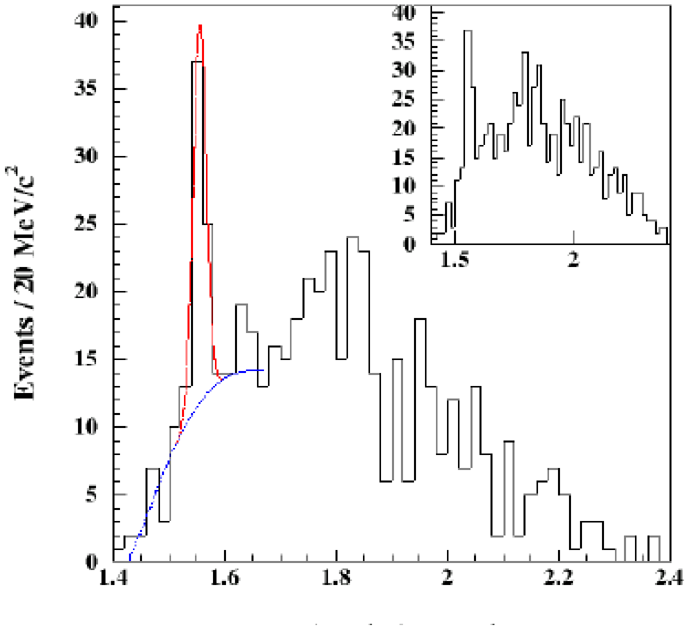
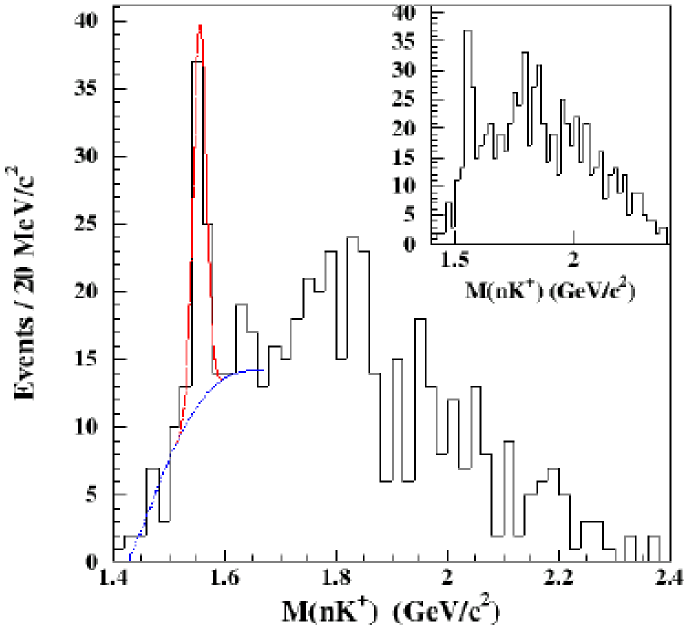
Is there evidence for a peak in this data?



Is there evidence for a peak in this data?



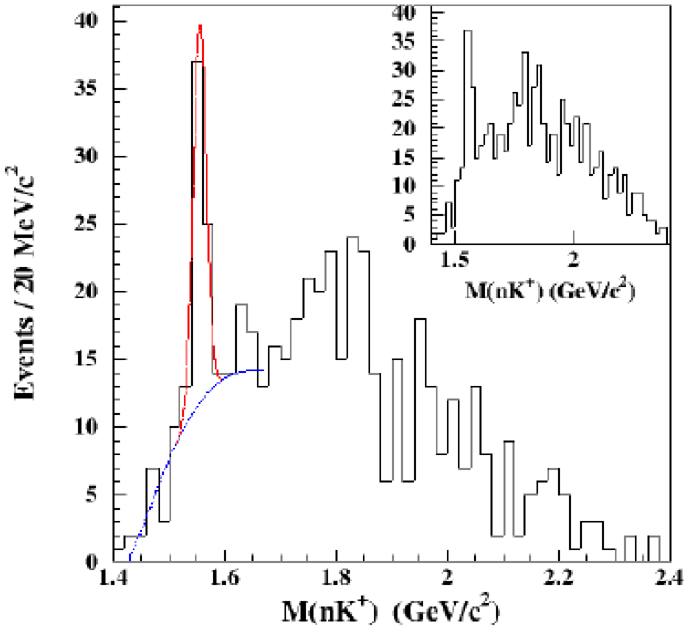
“Observation of an Exotic S=+1

Baryon in Exclusive Photoproduction from the Deuteron”

S. Stepanyan et al, CLAS Collab, Phys.Rev.Lett. 91 (2003) 252001

“The statistical significance of the peak is 5.2 ± 0.6 σ”

Is there evidence for a peak in this data?



“Observation of an Exotic S=+1 Baryon in Exclusive Photoproduction from the Deuteron”

S. Stepanyan et al, CLAS Collab, Phys.Rev.Lett. 91 (2003) 252001

“The statistical significance of the peak is 5.2 ± 0.6 σ”

“A Bayesian analysis of pentaquark signals from CLAS data”

D. G. Ireland et al, CLAS Collab, Phys. Rev. Lett. 100, 052001 (2008)

“The ln(RE) value for g2a (-0.408) indicates weak evidence in favour of the data model without a peak in the spectrum.”

Comment on “Bayesian Analysis of Pentaquark Signals from CLAS Data” Bob Cousins, http://arxiv.org/abs/0807.1330

Statistical issues in searches for New Phenomena:
p-values, Upper Limits and Discovery

Louis Lyons

IC and Oxford

l.lyons@physics.ox.ac.uk

CERN,

July 2014

See ‘Comparing two hypotheses’

http://www.physics.ox.ac.uk/users/lyons/H0H1_A~1.pdf

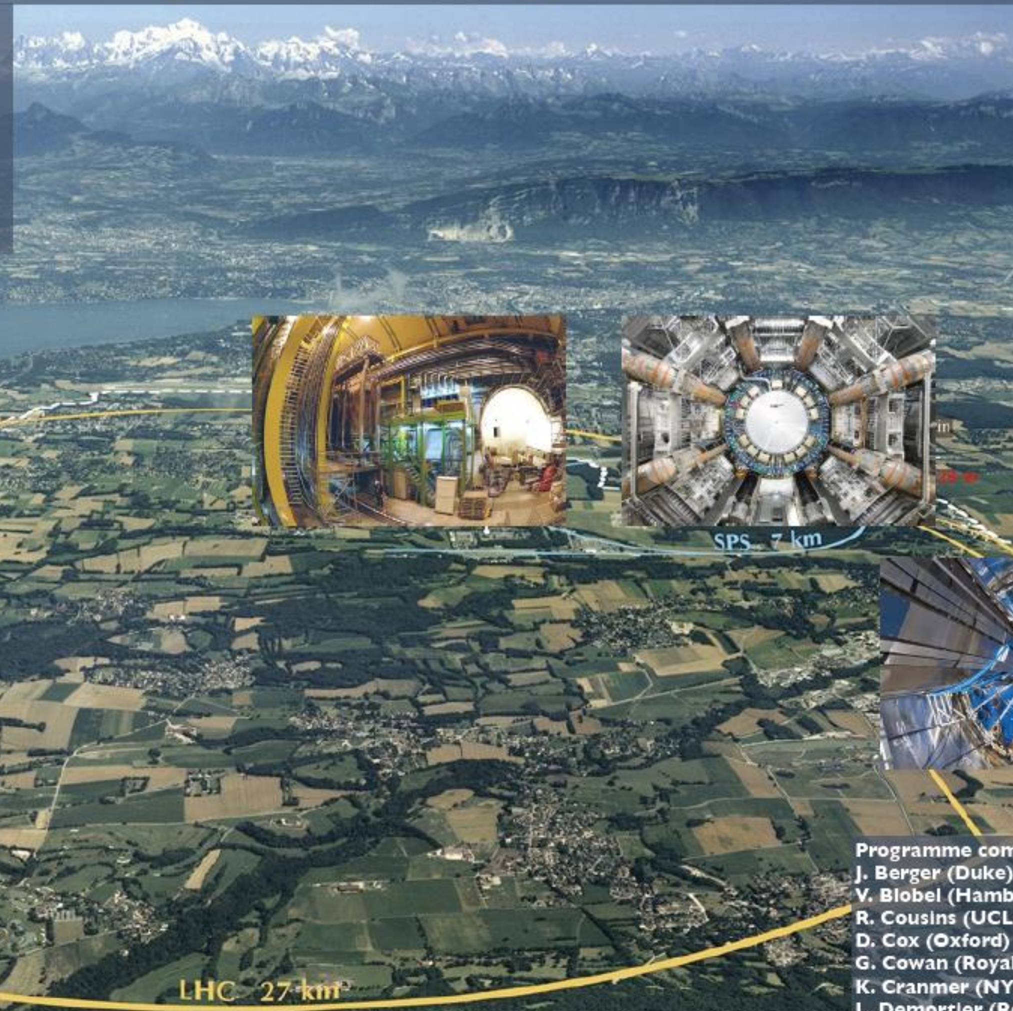
PHYSTAT 2011 Workshop at CERN, Geneva

17-20 January 2011

http://indico.cern.ch/event/phystat2011

Contacts:
Albert de Roeck <albert.de.roeck@cern.ch>
Louis Lyons <l.lyons1@physics.ox.ac.uk>

LPCC
LHC Physics Centre at CERN



Jan 17-19: Statistical Issues for Search Experiments

Statistical issues related to discovery claims in search experiments, with emphasis on those at the LHC.

Jan 20: Unfolding

Unfolding of detector effects from experimental distributions

Programme committee
- J. Berger (Duke)
- V. Blobel (Hamburg)
- R. Cousins (UCLA)
- D. Cox (Oxford)
- G. Cowan (Royal Holloway)
- K. Cranmer (NYU)
- L. Demortier (Rockefeller)
- A. de Roeck (CERN)
- B. Efron (Stanford)
- G. Flucke (DESY)
- E. Gross (Weizmann)
- D. Hand (Imperial College)
- J. Linnemann (MSU)
- R. Lockhart (Simon Fraser)
- L. Lyons (Imperial College)
- M.L. Mangano (CERN)
- S. Schmitt (DESY)
- M. Williams (Imperial College)

TOPICS

Discoveries

H0 or H0 v H1

p-values: For Gaussian, Poisson and multi-variate data

Goodness of Fit tests

Why 5σ?

Blind Analysis

Look Elsewhere Effect

What is p good for?

