

# Statistics for physics — part II and part III

---

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

April 27, 2022

*Future Flavours school 2022*

# The model

---

The **model** is the mathematical structure

$$p(\text{data} \mid \text{physics}) = p(x|m)$$

that incorporates all the physics, knowledge, intuition to best describe the relevant relations between observables  $x$  and unknown parameters  $m$ .

It is a **probability** model — *you don't know exactly what value of  $x$  would be observed even if you knew exactly the value of  $m$ .*

The model is the fundamental building block of most of HEP inference, both in Frequentist and Bayesian procedures. This is the step everyone agrees on.

# Inference

---

The model gives the probability to observe certain data assuming some physics

$p(\text{data} \mid \text{physics})$  is known from the model

That is, the “forward” process, **from physics to data**, which occurs in

- running experiments (physics true but unknown) and
- simulation (physics known but not necessarily true).

The “backward” process, **from data to physics, is inference**: objective quantitative statements on a population when only a sample of observations is available.

Not possible using the certainty of deductive logic. Unobservability of the parent distribution imposes assessments of **probability** (or confidence, or uncertainty)

# Set-theoretical axioms of probability

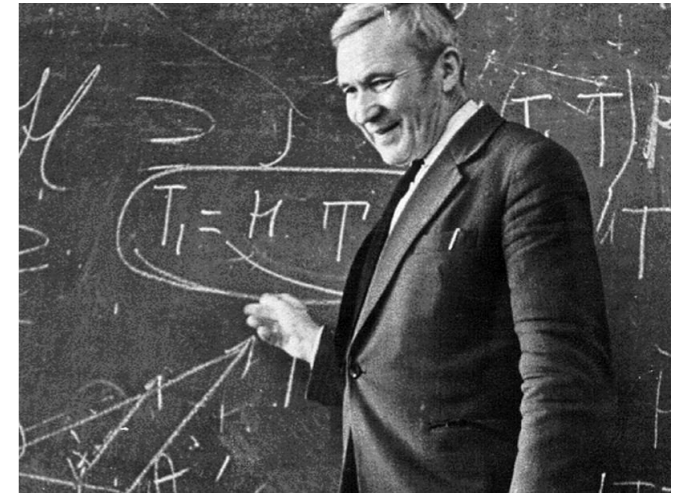
---

Define the set  $\Omega$  of all the possible mutually exclusive outcomes of a statistical experiment (sample space). An event  $A$  is a set containing one or more elementary outcomes.

Assume that probability  $P$  is an additive function on the set and it is measurable on a continuous scale so that it can be represented by a real number. Then

- $P(A)$  is non-negative for each possible outcome  $A$ .
- The probability sum over all possible outcomes (sample space  $\Omega$ ) is unity,  $P(\Omega) = 1$ .
- Probability for observing  $A$  or  $B$  is  $P(A)+P(B)$  if  $A$  and  $B$  are disjoint sets

**Abstract concept. Cannot be measured. Need an operational definition.**



Andrey N. Kolmogorov (1903-1987)

# Operational definitions of probability

---

**Combinatorial** (Laplace): an elementary event  $A$  can occur in  $N$  distinct and equally likely ways;  $n$  of these are favorable cases. Then the probability is defined as  $p(A) = n/N$

**Frequency theory** (Venn, Von Mises): we observe an event  $A$  occurring  $n$  times in  $N$  trials; the probability is defined as the limit for  $N \rightarrow \infty$  of  $p(A) = n/N$ . Used in most scientific work because it is “objective”: is the same for all observers and can be determined (in principle) to any desired accuracy.

**Subjective Bayesian** (Ramsey, De Finetti): the probability for an event  $A$  to occur is the measure of one’s belief for it to occur — akin to our own everyday’s thinking

All of the above heuristic definitions obey the set-theoretical, axioms based on the concept of measure, but they lead to quite differing probability notions, and therefore differing inferences.

# What is probability?

---

**Frequentist** — limiting frequency on independent, identically distributed trials

$$P(A) = \lim_{N \rightarrow \infty} (N_A/N)$$

- Uses information from observed data (and from data that could have been observed in other trials).
- Only applies to repeatable events. Data are random, theories are not.
- Hence, restricts deductions based on **p(data|theory)**: theories for which the set of observed data is more *usual* are *avored*

**Bayesian** — subjective degree of belief

- combines info from observed data with subjective judgment. Same data and different scientists may lead to inconsistent results. May change with time as prior information changes.
- Treating as random variables any unknown broadens the scope of application to include theories and hypotheses.
- Gets to **p(theory | data)**: the inductive reasoning one is interested to.

# In their own words

---

*“The probability of any event is the ratio between the value at which an expectation depending on the happening of an event ought to be computed, and the value of the thing expected upon its happening. [...] By chance I mean the same as probability” — T. Bayes (1763)*

*“Probability theory is nothing but common sense reduced to calculation” — Laplace (1818)*

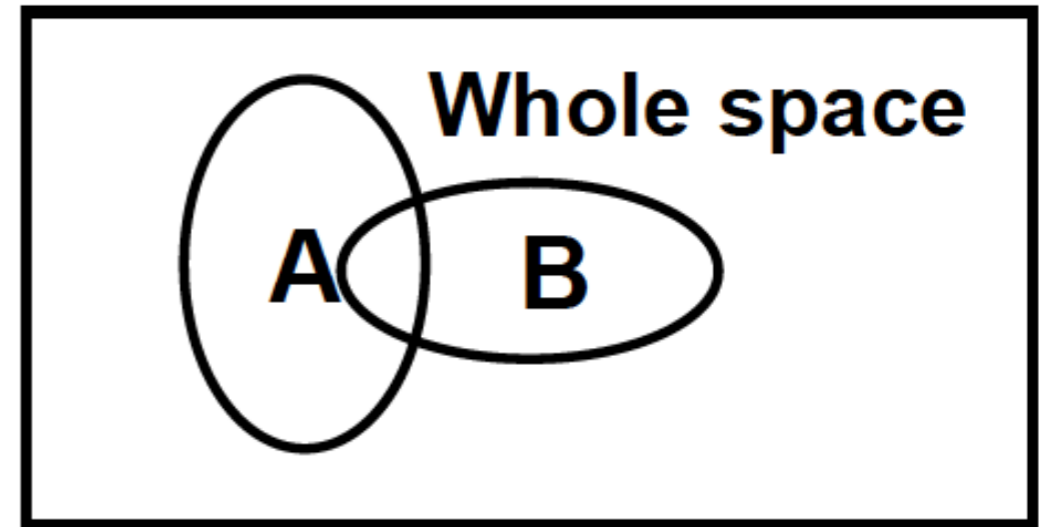
Frequentist use impeccable logic to deal with an issue of no interest to anyone.

Bayesians address the question everyone is interested in, by using assumptions no-one believes

# The sample (or whole) space

---

For probabilities to be well defined, the whole space or sample space need be defined (determines normalization)



*“90% of our flights arrive on time”*

Flight delayed several hours are canceled, not ‘delayed’, so they get excluded from our sample space.

*“Our survey shows that most people lose 5 Kg in a month on this diet”*

Happy customers who lost weight are most likely to respond to our survey. The ones who gained weight most likely threw away our survey postcard.

Whole space can be thought as the space of available possibilities given (i.e., conditional to) the assumptions associated with the model.



# Bayesian inference

# Conditional probabilities

---

$P(A|B)$  is the probability of A given that B has occurred. That is, the probability that A occurs *under the condition* that B has occurred.

