

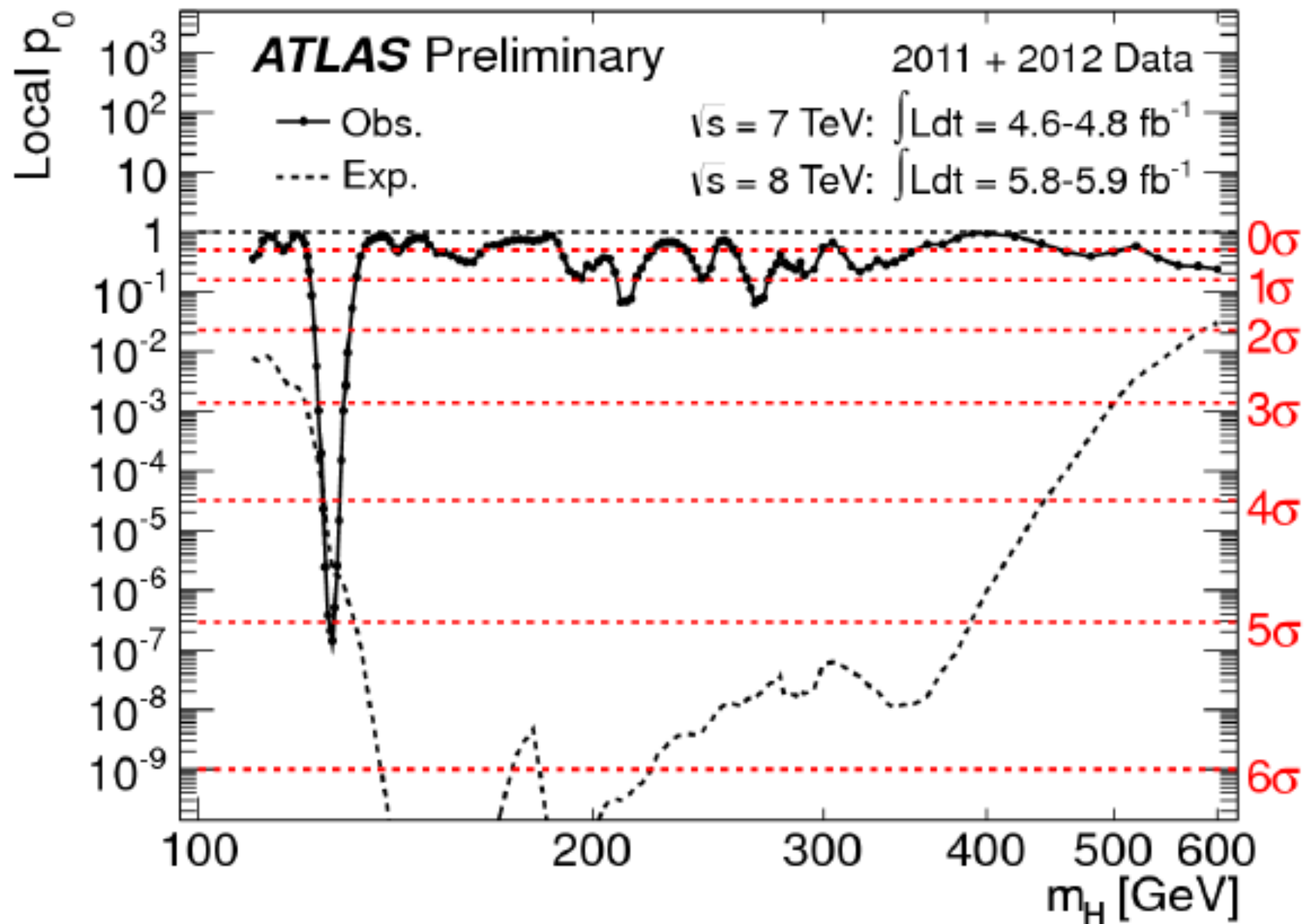
# Statistics for HEP (2/3)

---

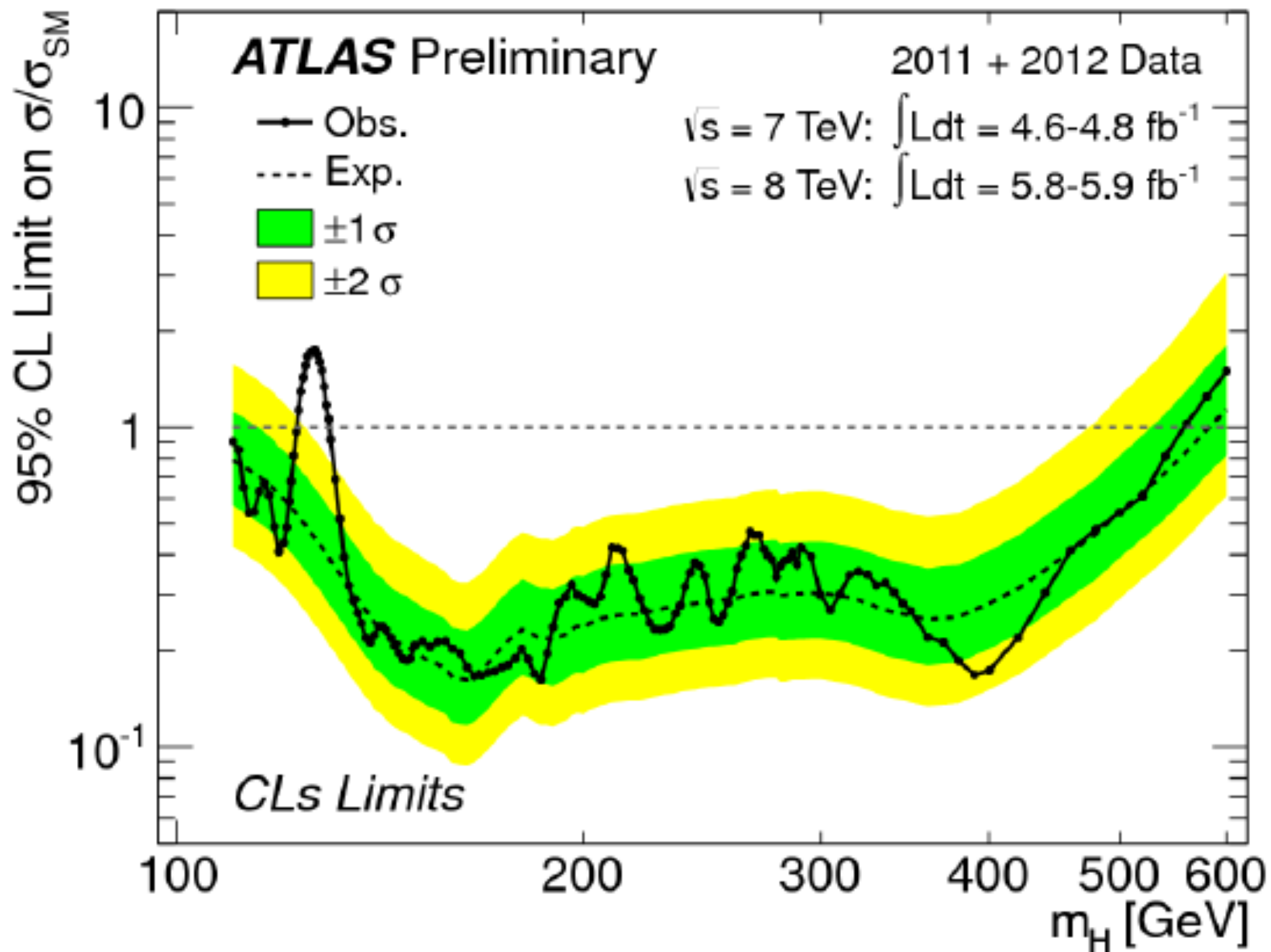
Diego Tonelli (INFN Trieste)  
[diego.tonelli@cern.ch](mailto:diego.tonelli@cern.ch)

*CERN-Fermilab HCP Summer School*  
*Aug 29, 2017*

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?



# Confidence intervals

# Confidence intervals

---

Mathematical procedure to address the question:

*Given a model  $p(x|m)$ , with unknown  $m$ , and observed data  $x_0$ , what are the values of  $m$  for which the observed value  $x_0$  is among the least extreme of all possible values of  $x$ ?*

# Confidence intervals

---

*What are the values of  $m$  for which the observed value  $x_0$  is among the least extreme possible values of  $x$ ?*

To define “extreme”, need an ordering principle. Rank the values of  $x$  for each possible value of  $m$ . High rank means not extreme (likely to be included in the interval). Low rank means extreme (likely to be outside of the interval).

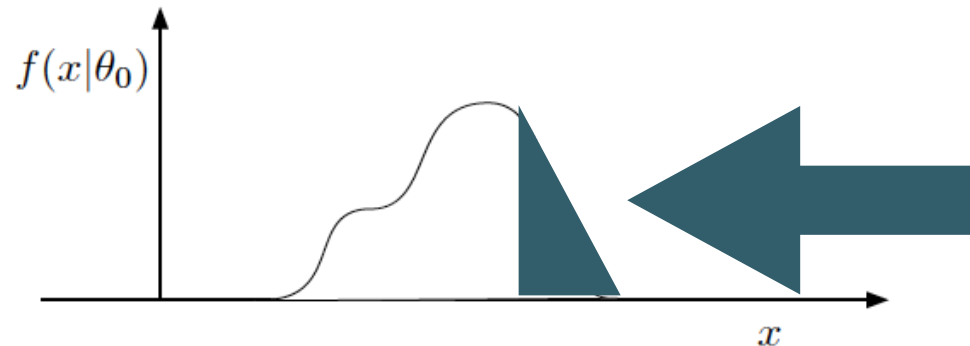
With that ordering, accumulate the values of highest-ranked (i.e., less extreme) values of  $x$  until you reach a predetermined fraction of  $x$  probability. Such fraction is the confidence level (CL). Typically 68%, 95%...

*Given a model  $p(x|m)$ , data  $x_0$ , an ordering, and a CL, the confidence interval  $[m_1, m_2]$  includes those values of  $m$  for which  $x_0$  aren't “extreme” at the chosen CL*

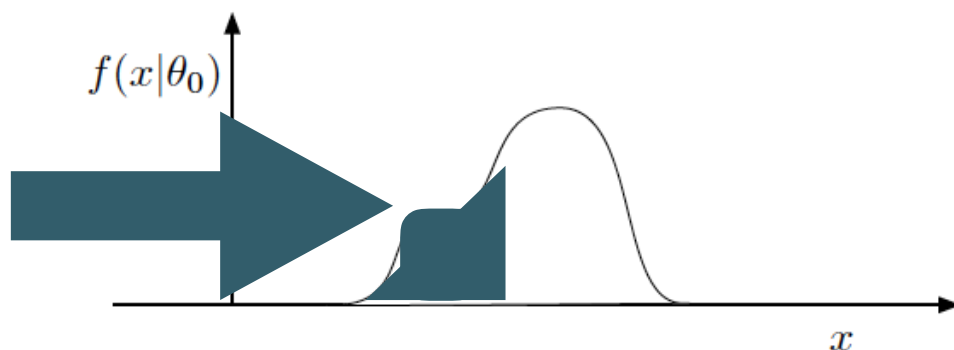
For example:  $[m_1, m_2]$  determined at 68% CL includes the values of  $m$  for which the observed data  $x_0$  belongs to the least extreme 68% values of  $x$

# One-sided, two-sided.

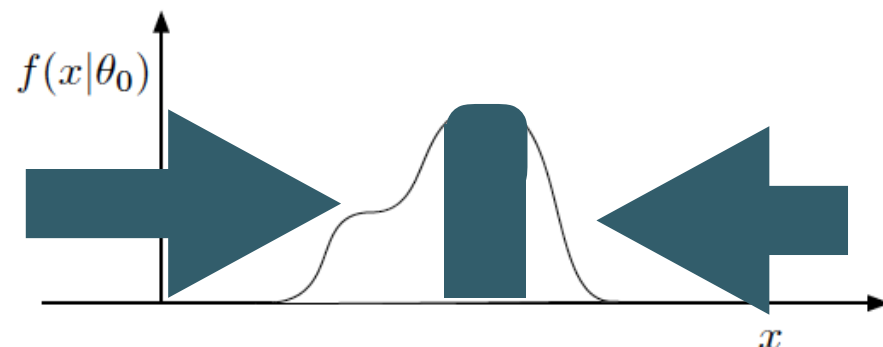
---



If “extreme” is defined as low-valued  $x$ , start accumulating from high values of  $x$ . Yields one-sided interval (upper limit on  $m$ )



If “extreme” is defined as high-valued  $x$ , start accumulating from low values of  $x$ . Yields one-sided interval (lower limit on  $m$ )



If “extremes” are high- and low-valued  $x$ , take the smallest central quantile. Yields central interval (lower limit on  $m$ )

(simplified interpretation applies only to one-dimensional  $x$ , and  $p(x|m)$  is such that higher values of  $m$  imply higher average  $x$ )

# CL

---

The confidence level is usually chosen to match the standard thresholds 68.3% ( $1\sigma$ ) 95.5% ( $2\sigma$ ) etc. Define also the lowest-ranked  $\alpha = 1 - \text{CL}$  fraction of the most extreme values

The endpoints of a central confidence interval at given CL can be determined from one-sided confidence intervals (lower and upper limits) at CL/2

A CL=84% upper limit  $m_2$  *excludes*  $m$  values for which  $x_0$  belongs to the set of lowest-valued that has 16% (1-CL) probability

A CL=84% lower limit  $m_1$  *excludes*  $m$  values for which  $x_0$  belongs to the set of highest-valued  $x$  set that has 16% (1-CL) probability

Then  $[m_1, m_2]$  includes the central 68% fraction of  $x$  values ordered from high to low: a  $1 - (16\% + 16\%) = 68\%$  central confidence interval



# Confidence intervals

---

## Confidence intervals for binomial parameter $\rho$ Directly relevant to efficiency calculation in HEP

Let  $\text{Bi}(n_{\text{on}} | n_{\text{tot}}, \rho)$  denote binomial probability of  $n_{\text{on}}$  successes in  $n_{\text{tot}}$  trials, each with **binomial parameter**  $\rho$ :

$$\text{Bi}(n_{\text{on}} | n_{\text{tot}}, \rho) = \frac{n_{\text{tot}}!}{n_{\text{on}}! (n_{\text{tot}} - n_{\text{on}})!} \rho^{n_{\text{on}}} (1 - \rho)^{(n_{\text{tot}} - n_{\text{on}})}$$

In repeated trials,  $n_{\text{on}}$  has **mean**  $n_{\text{tot}} \rho$  and

**rms deviation**  $\sqrt{n_{\text{tot}} \rho (1 - \rho)}$

With observed successes  $n_{\text{on}}$ , **the M.L. point estimate  $\hat{\rho}$  of  $\rho$  is**

$$\hat{\rho} = n_{\text{on}} / n_{\text{tot}} .$$

**What confidence interval  $[\rho_1, \rho_2]$  should we report for  $\rho$ ?**

# Confidence intervals

---

Suppose we observe 3 successes on 10 trials. What is our efficiency and its uncertainty?

It is tempting to replace  $\hat{p} = 0.30$  into  $\hat{\sigma} = (1/n_{\text{tot}})\sqrt{\hat{p}(1-\hat{p})}$  and obtain the interval  $[\rho_1, \rho_2] = \hat{p} \pm \hat{\sigma}$

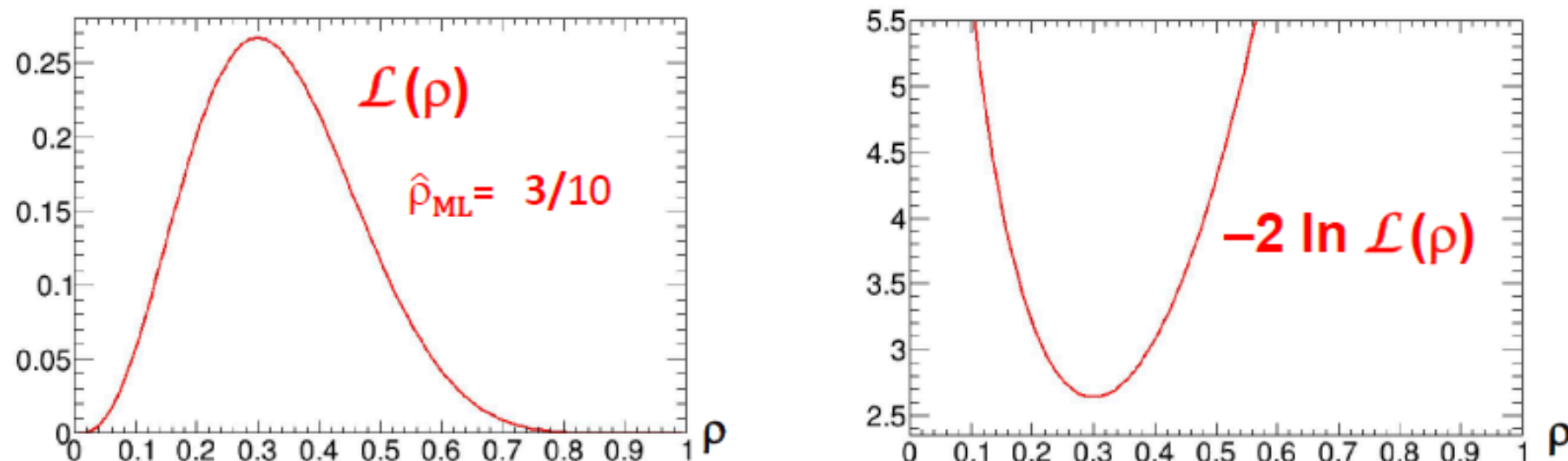
This is not a confidence interval since it does not follow the proper logic of a frequentist inference. In the construction of the interval each  $\sigma$  should be consistently associated with each  $p$

This is manifest for the cases in which  $n_{\text{on}} = n_{\text{tot}}$  or  $n_{\text{on}} = 0$ .

# Confidence intervals

## Confidence intervals for binomial $\rho$ (cont.)

Suppose  $n_{\text{on}}=3$  successes in  $n_{\text{tot}}=10$  trials.



Let's find exact 68% C.L.\* *central* confidence interval  $[\rho_1, \rho_2]$ .

Recall shortcut above for central intervals:

Find lower limit  $\rho_1$  with C.L. =  $1 - (1 - 68\%)/2 = 84\%$

I.e., Find  $\rho_1$  such that  $\text{Bi}(n_{\text{on}} < 3 \mid n_{\text{tot}}=10, \rho_1) = 84\%$

Find upper limit  $\rho_2$  with C.L. = 84%

I.e., Find  $\rho_2$  such that  $\text{Bi}(n_{\text{on}} > 3 \mid n_{\text{tot}}=10, \rho_2) = 84\%$

# Confidence intervals

$n_{\text{on}} = 3$  ,  $n_{\text{tot}} = 10$ .

Find  $\rho_1$  such that

$\text{Bi}(n_{\text{on}} < 3 \mid \rho_1) = 84\%$

$\text{Bi}(n_{\text{on}} \geq 3 \mid \rho_1) = 16\%$

(lower limit at 84% C.L.)

Solve:  $\rho_1 = 0.142$

And find  $\rho_2$  such that

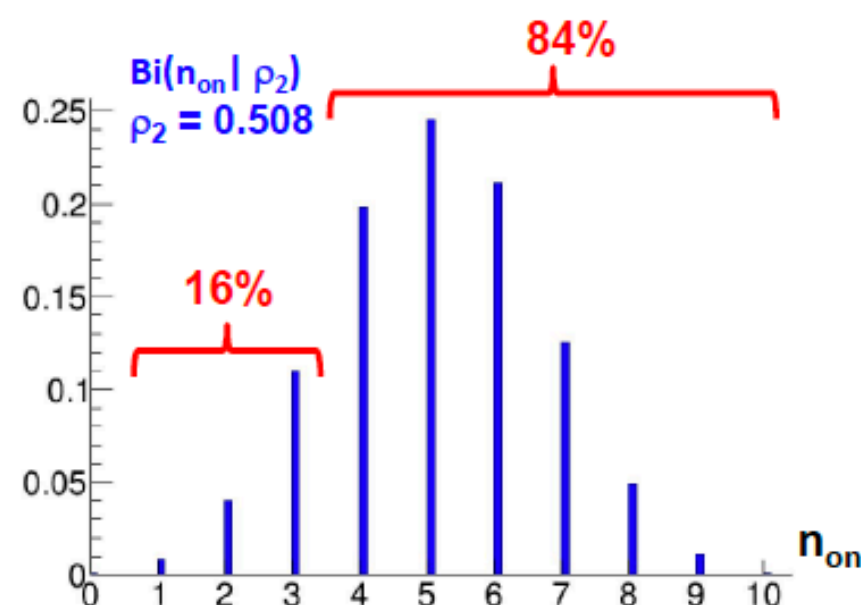
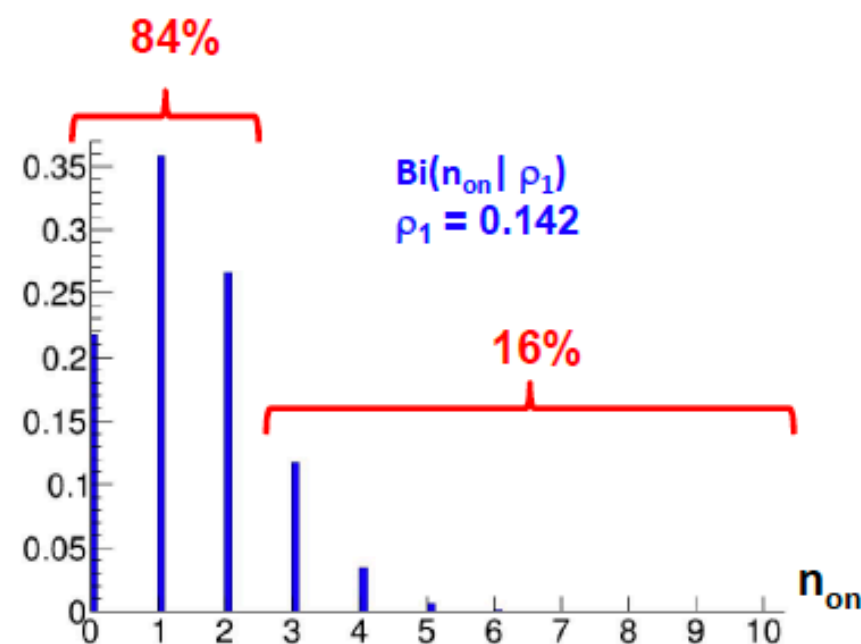
$\text{Bi}(n_{\text{on}} > 3 \mid \rho_2) = 84\%$

$\text{Bi}(n_{\text{on}} \leq 3 \mid \rho_2) = 16\%$

(upper limit at 84% C.L.)