Errors of 1st and 2nd kind

What a p-value is not

P(theory | data) ≠ P(data | theory)

Setting Limits

Case study: Search for Higgs boson

DISCOVERIES

"Recent" history:

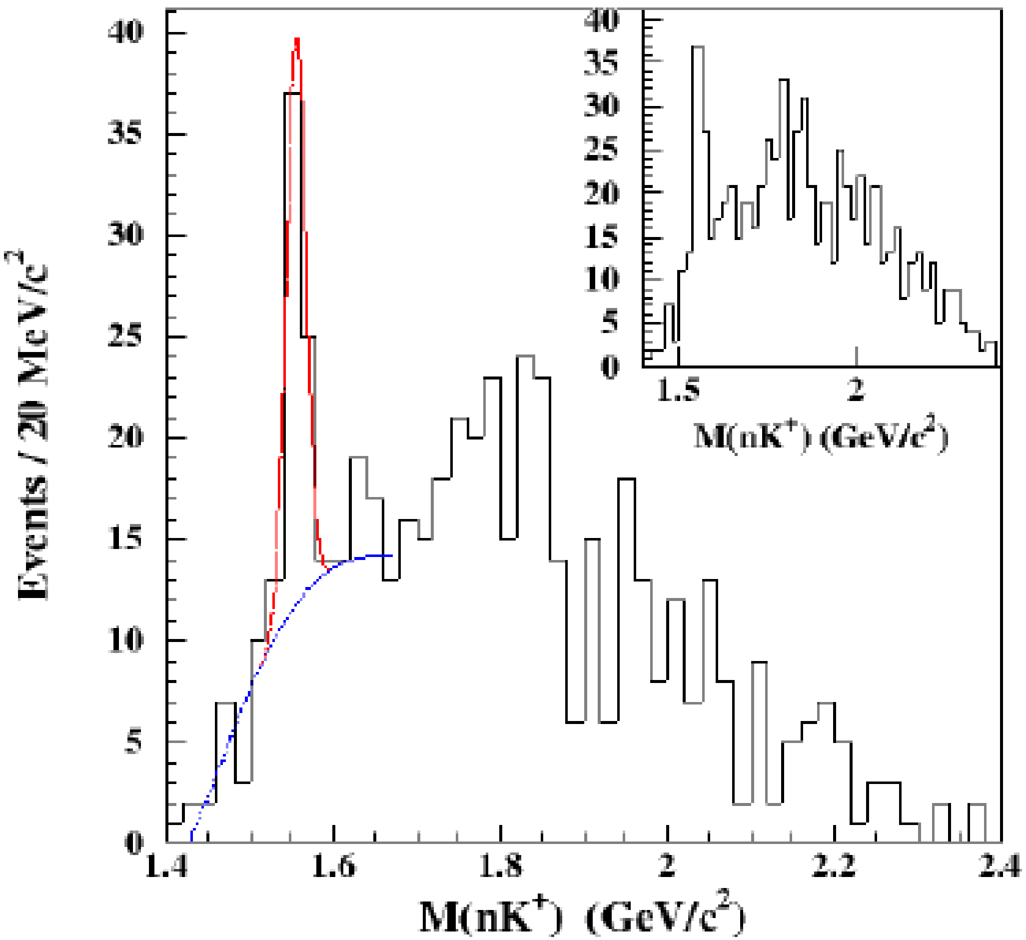
Charm	SLAC, BNL	1974
Tau lepton	SLAC	1977
Bottom	FNAL	1977
W, Z	CERN	1983
Top	FNAL	1995
{Pentaquarks	~Everywhere	2002 }
Higgs	CERN	2012
?	CERN	2015?

? = SUSY, q and l substructure, extra dimensions,
free q/monopoles, technicolour, 4^{th} generation, black holes,…..

QUESTION: How to distinguish discoveries from fluctuations?

Penta-quarks?

Hypothesis testing: New particle or statistical fluctuation?



H0 or H0 versus H1 ?

H0 = null hypothesis

e.g. Standard Model, with nothing new

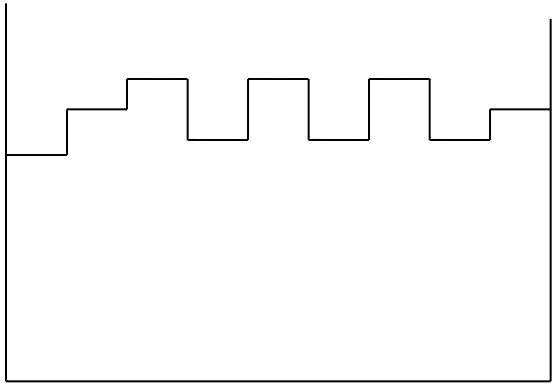
H1 = specific New Physics e.g. Higgs with M_H = 125 GeV

H0: "Goodness of Fit" e.g. χ^2, p-values

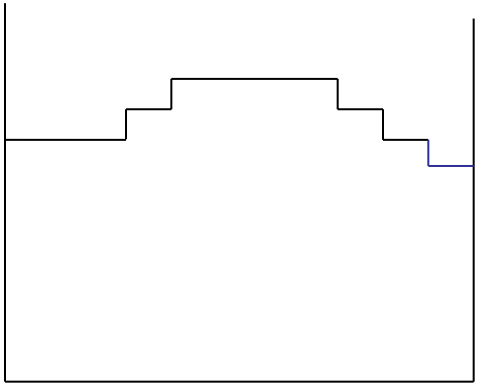
H0 v H1: "Hypothesis Testing" e.g. $\mathcal{L}$-ratio

Measures how much data favours one hypothesis wrt other

H0 v H1 likely to be more sensitive



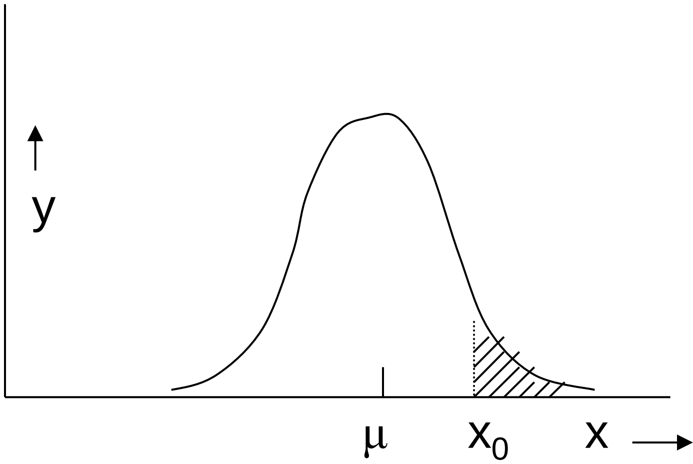
or



p-values

Concept of pdf

Example: Gaussian



y = probability density for measurement x

$y = 1/(\sqrt{(2\pi)}\sigma) \exp\{-0.5*(x-\mu)^2/\sigma^2\}$

p-value: probablity that $x \geq x_0$

Gives probability of "extreme" values of data (in interesting direction)

$(x_0-\mu)/\sigma$	1	2	3	4	5
p	16%	2.3%	0.13%	0. 003%	$0.3*10^{-6}$

i.e. Small p = unexpected

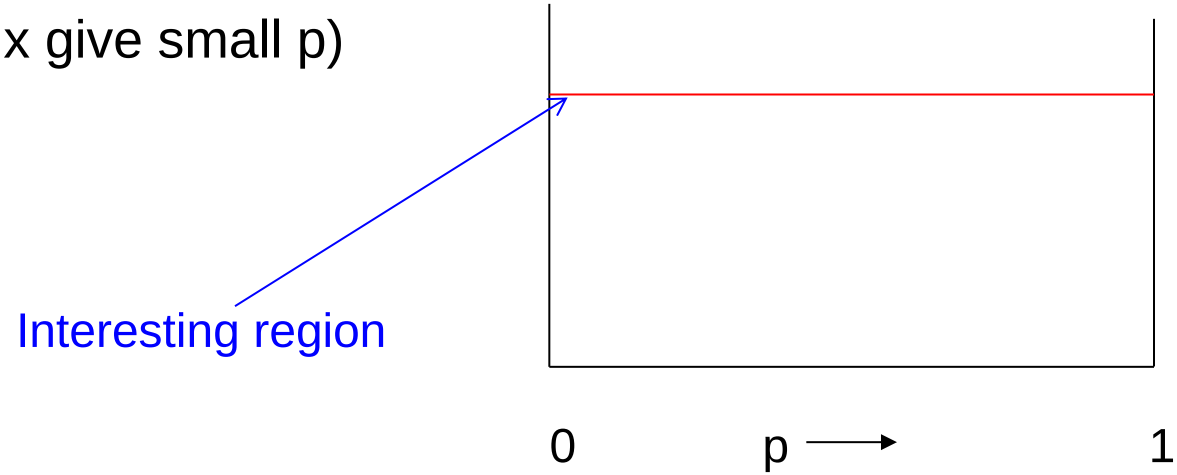
p-values, contd

Assumes:

- Gaussian pdf (no long tails)
- Data is unbiassed
- σ is correct

If so, Gaussian x ⟹ uniform p-distribution

(Events at large x give small p)

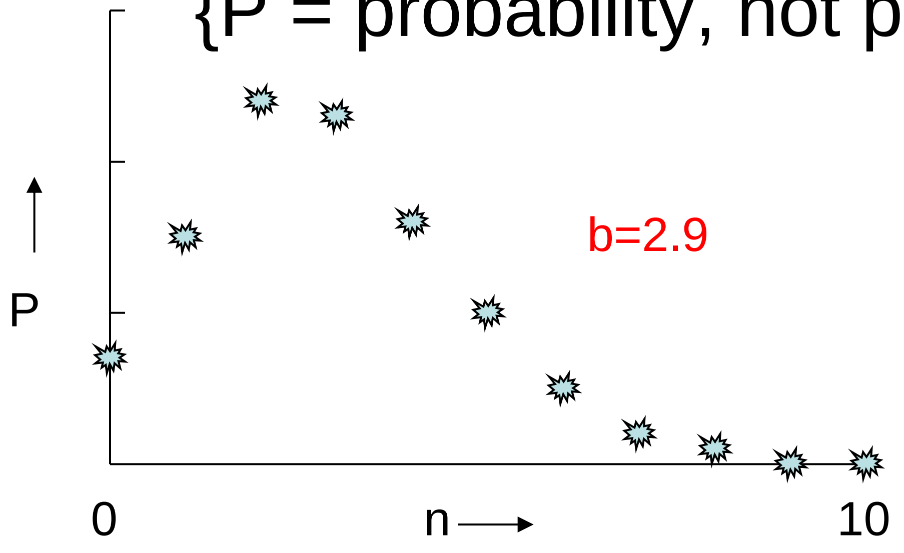


p-values for non-Gaussian distributions

e.g. Poisson counting experiment, bgd = b

$P(n) = e^{-b} * b^n/n!$

{P = probability, not prob density}



For n=7, p = Prob(at least 7 events) = P(7) + P(8) + P(9) +…….. = 0.03

p-values and σ

p-values often converted into equivalent Gaussian σ

e.g. $3*10^{-7}$ is "5σ" (one-sided Gaussian tail)

Does NOT imply that pdf = Gaussian

Significance

Significance = $S/\sqrt{B}$?   (or $S/\sqrt{(S+B)}$, etc)

Potential Problems:

- Uncertainty in B
- Non-Gaussian behaviour of Poisson, especially in tail
- Number of bins in histogram, no. of other histograms [LEE]
- Choice of cuts   (Blind analyses)
- Choice of bins   (……………….)

For future experiments:

- Optimising: Could give S =0.1, B = 10^{-4}, $S/\sqrt{B}$ =10

Look Elsewhere Effect

See 'peak' in bin of histogram

Assuming null hypothesis, p-value is chance of fluctuation at least as significant as observed ……….

