

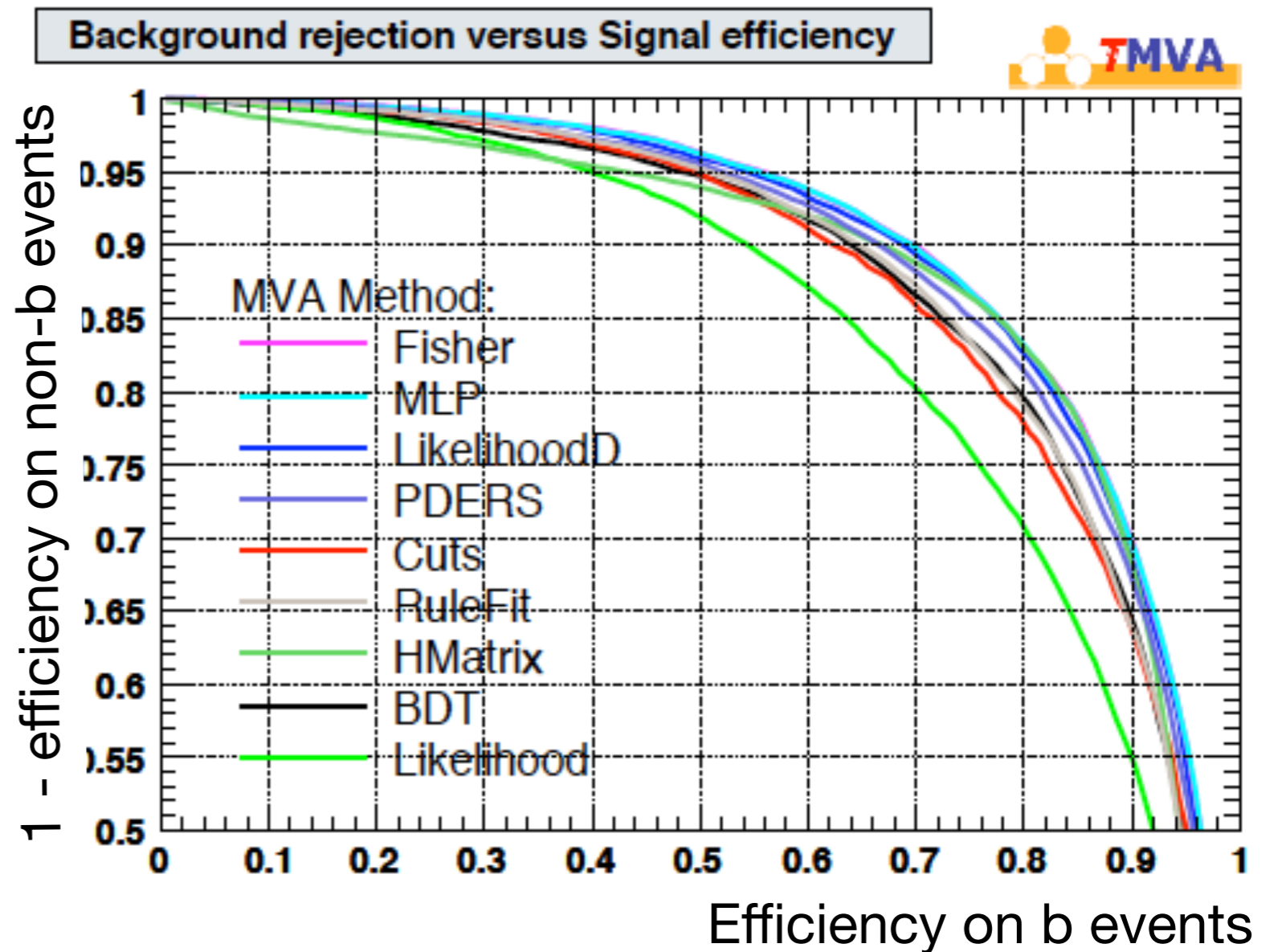
Statistics for HEP (2/3)

Diego Tonelli (INFN Trieste)
diego.tonelli@cern.ch

Pre-SUSY2017 School
Dec 7, 2017 - TIFR, Mumbai

Recap: role of prior

Cannot answer.



Need to know the fraction of b-jets in my sample, **that is the prior $p(\text{b-jet})$.**

Frequentist too believe in Bayes theorem

Application of Bayes' theorem to random events for which prior information is known is the most powerful way of exploiting all the available information.

Knowledge of the **probability distribution $p(x|m)$** and the **prior probabilities for m** (prior to the observation of x) is very powerful.

It allows to use the observation of x to update the prior knowledge and therefore determine the **posterior probability $p(m|x)$, that is the “backward process” probability** - which offers all information one might possibly want on m

Bayesian Inference

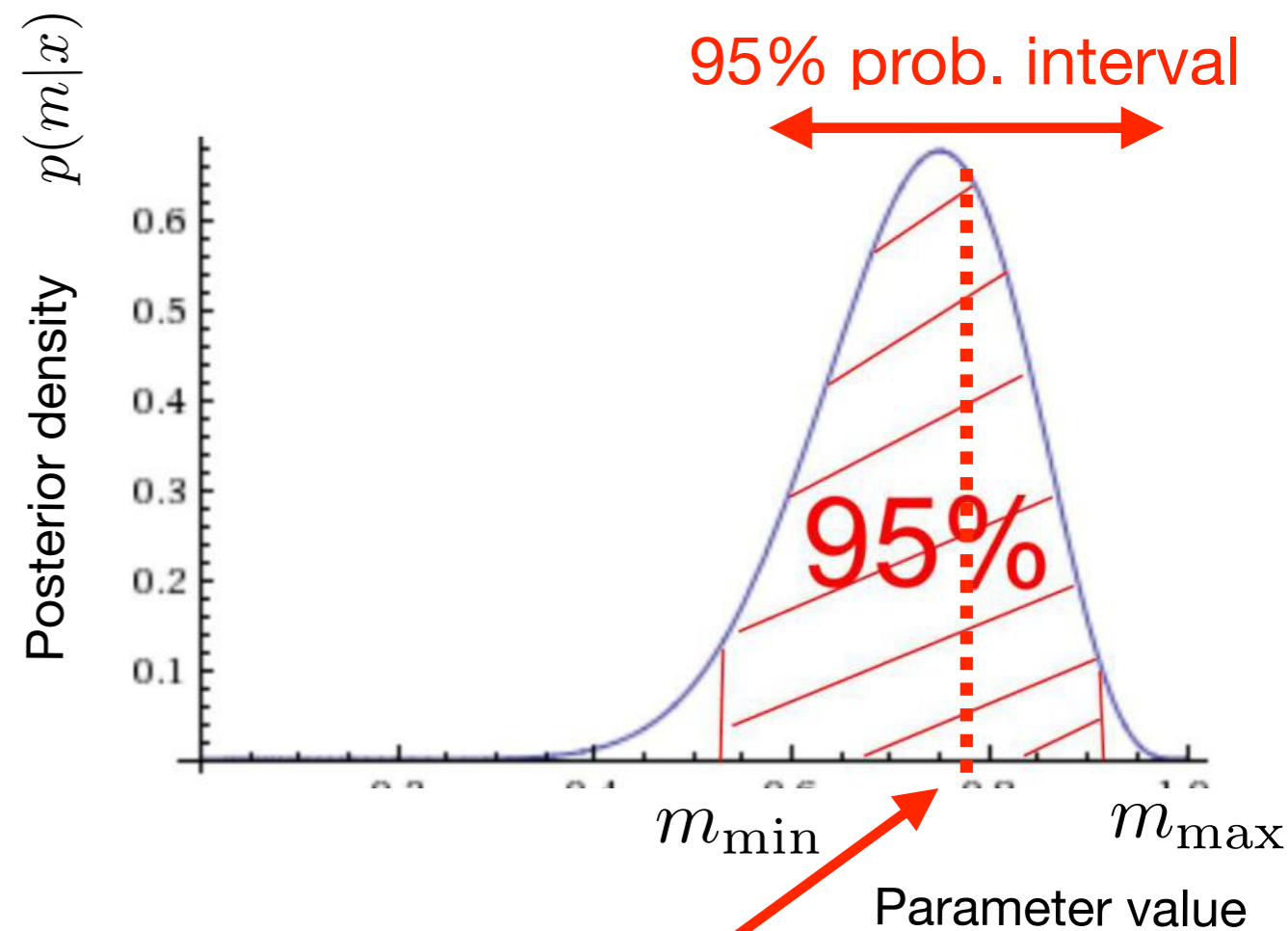
Thanks to the prior one determines $p(m|x)$: **the posterior probability density for the theory given the data.**

Once $p(m|x)$ is known, the rest is straightforward:

Point estimate Mean of $p(m|x)$, which minimizes the variance of m . Alternatively, value m_{best} that maximizes $p(m|x)$. But it depends on metric: differs if parameter is m or any function $f(m)$.

Interval estimate interval (not unique) of m values such that

$$\int_{m_{\min}}^{m_{\max}} p(m|x) dm = \alpha \quad (\text{e.g., } \alpha = 95\%)$$



Highest posterior density

What if priors aren't known or cannot be defined

- Frequentist: give up on getting $p(m|x)$. Revert to an estimate based only on data and the assumed model, not on prior knowledge.
- Bayesian: stick to Bayes' theorem by assuming a prior

Both options are though businesses, as priors do carry information. E.g, the posterior $p(m_0|x)$ is zero for any value $m=m_0$ for which $p(m_0)=0$ regardless of what are the observed data

Because HEP folks expect objective/repeatable results that are free from subjective input and can be interpreted in terms of coverage (more later), many Bayesian analyses make an effort toward using priors that have minimal influence on the result.

Flat priors

Uniform (“flat”) priors are commonplace in HEP papers. *“Knowing nothing about a parameter, I assign equal probabilities to all its possible values”* (the noninformative argument)

Sounds intuitively plausible and has attractive practical features: it’s easy and the parameter value that maximizes the posterior density is the same that maximizes the likelihood.

However, flat priors have serious issues: (i) cannot be normalized without a cutoff (ii) puts most of belief at infinity (iii) the noninformative argument is ill-defined, as any pdf can be transformed into a flat pdf and you’ll get a different answer if the prior is flat in m , $1/m$, $\log(m)$ etc..

All of this **exacerbates with increasing dimensionality of the space of parameters**

Lot of thinking (Jeffrey’s most notably) went into pursuing priors containing “as little information as possible”, so that the posterior is dominated by the data.

A better approach - assessing sensitivity to priors

Convincing support of Bayesian results is typically achieved through analysis' sensitivity studies.

Investigate the sensitivity of one's analysis on prior choices by, e.g., looking at the median expected results in simulated events, repeat the analysis with various choices for priors, or on smaller subsets of the sample.

Sensitivity analysis provides essential information on **how much of the final result $p(m|x)$ is driven by data ($p(x|m)$) and how much by the prior $p(m)$** and is therefore a very desirable “calibration” of any Bayesian result.

T. AALTONEN *et al.*

TABLE V. Summary of the sensitivity study. The 68% credibility interval on $\beta_s^{J/\psi\phi}$ is given for the unconstrained result and when $2|\Gamma_{12}^s|$ is constrained to its SM prediction.

Variation	Constrained	Unconstrained
Default	[0.09,0.32]	[0.11,0.41]
Flat $\sin 2\beta_s^{J/\psi\phi}$	[0.08,0.31]	[0.09,0.37]
Flat $\cos\delta_{\perp}$	[0.09,0.33]	[0.10,0.43]
Flat $\cos\delta_{\parallel}$	[0.09,0.32]	[0.11,0.41]
Previous three together	[0.07,0.31]	[0.09,0.39]
Flat in amplitudes	[0.09,0.32]	[0.11,0.41]
Gaussian mixing-induced CP violation	[0.09,0.34]	

Example from PRD 85, 072002 (2011)

What Can Be Computed without Using a Prior?

Not $P(\text{constant of nature} \mid \text{data})$.

- 1) *Confidence Intervals* for parameter values, as defined in the 1930's by Jerzy Neyman.
- 2) *Likelihood ratios*, the basis for a large set of techniques for point estimation, interval estimation, and hypothesis testing.

These can both be constructed using the frequentist definition of P .

The likelihood

Likelihood function

Model $p(x|m)$ evaluated at fixed data. Essential in any inference

- probability density function $p(x|m)$ of observing generic data x , given the unobservable value of the parameter m .
- Then take actual sample of observed data x_0 and evaluate $p(x_0|m)$
- The likelihood $L(m) = p(x_0|m)$ is a function of parameter m given your data

Connected to *probability for observing data x* for different choices of the value of the parameter m , **not** the probability that m has some value given the data.

Likelihood is a complete summary of the data information relevant to the estimate at hand. Ideally should be published as is.

A likelihood is not a pdf

The **probability density function** $p(x|m)$ is a parametric **function of the observable data x** .

The **likelihood function** $L(m)$ is a function of the **unobservable parameter m** .

The pdf, a probability density of the data (random variable), should be normalized to unity over the domain of the random variable.

$$\int_{\mathcal{X}} p(x|m) dx = 1$$

The likelihood, a function of the parameter m , obeys no specific normalization.

$$\int_{\mathcal{M}} p(x_0|m) dm = ?$$

In addition, **the function values $L(m)$ are invariant under reparametrization of m into $f(m)$** : $L(m) = L[f(m)]$. No Jacobians here, reinforcing the notion that $L(m)$ is not a pdf for m .

Maximum of the likelihood

The likelihood expresses the probability of observing the data you observed as a function of the parameter value m .

Given some data, parameter values m_{low} that make $L(m)$ small are disfavored: it would be unlikely for nature to generate that set of observed data, had the true value of m been m_{low} . Conversely, values m_{high} that make $L(m)$ large are favored

HEP usually deals with repeated observations x that are independent and identically distributed. If the likelihood for a single observation x' is

$$L(m) = p(x'|m),$$

the likelihood for the whole experiment is the product of the single-event likelihoods

$$L(m) = \prod p(x|m)$$

Example — exponential

Decay process. Assume exponential model. Pdf

$$p(t|\tau) = \frac{1}{\tau} e^{-t/\tau}$$

Probability density of survival after time t

Then we observe N decay times and infer the lifetime by maximizing the likelihood.

$$L_k(\tau) = p(t_k|\tau) = \frac{1}{\tau} e^{-t_k/\tau}$$

Likelihood of observation of $t = t_k$

$$L(\tau) = \prod_{k=1}^N \frac{1}{\tau} e^{-t_k/\tau} = \left(\frac{1}{\tau}\right)^N \exp\left(-\frac{\sum_{k=1}^N t_k}{\tau}\right)$$

Likelihood of observation of the full data set

Example - exponential (cont'd)

As high values of the likelihood are associated with favored values of the unknown parameter (lifetime tau here), set to zero derivative

$$\frac{dL(\tau)}{d\tau} = \left[\sum_{k=1}^N t_k (1/\tau)^{N+2} - N(1/\tau)^{N+1} \right] \exp \left(-\frac{\sum_{k=1}^N t_k}{\tau} \right)$$

$$dL(\tau)/d\tau = 0 \text{ implies } \hat{\tau} = \sum_{k=1}^N t_k / N \quad \text{tau corresponding to the average of observed decay times maximizes the likelihood}$$

Had I framed my inference in terms of natural width, $\Gamma = 1/\tau$

$$L(\Gamma) = \Gamma^N \exp \left(-\Gamma \sum_{k=1}^N t_k \right) \quad \hat{\Gamma} = N / \left(\sum_{k=1}^N t_k \right) = 1/\hat{\tau}$$

Because L is invariant under parameter transform, its maximum too is so.