We restrict the possible outcomes to the subset B, included in  $\Omega$ . So, B becomes the new sample space. Hence, if the outcome is in A, it is in the intersection of A and B. Because the maximum value of  $P(A \text{ and } B)$  is  $P(B)$ , the conditional probability, which as any probability cannot exceed 1, should meet

$$P(A|B) = P(A \text{ and } B) / P(B)$$

$$P(A \text{ and } B) = P(A | B) * P(B)$$

Probability  
that A and B  
occur jointly

= (Conditional) probability  
that A occurs given that B  
occurred X

(Marginal) probability  
that B occurs

$$\text{Analogously } P(A \text{ and } B) = P(B | A) P(A)$$

# Conditional probabilities

---

Probability for jointly observing A and B

$$P(A \text{ and } B) = \begin{cases} P(A|B) * P(B) \\ P(B|A) * P(A) \end{cases}$$

(Conditional probability for A given B)      (Marginal probability for B)

(Conditional probability for B given A)      (Marginal probability for A)

# Bayes' theorem

---

Yields a key relation between conditional and marginal probabilities.



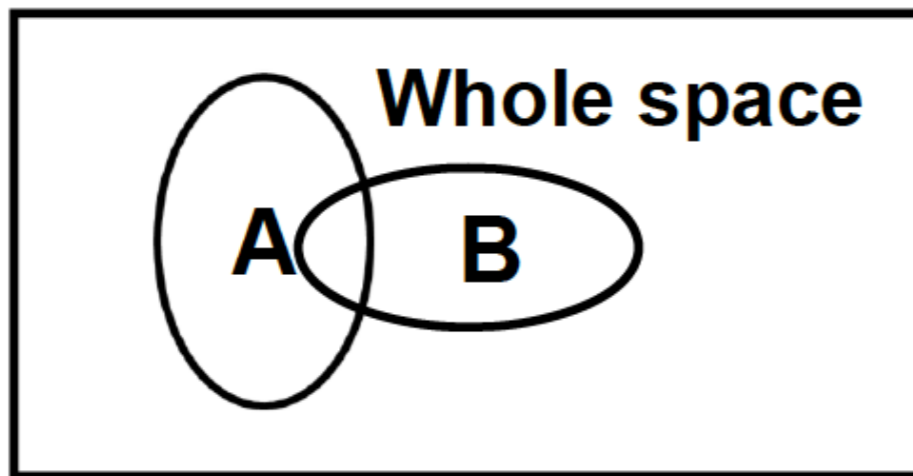
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\text{not}B)P(\text{not}B)}$$

T. Bayes  
(1702-1761)

- $P(B|A)$  is the conditional probability for B given A. Also called **posterior** because evaluated *after* fixing a specific value of A
- $P(A|B)$  is the conditional probability of A given B
- $P(B)$  is the **prior** probability for B, evaluated *before* knowing any information on A
- $P(A)$  is the marginal (or “prior”) probability for event A. Serves as normalization.

# In pictures

## P, Conditional P, and Derivation of Bayes' Theorem in Pictures



$$P(A) = \frac{\text{Area of } A}{\text{Area of Whole space}}$$

$$P(B) = \frac{\text{Area of } B}{\text{Area of Whole space}}$$

$$P(A|B) = \frac{\text{Area of } A \cap B}{\text{Area of } B}$$

$$P(B|A) = \frac{\text{Area of } A \cap B}{\text{Area of } A}$$

$$P(A \cap B) = \frac{\text{Area of } A \cap B}{\text{Area of Whole space}}$$

$$P(A) \times P(B|A) = \frac{\text{Area of } A}{\text{Area of Whole space}} \times \frac{\text{Area of } A \cap B}{\text{Area of } A} = \frac{\text{Area of } A \cap B}{\text{Area of Whole space}} = P(A \cap B)$$

$$P(B) \times P(A|B) = \frac{\text{Area of } B}{\text{Area of Whole space}} \times \frac{\text{Area of } A \cap B}{\text{Area of } B} = \frac{\text{Area of } A \cap B}{\text{Area of Whole space}} = P(A \cap B)$$

$$\Rightarrow P(B|A) = P(A|B) \times P(B) / P(A)$$

# Remember !!!

---

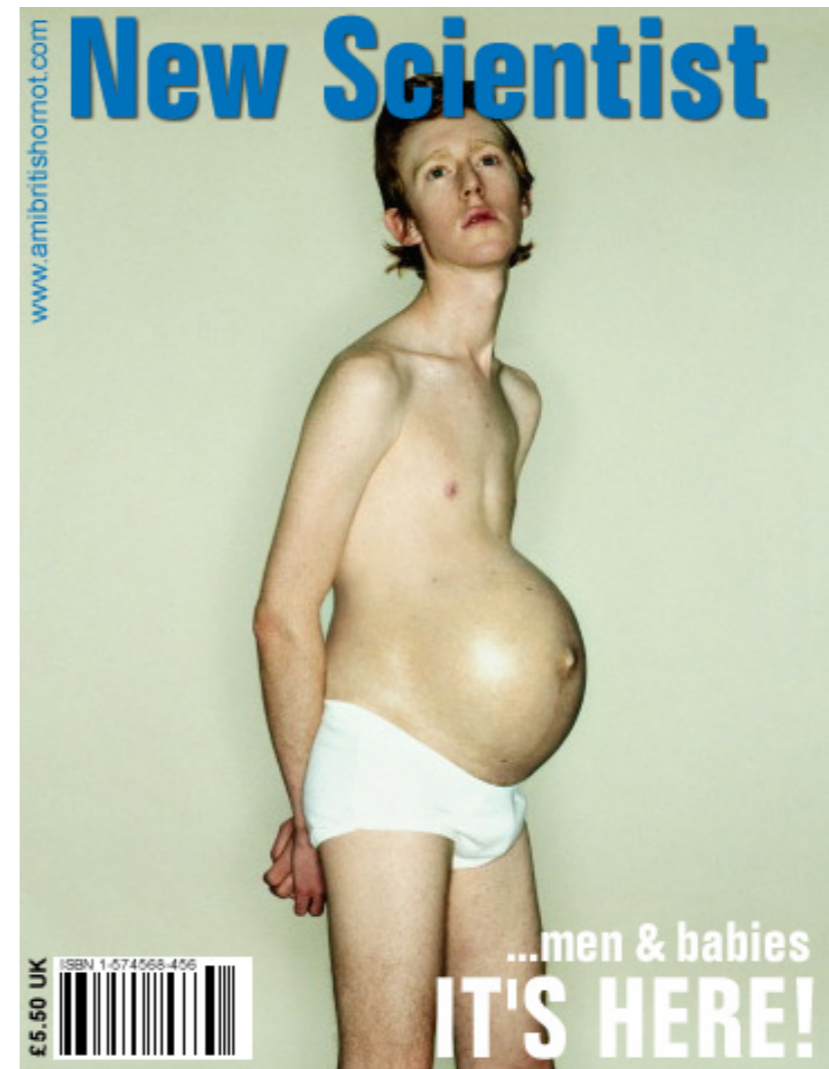
$P(A|B)$  is NOT equal to  $P(B|A)$ .

Variable A: “pregnant”, “not pregnant”

Variable B: “male”, “female”.

$P(\text{pregnant} | \text{female}) \sim 3\%$  but

$P(\text{female} | \text{pregnant}) \gg 3\%$  !



