we correct for these as well?

How large is the testing space to base our correction on?

Should we go back and correct previously published p-values when new analyses are completed?



(Arbitrary) guidance (consensus at the Banff 2010 Statistics Workshop): limit the testing space to models (i.e, plots) that are inspected within a single published analysis

The conventional “5σ rationale”

HEP experimenters conventionally agreed to deal with the LEE by setting a rather extreme standard for p-values to justify claims of new effects.

One requires the null to be rejected with significance of 3.5σ (for “evidence”) and 5σ (“observation”), corresponding to very small p-values (fluctuations that occur 3 times every 10 million trials). (See www.huffingtonpost.com/victor-stenger/higgs-and-significiance_b_1649808.html for an historical recollection)

The loose rationale is that such high thresholds should protect from the shortcomings discussed above.

Does this actually protect?

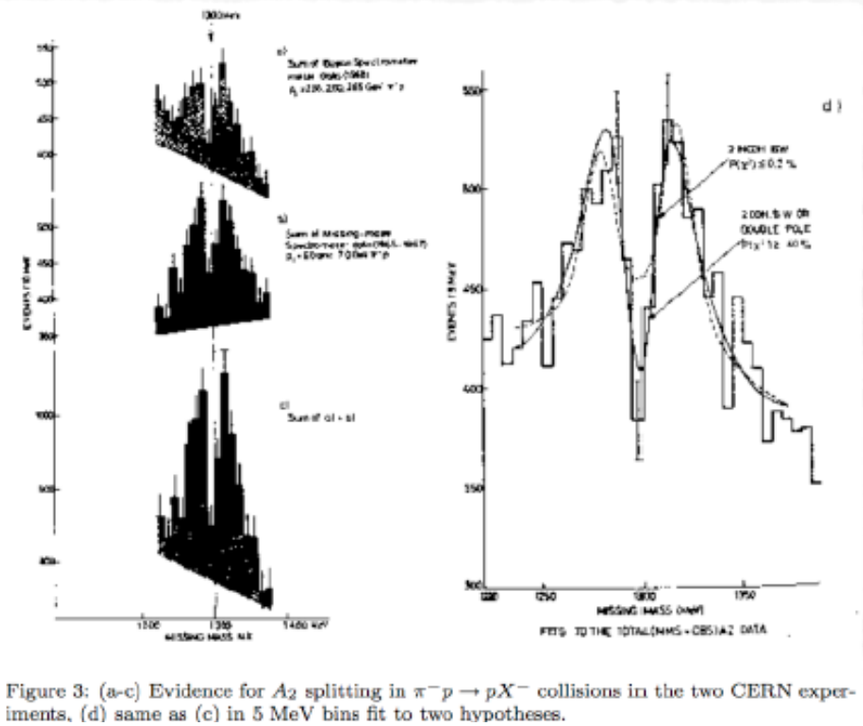
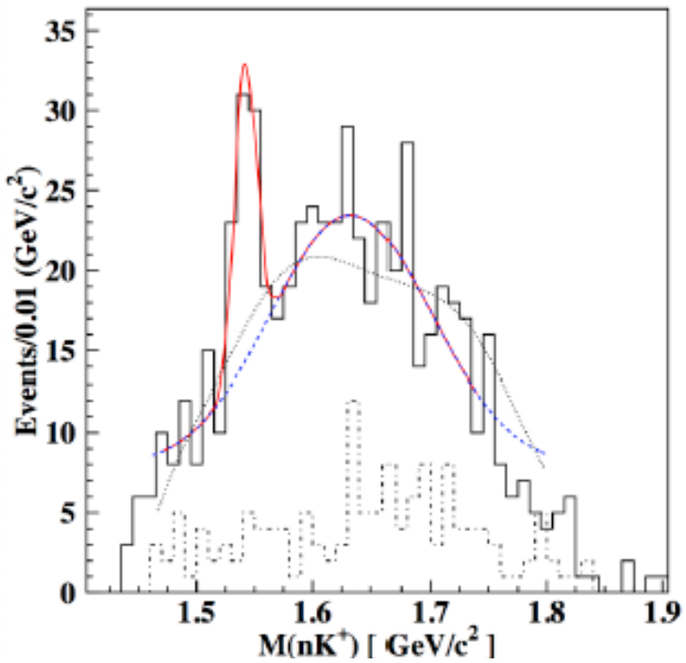
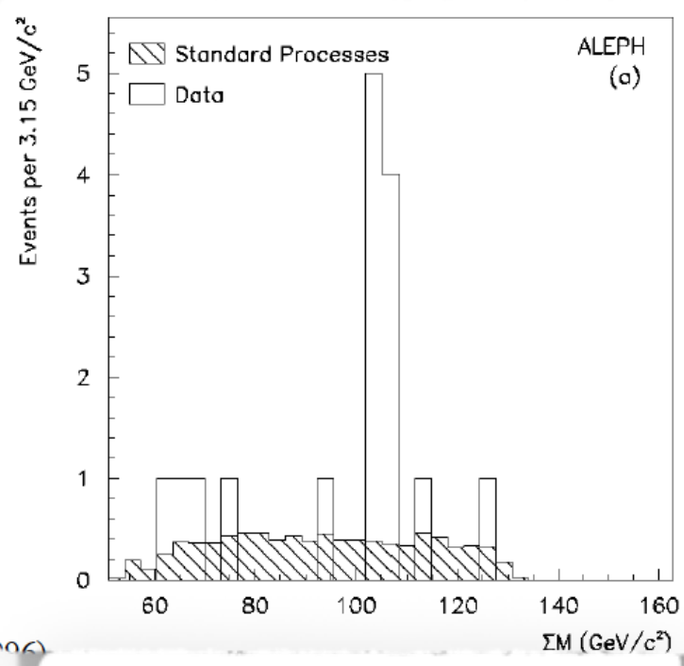


Figure 3: (a-c) Evidence for A_2 splitting in $\pi^- p \to pX^-$ collisions in the two CERN experiments. (d) same as (c) in 5 MeV bins fit to two hypotheses.

Split A2 resonance, CBS and MMS collaborations, CERN mid-60ies
http://arxiv.org/pdf/hep-ph/



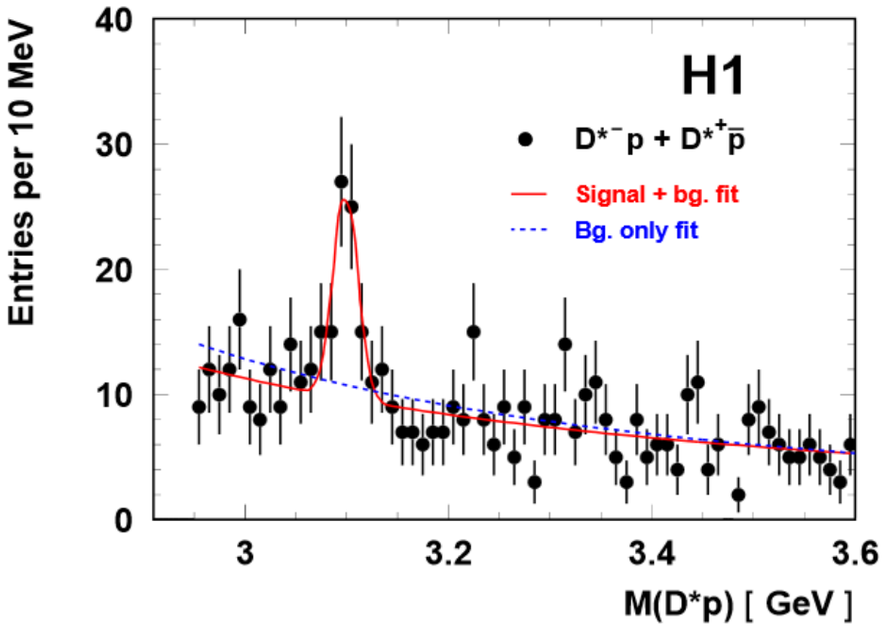
Observation of Pentaquarks
CLAS Collab. PRL 91 (2003) 252001