Example — Poisson

Model: Poisson-distributed signal, no background. $p(j|\mu) = \frac{\mu^j}{j!} e^{-\mu} = L(\mu)$

Observe $j = 5$. What's the maximum likelihood estimate for my Poisson mean?

Probability mass function

$$p(j|\mu) = \frac{\mu^j}{j!} e^{-\mu} :$$

(Discrete) function of data

Likelihood

$$L(\mu|j = 5) = \frac{\mu^5}{5!} e^{-\mu}$$

(Continuous) function of physics par.

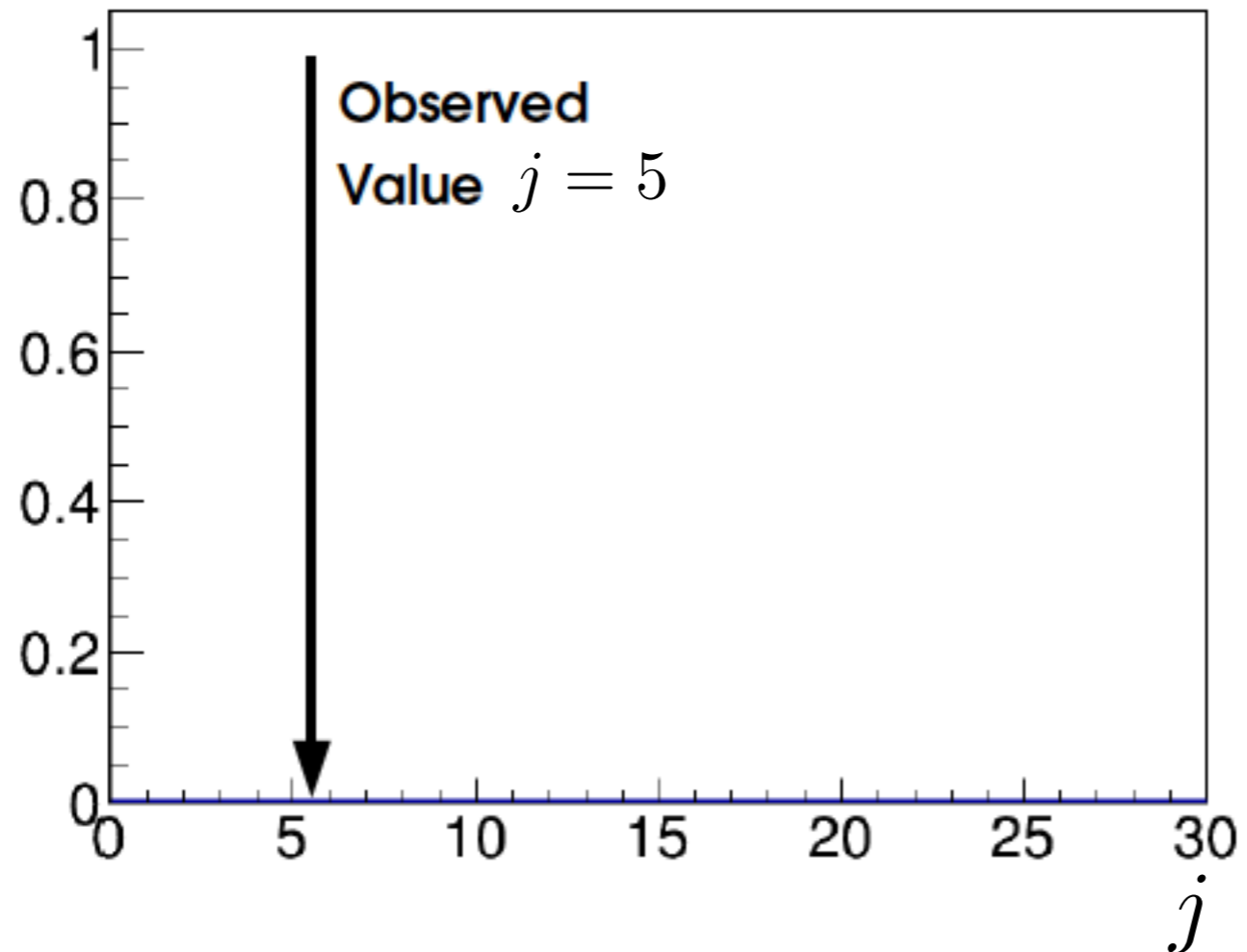
Minimize $-\ln L$. $-\frac{d}{d\mu} \ln L(\mu)|_{\hat{\mu}} = 0$ $-\frac{d}{d\mu} (\mu - j \ln \mu + \ln j!) = 1 - \frac{j}{\mu}$

Given observation j , the ML estimator of the mean rate of success μ is $\hat{\mu} = j$

Illustrated

Model: Poisson-distributed signal, no background.

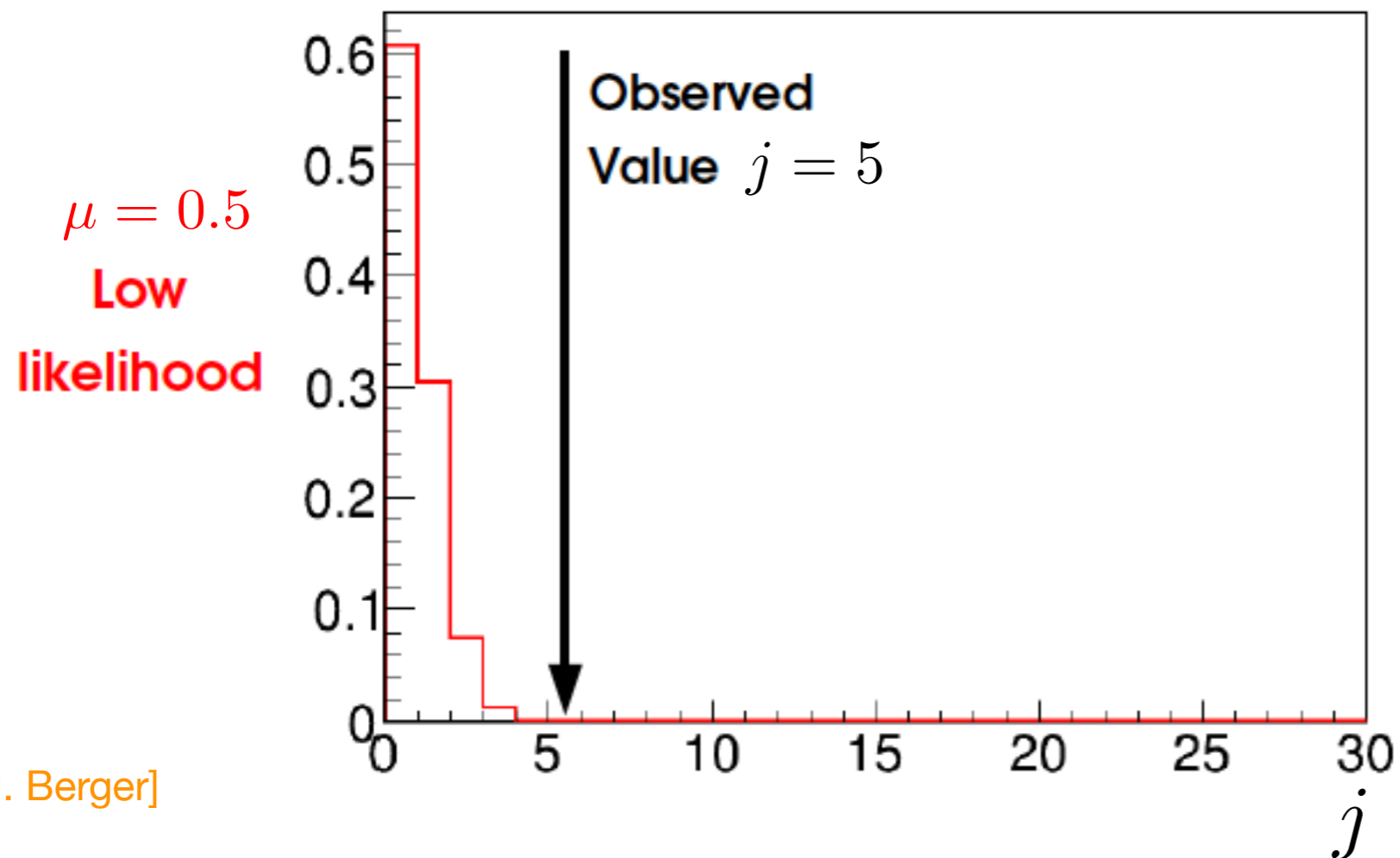
Observe $j = 5$.



Poisson illustrated

Model: Poisson-distributed signal, no background.

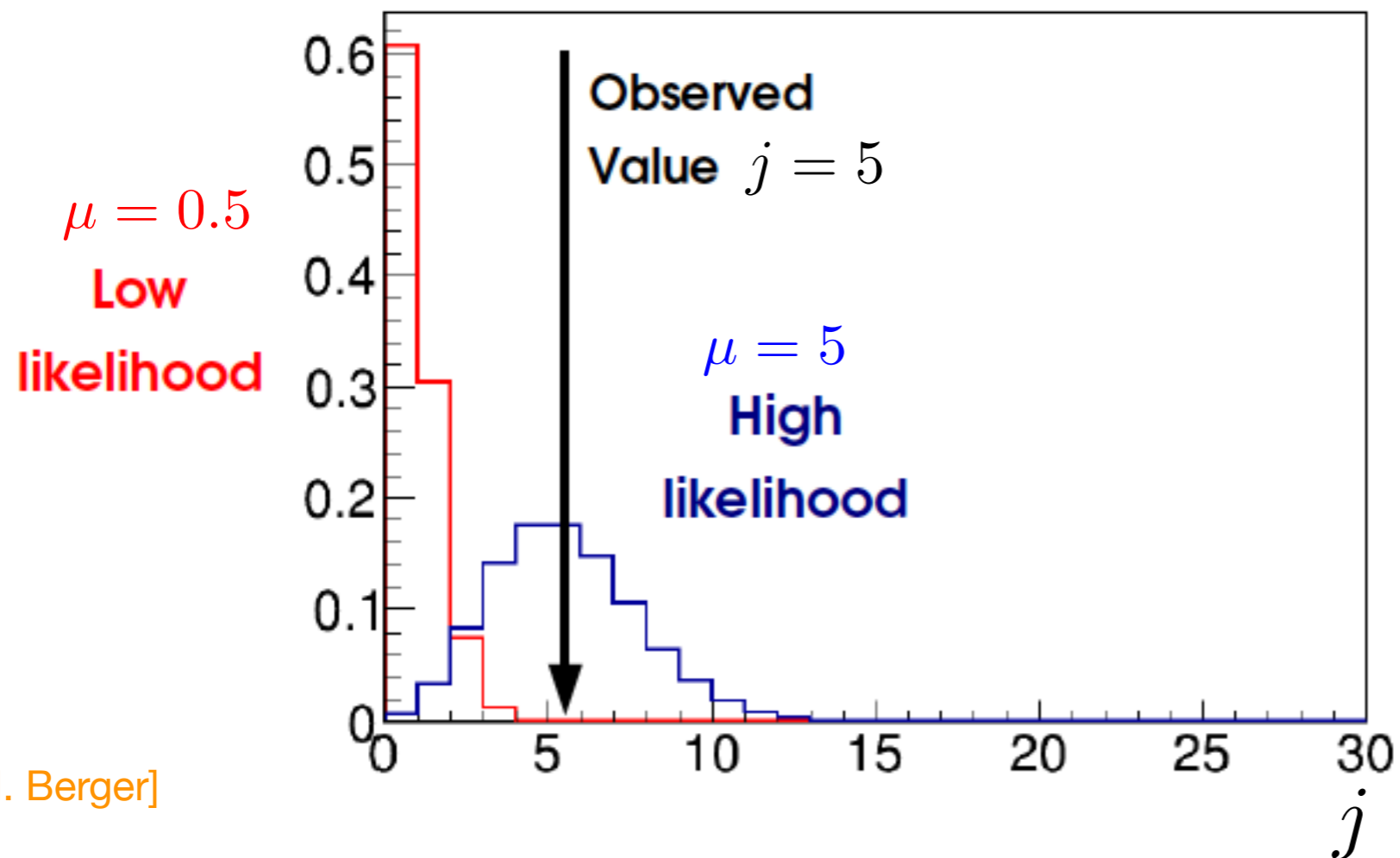
Observe $j = 5$.



Poisson illustrated

Model: Poisson-distributed signal, no background.

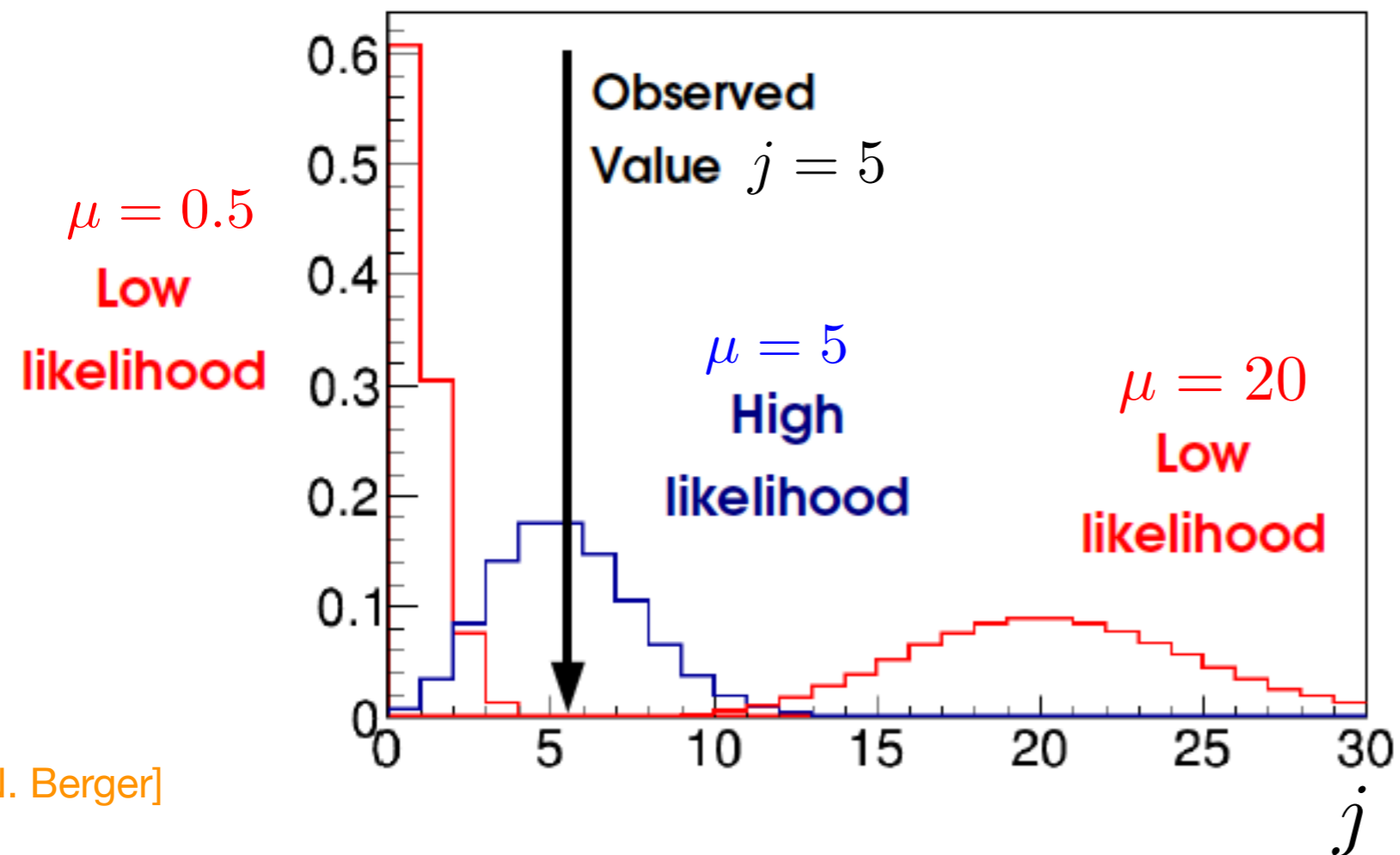
Observe $j = 5$.