1) at the position observed in the data; or
2) anywhere in that histogram; or
3) including other relevant histograms for your analysis; or
4) including other analyses in Collaboration; or
5) In any CERN experiment; or

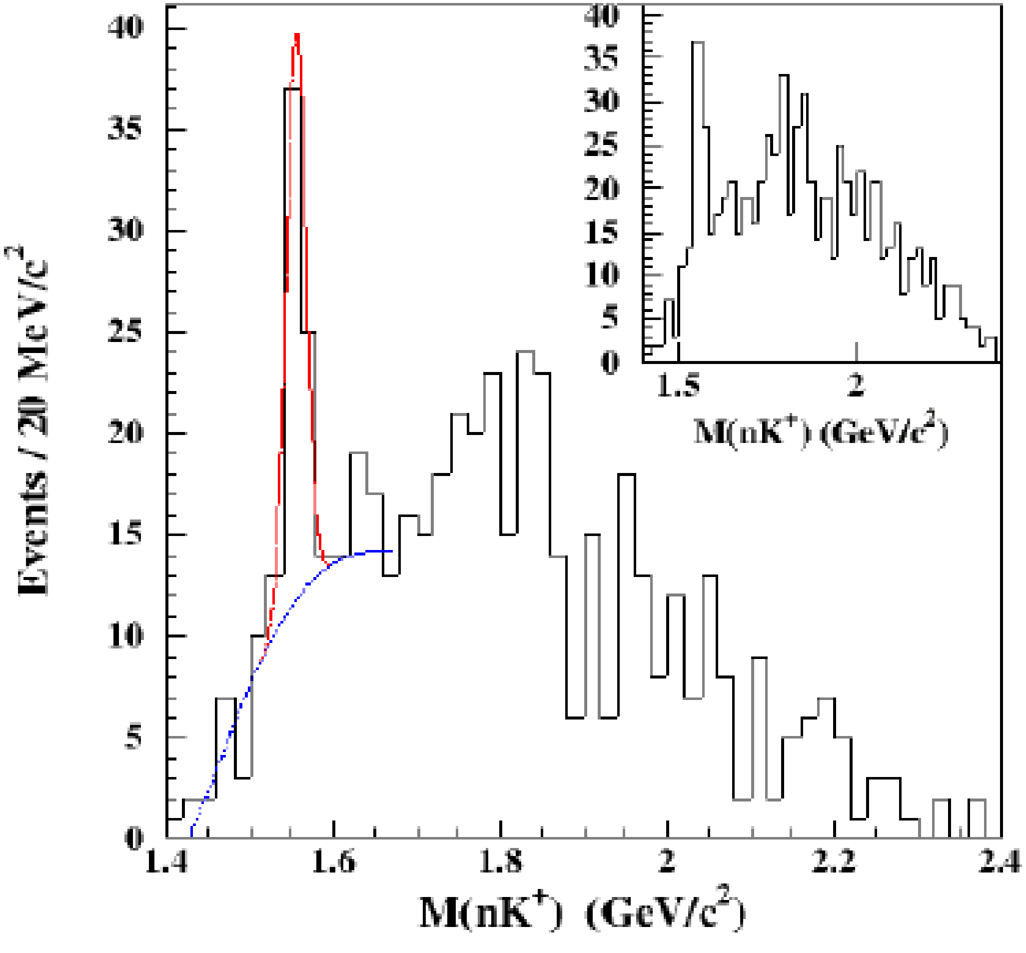
etc.

Contrast local p-value with 'global' p-value

Specify what is your 'global'

Penta-quarks?

Hypothesis testing: New particle or statistical fluctuation?



Goodness of Fit Tests

Data = individual points, histogram, multi-dimensional, multi-channel

χ^2 and number of degrees of freedom

$\Delta\chi^2$ (or *ln*$\mathcal{L}$-ratio): Looking for a peak

Unbinned $\mathcal{L}_{max}$?

Kolmogorov-Smirnov

Zech energy test

Combining p-values

Lots of different methods.

R. B. D'Agostino and M. A. Stephens, 'G of F techniques' (1986, Dekkar)

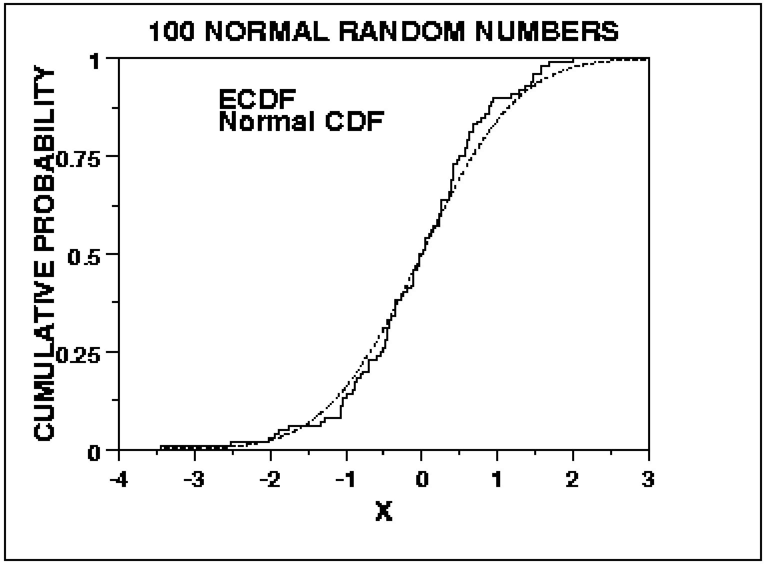
M. Williams, 'How good are your fits? Unbinned multivariate goodness-of-fit tests in high energy physics', http://arxiv.org/abs/1006.3019

Goodness of Fit: Kolmogorov-Smirnov

Compares data and model cumulative plots

Uses largest discrepancy between dists.

Model can be analytic or MC sample



Uses individual data points

Not so sensitive to deviations in tails

(so variants of K-S exist)

Not readily extendible to more dimensions

Distribution-free conversion to p; depends on n

(but not when free parameters involved – needs MC)

Combining different p-values

******* Better to combine data ************

Several results quote independent p-values for same effect:

p_1, p_2, p_3….. e.g. 0.9, 0.001, 0.3 ……..

What is combined significance? A nswer not unique

Not just $p_{1*}p_{2*}p_3$…..

(If 10 expts each have p ~ 0.5, product ~ 0.001 and is clearly **NOT** correct combined p)

$$S = z_* \sum_{j=0}^{n-1} (-\ln z)^j / j! \quad , \qquad z = p_1 p_2 p_3 \ldots\ldots$$

(e.g. For 2 measurements, $S = z_* (1 - \ln z) \geq z$)

Slight problem: Formula is not associative

Combining {{p_1 and p_2}, and then p_3} gives different answer

from {{p_3 and p_2}, and then p_1} , or all together

Due to different options for "more extreme than x_1, x_2, x_3".

Combining different p-values

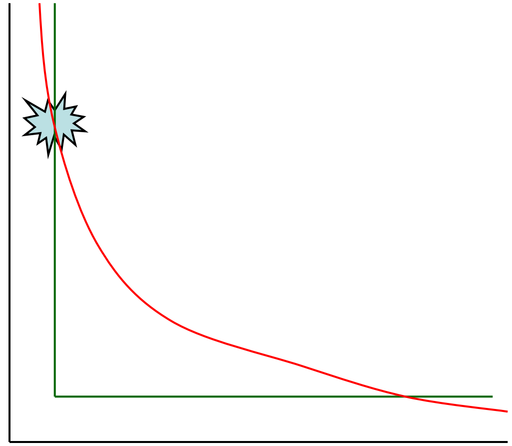
Conventional:

Are set of p-values consistent with H0?

SLEUTH:

How significant is smallest p?

$$1\text{-}S = (1-p_{smallest})^n$$



	$p_1 = 0.01$		$p_1 = 10^{-4}$	
	$p_2 = 0.01$	$p_2 = 1$	$p_2 = 10^{-4}$	$p_2 = 1$
Combined S				
Conventional	$1.0\ 10^{-3}$	$5.6\ 10^{-2}$	$1.9\ 10^{-7}$	$1.0\ 10^{-3}$
SLEUTH	$2.0\ 10^{-2}$	$2.0\ 10^{-2}$	$2.0\ 10^{-4}$	$2.0\ 10^{-4}$

******* N.B. Problem does not have a unique answer ***********

Why 5σ?

- Past experience with 3σ, 4σ,… signals
- Look elsewhere effect:
  - Different cuts to produce data
  - Different bins (and binning) of this histogram
  - Different distributions Collaboration did/could look at
  - Defined in SLEUTH
- Worries about systematics
- Bayesian priors:

$$\frac{P(H0|data)}{P(H1|data)} = \frac{P(data|H0) * P(H0)}{P(data|H1) * P(H1)}$$

Bayes posteriors — Likelihoods — Priors

Prior for {H0 = S.M.} >>> Prior for {H1 = New Physics}

Why 5σ?

BEWARE of tails,

especially for nuisance parameters

Same criterion for all searches?

- Single top production
- Higgs
- Highly speculative particle
- Energy non-conservation

How many σ's for discovery?

SEARCH	SURPRISE	IMPACT	LEE	SYSTEMATICS	No. σ
Higgs search	Medium	Very high	M	Medium	5
Single top	No	Low	No	No	3
SUSY	Yes	Very high	Very large	Yes	7
B_s oscillations	Medium/Low	Medium	Δm	No	4
Neutrino osc	Medium	High	$\sin^2 2\vartheta$, Δm^2	No	4
$B_s \rightarrow \mu\mu$	No	Low/Medium	No	Medium	3
Pentaquark	Yes	High/V. high	M, decay mode	Medium	7
$(g-2)_\mu$ anom	Yes	High	No	Yes	4
H spin ≠ 0	Yes	High	No	Medium	5
4^{th} gen q, l, ν	Yes	High	M, mode	No	6
Dark energy	Yes	Very high	Strength	Yes	5
Grav Waves	No	High	Enormous	Yes	8

Suggestions to provoke discussion, rather than `delivered on Mt. Sinai'

Bob Cousins: "2 independent expts each with 3.5σ better than one expt with 5σ"

What is p good for?

Used to test whether data is consistent with H0

Reject H0 if p is small : $p \leq \alpha$ (How small?)

Sometimes make wrong decision:

Reject H0 when H0 is true: Error of 1st kind

Should happen at rate α

OR

Fail to reject H0 when something else (H1,H2,…) is true: Error of 2nd kind

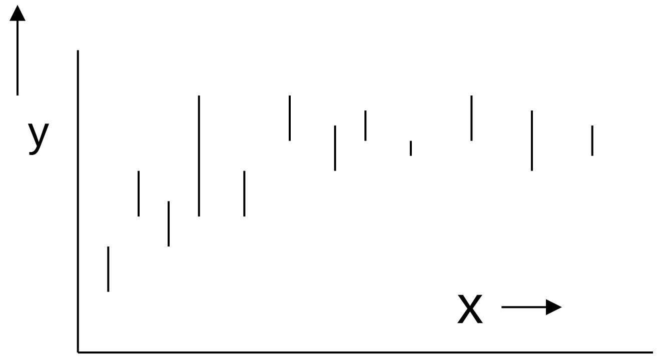
Rate at which this happens depends on……….

Errors of 2nd kind: How often?

e.g.1. Does data line on straight line?

Calculate χ^2

Reject if $\chi^2 \geq 20$



Error of 1st kind: $\chi^2 \geq 20$ Reject H0 when true

Error of 2nd kind: $\chi^2 \leq 20$ Accept H0 when in fact quadratic or..

How often depends on:

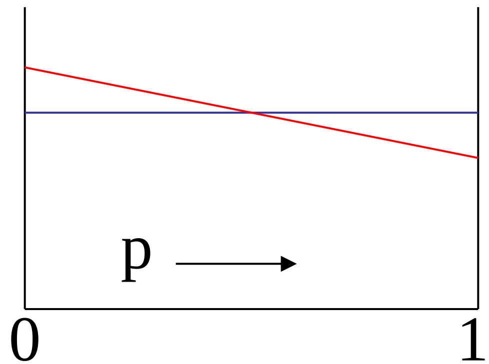
- Size of quadratic term
- Magnitude of errors on data, spread in x-values,…….
- How frequently quadratic term is present

Errors of 2nd kind: How often?

e.g. 2. Particle identification (TOF, dE/dx, Čerenkov,…….)

Particles are π or μ

Extract p-value for H0 = π from PID information



π and μ have similar masses

Of particles that have p ~ 1% ('reject H0'), fraction that are π is

a) ~ half, for equal mixture of π and μ

b) almost all, for "pure" π beam

c) very few, for "pure" μ beam

What is p good for?