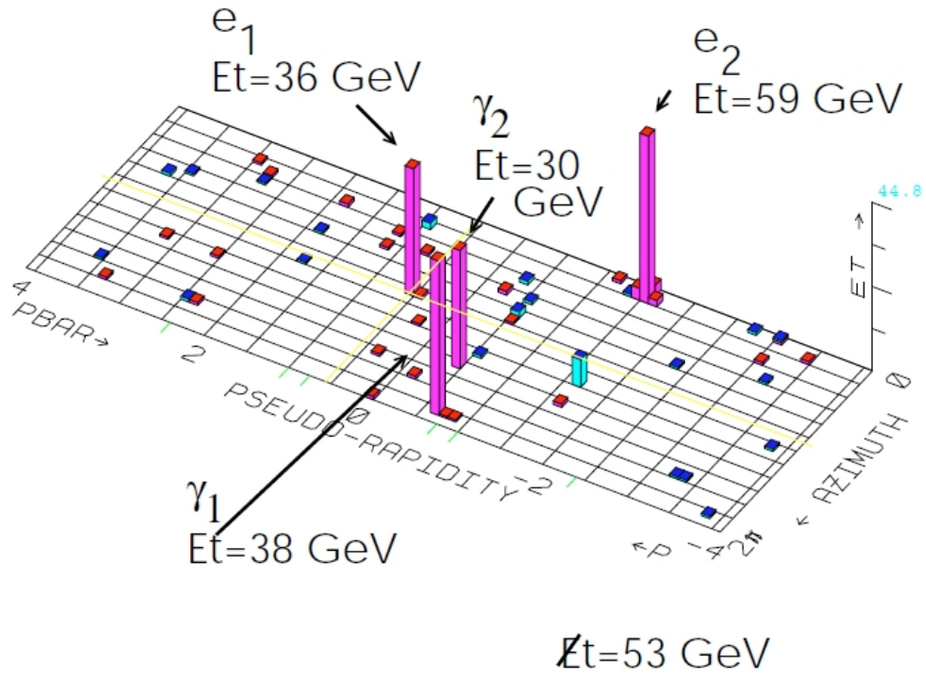
"the width of the bins is designed to correspond to twice the expected resolution ... and their origin is deliberately chosen to maximize the number of events found in any two consecutive bins"

ALEPH collaboration, CERN,
Z. Phys. C71 (1996) 179

....?



H1 pentaquarks PLB 588 (2004) 17



CDF impossible event (1995)

An emerging pattern?

Claim	Claimed Significance				Verified or Spurious
Top quark evidence	3				True
Top quark observation			5		True
CDF bbγ signal		4			False
CDF eeggMEt event				6	False
CDF superjets				6	False
Bs oscillations			5		True
Single top observation			5		True
HERA pentaquark				6	False
ALEPH 4-jets		4			False
LHC Higgs evidence	3				True
LHC Higgs observation			5		True
OPERA v>c neutrinos				6	False
CDF Wjj bump		4			False
LHC 750 GeV diphoton		4			False

A one-size-fits-all threshold seems not to fully encapsulate the complexity of the problem. Should one tune/correct the threshold based on the “a priori” expectation for the effect? (is 5σ adequate a threshold for significance of life on Mars?)

Another major issue: how does one include systematic uncertainties in p-values?

What is systematics?

Hard to find a precise, rigorous definition.

In experimental physics one assesses systematic uncertainties all the time, but when it comes to define them only semi-empiric definitions exist, based on examples.

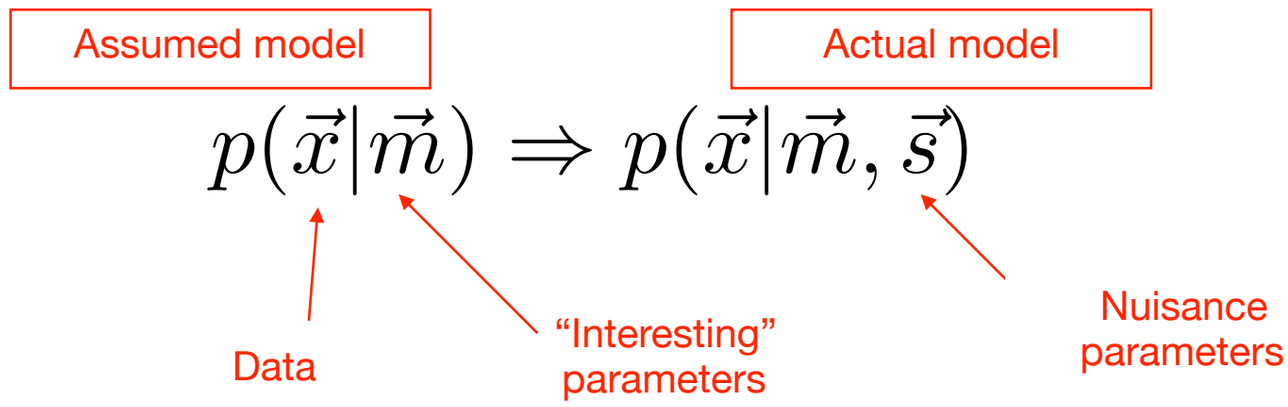
Any statistical inference is based on p(x|m): observe x to extract information about m, assuming to know the distribution p(x|m), that is “the model”.

The systematic uncertainty is that component of the uncertainty that is due to the imperfect knowledge about the shape of the probability distribution p(x;m).

G. Punzi

Nuisance parameters

Assume model p(x|m), which in general differs from the actual model. Difference is parametrized by introducing an additional dependence on unknown nuisance parameters. Parameters that are not interesting for the measurement at hand but do influence its outcome.



The width of p(x|m) connects with the statistical uncertainty. The shape, which depends on nuisance parameters s, with the systematic uncertainty.

Not only we don't know exactly what value of x would be observed if m had some definite value; we don't even know exactly how probable each possible value of x is. Cannot define standard deviation for s; would imply knowing the distribution p(s). But then s wouldn't be any longer a nuisance and would get embedded in the model! Can only estimate an allowed range for s, and ensure that any result of the inference hold for any s in that range.

Bayesian approach

For Bayesians, s is just another parameter. Assume an a priori distribution for s that allows “integrating it out” through marginalization and use the result p(x|m) as model for any subsequent (Bayesian) inference.

$$p(\vec{x};\vec{m}) = \int p(\vec{x};\vec{m},\vec{s})p(\vec{s})d\vec{s}$$

- A significant dependence of results on the chosen prior p(s) may occur

- Results from multiple measurements that are based on independent data but share nuisance parameters may get correlated (through common priors)

Avoid mixing frequentist and Bayesian approaches. E.g., don’t use marginalized p(x;m) to get frequentist confidence intervals. Hybrid results are harder to interpret. If the distribution of parameter s is assumed known, we are in Bayesian realm, hence rather assume known the distributions of *all* parameters

Incorporating systematic uncertainties in p-values

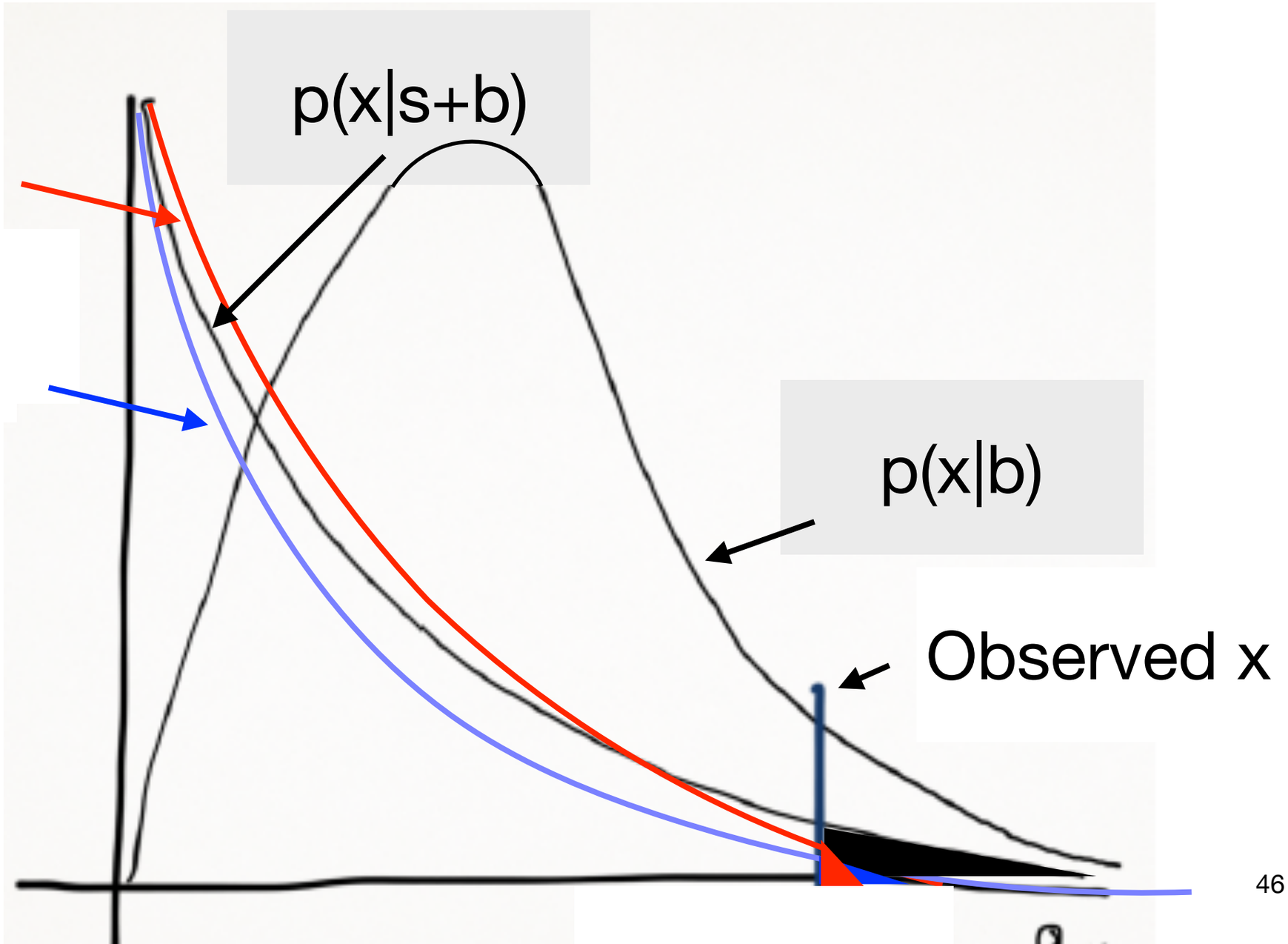
In searches, typically the uncertainty is dominated by the statistical component associated with the small size of the event sample and/or the small signal-to-background ratio.