Poisson illustrated

Model: Poisson-distributed signal, no background.

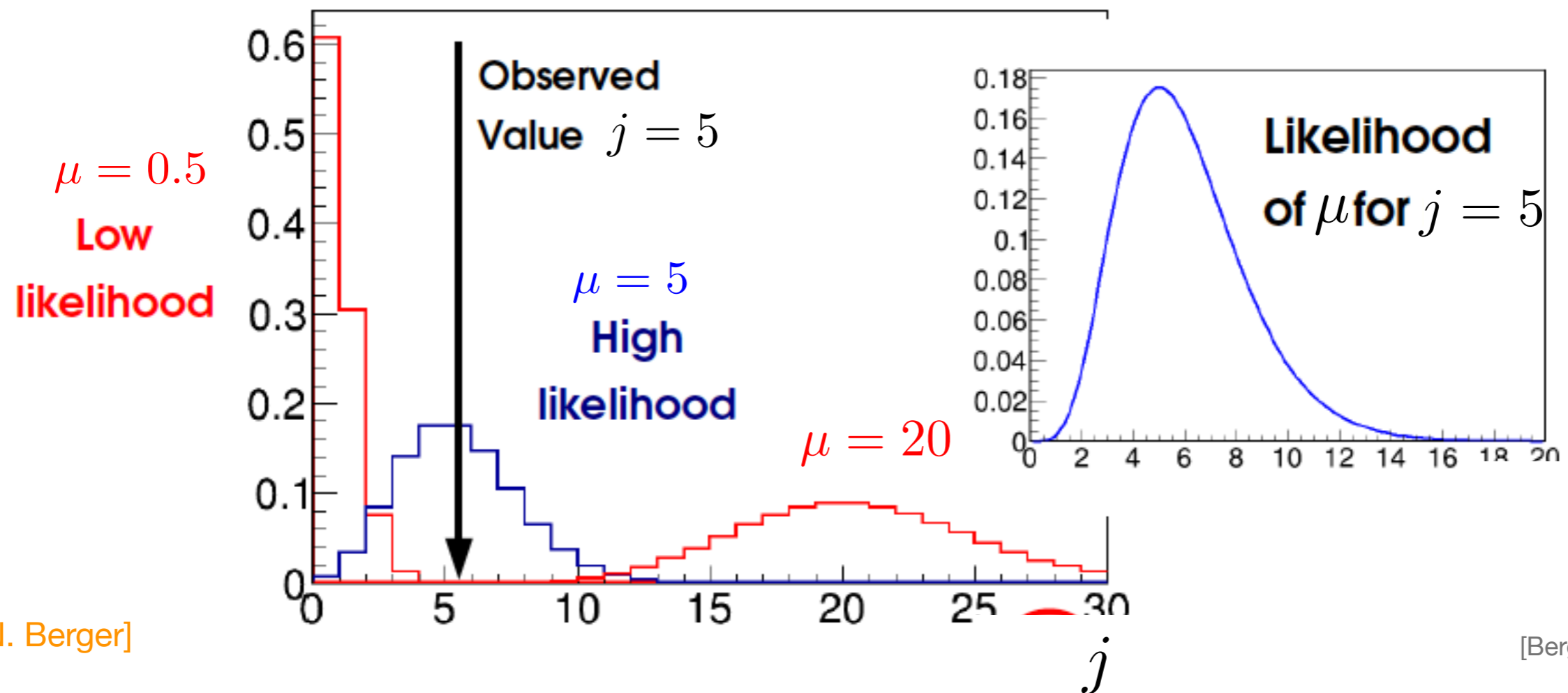
Observe $j = 5$.



Poisson illustrated

Model: Poisson-distributed signal, no background.

Observe $j = 5$.



Extended likelihood

Sometimes the number of events of the sample N is itself part of the inference, e.g., measure a production cross sections.

The result of the experiment is $N, \vec{x}_1, \vec{x}_2, \dots, \vec{x}_N$, it is convenient to use **the extended likelihood**, where addition of a Poisson term (due to total event count) properly accounts for the fluctuations on N

$$L(\nu, m) = \frac{\nu^N}{N!} e^{-\nu} \prod_{i=1}^N p(x_i; m)$$

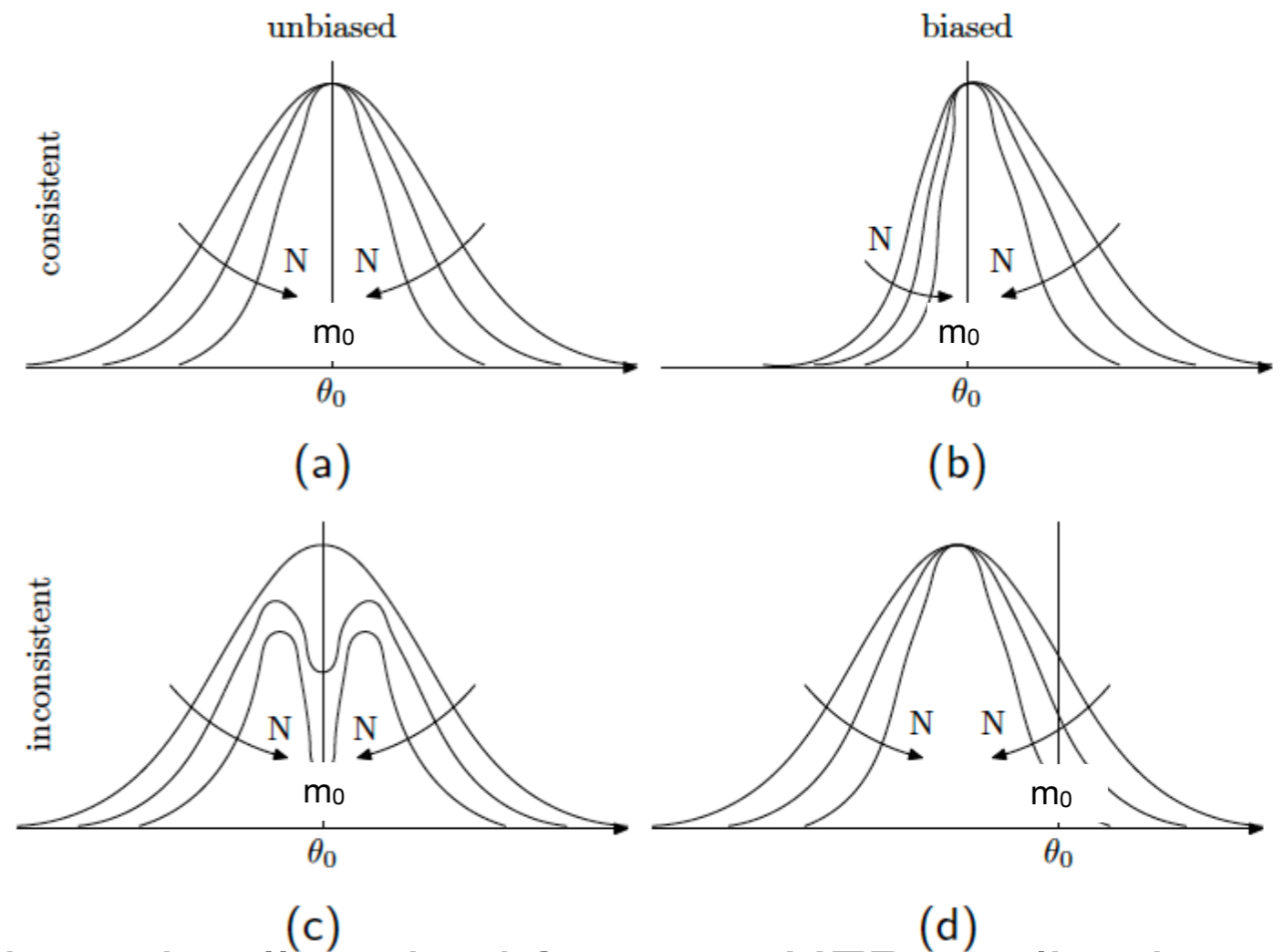
Besides the uncertainties in the proportions of each class of events in the sample, the Poisson term accounts for the global fluctuation on N

Estimators

The maximum likelihood is an estimator.

Estimator — a function of the data $e(x)$ used to provide an *estimate* (“a measurement”) of a parameter. [James]

Estimators are functions of data (random variables), hence estimators are random variables with their own probability distributions. An estimator's performance depends on the properties of its distribution.



The maximum likelihood estimator is optimally suited for most HEP applications and we won't discuss other estimators.



Information on a parameter brought by data

(If it exist) the Fisher information of an observation x on the parameter m , related by the likelihood $p(x|m) = L_x(m)$ is

$$I_x(m) = E \left[\left(\frac{\partial \log(L_x(m))}{\partial m} \right)^2 \right] \quad [I_x(m)]_{ij} = E \left[\frac{\partial \log(L_x(m))}{\partial m_i} \frac{\partial \log(L_x(m))}{\partial m_j} \right]$$

1 parameter many parameters

If (i) the possible values of x do not depend on m and (ii) the likelihood is twice differentiable and derivatives in m and integrals in x commute

$$[I_x(m)]_{ij} = -E \left[\frac{\partial^2 \log(L_x(m))}{\partial m_i \partial m_j} \right]$$

See Eq 28 in <https://arxiv.org/pdf/1007.1727.pdf> for a convenient approximation of the Fisher's information

As for N observations the Fisher information is proportional to N , the precision of cannot improve faster than $1/\sqrt{N}$



Minimum variance bound

An attractive property of an estimator is its precision (variance). Can it be made arbitrarily small at given number of observations N ?

No.
$$\text{Var}(\hat{m}) = E[(\hat{m} - E[\hat{m}])^2] \geq \frac{(1 + db/dm)^2}{I_{\hat{m}}(m)} \geq \frac{(1 + db/dm)^2}{I_x(m)}$$

where \hat{m} estimator of m , $b = E[\hat{m}] - m$ is its bias and $I_x(m)$ is the Fisher information

If inequalities become equalities, \hat{m} **reaches minimum variance: efficient estimator.** Implies that once \hat{m} is known, no further information is brought by complete knowledge of all data x .

See Eq 28 in <https://arxiv.org/pdf/1007.1727.pdf> for a convenient approximation of the Fisher's information

Under weak conditions, the maximum likelihood estimator is asymptotically ($N \rightarrow \infty$) consistent, efficient, and normal (i.e., has Gaussian uncertainties).

NB: does not apply if the range of the observations or the dimensionality of the likelihood depend on the parameter being estimated.

Maximum likelihood variance (“fit error”)

The minimum variance bound offers an approximated estimate of the variance as the curvature (2nd derivative) of the log-likelihood at its maximum.

$$[I_x(m)]_{ij} = -E \left[\frac{\partial^2 \log(L_x(m))}{\partial m_i \partial m_j} \right]$$

$$\begin{aligned} \hat{V}(\hat{m}) &\approx -1/E \left[\frac{\partial^2 \ln L}{\partial m^2} \right] \\ &\approx - \left(\frac{\partial^2 \ln L}{\partial m^2} \right)^{-1} \Big|_{m=\hat{m}} \end{aligned}$$

This is the symmetric uncertainty MINUITs computes after MIGRAD/HESSE
Accurate only for linear problems (Gaussian likelihood).

No guarantee that for N finite the estimator has reached minimum variance. The number of observations needed to approximate asymptotic regime depend on the problem at hand. If in doubt check with toys.

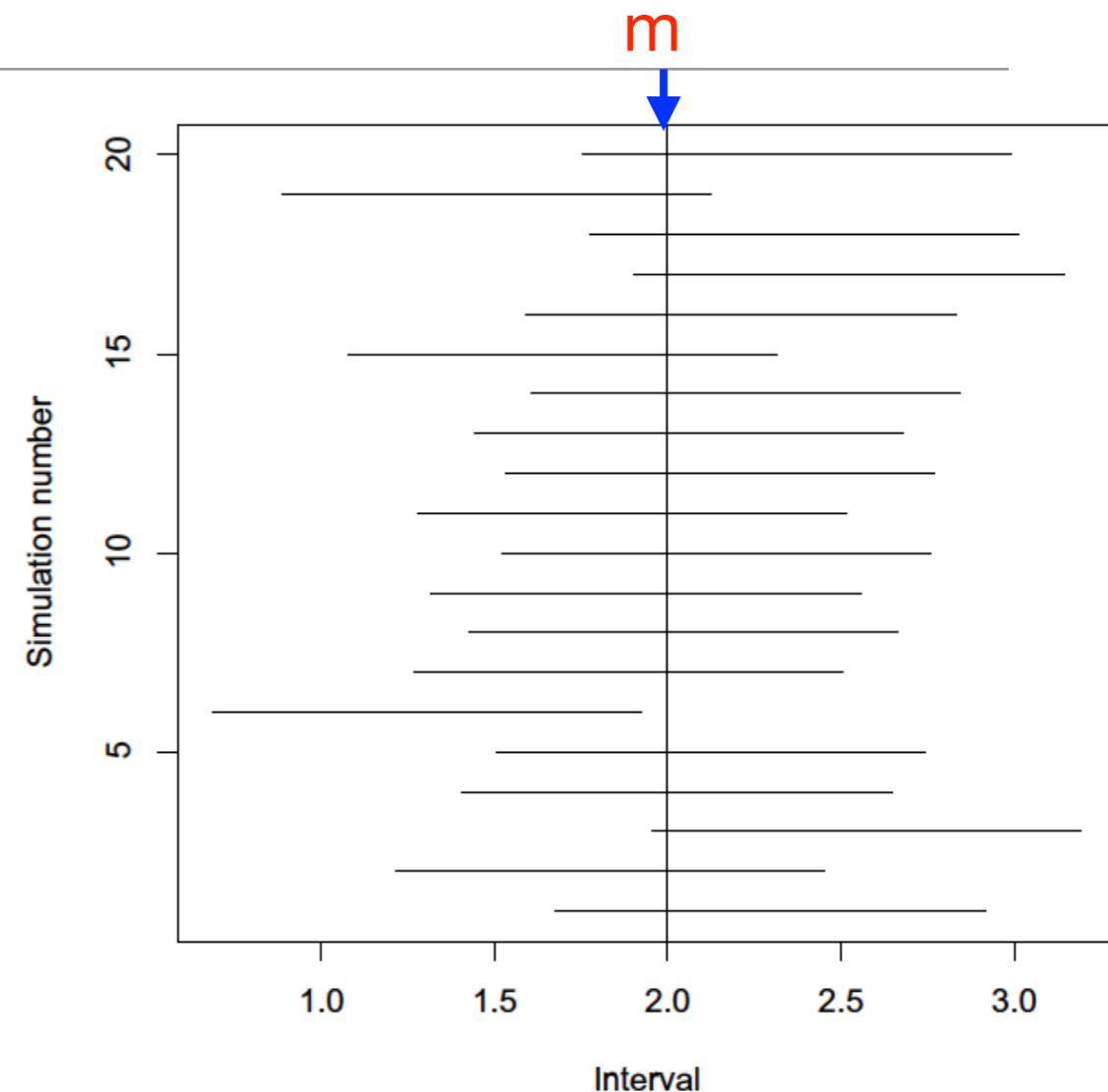
Statistical uncertainty

Repeating our experiment many times, 68.3% of the resulting $[\hat{m}-\sigma, \hat{m}+\sigma]$ intervals include the true value of the parameter

This differs from stating “in 68.3% of the experiments the true value is the $[\hat{m}-\sigma, \hat{m}+\sigma]$ range” or “there is 68.3% probability that the true value is in the $[\hat{m}-\sigma, \hat{m}+\sigma]$ range”

Language is subtle and important. **The true value is not random.** Cannot move around or have a probability.