Solve:  $\rho_2 = 0.508$

Then  $[\rho_1, \rho_2] = (0.142, 0.508)$   
is *central* confidence interval  
with 68% C.L. Same as  
Clopper and Pearson (1934)



# Neyman construction

---

J. Neyman came up with a mathematically rigorous 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

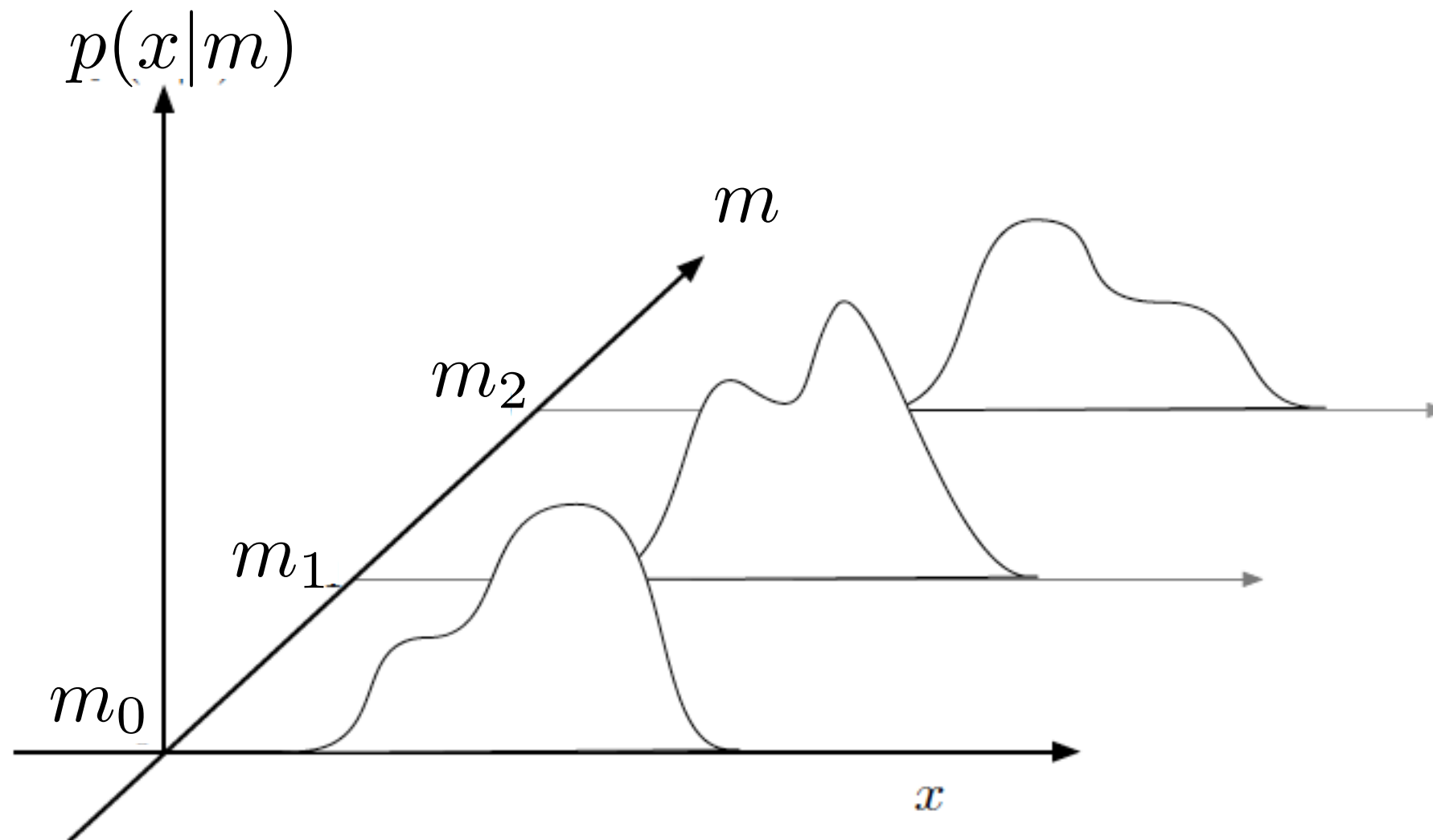
*By* J. NEYMAN

*Reader in Statistics, University College, London*

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

# Neyman construction illustrated

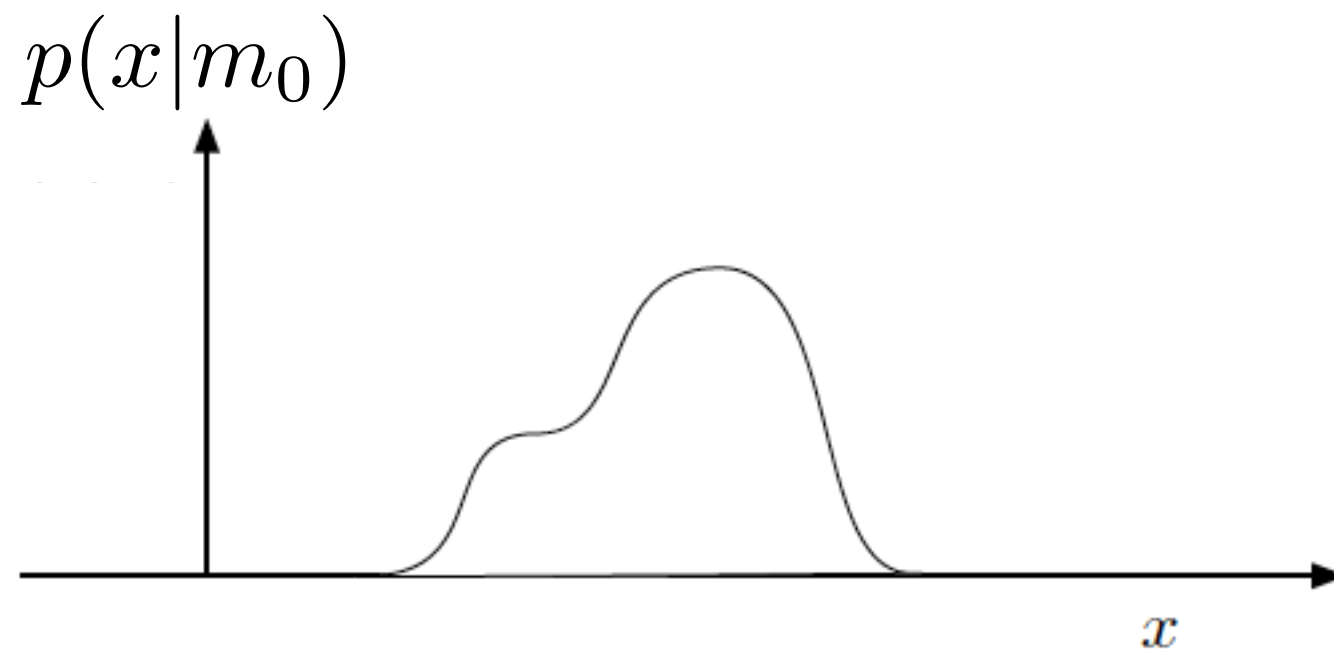
*Prior to looking at data*, for each possible true value of parameter  $m$ , consider  $p(x|m)$ . Its shape can vary as a function of  $m$ .



# Neyman illustrated I

---

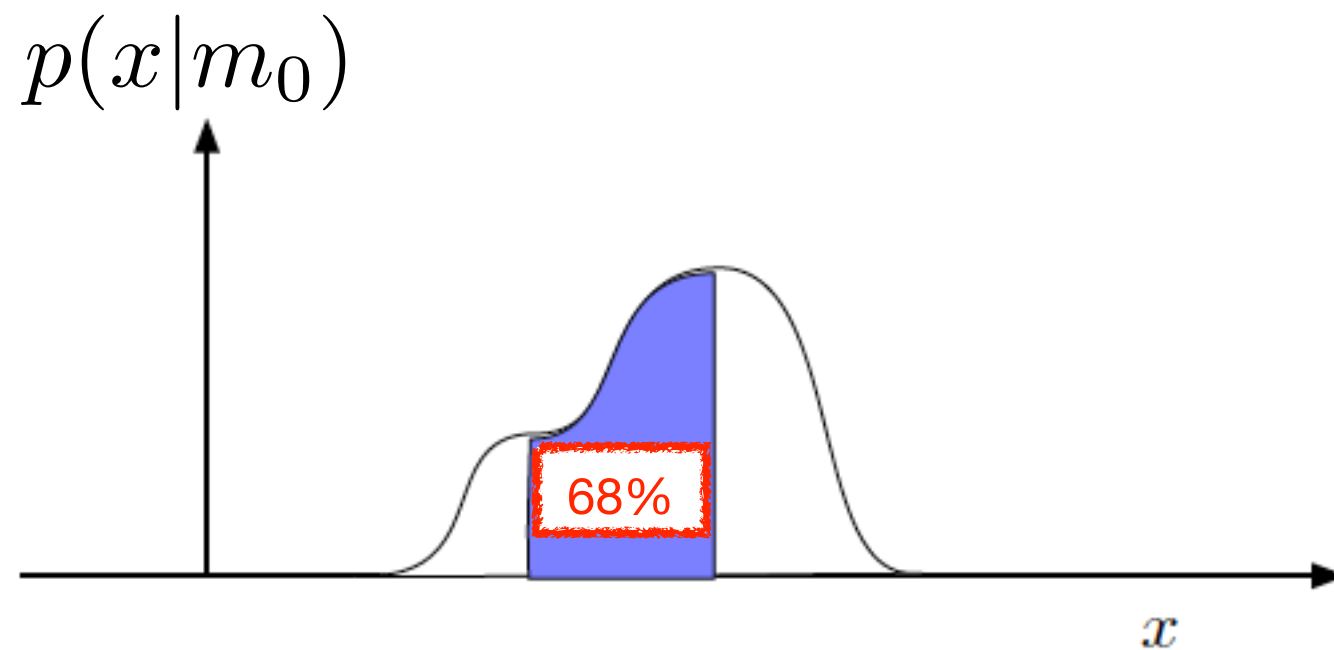
Take a specific value  $m_0$  of the parameter



# Neyman illustrated II

---

Use  $p(x|m_0)$  to define an acceptance range in  $x$ , such that  $p(x \in \text{range} \mid m_0) = 68\%$ .





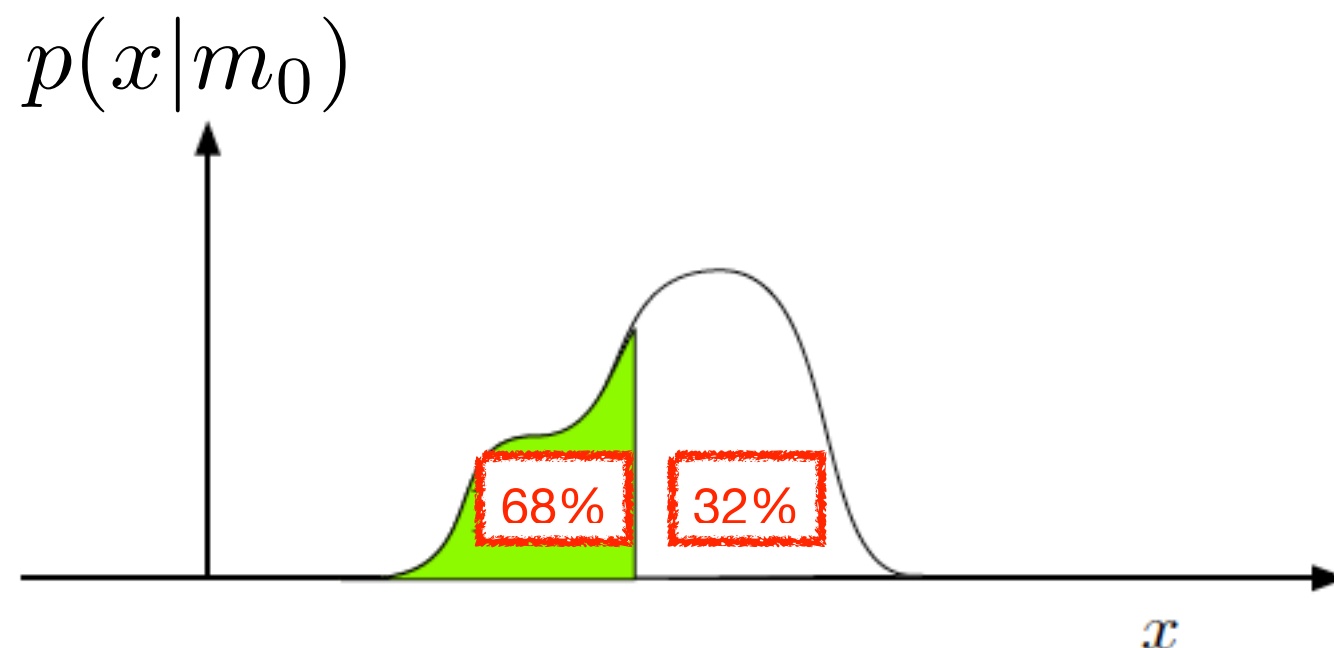
# Neyman illustrated III

---

The definition of the acceptance range is not unique

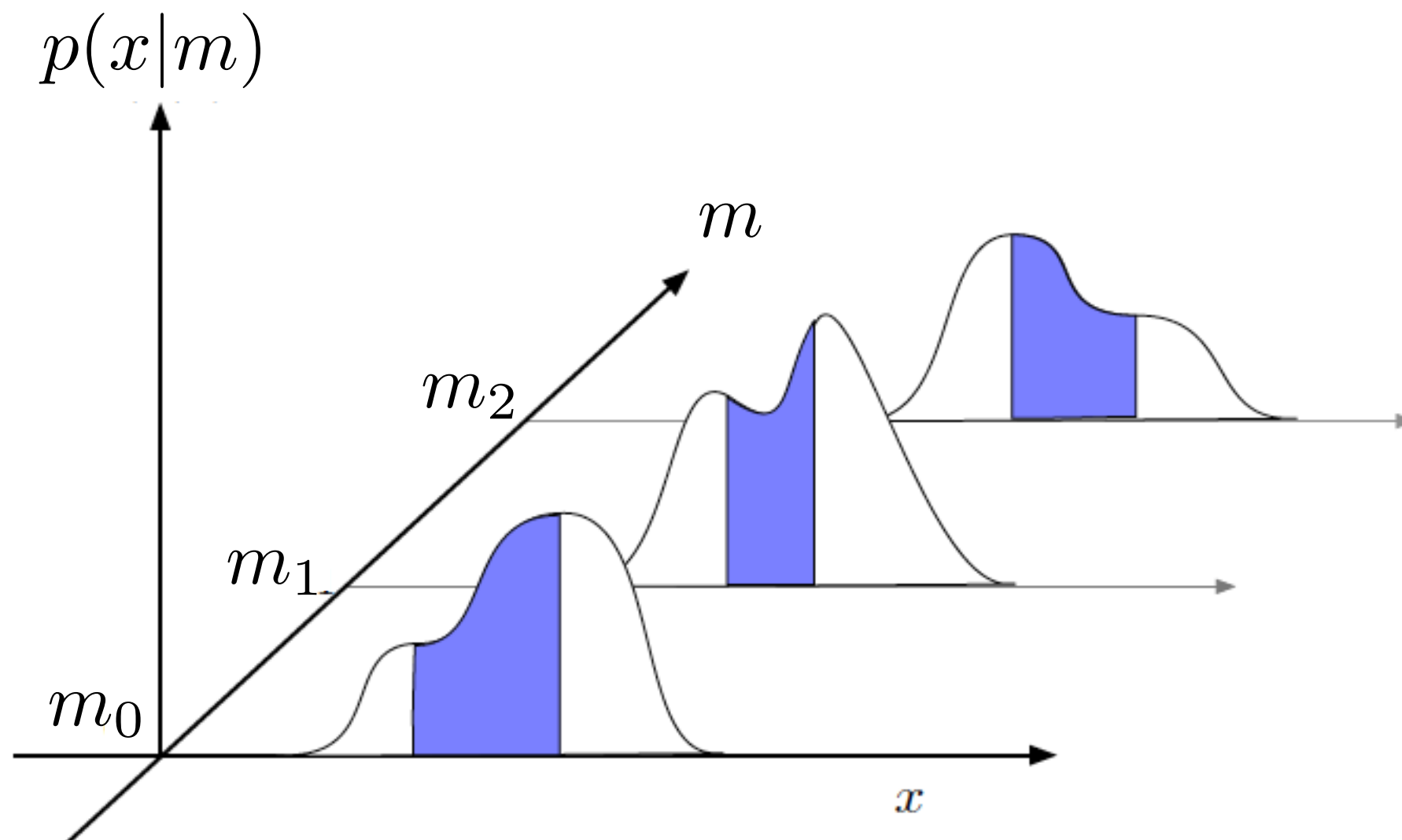
The criterion to choose of the region is chosen is the *ordering rule*

The rule defining the *order* of accumulation of the elements along  $x$  until the desired amount of probability, corresponding to the chosen confidence level (68%, in our example), is accumulated.



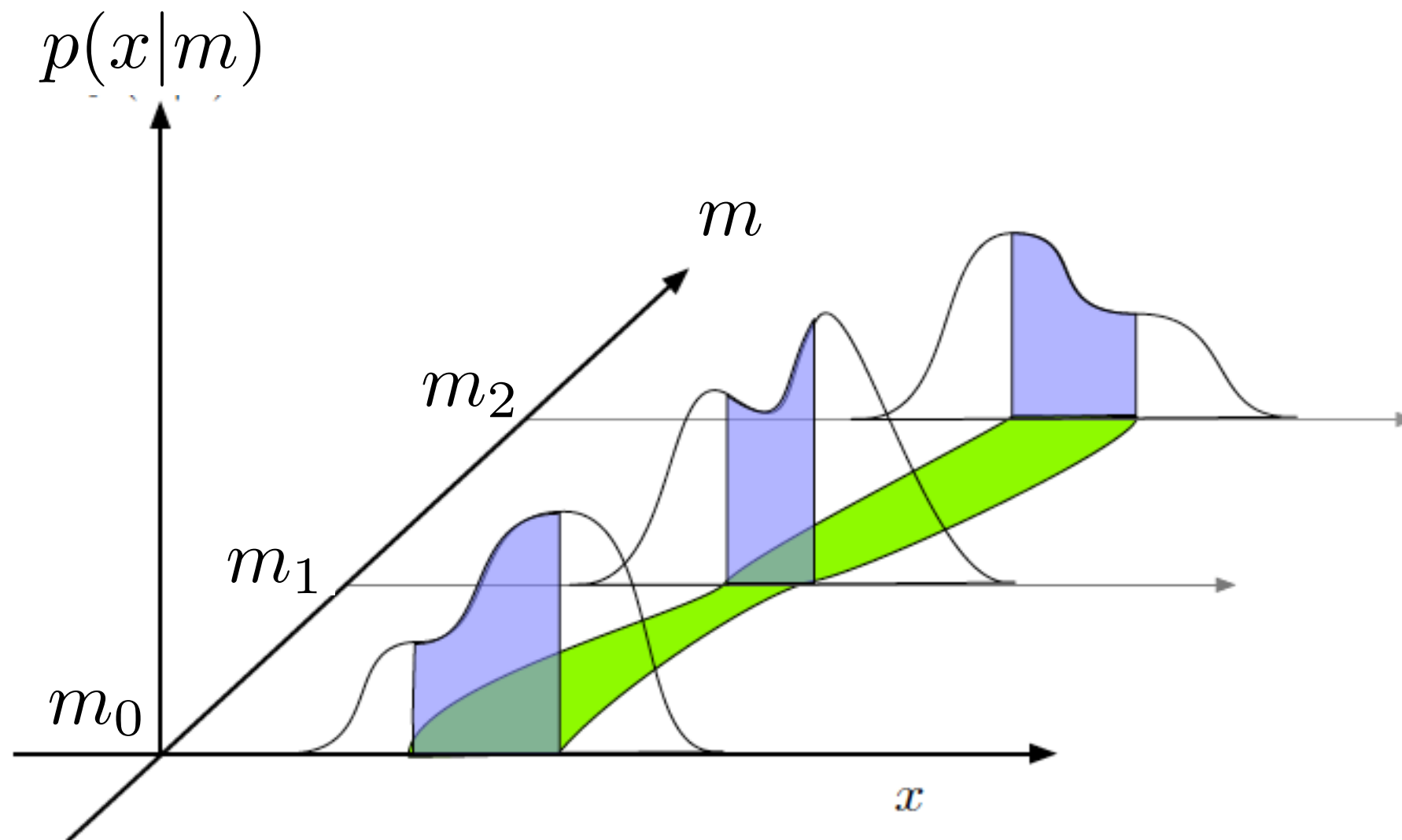
# Neyman illustrated V

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



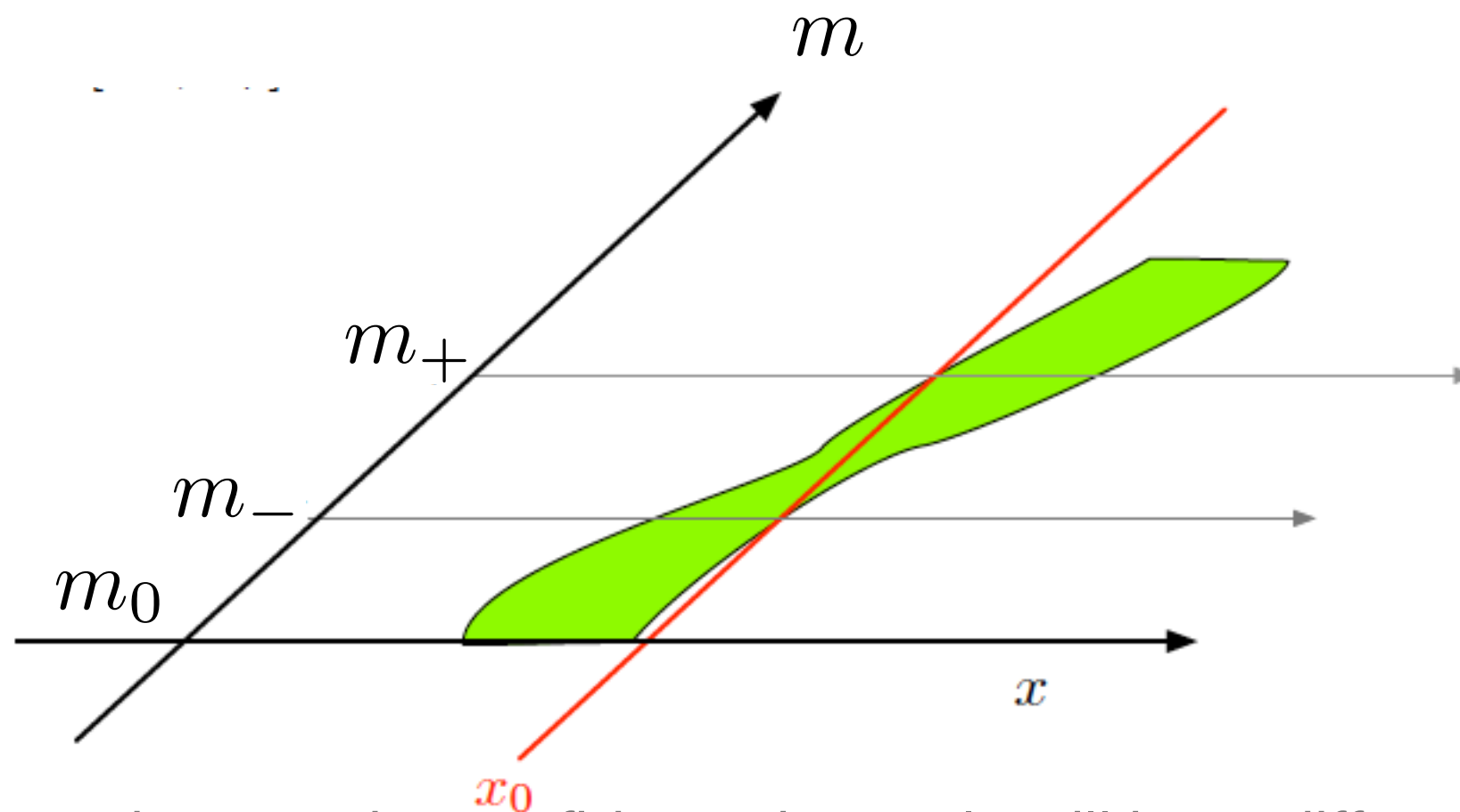
# Neyman illustrated VI

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  $x$ )