However, systematic uncertainties are there as well, and incorporating them into the p-value evaluation is needed, especially if there is the potential for a discovery (“extraordinary claims require extraordinary evidence”)

How one does incorporate systematic uncertainties into p-values?

The problem

I don’t know which of the three curves approximates better the real s+b distribution and the p-value depends on which curve I use.
Should I use this  ? Or this  ? Or this  ?



NB: for simplicity assuming nuisance on p(x|s+b) only. In most cases p(x|b) is also affected

Options

Supremum p-value: calculate the p-value for any allowed value of nuisance parameters and quote the least significant p-value (black curve in our case).
Pros: gets frequentist coverage whatever the value of nuisance parameters.
Cons: if the space of nuisance parameters is multidimensional, lots of CPU needed because need to construct many predicted distributions of the test statistics p(x|s+b) and p(x|b), one for each choice of nuisance parameters. Also, can "spoil" the sensitivity of the measurement as implausible choices of nuisance parameters could make p(x|s+b) very close to p(x|b).

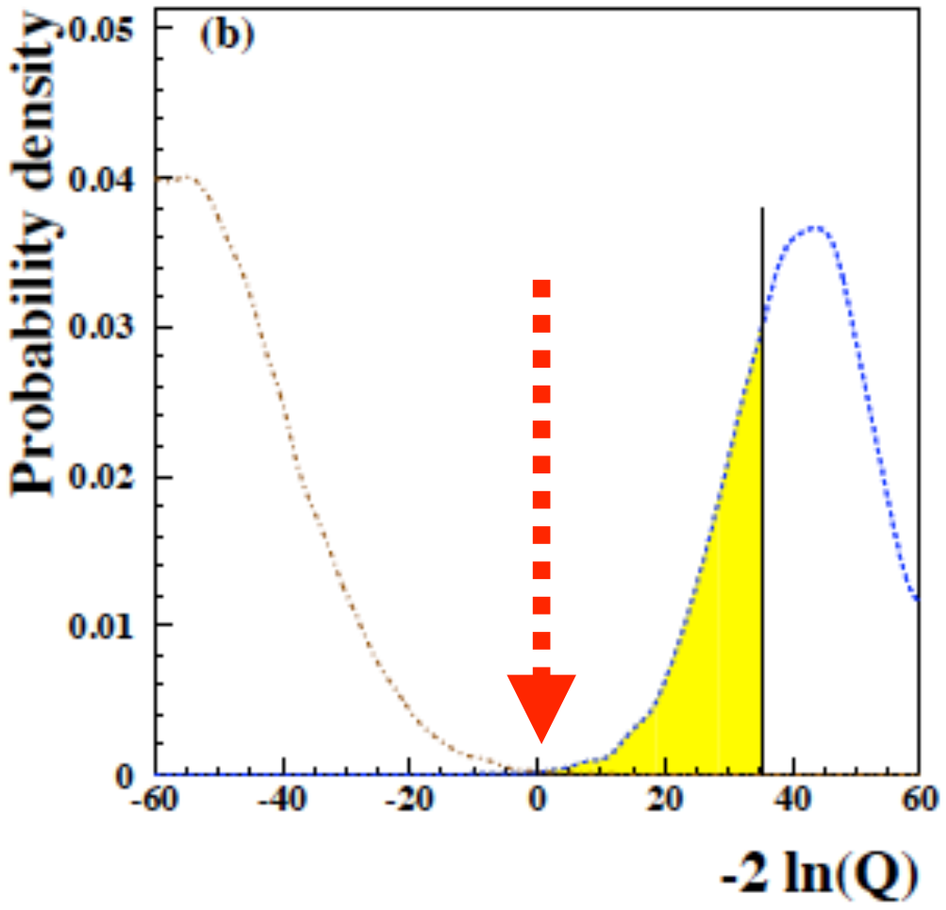
Supremum p-value with Berger-Boos restriction as above, but with nuisance parameters restricted to a subspace based on their determination in data. Use data twice: once to calculate intervals for nuisance parameters, and another to calculate supremum p-values in that interval, then correct for the chance that the nuisance is outside the interval. Pros: mitigates the cons of the pure supremum.
Cons: not obvious if no experimental determination is associated with nuisance parameters.
R. L. Berger and D.D. Boos, J. of the American Stat. Assoc. 89, 427 (1994), 1012

(Most popular LHC) options

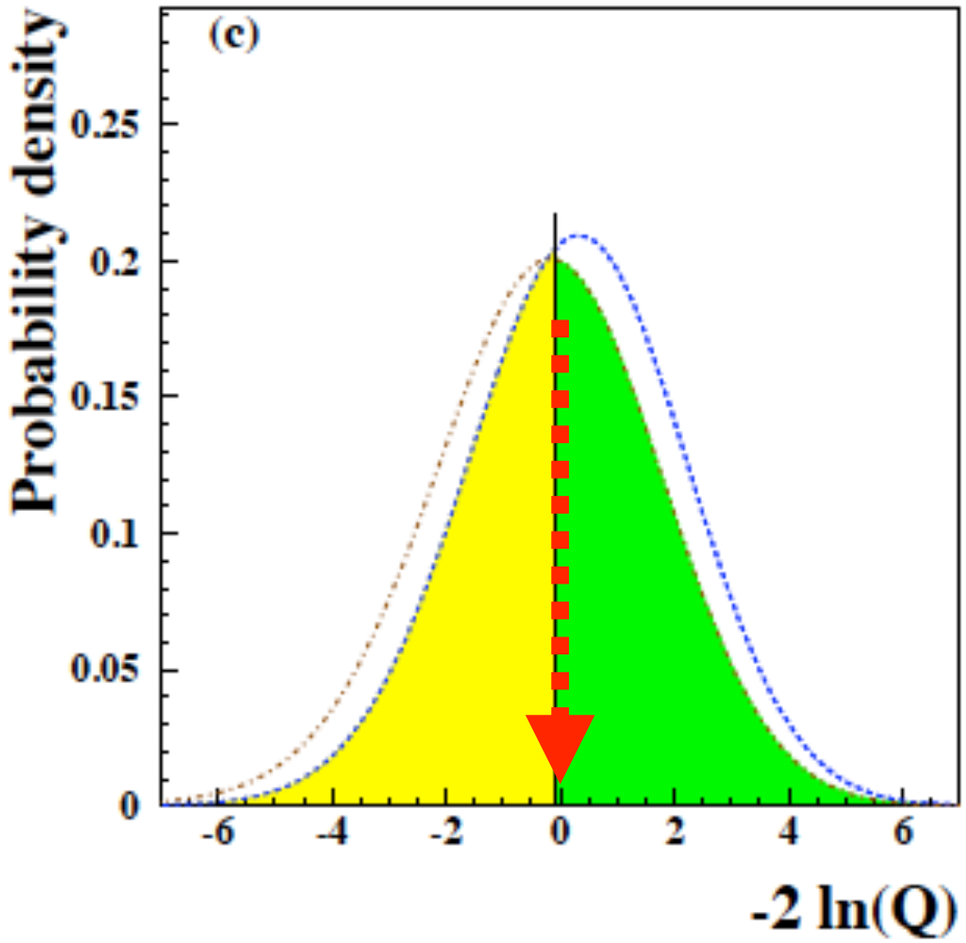
Plugin p-value: determine the central values of nuisance parameters in data (e.g., with a fit) and calculate the p-value for that choice of nuisance parameters.
Pros: computationally fast.
Cons: coverage not guaranteed, subject to the major assumption that true values of nuisance parameters in nature are those determined by the fit.

