

Statistical methods for the Large Hadron Collider

Diego Tonelli/INFN



Seminario per il corso di statistica per fisici — Universita' di Trieste
Dec, 22, 2015

The LHC

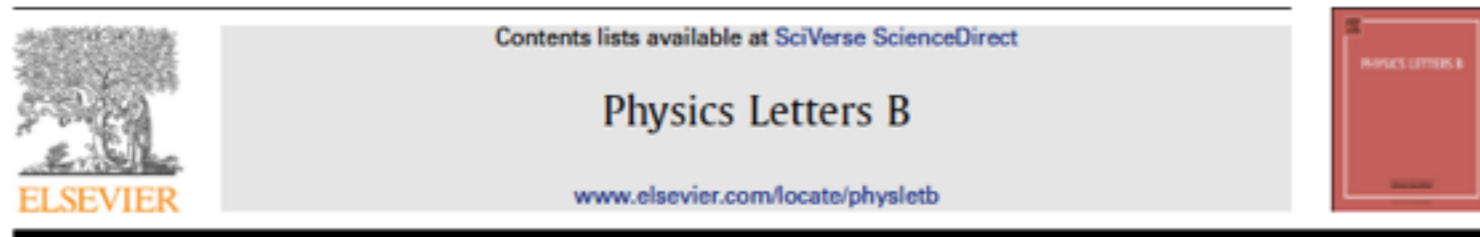


An accelerator that collides, 40 million times per second, protons against protons at center-of-momentum energies of 7 to 13 TeV.

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.

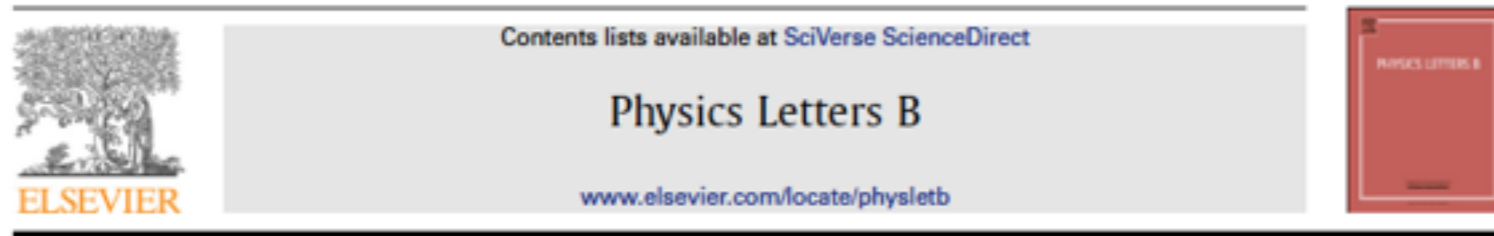
Done



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

CMS Collaboration^{*}

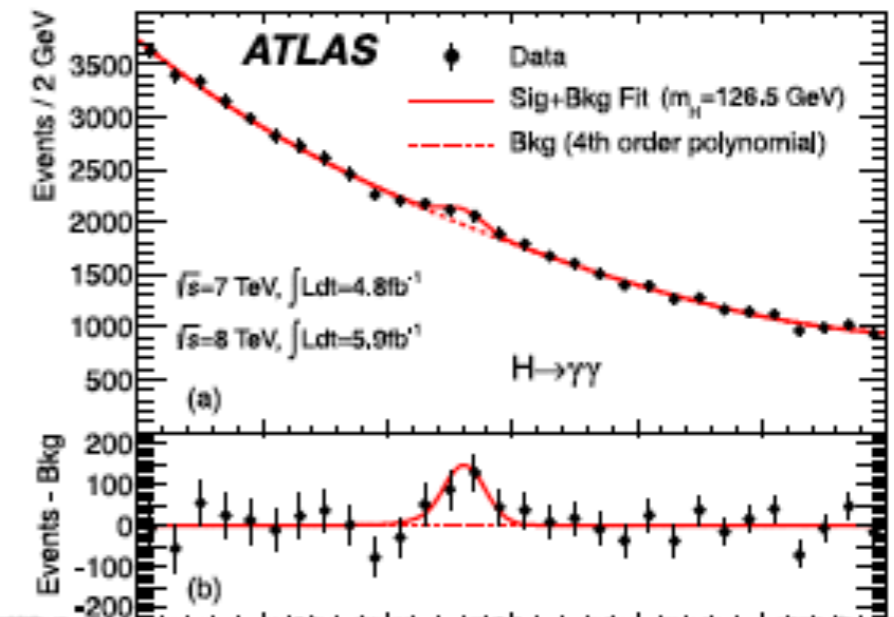
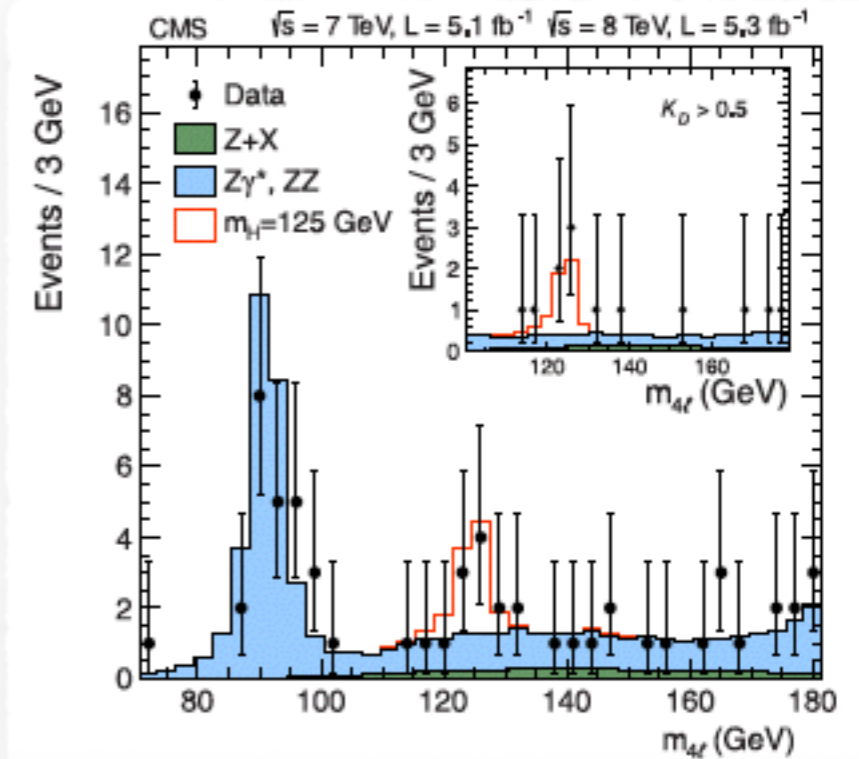
Physics Letters B 716 (2012) 1–29



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

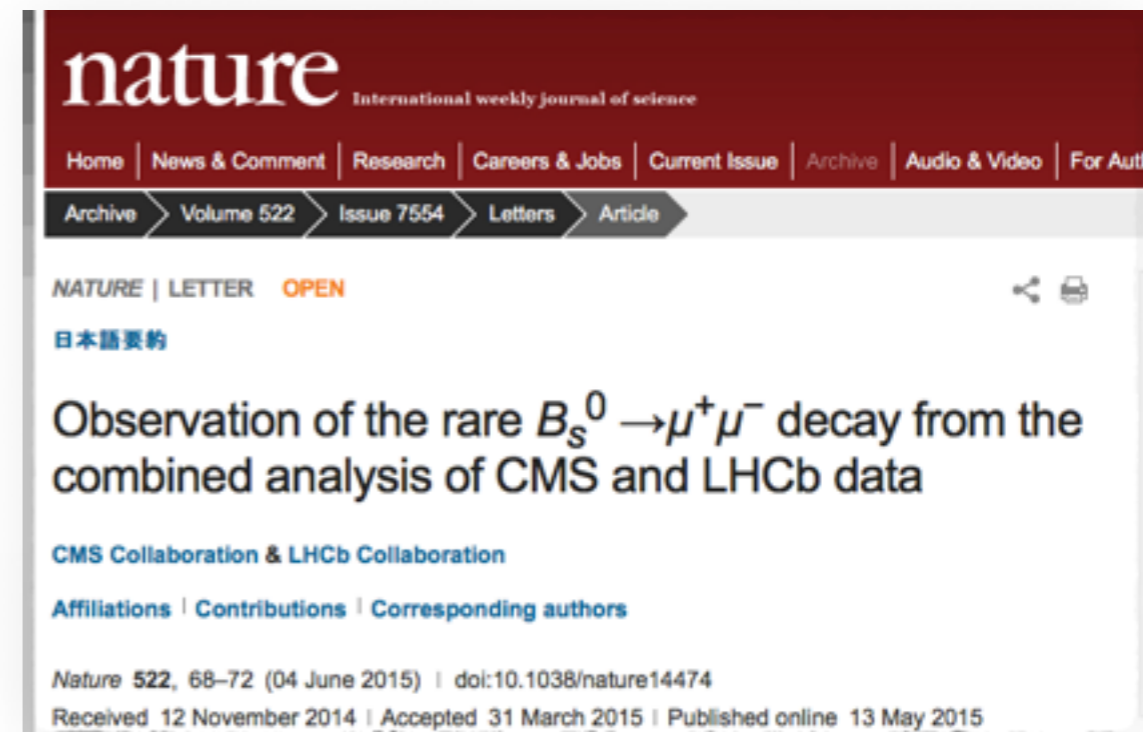
Hadron colliders at the energy frontier are machines with a broad discovery potential.

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

With some luck, the effort could result in discoveries. Otherwise, one reaches an improved understanding of known phenomena, useful to inform/guide future scientific decisions.

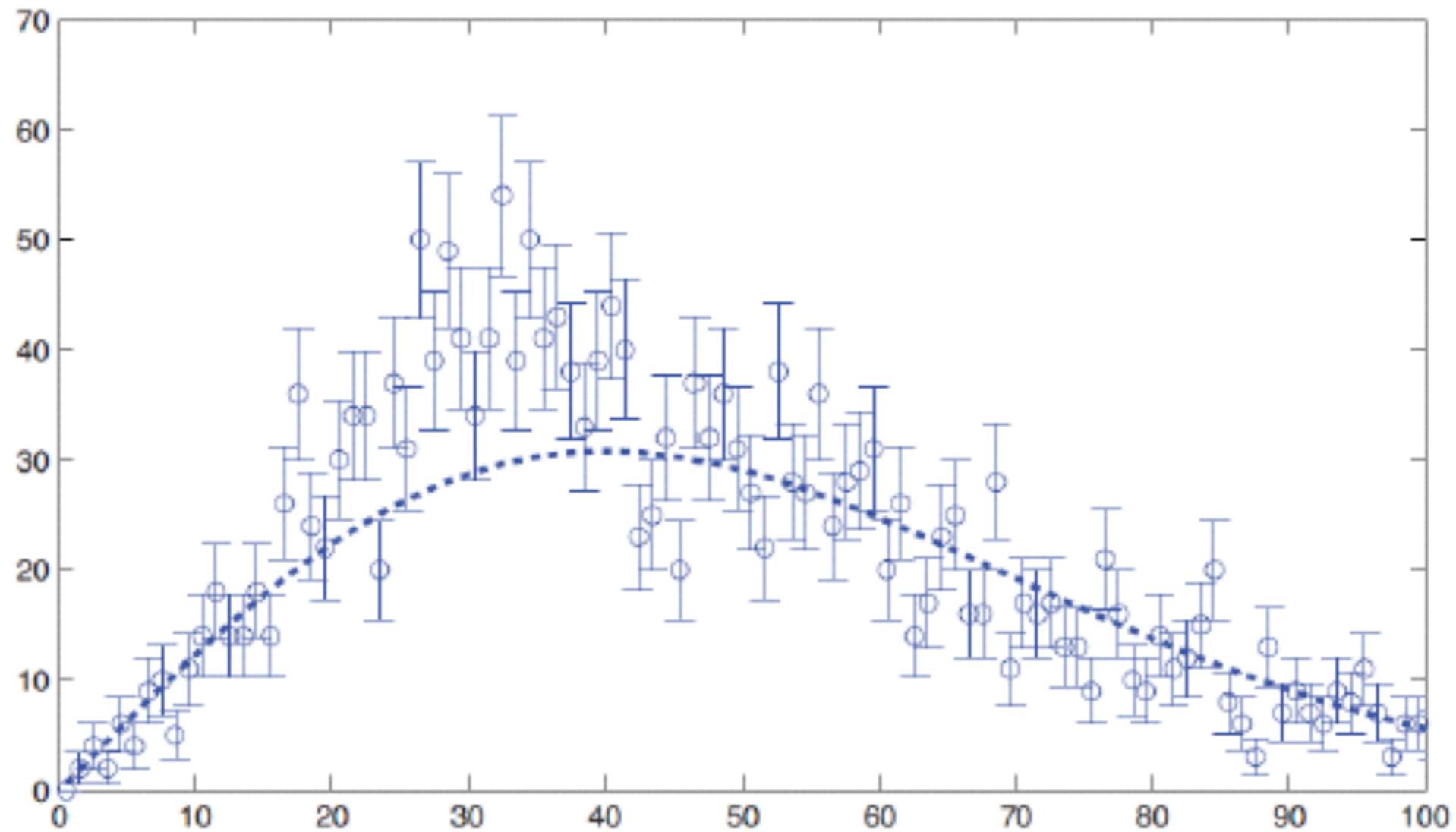
LHC experiments produce O(1000) physics measurements each year.

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.



The chief LHC statistical challenge

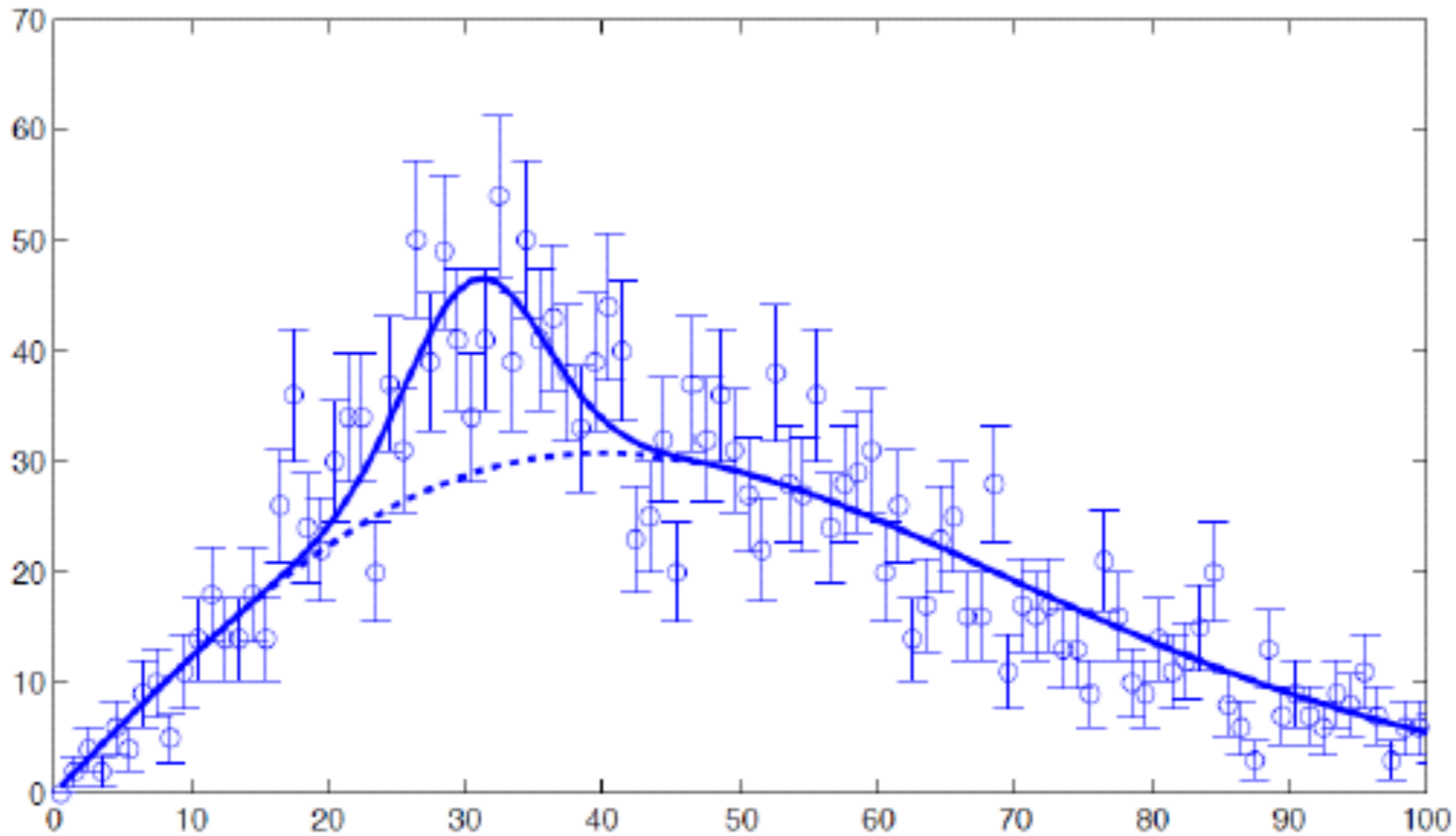
Background only?



Huge number of collisions, reconstructed with complex infrastructures.

However, at the end of the day lots of analyses boil down to studying whether a data distribution shows compatibility with what is expected from known processes only (“background”) or if it indicates presence 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 ?

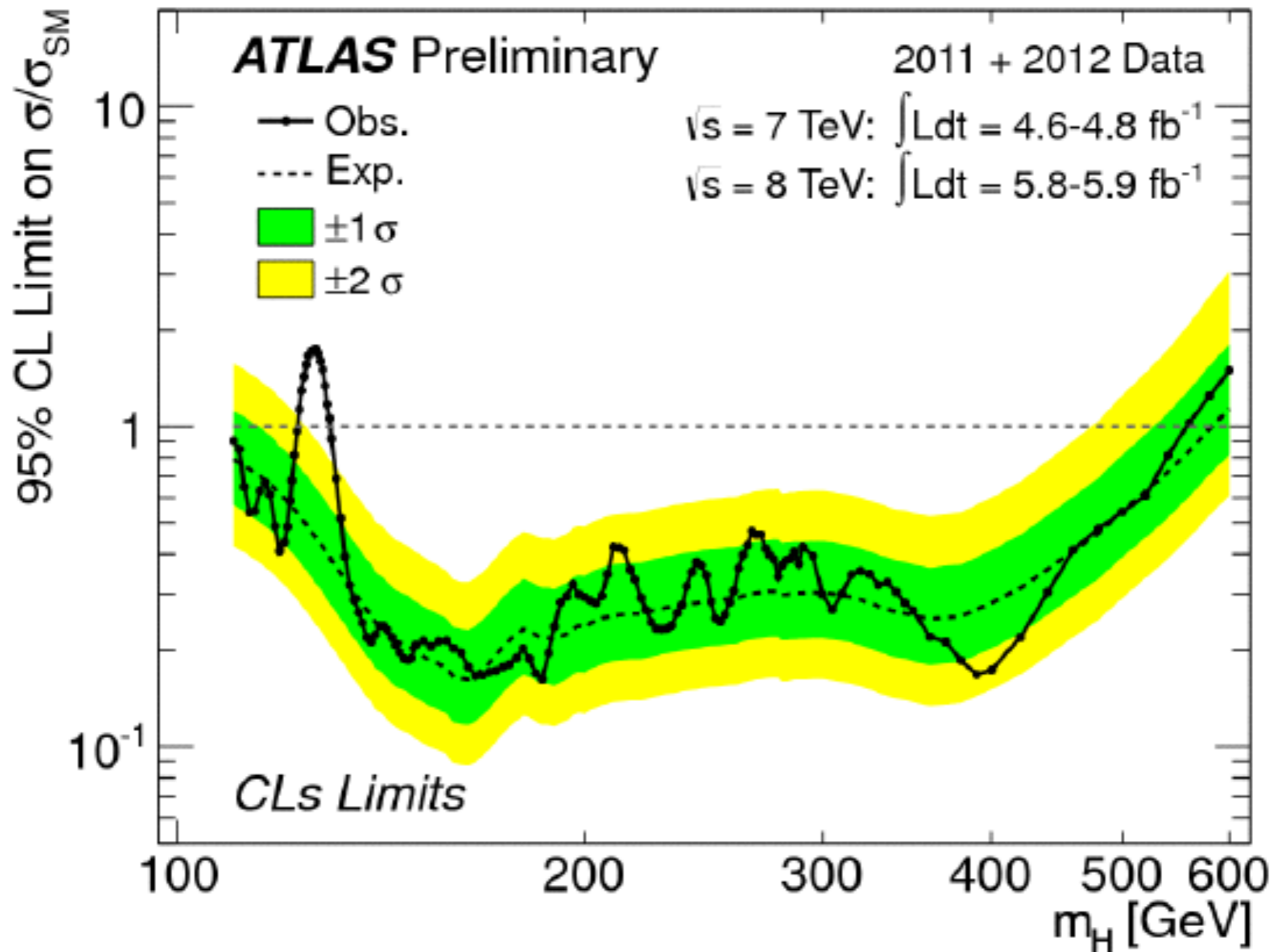
Today

- p-values, look-elsewhere-effect, 5-sigma and all that
- Role of modeling
- Systematic uncertainties
- Confidence regions with nuisance parameters.
- Punzi-effect and other miscellanea

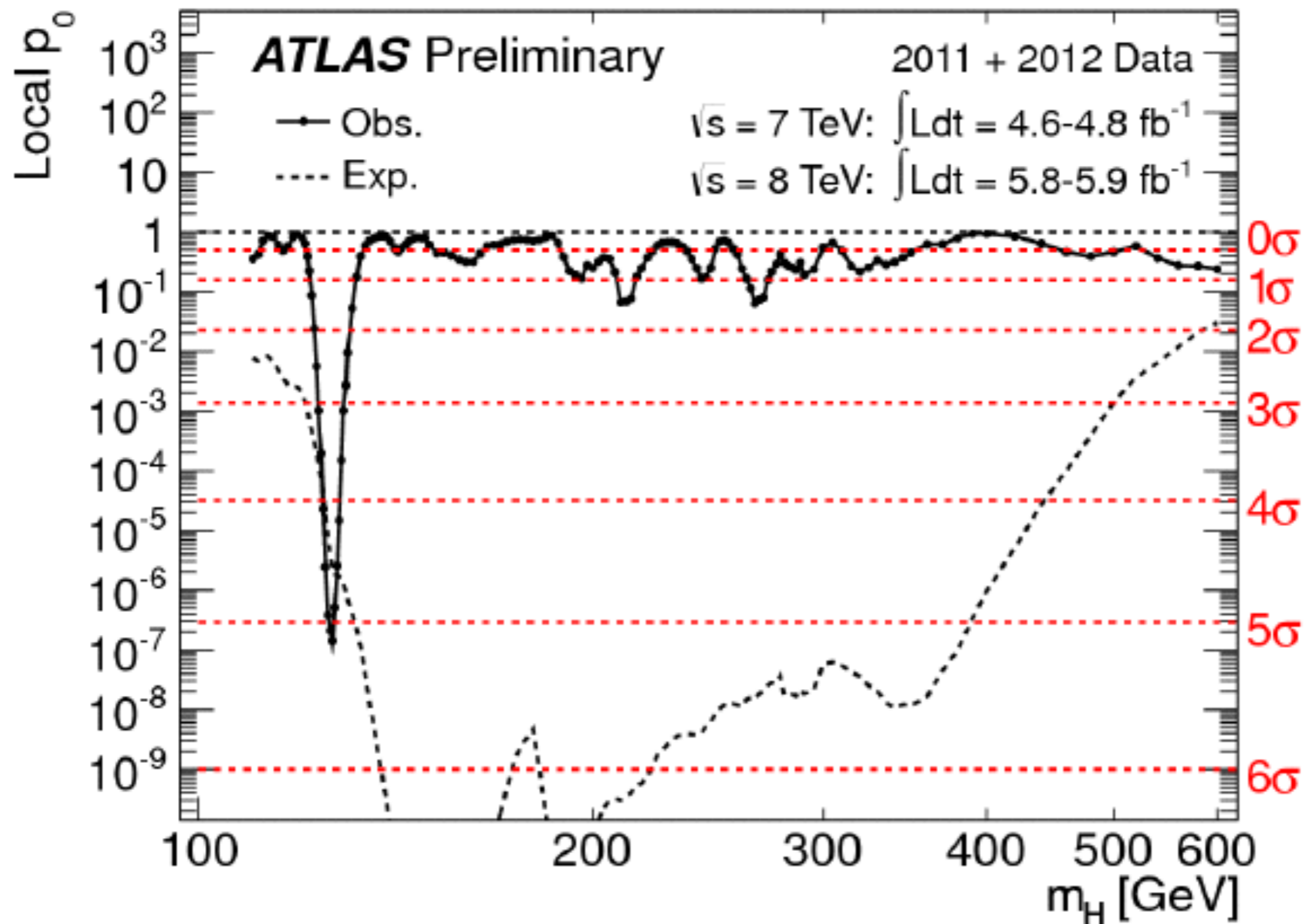
In god we trust, all others bring data

W. E. Deming

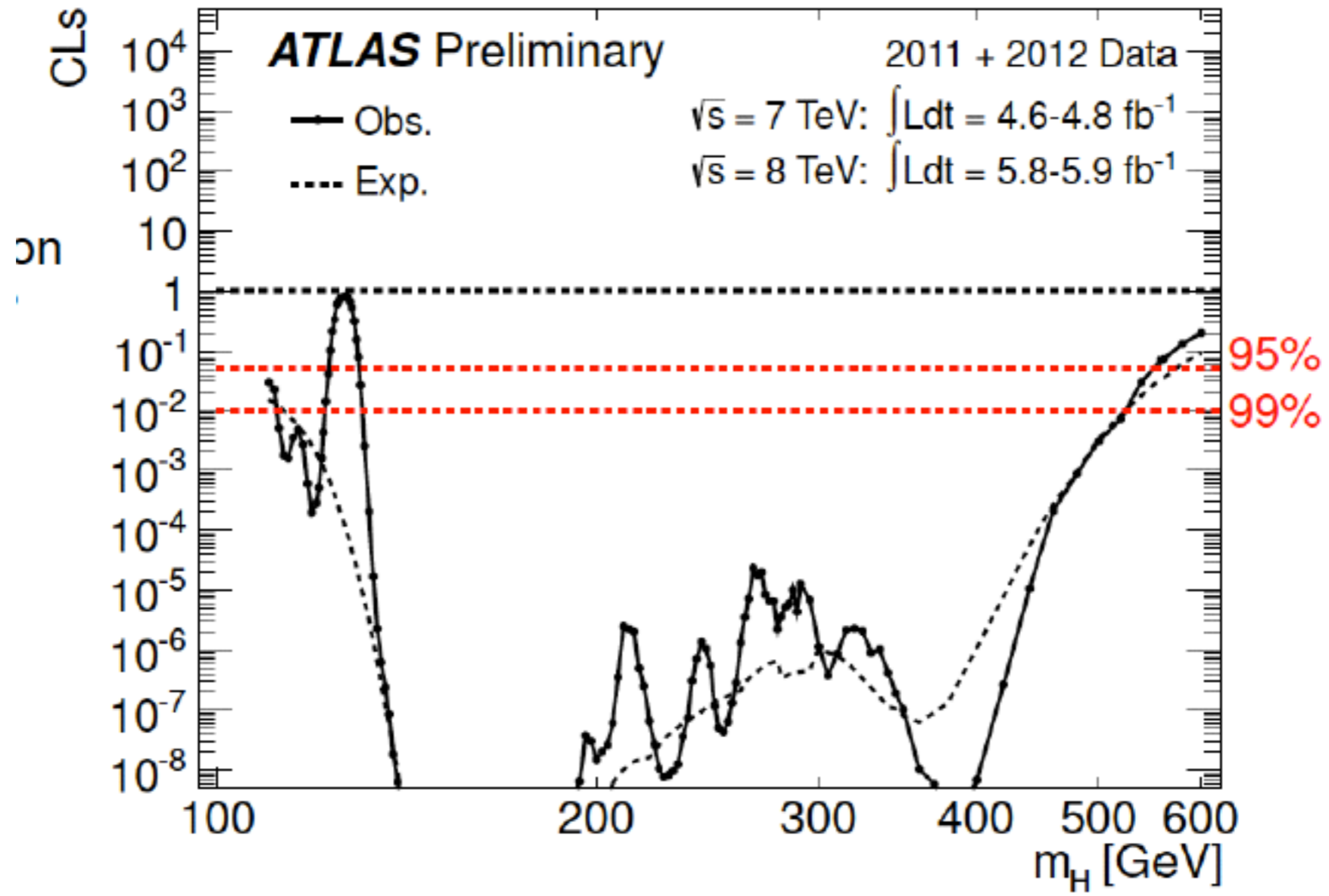
What does the “Brazil plot” mean?



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



What is CLs?