Only data, that is **the interval extremes**, are **random** and fluctuate around the true value.



95.5% confidence intervals resulting from 20 identical measurements of a true value of 2.0

Coverage

The capability for an inference procedure to yield uncertainties that *cover* the true value with the stated *confidence level* is a fundamental requirement in frequentist inference.

It is also generally desired/expected in HEP (even in Bayesian measurements).

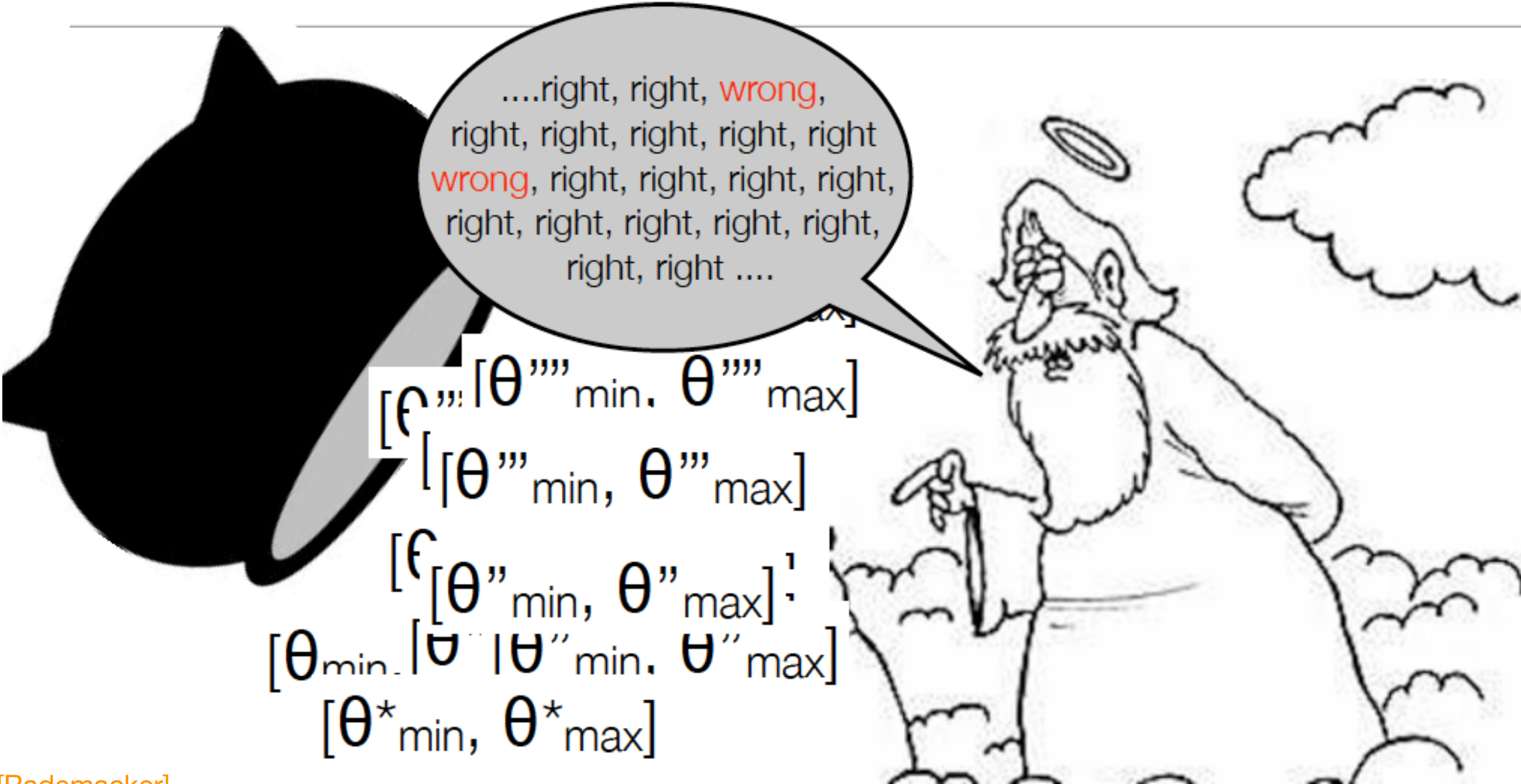
Coverage is a feature of the procedure used, not of a single measurement.

The single interval resulting from a specific measurement may contain or not the true value.

Like in linear algebra one defines a vector as an element of a vector space with some properties, a confidence interval is an element of a confidence set of intervals that have coverage under repeated sampling [Cousins]

Coverage

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

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

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

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

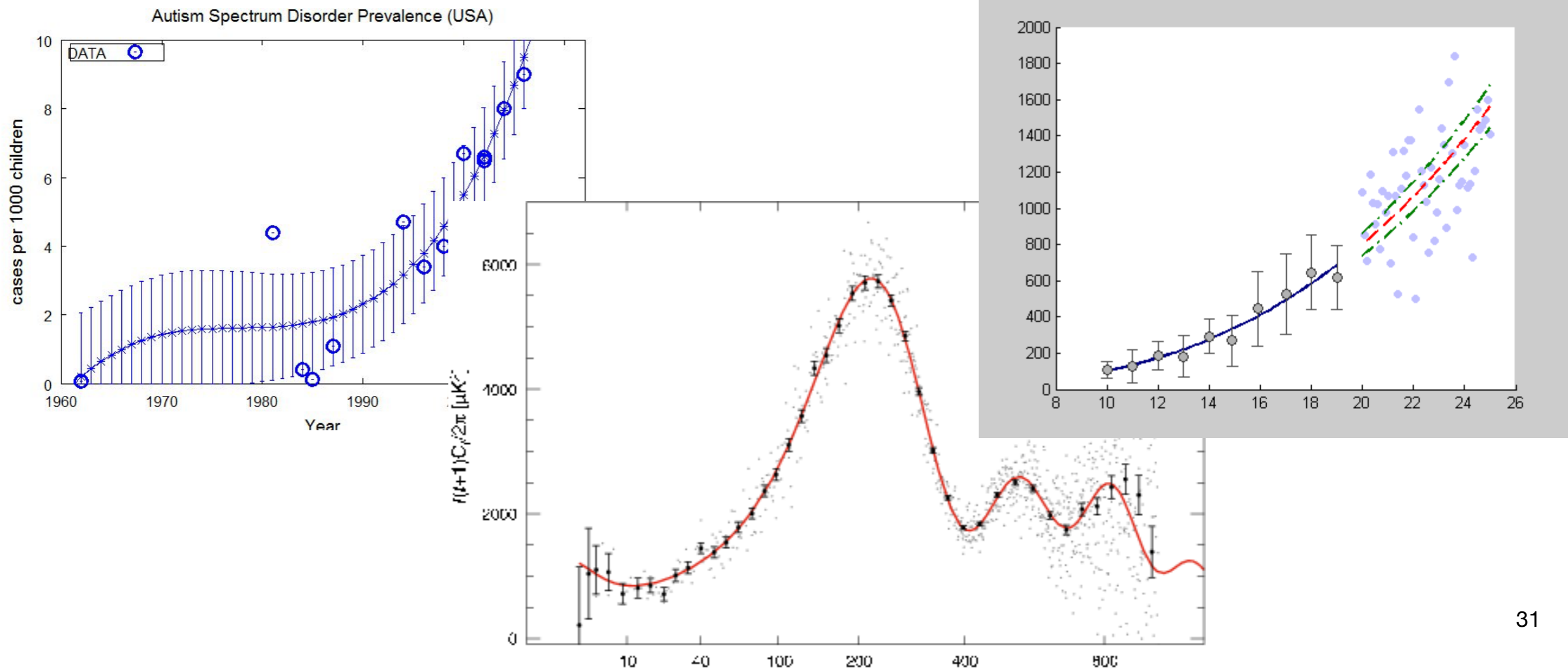
$$[\theta_{\min}, \theta_{\max} | \theta''_{\min}, \theta''_{\max}]$$

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

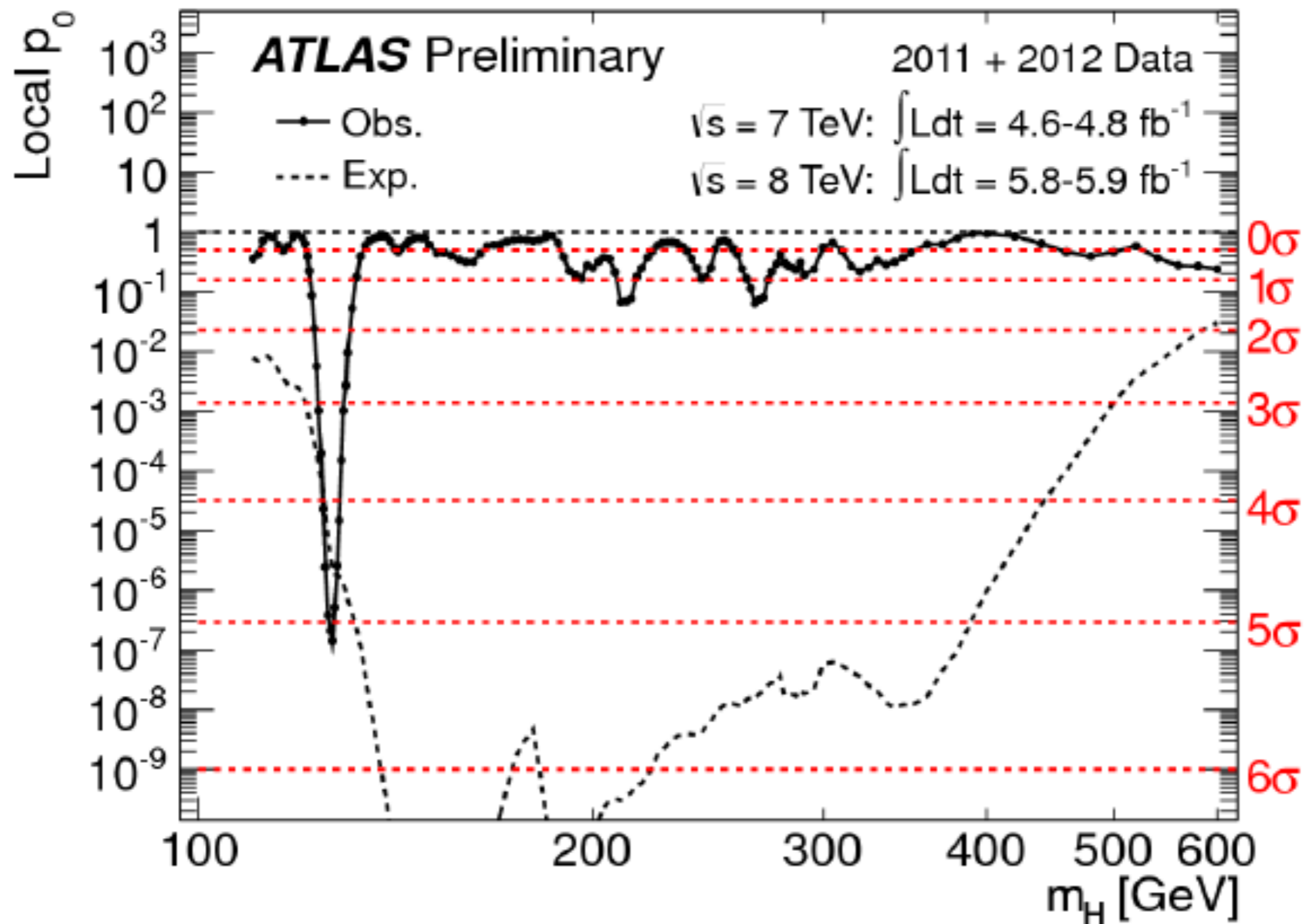
1σ implies that $\sim 1/3$ of points should be off!

One-sigma corresponds to 68.3% confidence level.

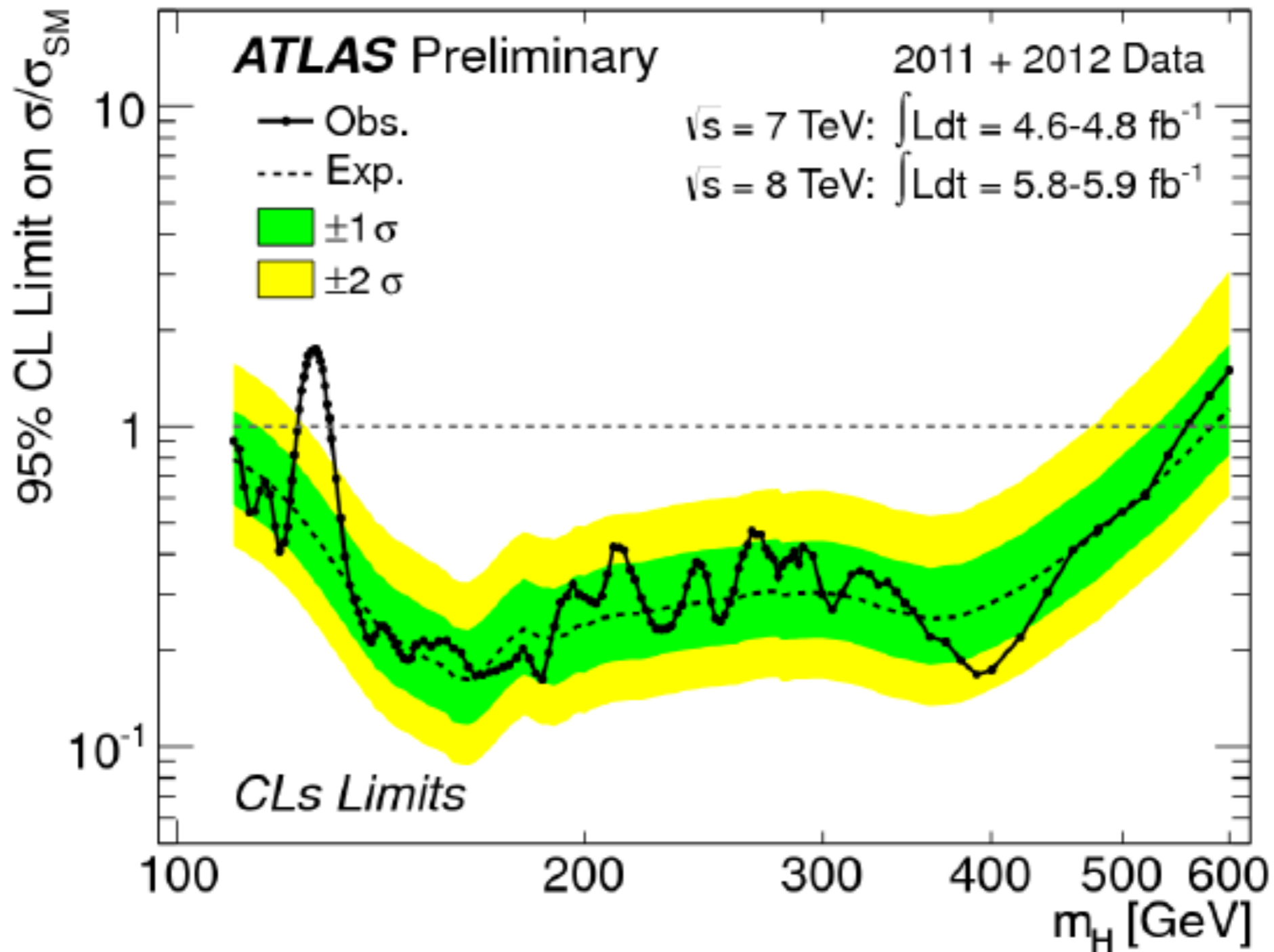
The scatter of points should bring roughly one out of three points farther than the error bars of the others.