Cousins-Highland p-value: When constructing the predicted distributions of the test statistics , vary the nuisance parameters according to their prior distributions.
Pros: computationally fast.
Cons: coverage not guaranteed, also admixture of Bayesian and frequentist reasoning which complicates interpretation. R.D. Cousins and V.L. Highland, Nucl. Instrum. Meth., A320, 331 (1992).

Issues with p-values



Possible to get an observation that rejects both the null and the signal hypotheses



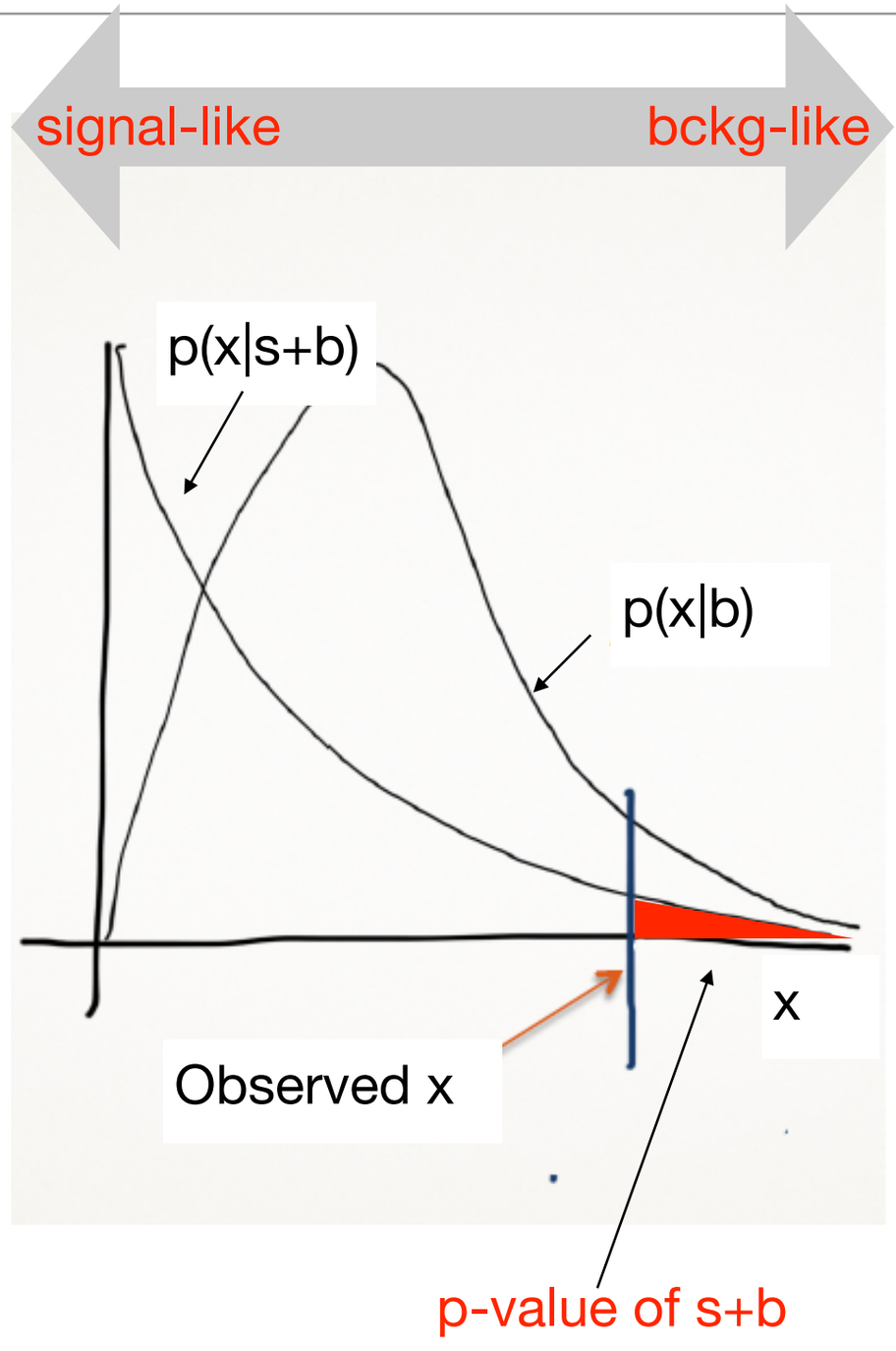
When searching for small signals with poor S-vs-B separation, sensitivity is low, which means that distributions of test statistics are nearly equal. Can make no statement about the signal, regardless of the outcome

The problem of spurious exclusion

Use the likelihood ratio x

Test the hypothesis of the presence of a signal (s+b).

Typically, if p-value of the hypothesis s+b is smaller than 5%, signal gets excluded with 95% CL.



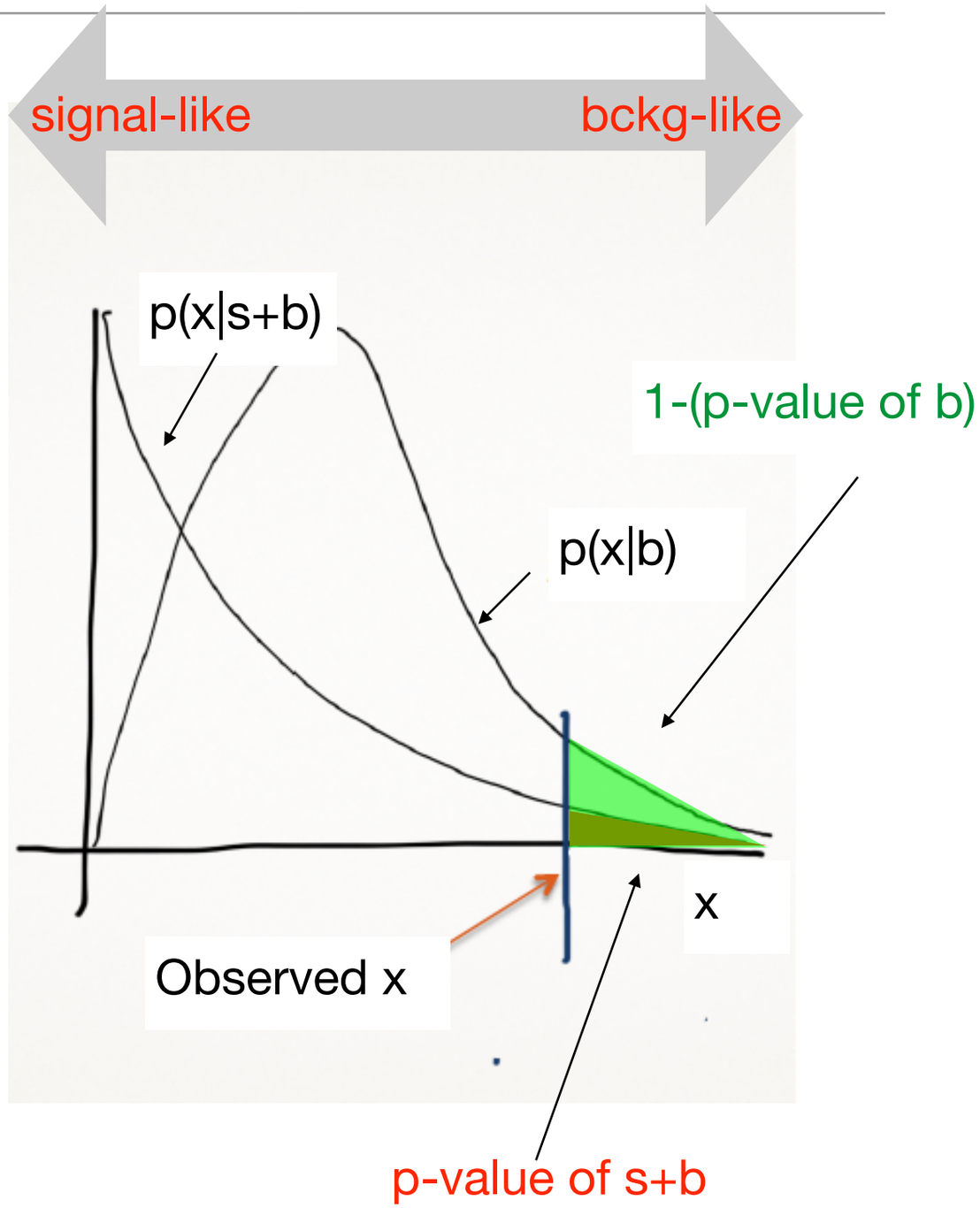
Spurious exclusion

Use the likelihood ratio x

Test the hypothesis of the presence of a signal (s+b).