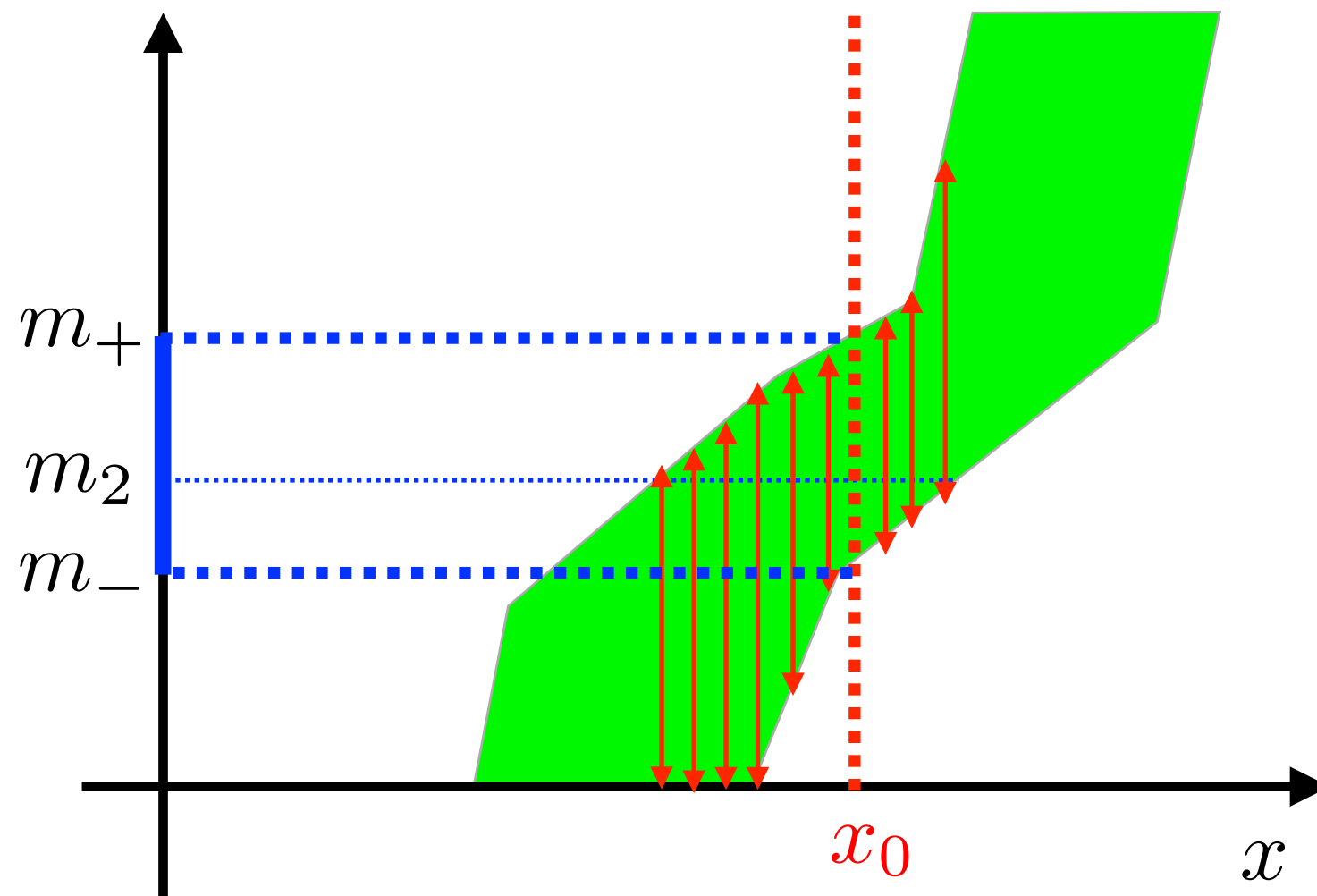


[Cranmer]

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.



“projection of the acceptance region onto the space of parameters” — a set-theory union, not an integral.

# Toy example

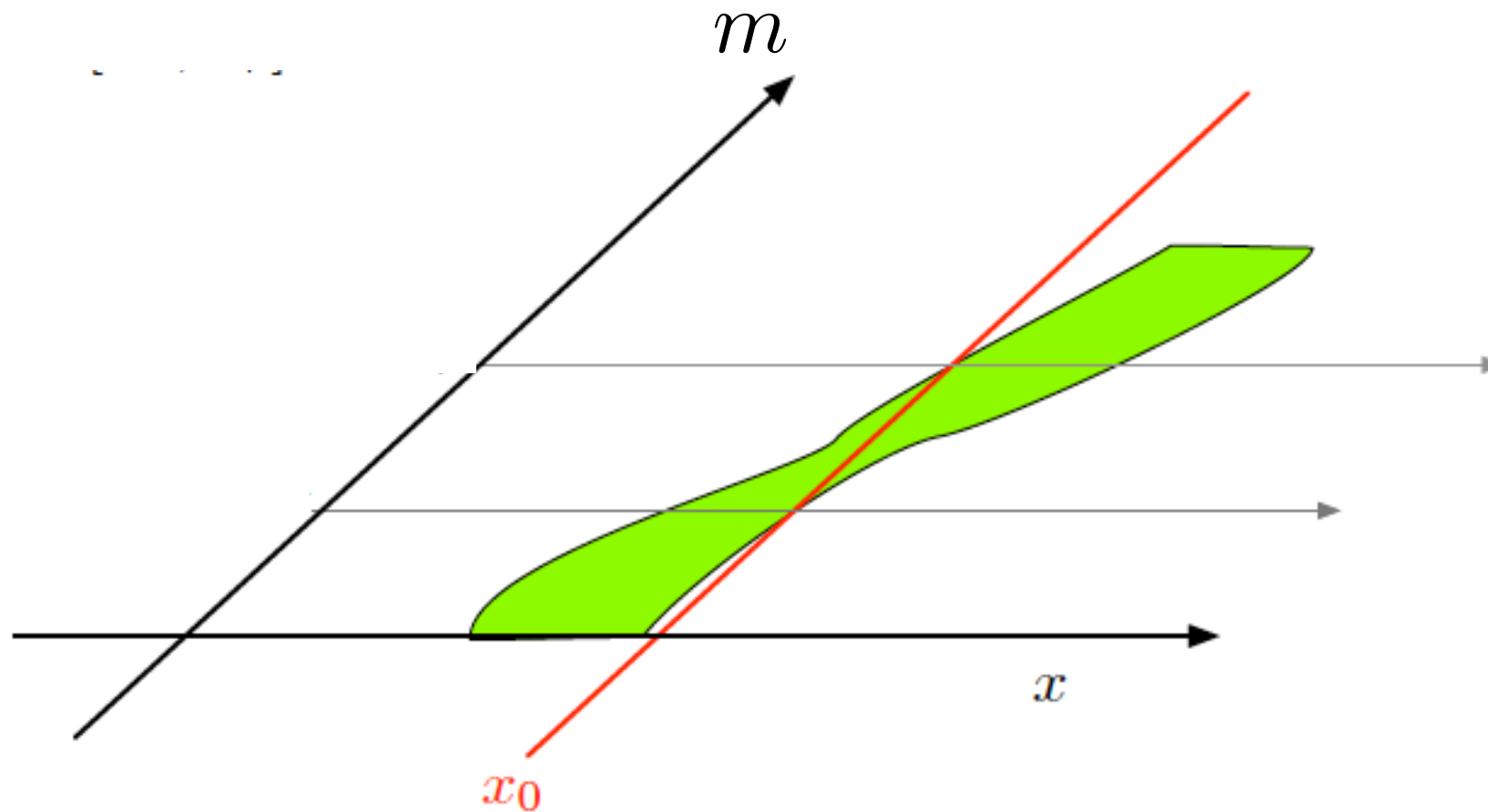
Bags of various classes: each class contains a different fraction of white balls (1%, 5%, 50%, 95%, and 99%). Extract 5 balls from a bag and infer to which class the bag belongs

True fraction of white balls						
	Class A = 1%	Class B = 5%	Class C = 50%	Class D = 95%	Class E = 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%	$3 \cdot 10^{-7}$	$10^{-10}$

# Note

---

For simplification purposes, examples discussed have one-dimensional space of parameter and one-dimensional space of observables and  $p(x|m)$  such that the higher the  $m$  the higher the  $x$ .



In general,  $x$  and  $m$  are  $\vec{x}$  and  $\vec{m}$  and they need not to have same ranges, units, or dimensionality

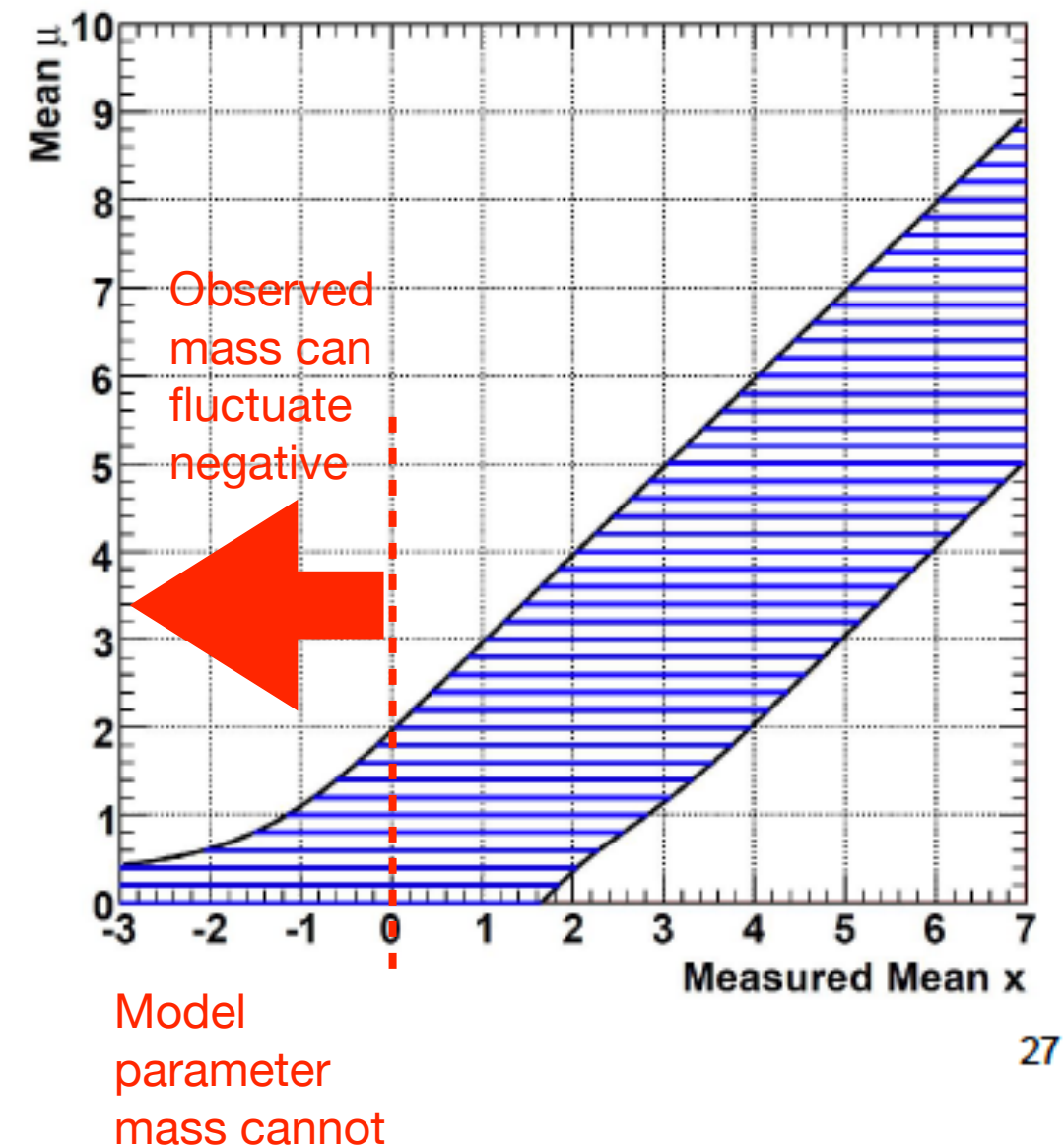
# “Non-physical” ranges

It's frequent to get confused about the ranges for the confidence band construction.

Example: measurement of a small mass  $m$ . using a Gaussian  $p(x|m)$  with  $x$  observed mass.

Keep distinct

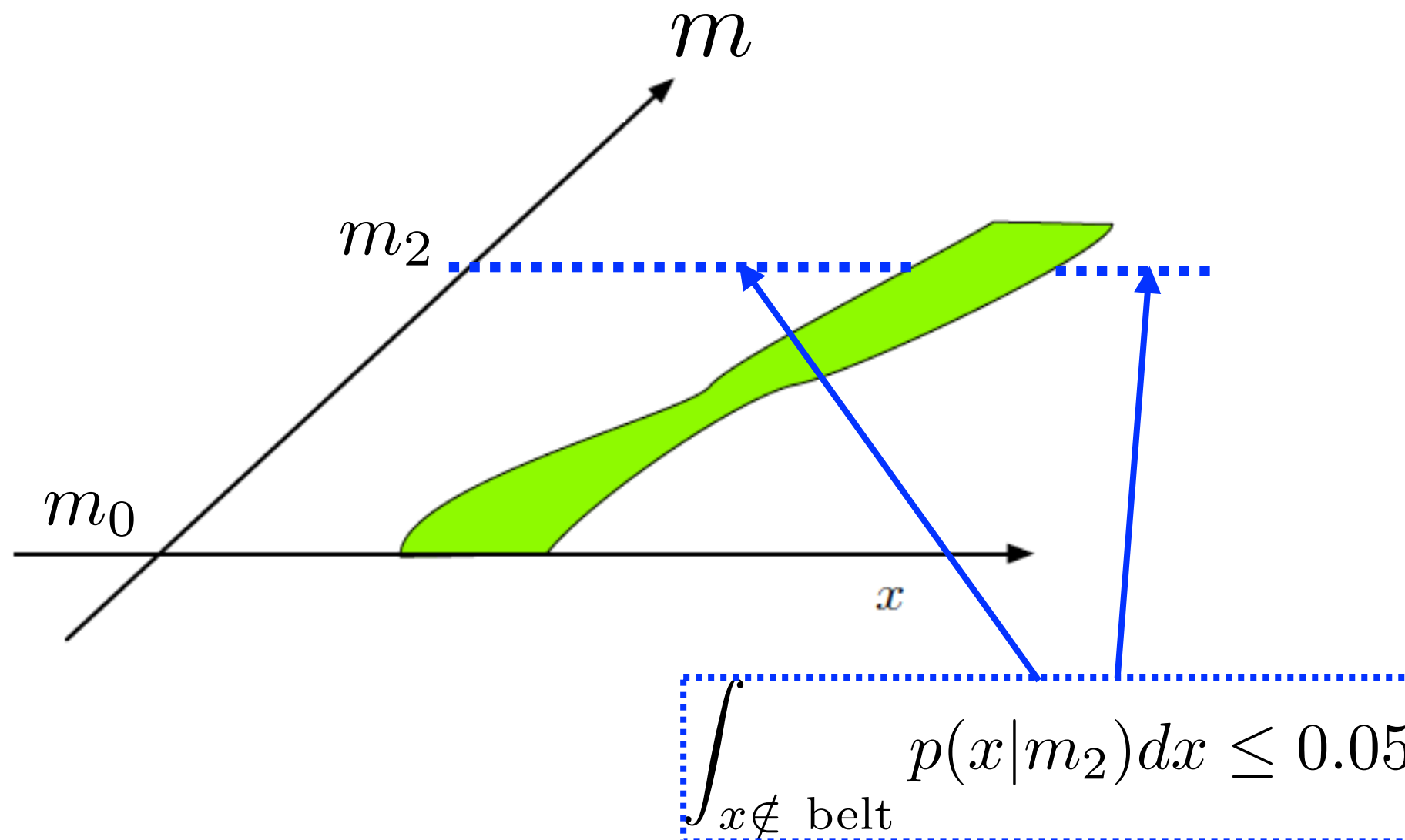
- data  $x$  which, due to resolution, could fluctuate negative
- the mass parameter  $m$ , for which negative values do not exist in the model





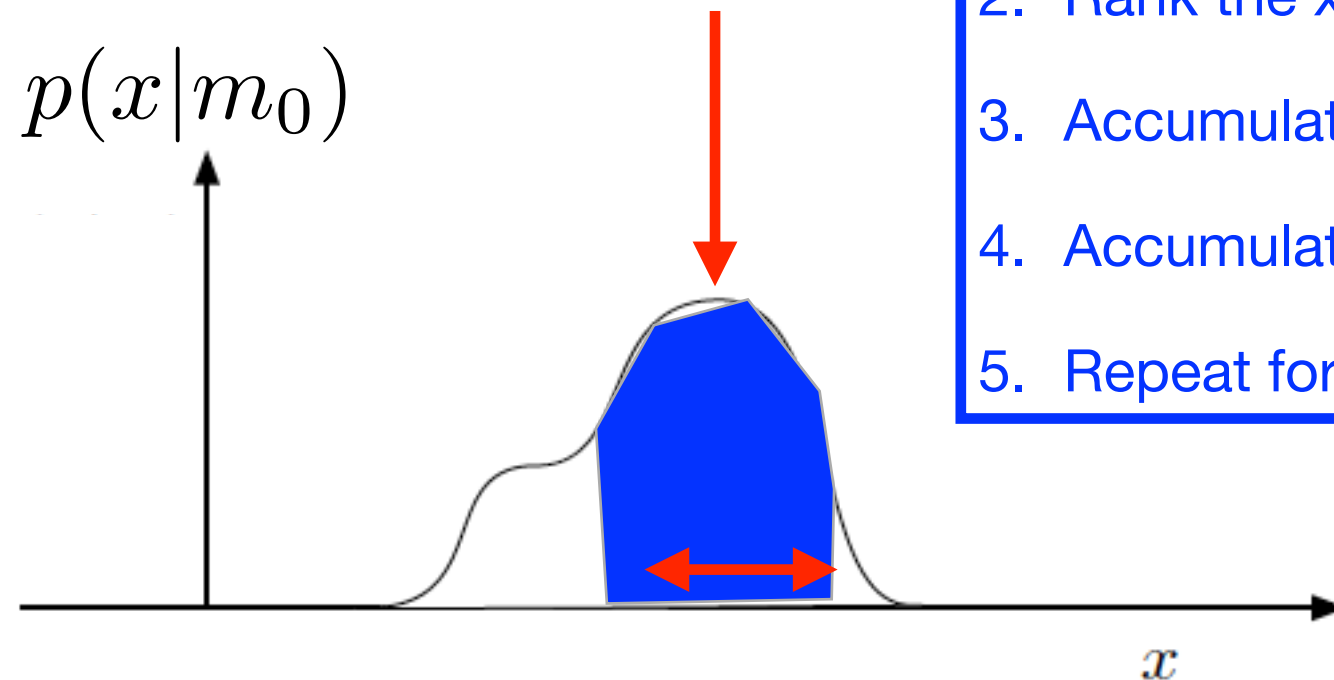
# Ordering

The ordering algorithm is arbitrarily chosen, provided that (i) has been **defined prior to look at the data** (ii) for each value  $m$  of the parameter, the integral of the pdf along the  $x$  region outside of the belt does not exceed  $1-CL$ .