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

Given a model $p(x|m)$, with

- unknown value of the parameter m , and*
- known observed data x_0 ,*

The confidence interval construction is a mathematical procedure to address the question:

What are the values of m for which the observed data 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 ?

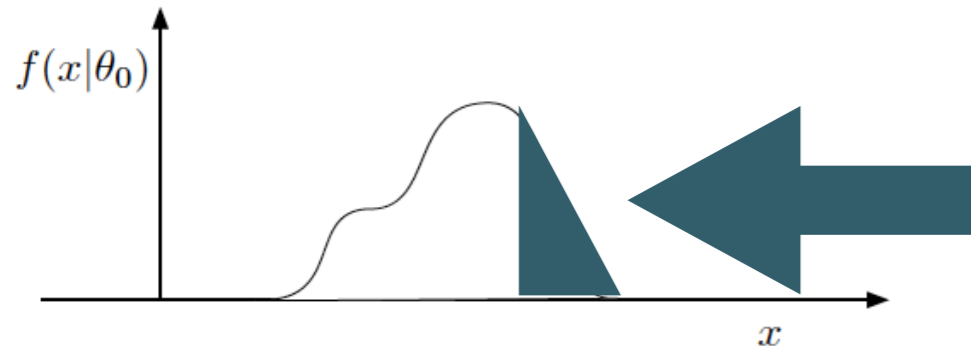
To define “extreme”, need an ordering principle. Rank the values of x for each possible value of m . High rank means not extreme (likely to be included in the interval). Low rank means extreme (likely to be outside of the interval).

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%...

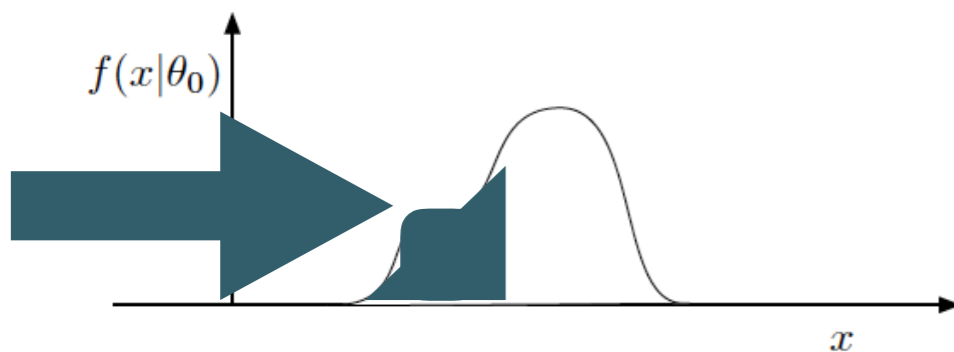
Given a model $p(x|m)$, data x_0 , an ordering, and a CL, the confidence interval $[m_1, m_2]$ includes those values of m for which x_0 aren't “extreme” at the chosen CL

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

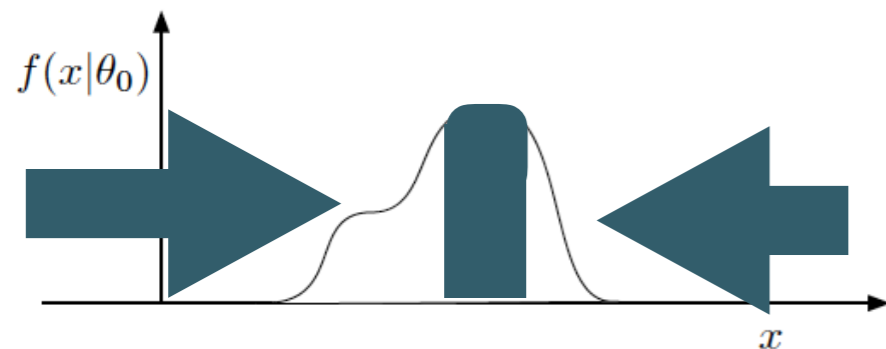
One-sided, two-sided.



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% (1σ) 95.5% (2σ) etc. Define also the lowest-ranked $\alpha = 1 - \text{CL}$ fraction of the most extreme values

Convenient practical trick: 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 values of* m for which x_0 belongs to the set of lowest-valued x that has 16% ($1 - \text{CL}$) probability
- A CL=84% lower limit m_1 *excludes* m values for which x_0 belongs to the set of highest-valued x that has 16% ($1 - \text{CL}$) 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 ρ Directly relevant to efficiency calculation in HEP

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

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

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

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

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

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

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

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 proper confidence interval.

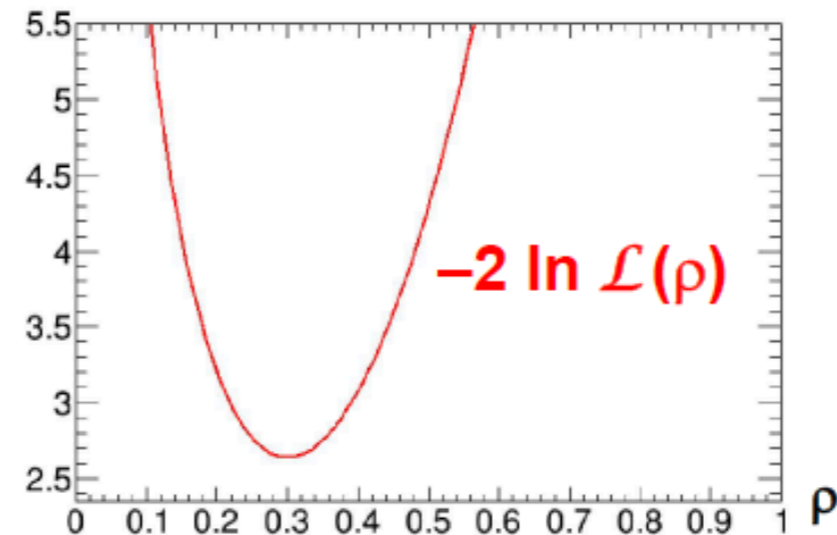
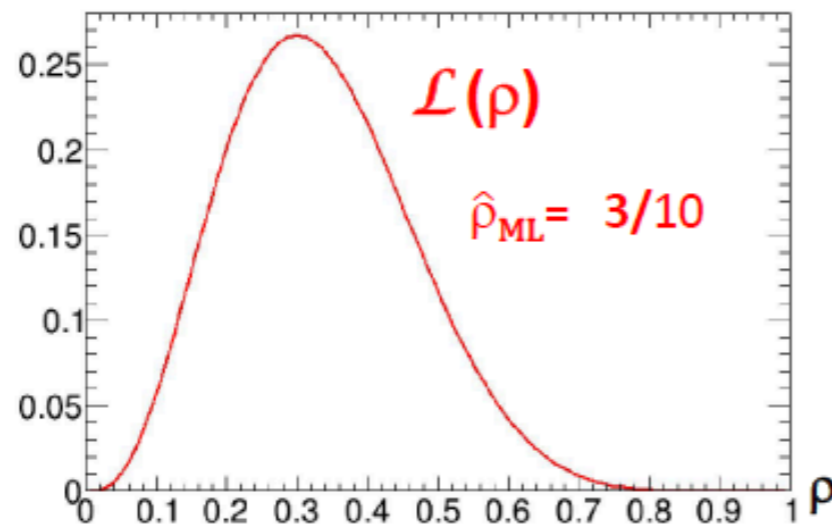
It does not follow the proper logic of frequentist inference: in the construction of the interval each σ should be consistently associated with each ρ value, while here I am using the same σ for all ρ values in the interval.

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

Confidence intervals

Confidence intervals for binomial ρ (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 ρ_1 with C.L. = $1 - (1 - 68\%)/2 = 84\%$

I.e., Find ρ_1 such that $\text{Bi}(n_{\text{on}} < 3 \mid n_{\text{tot}}=10, \rho_1) = 84\%$

Find upper limit ρ_2 with C.L. = 84%

I.e., Find ρ_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 ρ_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 ρ_2 such that

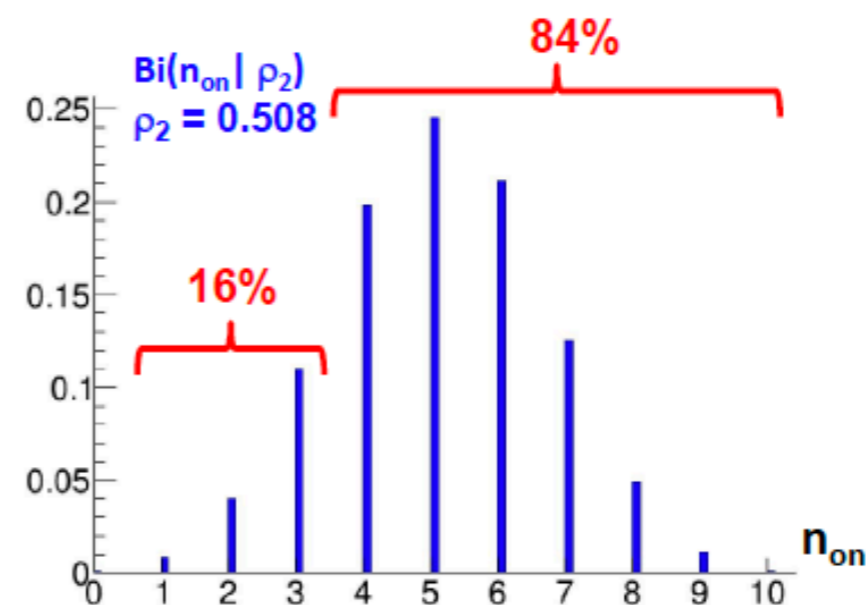
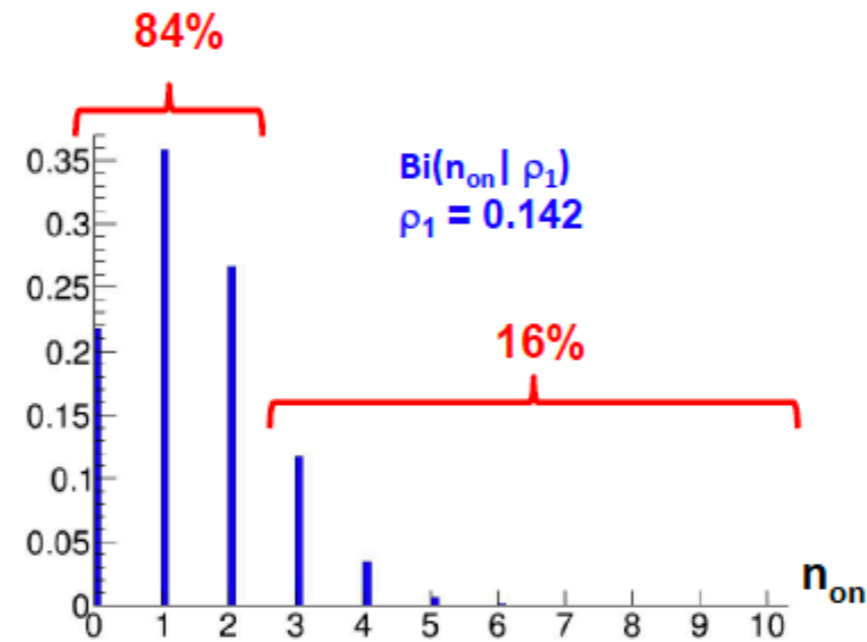
$\text{Bi}(n_{\text{on}} > 3 \mid \rho_2) = 84\%$