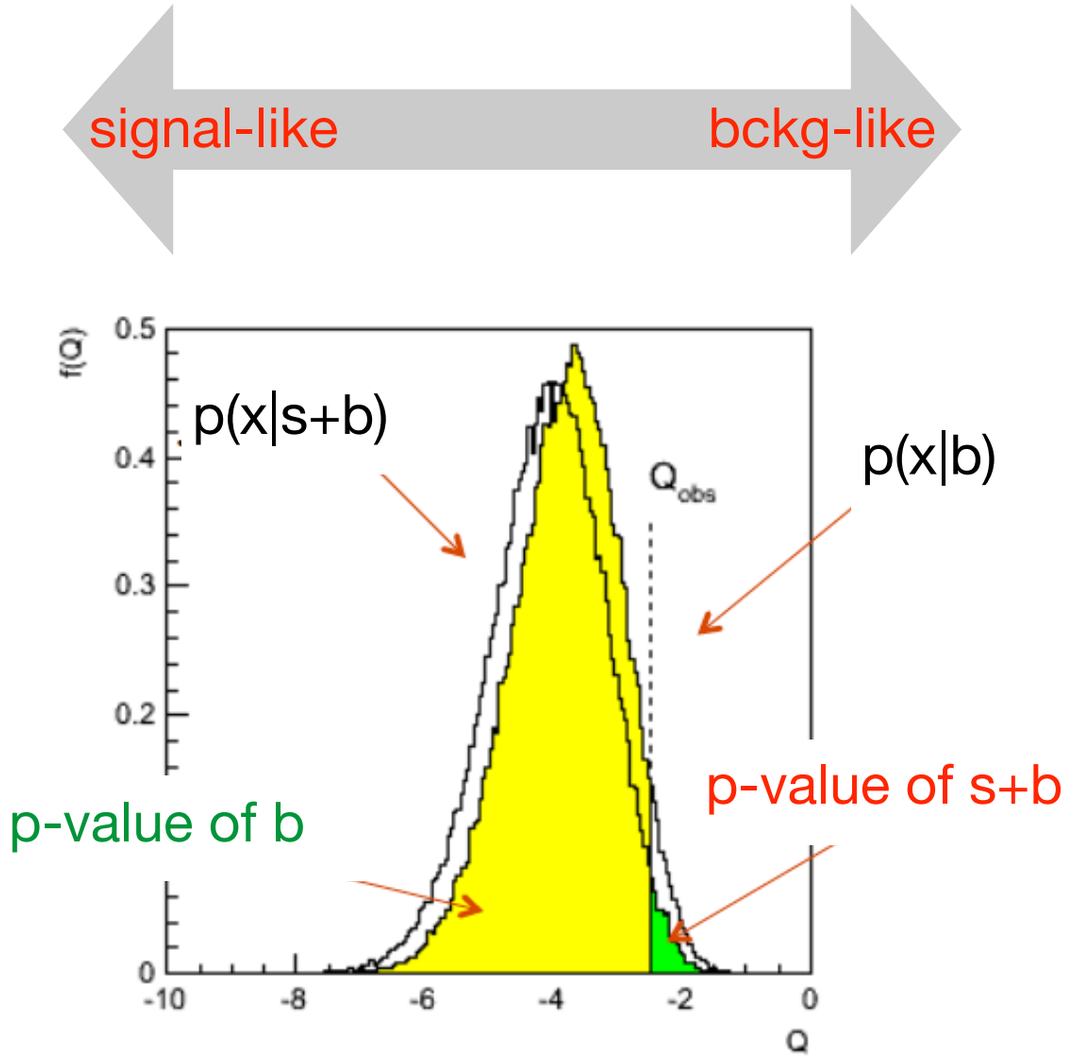
Typically, if p-value of the hypothesis s+b is smaller than 5%, signal gets excluded with 95% CL.

However, when the distributions of the test statistic are similar, 1-pvalue of the background hypothesis is just marginally higher than p-value of s+b.



The CLs method

Modified p-value with no rigorous statistical foundations but “works” fairly well. Allows for treating simultaneously exclusion and discovery and prevents from excluding hypotheses to which there is no sensitivity.



Base test on the pvalue for the s+b hypothesis scaled by (1-pvalue of b). Exclude only if

CLs = [pvalue for s+b] / [1 - pvalue of b]

is small. Denominator increases the CLs thus preventing excluding signals for which there is no sensitivity.

A conditional method inspired by similar methods (Zech, Roe&Woodroofe) developed for counting experiments.

A Poisson example

$$P(n_o \le n_{s+b} \,|\, n_b \le n_o, s+b) = \frac{P(n \le n_o \,|\, s+b)}{P(n \le n_o \,|\, b)}$$

- Suppose $<n_b>=100$
- $s(m_{H1})=30$
- Suppose $n_{obs}=102$
- $s+b=130$
- $\text{Prob}(n_{obs} \le 102 \,|\, 130) < 5\%$, m_{H1} is excluded at >95% CL

- Now suppose $s(m_{H2})=1$, can we exclude m_{H2}?
- If $n_{obs}=102$, obviously we cannot exclude m_{H2}
- Now suppose $n_{obs}=80$, $\text{prob}(n_{obs} \le 80 \,|\, 101) < 5\%$, we looks like we can exclude m_{H2}… but this is dangerous, because what we exclude is $(s(m_{H2})+b)$ and not s……
- With this logic we could also exclude b (expected b=100)
- To protect we calculate a modofoed p-value $\dfrac{\Pr ob(\text{nobs} \le 80 \,|\, 101)}{\Pr ob(\text{nobs} \le 80 \,|\, 100)} \sim 1$
- We cannot exclude m_{H2}

CLs references and code

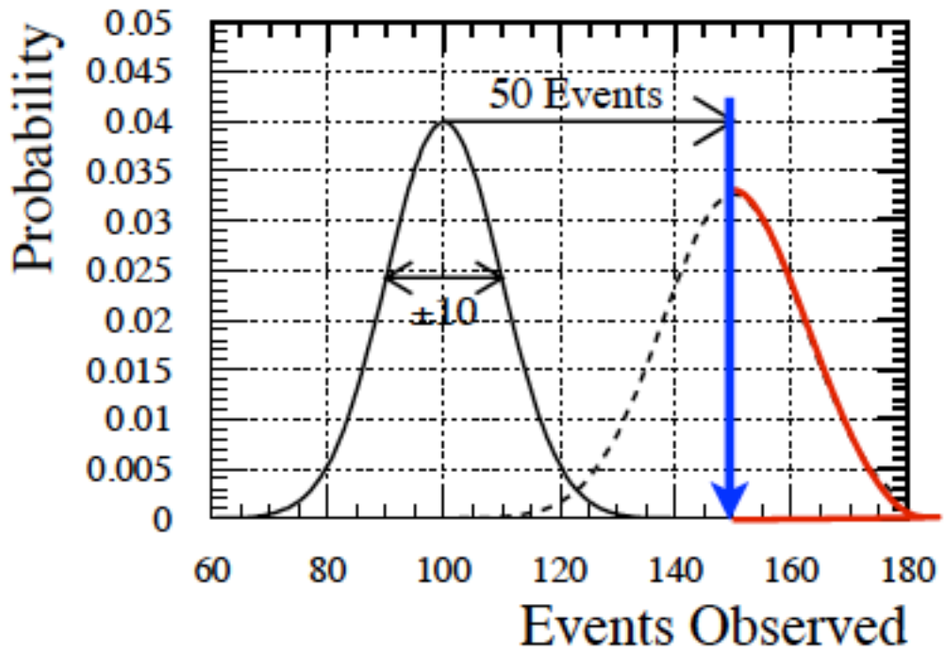
Popular references for CLs are A.L. Read, J. Phys. G Nucl. Part. Phys. 28 (2002), 2693 and T. R. Junk, Nucl. Instr. and Methods in Phys. Res. A 434 (1999), 435

If you ever wanna wet your feet with CLs here is some code and documentation that many turn out to be convenient http://www-cdf.fnal.gov/~trj/mclimit/production/mclimit.html (CLs limits using Bayesian marginalization for the nuisance parameters — more on this later)

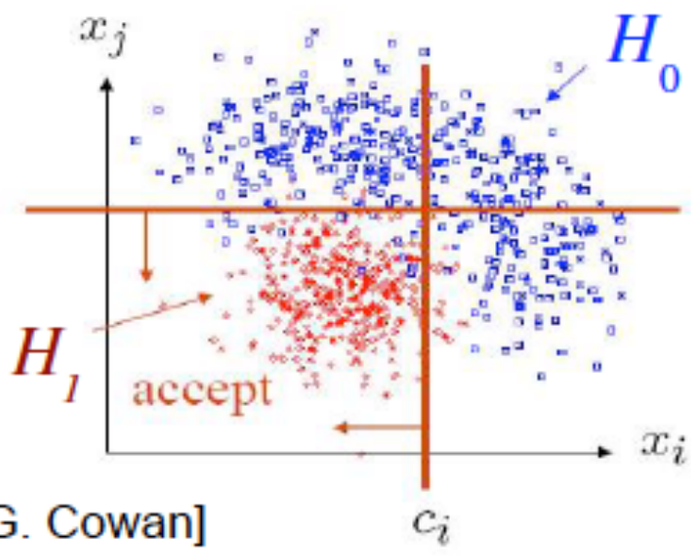
Which function of the observables x to choose?