# Using Bayes theorem for inference

---

- Suppose both  $x$  and  $m$  are random variables, with known probability distribution  $p(x|m)$ .  $x$  is observable and  $m$  inobservable.
- Observe  $x$  (“perform a measurement of  $x$ ”), what can I say about  $m$ ? I basically wanna know  $p(m|x)$ .
- Bayes’ theorem tells me all I possibly need. It allows determining the “a posteriori” probability for any value of  $m$

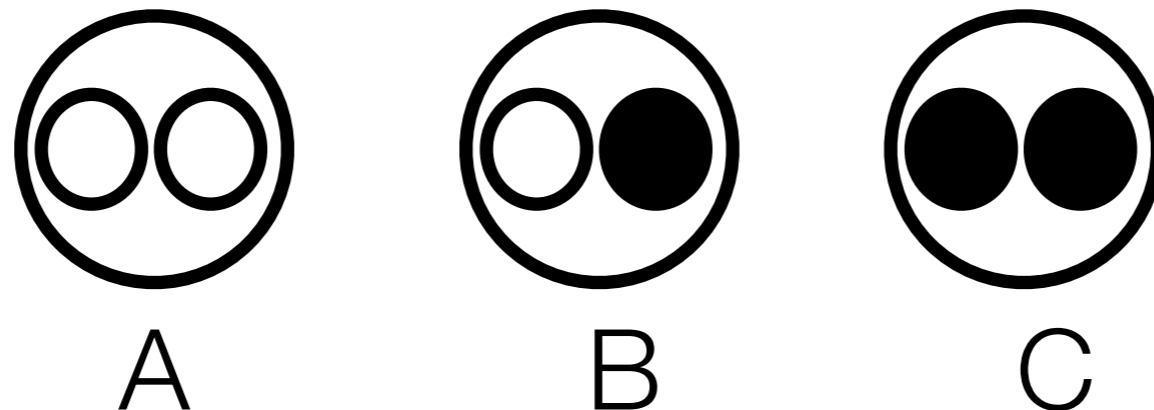
$$P(A | B) = \frac{P(B | A)P(A)}{P(B)} \Rightarrow p(m | x) = \frac{p(x | m)p(m)}{p(x)} = \frac{L(m)p(m)}{p(x)}$$

- posterior probability for the parameter  $m$  = Likelihood \* prior/normalization

# Inference — elementary example

---

- Three identical bags with two balls each, which can be black or white

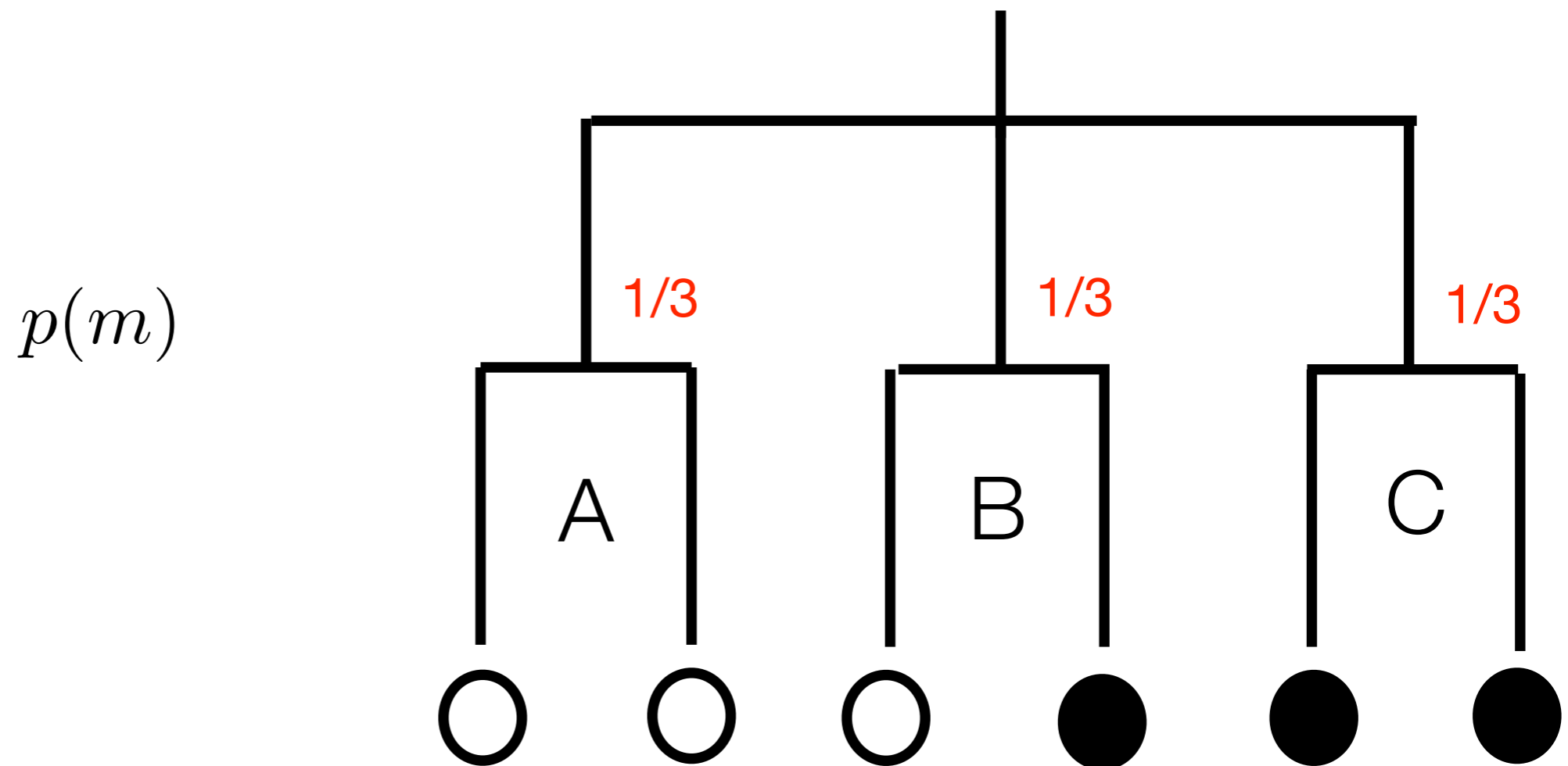


- Pick a random bag ( $m$ , **unobservable**) and a random ball inside it ( $x$ , **observable**)
- Ball is white ( $x = w$ ); what can one say about the chosen bag?

Want to know  $p(m|w)$ , the probability I picked each bag, given that the ball is white