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

(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

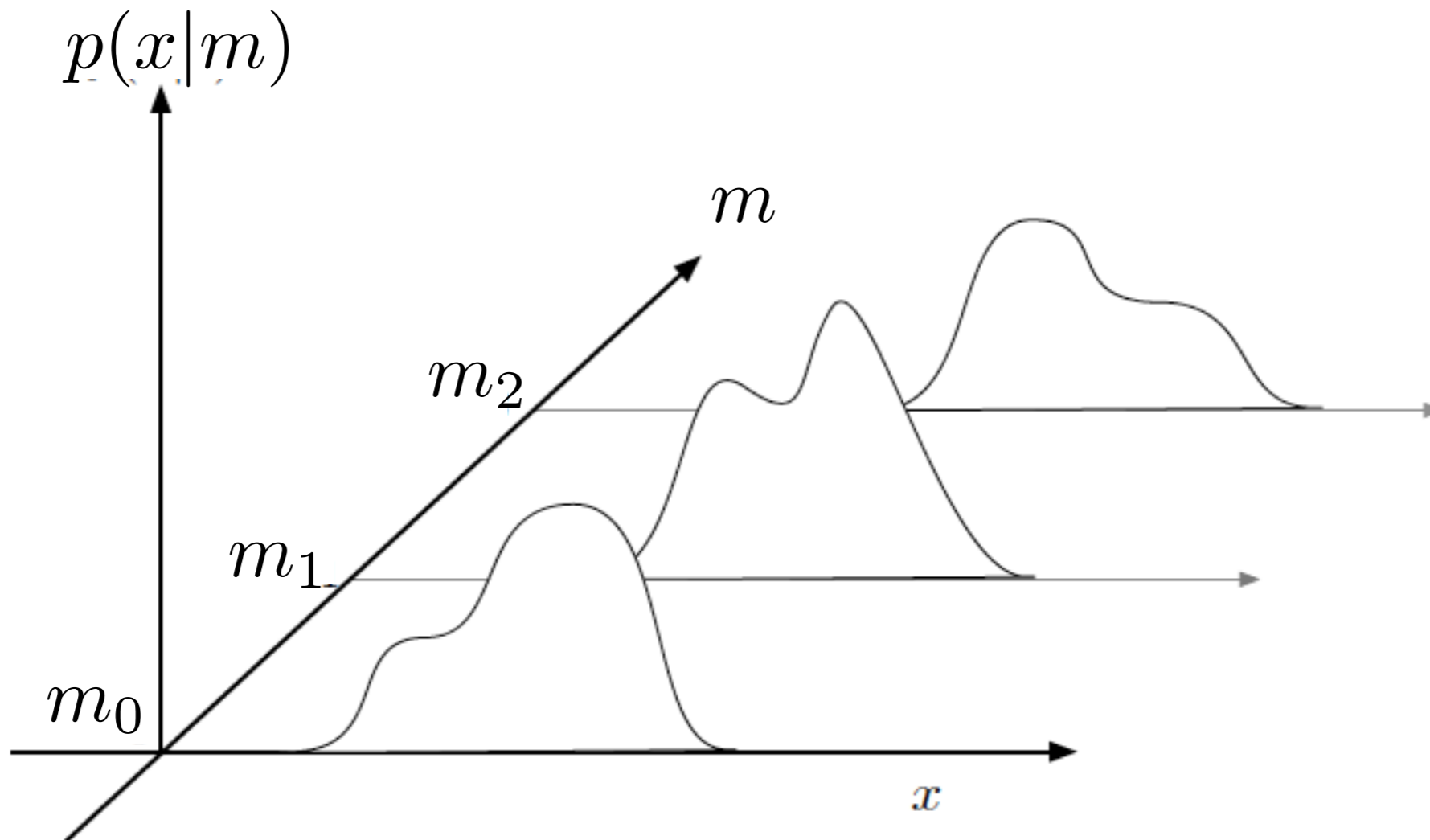
By J. NEYMAN

Reader in Statistics, University College, London

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

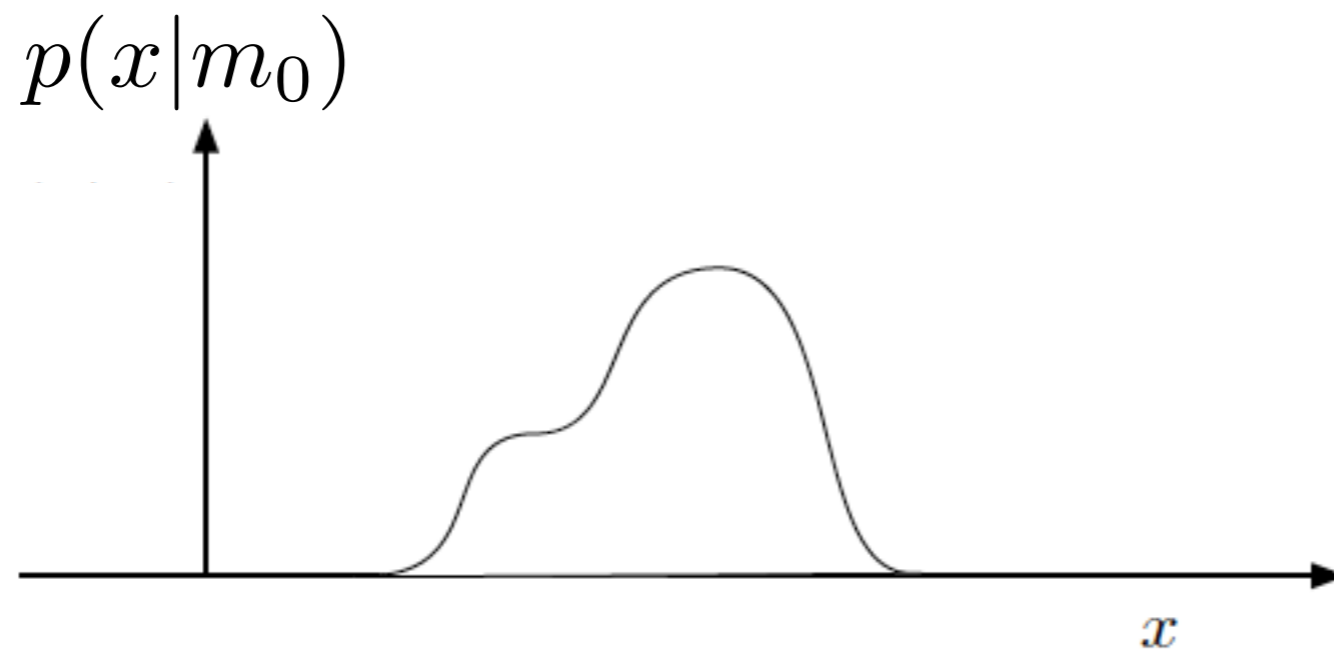
Neyman construction illustrated

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 .



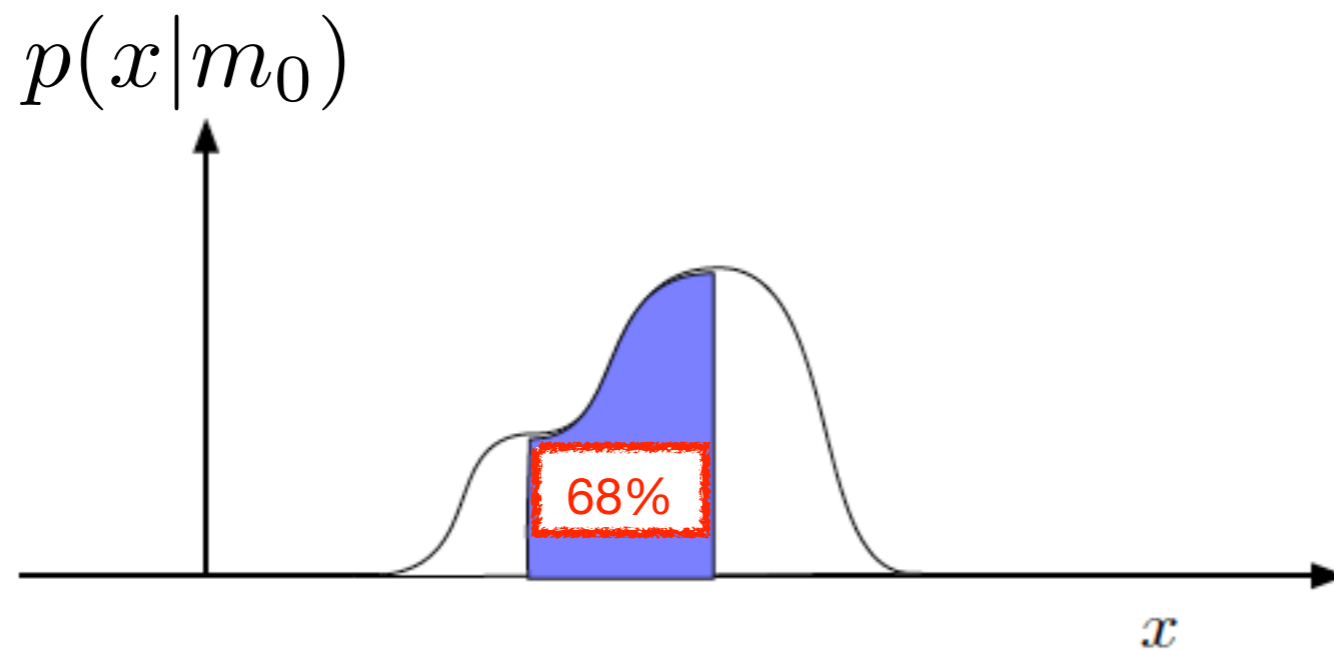
Neyman illustrated I

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 $p(x \in \text{range} \mid m_0) = 68\%$.

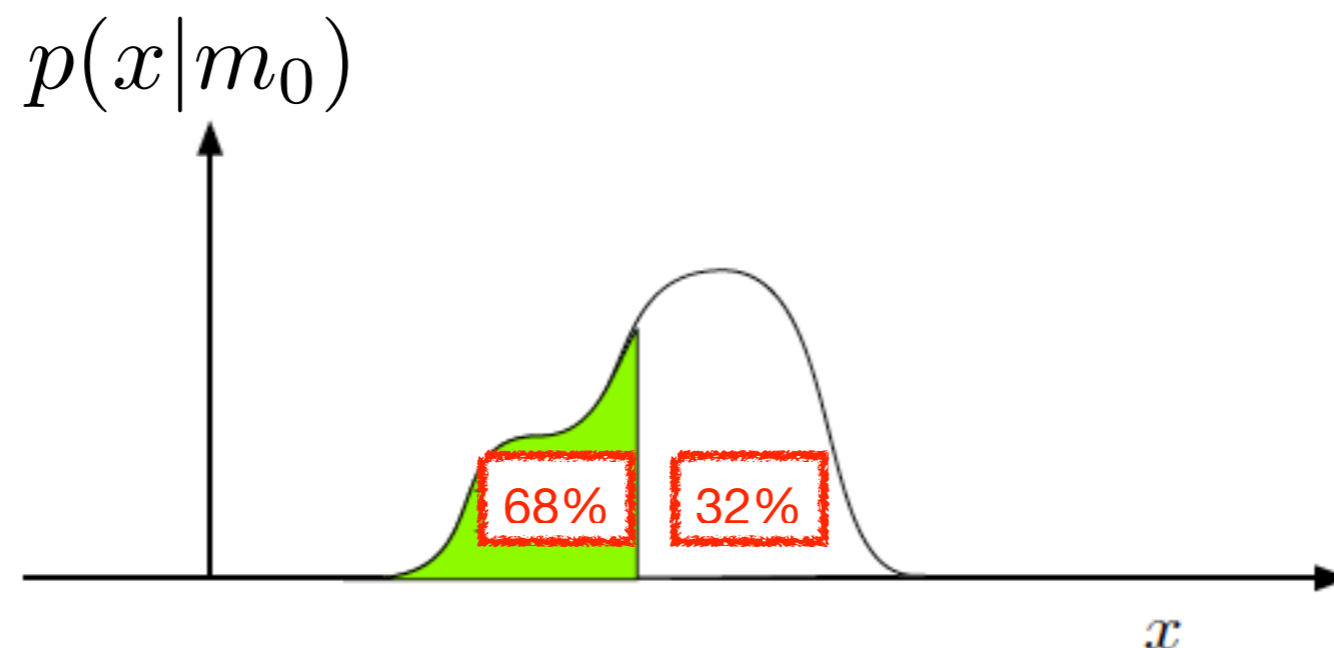


Neyman illustrated III

The definition of the acceptance range is not unique

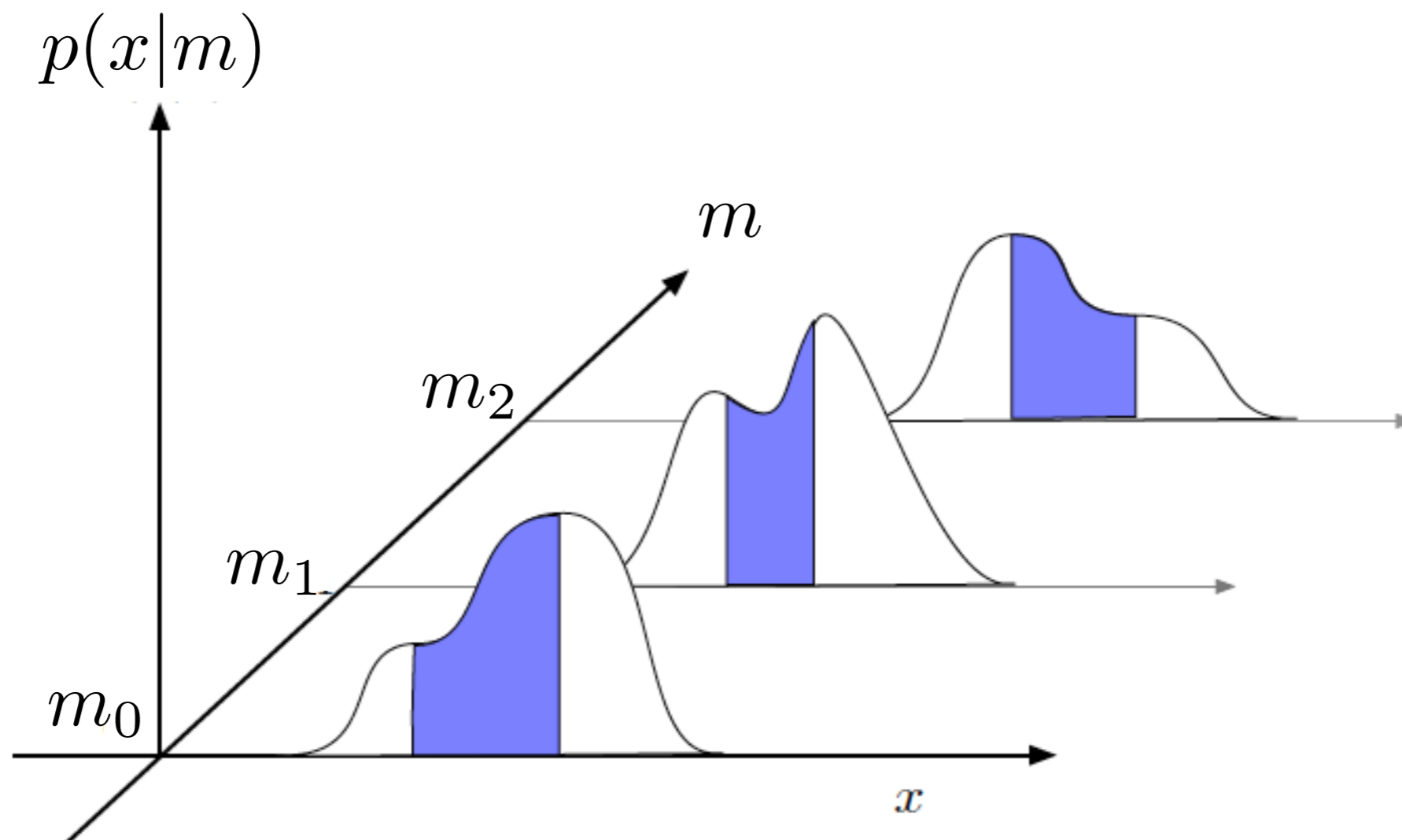
The criterion to choose of the region is chosen is the *ordering rule*

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.