# Probability ordering

In the past, many tried to get the shortest possible interval, so that the resulting confidence intervals were likely narrower yielding more precise measurements. (this is the probability ordering or “Crow-Gardner ordering”)

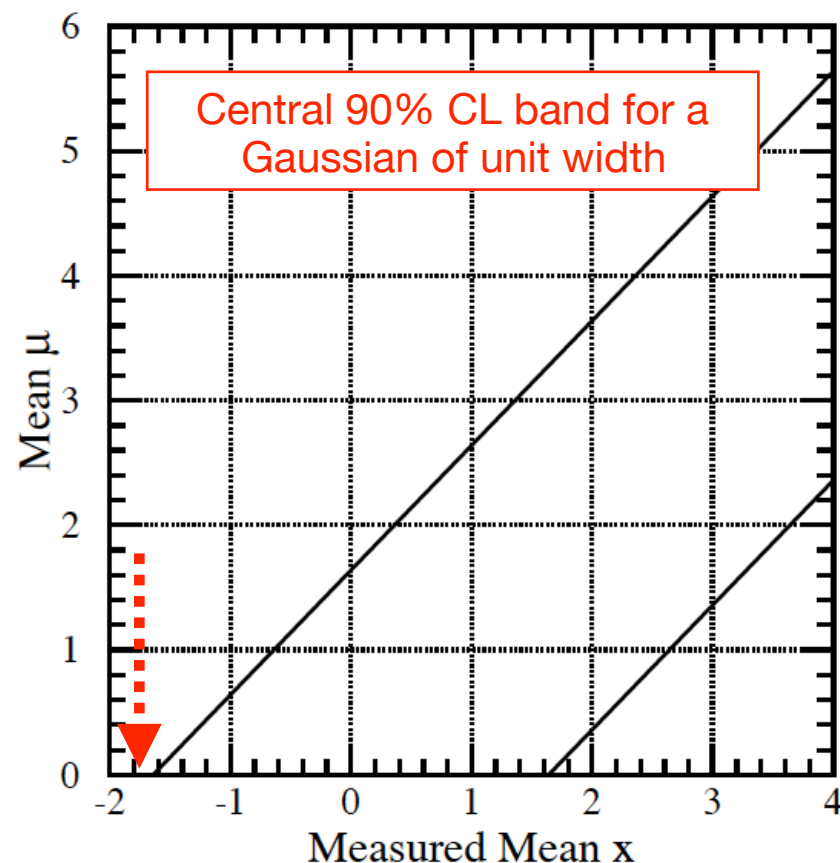


1. Choose one value for  $m$ ,  $m_0$ , and look at  $p(x|m_0)$
2. Rank the  $x$  values in decreasing order of  $p(x|m_0)$
3. Accumulate  $x$  starting from the  $x$  with highest probability
4. Accumulate all other  $x$  until the desired CL is reached.
5. Repeat for all  $m$

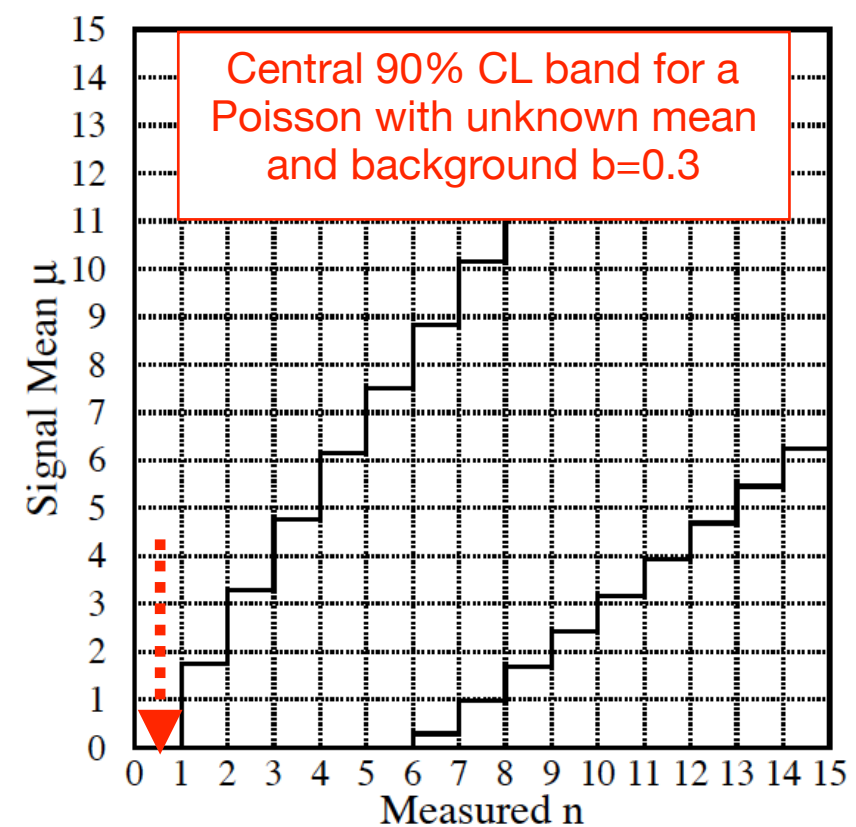
This is ill-defined: as probability **depends on the metric** for the observable  $x$ , the shortest interval in one metric isn't shortest in others.

# Issues

Long-standing inconsistencies found in Neyman constructions based on simplistic ordering criteria (i) Gaussian measurement resolution near a physical boundary (e.g., like a measurement of neutrino mass square close to zero) (ii) measurements of a Poisson signal in the presence of background when observed number of events fluctuates below the expected background count.



What if one observes  $x = -1.8$ ? or  $n = 0$ ?



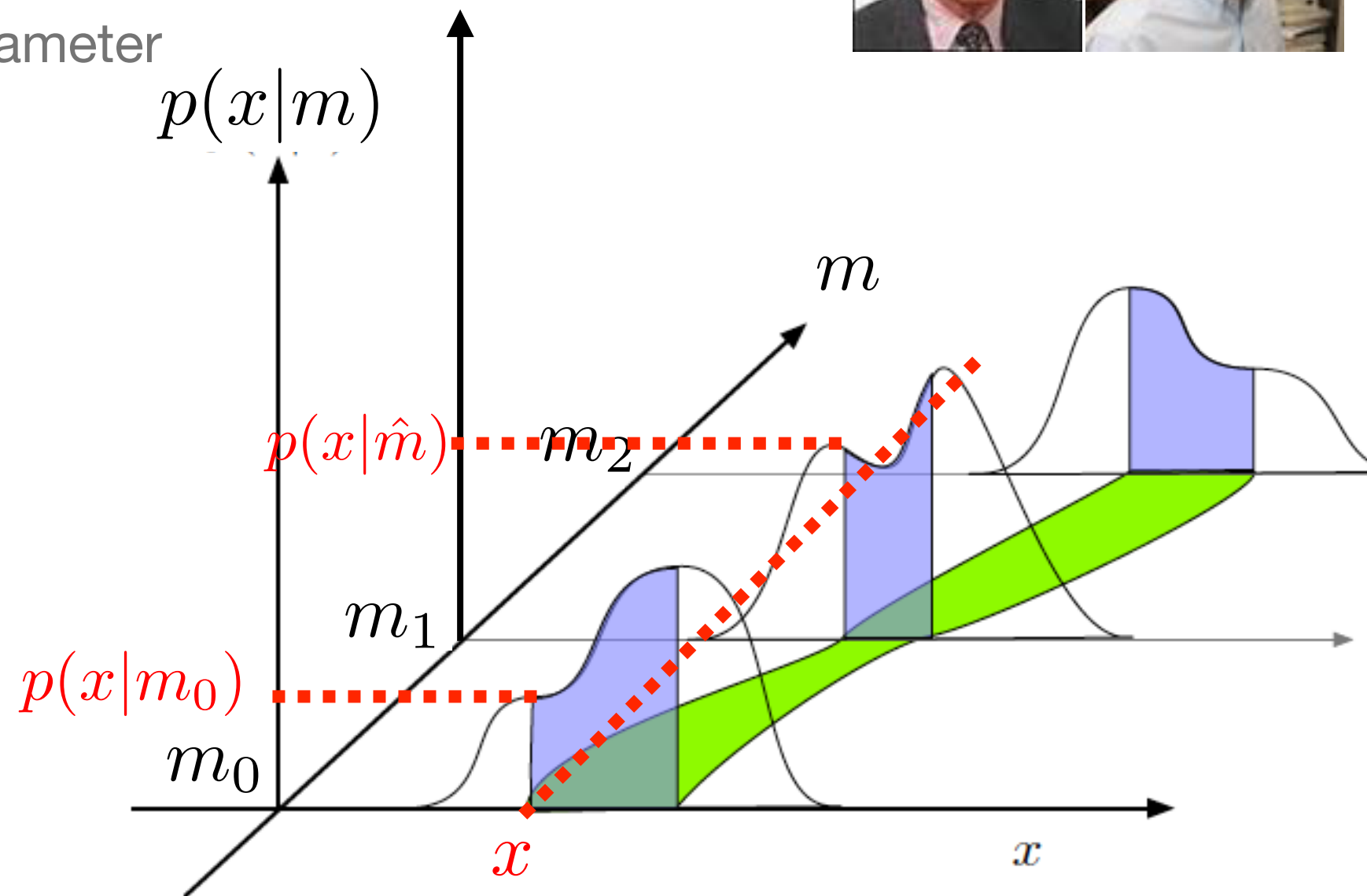
The resulting confidence regions are empty, which is clearly indicative of a problem.

# Likelihood-ratio ordering (“Feldman and Cousins”)

Those issues were solved by adapting a more ordering, based on **the likelihood ratio**

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

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



The “accumulation score” of each element in  $x$ , no longer depends only on  $p(x|m_0)$  but also on  $p(x|m)$  at other  $m$  values



# Likelihood-ratio ordering

---

1. Choose one value for  $m$ ,  $m_0$  and generate simulated 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 likelihood  $L(\hat{m})$  over the space of  $m$  values.
3. Rank all  $x$  in decreasing order of likelihood ratio  $LR=L_x(m_0)/L_x(\hat{m})$ .
4. Accumulate starting from the  $x$  with higher LR until the desired CL is reached.
5. Repeat for all  $m$

As the likelihood is metric-invariant so is the ratio of likelihoods. Therefore LR-ordering preserves the metric, mostly avoids empty confidence regions and has several other attractive features. By far the most popular ordering in HEP.

Take LR-ordering as default option unless there are strong motivations against it.

# Likelihood-ratio ordering practice

It is instructive to trying to reproduce LR bands as per the original paper. <http://arxiv.org/pdf/physics/9711021v2.pdf>. Further useful and interesting info in <http://users.physics.harvard.edu/~feldman/Journeys.pdf>

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

$n$	$P(n \mu)$	$\mu_{\text{best}}$	$P(n \mu_{\text{best}})$	$R$	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		✓	

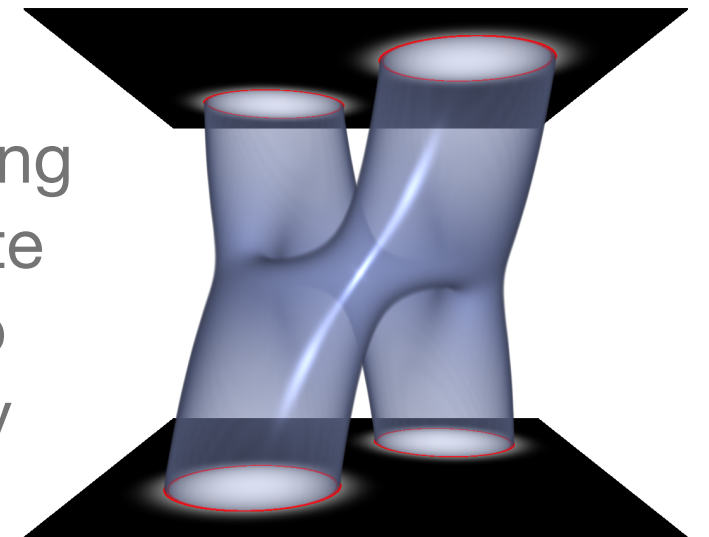
Observed count	$L(\mu = 0.5)$ of observed count	$\hat{\mu}$ that maximizes $L$ of observed count	$L(\hat{\mu})$ of observed count	Likelihood ratio $L(\mu = 0.5)/L(\hat{\mu})$ (ordering score)
-------------------	---	---	---	---

# Real life — high dimensions

---

In most real problems likelihoods are complicated multidimensional functions that cannot be analytically maximized. Two main issues:

Constructing confidence intervals is a significant computing burden: for each test value of the parameter  $m$ , (i) generate many samples of pseudodata, (ii) fit, and (iii) then move to another  $m$  value etc.. Diverges quickly with dimensionality



With highly-dimensional likelihoods the “projection” of the full-dimensional confidence band into the lower-dimensional subspace of interest leads to information loss: structure in the full dimensional space is lost when projected. The resulting confidence interval is bigger (less precise results).

# Real life — nuisance parameters

---


In most problems, systematic uncertainties complicate the interval determination  
Neither Neyman nor Feldman-Cousins have a prescription for that.

Parametrize the uncertainty in the shape of the model by **unknown nuisance parameters**. Not interesting for the measurement but do influence the result.

Assumed model

Reality

$$p(\vec{x}|\vec{m}) \Rightarrow p(\vec{x}|\vec{m}, \vec{s})$$

Nuisance parameters of unknown values

Typically cannot define a probability distribution for  $s$  (otherwise they'd be part of the model) but just a range. Goal: a procedure that **guarantees coverage** *whatever* is the value of the nuisance parameters within such ranges

Rigorous frequentist confidence intervals in the presence of nuisance parameters is a complicated problem for which no universal prescription yet exists.



# Profile-likelihood ratio ordering

---

A promising approach: profile-likelihood-ratio (PLR) ordering.

FC ratio-ordering applied **to likelihoods profiled (i.e., maximized) with respect to the uninteresting parameters**. The profile-likelihood *is not a likelihood*. It is a lower-dimensional derivation of it obtained by maximizing the likelihood wrt to the nuisance parameters. However, it preserves some of the nice features of the likelihood ratio: its asymptotic distribution is known and independent of  $m$

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

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 $L(x m = m_0, s)$

# In practice

---

Generate pseudodata that sample the full multidimensional space of the parameters. **fit** each sample **twice**, one with all parameters (physics and nuisance) floating, and another one with physics parameters fixed to their test value  $m_0$ .

1. Choose one value  $m_0$  for  $m$  and one value  $s_0$  for  $s$ , and generate pseudodata  $x$  accordingly
2. For each sample  $x$  (i) maximize  $p(x|m=m_0,s)=L(m=m_0,s)$  with respect to  $s$  to get  $L(m=m_0,\hat{s}^*)$  and (ii) maximize the likelihood  $L(m,s)$  over the space of  $m$  and  $s$  to obtain  $L(\hat{m},\hat{s})$
3. Rank all  $x$  in decreasing order of profile likelihood ratio  $PLR=L(m=m_0,\hat{s}^*)/L(\hat{m},\hat{s})$
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 the whole range]

Step 6 is essential to ensure the procedure has coverage for all values of the nuisance parameters.

Sometimes circumvented using the “*plugin method*”: only generate pseudodata at the  $s$  values estimated on data. Likely to spoil coverage

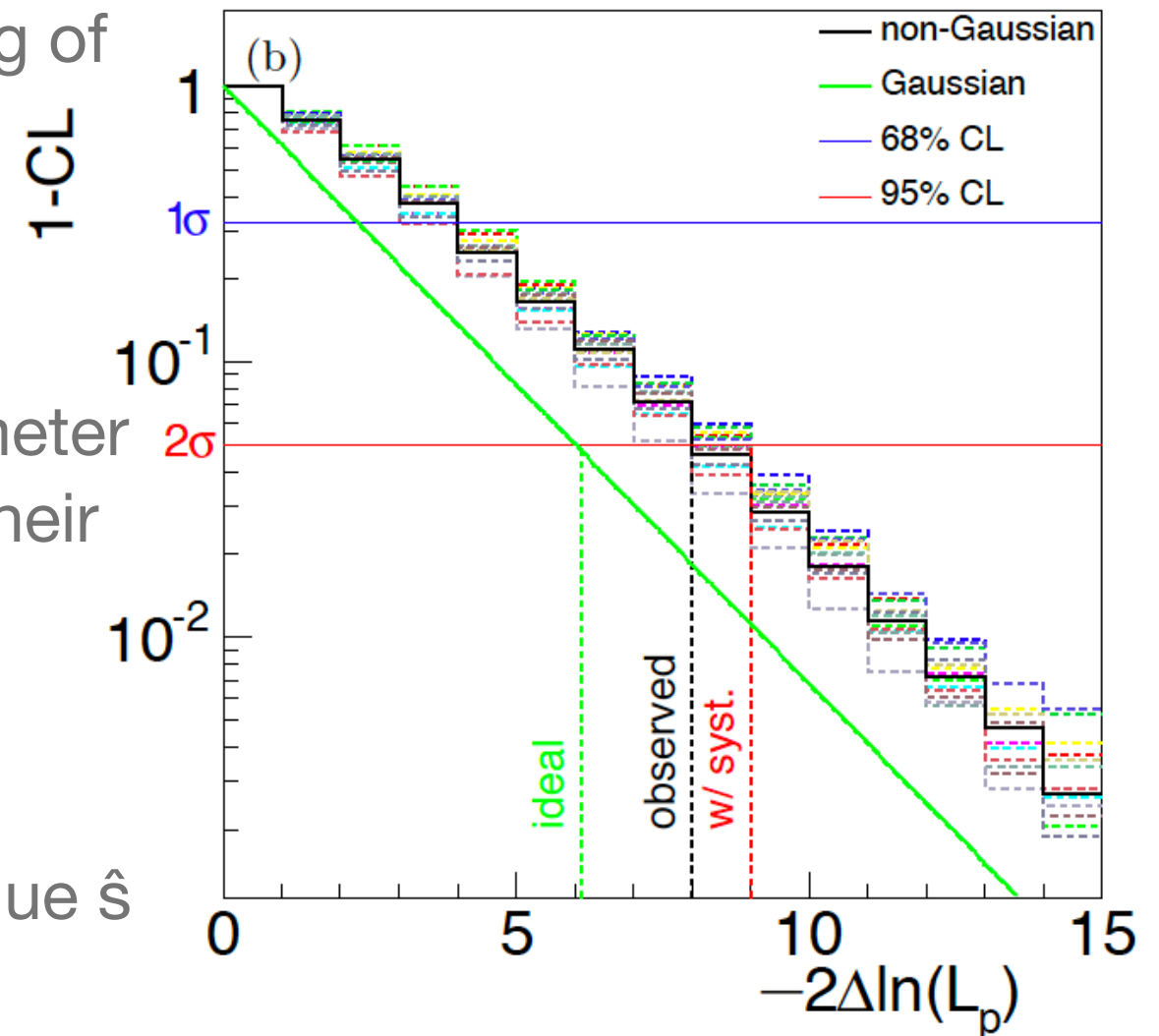
# Sampling the nuisance parameter space

Perform the procedure on a sufficient sampling of the full space of  $\vec{s}$  is expensive.

Midway between simplistic plugin and full treatment: restrict the sampling of the  $\vec{s}$  parameter space to a plausible subvolume centered on their estimates in data.

E.g., Berger and Boos: sample along each dimension  $s_i$  a range around the estimated value  $\hat{s}$  with CL much larger than the target CL of the profile-likelihood interval. (e.g, when constructing a 68% CL band in  $m$ , sample a 99.7% CL range in each dimension in  $s$  space)

JASA, 89, 427 (1994)



Application in a 27-dimensional case

<https://arxiv.org/pdf/0810.3229.pdf>

Phys. Rev. Lett 100 161802,

Phys Rev D 85, 072002

Phys. Rev. Lett. 109, 171802,

# The main burden

---



By this point you probably have realized that in a confidence interval construction, most of the time and effort is spent in generating a fitting simulated data sampled from  $p(x|m)$ . Effort and the time needed when likelihoods are highly multidimensional can be disconcerting.

Any way of avoiding this?

# Wilks' theorem

---

Asymptotically (large N), the distribution of the likelihood ratio

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

approaches a  $\chi^2$  distribution with # of degrees of freedom equal to # of additional free parameters in the denominator wrt the numerator



Samuel S. Wilks (1906-1964)

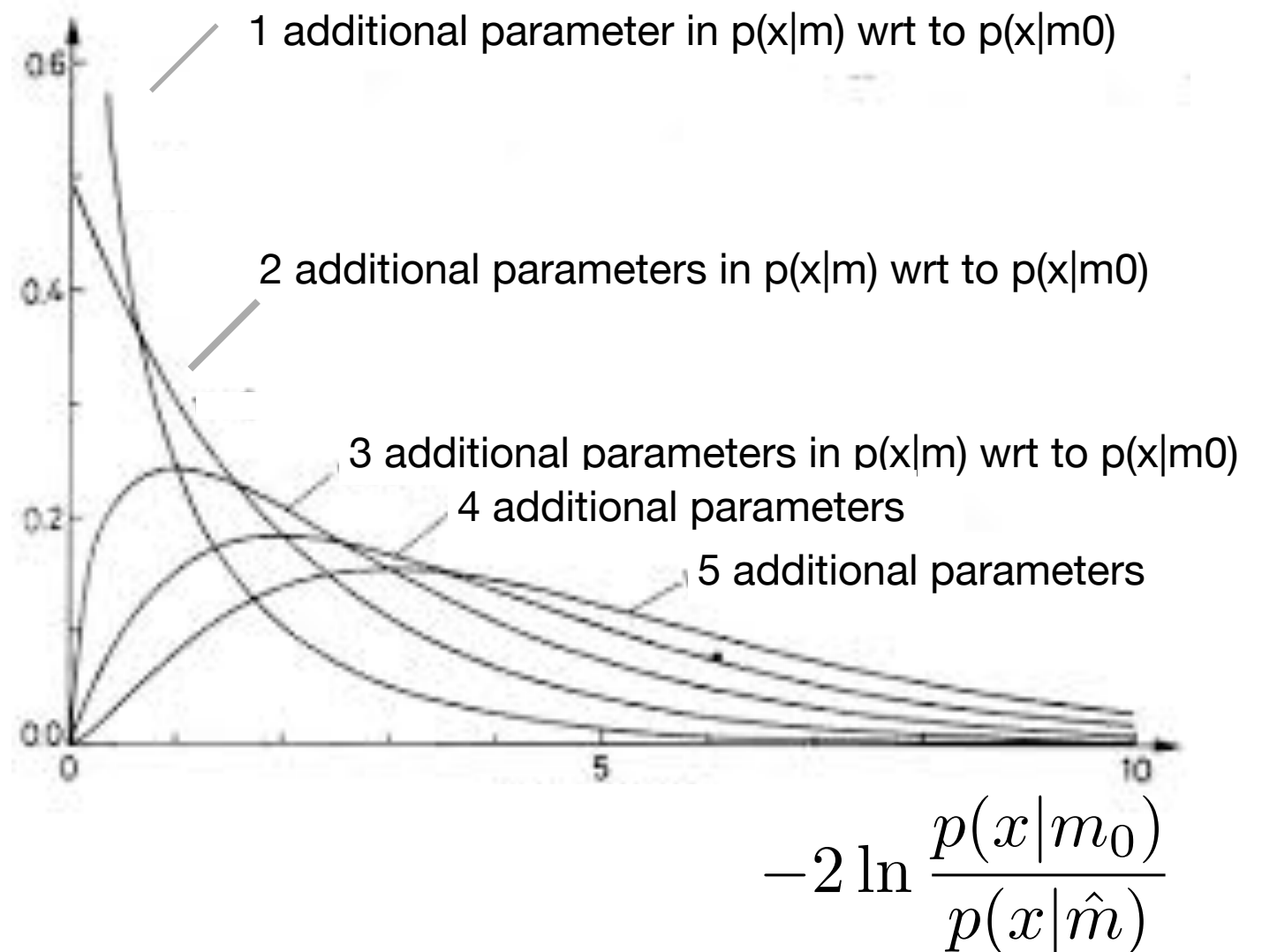
This holds independently of the shape of  $p(x|m)$  and on the value of  $m$ .

Great helps in 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.

# Wilks' theorem

No need to generate the sampling distributions of the ordering statistic.

Just look at where the (profile)-likelihood ratio observed in my data falls along the appropriate curve (determined by the number of degrees of freedom)



How do I know if L is asymptotic? Look at a few samples of pseudodata, and compare with above.