# Inference — elementary example



$p(m)$

$1/3$

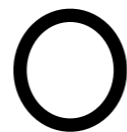
$1/3$

$1/3$

A

B

C



1

$1/2$

0

$1/3$

$1/6$

0

$2/3$

$1/3$

0

$$p(w|m)$$

$$p(w, m) = p(w|m) p(m)$$

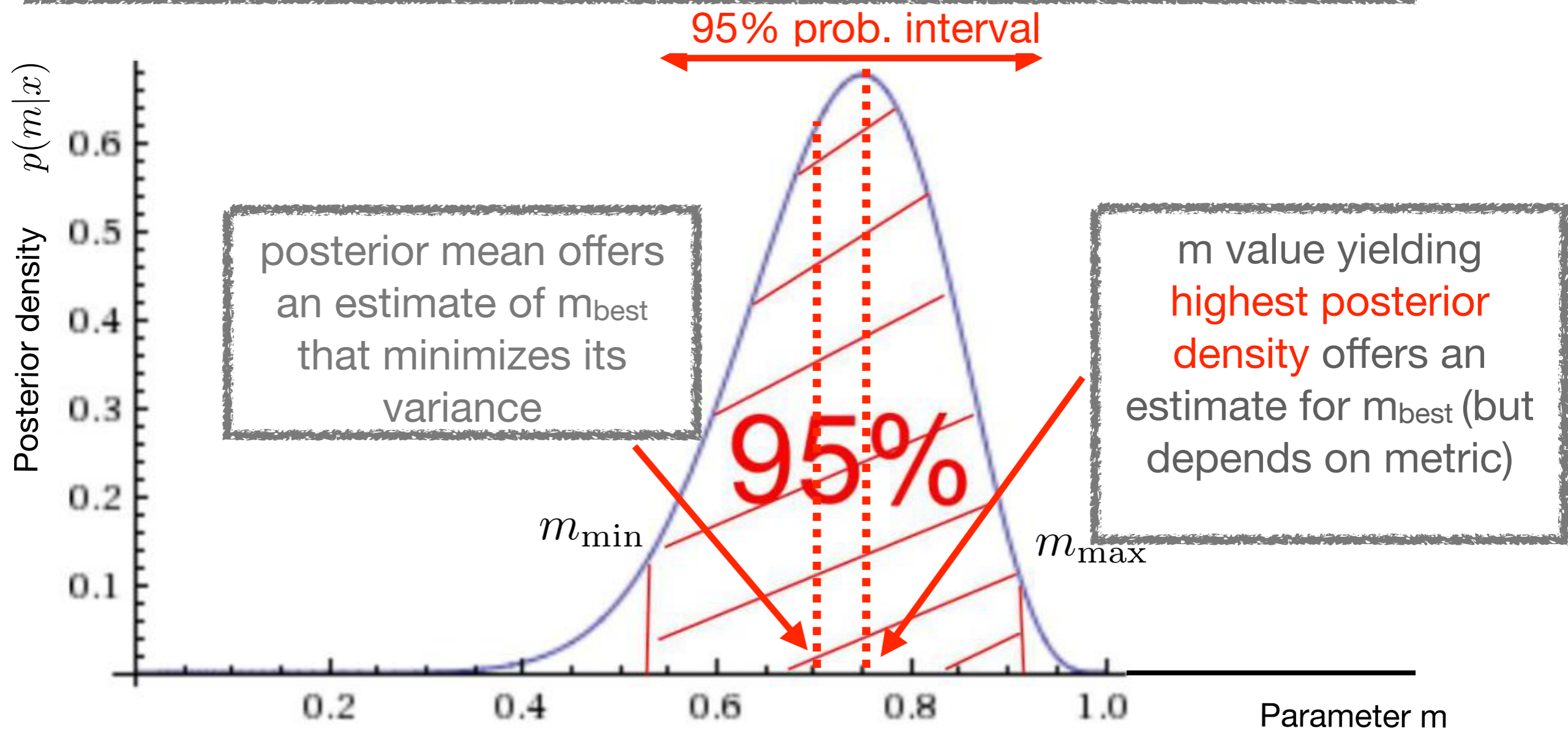
$$p(m|w) = p(w|m) p(m) / p(w)$$

The posterior probabilities are  $p(A|w) = 66\%$ ,  $p(B|w) = 33\%$ ,  $p(C|w) = 0$ .

# Bayesian inference — posterior density says it all

The posterior probability density  $p(m|x)$  is a function of  $m$  that provides any inference we desire:

(Not unique) **interval** of  $m$  values such that  $\int_{m_{\min}}^{m_{\max}} p(m|x) dm = \alpha$  (e.g.,  $\alpha = 95\%$ )



# Frequentists 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 **prior probabilities for  $m$**  (prior to the observation of  $x$ ) allows to use the observation  $x$  to update the prior knowledge and therefore determine the **posterior probability density  $p(m|x)$** .

**That is the “backward process” probability** - which offers all information one might possibly want on  $m$

# What if priors aren't known (or cannot be defined) ?

---

In examples seen so far, both  $x$  and  $m$  were random variables.

But what happens if “ $m$ ” is not a random variable (as in frequentist view)?

Application of the Bayes' theorem to hypotheses or theories “ $m$ ”, to which associating probabilities is nontrivial is less straightforward.

This is where the two schools part.

- ❑ Frequentist: give up on getting  $p(m|x)$ . Revert to an estimate based only on data and the assumed model  $p(x|m)$ , not on prior knowledge.
- ❑ Bayesian: stick to Bayes' theorem by assuming priors. Because physicists typically expect objective/reproducible results free from subjective input, many Bayesian analyses in HEP strive to use priors that have minimal influence on the result.

# Uniform 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 (in one dimension)

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 - sensitivity studies

---

Support Bayesian results through sensitivity studies: investigate the sensitivity of one's analysis on prior choices by, e.g., repeating the analysis with various choices for priors, or on smaller subsets of the sample.

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)

Sensitivity analysis provides reliable 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.

# What if I don't look at theory $m$ as a random variable?

---

What can I compute without priors?

Not  $p(m|x)$ . That is, not  $p(\text{theory given data})$ .

Use  $p(x|m)$  to favor/disfavor a certain theory based on how likely that theory is to generate the data I observe.

*The likelihood*



# The likelihood function $L(m) = p(x|m)$

---

- With parameter  $m$  fixed at a specific value  $m'$ , the model  $p(x|m')$  is the probability density function of observing generic data  $x$ ,
- With data  $x$  fixed at the specific set that was observed  $x_0$ , the model  $L(m) = p(x_0|m)$  is the likelihood function of the  $m$  parameters given your data

The likelihood is not a probability. But it is connected to the *probability for observing data  $x$*  for each choice of parameter  $m$ . This is **not** to the probability that  $m$  has some value given the data. (remember, hypotheses or theories or physical constants are not random variables)

The likelihood is a complete summary of the data relevant to the estimate at hand. Ideally should be published as is (hopefully we'll get there in HEP).

The likelihood (that is, the model) is the single strongest driver of inference performance: improving the model is the best way of improving the inference.

# A likelihood is NOT a pdf

---

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

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  $m$ ; it obeys no specific normalization.

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

In addition, the likelihood maximum  $L(\hat{m})$  is invariant under reparametrization of  $m$  into  $f(m)$ .

If  $\hat{m}$  is a MLE of  $m$ , then  $f(\hat{m})$  is a MLE of  $f(m)$ .

No Jacobians here, reinforcing the notion that  $L(m)$  is not a pdf for  $m$ .

# What does the likelihood mean?

---

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

Given the observed set of data  $x_0$ ,

- parameter values  $m_{\text{low}}$  that decrease  $L(m|x_0)$  are disfavored: it would be unlikely for nature to generate that set of observed data, had the true value of  $m$  been  $m_{\text{low}}$ .
- Conversely, values  $m_{\text{high}}$  that increase  $L(m|x_0)$  are favored

The value of  $m$  that maximizes the likelihood is **not** the “most likely value of  $m$ ”, It is the value of  $m$  that makes your data most likely.

Physics usually deals with repeated observations  $x$  that are independent and identically distributed.