Caveats

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

Please interrupt me to ask questions, will help keeping all of us awake. Also, feel free to follow-up at diego.tonelli@cern.ch

Slides will be available shortly at www.pi.infn.it/~dtonelli/StatLHC

Notation

Probably my notation won't match the one you are used from Prof. Milotti's course. Apologies: I know this might be confusing, did not have the time to uniformize.

Will try to stick to minimal amount of notation. You should be able to follow most of today's talk by keeping in mind that I usually use

- **x are observed data** (e.g., diphoton invariant mass) or any function of them (e.g., likelihood ratio). It's irrelevant whether x is one- or multi-dimensional.
- **m are physics parameter to be measured** (e.g., Higgs mass). Irrelevant whether m is one- or multidimensional
- **$p(x|m)$ is the probability density function for x given m.**

Testing hypotheses

P-values, look-elsewhere effect, and all that

Significant deviation?

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

The first thing you do if you suspect you may have a discovery.

At LHC (and in particle physics at large) this is mostly addressed using “p-values”

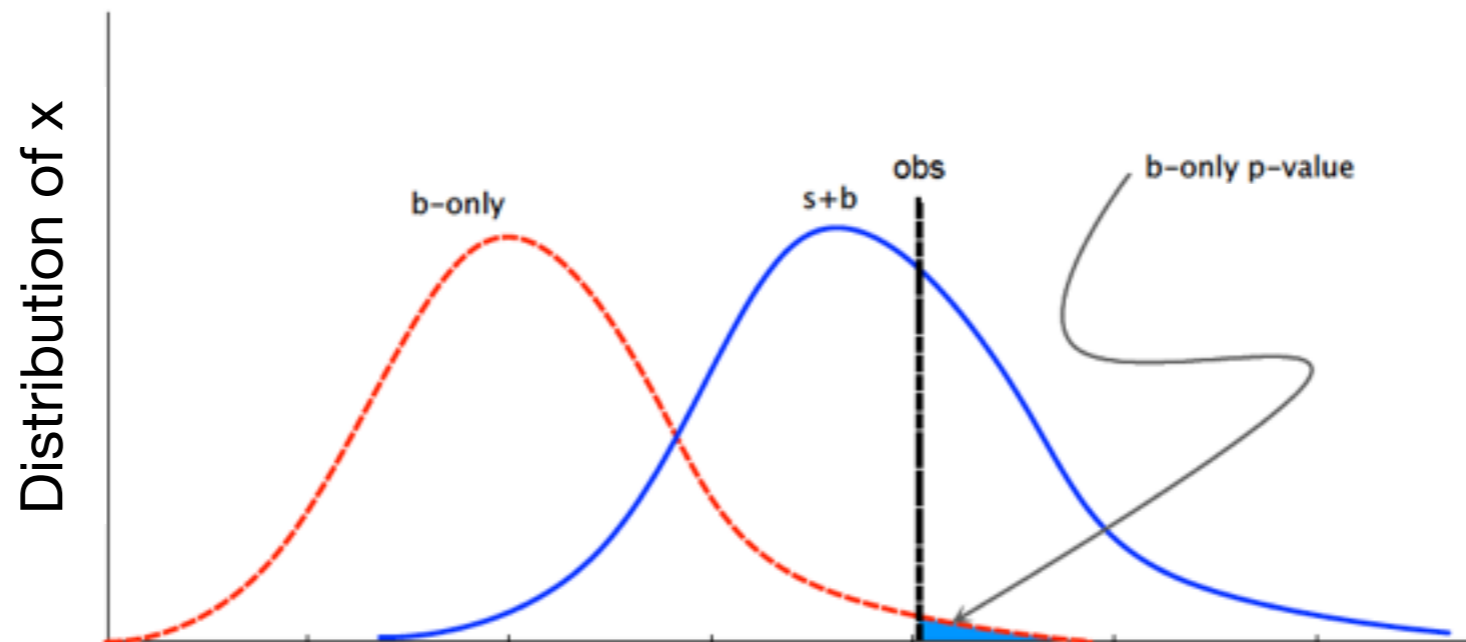
A p-value is a random variable that provides a quantitative evaluation of the probabilities to be observing a genuine anomaly or a fluke.

(Check this out for a funny piece about origin of p-values

<http://priceconomics.com/the-guinness-brewer-who-revolutionized-statistics/>)

Testing “signal+background” vs “background”

Need two hypotheses. Null hypothesis: only known phenomena contribute (“background”). Signal hypothesis: new phenomena (“signal”) contribute as well.



Arbitrary function x of the data that allows separating between the two hypotheses

Devise a function x of the data (e.g., signal event count), whose pdf under the null hypothesis $p(x|m_0)$ “differs” from the pdf under the signal hypothesis $p(x|m_1)$.

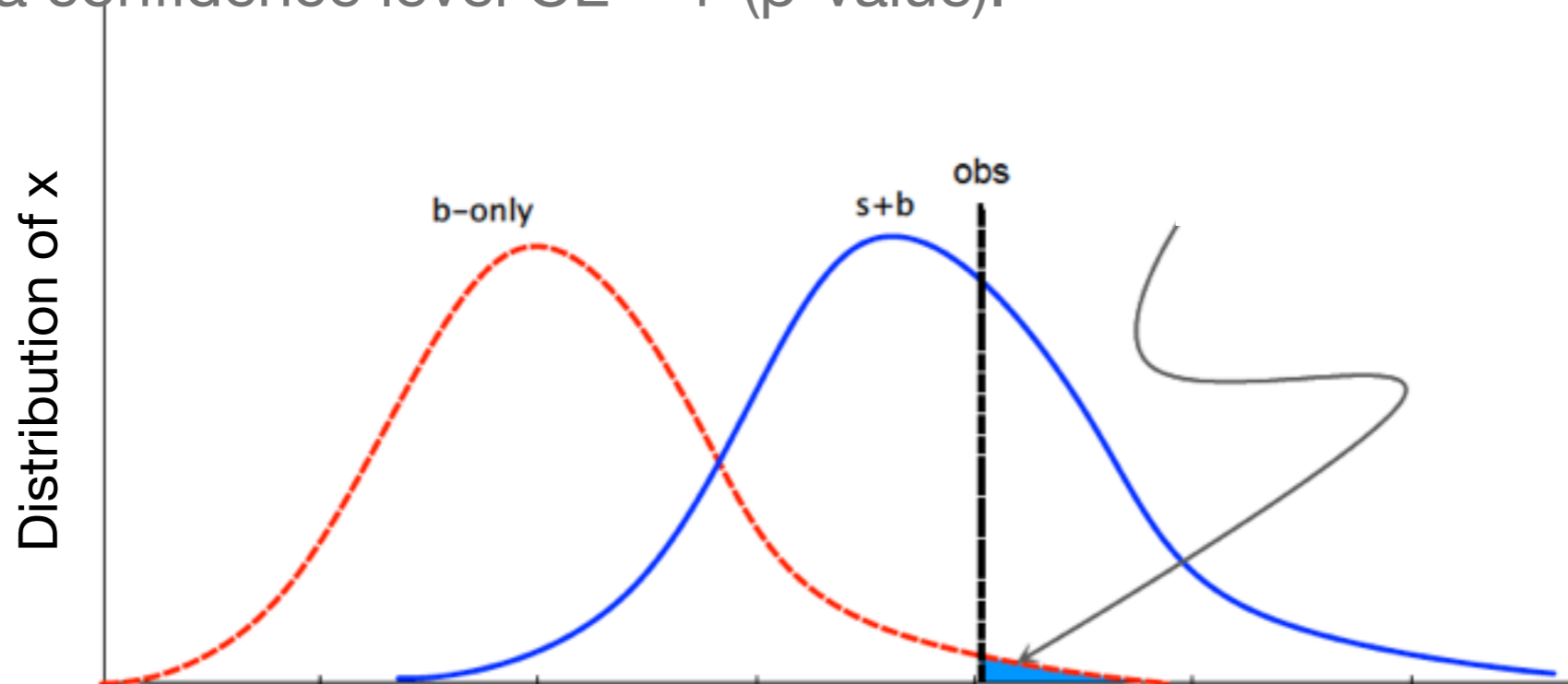
Predict distributions of x under the two hypotheses (typically done using simulation)

Decide and fix prior to observation the false-positive rate: how far the observed value of x should be from the core of $p(x|m_0)$ to exclude the null (i.e., favor signal.)

p-values

Make the observation. The “relative location” of the observation x with respect to the two shapes offers 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 null 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 hypothesis. If $p\text{-value} < \text{false-positive rate}$, the null is excluded at a confidence level $CL = 1 - (p\text{-value})$.

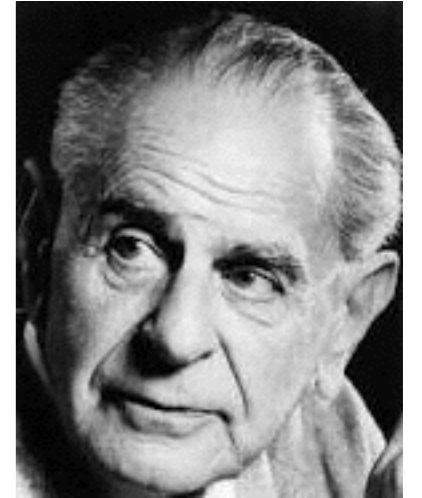


Arbitrary function x of the data that allows for separation between the two hypotheses

Model testing the Popperian way

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

“Discover” a signal by excluding its absence to an high-level of significance (that is by excluding that only background contributes).



Set limits to the existence of a signal by excluding to an high level of significance is presence.

Karl Popper
(1902-1994)

p-value is not a probability!

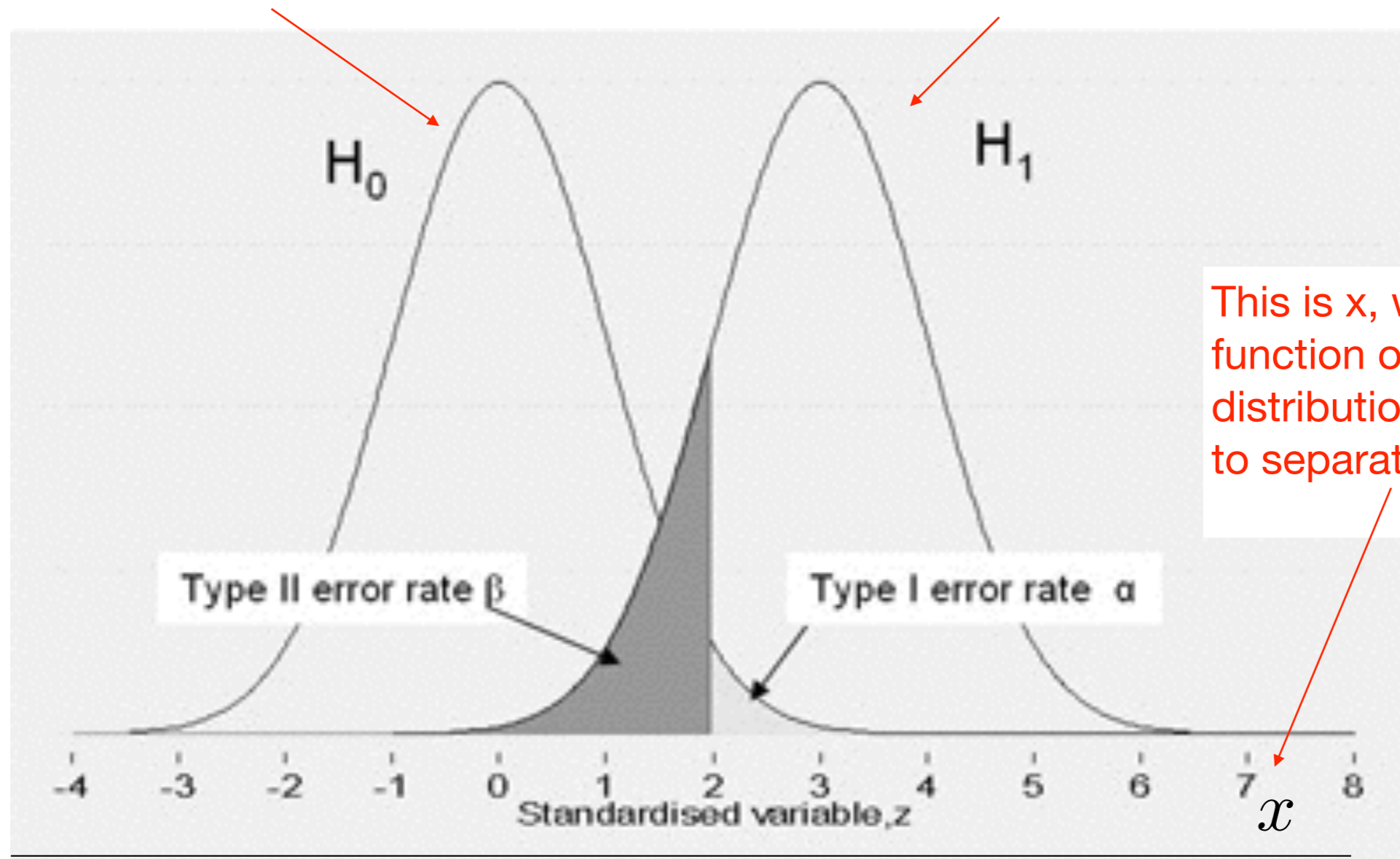
It is a random variable (function of the data) that is distributed uniformly if the tested hypothesis is true.

It does not express the probability that an hypothesis is true or false! It relates to the probability that, if an hypothesis were true, one would observe x or a more extreme value.

In one slide

This is $p(x|m_0)$, the distribution of x under the null hypothesis

This is $p(x|m_1)$, the distribution of x under the signal hypothesis



Symbol	Meaning
α	Rate of false positives (Type I error: reject H_0 , while it was true)
β	Rate of false negatives (Type II error: reject H_1 , while it was true)
$1 - \beta$	Power of the test

Folklore

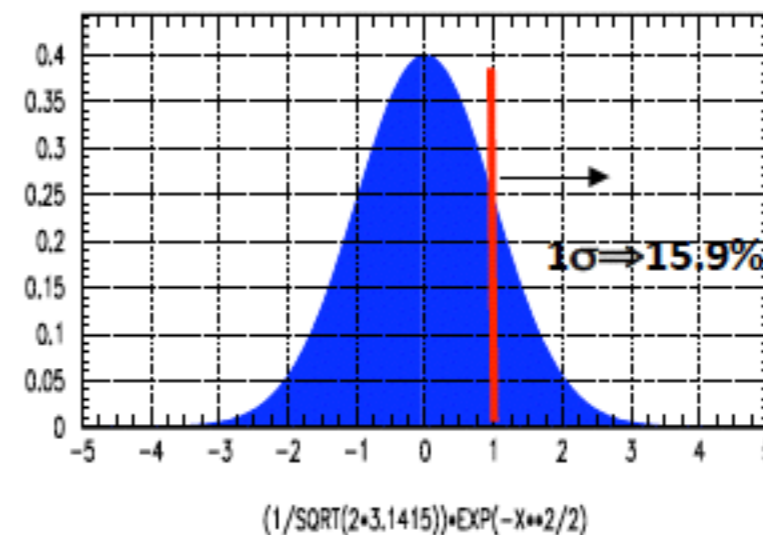
Physicists' lingo goes like "at how many sigma such and such result is significant"
We have less feel for p-values.

The "number of sigma" (or z-value) is just a translation of p-values using the integral 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)

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

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



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 $n=17, 18, 19, 20, 0, 1, 2, 3$.

$P(n=0,1,2,3,17,18,19,\text{or }20) = 0.26\%$ 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.

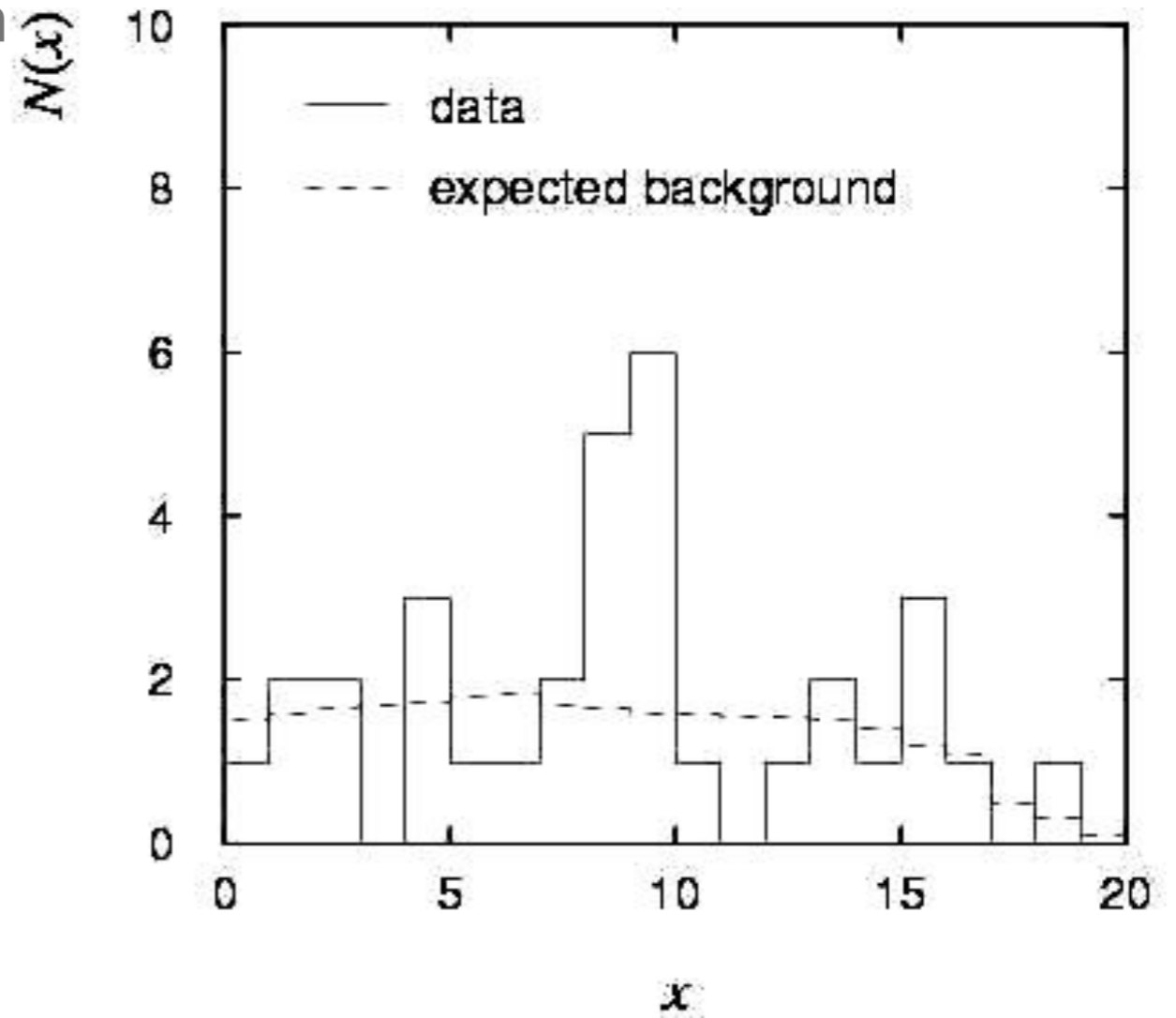
The count in each bin is a Poisson random variable, whose mean in the H_0 hypothesis is given by the dashed line

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

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

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

Is this evaluation fair or biased?



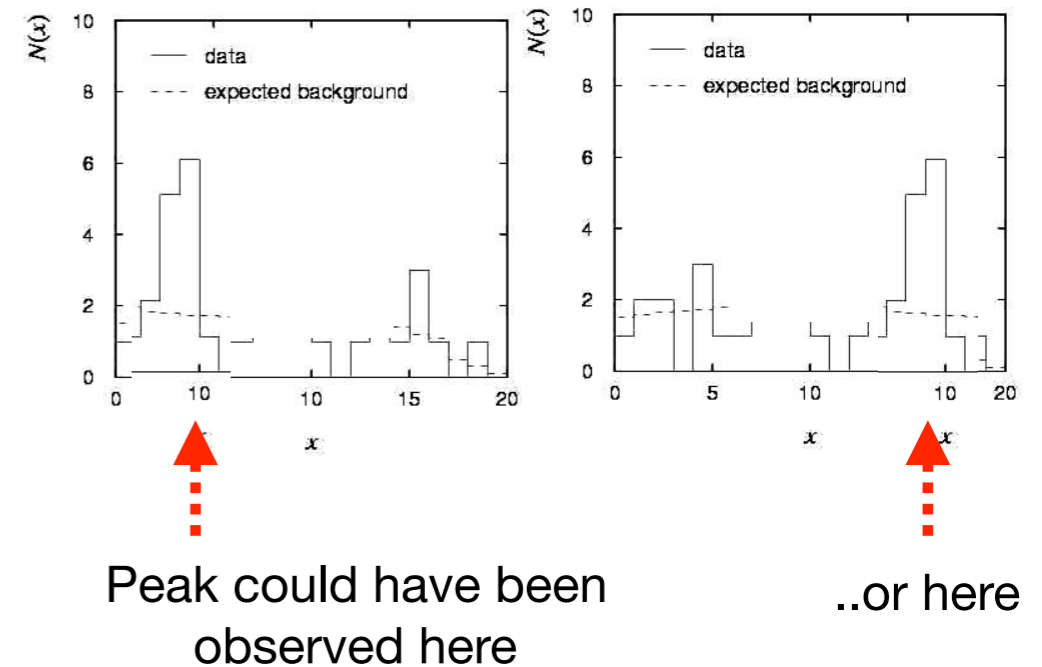
“Look elsewhere” effect

It does not account for the chances that an excess could have arisen in any pair of adjacent bins. With 20 bins (10 pairs of adjacent bins) the p-value gets multiplied by ≈ 10 .

Lots of bins, lots of chances at fluctuations.

Need to correct for the **effect of multiple testing** (i.e., need to account that we are also “looking elsewhere” from where we see an anomaly).

When quoting p-values, need to account for the size of the test space. The larger the size, the higher the probabilities to observe rare fluctuations. Otherwise significances may be grossly overestimated.



Short aside (not LHC stuff)

The discovery of the Ooops-Leon particle

Leon

Leon Lederman is a living legend. In the HEP golden age of the '60-'70 he did many of the key experiments that laid the foundations of the Standard Model. In 1988, he got the Nobel prize in physics for the discovery of the muon neutrino.



In 1976, his group announced the observation of a new particle produced by a beam of protons on Beryllium and decaying into $e^+ e^-$ pairs, with a mass of about 6 GeV.

Observation of High-Mass Dilepton Pairs in Hadron Collisions at 400 GeV

D. C. Hom, L. M. Lederman, H. P. Paar, H. D. Snyder, J. M. Weiss, and J. K. Yoh
*Columbia University, New York, New York 10027**

and

J. A. Appel, B. C. Brown, C. N. Brown, W. R. Innes, and T. Yamanouchi
Fermi National Accelerator Laboratory, Batavia, Illinois 60510†

and

D. M. Kaplan
*State University of New York at Stony Brook, Stony Brook, New York 11794**
(Received 28 January 1976)

We report preliminary results on the production of electron-positron pairs in the mass range 2.5 to 20 GeV in 400-GeV p -Be interactions. 27 high-mass events are observed in the mass range 5.5–10.0 GeV corresponding to $\sigma = (1.2 \pm 0.5) \times 10^{-36}$ cm² per nucleon. Clustering of 12 of these events between 5.8 and 6.2 GeV suggests that the data contain a new resonance at 6 GeV.

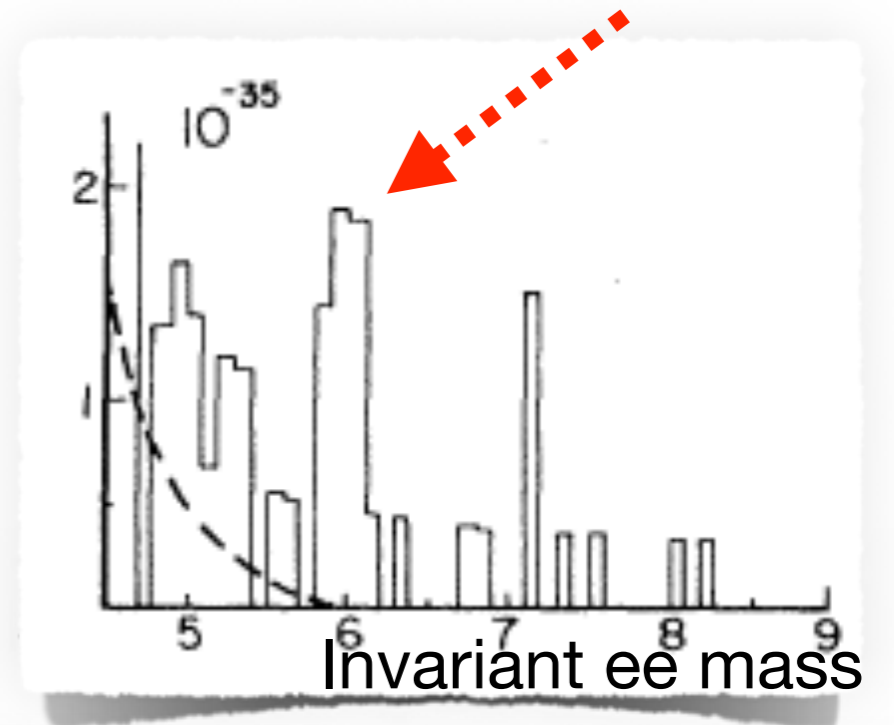
The “Oops-Leon” particle

This was published and provided a very strong candidate for the Upsilon, the bound state of a (then still unobserved) fifth quark.

More data did not confirm the finding.

[Embarrassment...]

The erroneous first claim has been later tracked down to a mistake in the statistical evaluation of the significance of the signal, which did not properly account for the LEE.



a linear A dependence.⁷ We have studied the probability for a clustering of events as is observed here to result from a fluctuation in a smooth distribution, e.g., Eq. (3). To avoid the difficult problems involved in the statistical theory associated with small numbers of events per resolution bin, a Monte Carlo method was used. Histograms were generated by throwing events according to a variety of smooth distributions, modulated by the mass acceptance, over the mass range 5.0 to 10.0 GeV. Clusters of events as observed occurring anywhere from 5.5 to 10.0 GeV appeared less than 2% of the time.⁸ Thus the statistical case for a narrow (< 100 MeV) resonance is strong although we are aware of the need for confirmation. These data, at a level of

PS

A couple of years later, the same group found the real Upsilon meson, at 9.5 GeV using muon pairs and nobody cared too much about the 6 GeV fluke, which someone dubbed “Oops-Leon” in a pun over Lederman’s name.



Observation of a Dimuon Resonance at 9.5 GeV in 400-GeV Proton-Nucleus Collisions

S. W. Herb, D. C. Hom, L. M. Lederman, J. C. Sens,^(a) H. D. Snyder, and J. K. Yoh
Columbia University, New York, New York 10027

and

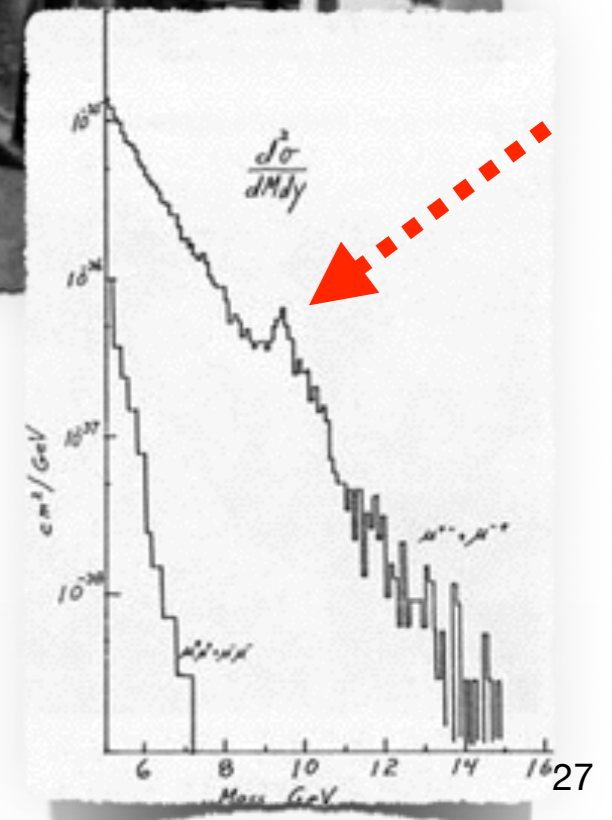
J. A. Appel, B. C. Brown, C. N. Brown, W. R. Innes, K. Ueno, and T. Yamanouchi
Fermi National Accelerator Laboratory, Batavia, Illinois 60510

and

A. S. Ito, H. Jöstlein, D. M. Kaplan, and R. D. Kephart
State University of New York at Stony Brook, Stony Brook, New York 11974
(Received 1 July 1977)

Accepted without review at the request of Edwin L. Goldwasser under policy announced 26 April 1976

Dimuon production is studied in 400-GeV proton-nucleus collisions. A strong enhancement is observed at 9.5 GeV mass in a sample of 9000 dimuon events with a mass $m_{\mu^+\mu^-} > 5$ GeV.