# Wilks' theorem at work — MINOS

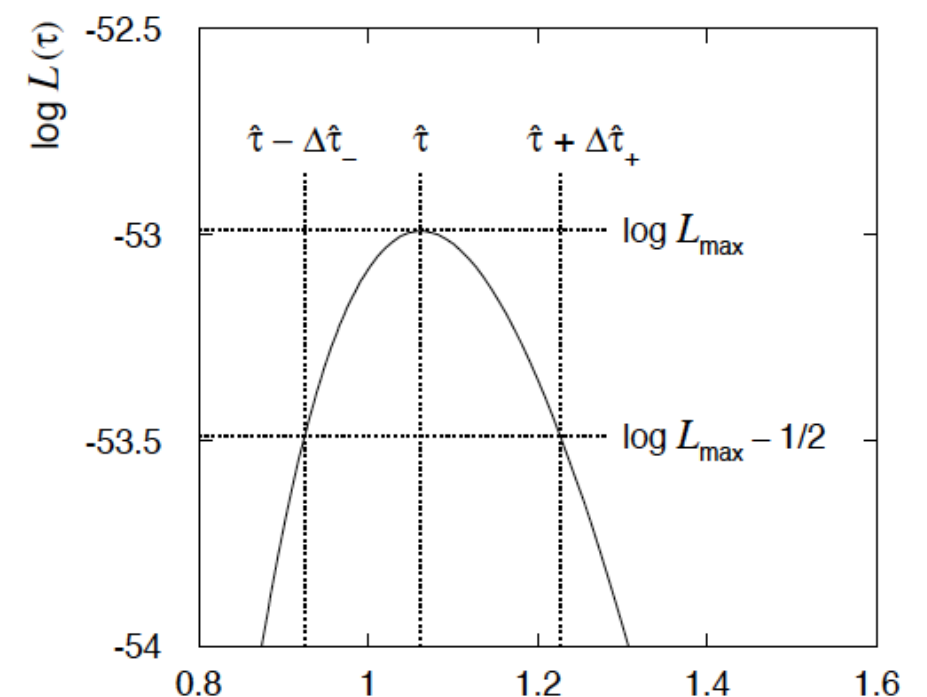
Moves down from the maximum  $L(\hat{m}, \hat{s})$  evaluating  $L(m_0, \hat{s}_m)$  at each point  $m_0$  by maximizing wrt parameters  $\vec{s}$  (i.e., likelihood of  $m$  profiled wrt  $\vec{s}$ ).

When  $L(m_0, \hat{s}_m)/L(\hat{m}, \hat{s})$  equals the threshold values tabulated from the  $\chi^2$  distribution the corresponding projection of the profile-likelihood onto the  $m$  space **approximates (large N) a Feldman-Cousins central confidence interval**

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

$\Delta$	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	$\Delta$				
	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
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

# The Asimov asymptotic formulas

---

Significant recent breakthrough allows generalizing the Wilks theorem and provides key asymptotic formulas for the distributions of profile likelihood ratios used in confidence intervals and hypothesis tests.

Eur. Phys. J. C (2011) 71: 1554  
DOI 10.1140/epjc/s10052-011-1554-0

---

**THE EUROPEAN  
PHYSICAL JOURNAL C**

---

Special Article - Tools for Experiment and Theory

## **Asymptotic formulae for likelihood-based tests of new physics**

**Glen Cowan<sup>1</sup>, Kyle Cranmer<sup>2</sup>, Eilam Gross<sup>3</sup>, Ofer Vitells<sup>3,a</sup>**

<sup>1</sup>Physics Department, Royal Holloway, University of London, Egham TW20 0EX, UK

<sup>2</sup>Physics Department, New York University, New York, NY 10003, USA

<sup>3</sup>Weizmann Institute of Science, Rehovot 76100, Israel

Received: 15 October 2010 / Revised: 6 January 2011 / Published online: 9 February 2011

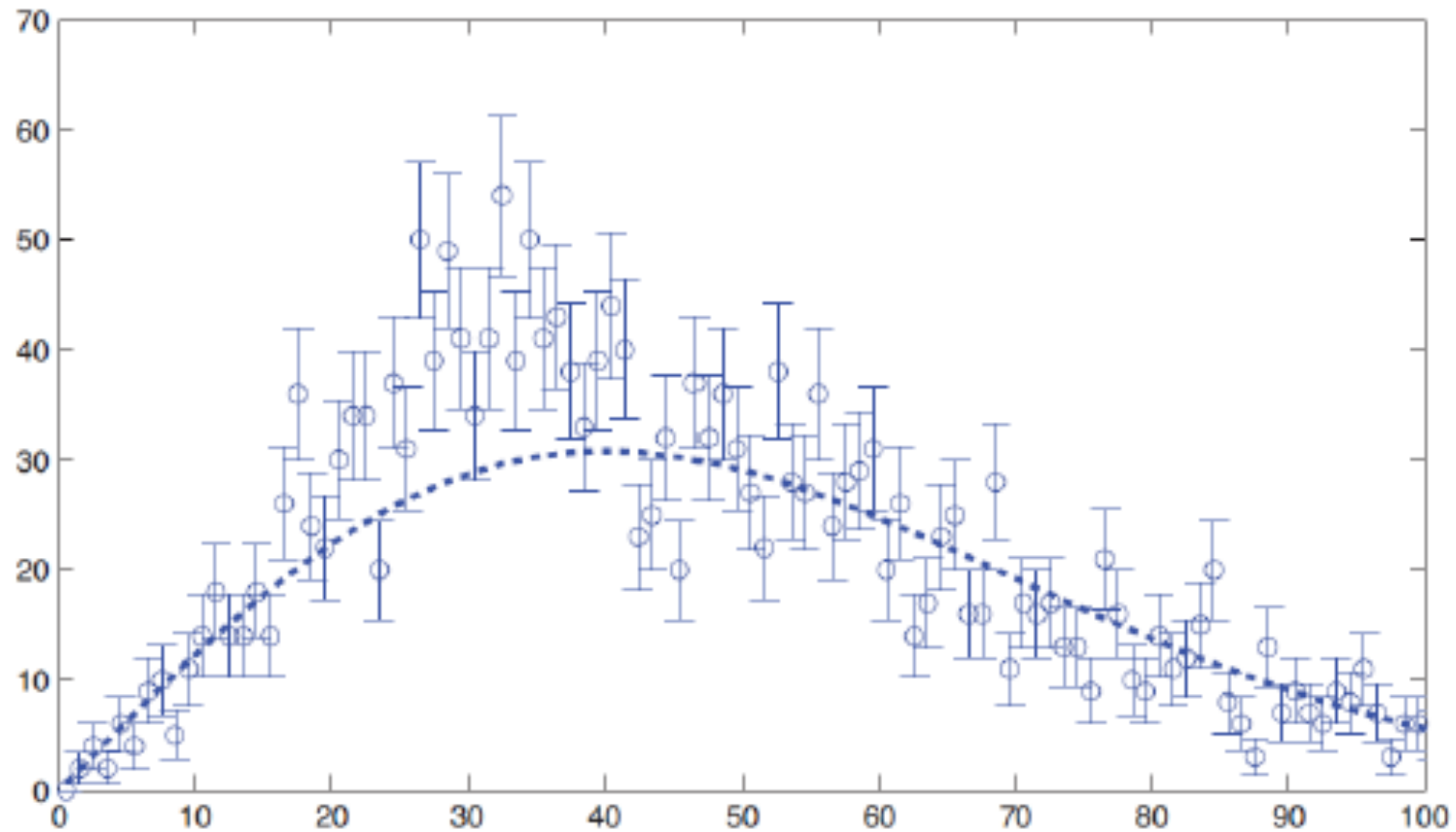
© The Author(s) 2011. This article is published with open access at Springerlink.com



# Hypothesis testing

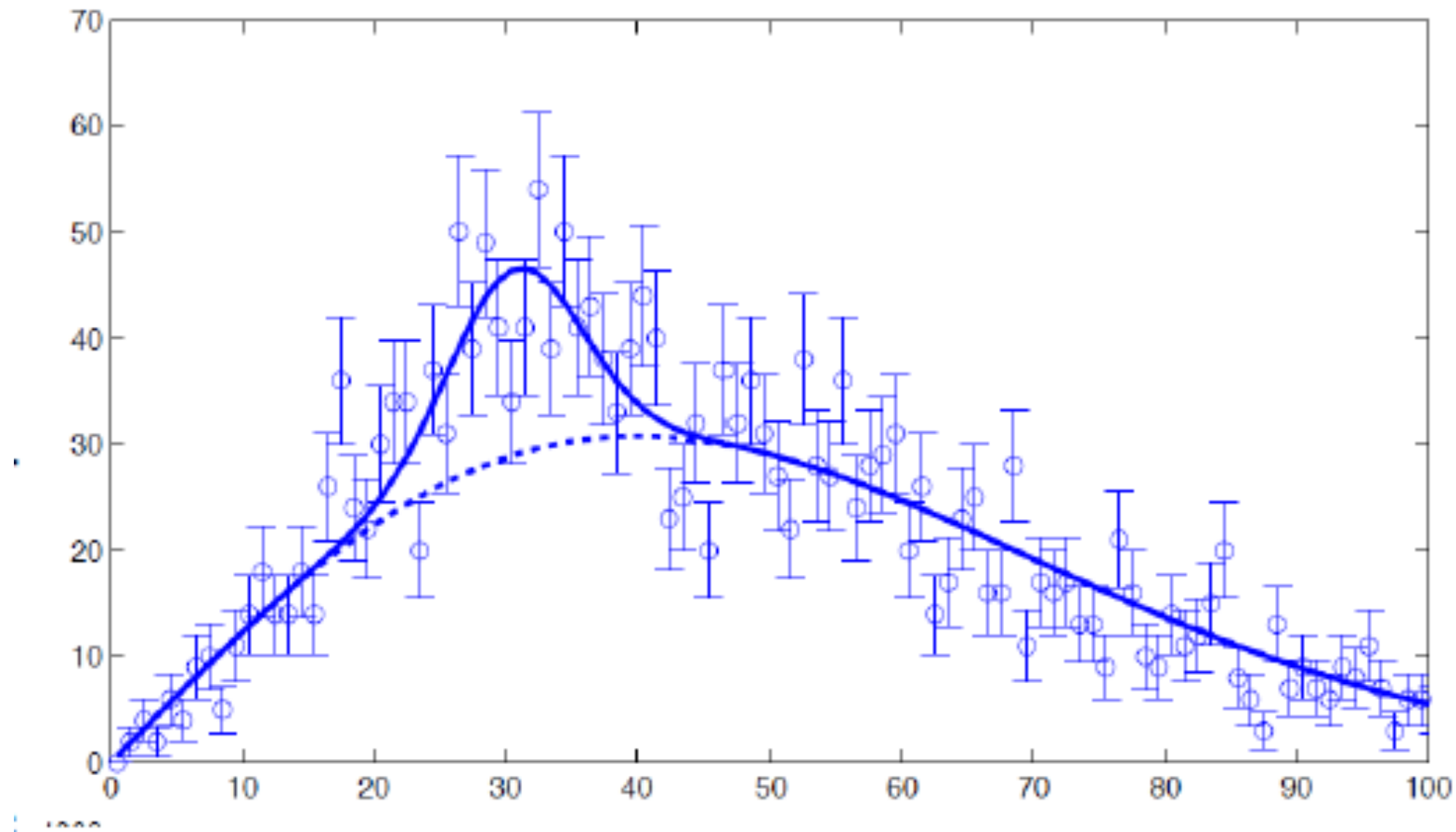
# Are my data compatible with background?

---



# Or they suggest the presence of a signal?

---

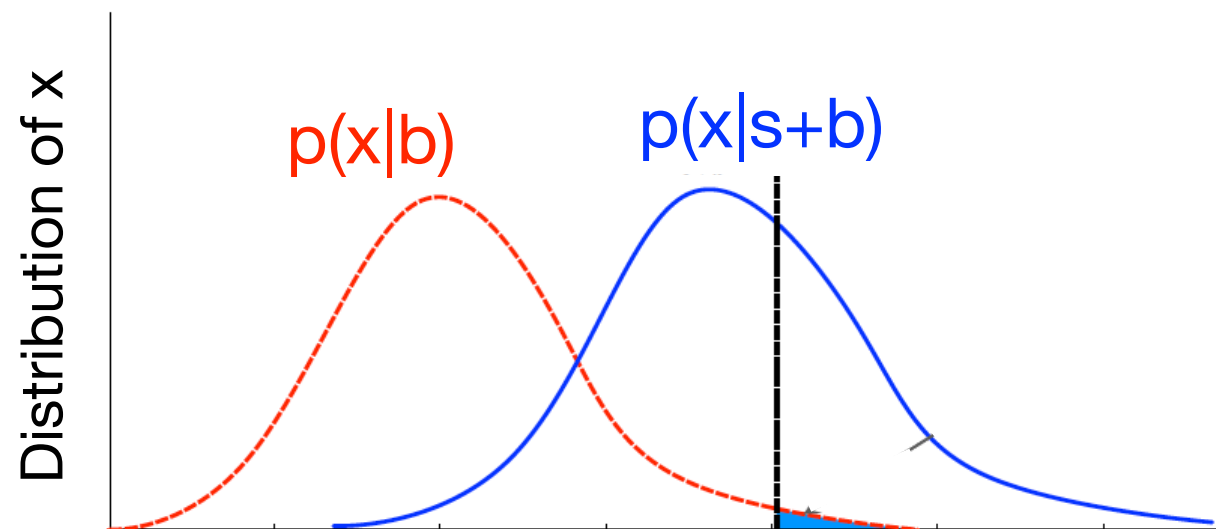


The p-value is a random variable that helps answering this question  
<http://priceconomics.com/the-guinness-brewer-who-revolutionized-statistics/>

# Ingredients

---

Need two hypotheses. **only known phenomena contribute** “null” or “background” )  
**new phenomena contribute too** (“alternate” or “signal”)



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

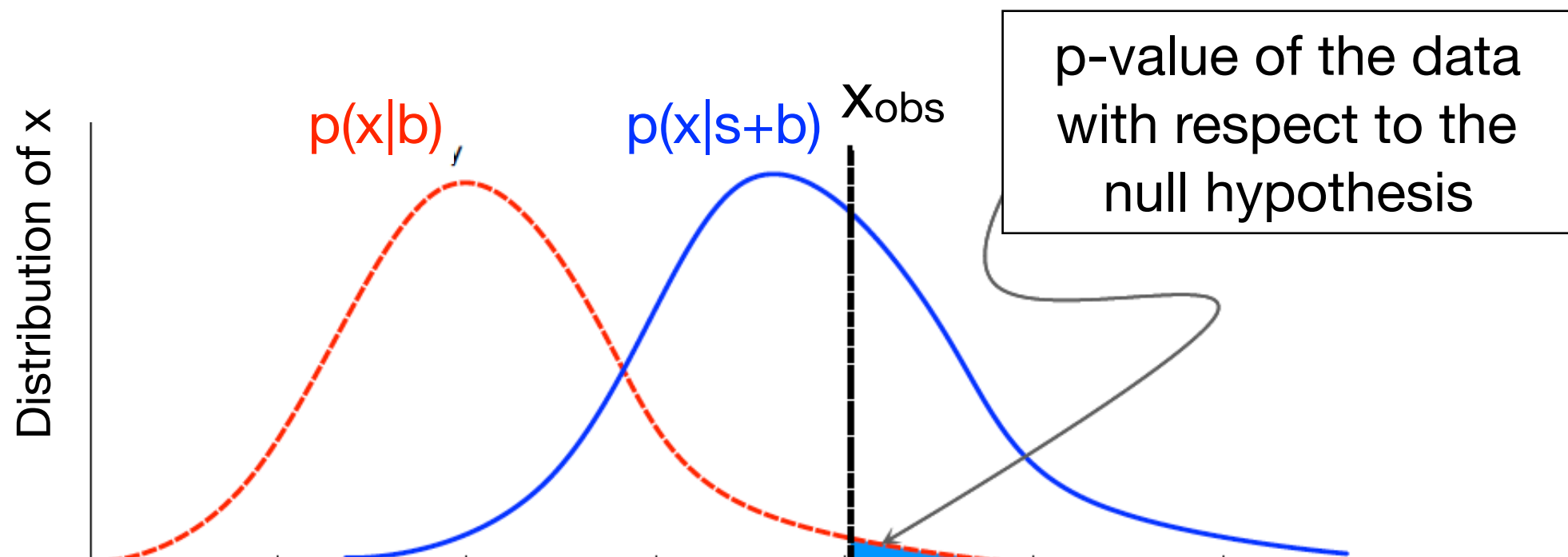
Devise a function  $x$  of the data (e.g., signal-event count), whose distribution under the null  $p(x|b)$  “differs” from that under the signal hypothesis  $p(x|s+b)$ . Generate these two distributions (labor intensive — typically done using simulation)

Set, prior to the observation, the false-positive rate: how much “signal-like” the observed value of  $x$  should be to exclude the background only hypothesis.

# p-values for discovering a new effect

Observe  $x_{\text{obs}}$ . The location of  $x_{\text{obs}}$  relative to the two pdf offers a quantitative measure of data compatibility with either hypotheses.

p-value: relative fraction of the integral of the null model over values of  $x$  as signal-like as those observed and more. The smaller the p-value, the stronger the evidence against the null hypothesis. If  $\text{p-value} < \text{false-positive rate}$ , exclude the background-only hypothesis at  $\text{CL} = 1 - (\text{p-value})$ .



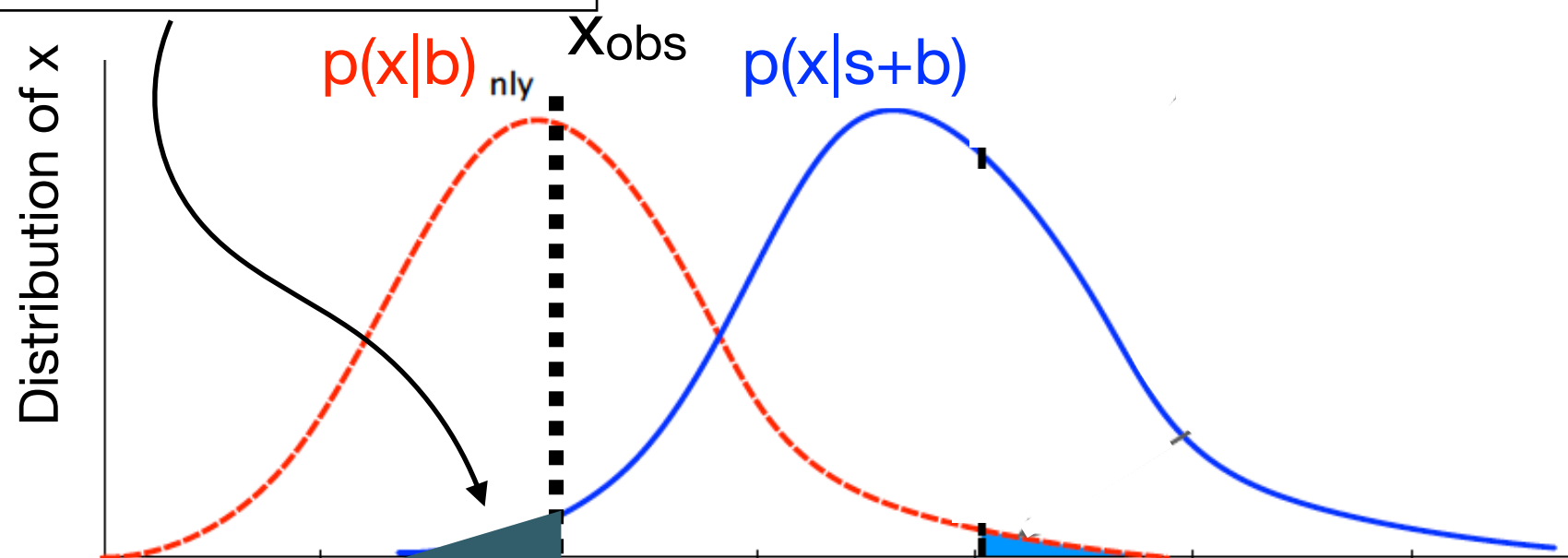
Arbitrary function  $x$  of the data that allows for separation between the two hypotheses

# p-values for excluding a new effect

If the purpose is to exclude a new effect, then 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



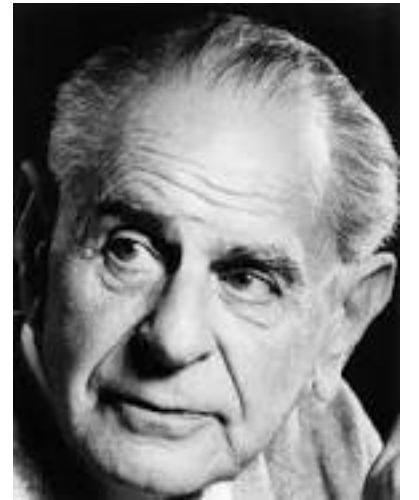
Arbitrary function  $x$  of the data that allows for separation between the two hypotheses

# Testing the Popperian way

---

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

“Discover” a signal by excluding its absence (that is, by excluding that only background contributes). Limit to the existence of a signal by excluding its presence.



Karl Popper (1902-1994)

A **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!**

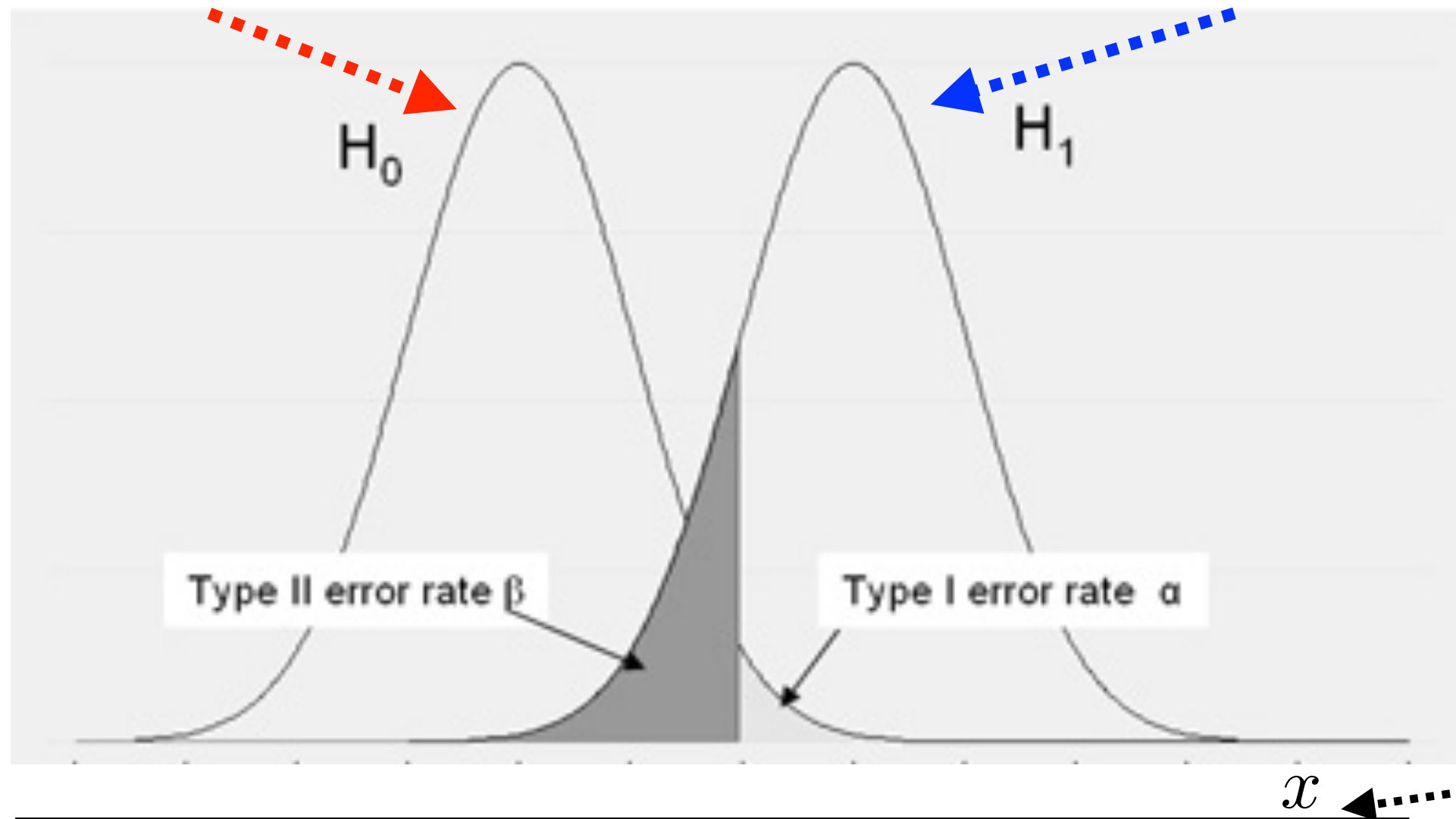
Wrong claim “The measurement shows that the probability for hypothesis blah is ..”