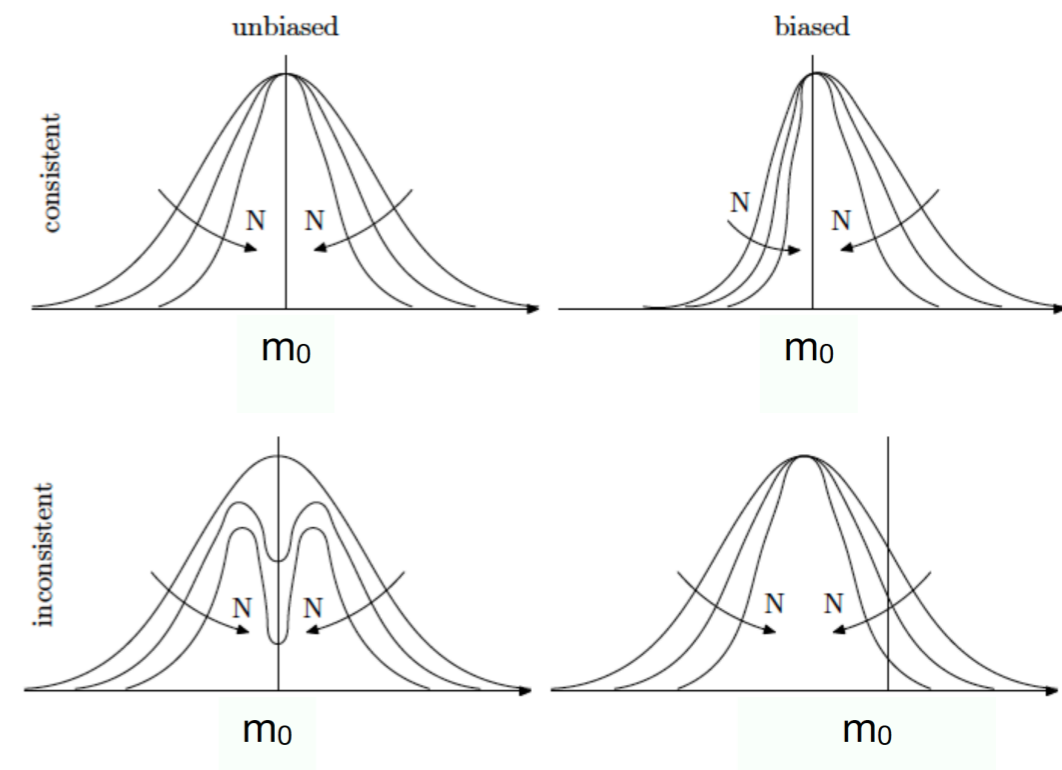
If the likelihood for a single observation  $x$  is  $L(m) = p(x|m)$ , the likelihood for the whole experiment is the product of the single-event likelihoods  $L(m) = \prod p(x|m)$

# Why do we insist on the likelihood concept?

Because likelihood-based estimates have desirable statistical properties — the go-to solution in most of the inferences you will ever do.

The maximum likelihood estimator under pretty weak conditions and asymptotically (for infinite observations  $N$ ...)

- is consistent (converges in probability to true value)
- is efficient (has minimal variance among all estimators)
- has a known distribution (Gaussian)
- is reparametrization-invariant



However, it's not perfect either: it's biased and does not provide goodness of fit information. Plus, in many practical cases the observed data are insufficient to consider the regime “asymptotic” and the above properties are not met.

# Example — exponential

---

Decay process. Assume exponential model. Pdf

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

Probability density of survival after time  $t$

Observe a set of  $N$  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  $k$ th event at  $t = t_k$

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

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 to maximize

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

$$dL(\tau)/d\tau = 0 \text{ implies } \hat{\tau} = \sum_{k=1}^N t_k / N$$

the value 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

---

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  $j$

Likelihood

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

(Continuous) function of physics parameter  $\mu$

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

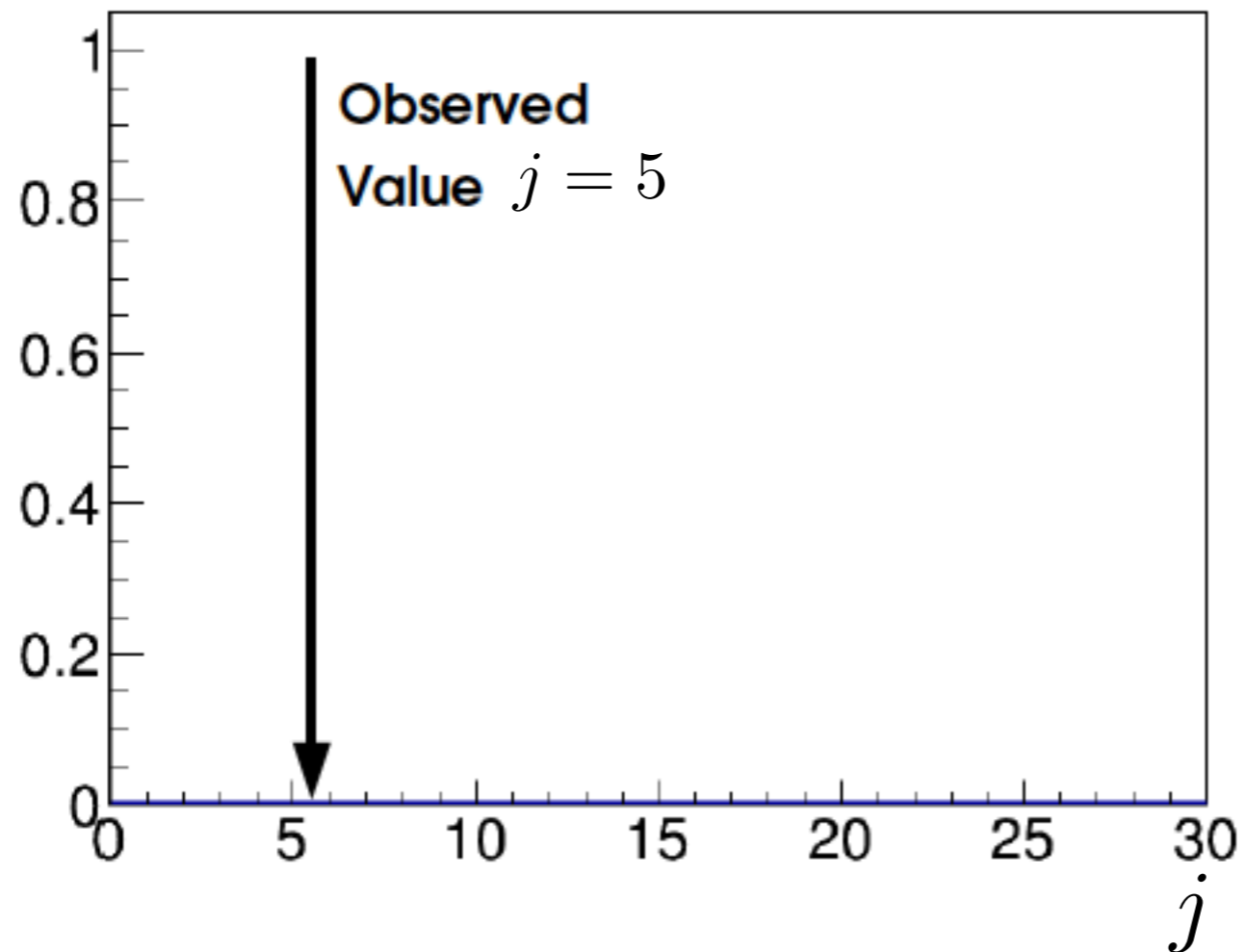
Given observation  $j$ , the ML estimator of the mean rate of success  $\mu$  is  $\hat{\mu} = j$

# Poisson illustrated

---

Model: Poisson-distributed signal, no background.

Observe  $j = 5$ .