How to deal with the effect of multiple testing

There are various semi-empiric recipes to determine an LEE-correction starting from a “local” p-value: width/resolution, Bonferroni, Dunn-Sidak, Gross-Vitells,

Most of these are only useful to provide a semiquantitative feel of the severity of the effect in simple cases. Cannot be applied to more complex analyses, where the final p-value is the result of a combination of analyses in various channels, each contributing different weight and with different experimental resolutions.

Ideal solution: a p-value of p-values, that is use the p-values as test statistics and look at the distribution of smallest p-values. Next to impossible (I don't think has ever been used in a realistic HEP analysis).

Note: interesting tradeoff between making your search as much broad as possible (thus increasing the chances to find something) but not so broad that the LEE spoils all sensitivity. Can an optimization be explored?

Where is “elsewhere”?

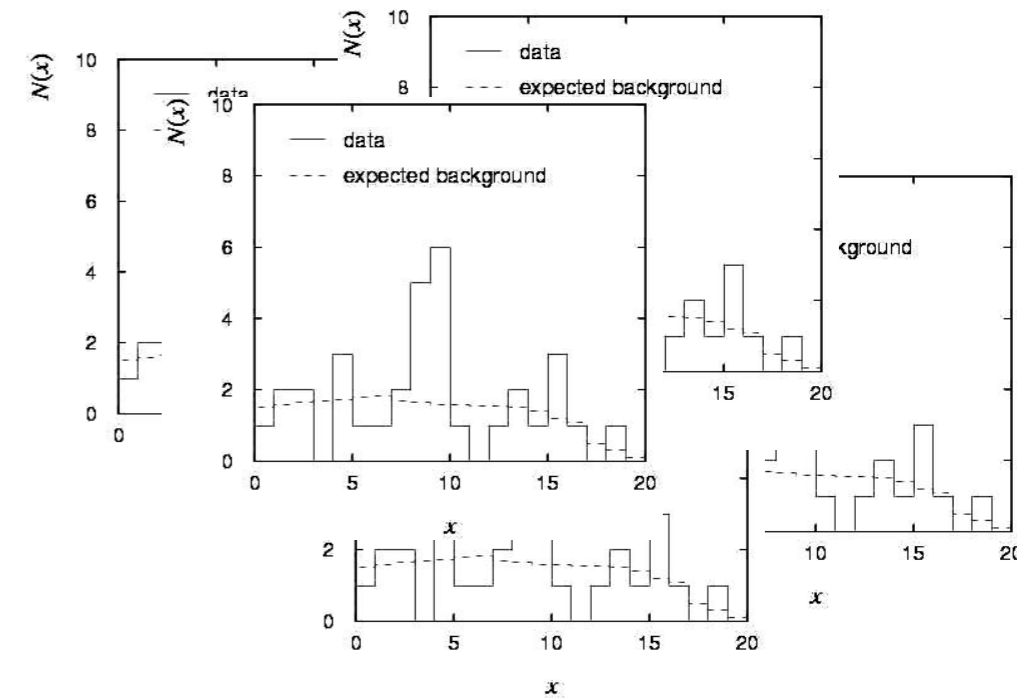
Also, it does not account for the possibility that tenths, or hundreds, or thousands of distributions may have been inspected, in the same analysis or in other analyses.

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?

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”

Forgetting about the LEE is hardly the only mistake one can do.

“Tuning” of the data selection to artificially enhance signal-like statistical fluctuations is a serious threat if the selection isn’t frozen before looking at the signal regions

Forgetting to include systematic uncertainties in p-value determinations could lead to false positives. Null should be rejected if the p-value is sufficiently small **for all allowed** values of nuisance parameters.

HEP folks conventionally agreed to deal collectively with these possible pitfalls by setting a rather high 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 http://www.huffingtonpost.com/victor-stenger/higgs-and-significance_b_1649808.html for an historical recollection)

Loose rationale: such high thresholds should protect from the effects above.

Still....

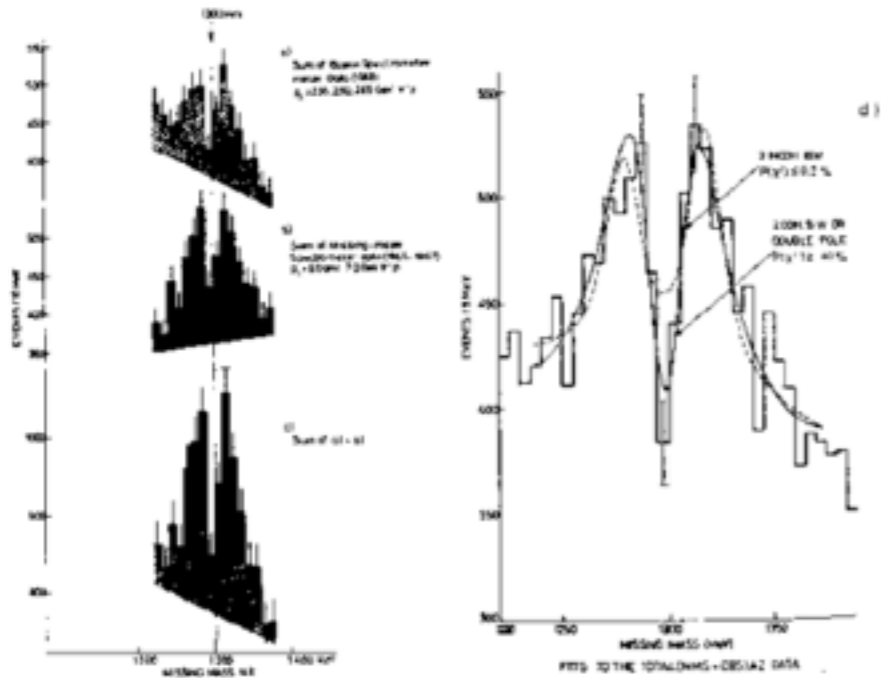
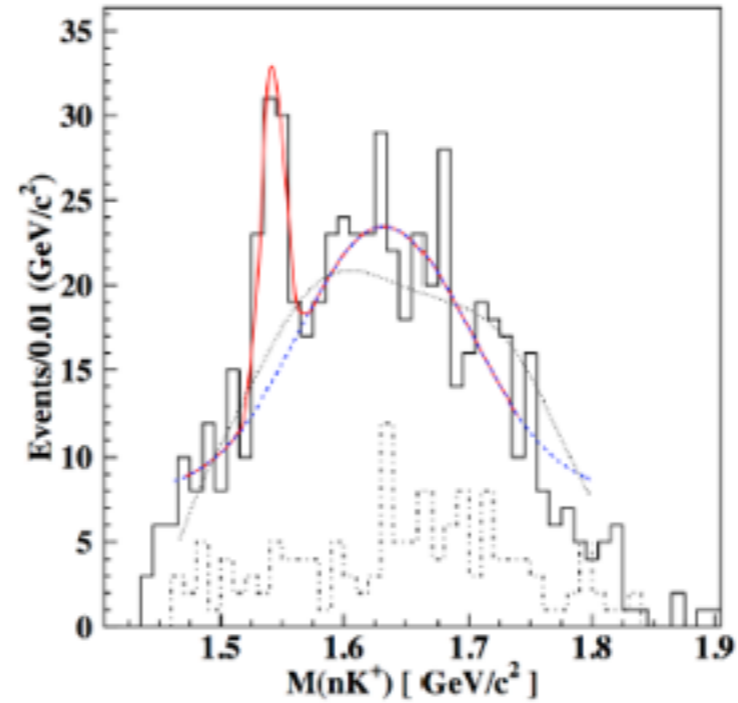
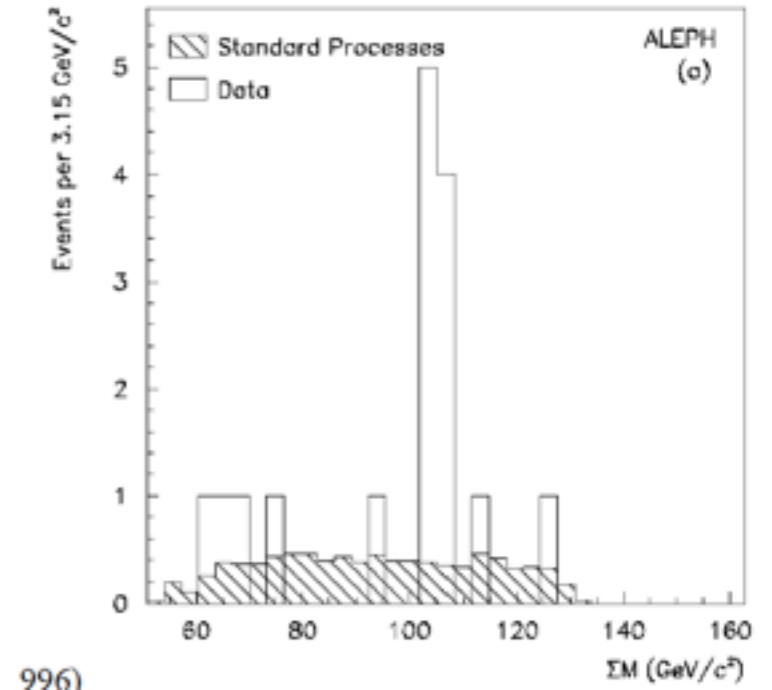


Figure 3: (a-c) Evidence for A_2 splitting in $\pi^- p \rightarrow p X^-$ collisions in the two CERN experiments. (d) same as (c) in 5 MeV bins fit to two hypotheses.



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

Significance = $5.2 \pm 0.6 \sigma$



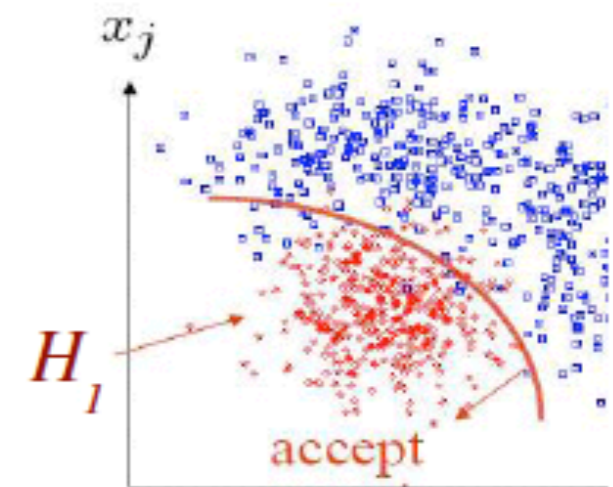
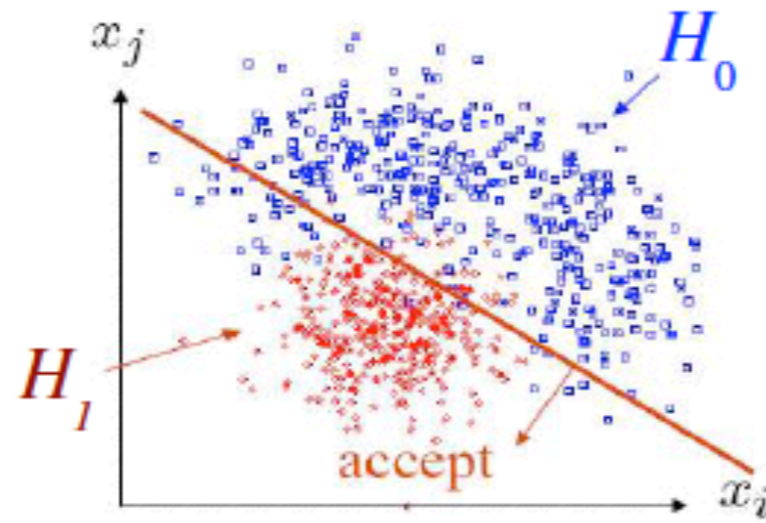
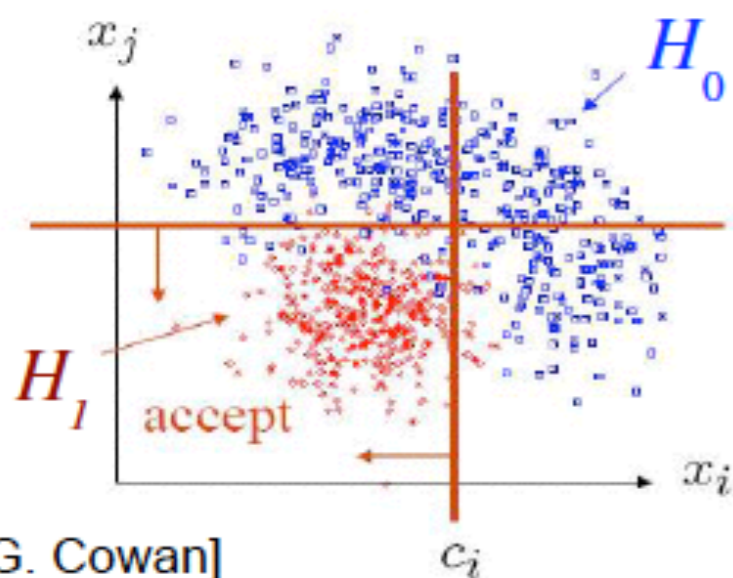
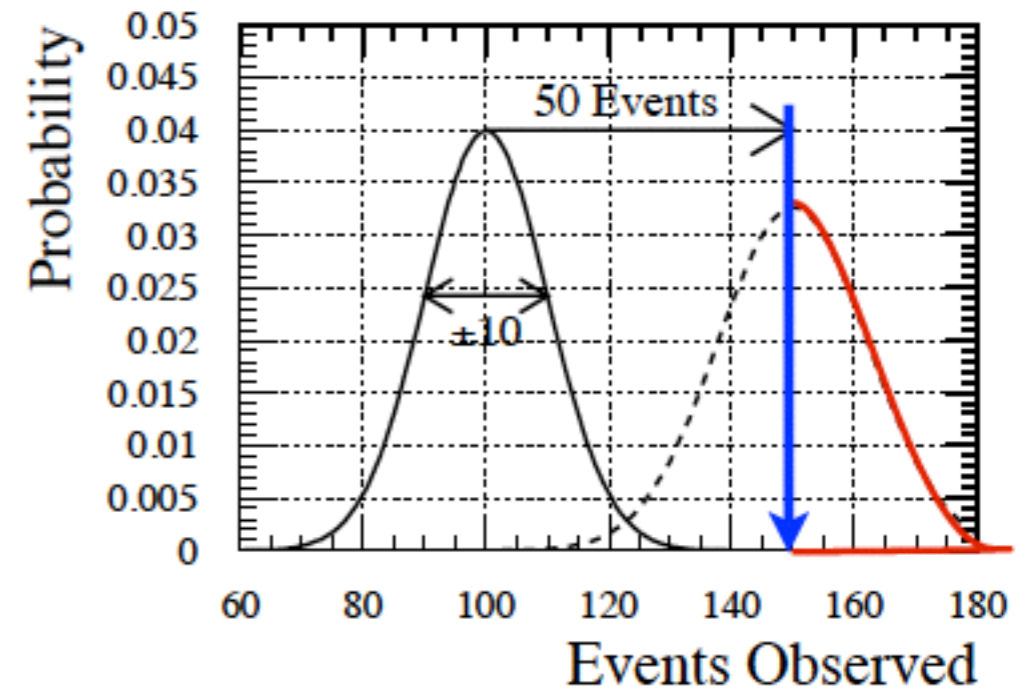
ALEPH collaboration,
CERN mid 90ies.

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

Which function of the observables x to choose?

Back to p-values. Arbitrariness in choosing the test quantity x . Need to find a function of the observables x that maximizes the power of my tests at fixed false-positive rate. Pretty obvious in simple counting experiments

Less obvious in multiple-dimensional nonlinear problems



[G. Cowan]

Neyman-Pearson Lemma

It can be demonstrated that such variable exist and it is the likelihood ratio (again!)

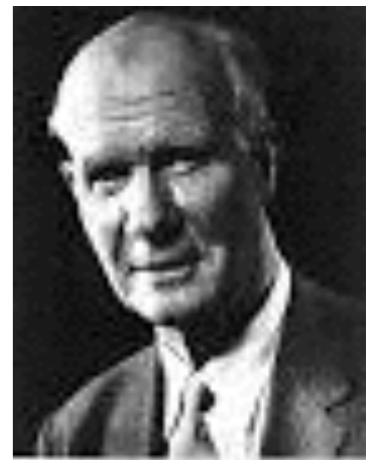
The region W of acceptance of the null which minimises the probability to accept the null when the signal hypothesis holds is a contour of the likelihood ratio

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

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



Jerzy Neyman
(1894-1981)

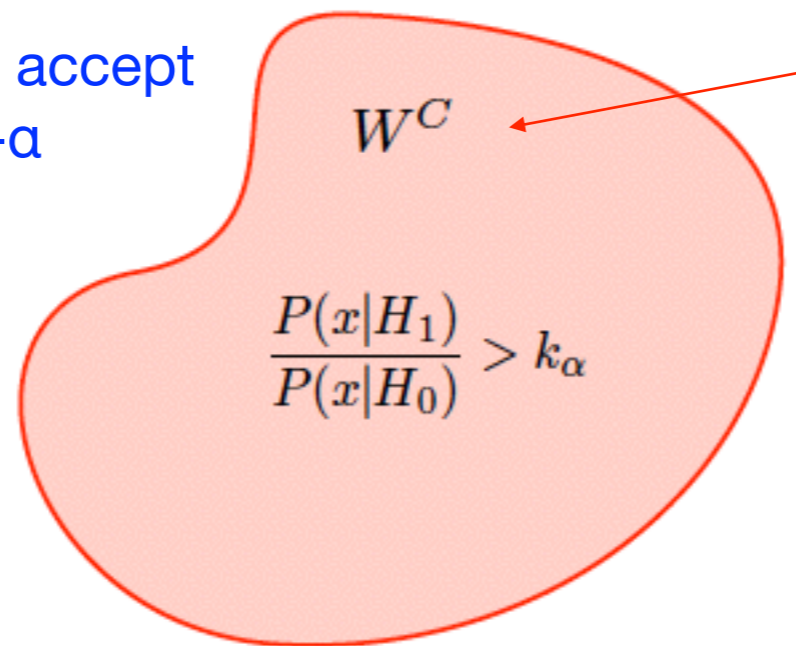


Egon S. Pearson
(1885-1980)

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 H_0

Region W : if data fall here we accept H_0 ; probability under H_0 is $1-\alpha$

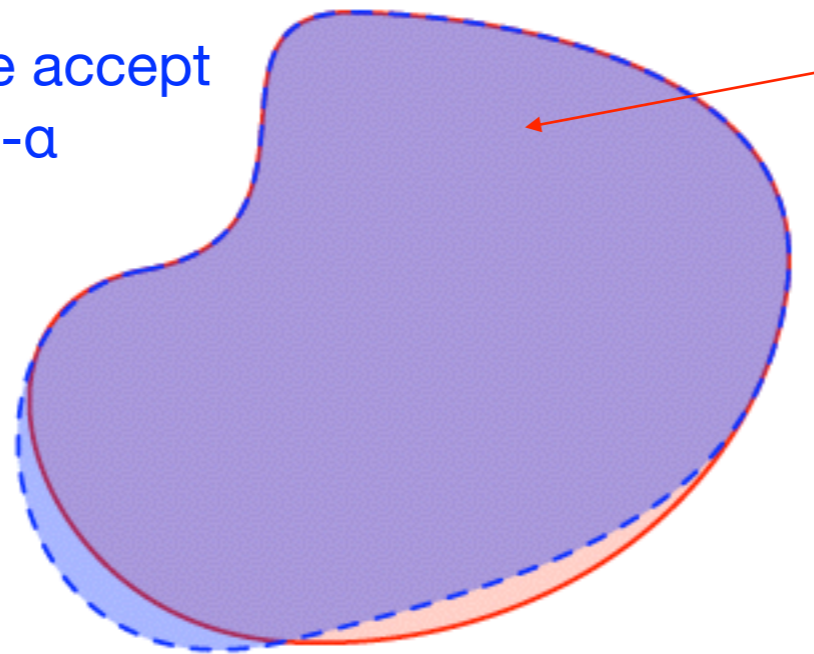


Region W^c : if data fall there we reject H_0 ; probability under H_0 is α

NP-lemma illustration

Take a variation that has the same rate α of false positives (same probability under H_0)

Region W : if data fall here we accept H_0 ; probability under H_0 is $1-\alpha$



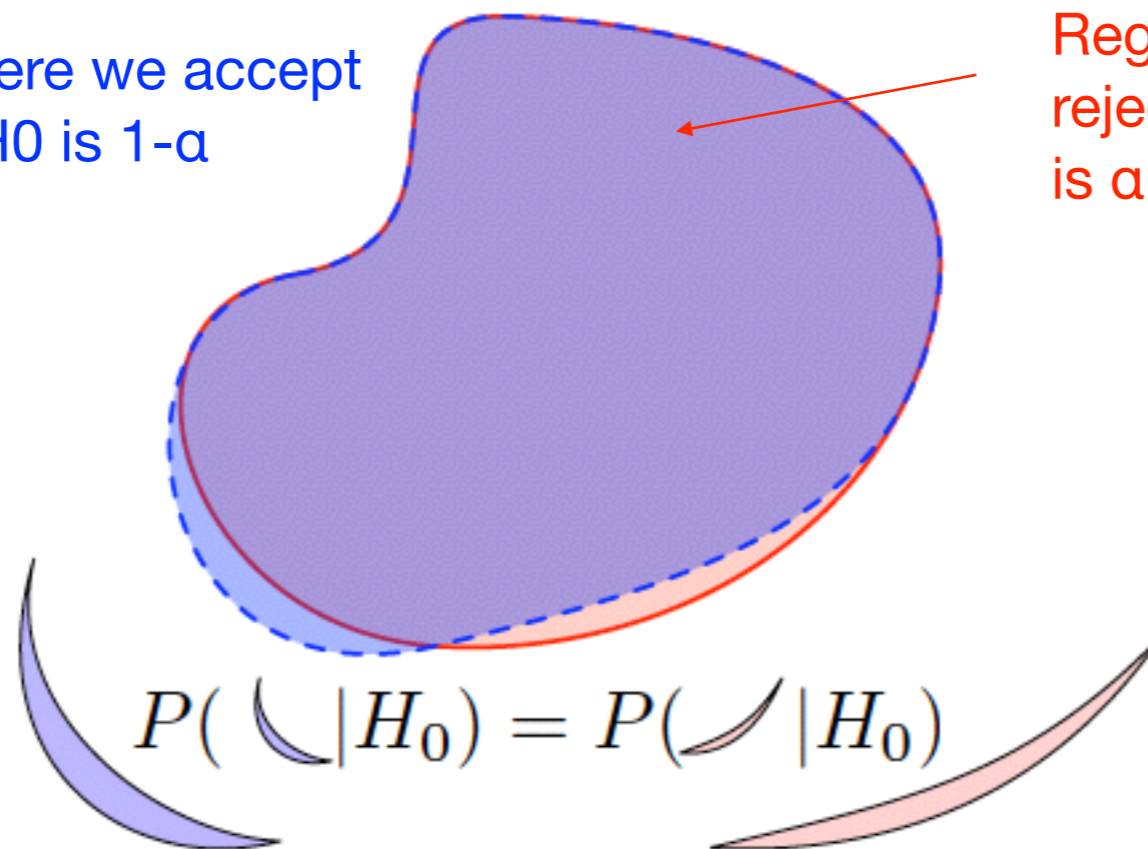
Region W^c : if data fall there we reject H_0 ; probability under H_0 is α

NP-lemma illustration

Take a variation that has the same rate α of false positives (same probability under H_0)

Region W : if data fall here we accept H_0 ; probability under H_0 is $1-\alpha$

Region W^c : if data fall there we reject H_0 ; probability under H_0 is α

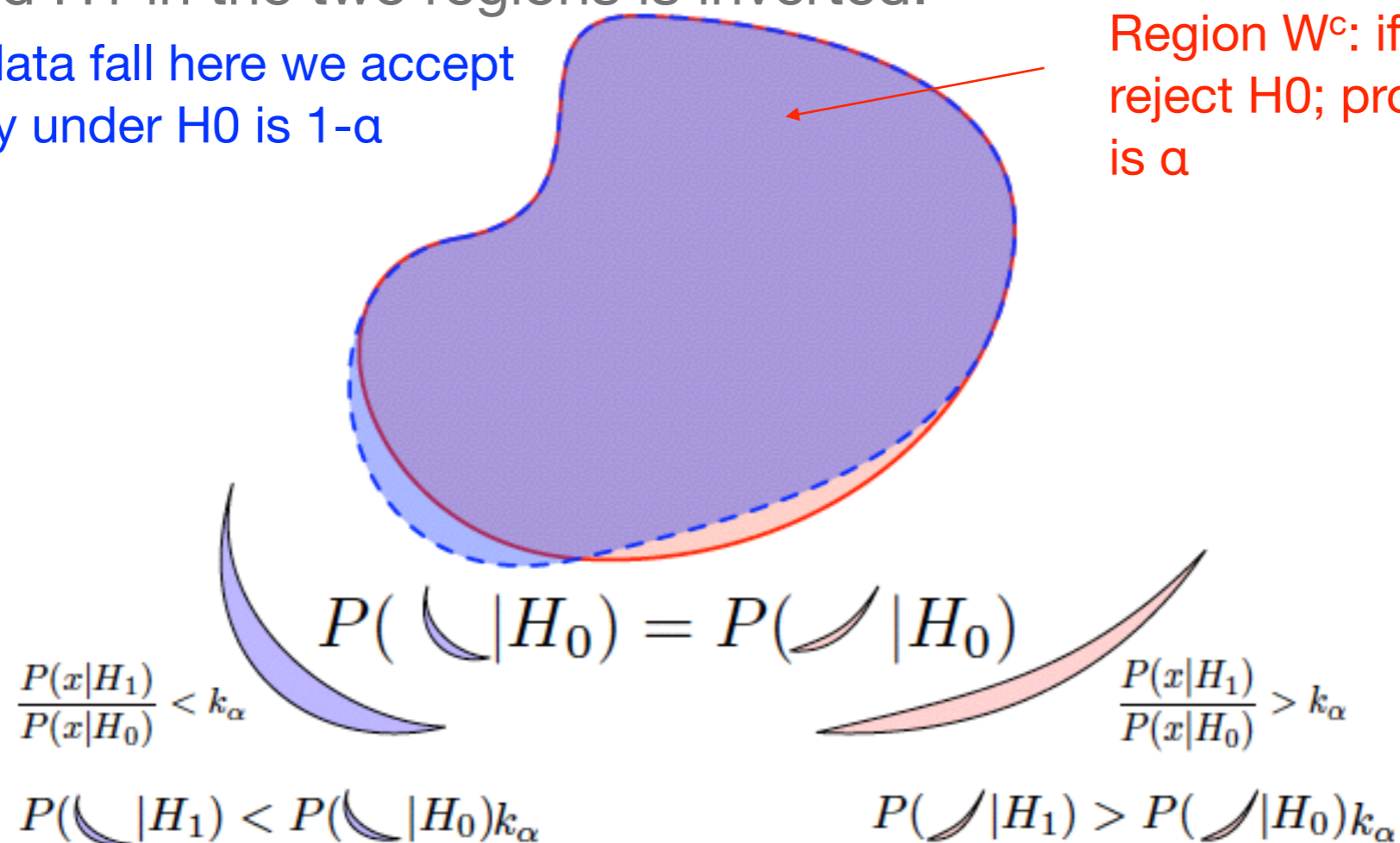


NP-lemma illustration

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 H_0 and H_1 in the two regions is inverted.