P-values connect to the probability to observe  $x_{obs}$  or a more extreme value *if a specific 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...”

# Nomenclature

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

This is  $p(x|s+b)$ , the distribution of  $x$  under the signal hypothesis



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

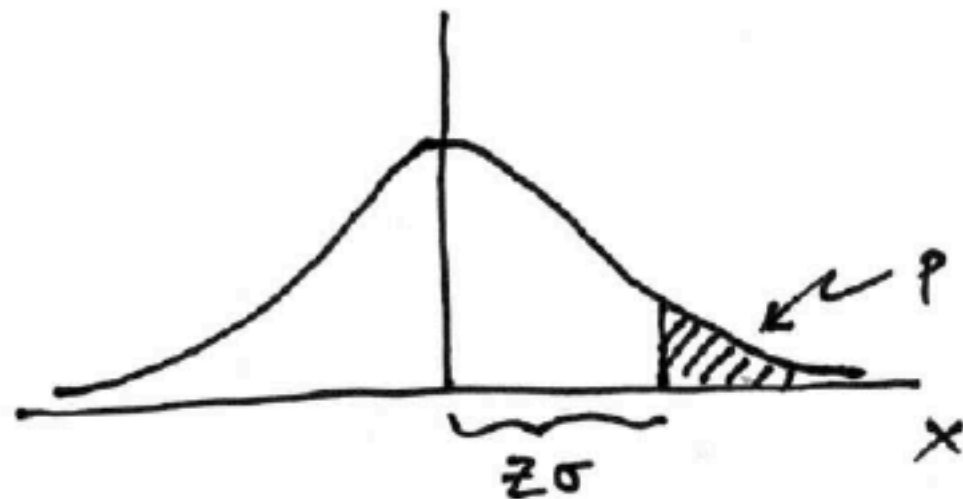


# “Significance”

---

“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



$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z) \quad \text{1 - TMath::Freq}$$

$$Z = \Phi^{-1}(1 - p)$$

`TMath::NormQuantile`

# 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

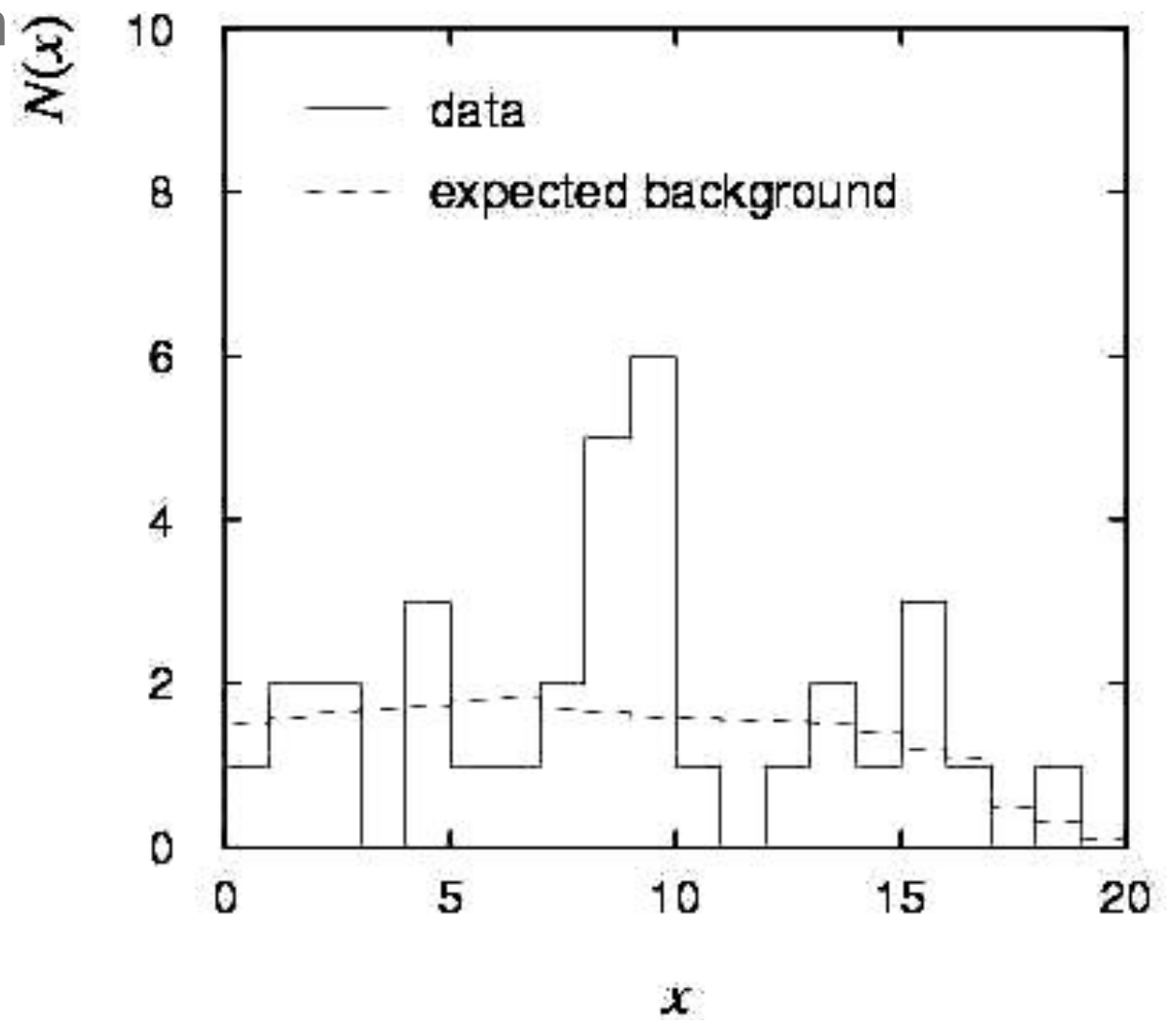
$$P(n; s, b) = \frac{(s + b)^n}{n!} e^{-(s+b)}$$

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?

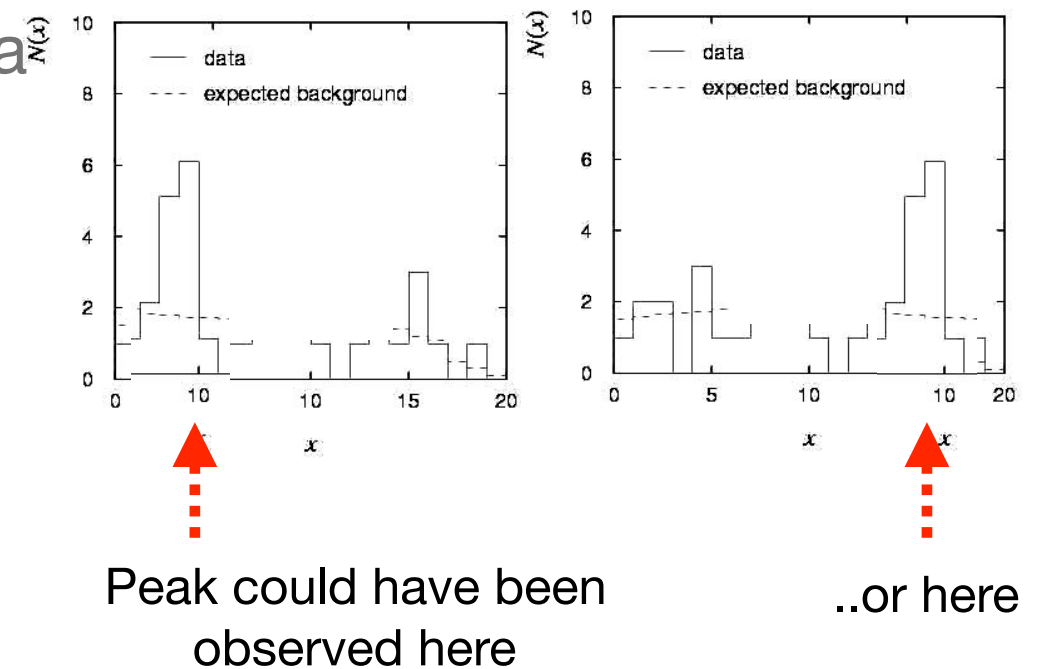
[Cowan]



# “Local” p-value and “look-elsewhere effect”

That evaluation only accounts for the chances of an upward fluctuation in that very position at  $x \sim 9$ . That's the “local p-value”.

“global p-value” need to 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 local p-value gets multiplied by  $\approx 10$ .



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

When quoting p-values, need to correct for the **effect of multiple testing** (i.e., account that we have also been “looking elsewhere” from where the anomaly is).

Use simulation, or approximate correction factors, e.g., in EPJ C70, 525 (2010)

# The conventional “ $5\sigma$ rationale”

---

HEP experimenters conventionally agree to deal with the LEE by setting a rather extreme standard for p-values to justify claims of new effects. (Originated by a survey of experimental results on “far-out hadrons” in 1968 — see backup)

One requires the null to be rejected with significance of  $3.5\sigma$  (for “evidence”) and  $5\sigma$  (“observation”), corresponding to very small p-values (fluctuations that occur 3 times every 10 million trials).

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

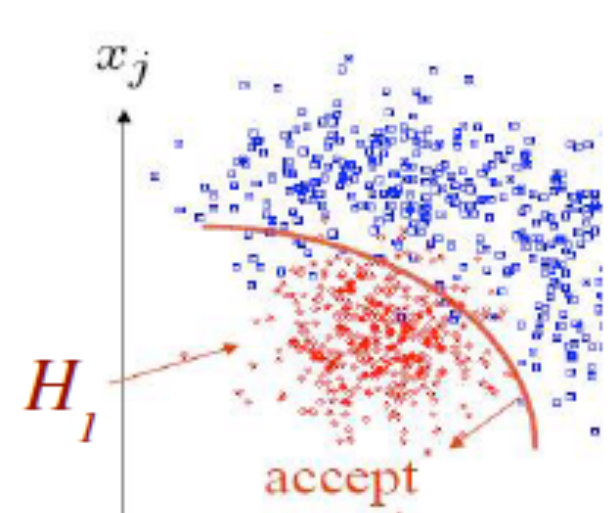
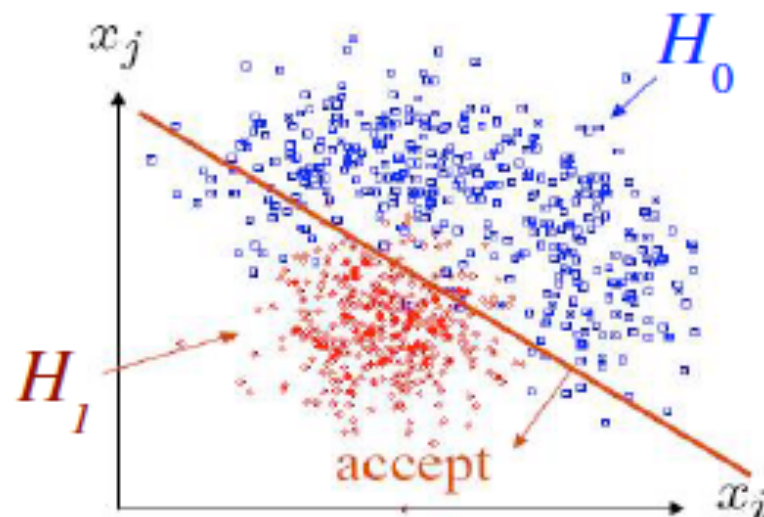
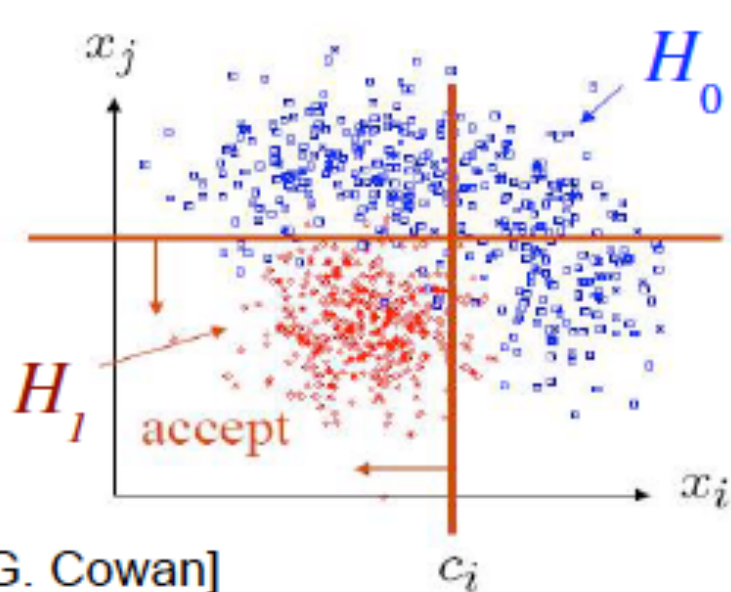
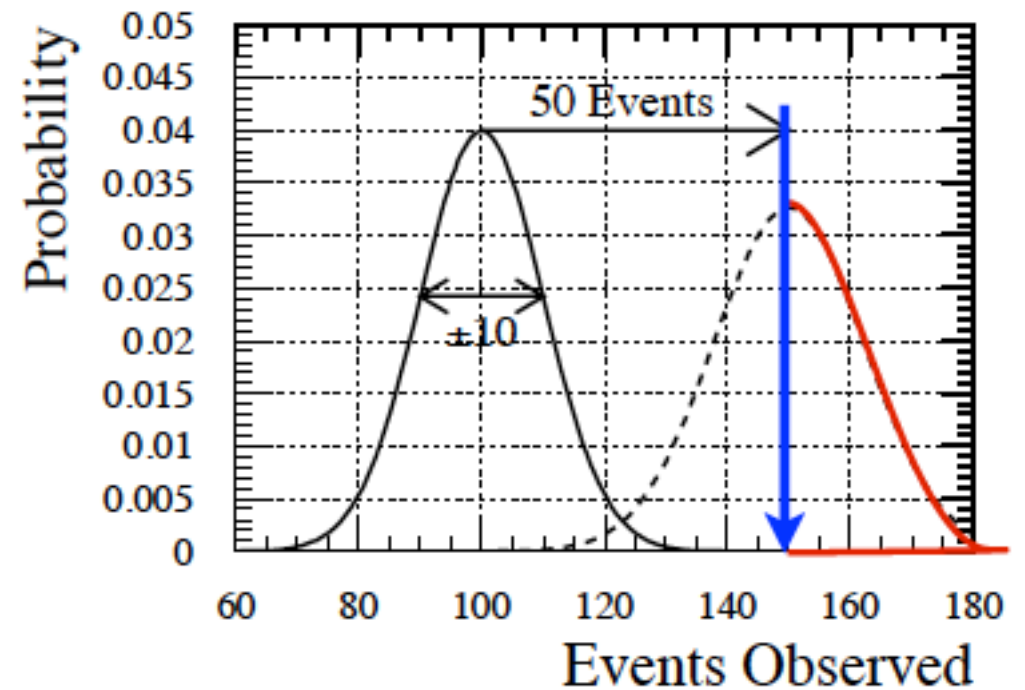
However, one-size-fits-all does not seem appropriate here.

# 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



# Neyman-Pearson lemma

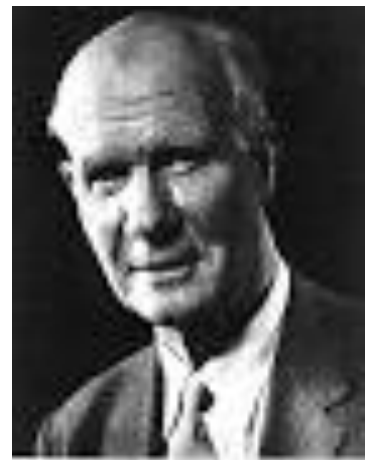
---

It does exist an universal statistic for optimal separation between the two hypotheses (for simple completely specified 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

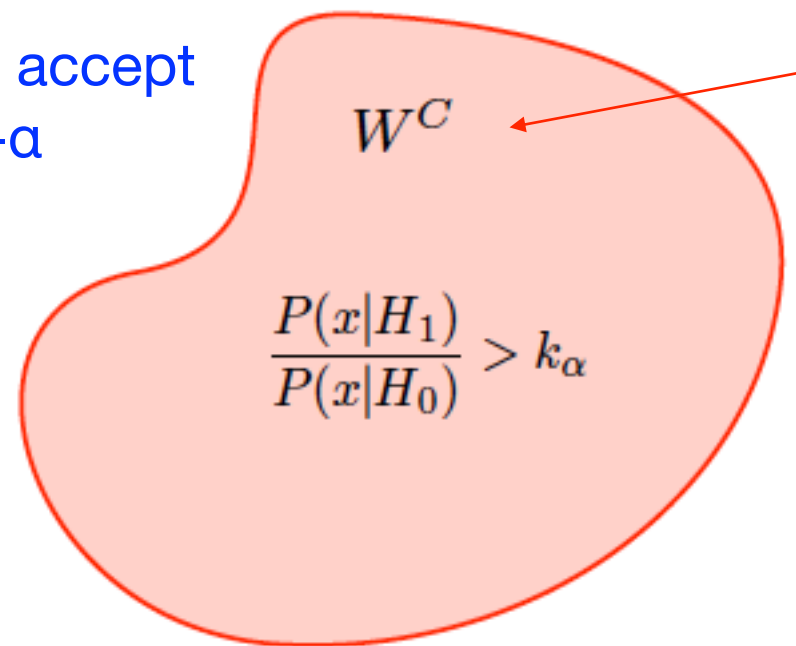
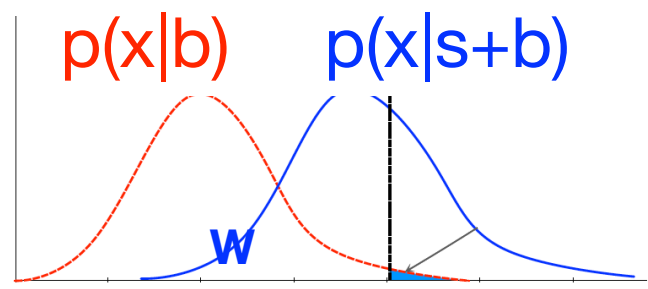
$$\frac{p(x|H_1)}{p(x|H_0)} > k_\alpha$$

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

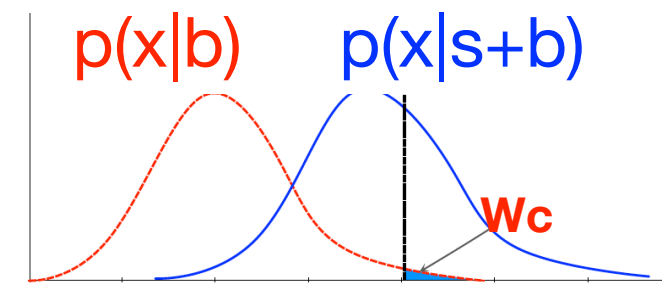
# NP-lemma illustrated proof

Take a contour of the likelihood ratio that has a given rate  $\alpha$  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  $1-\alpha$



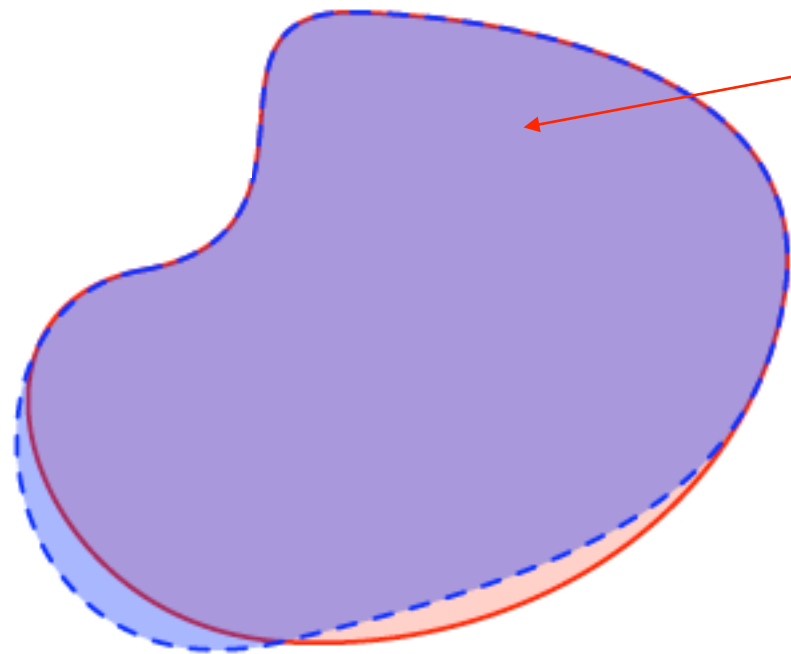
Region  $W^c$ : if data fall there we reject  $H_0$ ; probability under  $H_0$  is  $\alpha$