Selecting sample of wanted events

e.g. kinematic fit to select t $\underline{t}$ events

t→bW, b→jj, W→μν $\underline{t}$→$\underline{b}$W, $\underline{b}$→jj, W→jj

Convert χ^2 from kinematic fit to p-value

Choose cut on χ^2 (or p-value) to select t $\underline{t}$ events

Error of 1st kind: Loss of efficiency for t $\underline{t}$ events

Error of 2nd kind: Background from other processes

Loose cut (large χ^2_{max} , small p_{min}): Good efficiency, larger bgd

Tight cut (small χ^2_{max} , larger p_{min}): Lower efficiency, small bgd

Choose cut to optimise analysis:

More signal events: Reduced statistical error

More background: Larger systematic error

p-value is not ……..

Does **NOT** measure Prob(H0 is true)

i.e. It is **NOT** P(H0|data)

It is P(data|H0)

N.B. P(H0|data) ≠ P(data|H0)

P(theory|data) ≠ P(data|theory)

"Of all results with $p \leq 5\%$, half will turn out to be wrong"

N.B. Nothing wrong with this statement

e.g. 1000 tests of energy conservation

~50 should have $p \leq 5\%$, and so reject H0 = energy conservation

Of these 50 results, all are likely to be "wrong"

$$P\,(\text{Data};\text{Theory}) \neq P\,(\text{Theory};\text{Data})$$

Theory = male or female

Data = pregnant or not pregnant

P (pregnant ; female) ~ 3%

$$P(\text{Data};\text{Theory}) \neq P(\text{Theory};\text{Data})$$

Theory = male or female

Data = pregnant or not pregnant

P (pregnant ; female) ~ 3%

but

P (female ; pregnant) >>>3%

BLIND ANALYSES

Why blind analysis?

Selections, corrections, method

Methods of blinding

- Add random number to result *
- Study procedure with simulation only
- Look at only first fraction of data
- Keep the signal box closed
- Keep MC parameters hidden
- Keep unknown fraction visible for each bin

After analysis is unblinded, ……..

* Luis Alvarez suggestion re “discovery” of free quarks

Choosing between 2 hypotheses

Possible methods:

$\Delta\chi^2$

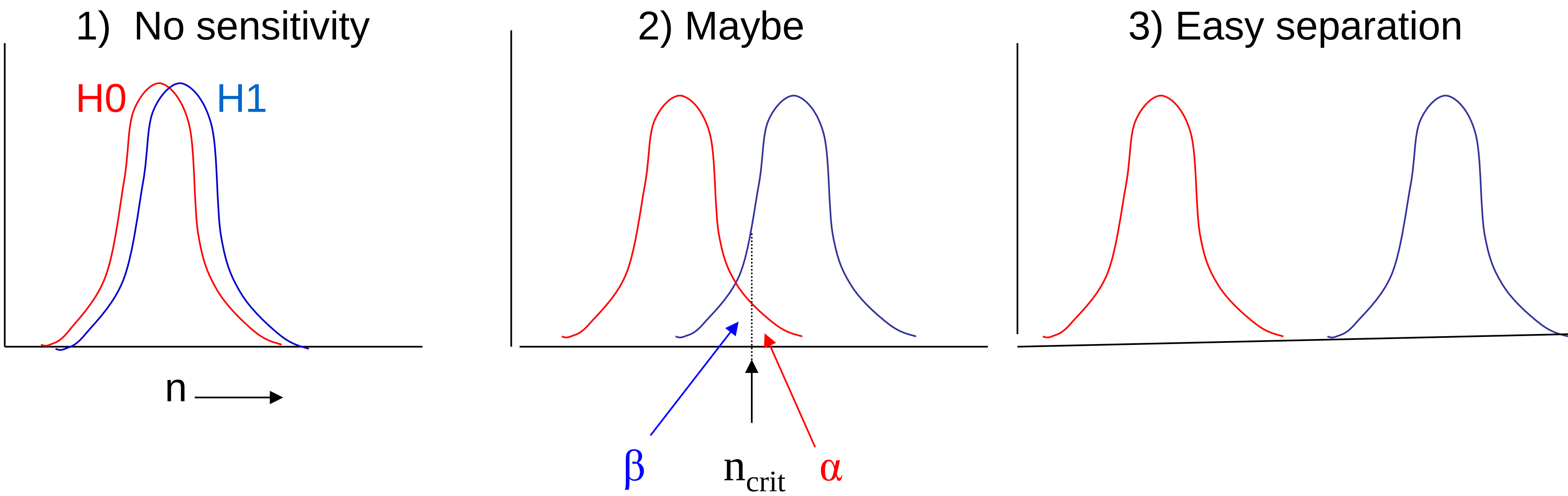
p-value of statistic →

ln$\mathcal{L}$–ratio

Bayesian:

- Posterior odds
- Bayes factor
- Bayes information criterion (BIC)
- Akaike …….. (AIC)

Minimise "cost"



Procedure: Choose α (e.g. 95%, 3σ, 5σ ?) and CL for β (e.g. 95%)

Given b, α determines n_{crit}

s defines β. For $s > s_{min}$, separation of curves → discovery or excln

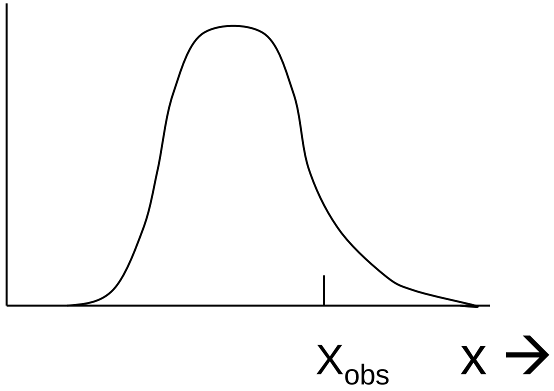
s_{min} = Punzi measure of sensitivity For $s \geq s_{min}$, 95% chance of 5σ discovery

Optimise cuts for smallest s_{min}

Now data: If $n_{obs} \geq n_{crit}$, discovery at level α

If $n_{obs} < n_{crit}$, no discovery. If $\beta_{obs} < 1 – CL$, exclude H1

p-values or $\mathcal{L}$ikelihood ratio?



$\mathcal{L}$ = height of curve

p = tail area

Different for distributions that

a) have dip in middle

b) are flat over range

Likelihood ratio favoured by Neyman-Pearson lemma (for simple H0, H1)

Use $\mathcal{L}$-ratio as statistic, and use p-values for its distributions for H0 and H1

Think of this as either

i) p-value method, with $\mathcal{L}$-ratio as statistic; or

ii) $\mathcal{L}$-ratio method, with p-values as method to assess value of $\mathcal{L}$-ratio

Why p $\neq$ Bayes factor

Measure different things:

p_0 refers just to H0; B_{01} compares H0 and H1

Depends on amount of data:

e.g. Poisson counting expt little data:

For H0, μ_0 = 1.0. For H1, μ_1 =10.0

Observe n = 10 $p_0 \sim 10^{-7}$ $B_{01} \sim 10^{-5}$

Now with 100 times as much data, μ_0 = 100.0 μ_1 =1000.0

Observe n = 160 $p_0 \sim 10^{-7}$ $B_{01} \sim 10^{+14}$

Bayes' methods for H0 versus H1

Bayes' Th: $P(A|B) = P(B|A) * P(A) / P(B)$

$$\frac{P(H0|data)}{P(H1|data)} = \frac{P(data|H0) * Prior(H0)}{P(data|H1) * Prior(H1)}$$

Posterior odds ratio — Likelihood ratio — Priors

N.B. Frequentists object to this

(and some Bayesians object to p-values)

Bayes' methods for H0 versus H1

$$\frac{P(H0|data)}{P(H1|data)} = \frac{P(data|H0) * Prior(H0)}{P(data|H1) * Prior(H1)}$$

Posterior odds   Likelihood ratio   Priors

e.g. data is mass histogram

H0 = smooth background

H1 = ……………………… + peak

1) Profile likelihood ratio also used but not quite Bayesian
(Profile = **maximise** wrt parameters.
Contrast Bayes which **integrates** wrt parameters)

2) Posterior odds

3) Bayes factor = Posterior odds/Prior ratio
(= Likelihood ratio in simple case)

4) In presence of parameters, need to integrate them out, using priors.
e.g. peak's mass, width, amplitude
Result becomes dependent on prior, and more so than in parameter determination.

5) Bayes information criterion (BIC) tries to avoid priors by
BIC = -2 *ln{$\mathcal{L}$ ratio} +k*ln{n}   k= free params; n=no. of obs

6) Akaike information criterion (AIC) tries to avoid priors by
AIC = -2 *ln{**L** ratio} + 2k

etc etc etc

LIMITS

- Why limits?
- Methods for upper limits
- Desirable properties
- Dealing with systematics
- Feldman-Cousins
- Recommendations

WHY LIMITS?

Michelson-Morley experiment → death of aether

HEP experiments: If UL on expected rate for new particle $<$ expected, exclude particle

CERN CLW (Jan 2000)

FNAL CLW (March 2000)

Heinrich, PHYSTAT-LHC, “Review of Banff Challenge”

SIMPLE PROBLEM?

Gaussian

$\sim \exp\{-0.5*(x-\mu)^2/\sigma^2\}$, with data x_0

No restriction on param of interest μ; σ known exactly

$\mu \leq x_0 + k\,\sigma$

BUT Poisson $\{\mu = s\varepsilon + b\}$

$s \geq 0$

ε and b with uncertainties

Not like : 2 + 3 = ?

N.B. Actual limit from experiment ≠ Expected (median) limit

Methods (no systematics)

Bayes (needs priors e.g. const, $1/\mu$, $1/\sqrt{\mu}$, μ, …..)

Frequentist (needs ordering rule,

possible empty intervals, F-C)

Likelihood (DON'T integrate your L)

χ^2 ($\sigma^2 = \mu$)

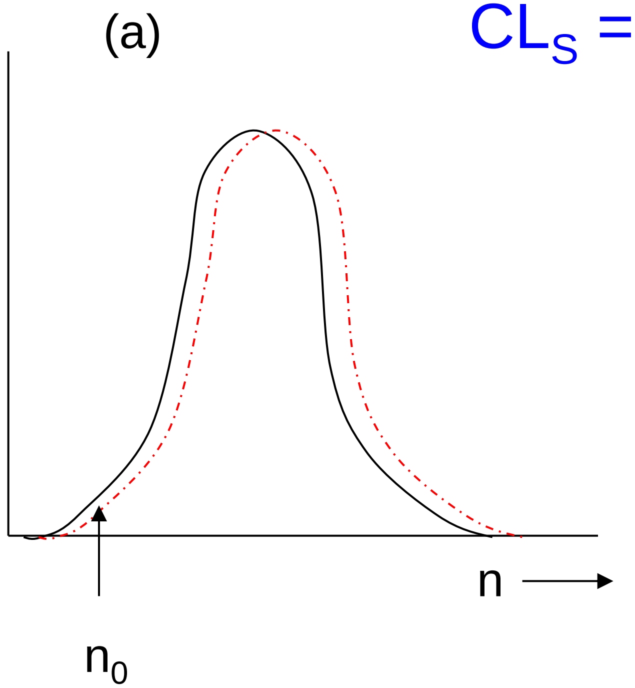
χ^2($\sigma^2 = n$)