Region W : if data fall here we accept H_0 ; probability under H_0 is $1-\alpha$

Region W^c : if data fall there we reject H_0 ; probability under H_0 is α



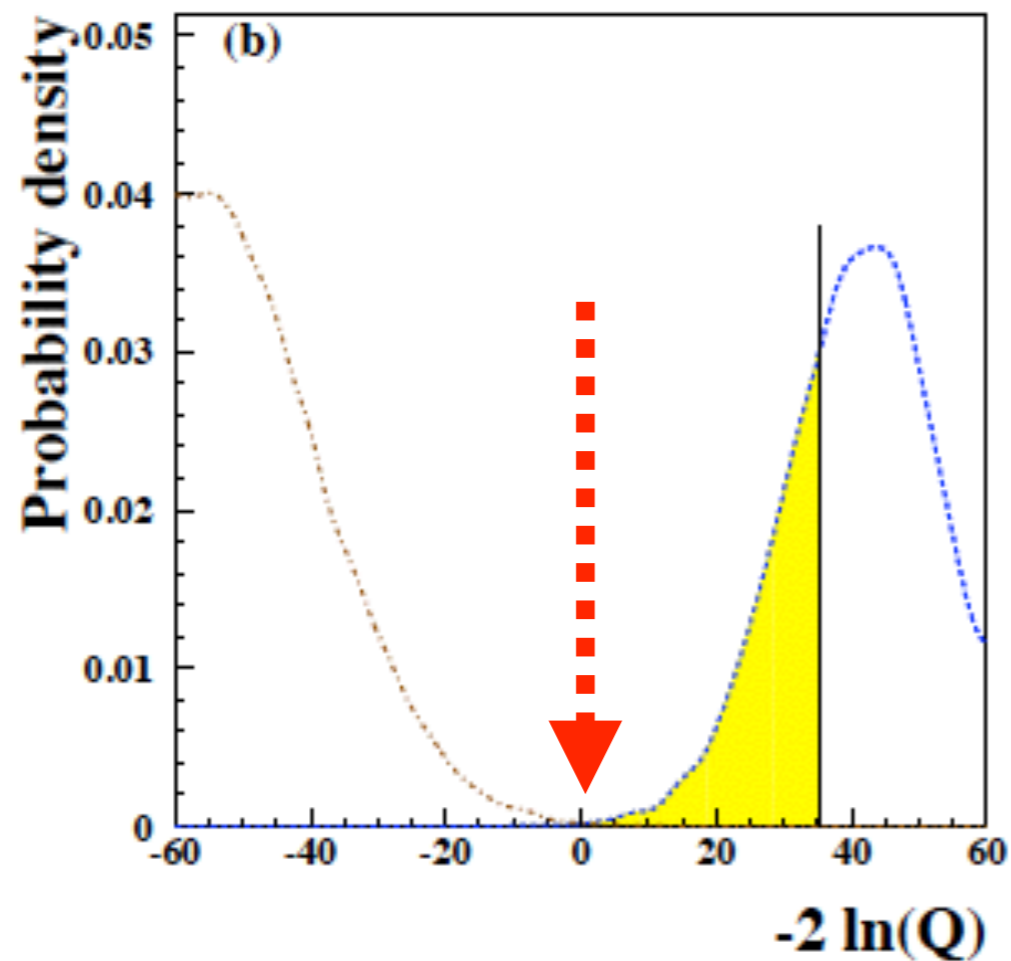
$$P(\cup | H_1) < P(\cup | H_1)$$

The new region region has less power.

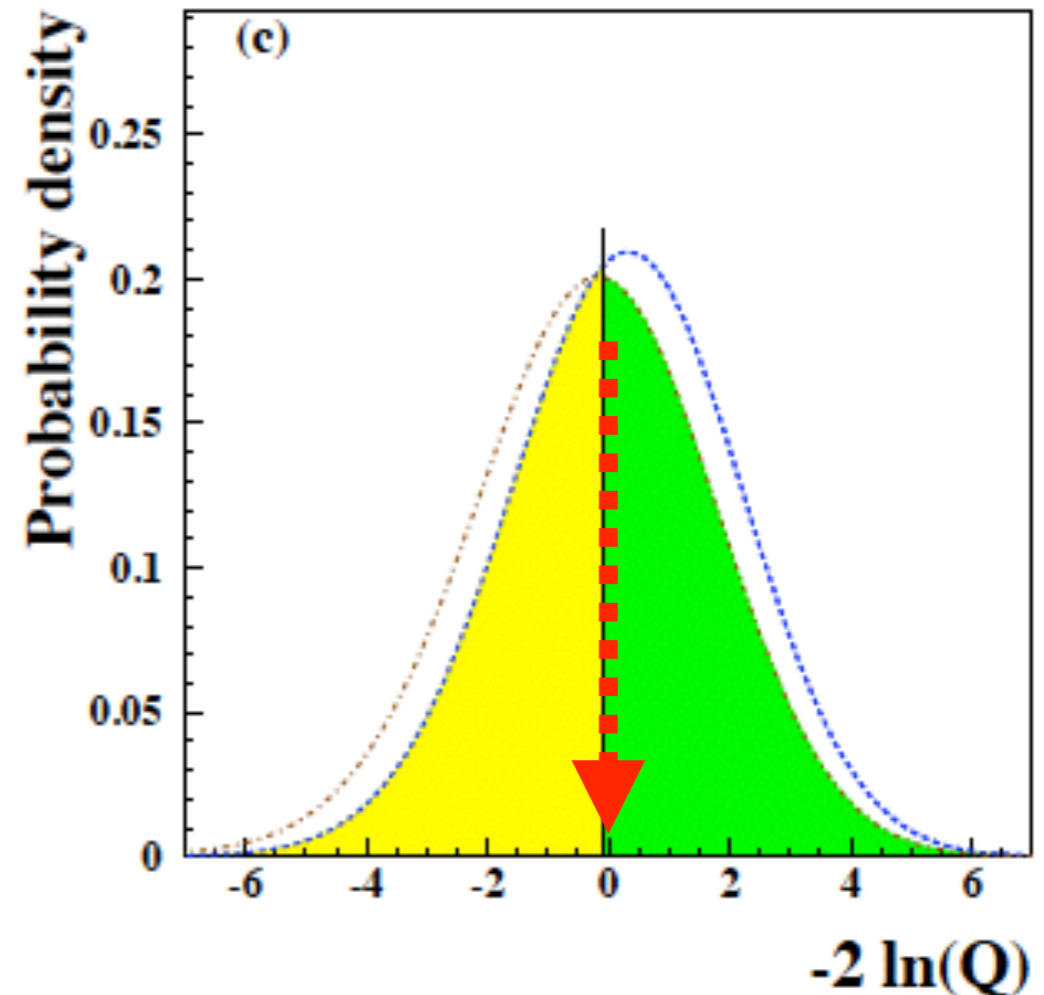
Likelihood ratio

In most of the limits and measurements made at the LHC, the test statistic of choice is the likelihood ratio, that is the ratio between the likelihood of the data under the null hypothesis and the likelihood of the data under the signal hypothesis.

Issues with p-values



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

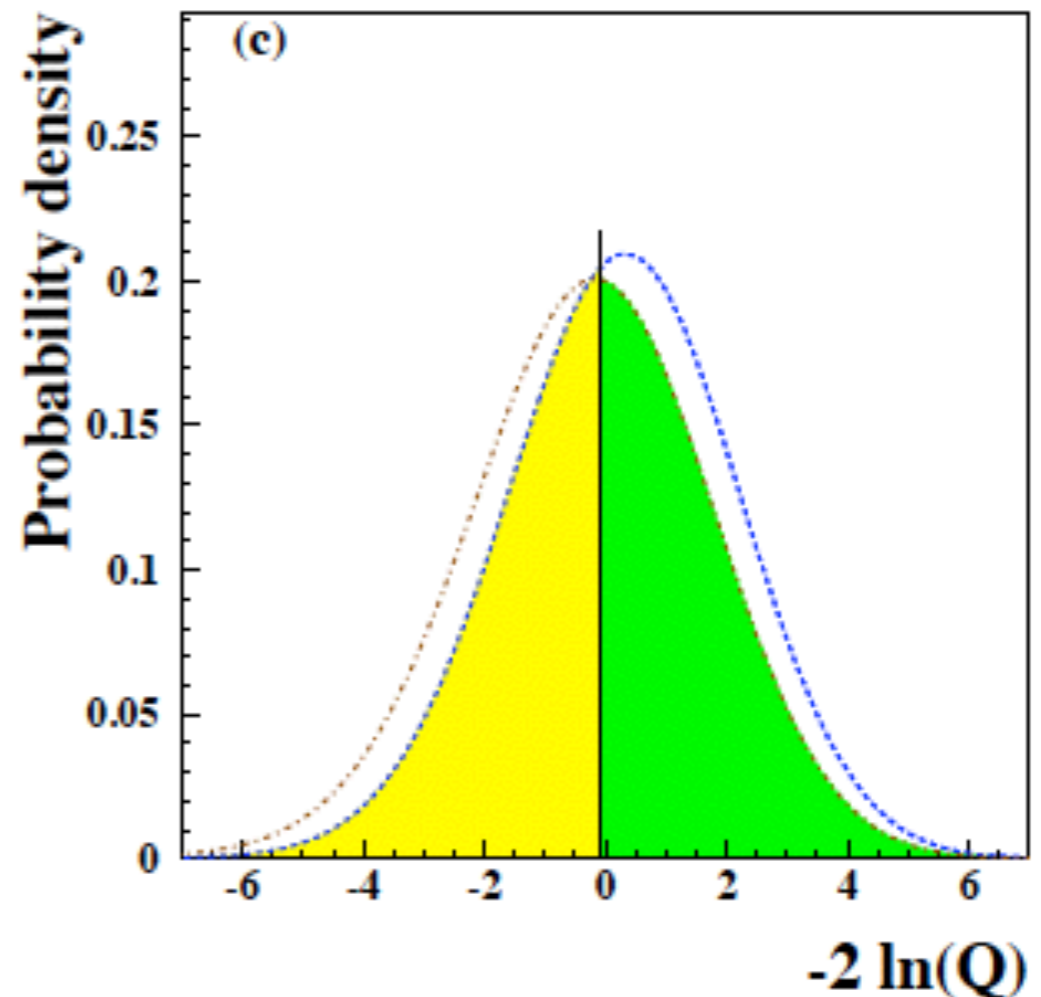


Can make no statement about the signal regardless of the outcome

Sensitivity

In searches for small signals with poor signal-to-background separation, the analysis has hard times in telling apart the possible presence of signal from the fluctuations of the background.

Analysis sensitivity is poor, implying that the distributions of the test statistics are similar for the hypotheses of signal +background and background only.

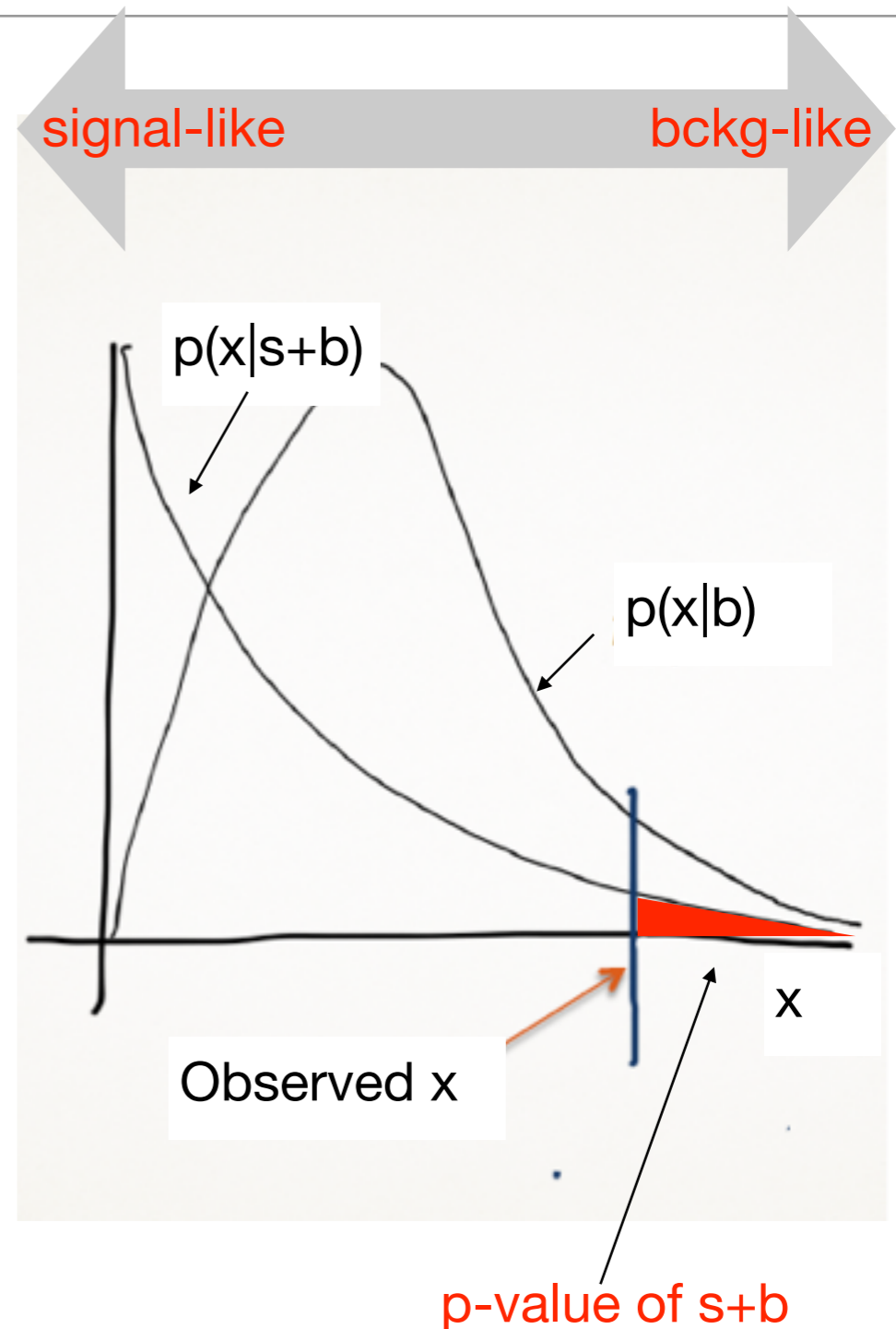


The problem of spurious exclusion

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.



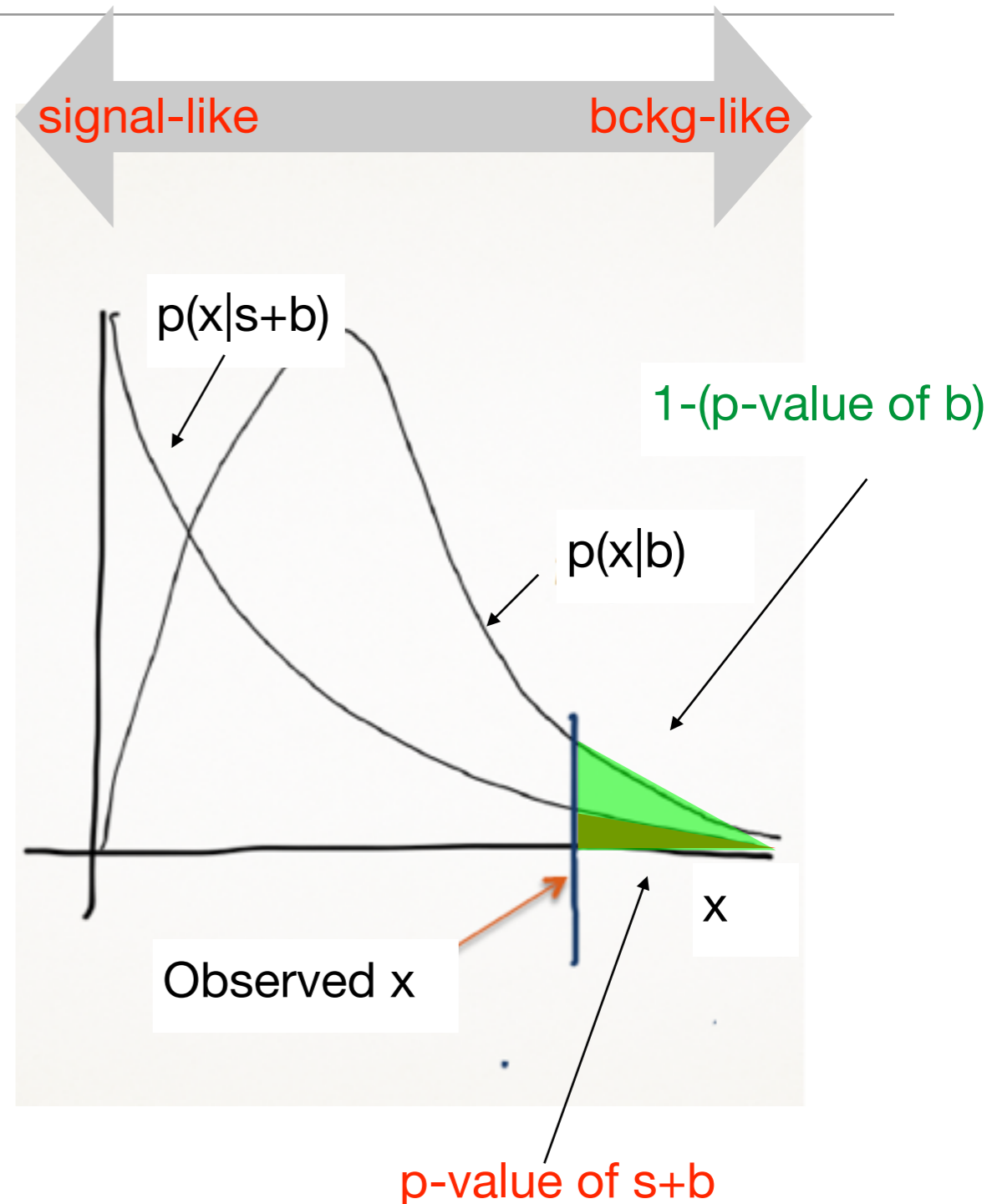
Spurious exclusion

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 $s+b$.



CLs

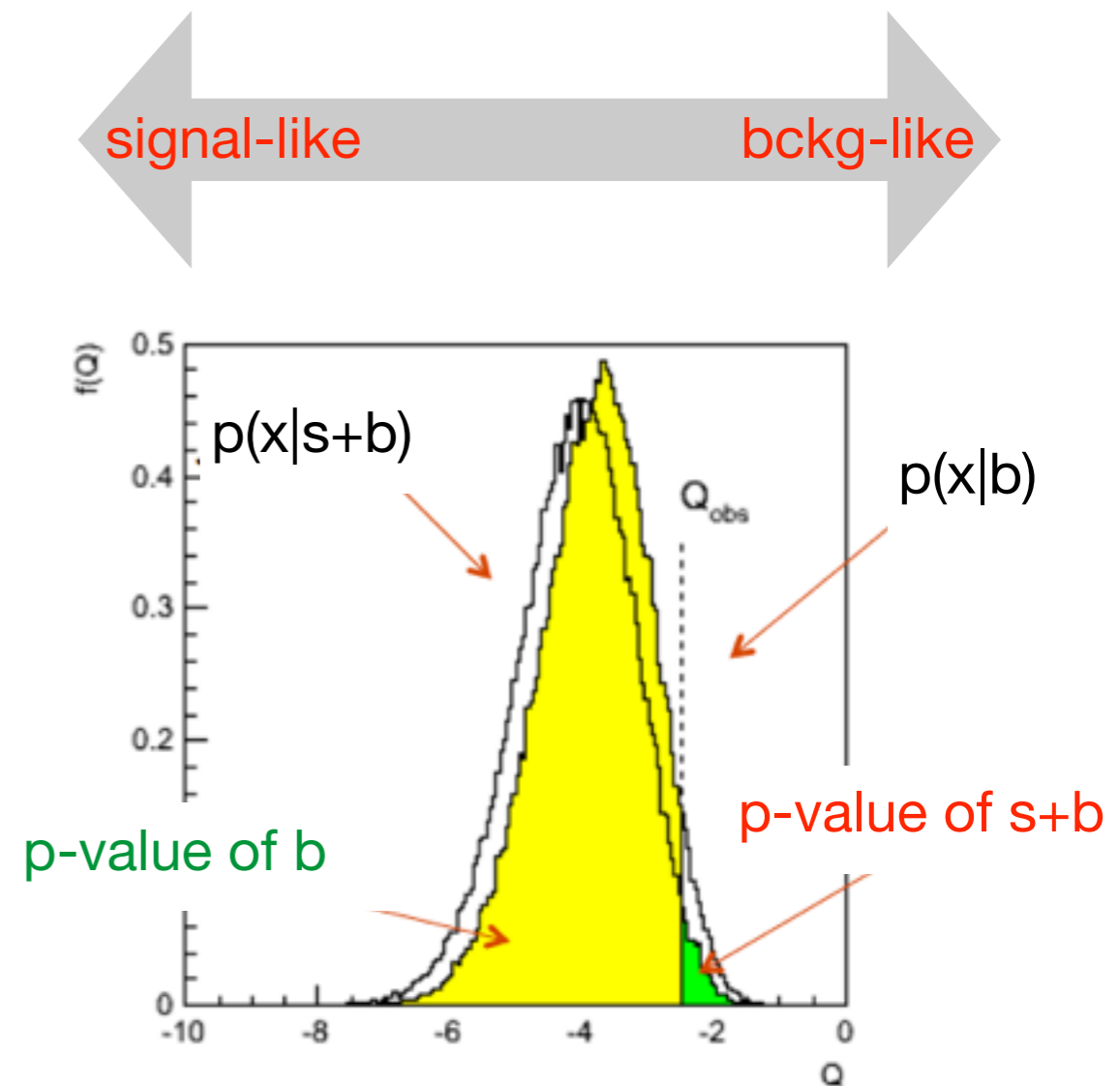
Modified p-value with no rigorous statistical foundations. “Works” fairly well allows for treating simultaneously exclusion and discovery and prevents from excluding hypotheses to which there is no sensitivity.

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

$$\text{CLs} = [\text{pvalue for } s+b] / [1 - \text{pvalue of } b]$$

is small. Denominator increases CLs if pvalue is small thus preventing excluding signal to which there is no sensitivity.

Inspired by similar methods (Zech, Roe&Woodroffe) developed for counting experiments.



A Poisson example

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

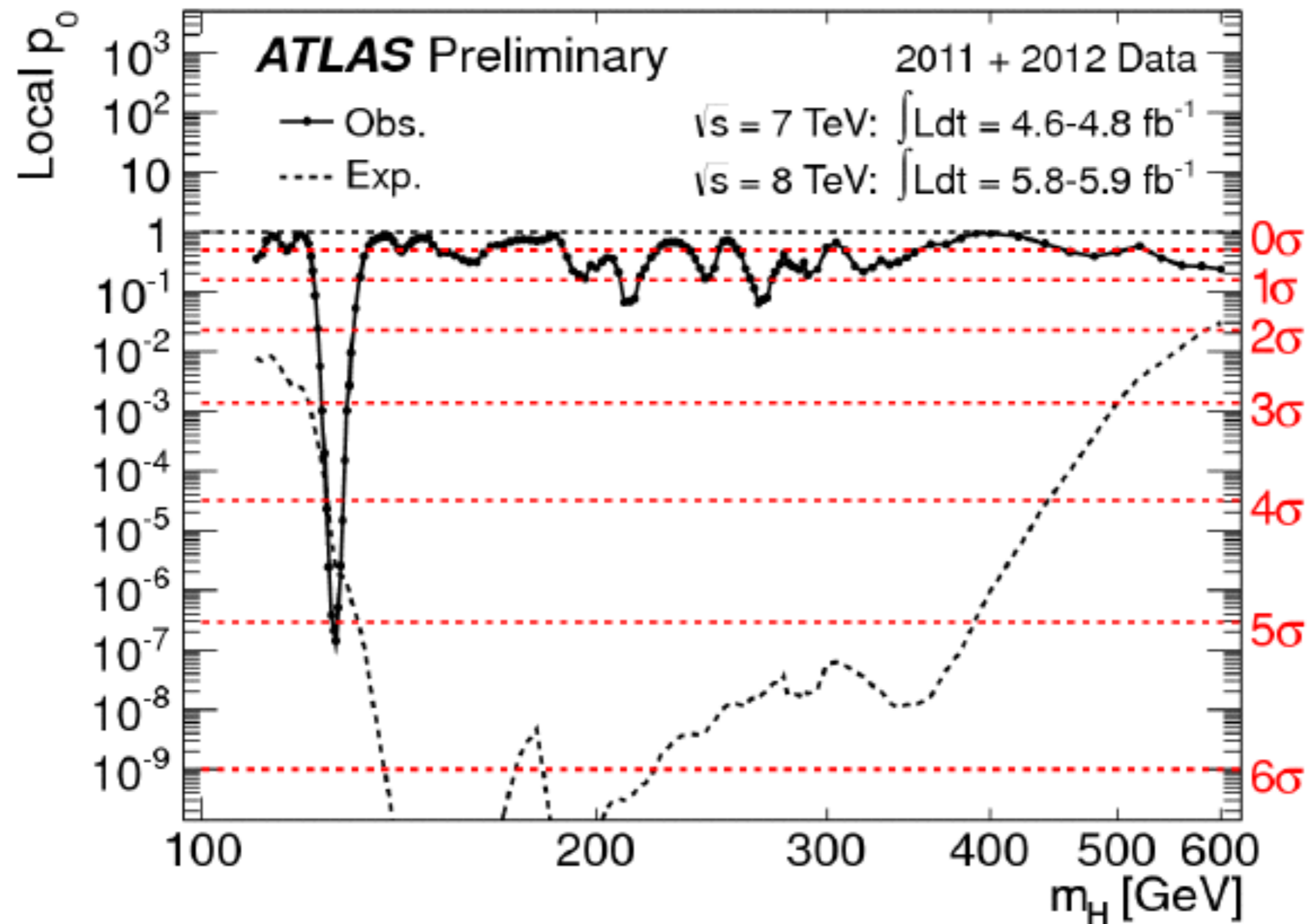
- Suppose $\langle n_b \rangle = 100$
- $s(m_{H1}) = 30$
- Suppose $n_{obs} = 102$
- $s+b = 130$
- $\text{Prob}(n_{obs} \leq 102 \mid 130) < 5\%$, m_{H1} is excluded at $>95\%$ CL
- Now suppose $s(m_{H2}) = 1$, can we exclude m_{H2} ?
- If $n_{obs} = 102$, obviously we cannot exclude m_{H2}
- Now suppose $n_{obs} = 80$, $\text{prob}(n_{obs} \leq 80 \mid 101) < 5\%$, we looks like we can exclude $m_{H2} \dots$ but this is dangerous, because what we exclude is $(s(m_{H2})+b)$ and not $s \dots \dots$
- With this logic we could also exclude b (expected $b = 100$)
- To protect we calculate a modofloed p-value $\frac{\text{Pr ob}(nobs \leq 80 \mid 101)}{\text{Pr ob}(nobs \leq 80 \mid 100)} \sim 1$
- We cannot exclude m_{H2}

Gauging the sensitivity of an analysis

Asimov approximations for median sensitivities

Blind analyses

So, now you should be able to understand this



Neyman construction

J. Neyman came up with a mathematically rigorous and very elegant procedure that allows constructing confidence intervals with the desired level of coverage



Jerzy Neyman (1894-1981)

X—Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability

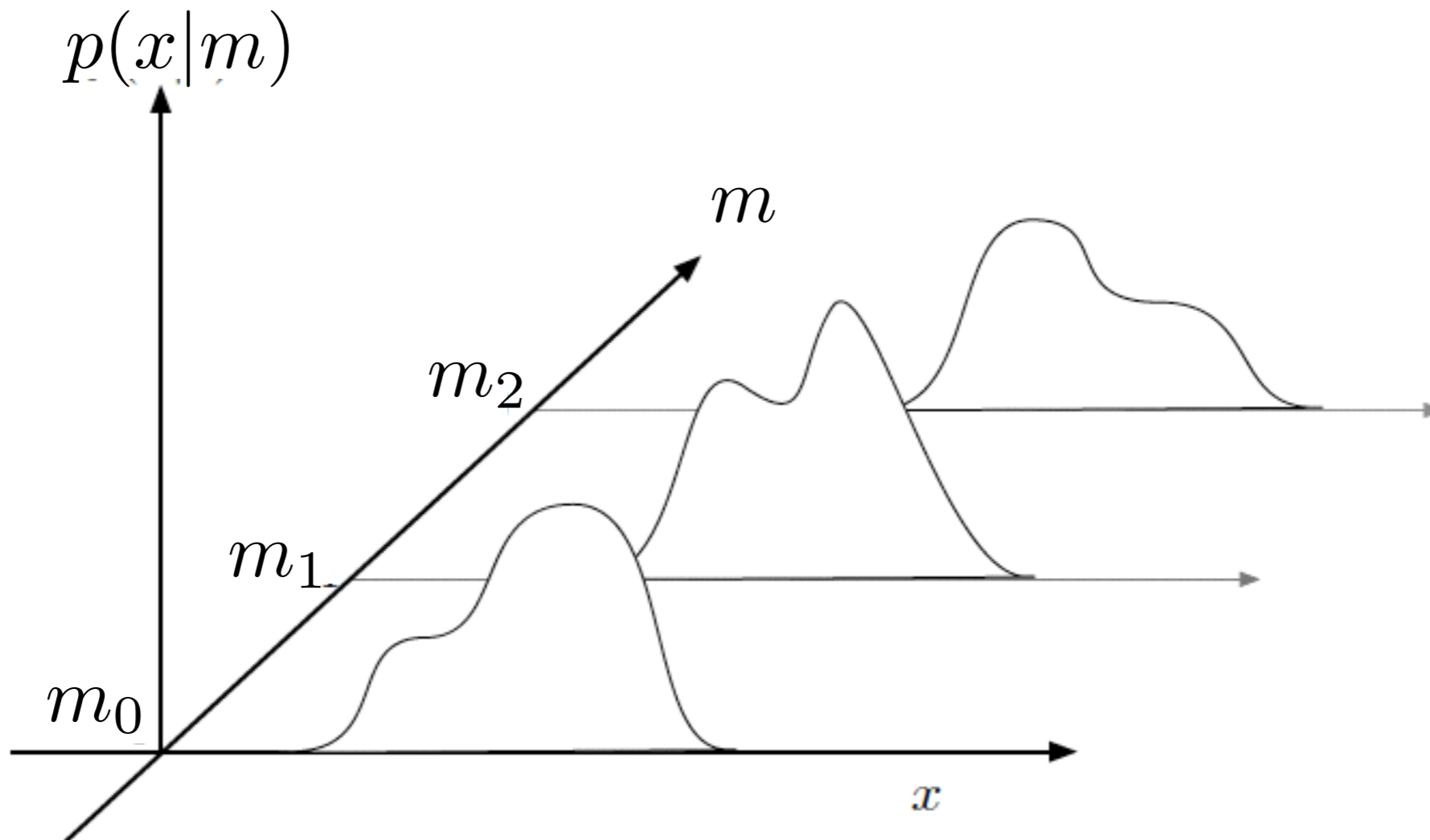
By J. NEYMAN

Reader in Statistics, University College, London

(Communicated by H. JEFFREYS, F.R.S.—Received 20 November, 1936—Read 17 June, 1937)

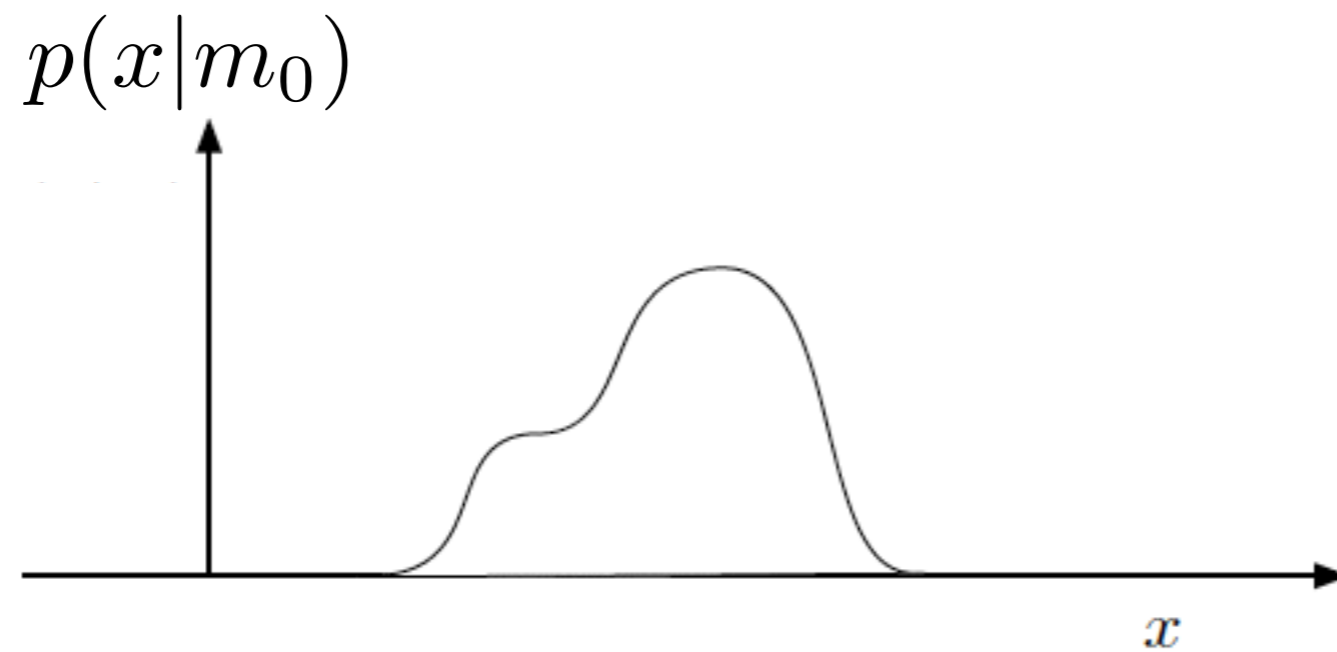
Neyman construction illustrated

For each possible true value of parameter m , consider $p(x|m)$. Its shape can vary as a function of m .