# NP-lemma illustration

---

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

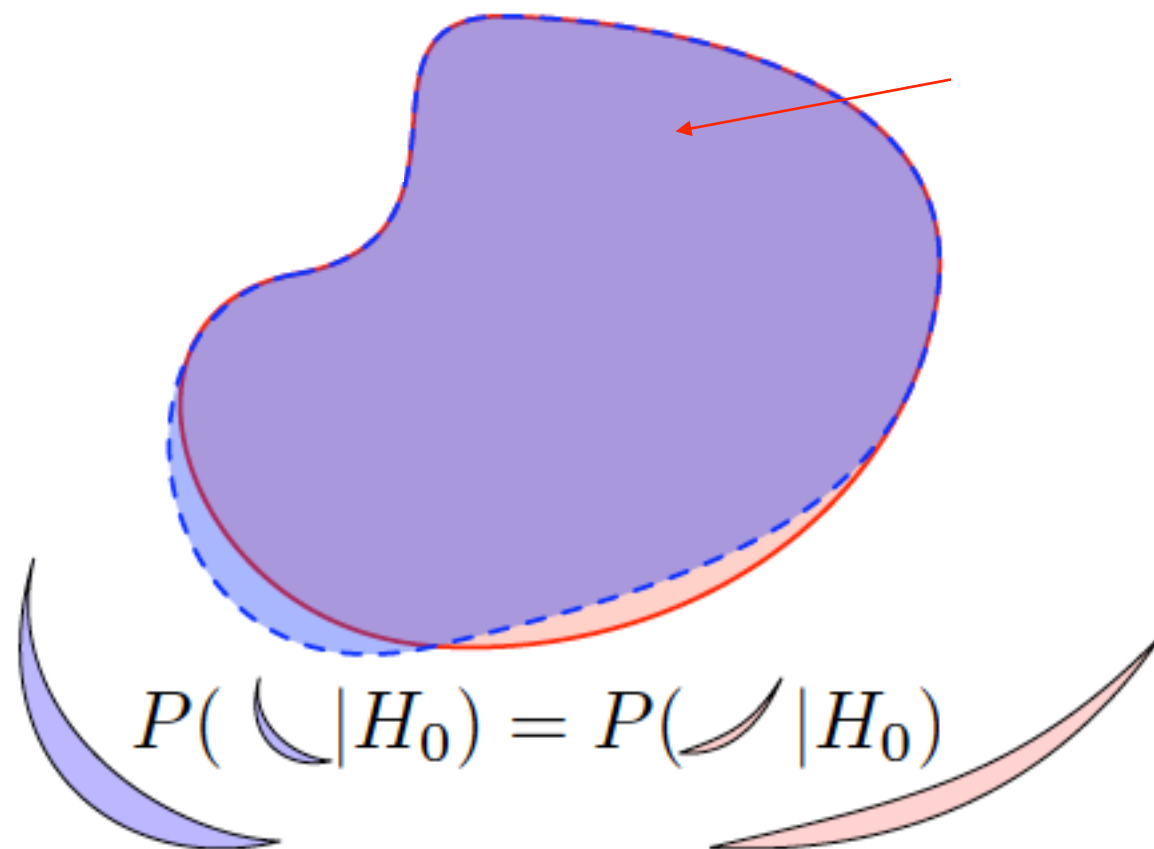




# NP-lemma illustration

---

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

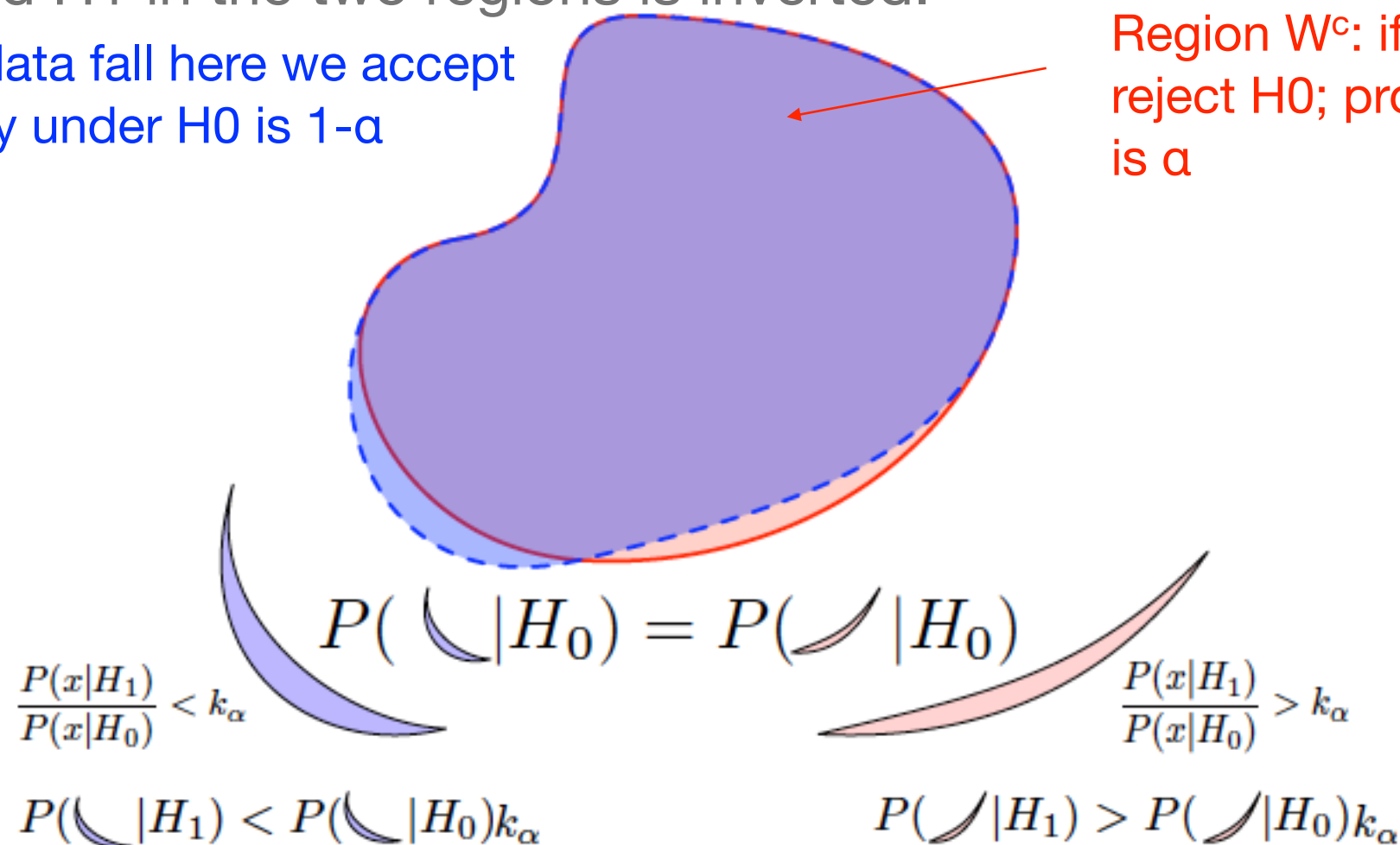


# NP-lemma illustration

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  $\alpha$



$$P(\text{blue curve} | H_1) < P(\text{red curve} | H_1)$$

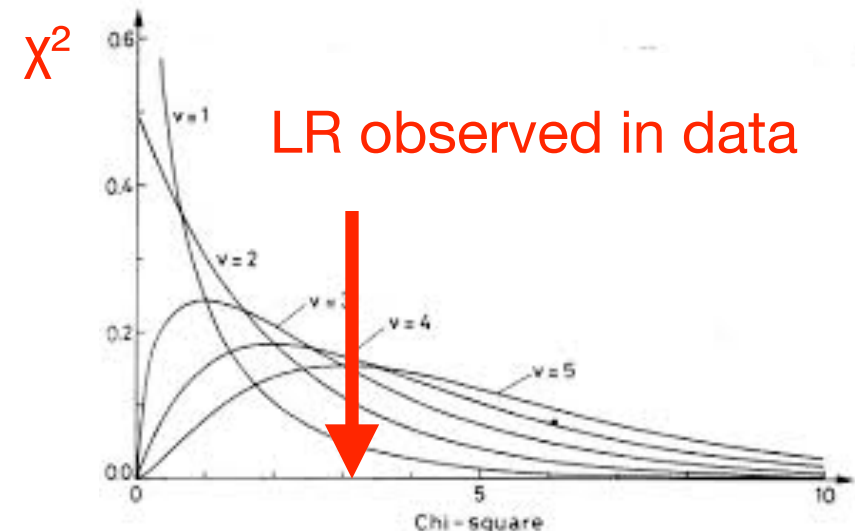
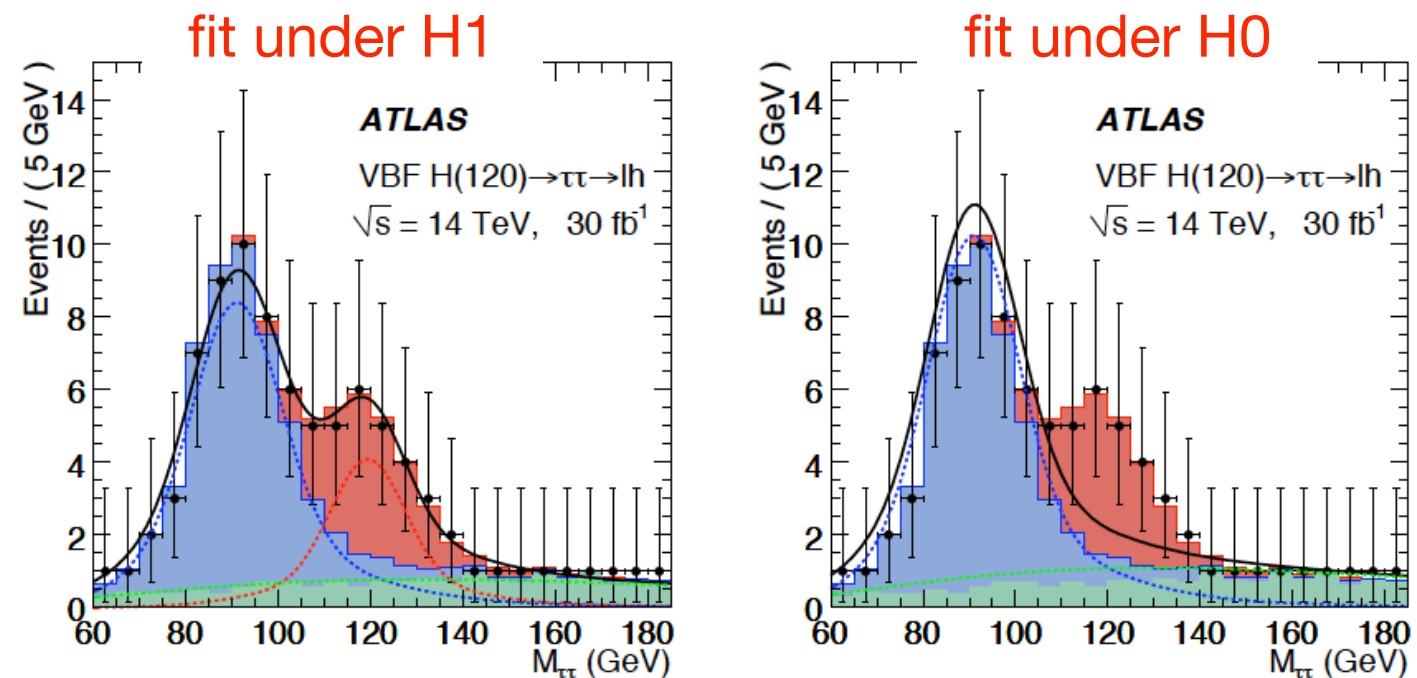
The new region region has less power.

# (profile) likelihood-ratio as a test statistic

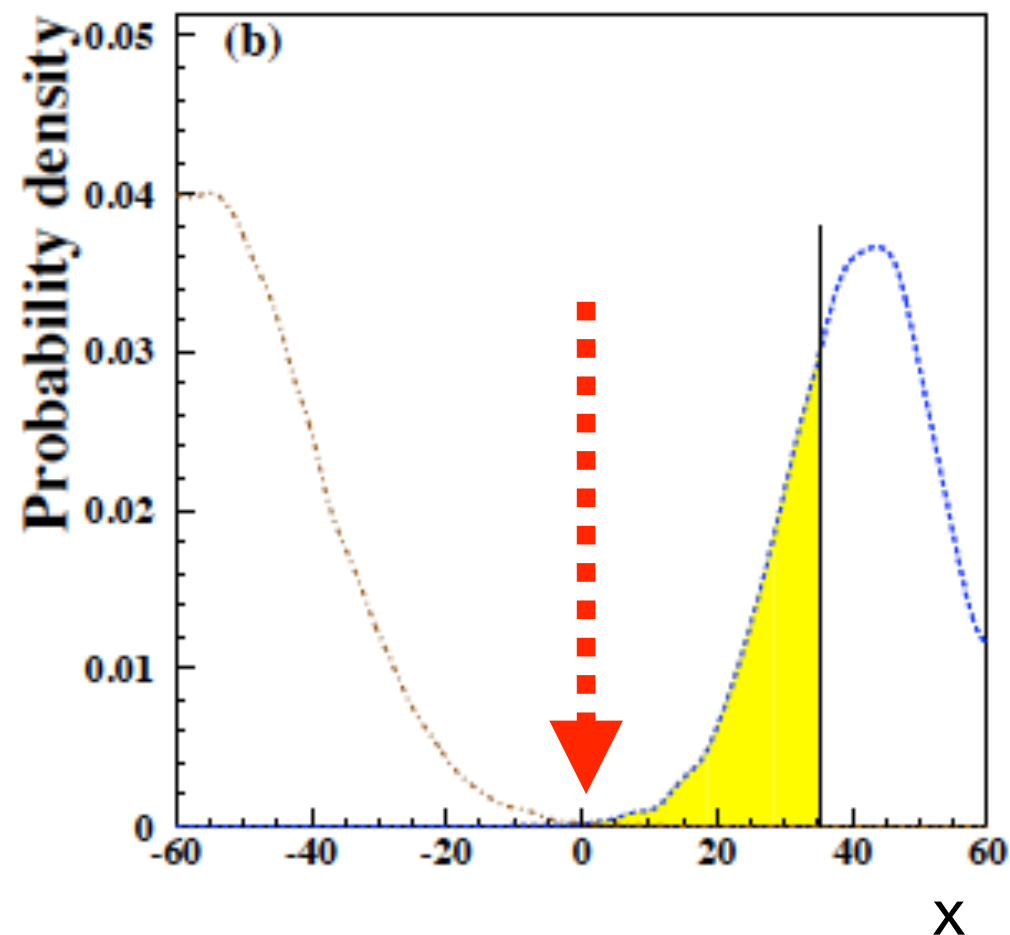
Convenient because (1) has optimal performance and (2) allows for testing with no need to laboriously construct distributions by generating and fitting pseudodata since its large-sample distribution is known ( $\chi^2$ )

[Cranmer]

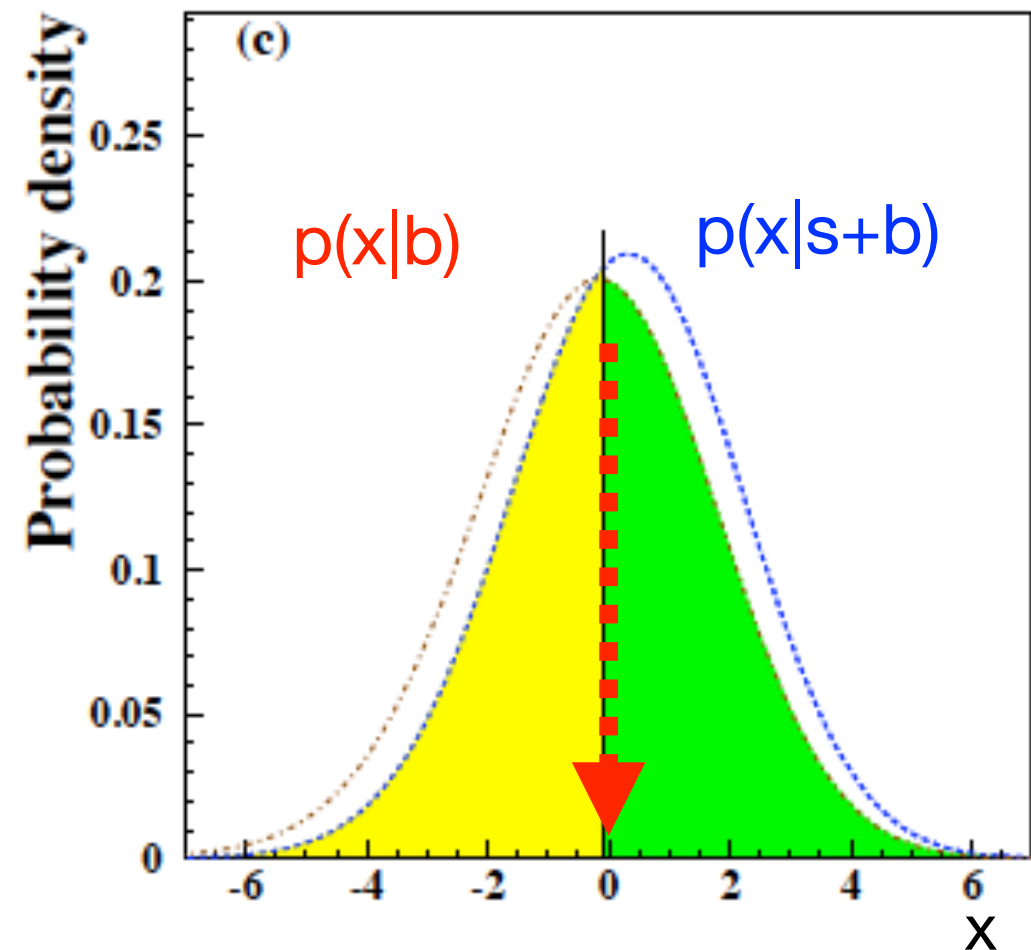
1. Fit data under  $H_0$ : i.e. with a likelihood that only has “background” parameters.
2. Fit data under  $H_1$ : i.e. with a likelihood that includes  $n$  additional “signal” free parameters
3. The ratio between the resulting values of the likelihood functions at their maxima is distributed as a  $\chi^2$  with  $n$  degrees of freedom.
4. Comparison of the ratio obtained in data with the relevant  $\chi^2$  distribution allows for testing  $H_1$  vs  $H_0$ .



# Issues with p-values



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



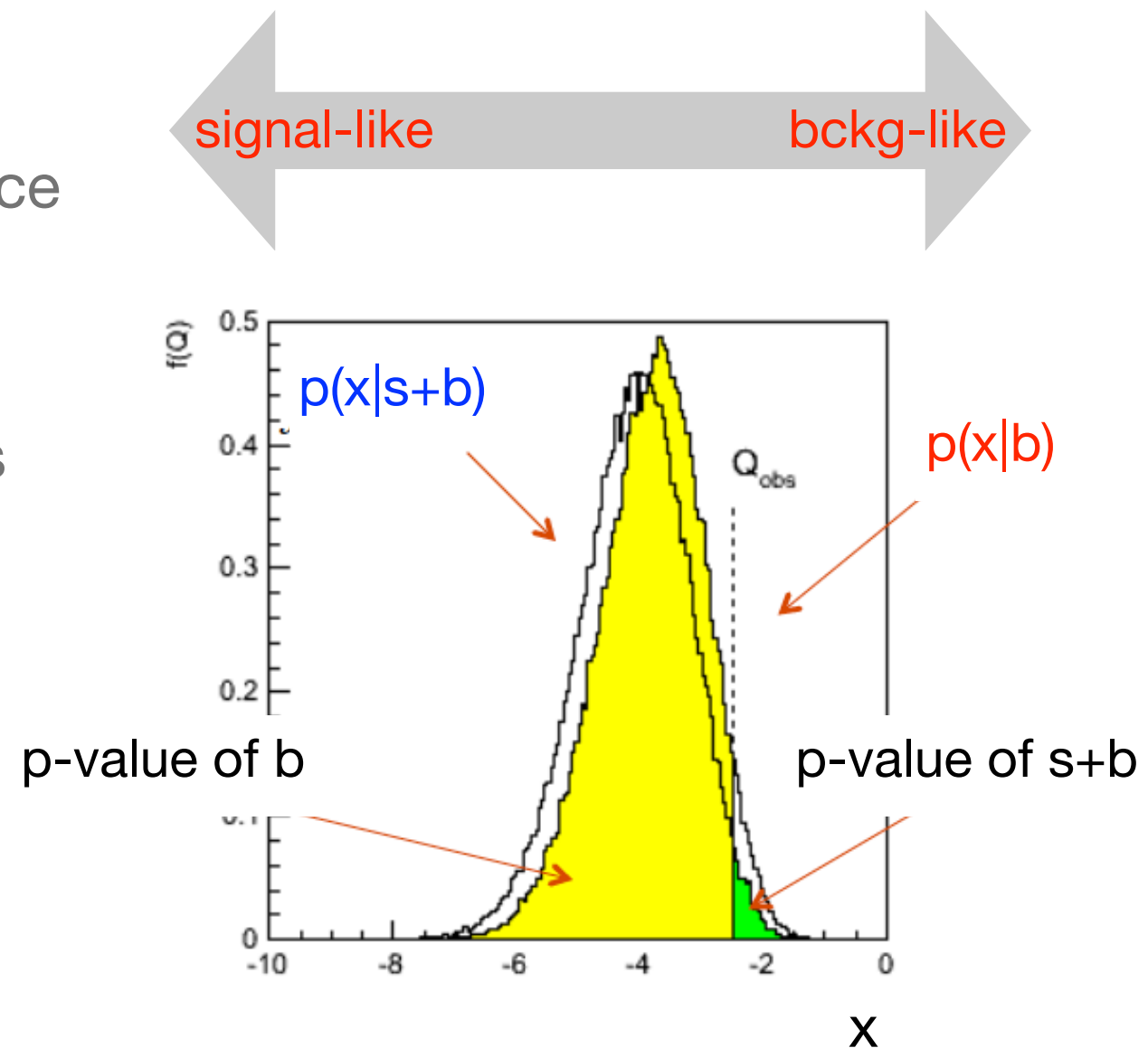
For small signals with poor S-vs-B separation, sensitivity is low, which means that distributions of test statistics are nearly equal. Can make no statement about the signal, regardless of the outcome

# Spurious exclusion

Use the likelihood ratio  $x$  to test the presence of a signal  $p(x|s+b)$ .

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  $s+b$ .