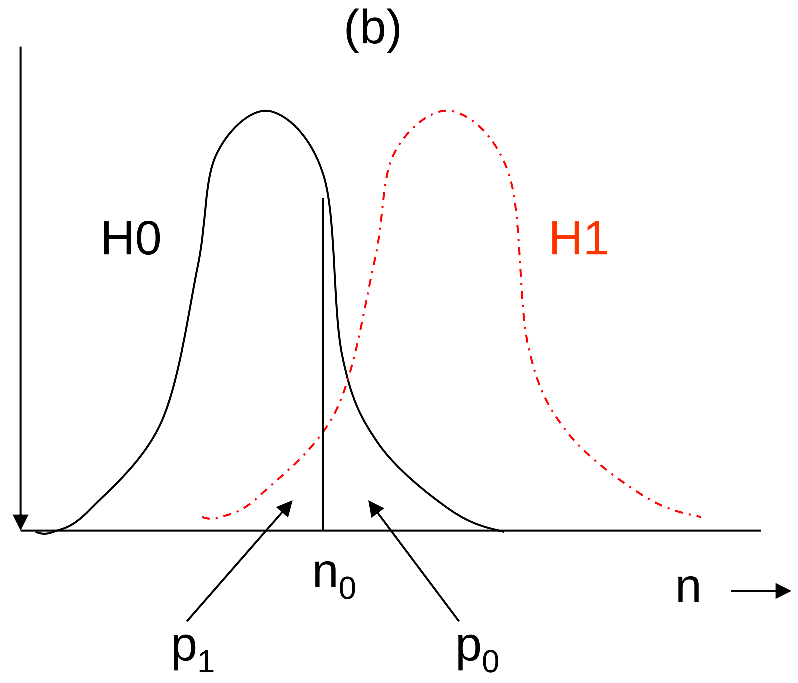
Recommendation 7 from CERN CLW: "Show your **L**"

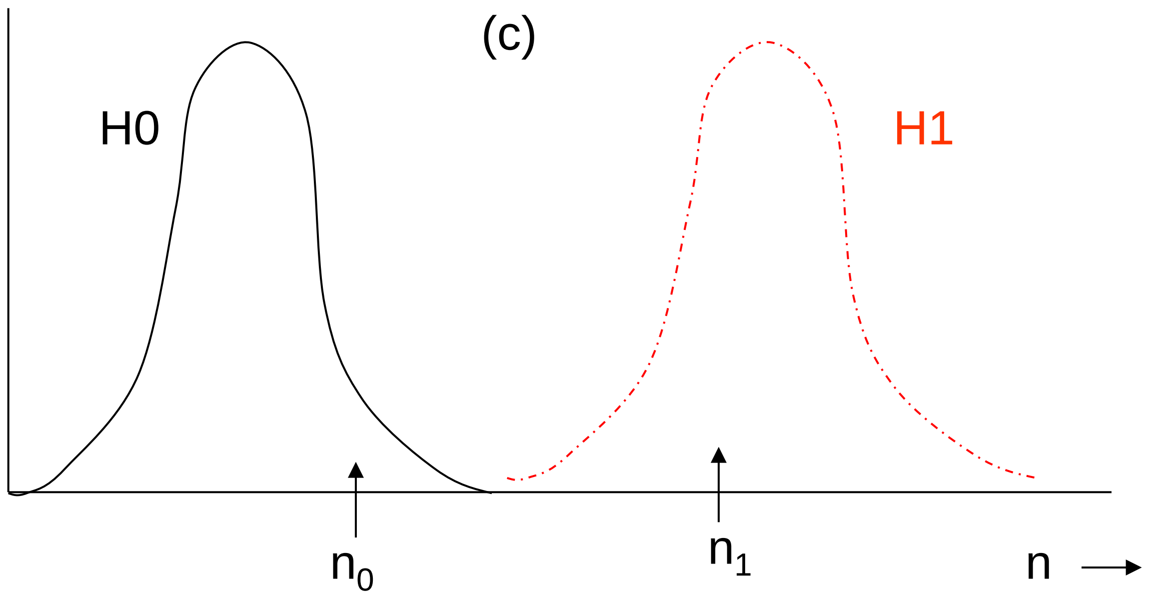
1) Not always practical

2) Not sufficient for frequentist methods

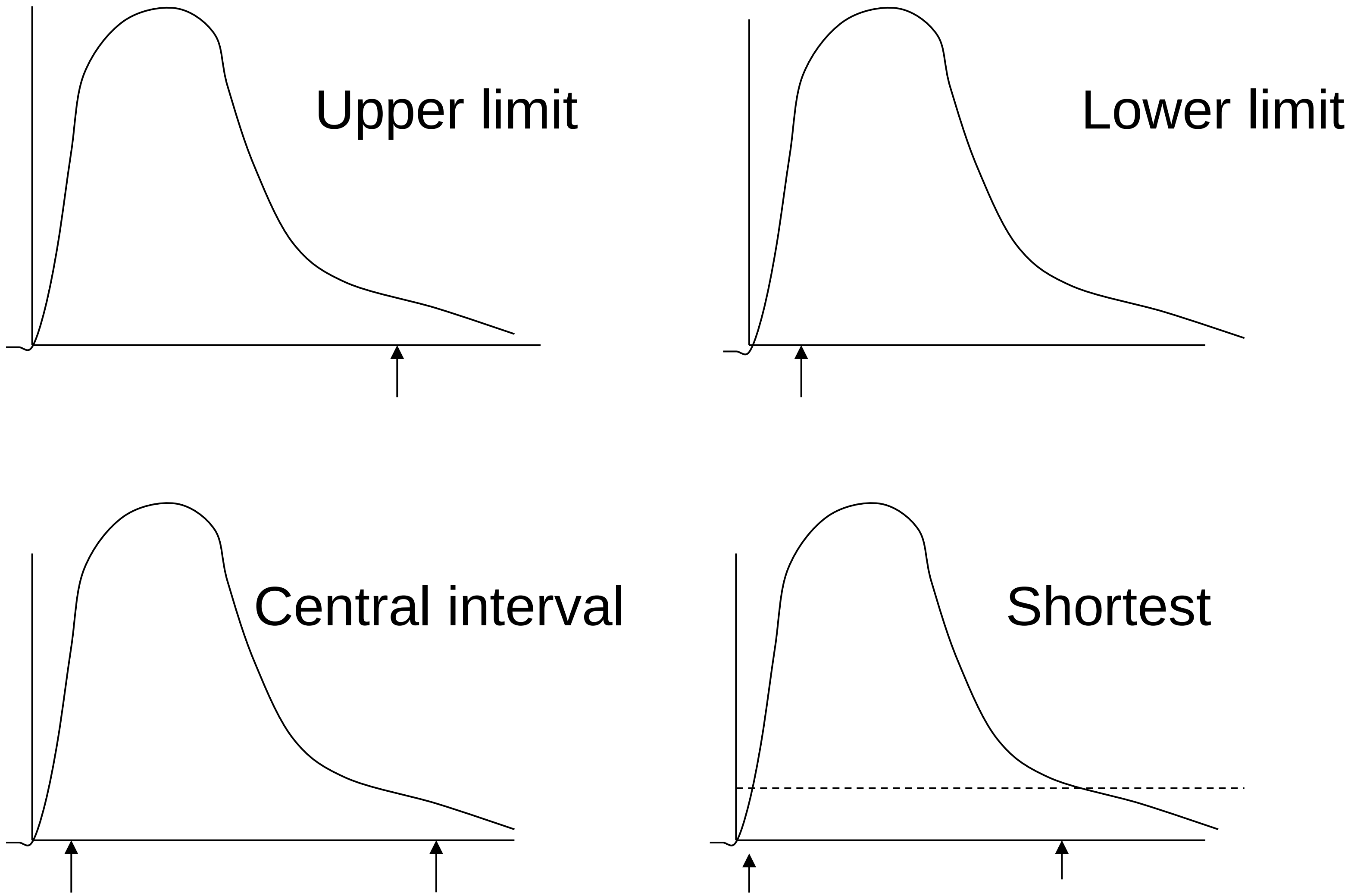
$CL_S = p_1/(1-p_0)$



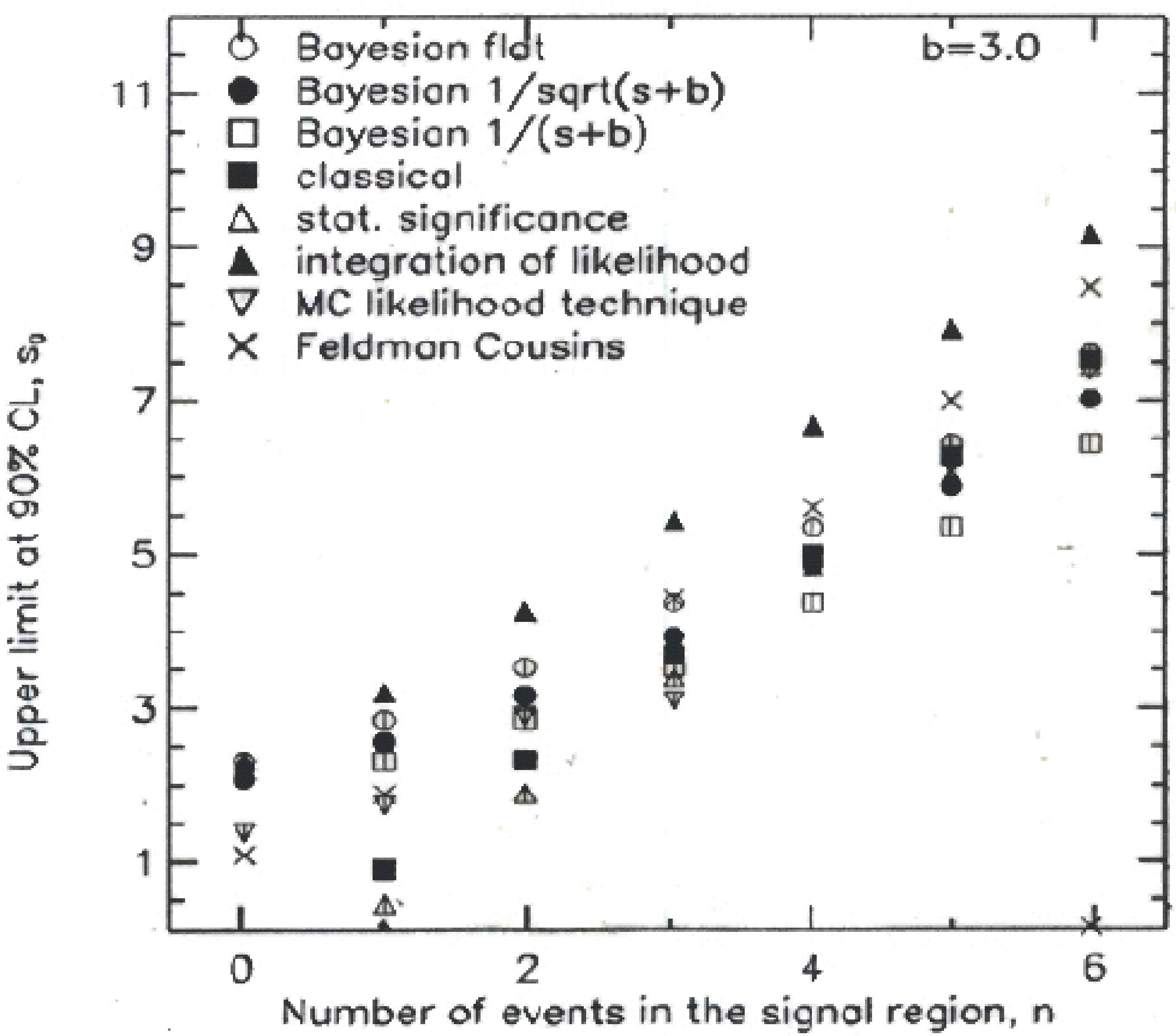




Bayesian posterior → intervals



Ilya Narsky, FNAL CLW 2000



DESIRABLE PROPERTIES

- Coverage
- Interval length
- Behaviour when $n < b$
- Limit increases as σ_b increases
- Unified with discovery and interval estimation