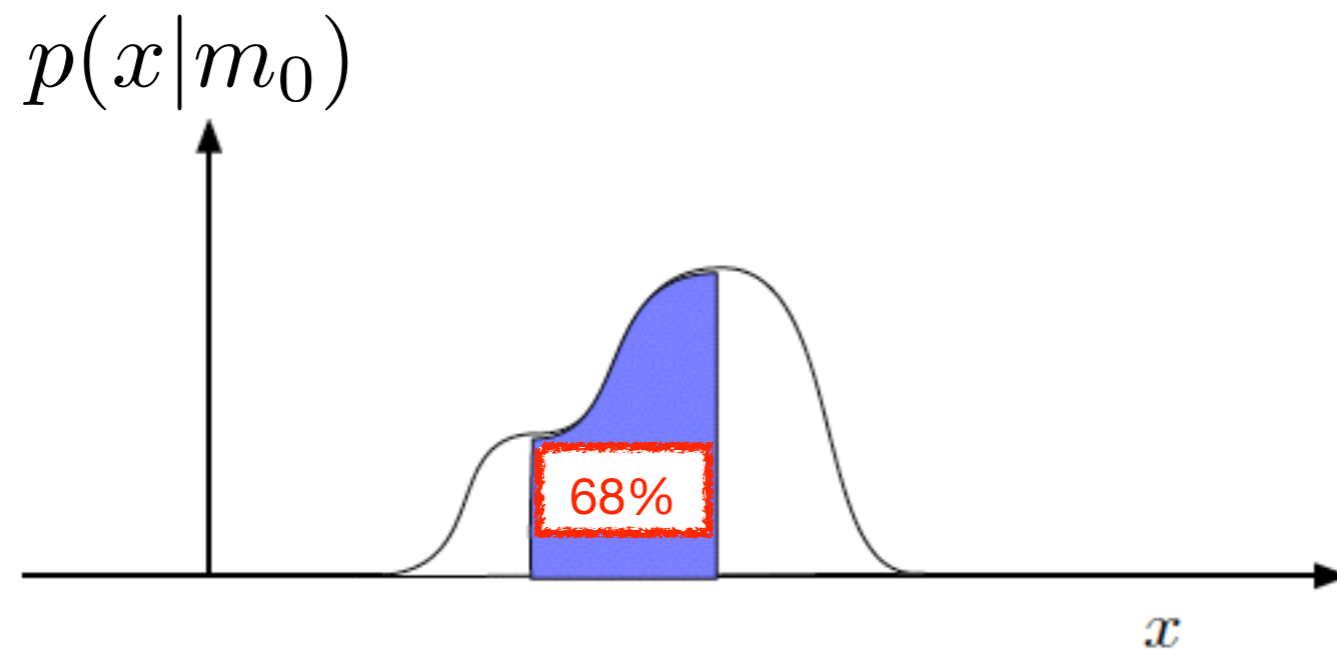
Neyman illustrated I

Take a specific value m_0 of the parameter



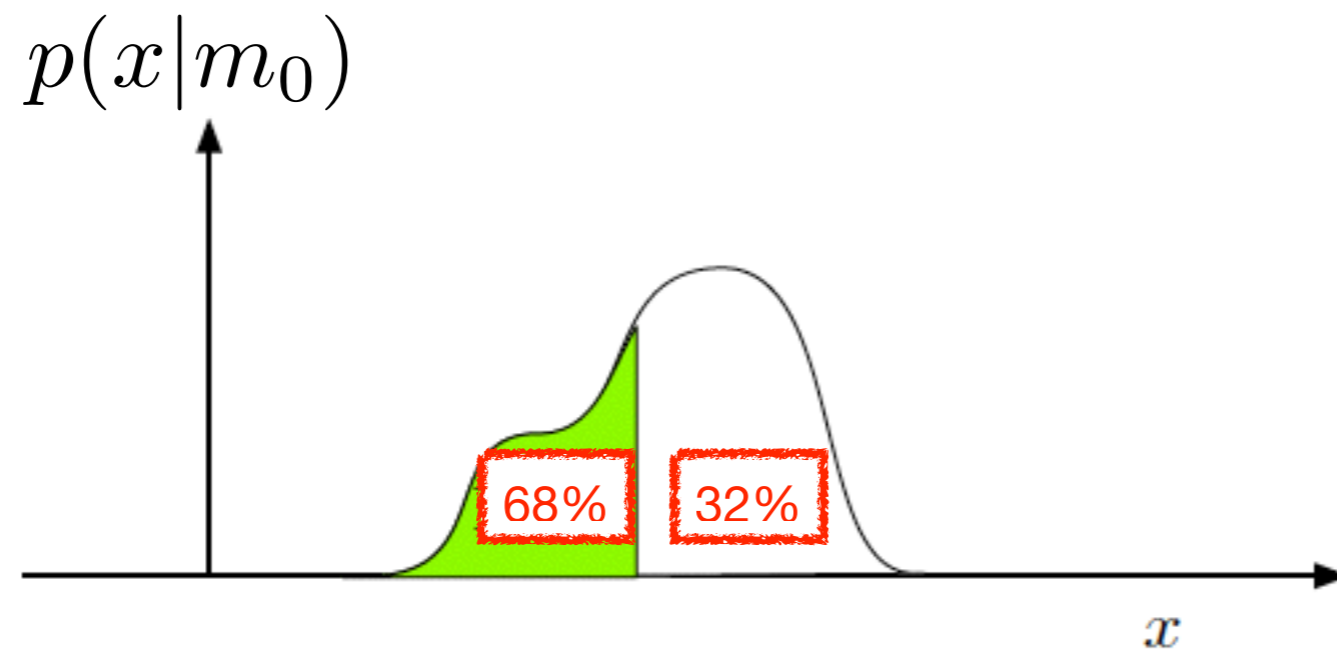
Neyman illustrated II

Define, using the $p(x|m_0)$ associated to that parameter, an acceptance range in x , such that $p(x \in \text{range} \mid m_0) = 68\%$.



Neyman illustrated III

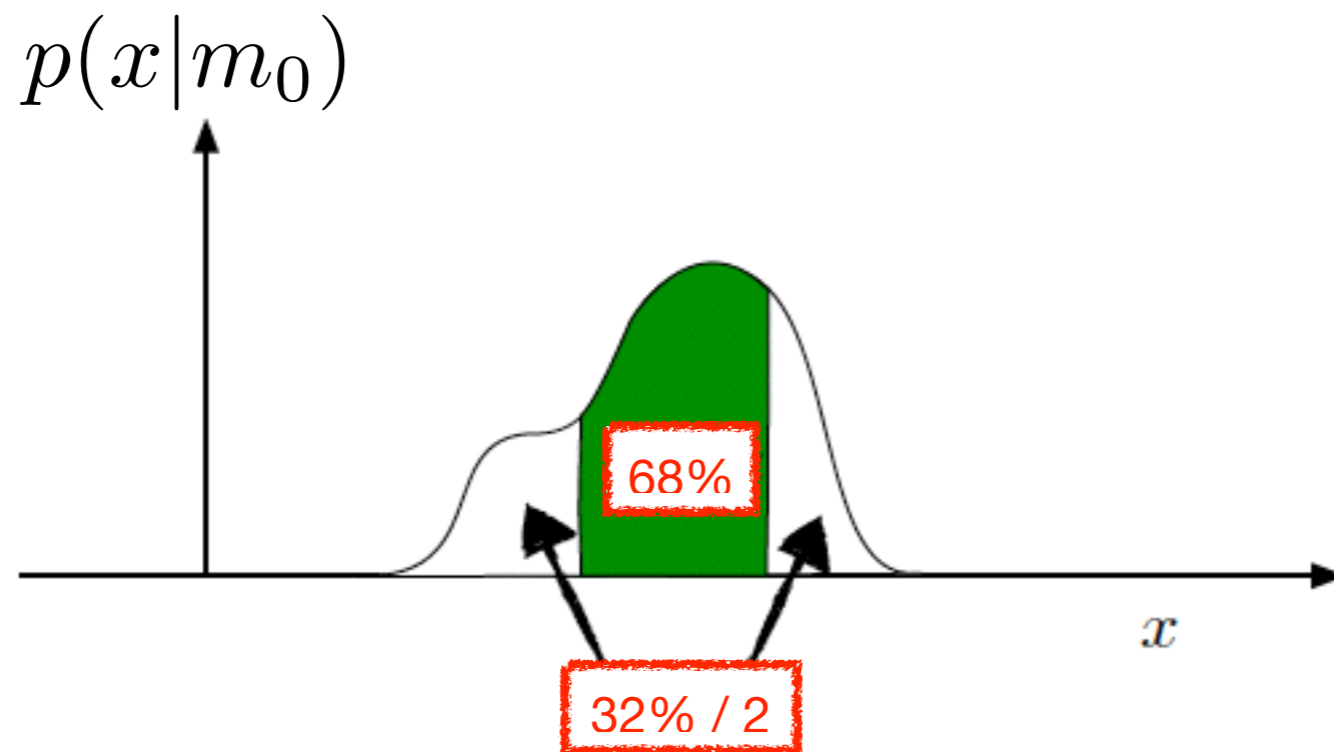
Note that such region is not unique. I could also have chosen to put an upper limit at 68% CL...



Neyman illustrated VI

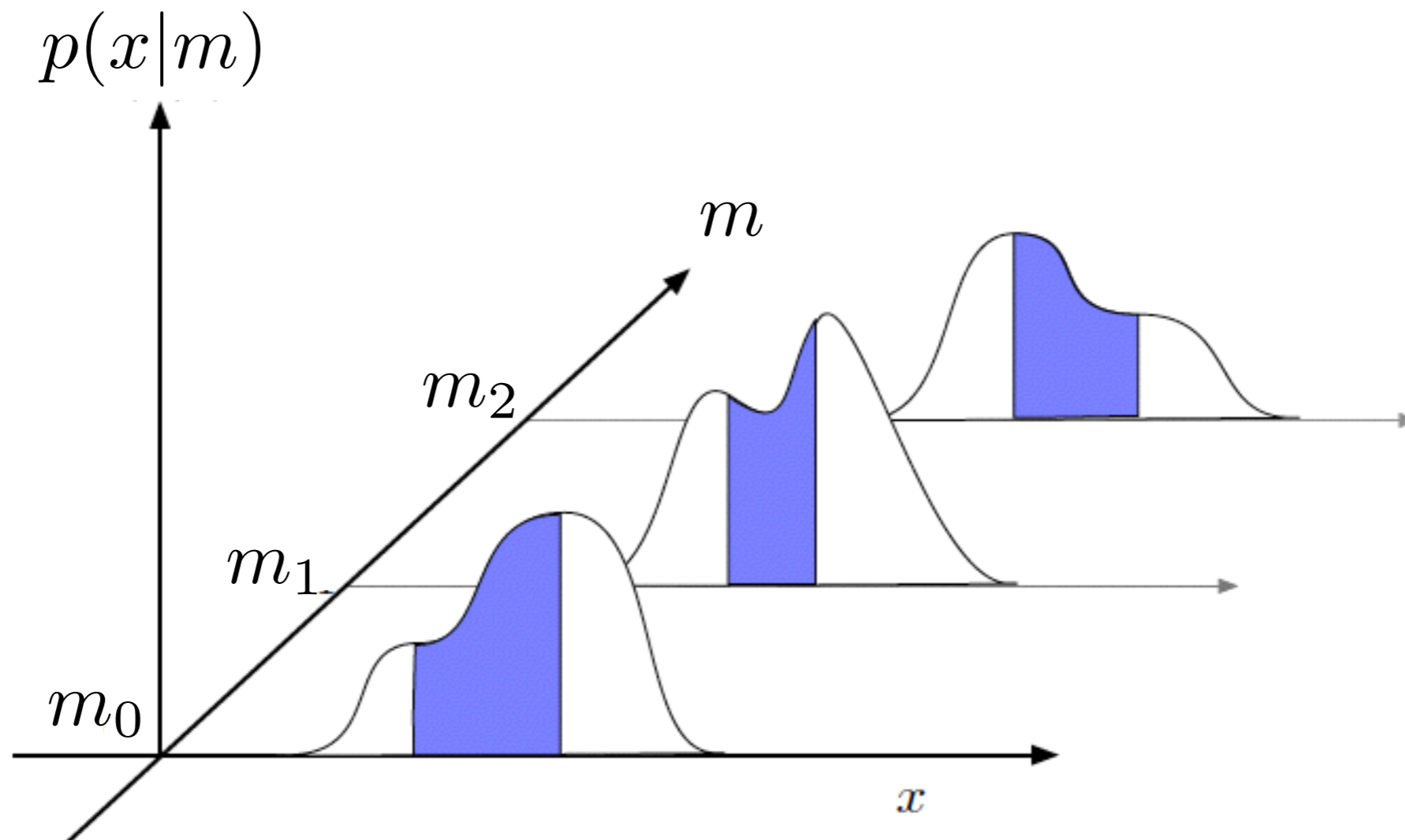
...or a central limit.

The criterion of choice of the region is chosen is the *ordering rule* (because it's the algorithm one uses to *order* the probability until an amount corresponding to the chosen confidence level (68%, in our example) is reached).



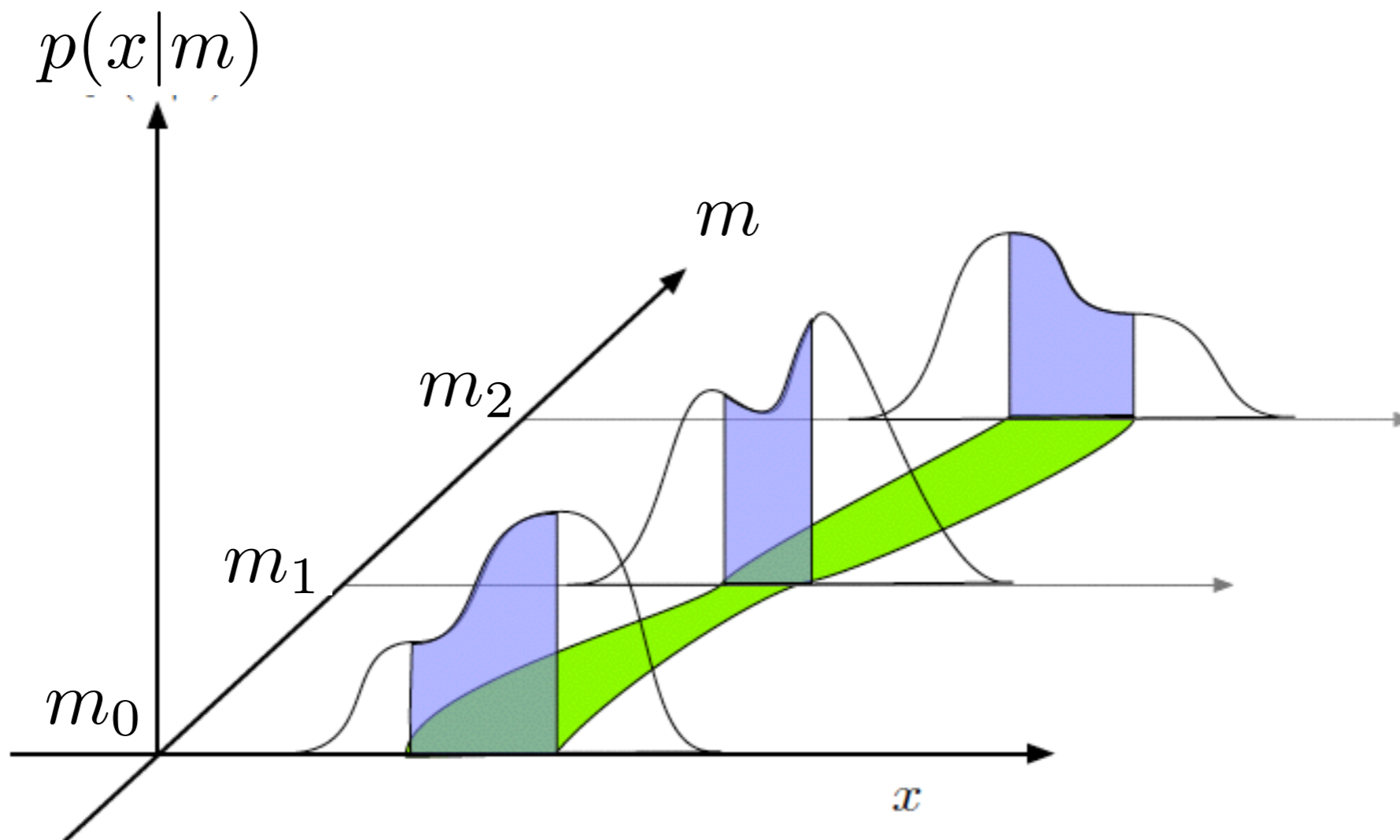
Neyman illustrated V

Derive the acceptance region for every possible true value of the parameter m



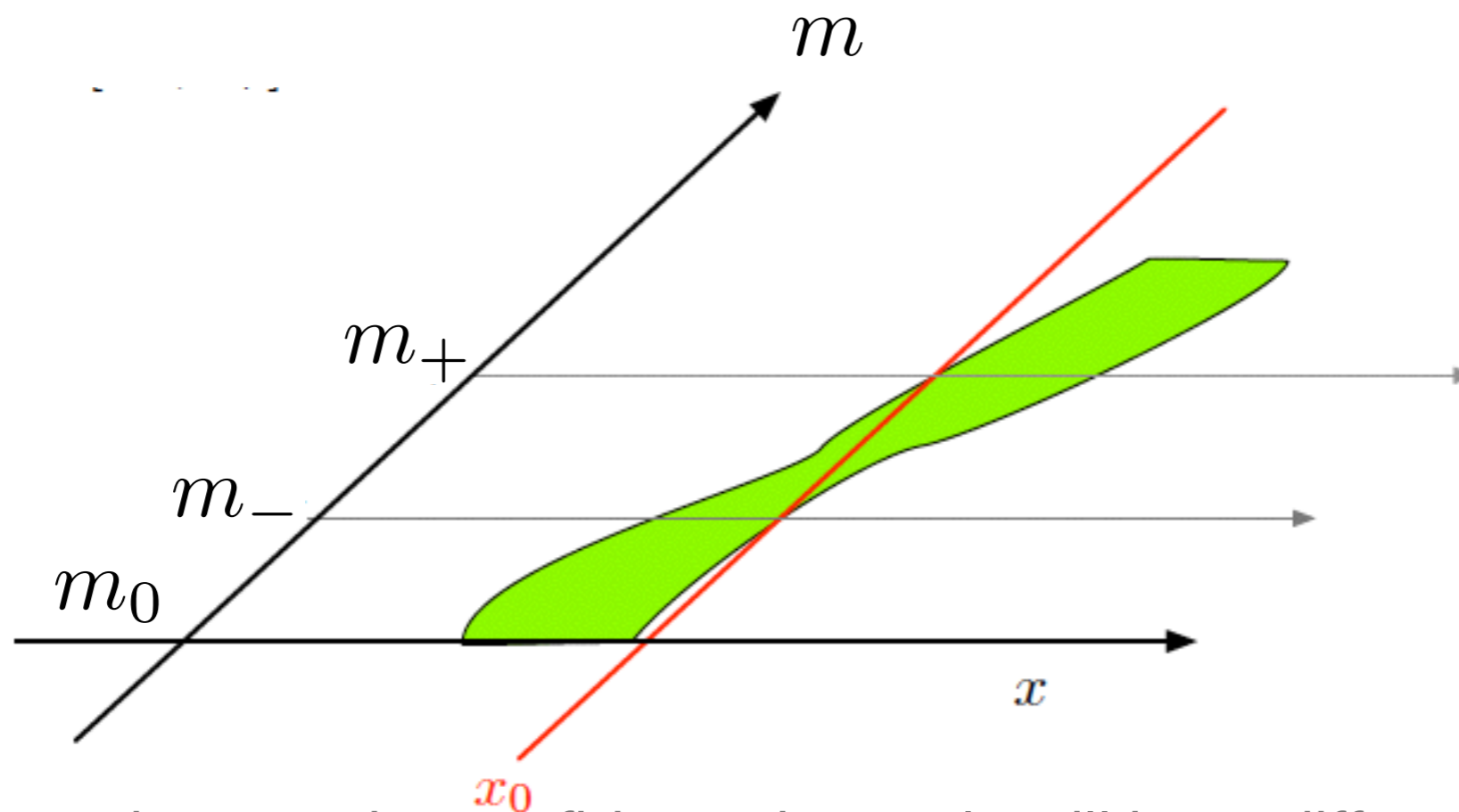
Neyman illustrated VI

This defines a confidence belt for m .



Neyman illustrated VII

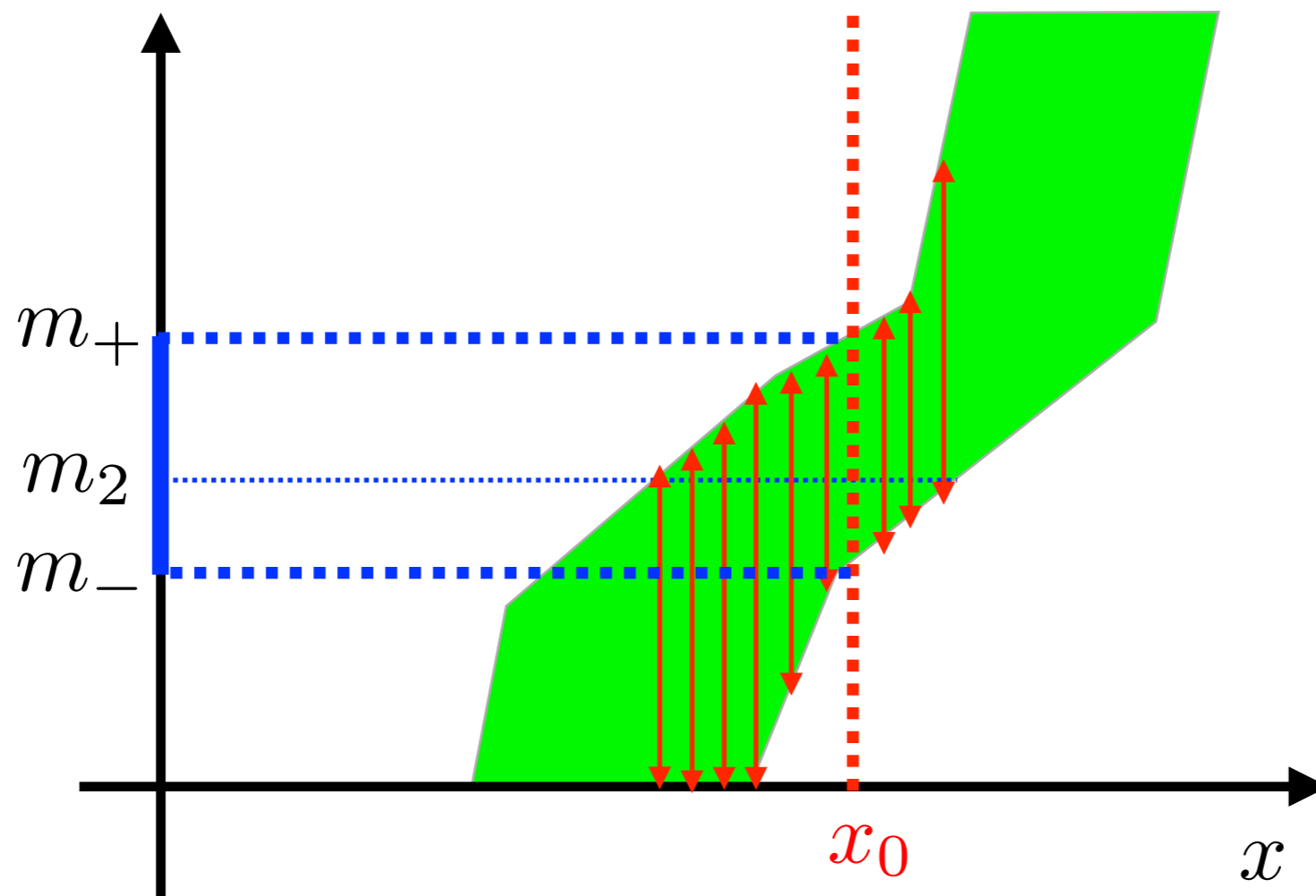
Then you do your analysis on data, and **observe a value x_0** . The observed value intersects the confidence belt. The union of all values of m for which acceptance ranges are intersected by the measurement defines the confidence interval $[m_-(x), m_+(x)]$ at the 68% CL for the parameter. Note that the extremes of the interval are random variables (functions of data x)



In repeated experiments, the confidence intervals will have different boundaries, but 68% of them will contain the (unknown) true value of the parameter m

Why does it work?