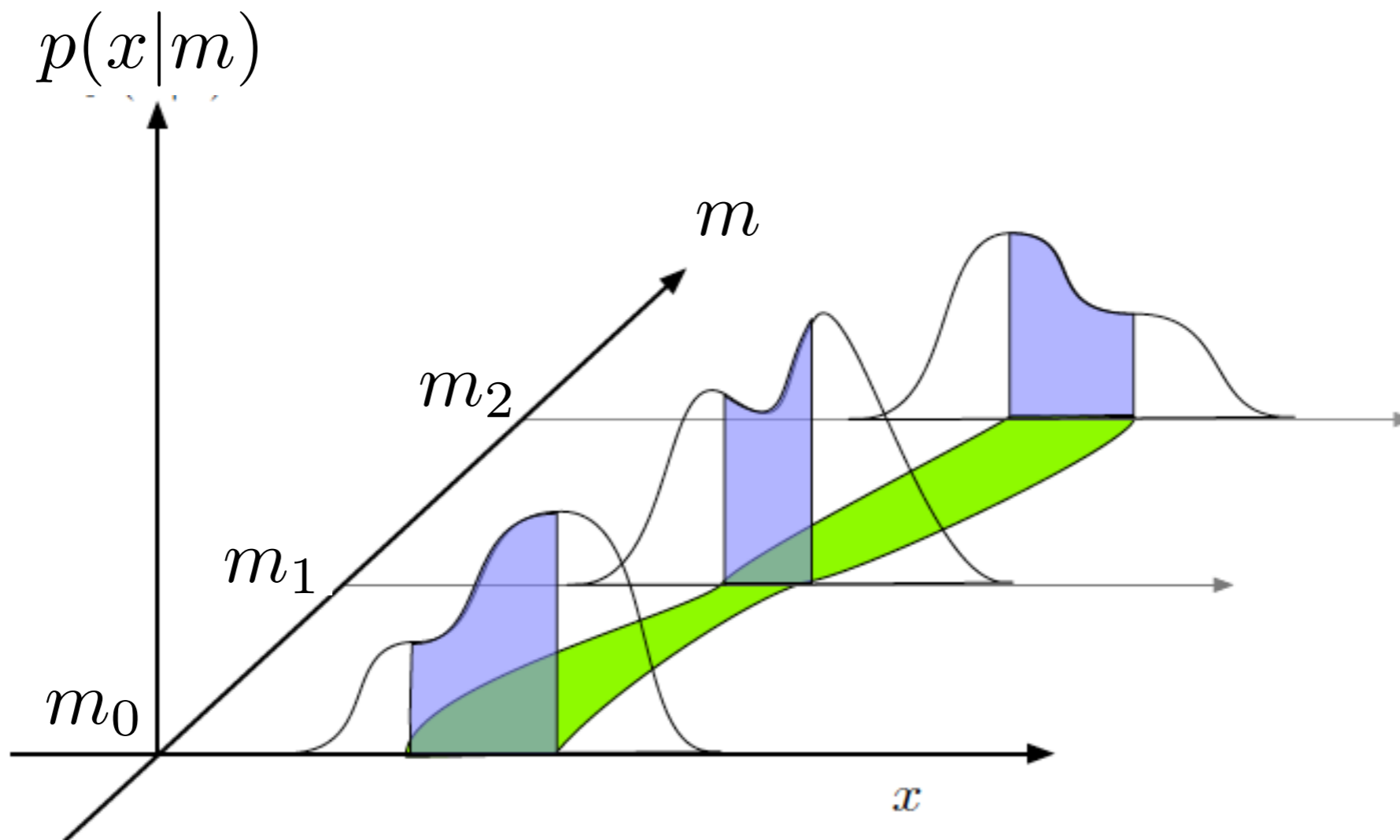
Neyman illustrated V

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



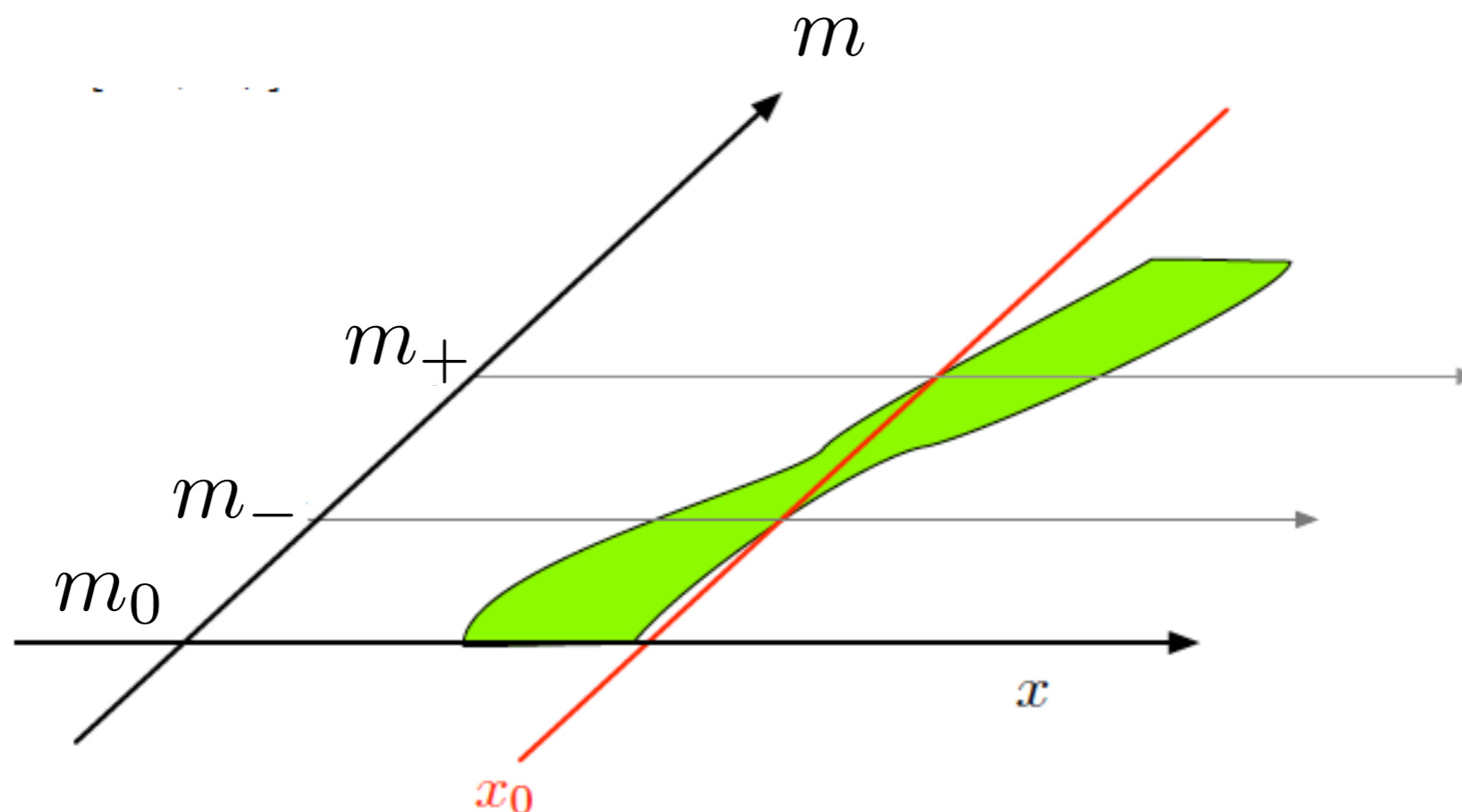
Neyman illustrated VI

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)

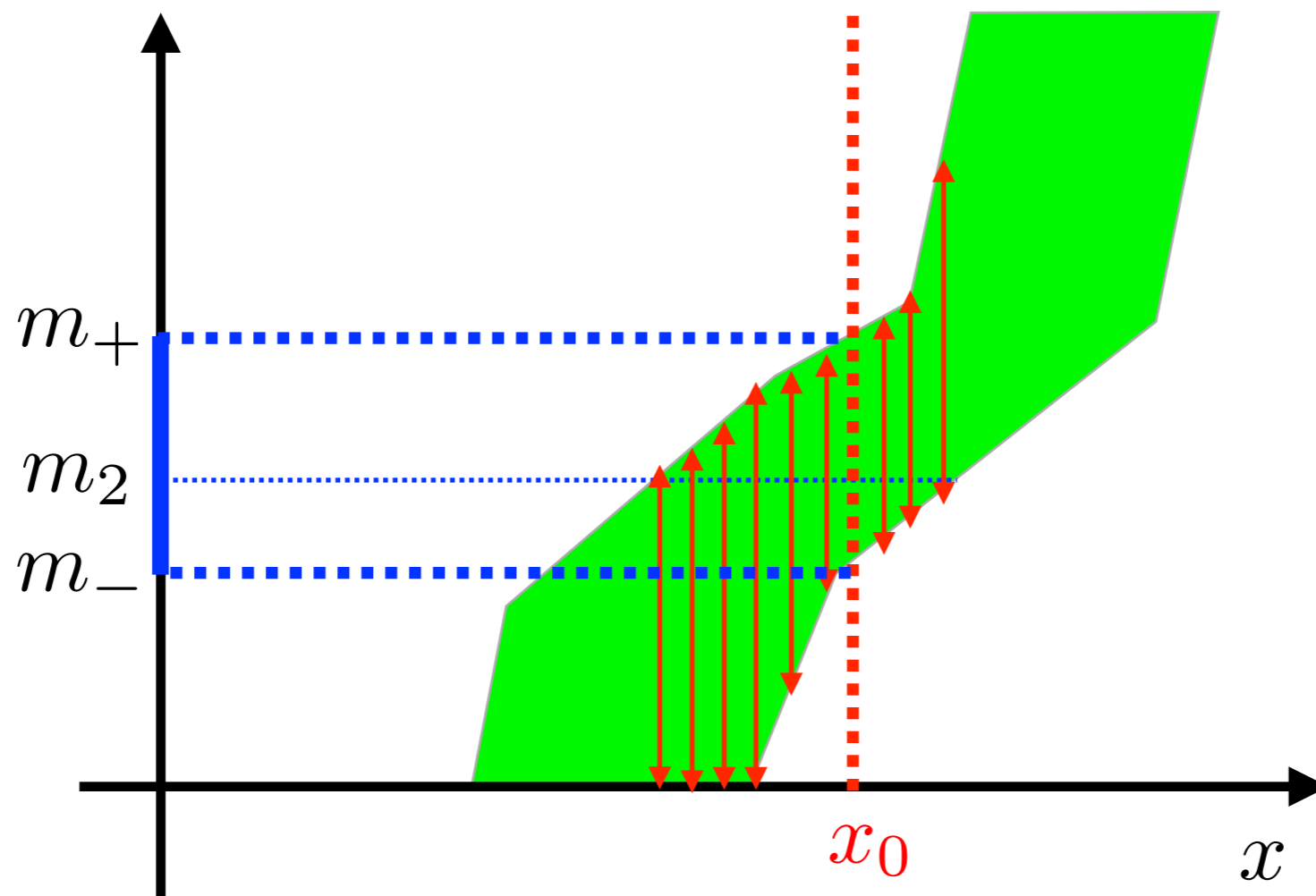


[Cranmer]

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.

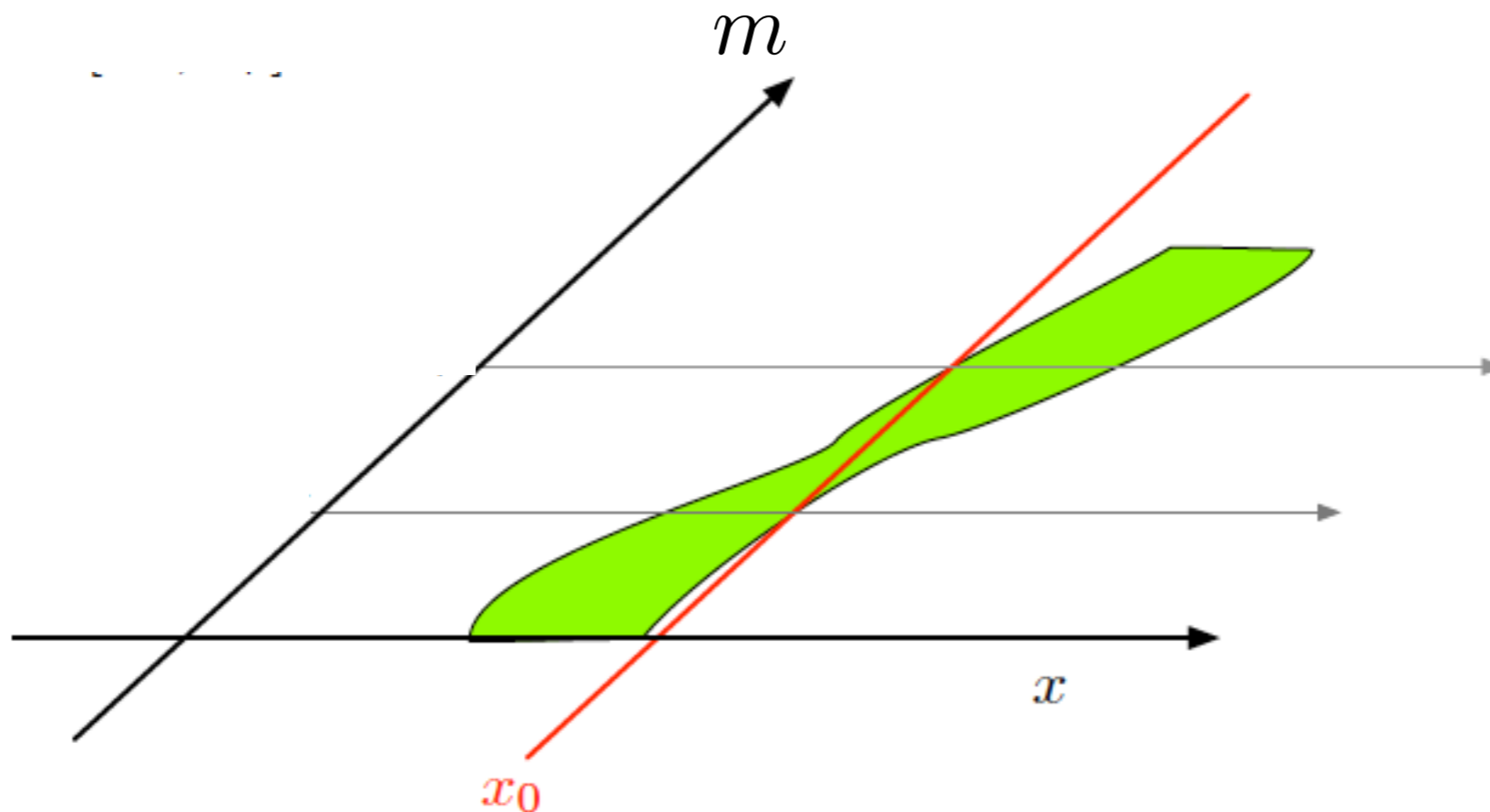
Toy example

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

Additional material

Confidence intervals

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

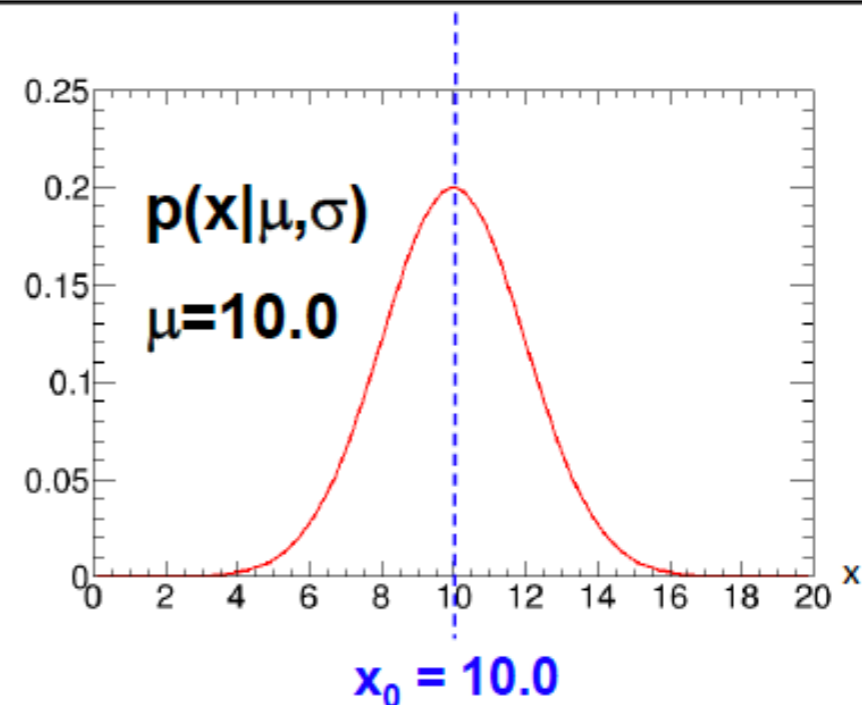
$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

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

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

Suppose $x_0 = 10.0$ is observed.

What can one say about μ ?