INTERVAL LENGTH

Empty → Unhappy physicists

Very short→ False impression of sensitivity

Too long→ loss of power

(2-sided intervals are more complicated because ‘shorter’ is not metric-independent:

e.g. 0 → 4 or 4 → 9 for x^2

cf 0 → 2 or 2 → 3 for x)

90% Classical interval for Gaussian

σ = 1 $\mu \geq 0$ e.g. $m^2(\nu_e)$

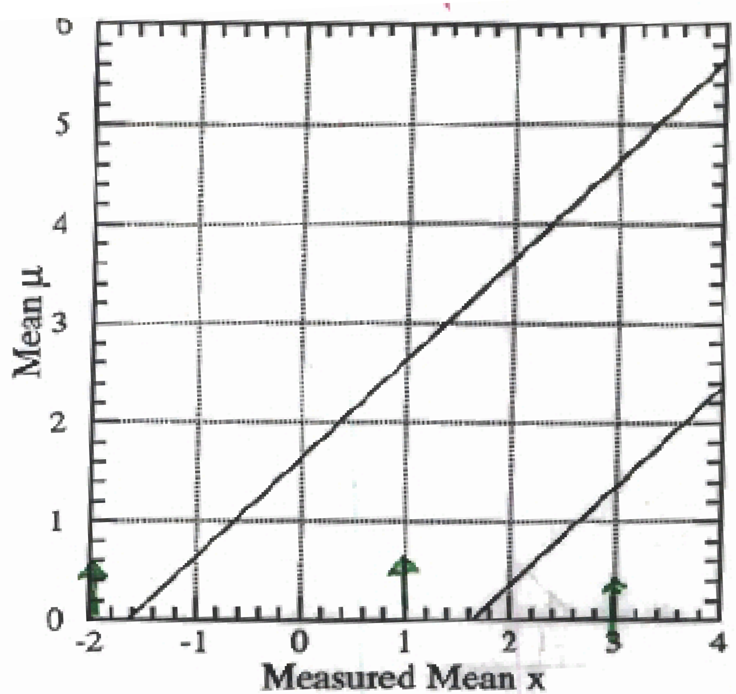


FIG. 3. Standard confidence belt for 90% C.L. central confidence intervals for the mean of a Gaussian, in units of the rms deviation.

$x_{obs} = 3$ Two sided limit

$x_{obs} = 1$ Upper limit

$x_{obs} = -2$ No region for μ

Behaviour when $n < b$

Frequentist: Empty for $n << b$

Frequentist: Decreases as n decreases below b

Bayes: For $n = 0$, limit independent of b

Sen and Woodroofe: Limit increases as data decreases below expectation

FELDMAN - COUSINS

Wants to avoid empty classical intervals →

Uses "$\mathcal{L}$-ratio ordering principle" to resolve ambiguity about "which 90% region?"

[Neyman + Pearson say $\mathcal{L}$-ratio is best for hypothesis testing]

Unified → No 'Flip-Flop' problem

Feldman-Cousins 90% conf intervals

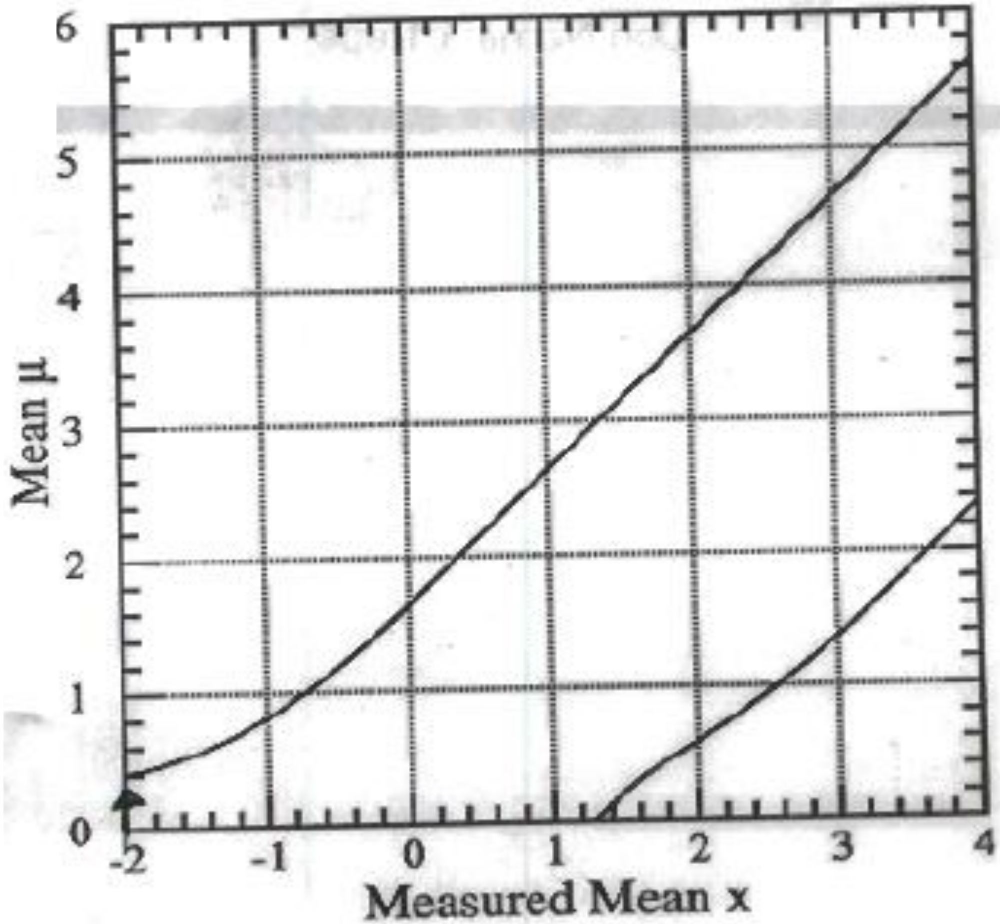


FIG. 10. Plot of our 90% confidence intervals for mean of a Gaussian, constrained to be non-negative, described in the text.

X_{obs} = -2 now gives upper limit

Recommendations?

CDF note 7739 (May 2005)

Decide method and procedure in advance

No valid method is ruled out

Bayes is simplest for incorporating nuisance params

- Check robustness
- Quote coverage

Quote sensitivity

Use same method as other similar expts

Explain method used

Case study: Successful search for Higgs boson

(Meeting of statisticians, atomic physicists, astrophysicists and particle physicist:
"What is value of H0?")