Make a measurement x_0 and determine the corresponding confidence interval. For every true value m of the parameter, say m_2 , included in the interval, 68% of the measurements would be in the acceptance region. Each of the measurements will lead to a confidence interval that contains m_2 . Hence, the interval contains the true value with 68% probability, $m \in [m_-, m_+]$ at the 68% CL.



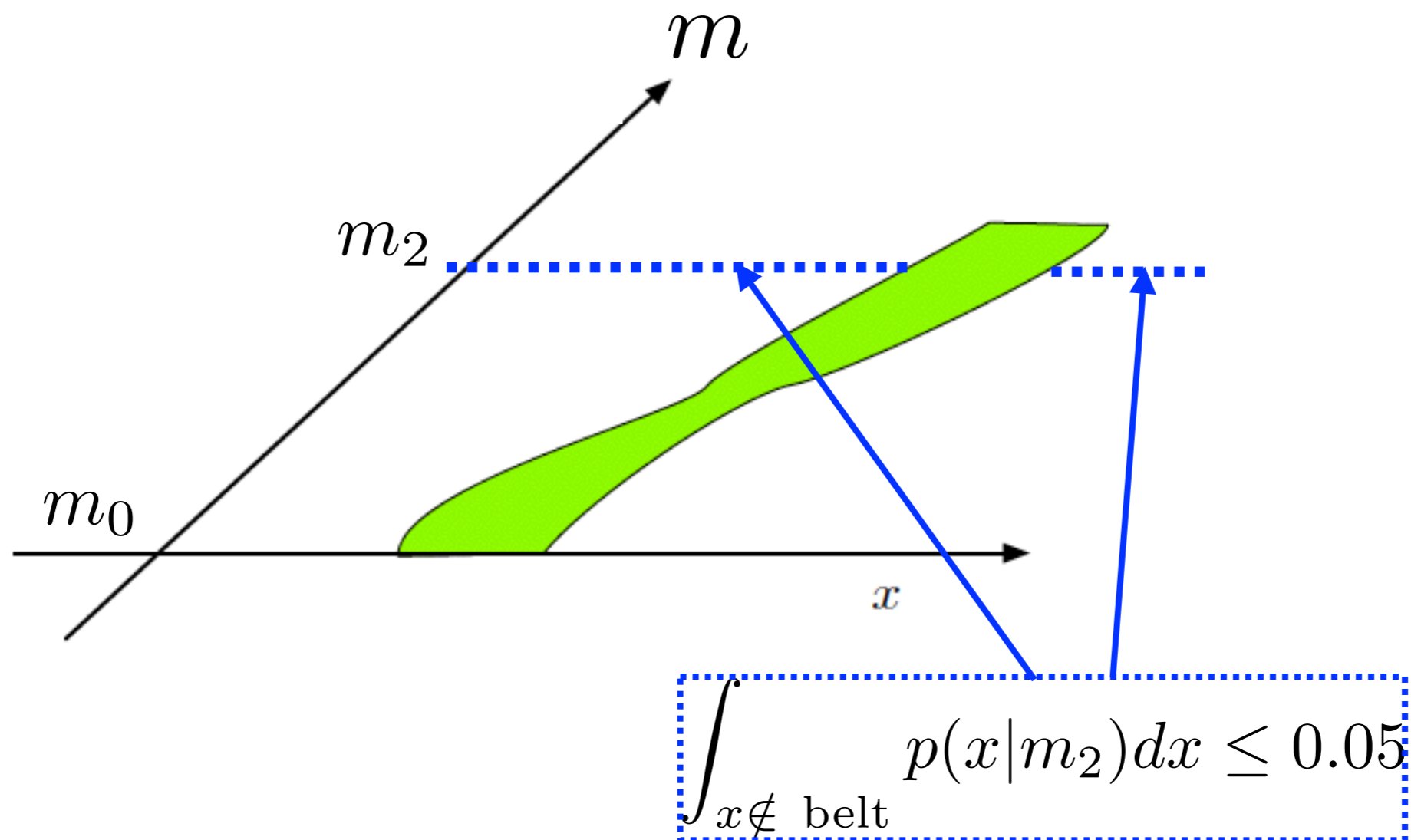
Simple numerical example

Bags of various classes: each class contains a different fraction of white balls (1%, 5%, 50%, 95%, and 99%). Extract N=5 balls from each bag. Would like to infer to which class each bag belongs

		True fraction of white balls				
		1%	5%	50%	95%	99%
Number of white balls observed	5	10^{-10}	$3 \cdot 10^{-7}$	3.1%	77.4%	95.1%
	4	$5 \cdot 10^{-8}$	$3 \cdot 10^{-5}$	15.6%	20.4%	4.8%
	3	10^{-5}	0.1%	31.3%	2.1%	0.1%
	2	0.1%	2.1%	31.3%	0.1%	10^{-5}
	1	4.8%	20.4%	15.6%	$3 \cdot 10^{-5}$	$5 \cdot 10^{-8}$
	0	95.1%	77.4%	3.1%	$3 \cdot 10^{-7}$	10^{-10}

Ordering

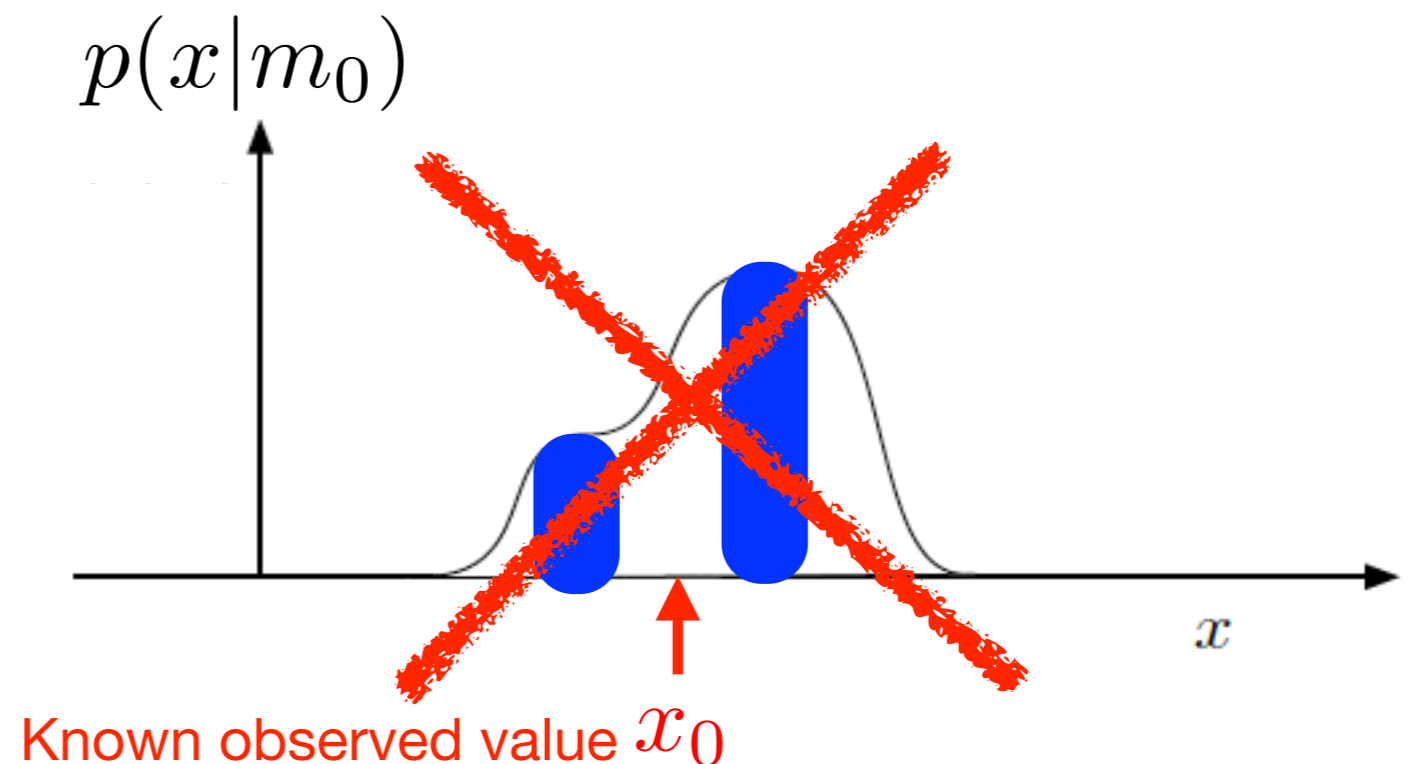
The ordering algorithm is arbitrarily chosen. The only constraint is that, for each value m of the parameter, the integral of the pdf along the x region outside of the belt does not exceed $1-CL$.



Ordering guidelines

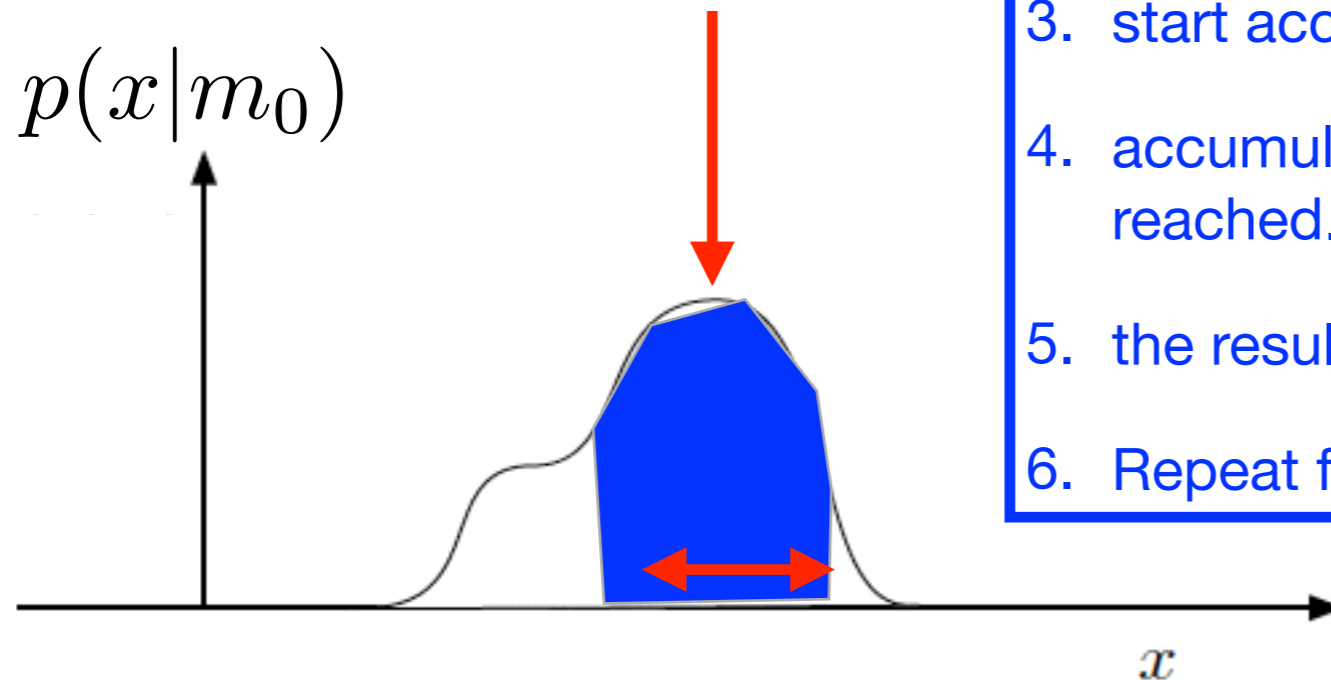
Despite arbitrariness, there are some standards and conventions that are usually followed in the construction of the region.

First and foremost: the ordering algorithm should be **decided and defined prior to look at the experimental data**. Otherwise one could artificially exclude the result of the experiment as long as the excluded area is less than $1-CL$. Also, usually one wants a connected region



Probability ordering

In the past, many tried to get the shortest possible interval, so that the resulting confidence intervals were likely to be more narrow (i.e., measurement more precise). (“probability ordering” or “Crow-Gardner ordering”)



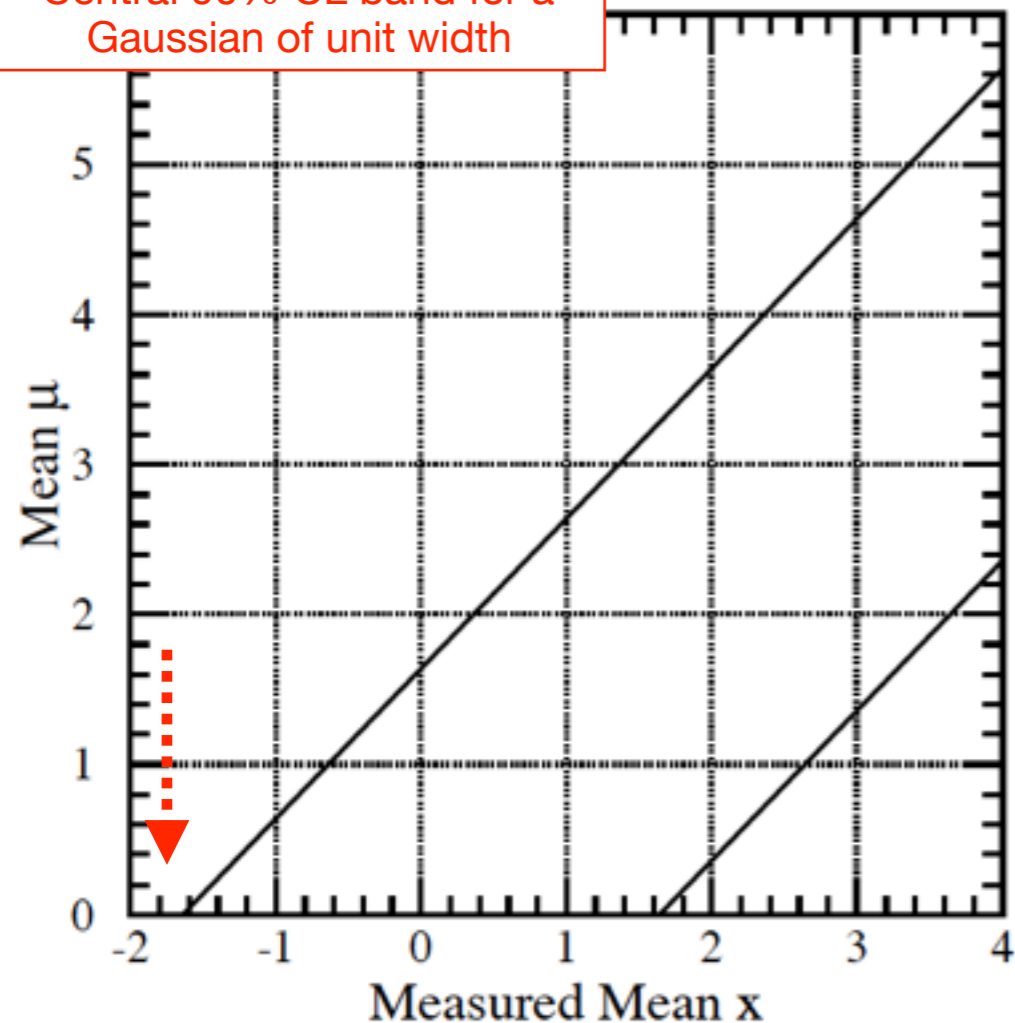
1. Choose one value for m , m_0 , and look at $p(x|m_0)$
2. rank the x values in decreasing order of probability
3. start accumulating from the x with highest probability
4. accumulate the other x values until the desired CL is reached.
5. the result is the shortest possible region for $m=m_0$
6. Repeat for all m

Unfortunately, such criterion is flawed, because the probability **depends on the metric** used for the observable x , so the shortest interval in one metric isn't shortest in others (think of looking at x or at $x' = \ln x$)

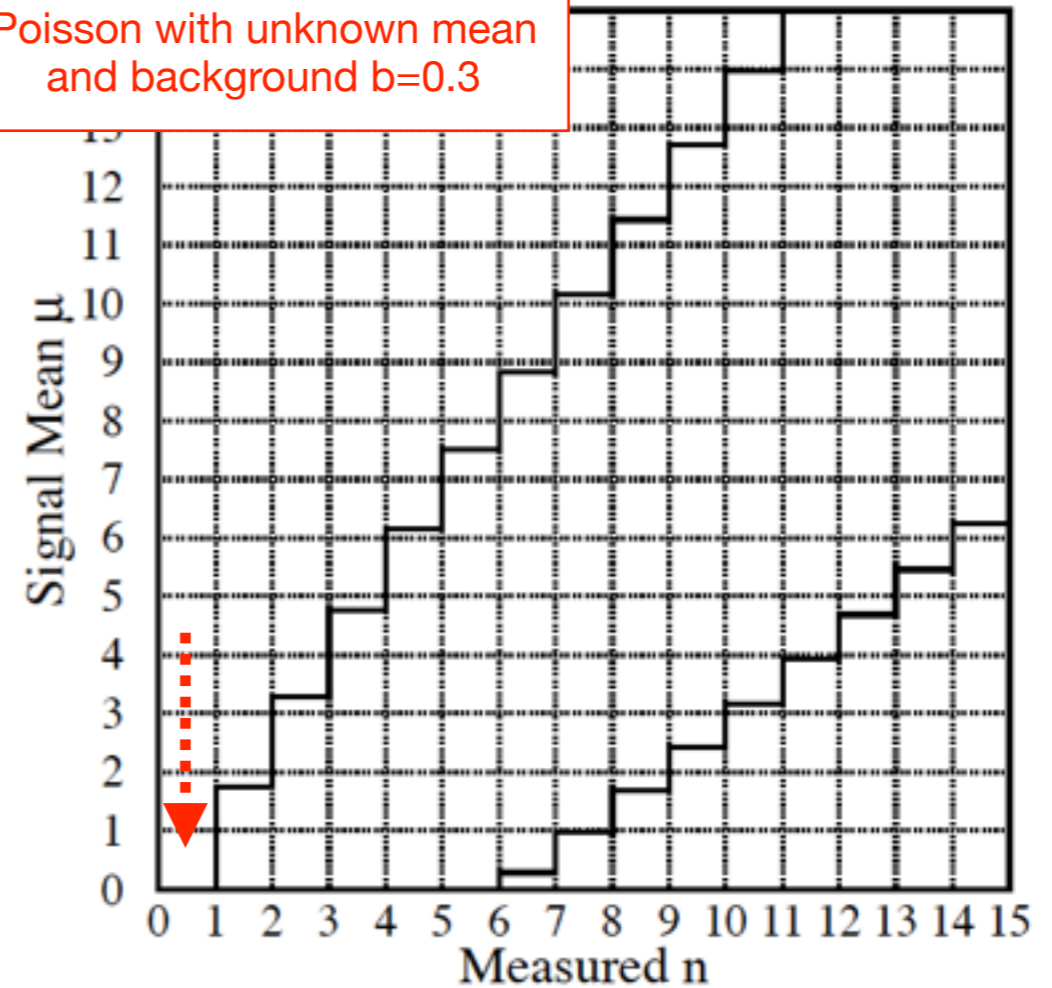
Issues

An number of inconsistencies were identified in Neyman constructions based on simplistic ordering criteria. Probably the worse were empty confidence regions

Central 90% CL band for a Gaussian of unit width



Central 90% CL band for a Poisson with unknown mean and background $b=0.3$



What if I observe $x = -1.8$? or $n = 0$? Resulting confidence regions are empty....

Likelihood-ratio ordering (Feldman and Cousins)

In the late 90ies, a better criterion was proposed. Idea parallels probability ordering, but rather than using the probability as ordering metric, **the likelihood ratio** is used.

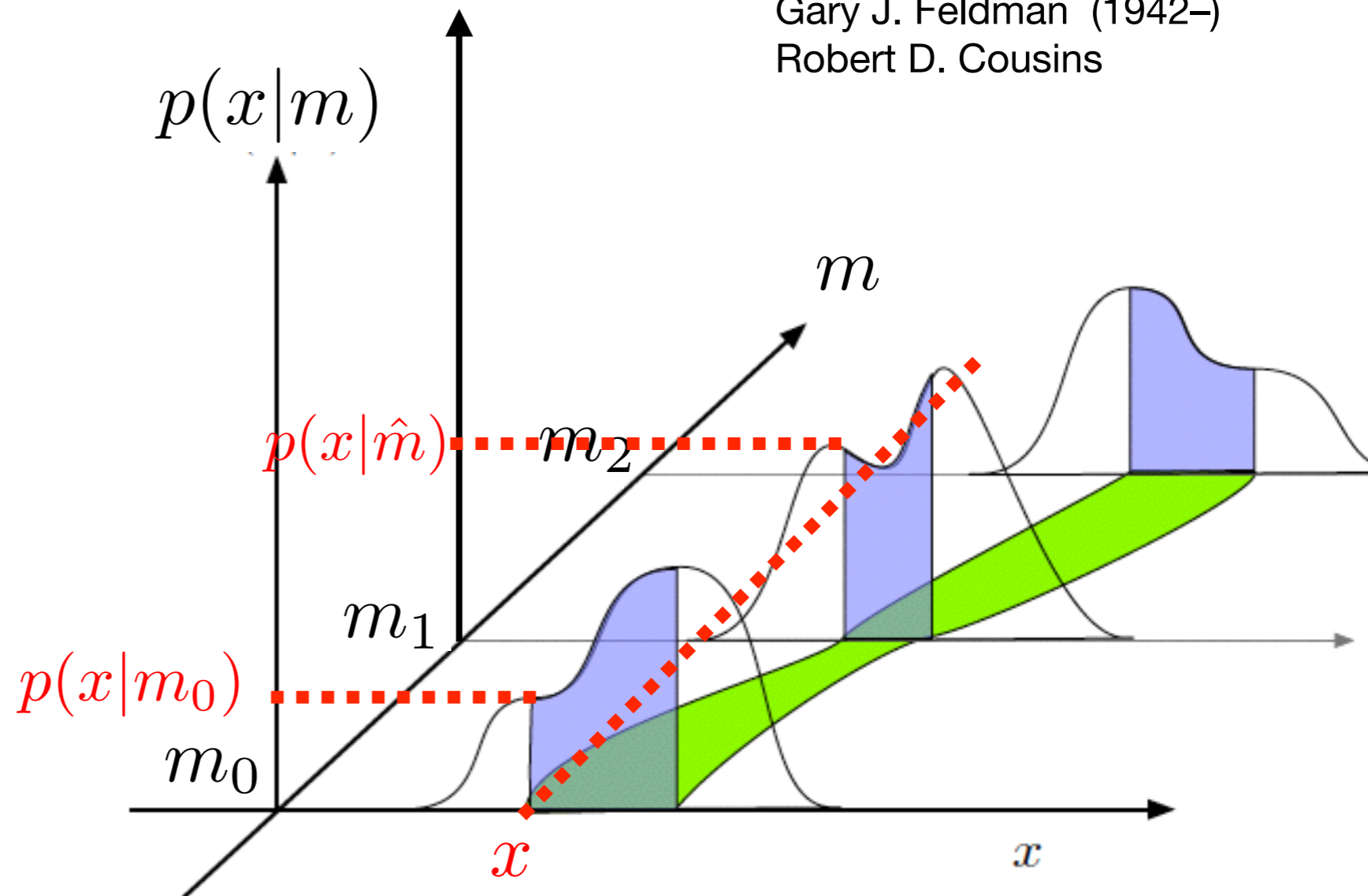


Gary J. Feldman (1942–)
Robert D. Cousins

Choose a value m_0 of the parameter and for each x calculate

$$\text{LR} = \frac{p(x|m_0)}{p(x|\hat{m})}$$

Criterion does not depend only on $p(x|m)$ at fixed m but also from $p(x|m)$ at other m values



Likelihood-ratio ordering

1. Choose one value for m , m_0 and generate simulated samples of pseudodata accordingly.
2. For each observation x calculate (i) the value of the likelihood at m_0 , $p(x|m_0)=L(m_0)$ and (ii) the maximum value of the likelihood $L(\hat{m})$ maximized over the space of m values.
3. Rank all x in decreasing order of likelihood ratio $LR=Lx(m_0)/Lx(\hat{m})$.
4. Start from the x with higher LR and accumulate the others until the desired CL is reached.
5. Repeat for all m

LR-ordering preserves the metric, mostly avoids empty confidence regions and has several other attractive features. It is the most widely used ordering in HEP.

It is not immune from founded criticism and may lead to paradoxes in specific problems, but if you ever will need to quote a confidence region in your analysis it is a good idea to take LR-ordering as default option unless there are strong motivations against it.

Likelihood-ratio ordering practice

Let's try to reproduce LR-bands. Use the original paper as a reference <http://arxiv.org/pdf/physics/9711021v2.pdf> and try to reproduce the confidence band in Fig 7. Useful and interesting information to understand the LR-ordering is also in <http://users.physics.harvard.edu/~feldman/Journeys.pdf>

TABLES

TABLE I. Illustrative calculations in the confidence belt construction for signal mean μ in the presence of known mean background $b = 3.0$. Here we find the acceptance interval for $\mu = 0.5$.

n	$P(n \mu)$	μ_{best}	$P(n \mu_{\text{best}})$	R	rank	U.L.	central
0	0.030	0.	0.050	0.607	6		
1	0.106	0.	0.149	0.708	5	✓	✓
2	0.185	0.	0.224	0.826	3	✓	✓
3	0.216	0.	0.224	0.963	2	✓	✓
4	0.189	1.	0.195	0.966	1	✓	✓
5	0.132	2.	0.175	0.753	4	✓	✓
6	0.077	3.	0.161	0.480	7	✓	✓
7	0.039	4.	0.149	0.259		✓	✓
8	0.017	5.	0.140	0.121		✓	
9	0.007	6.	0.132	0.050		✓	
10	0.002	7.	0.125	0.018		✓	
11	0.001	8.	0.119	0.006		✓	

Indifferent distribution

Suppose that $p(x|m) \approx f(x)$, which is nearly independent on m . Here, for ε arbitrarily small, observing x says very little about m .

	$m1$	$m2$
$x1$	$0.95 + \varepsilon$	$0.95 - \varepsilon$
$x2$	$0.05 - \varepsilon$	$0.05 + \varepsilon$

Any confidence region construction should not provide information about m .

Intuitively one would choose the band that covers the whole space

	$m1$	$m2$
$x1$	$0.95 + \varepsilon$	$0.95 - \varepsilon$
$x2$	$0.05 - \varepsilon$	$0.05 + \varepsilon$

LR-ordering unambiguously would choose this one instead

	$m1$	$m2$
$x1$	$0.95 + \varepsilon$	$0.95 - \varepsilon$
$x2$	$0.05 - \varepsilon$	$0.05 + \varepsilon$

LR here achieves something nearly magical: conclude something out of a nearly uncorrelated information.

Empty intervals with Likelihood-ratio ordering

There is more than that.

LR-ordering has been proposed by Feldman and Cousins as guaranteed not to yield empty intervals.

This holds in many practical applications but is not mathematically true by construction.