Minimum χ^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 σ a function of μ : $\sigma = 0.2 \mu$

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

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

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

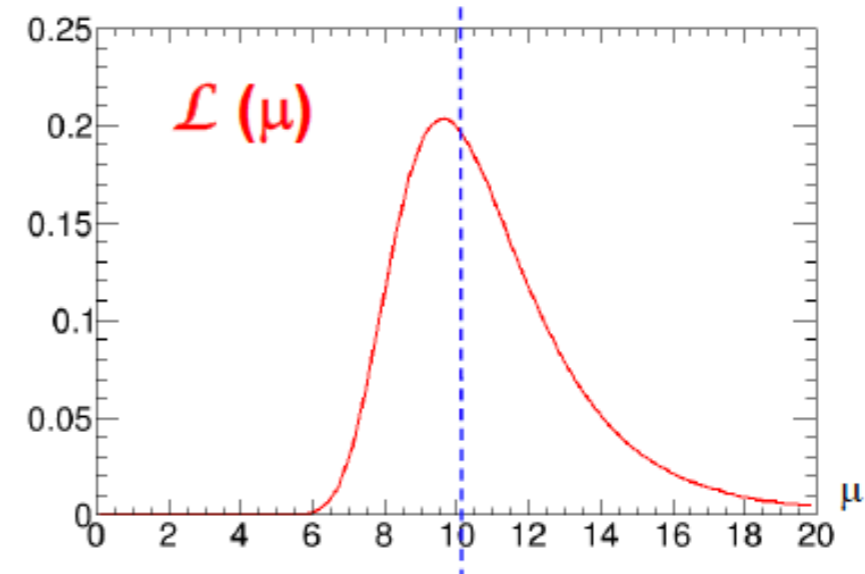
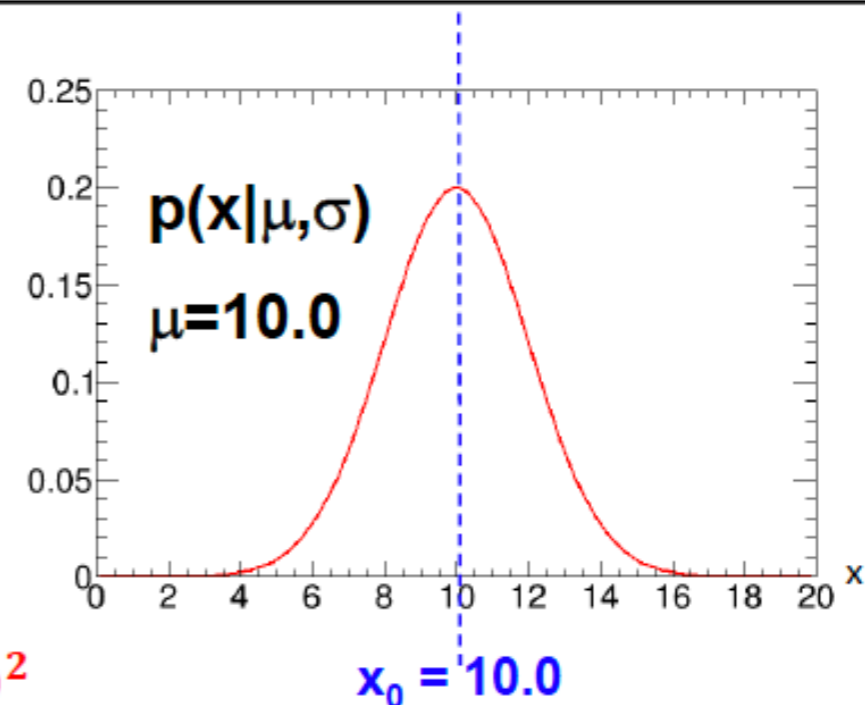
Suppose $x_0 = 10.0$ is observed.

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

$\mathcal{L}(\mu)$ for observed $x_0 = 10.0$:

$$\mu_{ML} = 9.63$$

What is confidence interval for μ ?



Confidence intervals

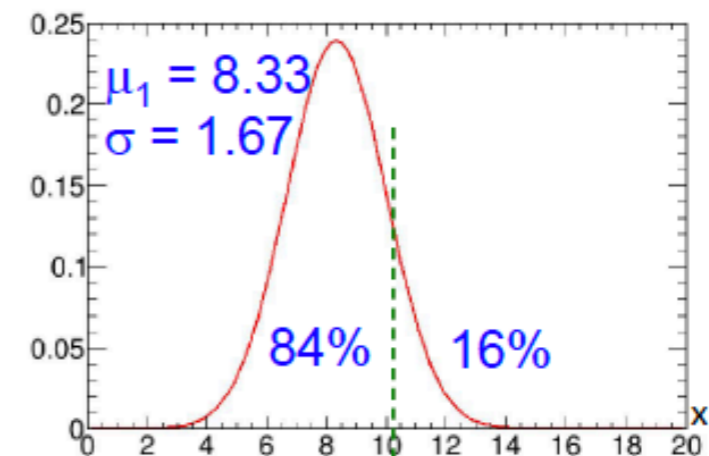
**Gaussian pdf $p(x|\mu,\sigma)$ with σ a function of μ : $\sigma = 0.2 \mu$
Observed $x_0 = 10.0$.**

Find μ_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

μ_1 is 84% C.L. lower limit for μ .

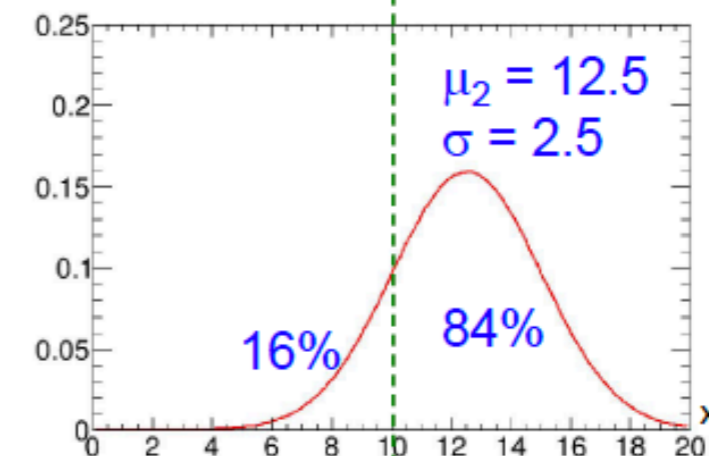


Find μ_2 such that 84% of $p(x|\mu_2,\sigma=0.2\mu_2)$ is above $x_0 = 10.0$; 16% of prob is below.

Solve: $\mu_2 = 12.5$.

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

μ_2 is 84% C.L. upper limit for μ .



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)

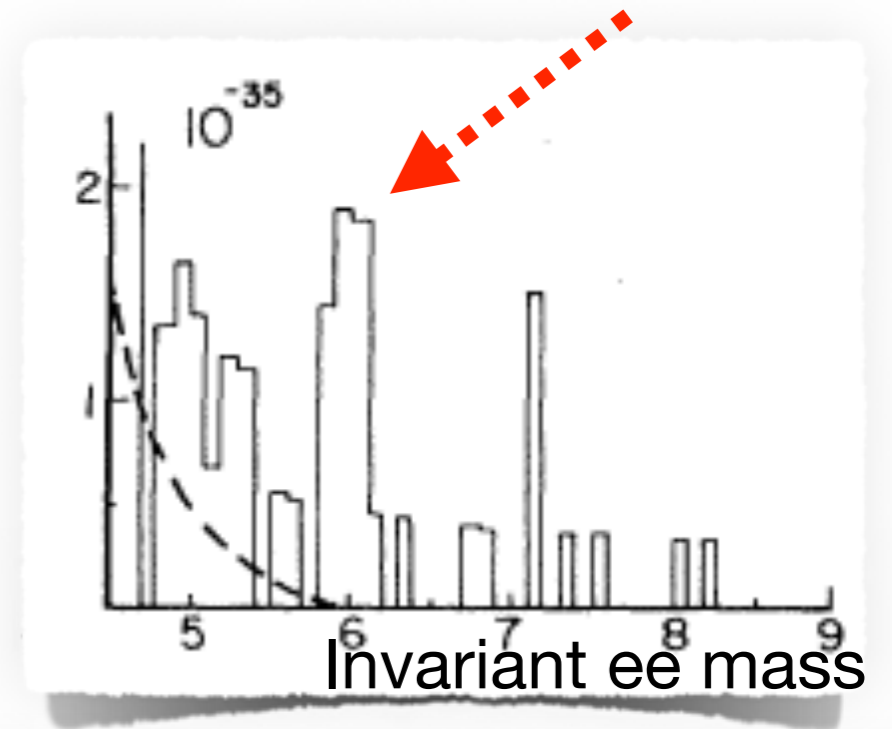
We report preliminary results on the production of electron-positron pairs in the mass range 2.5 to 20 GeV in 400-GeV p -Be interactions. 27 high-mass events are observed in the mass range 5.5–10.0 GeV corresponding to $\sigma = (1.2 \pm 0.5) \times 10^{-35}$ cm² per nucleon. Clustering of 12 of these 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.

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 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,^(a) H. D. Snyder, and J. K. Yoh
Columbia University, New York, New York 10027

and

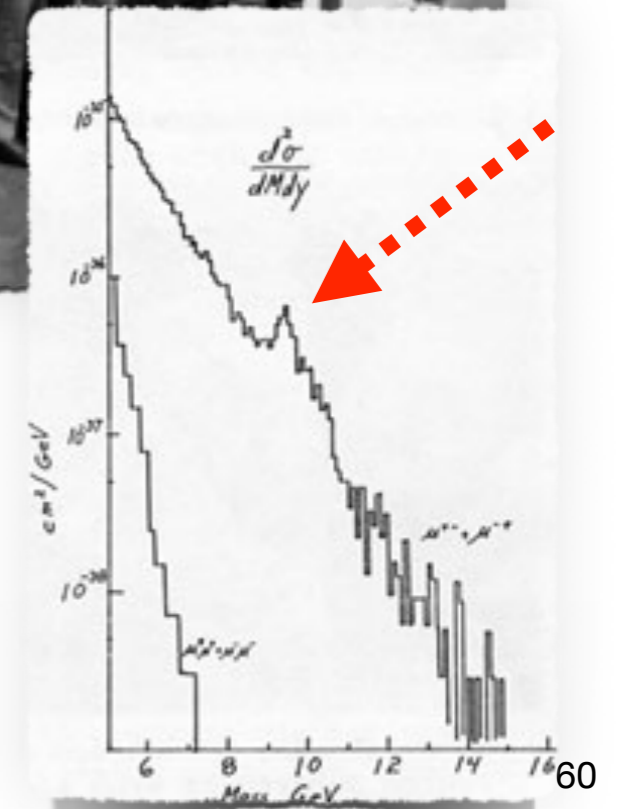
J. A. Appel, B. C. Brown, C. N. Brown, W. R. Innes, K. Ueno, and T. Yamanouchi
Fermi National Accelerator Laboratory, Batavia, Illinois 60510

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.



Where is “elsewhere”?

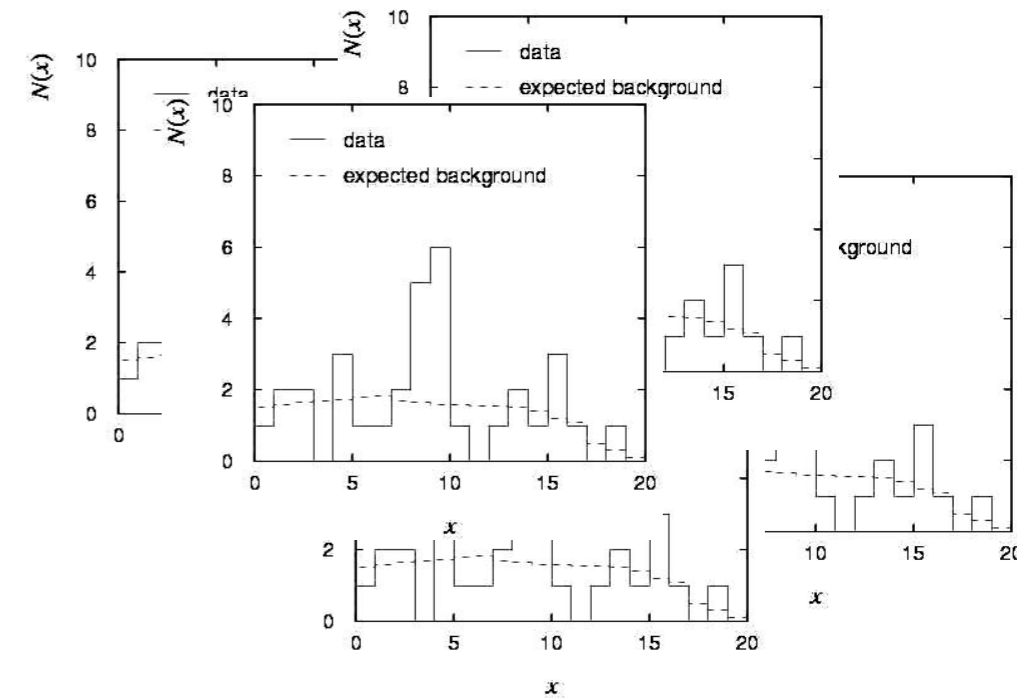
Tenths, or hundreds, or thousands of distributions may have been inspected, in the same analysis or in other analyses.

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?

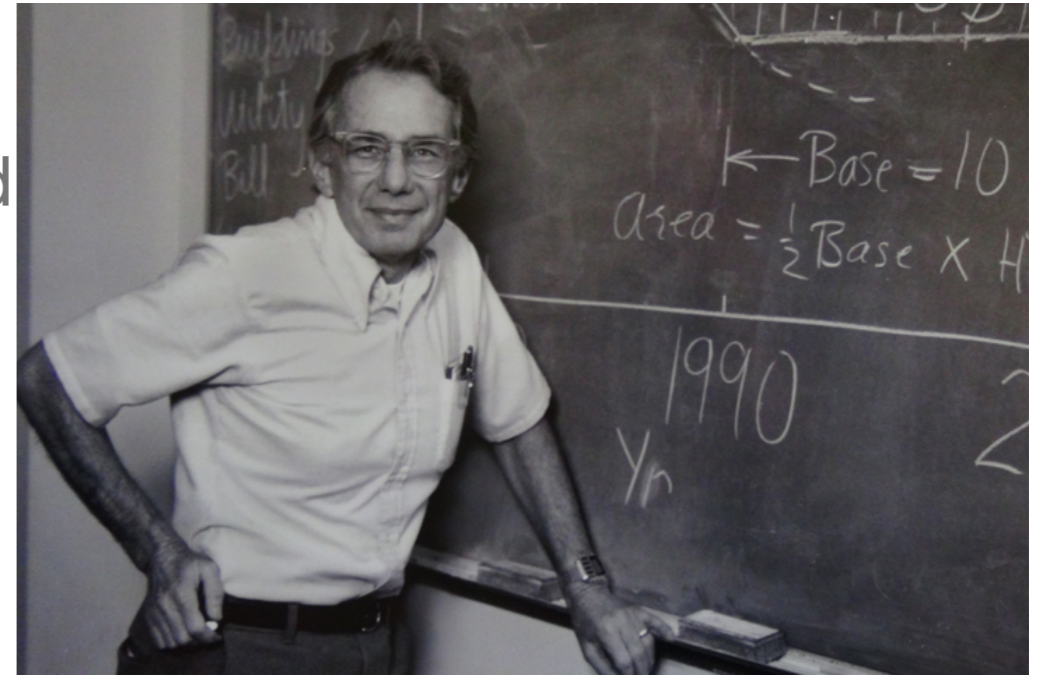
Guidance (consensus at the Banff 2010 Statistics Workshop): limit the testing space to models that are inspected within a single published analysis



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 5σ effects. For the experimental group who has spent a year of their time and perhaps a million dollars, the problem is harder... go ahead and publish... but they should realize that any bump less than about 5σ calls for a repeat of the experiment.”

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

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