H^0 very fundamental

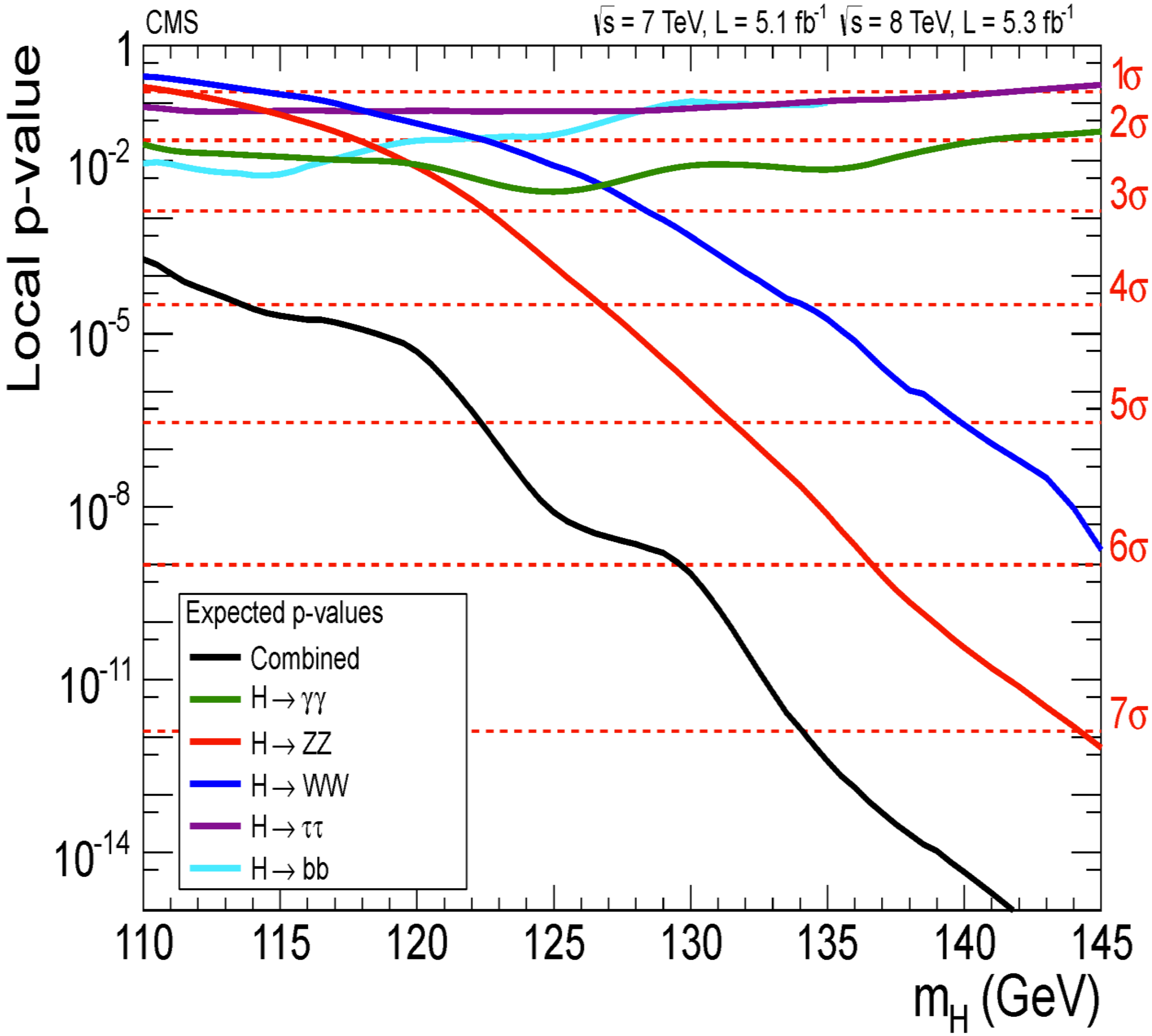
Want to discover Higgs,

but otherwise exclude

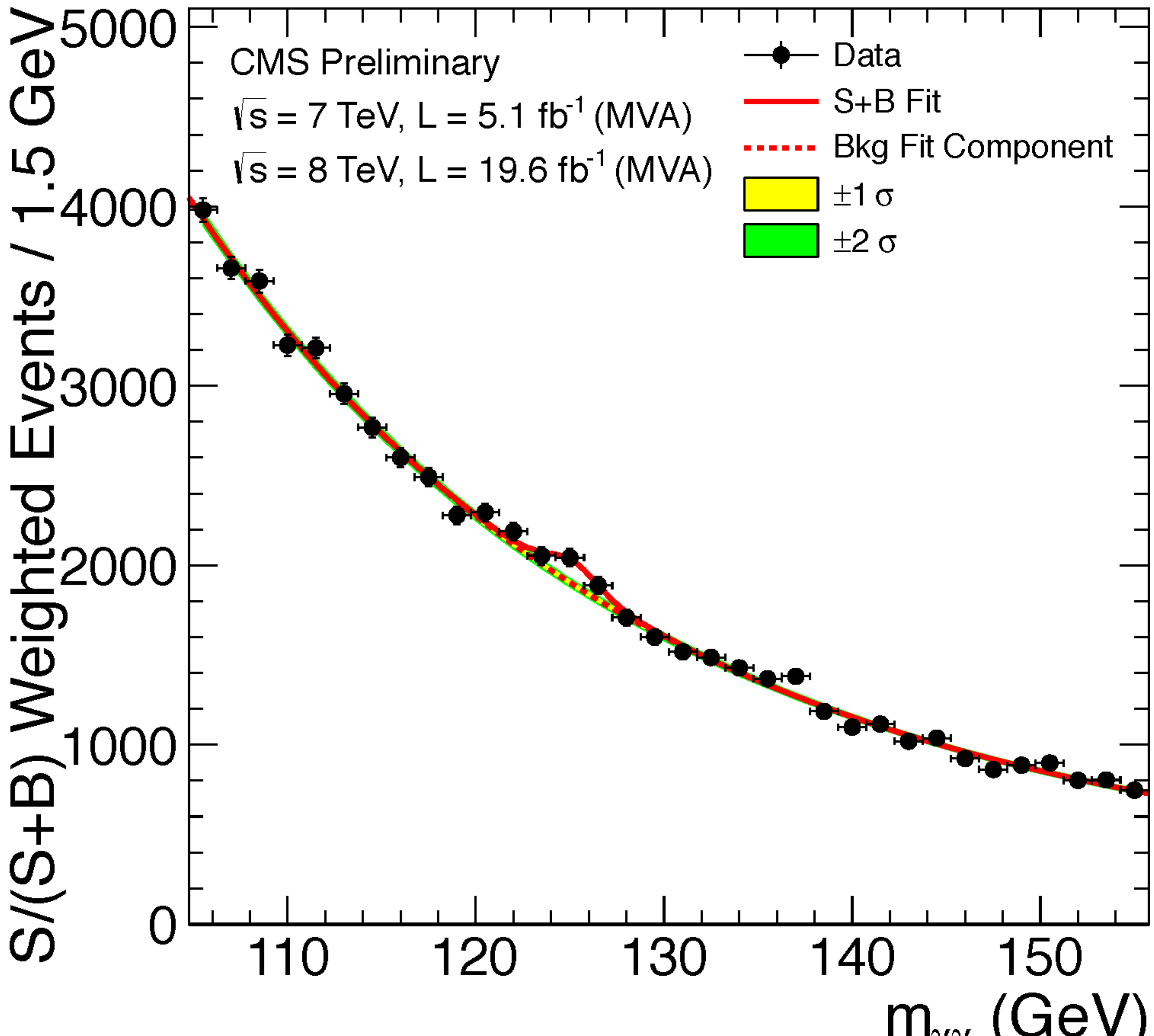
{Other possibility is 'not enough data to distinguish'}

Expected p-value as function of m_H

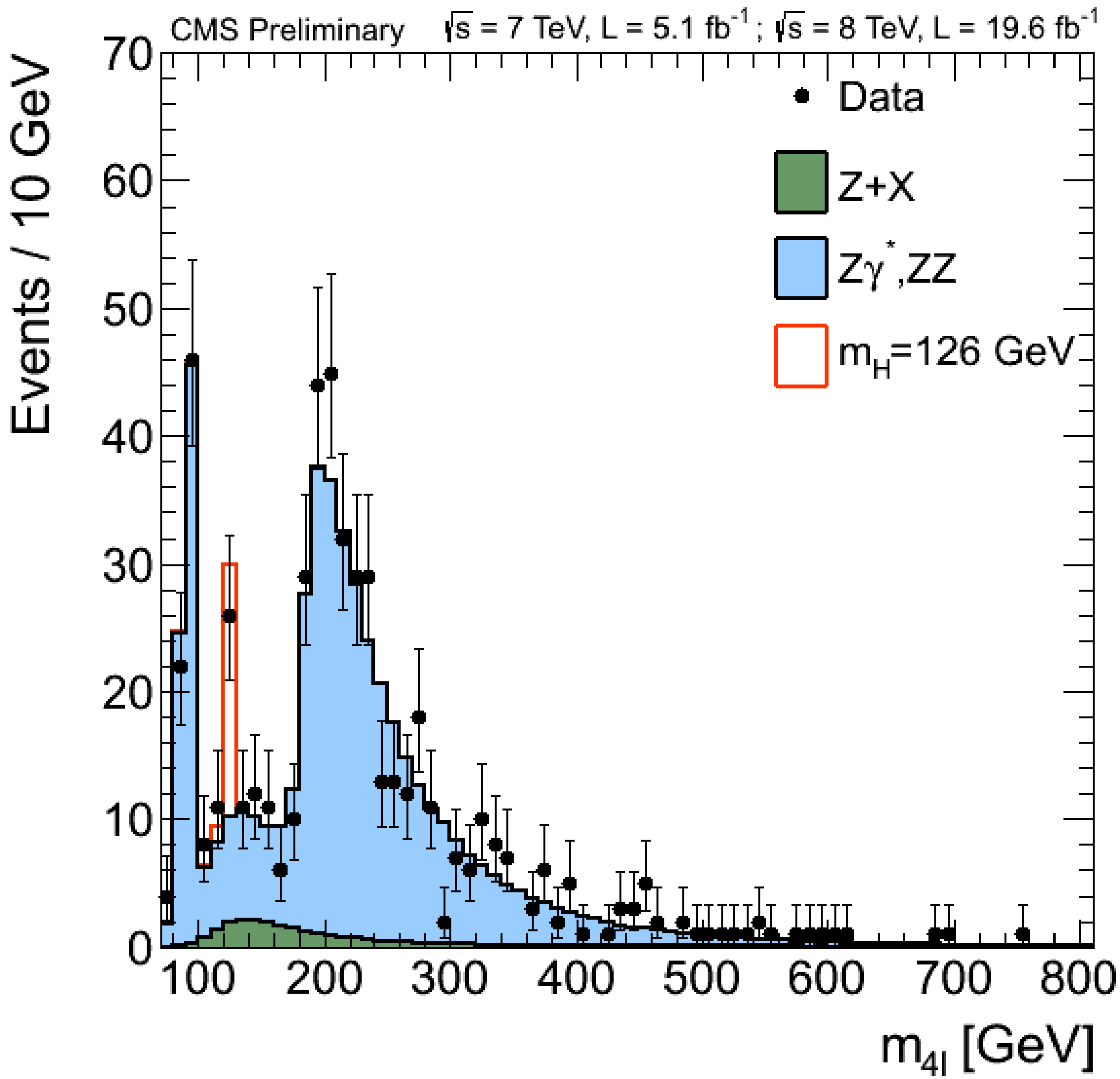
(For given m_H, prodn rate of S.M. H^0 is known)