<http://arxiv.org/pdf/hep-ex/9912048v3.pdf>

Confidence interval formalism

The goal is to find the range $m_-(x) < m < m_+(x)$ that contains the true value m with probability β (typically one chooses β large, like 68% or 90% or 95%)

Given observation x from a pdf $p(x|m)$, the probability content β of the region $[a,b]$ in x space is

$$\beta = p(a < x < b) = \int_a^b p(x|m) dx$$

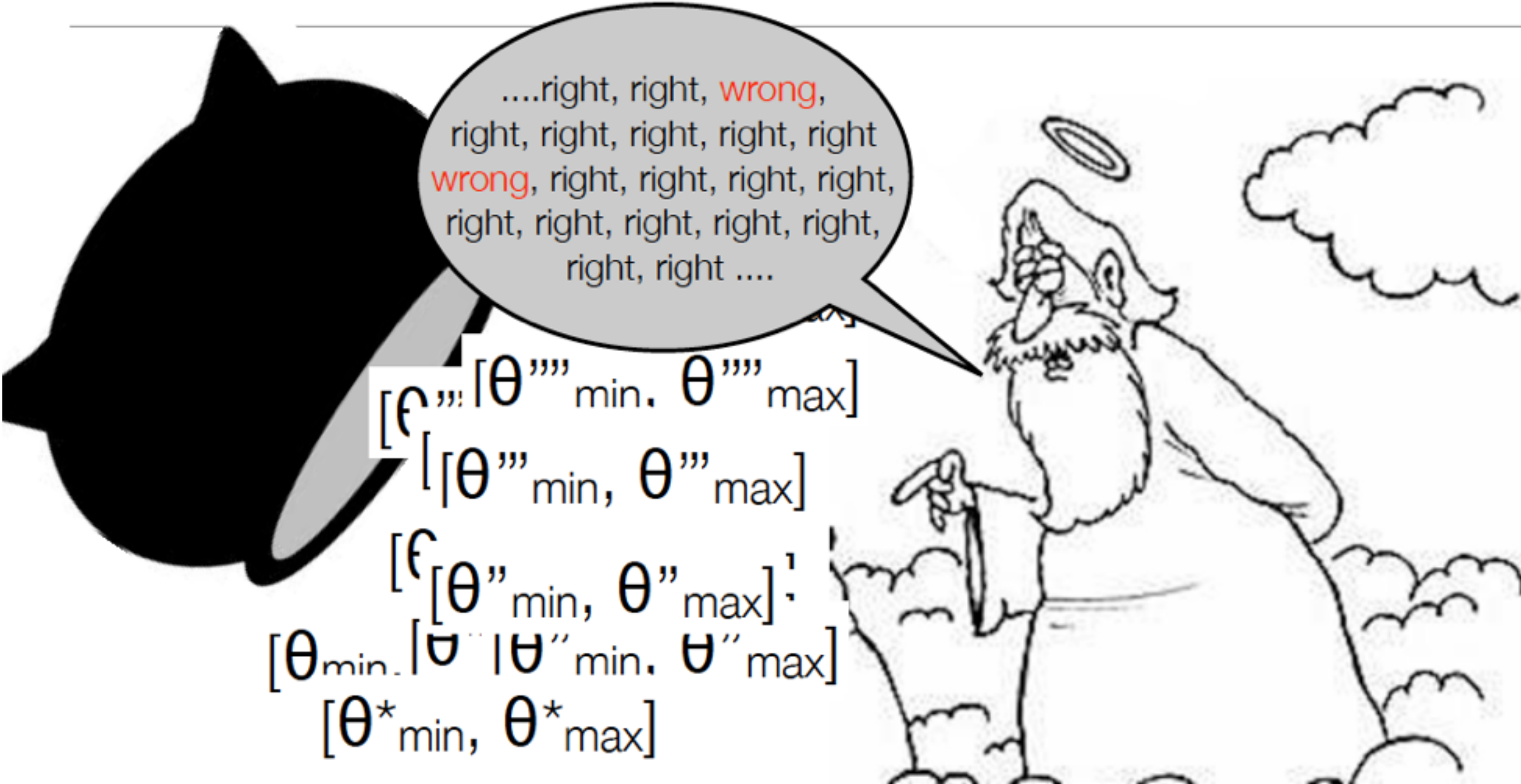
If the pdf and true value m are known one can get β given a and b . But when the true value of the parameter m is unknown, one has to find another random variable $z(x,m)$ such that its pdf is independent of the unknown value of m . If this can be found, then one can re-express the above equation into

$$p(m_-(x) < m < m_+(x)) = \beta$$

A method that yields such an interval possesses the property of coverage.

Coverage

Coverage is a property of the procedure, not of the single measurement.



....right, right, **wrong**,
right, right, right, right, right
wrong, right, right, right, right,
right, right, right, right, right,
right, right

$$[\hat{\theta}''', [\theta''''_{\min}, \theta''''_{\max}]]$$

$$[[\theta'''_{\min}, \theta'''_{\max}]]$$

$$[\hat{\theta}''[\theta''_{\min}, \theta''_{\max}]]$$

$$[\theta_{\min}, \hat{\theta}'' | \theta''_{\min}, \theta''_{\max}]$$

$$[\theta^*_{\min}, \theta^*_{\max}]$$

Inferring from data

Choice of the model

Inference

Given some data, to do inference I need to

1. Identify all the known observations x ;
2. Identify all the unknown parameters m ;
3. Construct a probability model for both

Model building

In all inferences we use **probability models** for the observables x and the unknown parameters m .

With model I mean the full structure $p(\text{data} \mid \text{parameters}) = p(x|m)$

- holding parameters fixed, gives us the probability density function of data, which provides the ability to generate pseudo-data via Monte Carlo.
- holding data fixed gives a likelihood function for parameters

$p(x|m)$ is often (always?) given as granted, but it usually entails major assumptions, where our physics knowledge, understanding, and intuition enter strongly.

Model building is necessary both in Frequentist and Bayesian procedures: this is the part everyone agrees on. Improving the model is the most efficient way of improving your inference

Model building

The model $p(x|m)$ is somehow a quantitative summary of the analysis.

If you had to explain or justify your analysis choices, you would tell a story about how and why you get to know what you know, based on previous results, auxiliary studies and so on.

The quality of the result is largely tied to how convincing and realistic this story is.

In HEP there are three main thrusts of motivation/justification for a model.

- Monte Carlo simulation of fundamental physics processes and their detection
- Data-driven model building
- Effective modelling

A real-life data analysis at LHC or elsewhere typically uses a mixture of the three.

Monte Carlo-based modeling

HEP enjoys the “standard model”: a quantum field theory of the elementary particles and their fundamental interactions. Encapsulates dynamics in a relatively simple Lagrangian density with 18 free parameters.

1. Phase space is sampled with Monte Carlo techniques. 2. The dynamics is simulated based on the Lagrangian. 3. Perturbation theory is used to systematically approximate the theory. 4. Interactions of particles in the detectors are simulated. 5. the results are subject to the same reconstruction algorithms as in experimental data

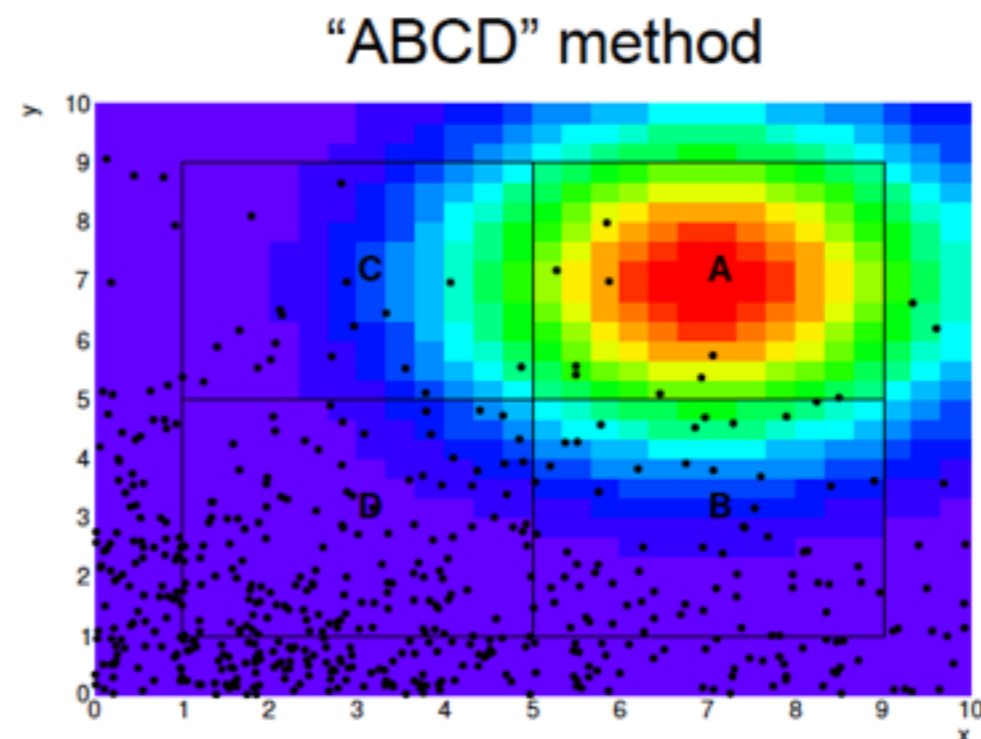
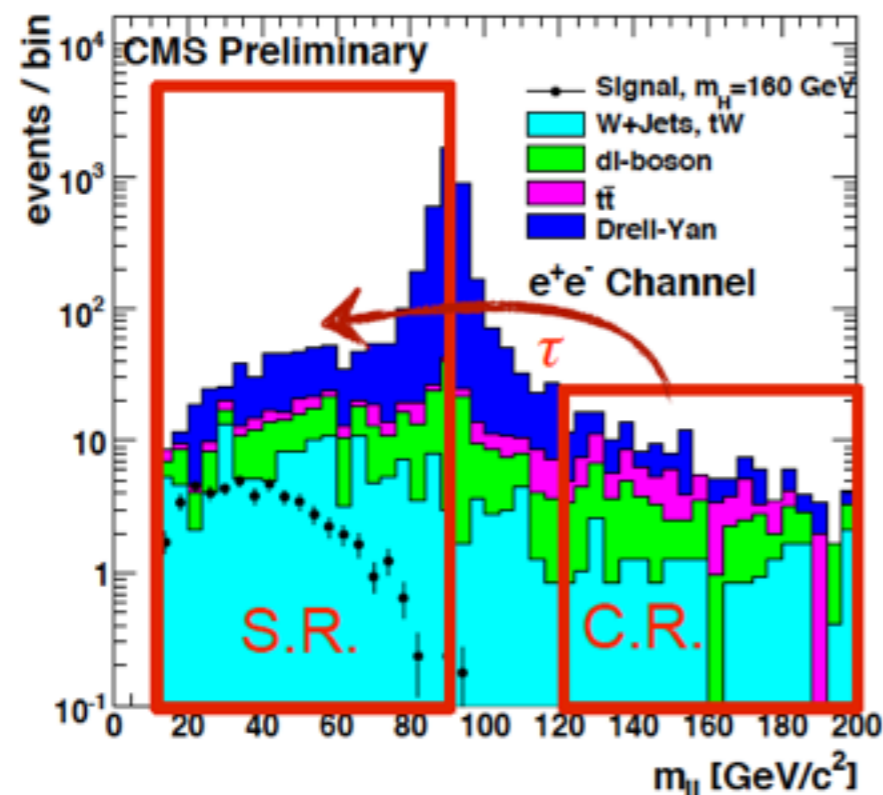
$$\begin{aligned}
 \mathcal{L}_{SM} = & \underbrace{\frac{1}{4} \mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
 & + \underbrace{\bar{L} \gamma^\mu (i\partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) L + \bar{R} \gamma^\mu (i\partial_\mu - \frac{1}{2} g' Y B_\mu) R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
 & + \underbrace{\frac{1}{2} |(i\partial_\mu - \frac{1}{2} g \boldsymbol{\tau} \cdot \mathbf{W}_\mu - \frac{1}{2} g' Y B_\mu) \phi|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{ and Higgs masses and couplings}} \\
 & + \underbrace{g'' (\bar{q} \gamma^\mu T_a q) G_\mu^a}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1 \bar{L} \phi R + G_2 R \phi_c L + h.c.)}_{\text{fermion masses and couplings to Higgs}}
 \end{aligned}$$

We can look at distributions of any observables we can measure in data

Data-driven modeling

Simulation may not be trustable to the desired precision for all processes. For those processes, typically backgrounds, for which simulation isn't realistic enough, try data-driven modeling.

Use subsets of events in data known to be dominated by the process of interest and to model its distributions. Use simulation to determine the coefficients needed to extrapolate from the control region to the signal region. Coefficient may have experimental and theoretical uncertainties.

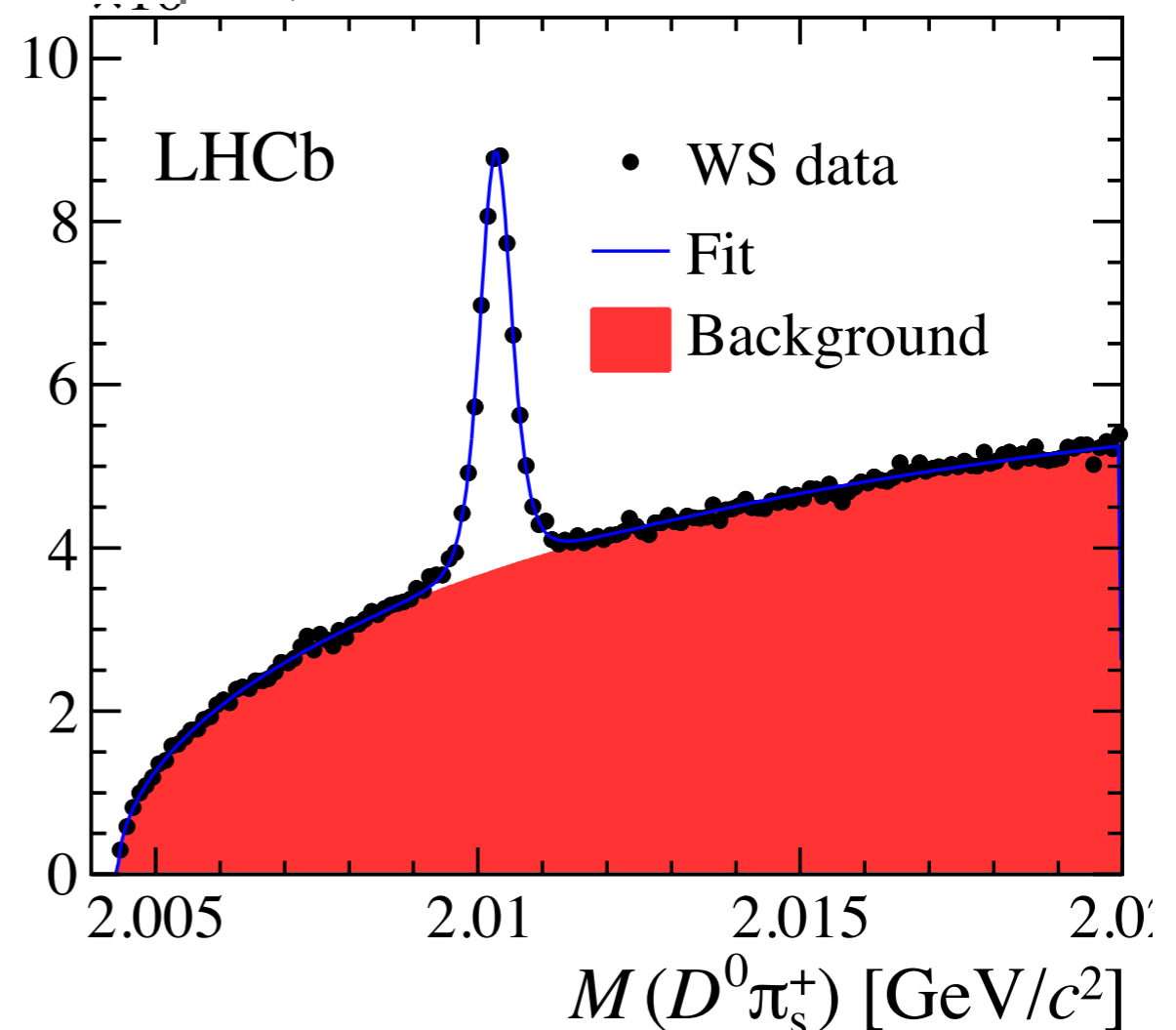


Effective (or empiric) modeling

When everything else fails because the processes are too complex to be reliably simulated and control regions in data are not available to determine their features, one resorts to empiric modeling.

Look at data distribution and try to guess shapes, features etc...

E.g., model the “combinatorial background” on the right with a square-root-like function that appears to adapt fairly well to the observed shape



Why?

Why is this relevant to our topic?

Because the discussion of modeling allows us to introduce a **new, and crucial,** discussion, related to **uncertainties** that are not statistical, but **systematic.**

Funny things happen if you've got your model wrong

The number of droplets produced by a charged particle in a cloud chamber is proportional to charge Q . A $Q=1$ particle have a Poisson distribution of droplets with $\mu = 229$ (known by counting droplets from known particles over the known length).

In 1969, out of a sample of 55000 particles, McCusker and Cairns observed a particle with only 110 droplets. As $p(j < 110 | \mu = 229) = 10^{-18} \ll 1/55000$, **they claimed evidence for free quarks (fractional charge).**

Shortly after Adair and Kasha pointed out that in each elementary scattering 4 droplets are produced in average. Therefore, assuming that always 4 droplets get produced, a much reasonable rate of occurrence would be expected from ordinary $Q=1$ particles

$$p(j < 110/4 \approx 28) \approx \sum_{j=1}^{28} (229/4)^j \frac{e^{-229/4}}{j!} \approx 10^{-5}$$



Funny things happen if you've got your model wrong

VOLUME 23, NUMBER 12

PHYSICAL REVIEW LETTERS

22 SEPTEMBER 1969

New York, 1964), pp. 3-13.

⁴Recommended Unit Prefixes; Defined Values and Conversion Factors; General Physical Constants, National Bureau of Standards Miscellaneous Publication No. 253 (U. S. Government Printing Office, Washington, D.C., 1963).

⁵C. V. Boys, Phil. Trans. Roy. Soc. (London), Ser. A 186, 1 (1895), and Proc. Roy. Inst. Gt. Brit. 14, 353 (1894).

⁶J. H. Poynting, "Gravitation," in Encyclopedia Britannica, 11th ed., and Collected Scientific Papers (The Macmillan Company, New York, 1920).

⁷P. Heyl, J. Res. Natl. Bur. Std. (U.S.) 5, 1243 (1930); P. Heyl and P. Chrzanowski, J. Res. Natl. Bur. Std.

(U.S.) 29, 1 (1942).

⁸J. W. Beams, A. R. Kuhlthau, R. A. Lowry, and H. M. Parker, Bull. Am. Phys. Soc. 10, 249 (1965).

⁹Procured from Micrometrical Division of Bendix Company.

¹⁰W. R. Towler and E. V. McVey, to be published.

¹¹R. V. Jones and J. C. S. Richards, J. Sci. Instr. 36, 90 (1959).

¹²J. H. Nash, A. C. Neeley, and P. J. Steger, Atomic Energy Commission Research and Development Report No. Y-1654 (unpublished), Oak Ridge Y-12 Plant, Union Carbide Company Nuclear Division.

¹³J. W. Beams, D. M. Spitzer, and J. P. Wade, Jr., Rev. Sci. Instr. 33, 131 (1962).

EVIDENCE OF QUARKS IN AIR-SHOWER CORES*

C. B. A. McCusker and I. Cairns

Cornell-Sydney University Astronomy Center, Physics Department, The University of Sydney, Sydney, Australia

(Received 3 September 1969)

In a study of air-shower cores using a delayed-expansion cloud chamber, we have observed a track for which the only explanation we can see is that it is produced by a fractionally charged particle.

VOLUME 23, NUMBER 23

PHYSICAL REVIEW LETTERS

8 DECEMBER 1969

ANALYSIS OF SOME RESULTS OF QUARK SEARCHES

R. K. Adair

Yale University, New Haven, Connecticut

and

H. Kasha

Brookhaven National Laboratory, Upton, New York 11973

(Received 31 October 1969)

The interpretation of the results of Cairns, McCusker, Peak, and Woolcott, indicating a discovery of quarks in the cores of very energetic extensive air showers, is shown to be extremely difficult to reconcile with the results of other negative experiments. Alternative explanations of their results are then suggested.

Inferring from data

Systematic uncertainties

What is systematics?

Hard to find any 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.



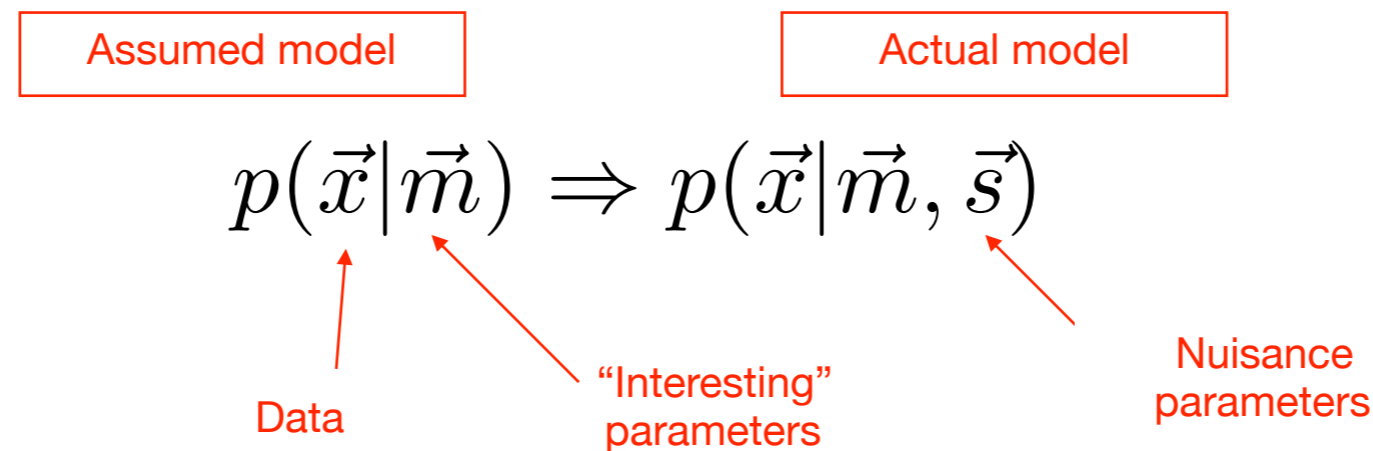
Any statistical inference is based on $p(x|m)$: observe x to extract information about m , assuming to know the distribution $p(x|m)$, that is “the model”.

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, 2001

Nuisance parameters

Assume model $p(x|m)$. But the actual model realized may differ. The difference is parametrized by the additional dependence on some **unknown nuisance 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 based on independent data but sharing nuisance parameters may get correlated (through common priors)

Typically good avoiding mixing frequentist and Bayesian approaches. E.g., don't use marginalized $p(x;m)$ to get Neyman confidence intervals. Hybrid results are hard to interpret. If you assume the distribution of parameter s known, you enter the Bayesian realm, hence should rather assume known the distributions of *all* parameters.

Frequentist approach — interval estimates

The goal remains to devise a procedure that **guarantees coverage** *whatever* is the value of the unknown nuisance parameters.

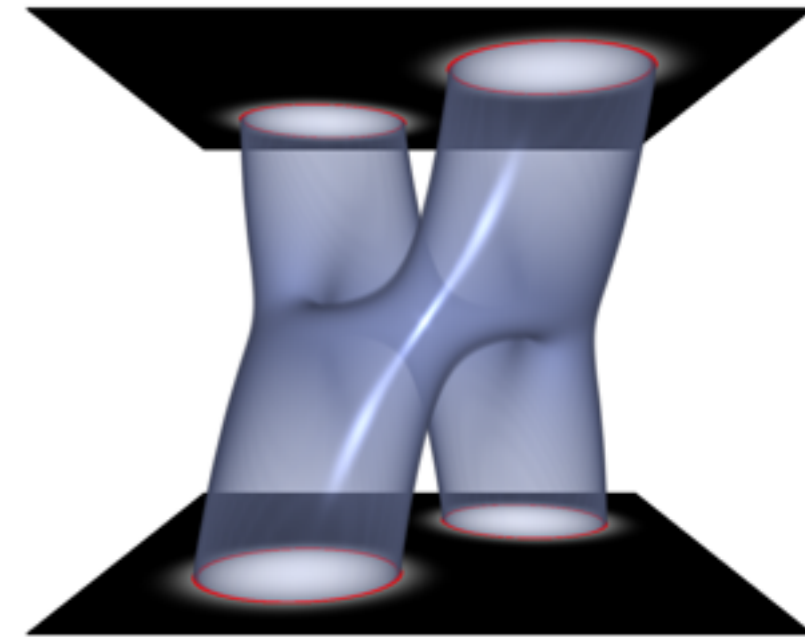
Because the allowed and plausible range of the possible true values of the nuisance parameters might be large, the coverage requirement may result into overcoverage for some values of s .

Finding the optimal procedure to obtain rigorous frequentist confidence intervals in the presence of nuisance parameters is the object of current active research in the statisticians/physicist community (see papers by, e.g., Rolke, Lopez and Conrad, K. Cranmer)

Consensus not yet reached on what is best. However, at the LHC the generalization of the LR ordering using profiled likelihoods gains popularity. This is the only method we'll discuss here.

Why not just multidimensional LR-ordering?

Can one generalize the construction based on likelihood-ratio ordering to multidimensional likelihoods, $L(\vec{m}, \vec{s})$? I.e., construct a multidimensional confidence belt whose intersection with observed value \vec{x} is “projected” onto the subspace of interesting parameters \vec{m} to get the interval?



One can. But has to deal with two **serious issues**.

- Due to geometry, projections of higher dimensional structures into lower-dimensional subspaces lead to information loss. Structures in the multidimensional space overlap and get “shadowed”, leading to broader (i.e. lower-precision) intervals that may extend over the whole domain of m , making the inference non informative.
- A LR-based construction implies generation and fit of pseudodata that achieve an adequate sampling of the space of parameters. If such space is multidimensional, the needed computing power quickly diverges,

Profile-likelihood ratio ordering

An attractive and promising approach is to perform ratio-ordering on the likelihood profiled with respect to its nuisance parameters. This is called a profile-likelihood,

Not a likelihood, but a lower-dimensional derivation of it: the likelihood is a multidimensional function of the physics and nuisance parameters; the profile likelihood is a function of the physics parameters only obtained by maximizing the likelihood wrt to the nuisance parameters.

Variable	Meaning
m	Parameters of interest ("physics parameters")
s	Nuisance parameters
\hat{m}, \hat{s}	Parameters that maximize $L(x m, s)$
\hat{s}^*	Parameter that maximizes $L(x m = m_0, s)$

$$\text{PLR} = \frac{L(x|m=m_0, \hat{s}^*)}{L(x|\hat{m}, \hat{s})}$$

Profile-likelihood ratio ordering

Algorithm is similar to that of the LR ordering.

1. Choose one value m_0 for m and one value s_0 for s , and generate pseudodata accordingly
2. For each observation x (i) maximize $p(x|m=m_0,s)=L(m=m_0,s)$ with respect to s to get $L_x(m=m_0,\hat{s}^*)$ and (ii) maximize the likelihood $L(m,s)$ over the space of m and s to obtain $L_x(\hat{m},\hat{s})$
3. Rank all x in decreasing order of profile likelihood ratio $PLR=L_x(m=m_0,\hat{s}^*)/L_x(\hat{m},\hat{s})$
4. Start from the x with higher PLR and accumulate the others until the desired CL is reached.
5. Repeat for all values of m
6. Repeat for values of s sampled in a plausible range

Need to **fit** each sample **twice**, one with all parameters (physics and nuisance) floating, and another one with physics parameters fixed to their test value m_0 .

It has been shown that the PLR is independent of the true values of the nuisance parameters s and, asymptotically, its distribution too gets independent of them.

When solving a given problem, try to avoid solving a more general problem as an intermediate step.

V.I. Vapnik

Inferring from data

Asymptotic properties of (profile) likelihood ratios

LR- and PLR-ordering a CPU nightmare

Neyman constructions using (P)LR-ordering imply generation and fit of large numbers of simulated experiments. These are necessary to determine the (P)LR distributions needed to build the confidence belt, for a set of true values of physics and nuisance parameters of the likelihood that adequately sample its n -dimensional space.

Given a target sampling density d , needed computing power grows like d^n and becomes soon unmanageable (I once dealt with a 27-dimensional case, and that was a nightmare, <http://arxiv.org/pdf/0810.3229v2.pdf>)

Can we use good-enough approximations of the (P)LR distributions for every value of the likelihood parameters, and that are based only on the observed likelihood function?

Wilks' theorem

Asymptotically (i.e. for large N), **the distribution of likelihood ratio**

$$-2 \ln \text{LR}(m_0) = -2 \ln \frac{p(x|m_0)}{p(x|\hat{m})}$$

approaches a χ^2 distribution with a number of degrees of freedom corresponding to the dimensionality of m .



Samuel S. Wilks (1906-1964)

This holds independently of the shape of $p(x|m)$ and on the value of m .

Facilitates enormously usage of likelihood- and profile-likelihood-ratio as ordering quantities in the construction of intervals. If the likelihood is regular enough to be in asymptotic regime, one can avoid massive production of simulated experiments.