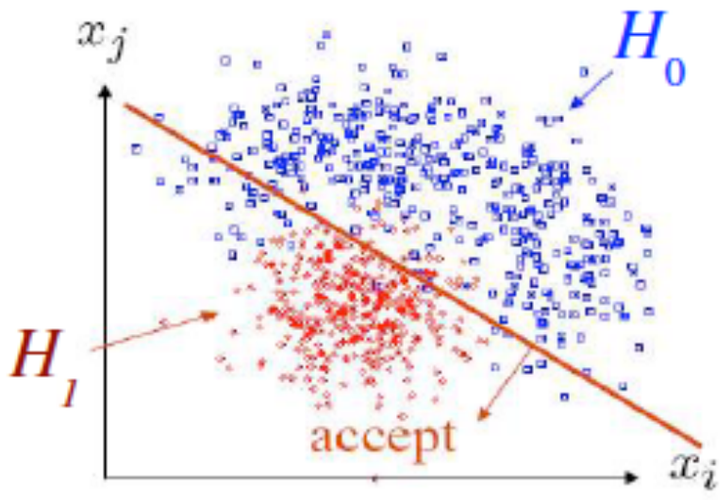
Back to p-values.

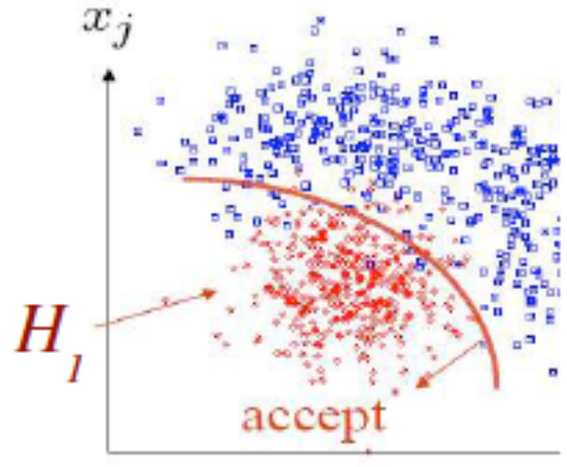
Can we exploit the arbitrariness in choosing the test quantity x? Can we devise a function of the observables x that maximizes the power of my test at fixed false-positive rate.



Pretty obvious in simple counting experiments.
Less obvious in multiple-dimensional nonlinear problems







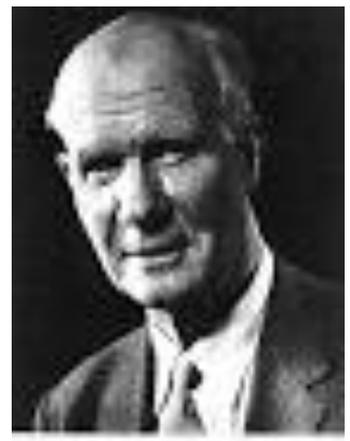
[G. Cowan]

Neyman-Pearson lemma

It does exist an universal statistic for optimal separation between the two hypotheses:

Ratio between the likelihood for the signal+background hypothesis (H1) and the likelihood for the background-only hypothesis (H0)

Jerzy Neyman
(1894-1981)



Egon S. Pearson
(1885-1980)

The region W of acceptance of the null which minimises the probability to accept the null when the signal hypothesis is true is the contour

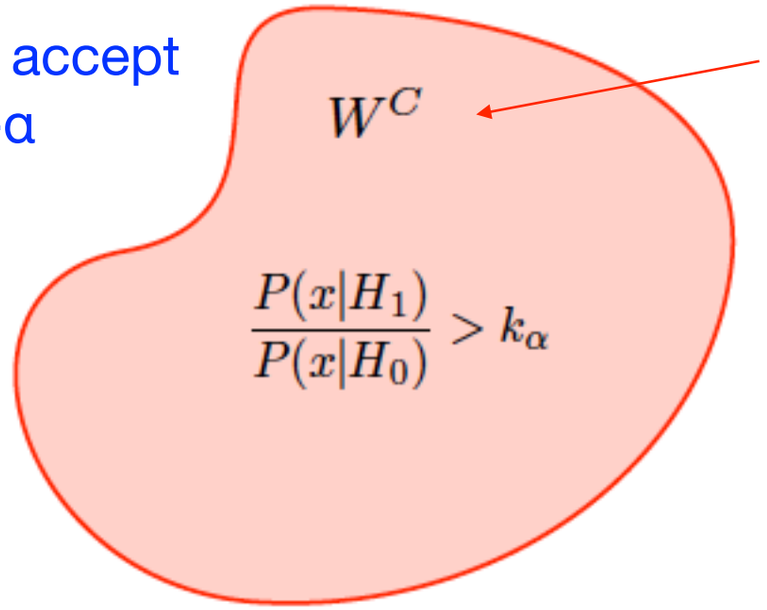
$$\frac{p(x|H_1}{p(x|H_0)} > k_\alpha$$

Any region that has the same false-positive rate would have higher rate of false negatives (technically, less power)

NP-lemma illustrated proof

Take a contour of the likelihood ratio that has a given rate α of false positives, that is a given probability under H0

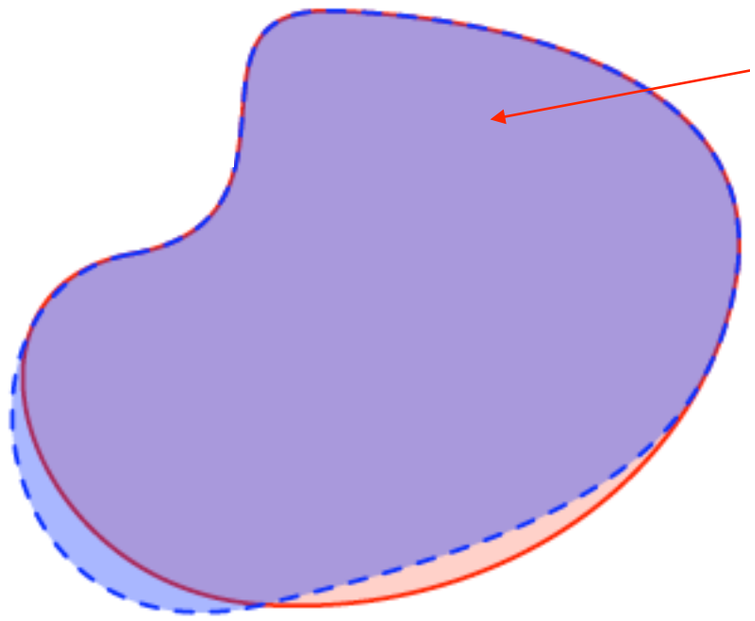
Region W: if data fall here we accept H0; probability under H0 is 1-α



Region W^c: if data fall there we reject H0; probability under H0 is α

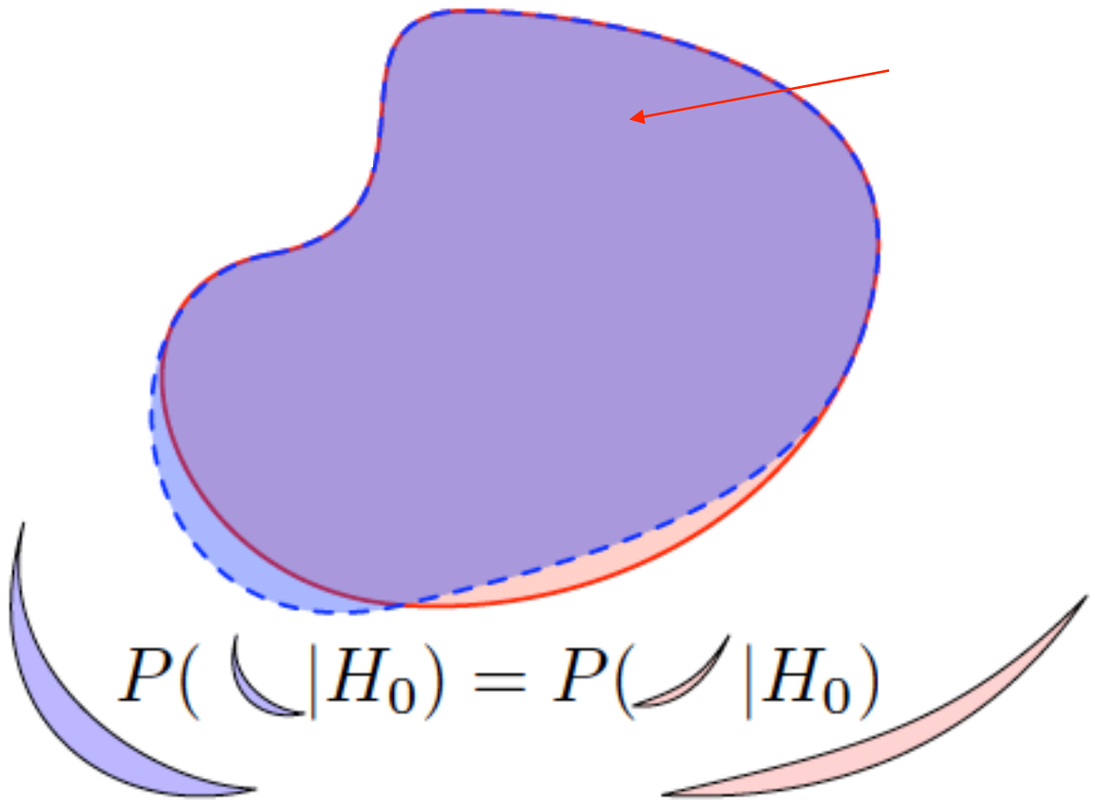
NP-lemma illustration

Take a variation that has the same rate α of false positives (same probability under H0)