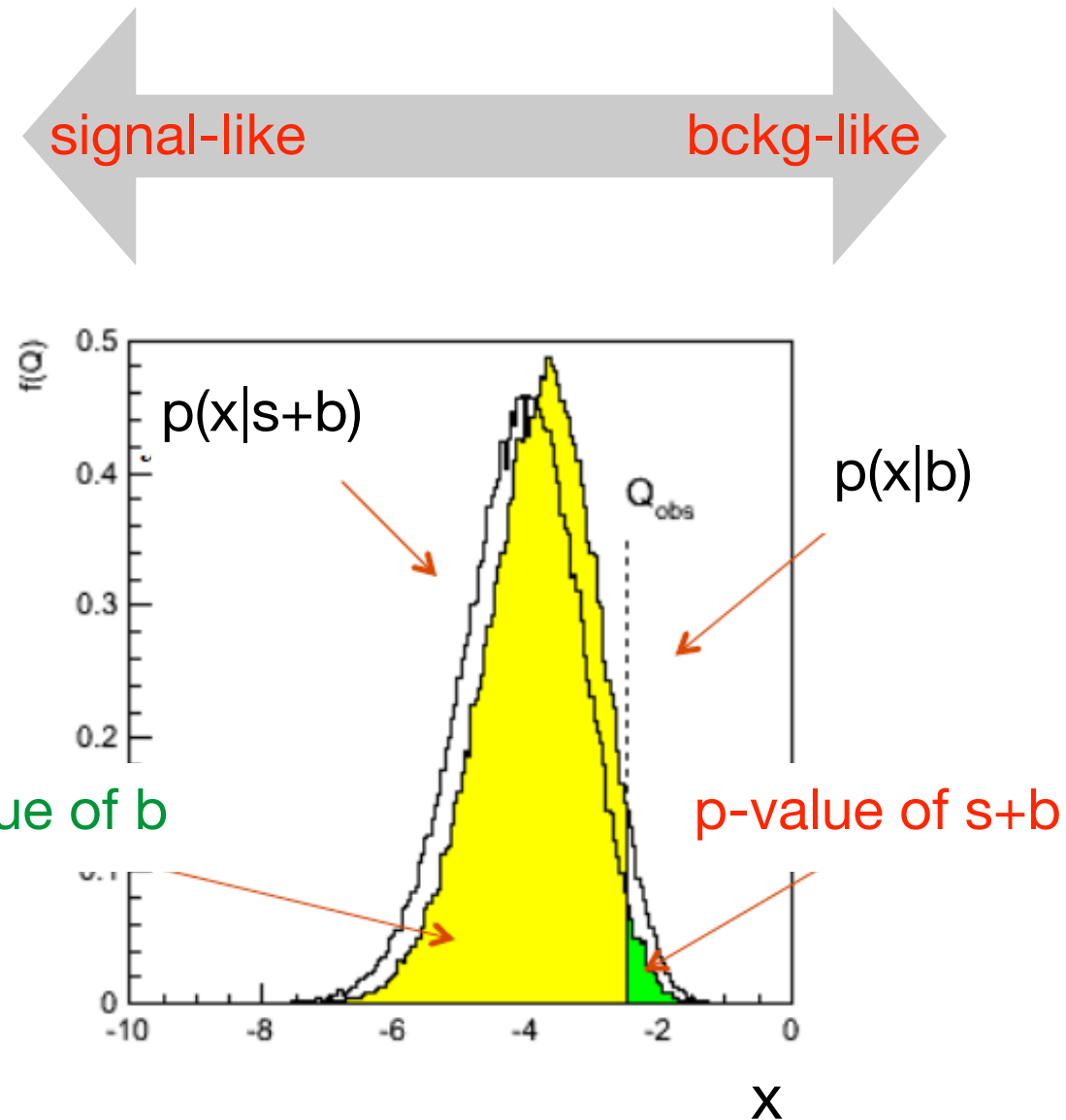
# The CLs method

Scaling the p-value 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

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

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



When quoting limits, it's good practice to assess the **analysis sensitivity** in terms of median expected limits based on ensembles of simulated experiments or asymptotic formulae if applicable to your case

# Duality

---

Given an ordering, there is a one-to-one correspondence between hypothesis testing and construction of confidence intervals.

It is the same problem.

Testing if parameter  $m$  equals  $m_0$  or rather any other value, with a chosen false-positive rate = pvalue, corresponds to checking if  $m_0$  is included in the confidence interval for  $m$  with  $CL=1-(pvalue)$

Subtends why the likelihood-ratio based ordering of Feldman and Cousins is a generalized and powerful criterion for construction of confidence intervals, thanks to the NP-lemma.

---

Additional material



# Confidence intervals

**Gaussian pdf  $p(x|\mu,\sigma)$  with  $\sigma$  a function of  $\mu$ :  $\sigma = 0.2 \mu$**

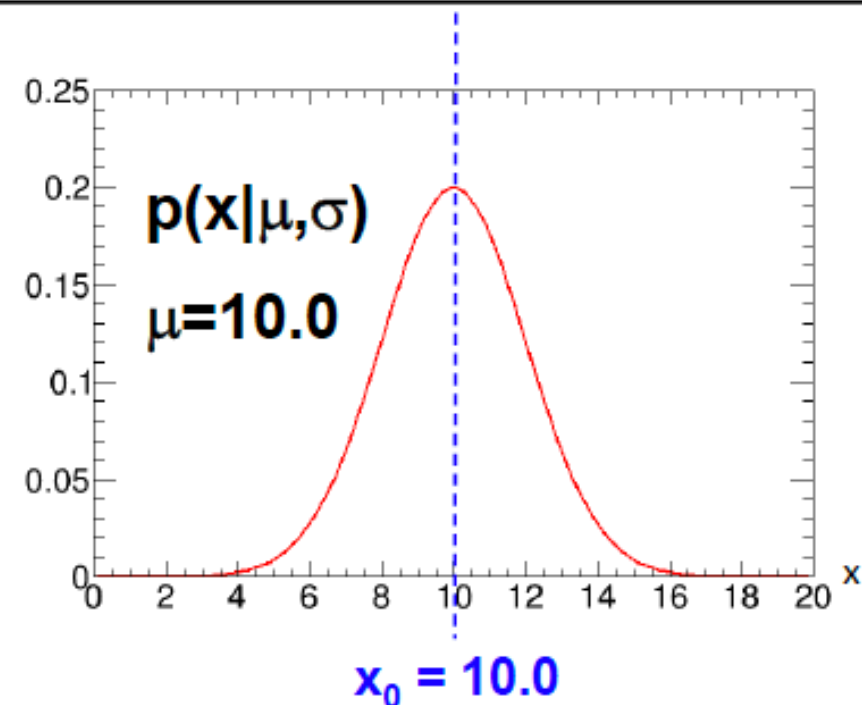
$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\sigma(\mu) = (0.2) \mu$$

**$p(x|\mu,\sigma)$  with  $\mu=10.0$ ,  $\sigma = 0.2$  :**

**Suppose  $x_0 = 10.0$  is observed.**

**What can one say about  $\mu$  ?**



Minimum  $\chi^2$  for a single observation of 10, yields  $\hat{\mu} = 10$ . Then estimate  $\hat{\sigma} = 0.2 \times \hat{\mu} = 0.2 \times 10 = 2.0$

Therefore  $\hat{\mu} \pm \hat{\sigma} = [8.0, 12.0]$

# Confidence intervals

**Gaussian pdf  $p(x|\mu,\sigma)$  with  $\sigma$  a function of  $\mu$ :  $\sigma = 0.2 \mu$**

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\sigma(\mu) = (0.2) \mu$$

$p(x|\mu,\sigma)$  with  $\mu=10.0$ ,  $\sigma = 0.2$  :

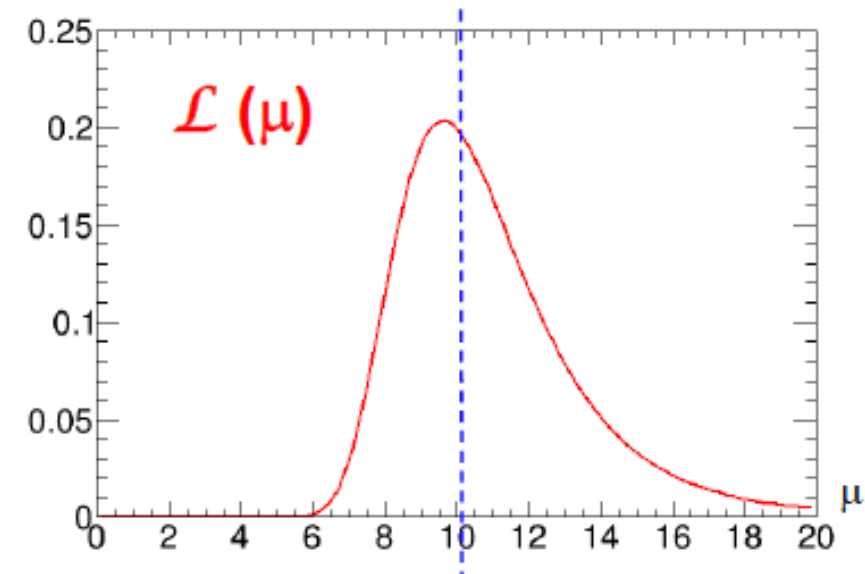
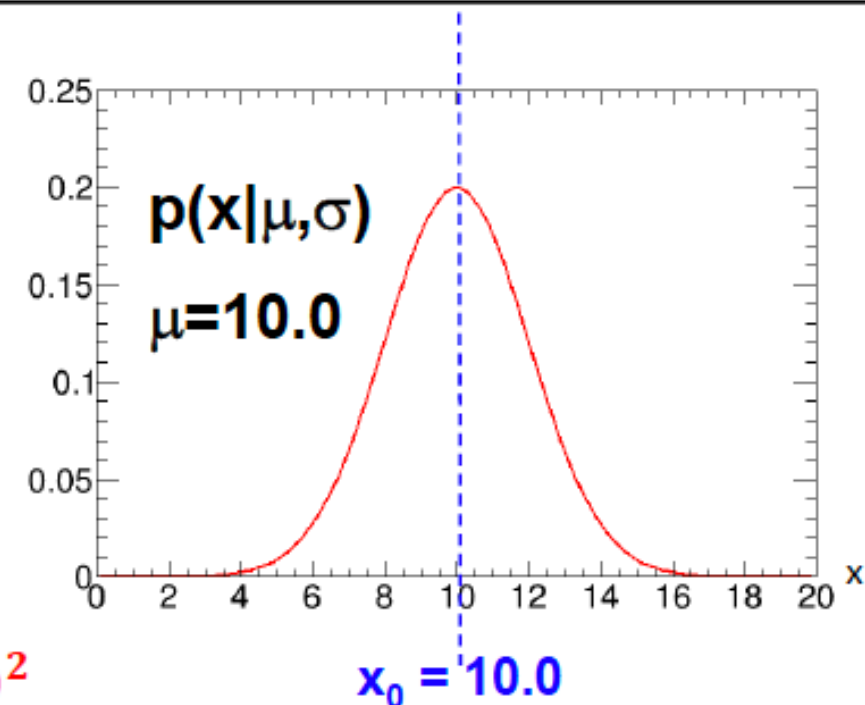
Suppose  $x_0 = 10.0$  is observed.

$$\mathcal{L}(\mu) = \frac{1}{\sqrt{2\pi(0.2\mu)^2}} e^{-(x-\mu)^2/2(0.2\mu)^2}$$

$\mathcal{L}(\mu)$  for observed  $x_0 = 10.$  :

$$\mu_{ML} = 9.63$$

What is confidence interval for  $\mu$ ?



# Confidence intervals

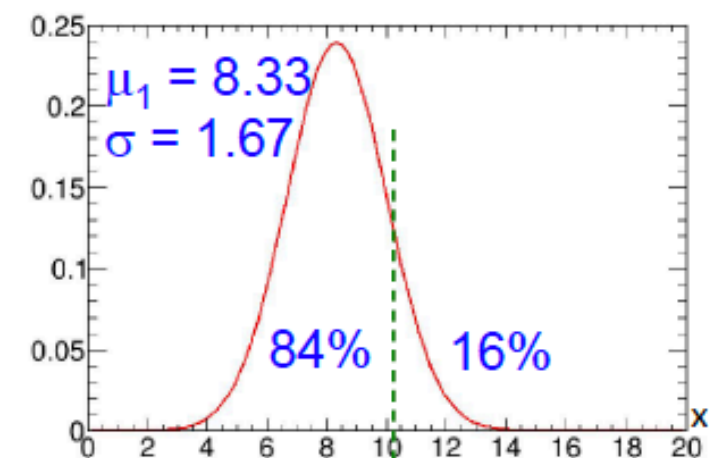
**Gaussian pdf  $p(x|\mu,\sigma)$  with  $\sigma$  a function of  $\mu$ :  $\sigma = 0.2 \mu$   
Observed  $x_0 = 10.0$ .**

Find  $\mu_1$  such that 84% of  $p(x|\mu_1, \sigma=0.2\mu_1)$  is below  $x_0 = 10.0$ ; 16% of prob is above.

Solve:  $\mu_1 = 8.33$ .

$[\mu_1, \infty]$  is 84% C.L. confidence interval

$\mu_1$  is 84% C.L. *lower* limit for  $\mu$ .

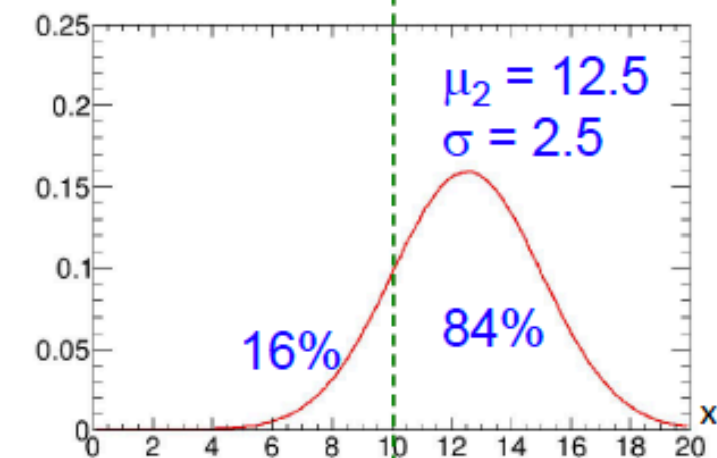


Find  $\mu_2$  such that 84% of  $p(x|\mu_2, \sigma=0.2\mu_2)$  is above  $x_0 = 10.0$ ; 16% of prob is below.

Solve:  $\mu_2 = 12.5$ .

$[-\infty, \mu_2]$  is 84% C.L. confidence interval

$\mu_2$  is 84% C.L. *upper* limit for  $\mu$ .



Then 68% C.L. *central* confidence interval is  
 $[\mu_1, \mu_2] = [8.33, 12.5]$ .

# LEE at Fermilab, the “Oops-Leon” discovery

---

Leon Lederman in the '60-'70 led many of the key experiments that laid the foundations of the standard model.



In 1976, Lederman's group announced the observation of a new particle produced in collisions 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)

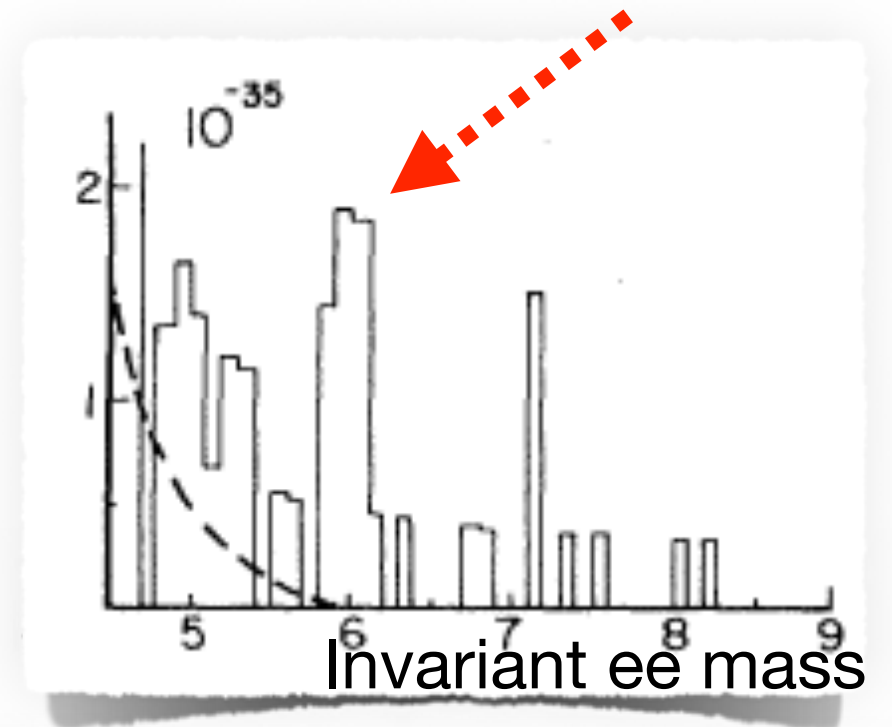
We report preliminary results on the production of electron-positron pairs in the mass range 2.5 to 20 GeV in 400-GeV  $p$ -Be interactions. 27 high-mass events are observed in the mass range 5.5–10.0 GeV corresponding to  $\sigma = (1.2 \pm 0.5) \times 10^{-35} \text{ cm}^2$  per nucleon. Clustering 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.

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.<sup>7</sup> 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.<sup>8</sup> 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 using muon pairs found the actual Upsilon meson, at 9.5 GeV.

Nobody cared too much about the 6 GeV fluke, which someone dubbed “Oops-Leon” in a pun over Lederman’s and the Upsilon’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,<sup>(a)</sup> H. D. Snyder, and J. K. Yoh  
*Columbia University, New York, New York 10027*

and

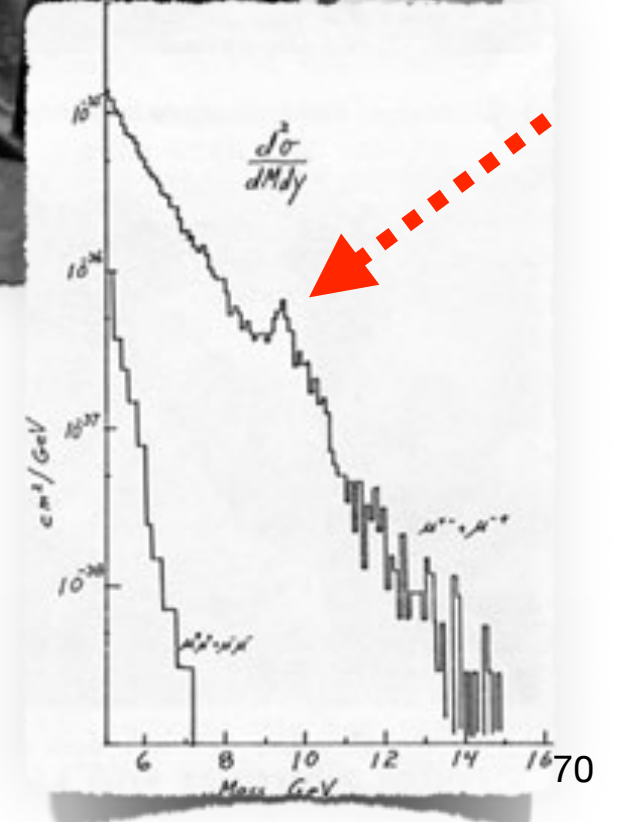
J. A. Appel, B. C. Brown, C. N. Brown, W. R. Innes, K. Ueno, and T. Yamanouchi  
*Fermi National Accelerator Laboratory, Batavia, Illinois 60510*

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.



# Where is “elsewhere”?

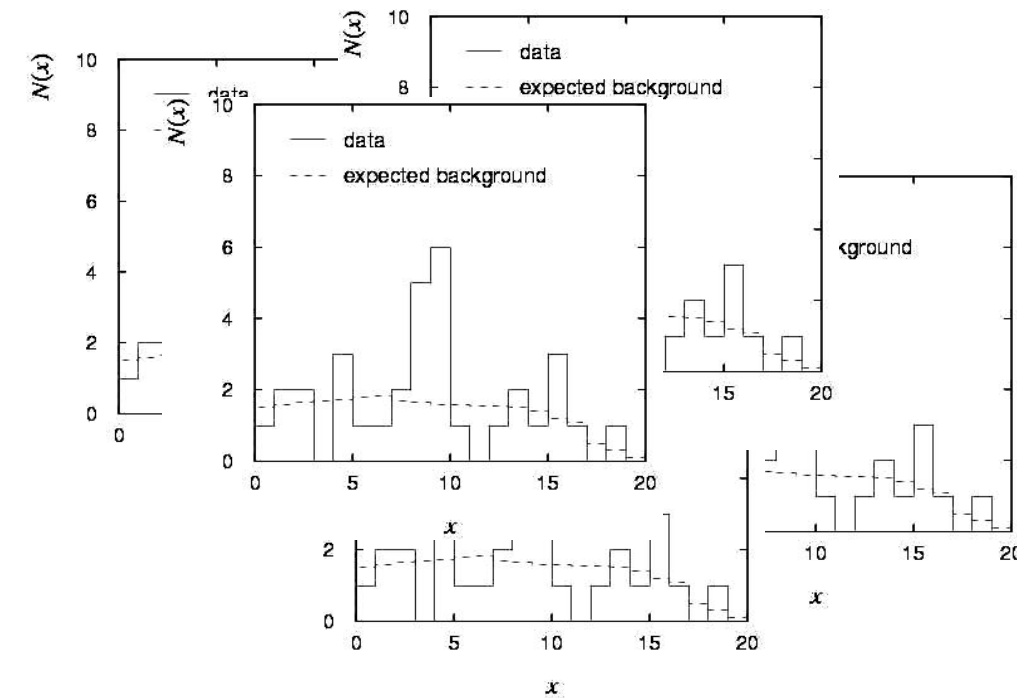
Tenths, or hundreds, or thousands of distributions may have been inspected, in the same analysis or in other analyses.

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?

**Guidance (consensus at the Banff 2010 Statistics Workshop): limit the testing space to models that are inspected within a single published analysis**

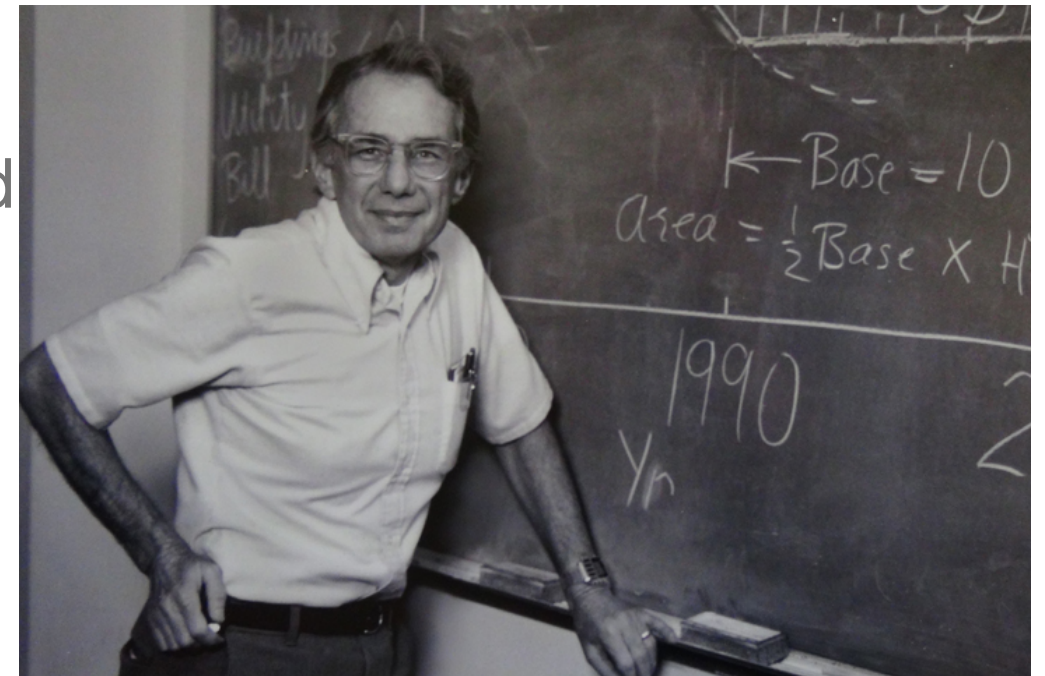


# Far-out hadrons

---

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 multiple 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..."*

*"[Dorigo]*



# 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  $5\sigma$  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  $5\sigma$  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.”*