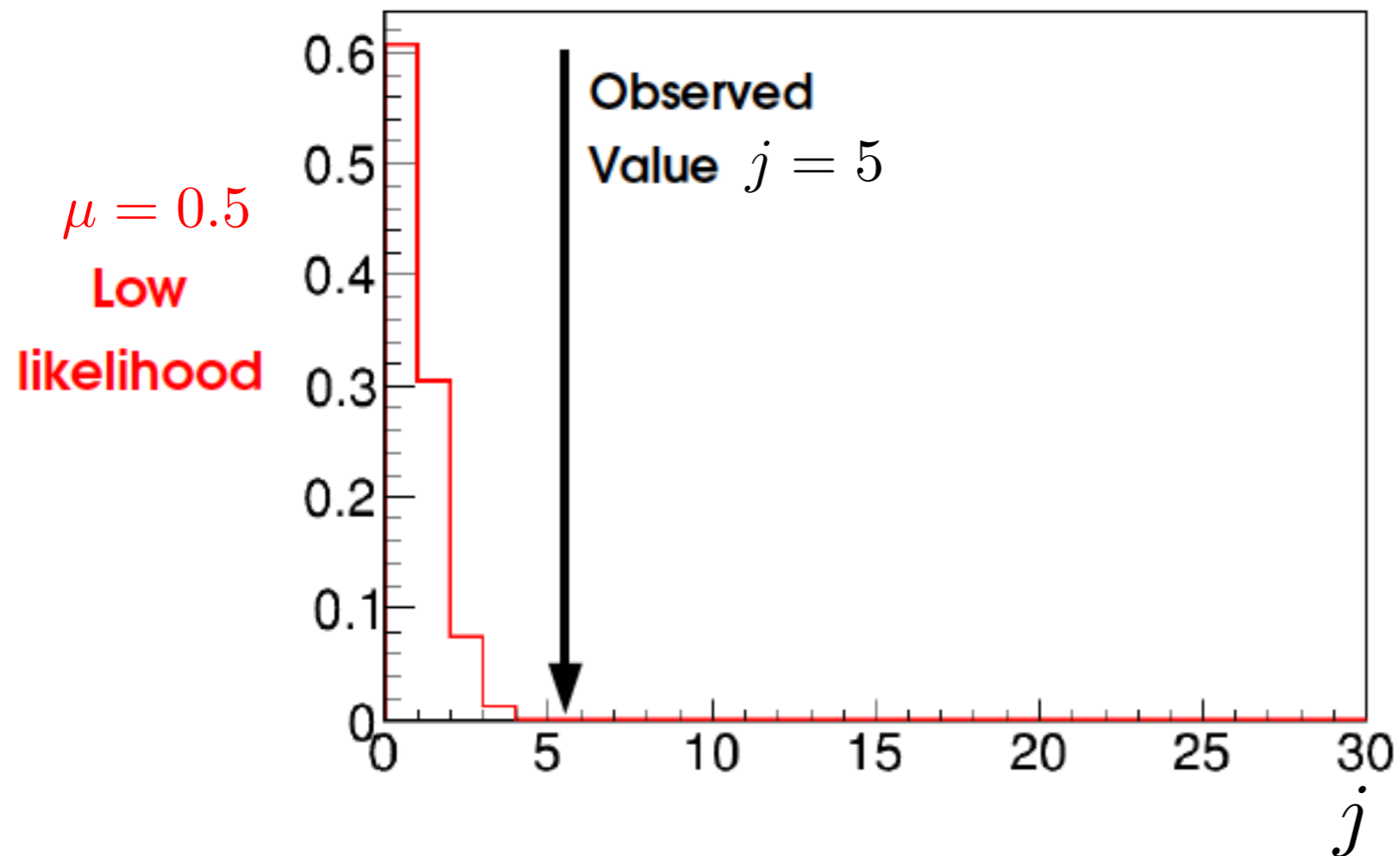


# Poisson illustrated

---

Model: Poisson-distributed signal, no background.

Observe  $j = 5$ .

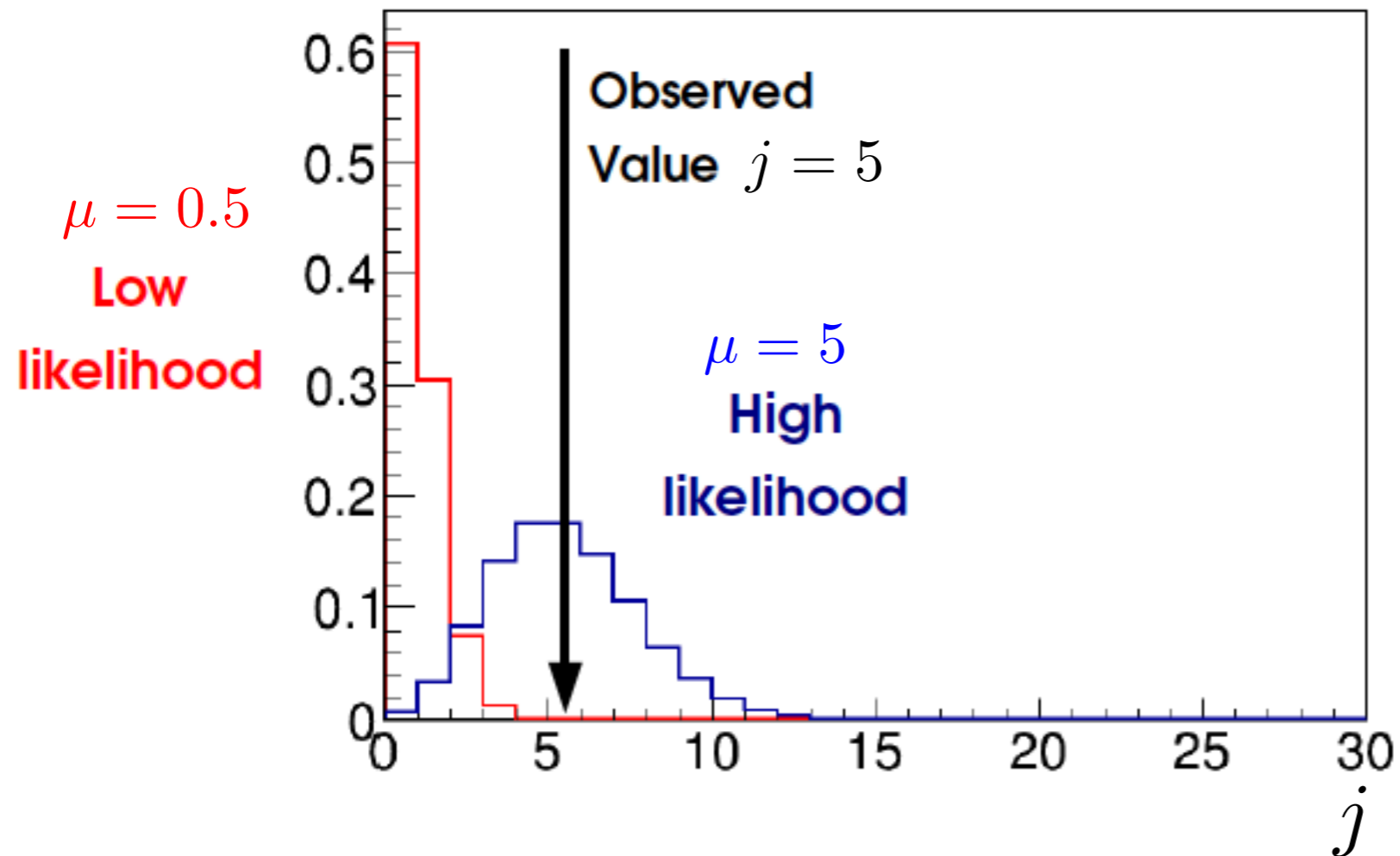


# Poisson illustrated

---

Model: Poisson-distributed signal, no background.

Observe  $j = 5$ .