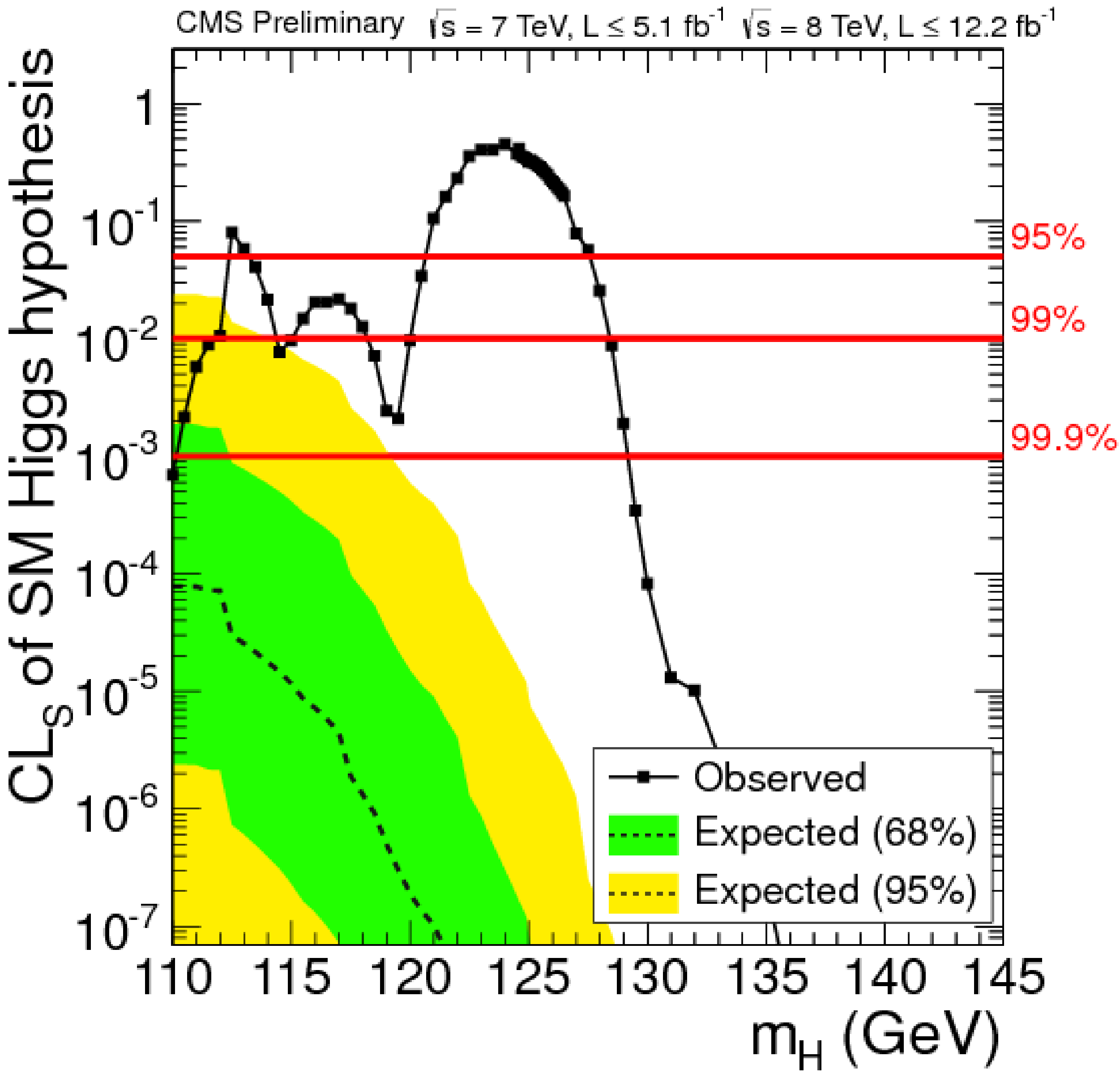
H→ γ γ: low S/B, high statistics



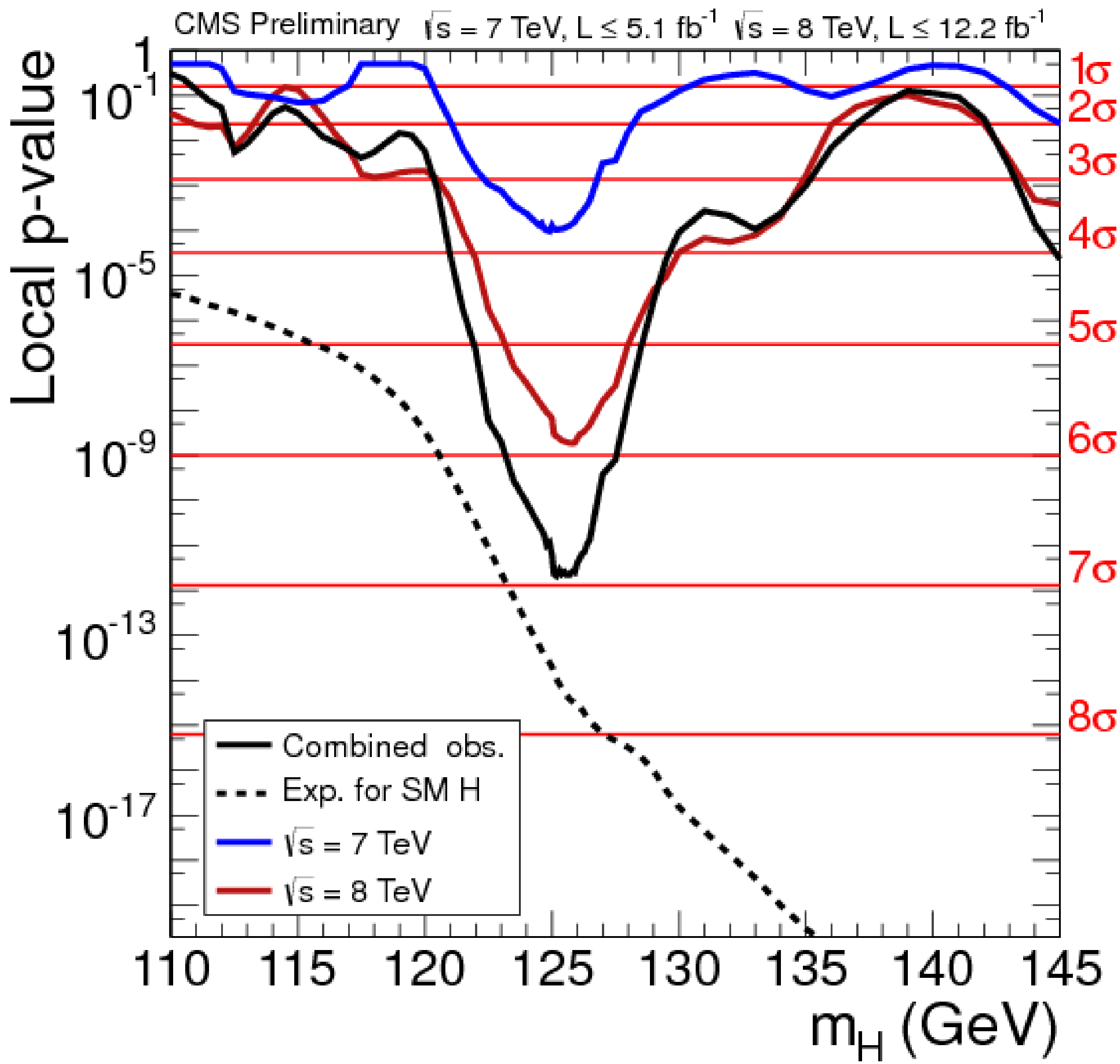
H→Z Z → 4 l: high S/B, low statistics



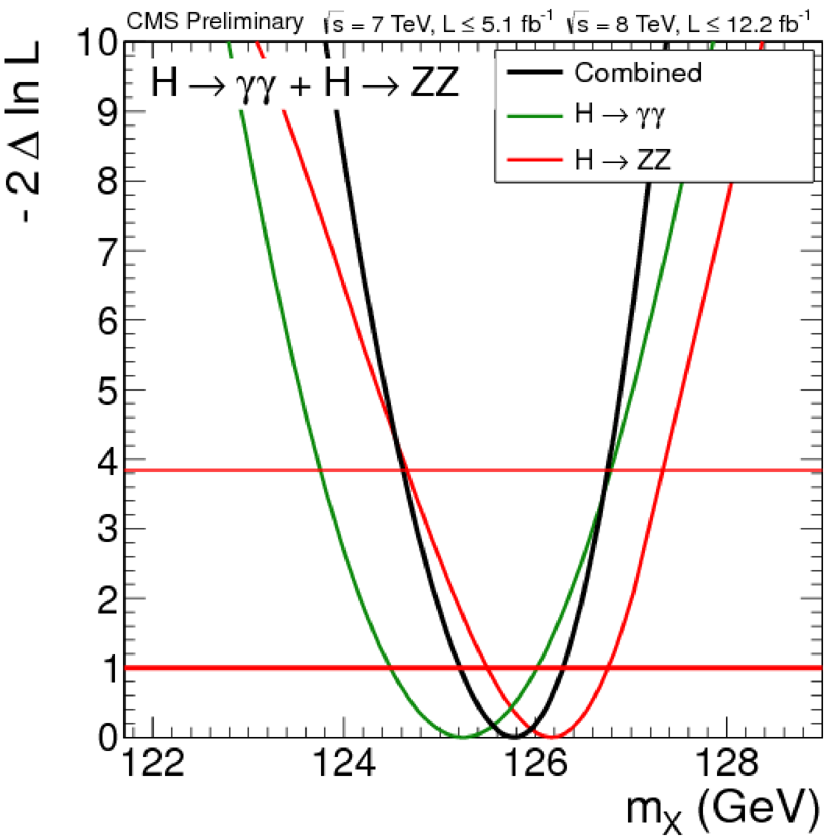
Exclusion of signal (at some masses) via CL_s

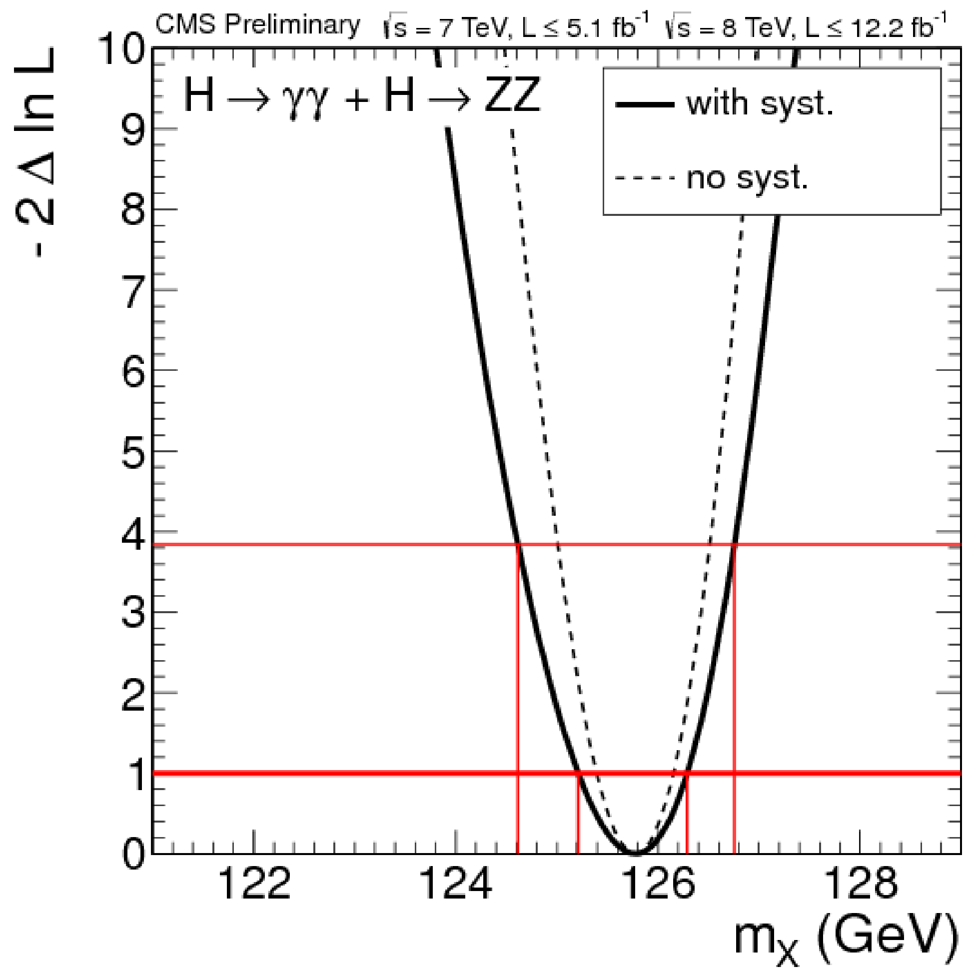


p-value for 'No Higgs' versus m_H

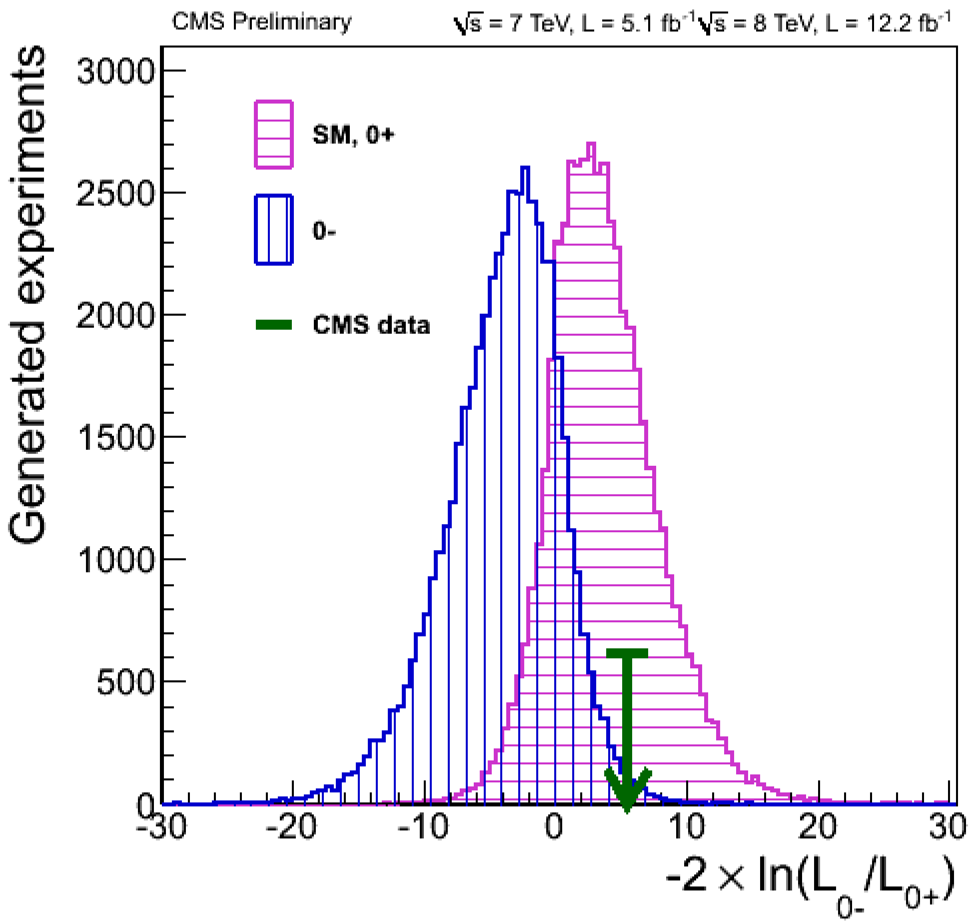


Likelihood versus mass





Comparing 0^+ versus 0^- for Higgs



http://cms.web.cern.ch/news/highlights-cms-results-presented-hcp

Summary

- P(H0|data) ≠ P(data|H0)
- p-value is NOT probability of hypothesis, given data
- Many different Goodness of Fit tests

  Most need MC for statistic → p-value
- For comparing hypotheses, $\Delta\chi^2$ is better than χ^2_1 and χ^2_2
- Blind analysis avoids personal choice issues
- Different definitions of sensitivity
- Worry about systematics
- H0 search provides practical example

PHYSTAT2011 Workshop at CERN, Jan 2011 (pre Higgs discovery)

"Statistical issues for search experiments"

Proceedings on website http://indico.cern.ch/conferenceDisplay.py?confId=107747