NP-lemma illustration

Take a variation that has the same rate α of false positives (same probability under H0)

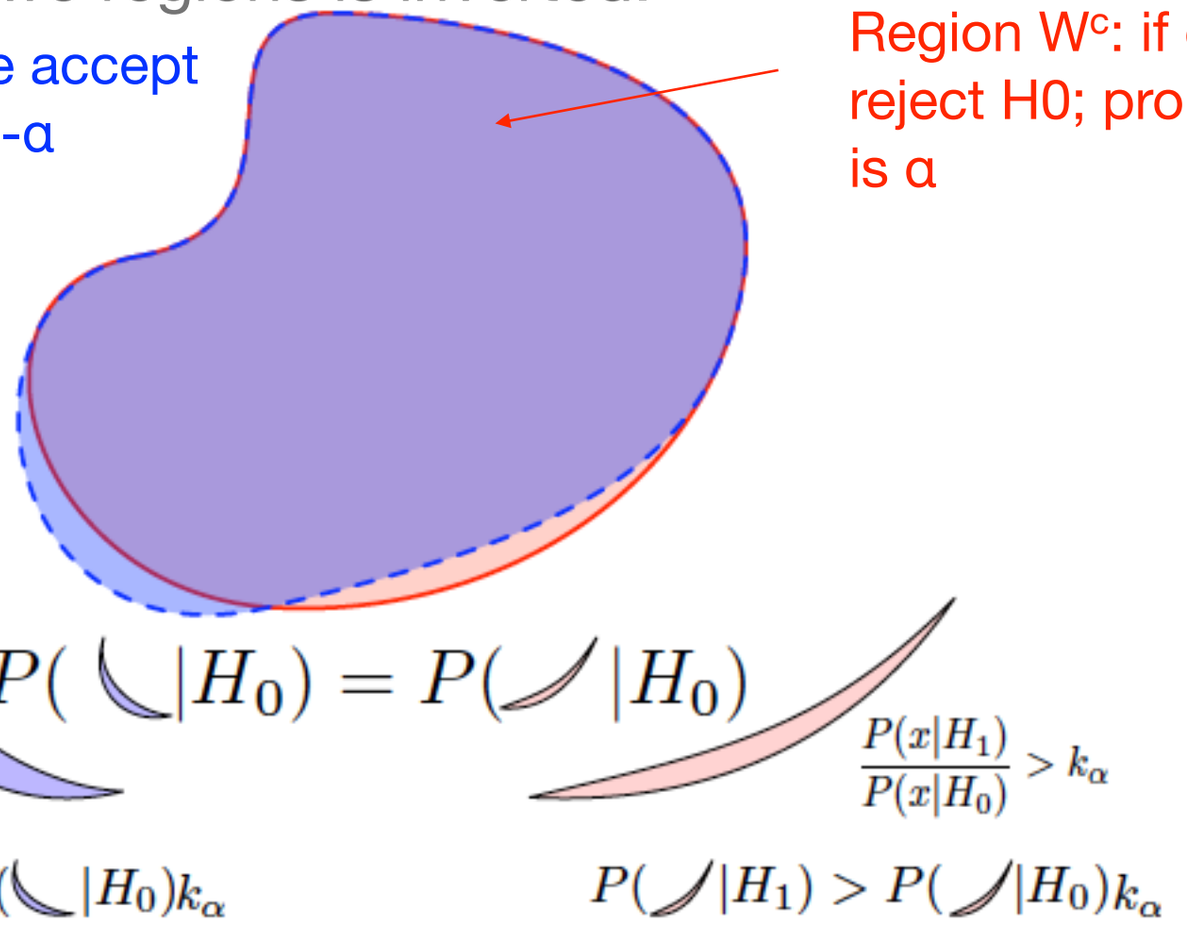


NP-lemma illustration

Because the region gained with the new contour was outside of the likelihood ratio contour and the region lost lost was inside it, the hierarchy between probabilities under H0 and H1 in the two regions is inverted.

Region W: if data fall here we accept H0; probability under H0 is 1-α

Region W^c: if data fall there we reject H0; probability under H0 is α



$$P(\ |H_1) < P(\ |H_1)$$

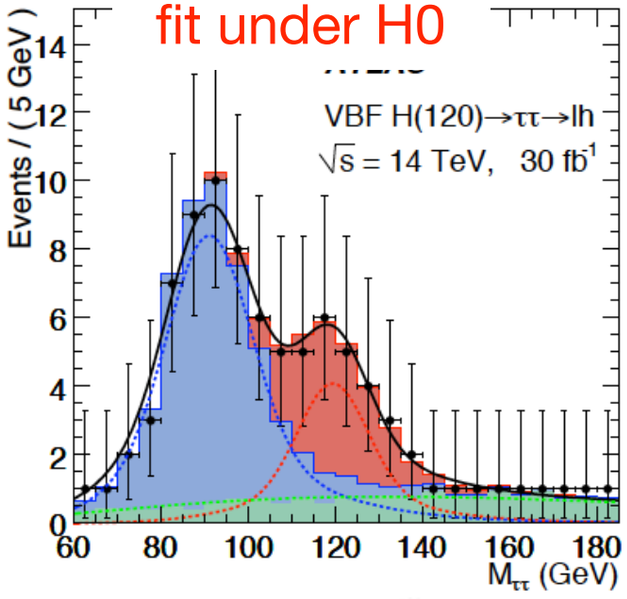
The new region region has less power.

Likelihood-ratio is LHC's most popular test statistic

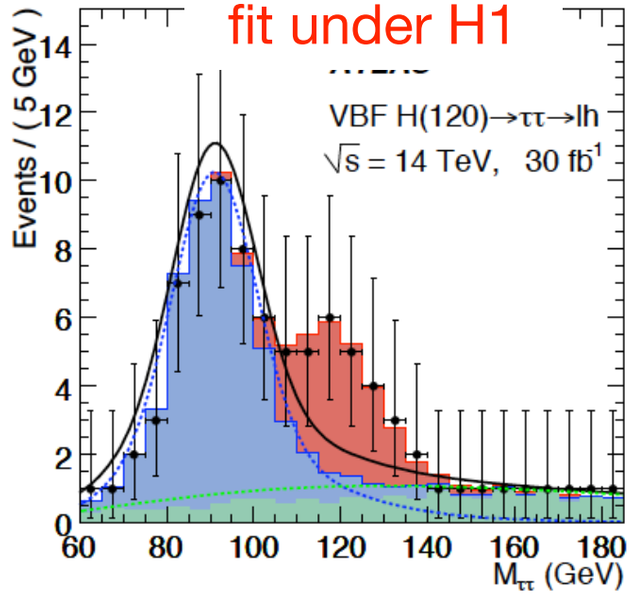
LR is convenient because (1) has optimal performance and (2) allows for testing with no need to laboriously construct distributions by generating and fitting pseudodata since its large-sample distribution is known (χ^2)

1. Fit data under H0: i.e. with a likelihood that only has "background" parameters.
2. Fit data under H1: i.e. with a likelihood that includes n additional "signal" free parameters
3. The ratio between the resulting values of the likelihood functions at their maxima is distributed as a χ^2 with n degrees of freedom.
4. Comparison of the ratio obtained in data with the relevant χ^2 distribution allows for testing H1 vs H0.