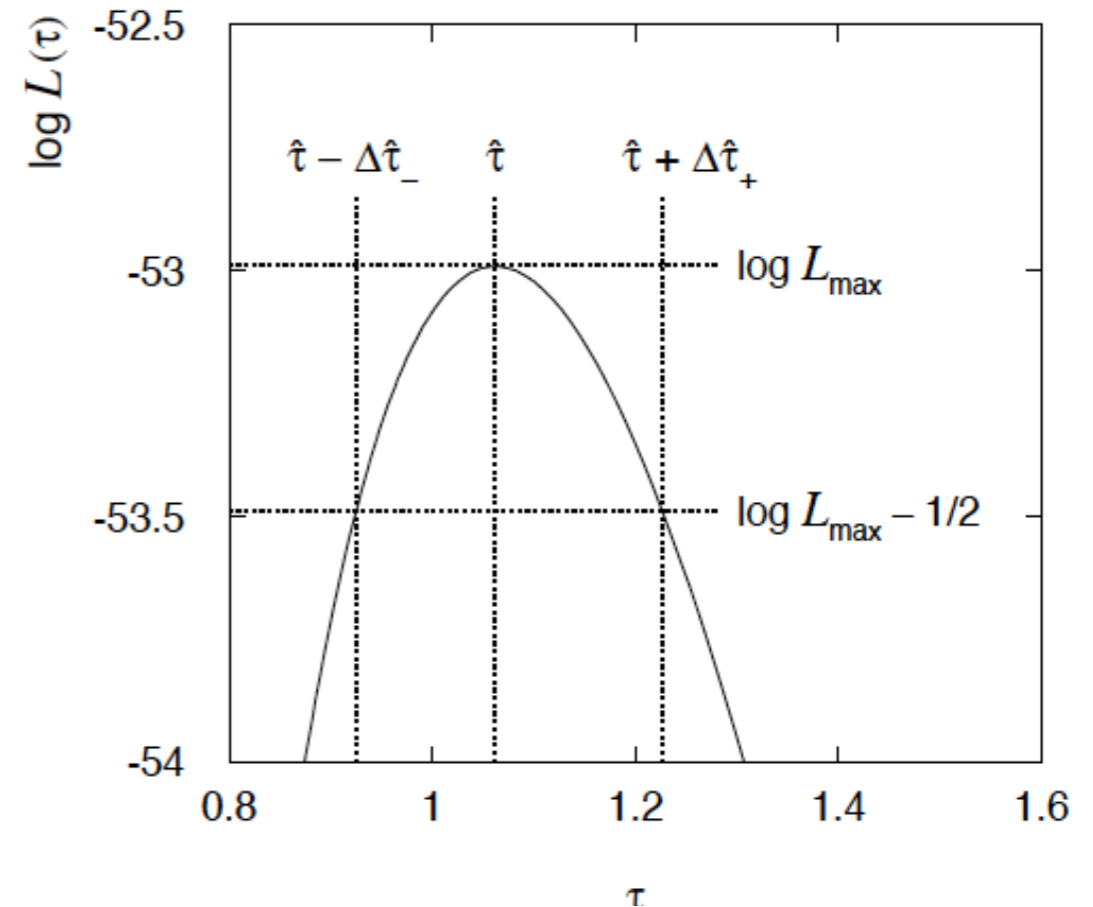
One can use few samples of pseudodata to check if the likelihood is already asymptotic. If it is, then trust Wilks' theorem; if not, only pseudodata allow determining the needed distributions

Wilks' theorem at work

Remember the graphical construction for the variance of a one-dimensional ML estimator?

Wilks' theorem tells us we can use this in any number n of likelihood dimensions to find **approximate central acceptance regions** in Neyman constructions that use (P)LR-ordering

$$-2 \ln \text{LR}(m_0) = -2 \ln \frac{p(x|m_0)}{p(x|\hat{m})} = \Delta$$



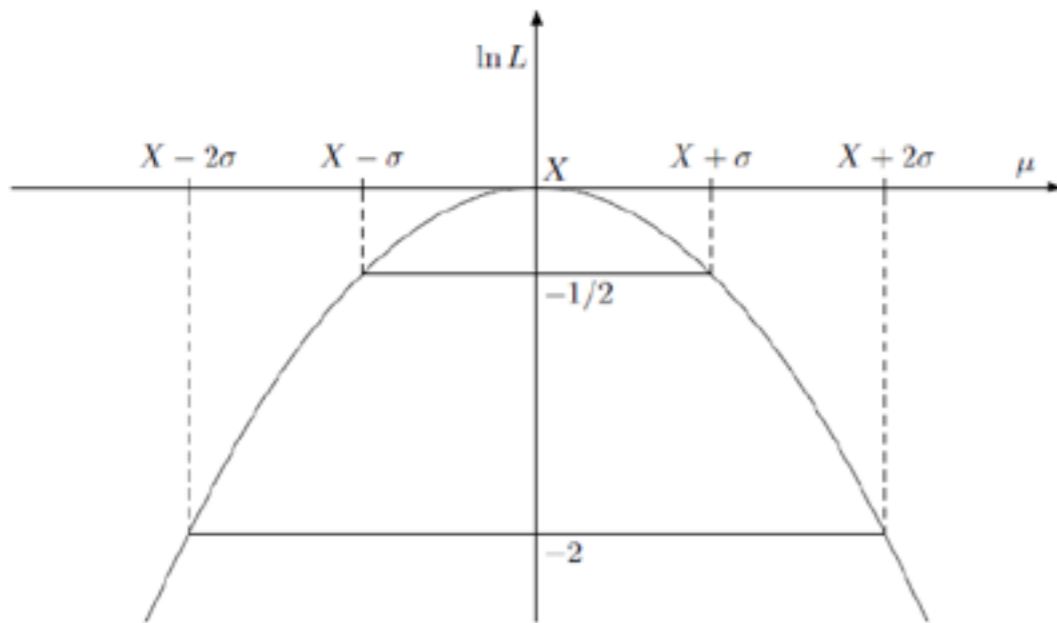
Δ	CL				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

projection onto the space of parameters of a 1(2)-dimensional likelihood at the point where $-2\ln\text{LR}$ varies by 1.0 units identifies a 1(2)-dimensional 68(39)% CL central interval

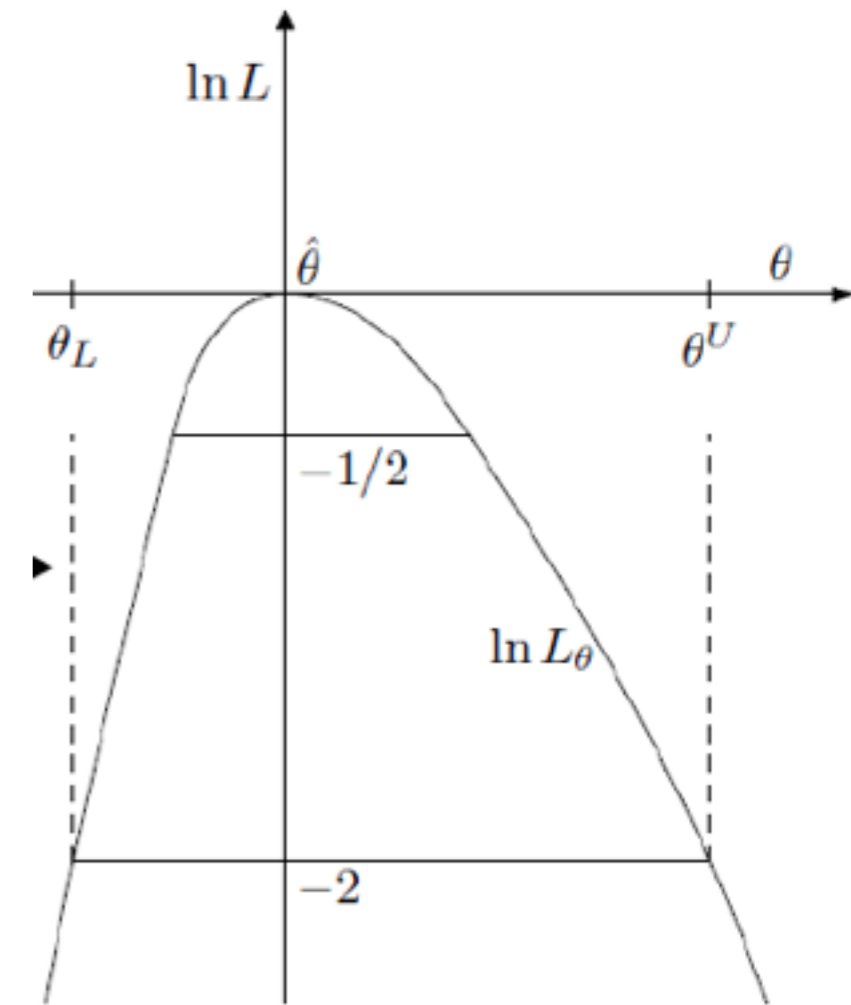
CL	$\hat{\Delta}$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

projection onto the space of parameters of a 3-dimensional likelihood at the point where $-2\ln\text{LR}$ varies by 6.25 units identifies a 3-dimensional 90%CL central interval

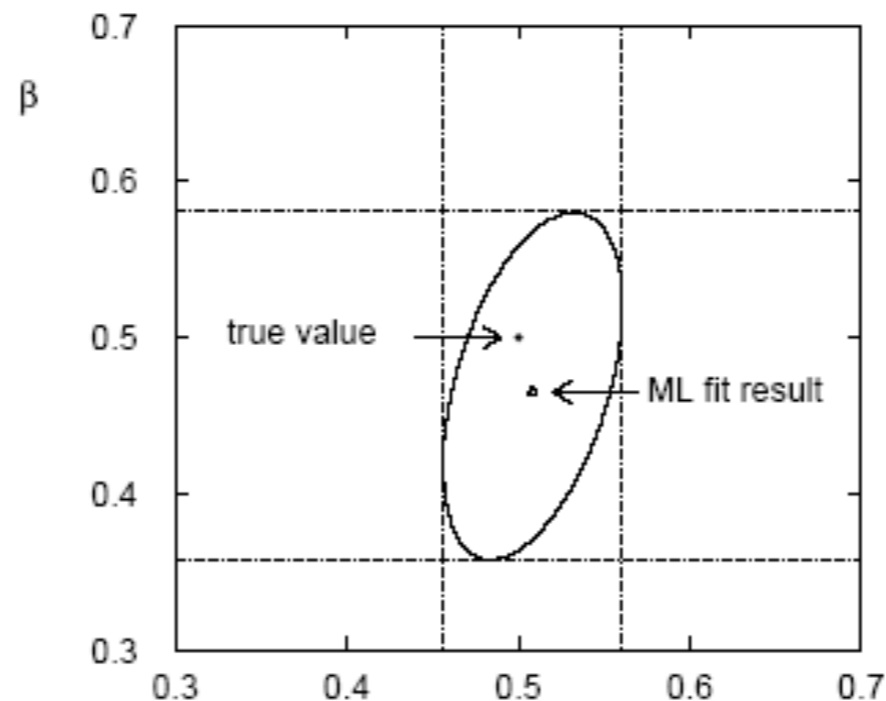
Wilks' theorem at work (F. James)



One-dimensional Gaussian likelihood



One-dimensional non-Gaussian likelihood

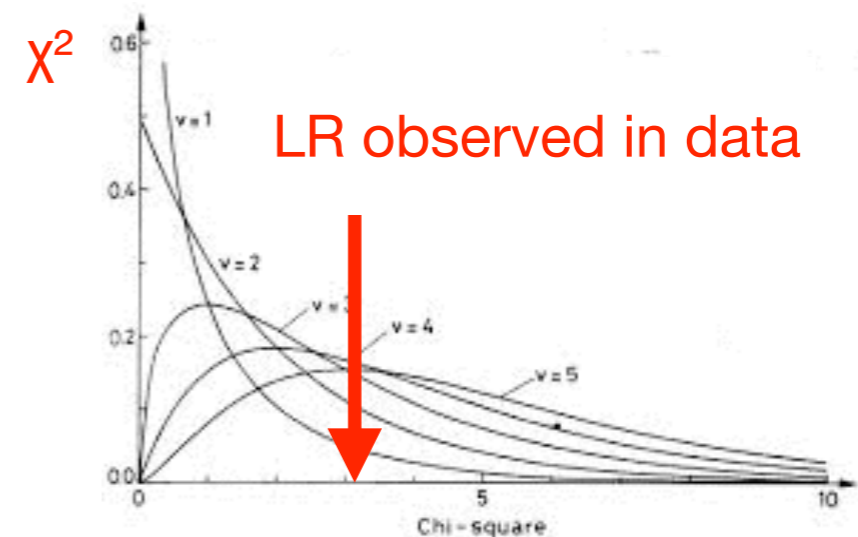
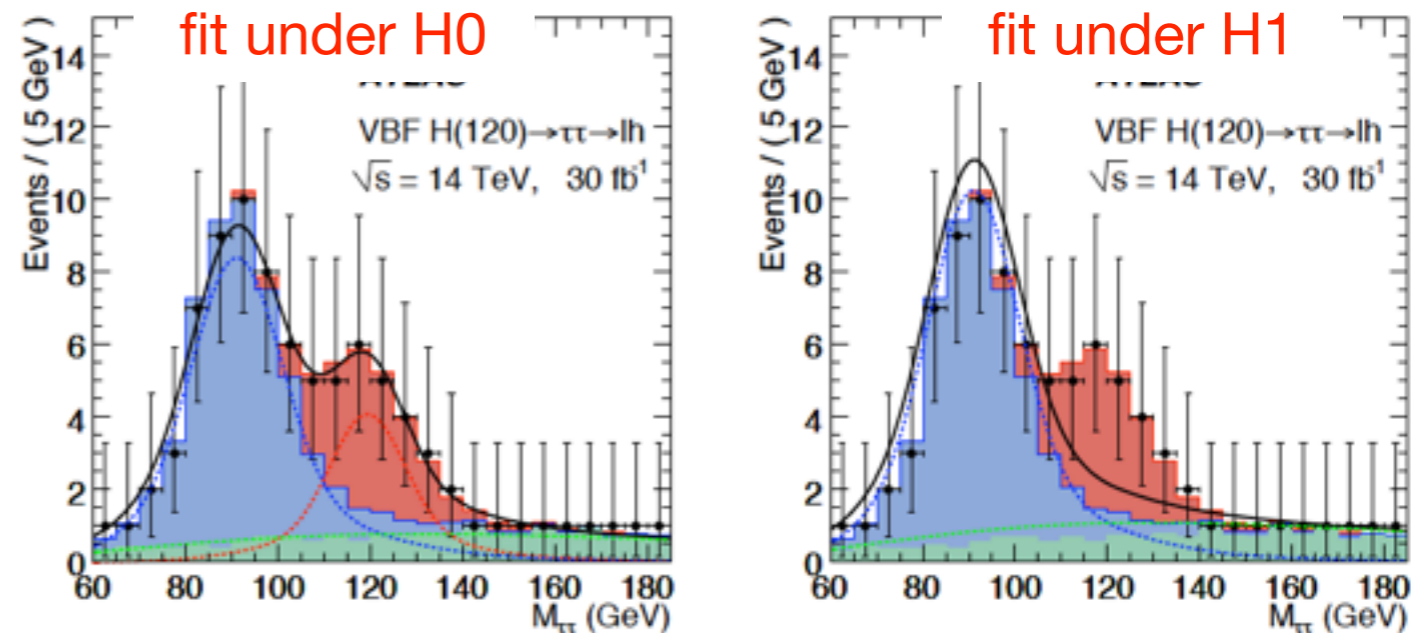


Two-dimensional Gaussian likelihood

(profile) likelihood-ratio as a test statistic

Interest in the NP lemma is not only theoretical. In many cases where testing an hypothesis is needed, one uses the (profile) likelihood ratio as a test statistic. It's known (χ^2) asymptotic distribution allow testing with no need to laboriously construct LR distributions by generating and fitting pseudodata.

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.

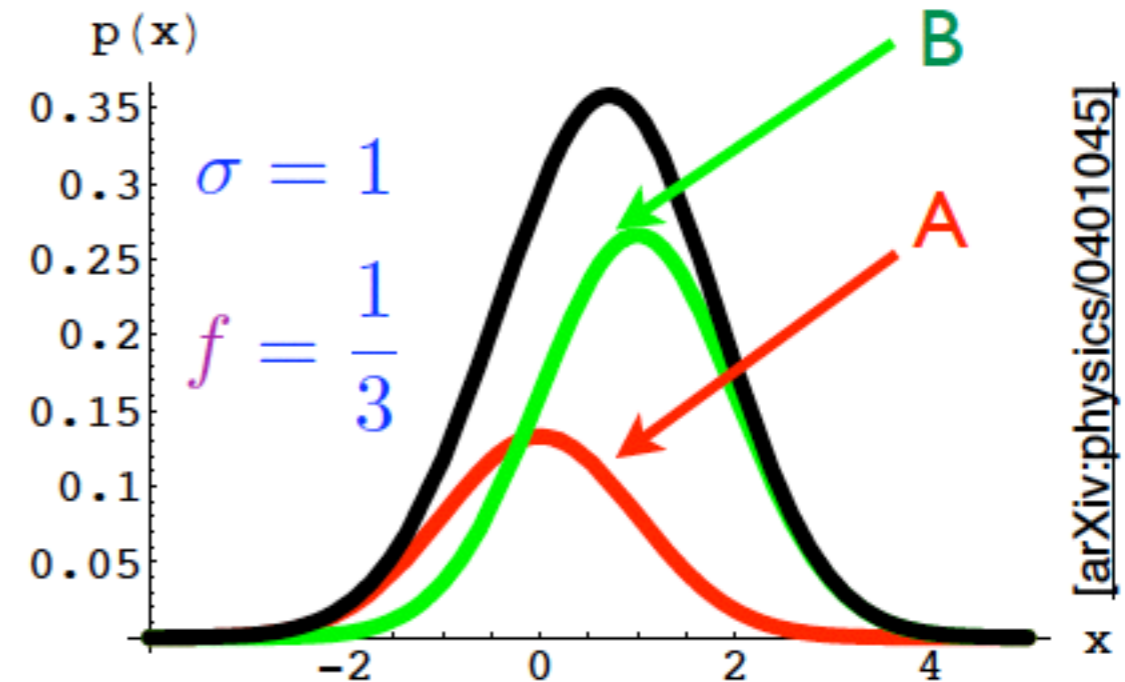


Inferring from data

Biased MLE bias from incomplete pdf (known at the LHC as “Punzi effect”)

A simple fit of sample composition

- Measure a variable x .
- Two classes of events, **A** and **B**
- For each class, x is distributed Gaussian



- What is f , the fraction of A-type events?

$$P(x|A) = g(x; 0, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-0)^2}{2\sigma^2}}$$

- Likelihood fit...

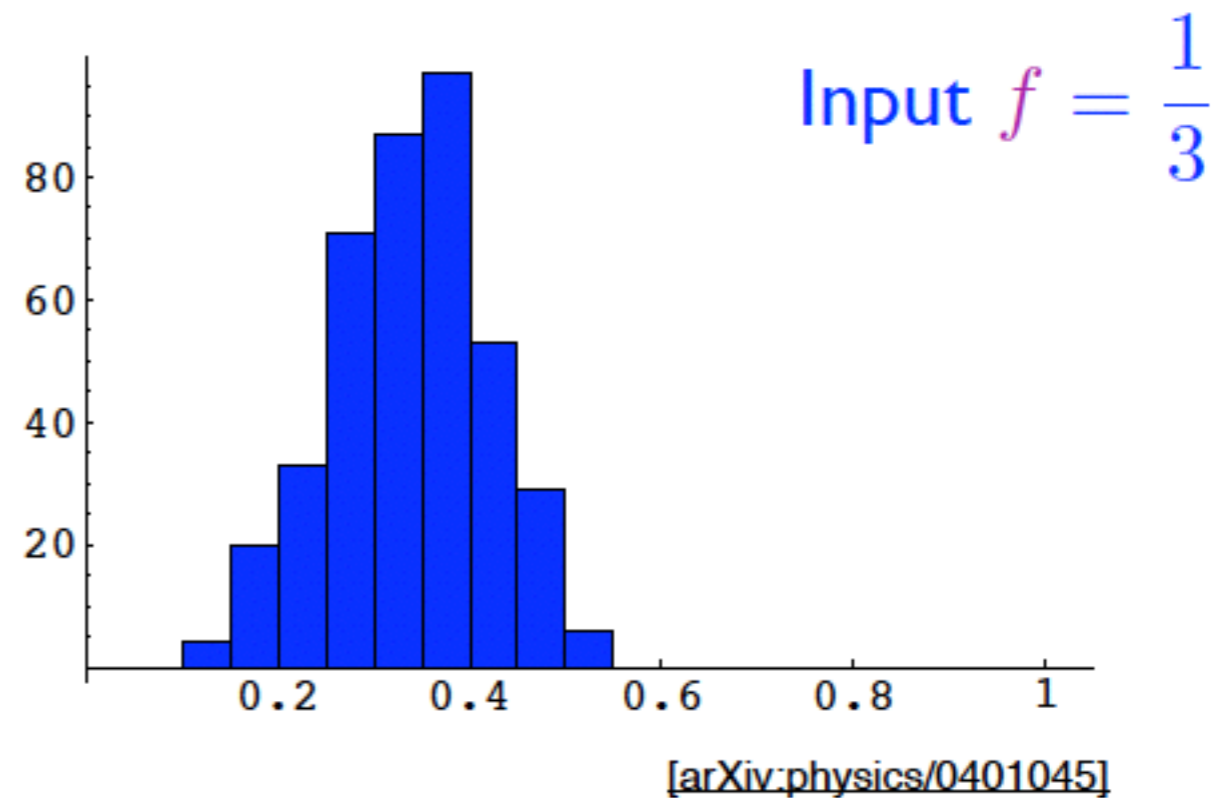
$$P(x|B) = g(x; 1, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-1)^2}{2\sigma^2}}$$

$$\mathcal{L}(f) = \prod_{i=\text{all evts}} \{f g(x_i; 0, \sigma) + (1 - f) g(x_i; 1, \sigma)\}$$

Results

- A few hundred toy experiments, 150 events each, generated with $f = 1/3$.
- Result:
 - Mean $f = 0.3337 \pm 0.0004$
 - $\sigma = 0.083$
- All good.


Fit results for f



Improving the fit with event-by-event resolution

prev.: $\mathcal{L}(f) = \prod_{i=\text{all evts}} \{f g(x_i; 0, \sigma) + (1 - f) g(x_i; 1, \sigma)\}$

now: $\mathcal{L}(f) = \prod_{i=\text{all evts}} \{f g(x_i; 0, \sigma_i) + (1 - f) g(x_i; 1, \sigma_i)\}$

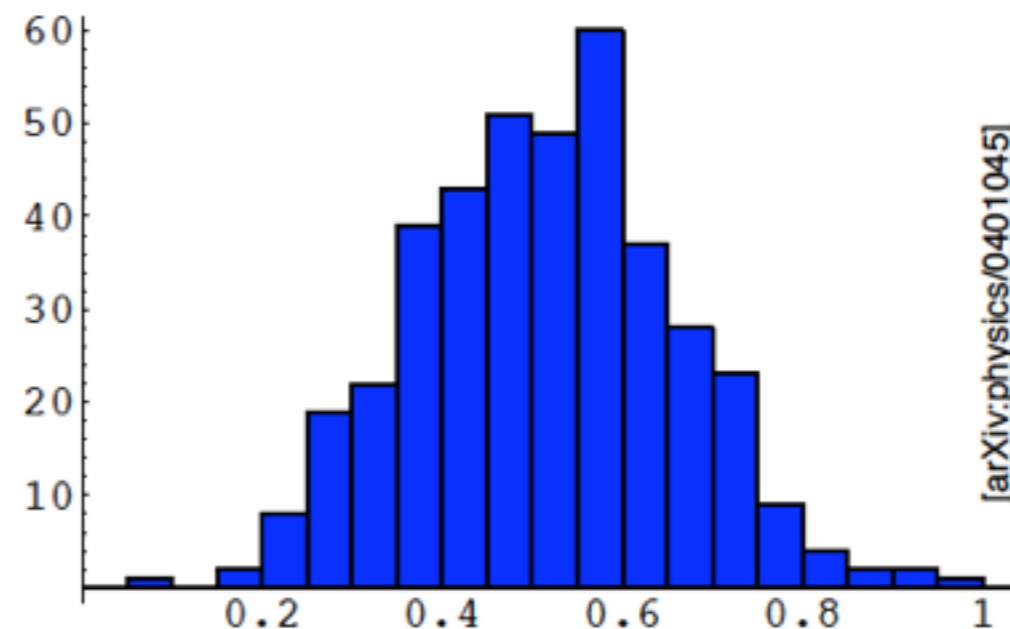


Use more information - expect better result

Results ????

$$\mathcal{L}(f) = \prod_{i=\text{all evts}} \{f g(x_i; 0, \sigma_i) + (1 - f) g(x_i; 1, \sigma_i)\}$$

- Input: $f = 1/3$.
- Result:
 - Mean $f = 0.514 \pm 0.007$
 - $\sigma = 0.14$



- What's gone so badly wrong?
- The expression $f g(x_i; 0, \sigma_i) + (1 - f) g(x_i; 1, \sigma_i)$ is not the probability to find x_i , $P(x_i)$. It is the probability to find x_i given σ_i , $P(x_i|\sigma_i)$.
- But: $P(x) \neq P(x|y)$

Correcting the pdf

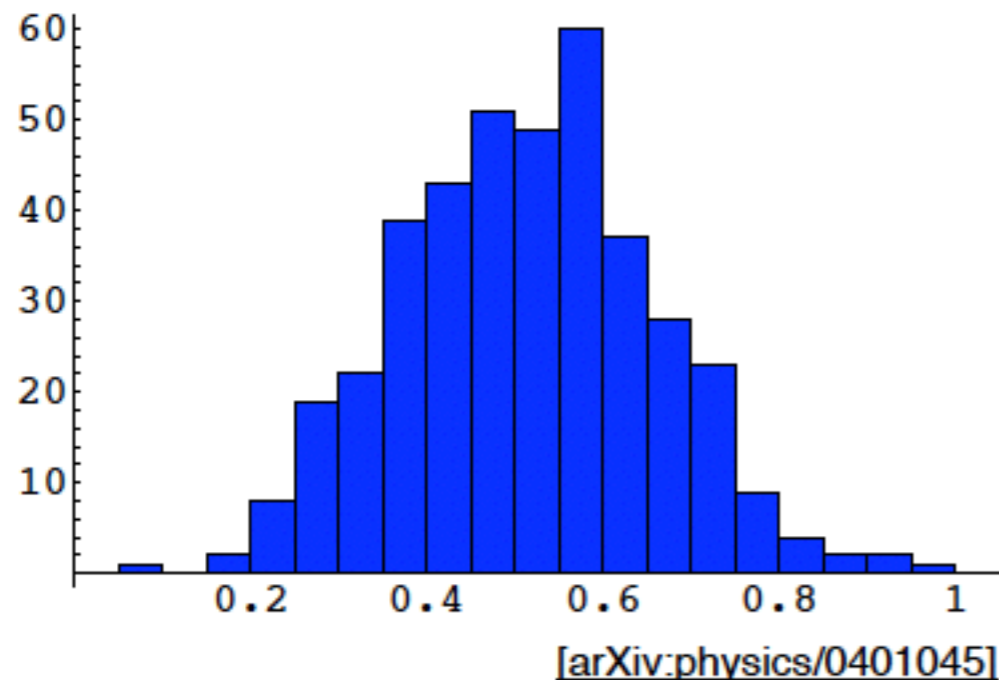
- And hence the correct likelihood

$$\mathcal{L}(f) = \prod_i \{f P(x_i|\sigma_i, A) \cdot P(\sigma_i|A) + (1-f) P(x_i|\sigma_i, B) \cdot P(\sigma_i|B)\}$$

Wrong Likelihood

Mean=0.514 ± 0.007

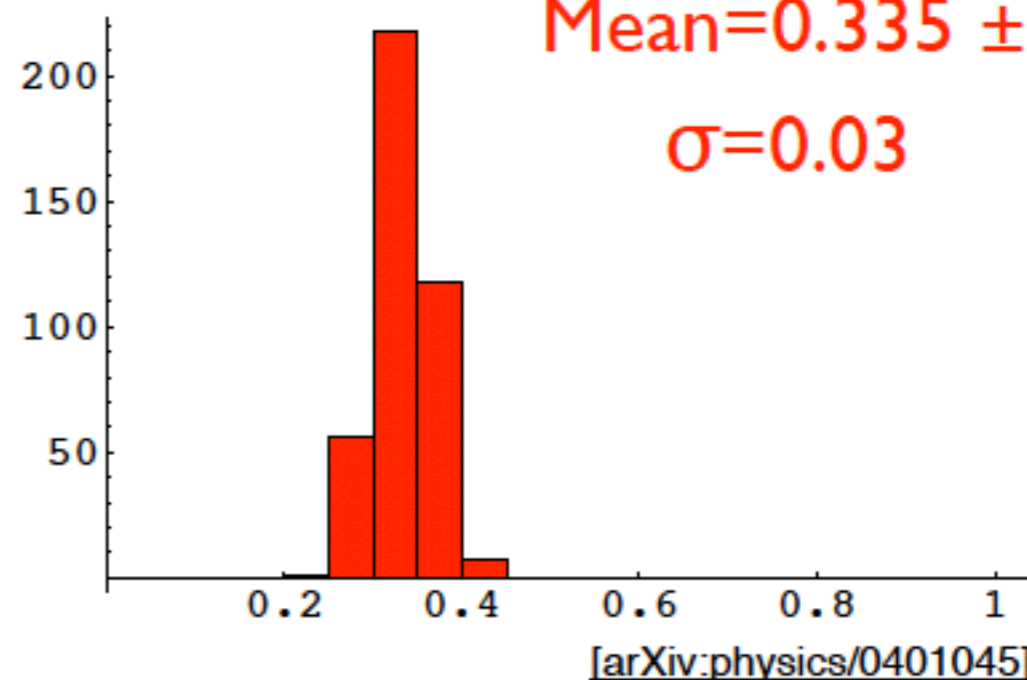
$\sigma=0.14$



Correct Likelihood

Mean=0.335 ± 0.002

$\sigma=0.03$



MLE biases

The pdf should explicitly include the distribution of any observable that enters the event-by-event.

If you don't include explicitly, the fit will do it for you implicitly, assuming that such distribution is the same for all classes of events in your sample.

If that's not the case, your fit results might get strongly biased.