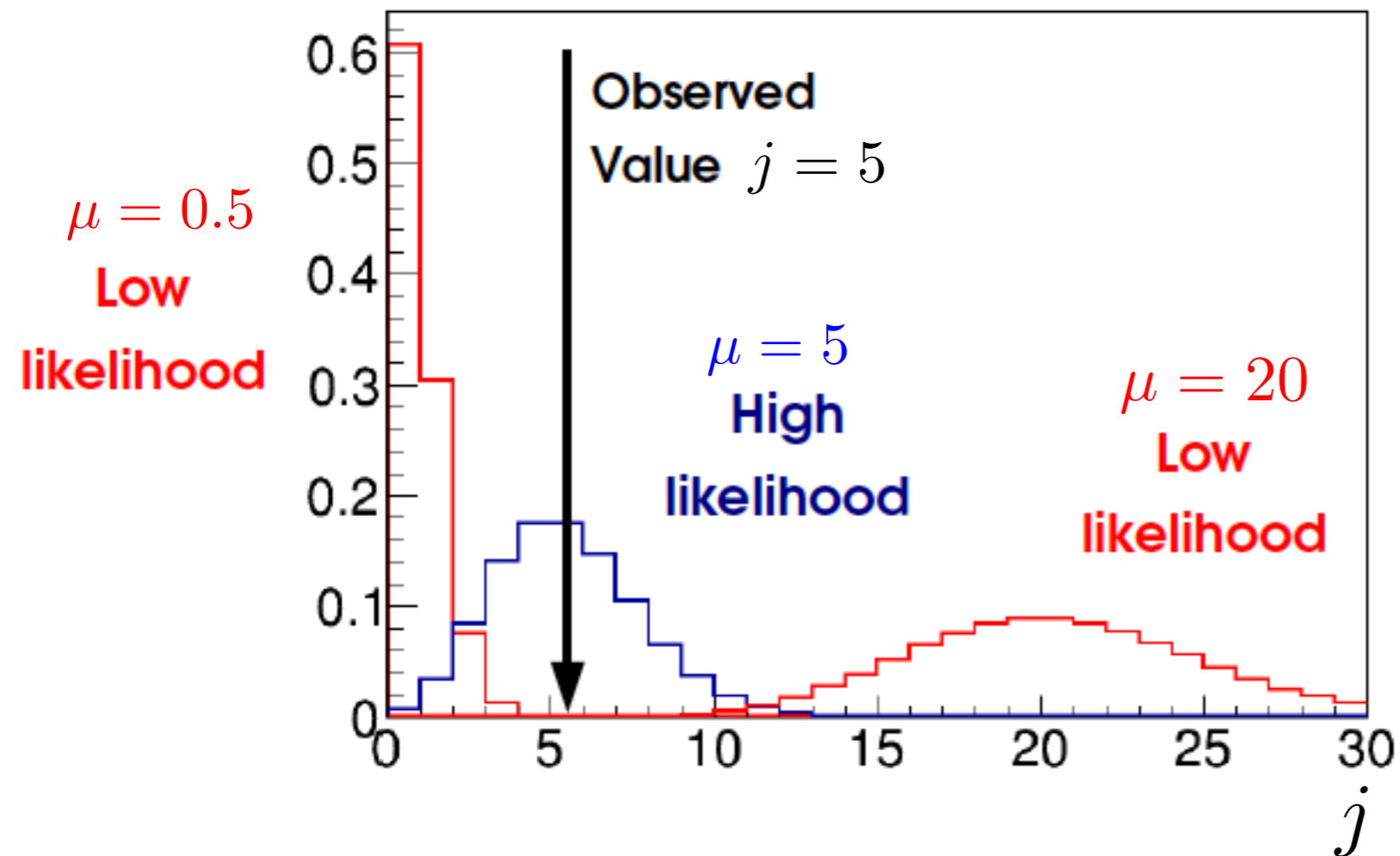


# Poisson illustrated

---

Model: Poisson-distributed signal, no background.

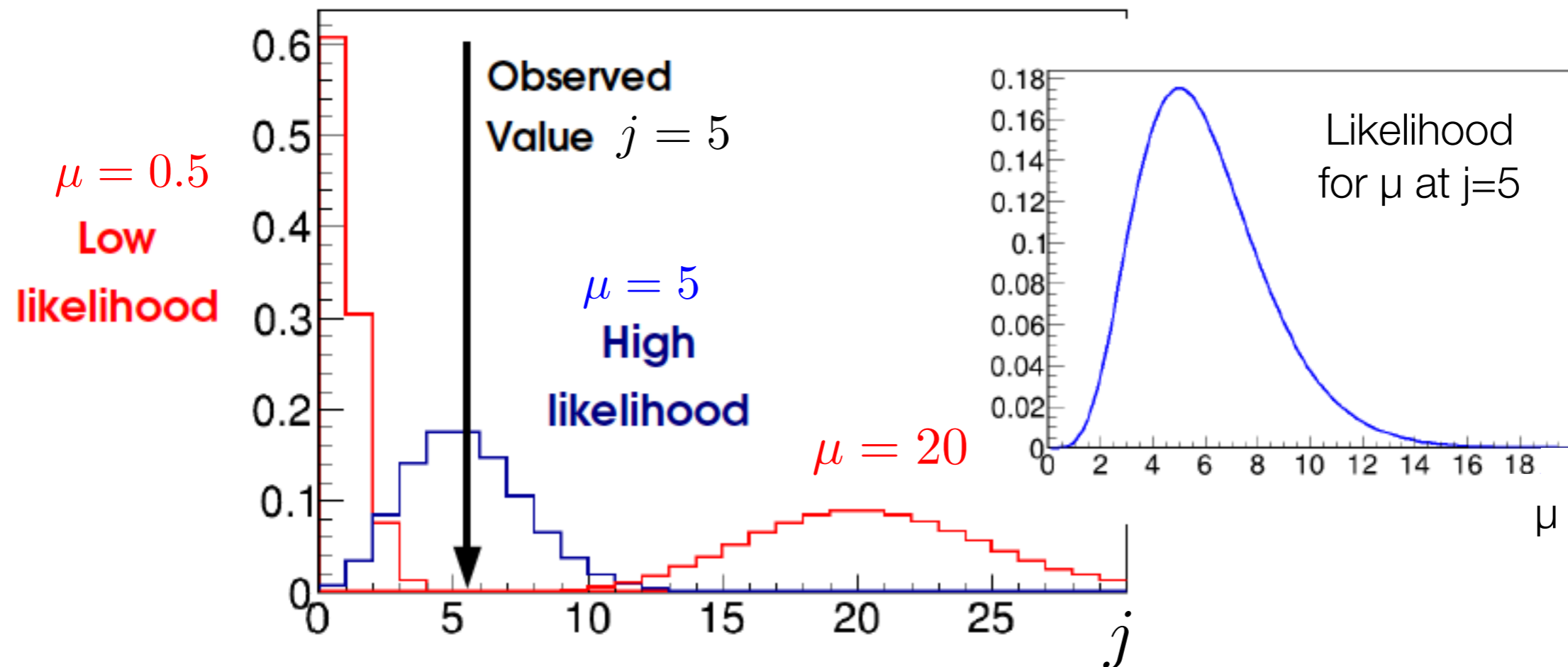
Observe  $j = 5$ .



# Poisson illustrated

Model: Poisson-distributed signal, no background.

Observe  $j = 5$ .



# What about the (statistical) uncertainty?

Given observations  $x_0$ , and assuming  $L(m) = p(x_0|m)$ , the value  $\hat{m}$  that maximizes  $L$  offers an estimate (“central” or “best fit”) of the true value of  $m$ .