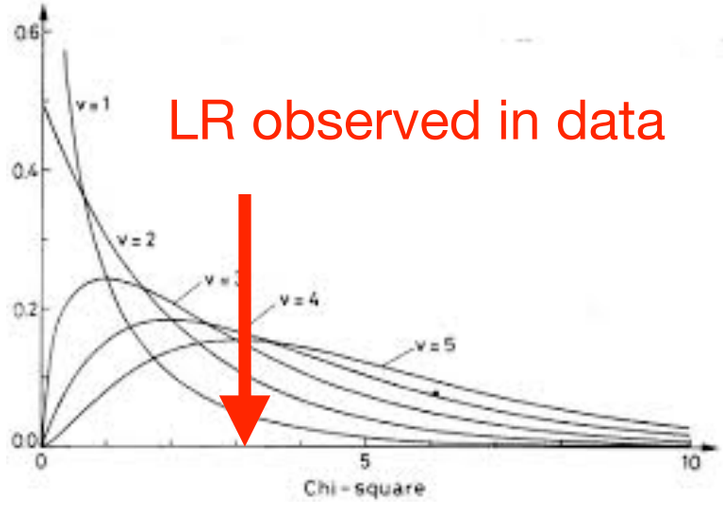
fit under H0



fit under H1

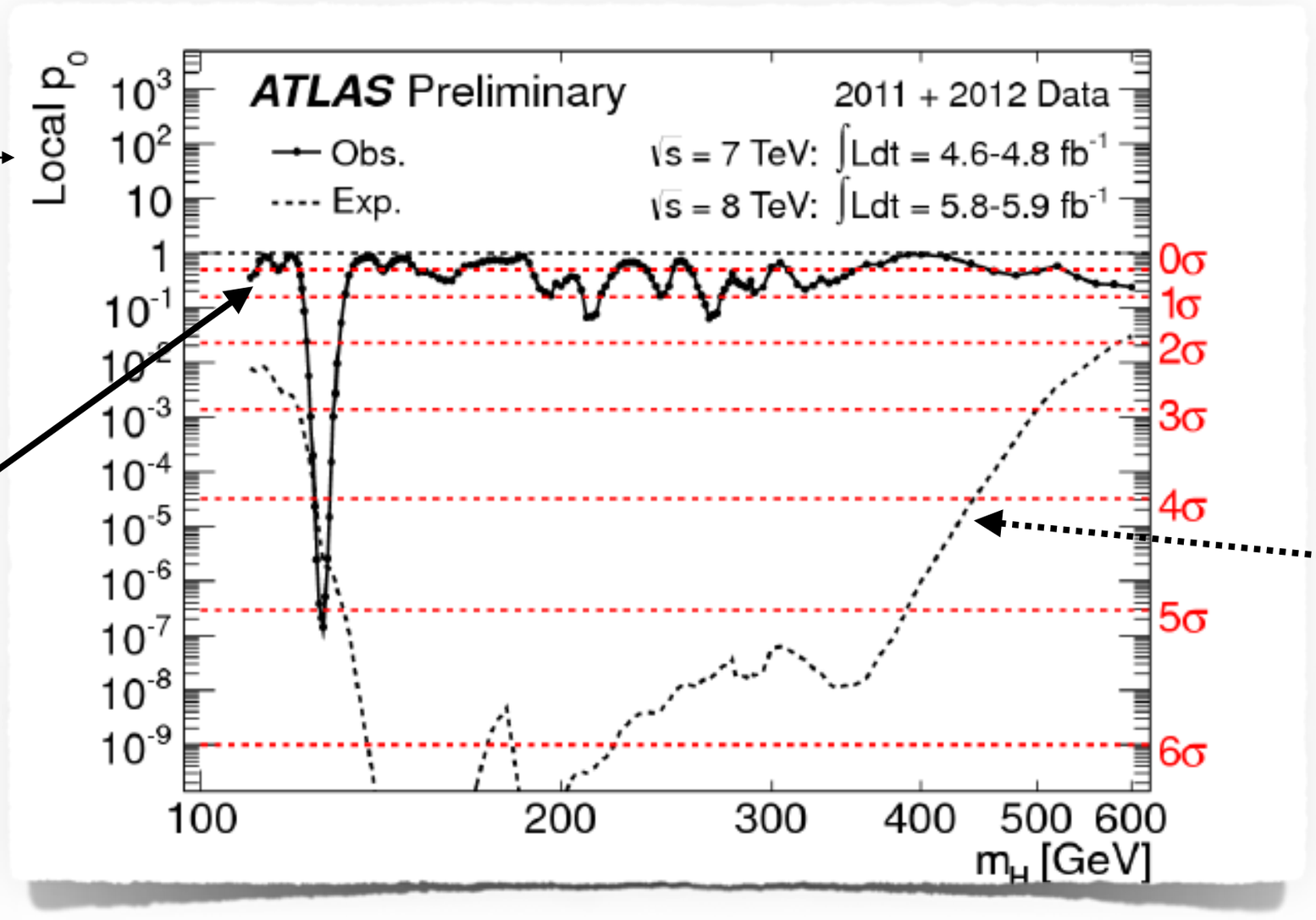


χ^2



So, now you should be able to understand this

Local p-value evaluated at various values of possible Higgs mass

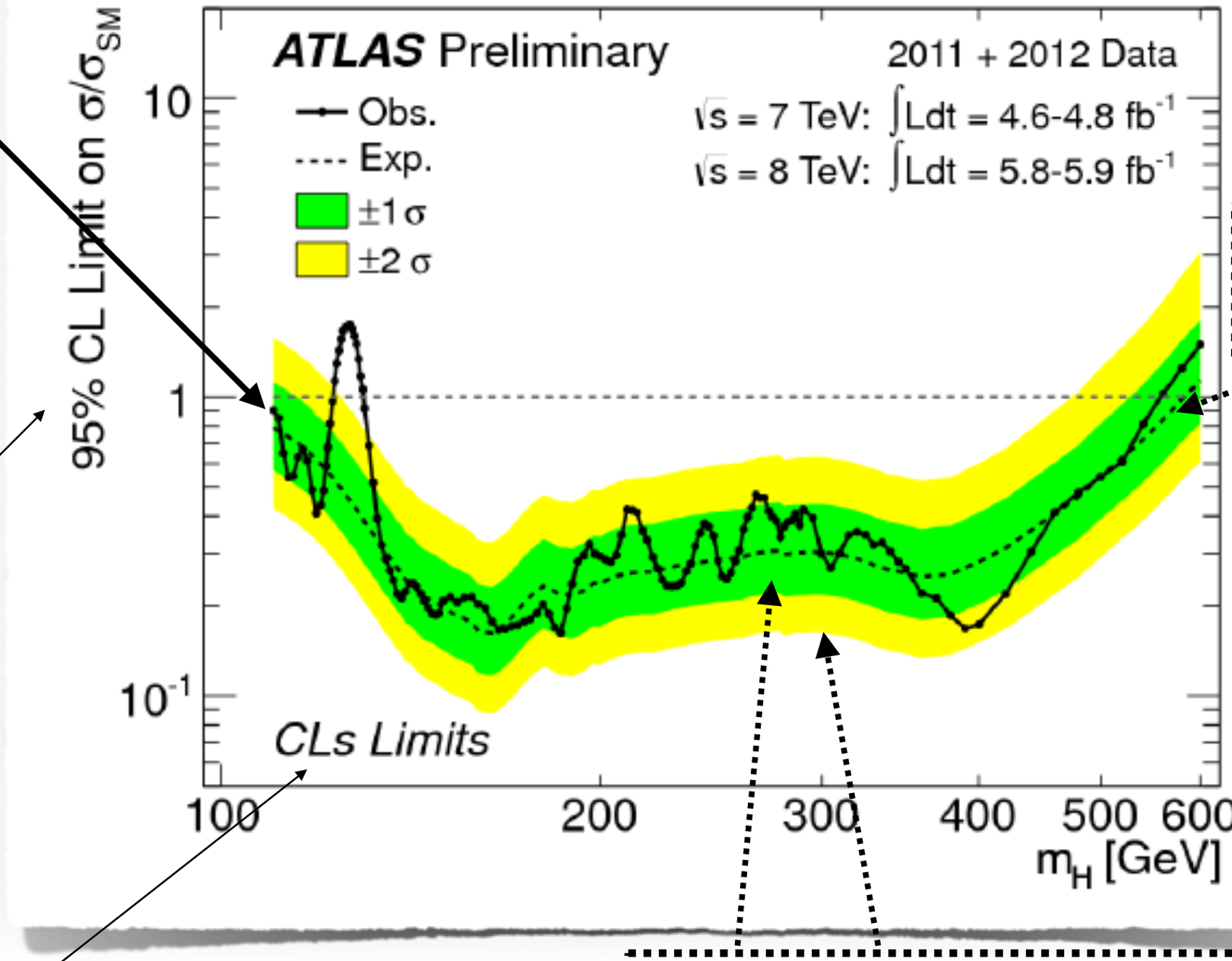


Observed local p-value for the background-only hypothesis

Median expected p-value for the signal +background hypothesis

..and this

Observed limit



Median expected limits in the absence of signal

Exclusion limit for the Higgs signal strenght (cross-section/SM cross section) as a function of Higgs mass

These limits are based on CLs

68.3% and 95.5% of the expected limits in the absence of signal

Grazie

a voi per la vostra attenzione ed al Prof. E. Milotti per questa opportunita'.