How about the uncertainty? Depends on the estimator’s variance — spread of results in repeated experiments

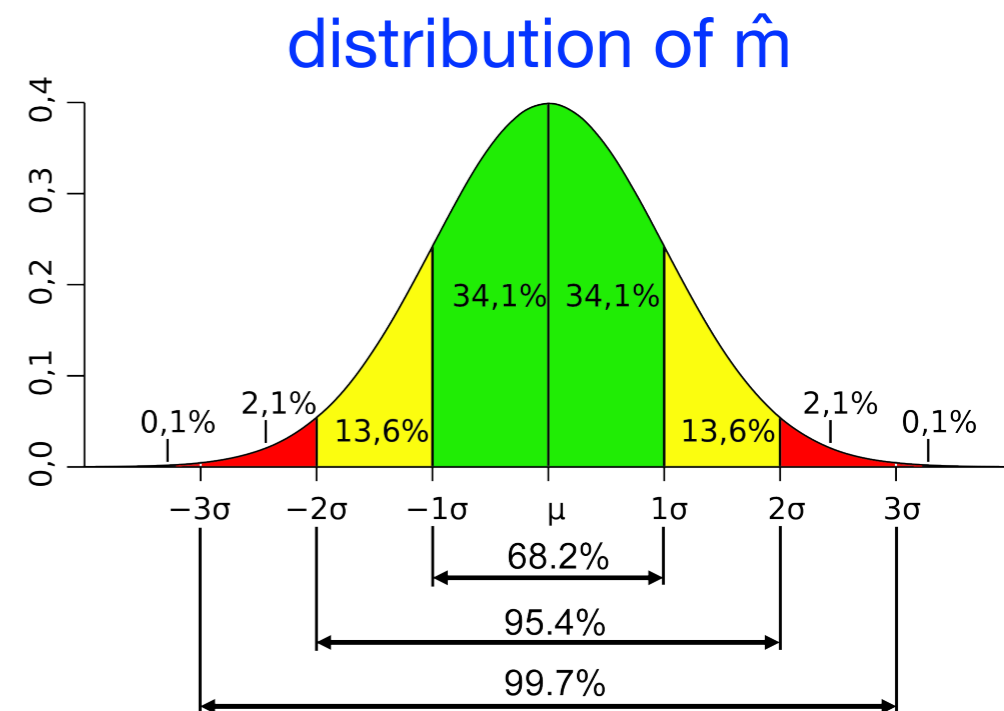
Recall: the best-fit value  $\hat{m}$  is function of observed data  $x_0$ ,  $\hat{m} = \hat{m}(x_0)$ . Other data  $x_1$  yield different best-fit value  $\hat{m}' = \hat{m}'(x_1)$

By repeating over many data sets, get distribution of best-fit values. Uncertainty is its standard dev.

Analytical calculation of  $E[(\hat{m}(x) - E[\hat{m}(x)])^2]$  requires analytical form of  $p(x|m)$  and tractable integrals.

Rarely applicable except in textbook examples

- Approximate uncertainty with its ideal lower limit.
- Get it brute-force from simulation.



Gaussian only in the ideal asymptotic limit (infinite observations  $N$  for each experiment)

# MLE variance from the minimum variance bound

---

The Fisher information is a measure of the information carried by an observation  $x$  over a parameter  $m$ , which is connected to  $x$  by the model  $p(x|m) = L_x(m)$

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

The variance of an estimator is as high or higher than the inverse of the Fisher information.

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

Hence, approximate the variance as the curvature (2nd derivative) of the log-L at its maximum. This is the uncertainty MINUIT returns after MIGRAD/HESSE.

Accurate only for linear problems. 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. If in doubt check with toys.

# MLE variance from simulation

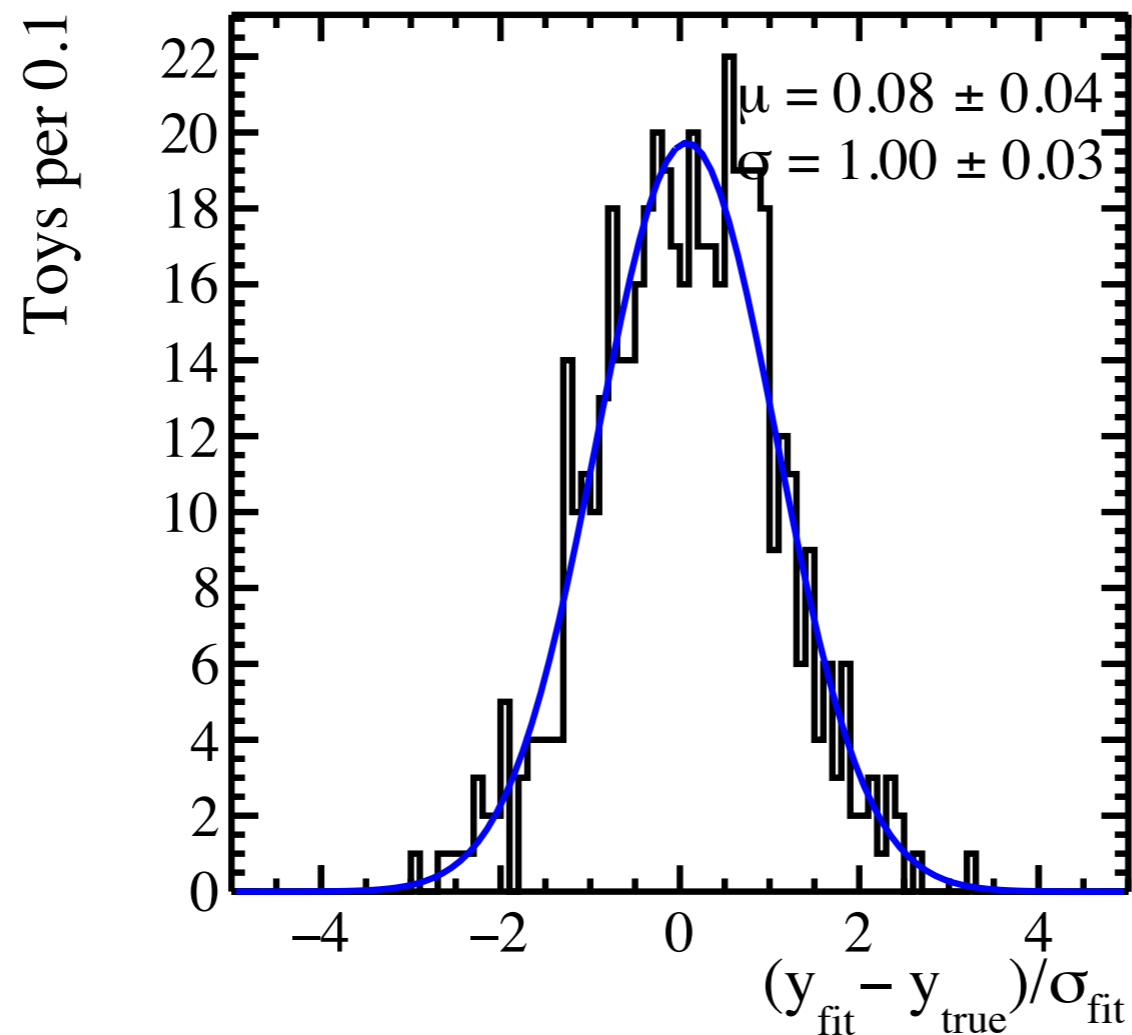
---

Use simplified simulated experiments (“toy Monte Carlo”) drawn from the likelihood to understand the distribution of the ML estimators and their properties prior to applying them to data.

- Choose a plausible true value of the relevant parameter  $m$
- Feed it into  $L(m)$  and generate several sets of simulated data  $x$  from random numbers distributed according to  $p(x|m)$
- Maximize the likelihood in each set (that is, repeat the experiment) and look at the distribution of the estimator
- Repeat for all relevant choices of true value  $m$  (important and often overlooked)

# A standard diagnostics - “fit pulls”

Each entry based on the “result of the measurement” in a simulated experiment, generated with the same set of true parameters



ML estimator of  $y$  perhaps biased. Uncertainty seems OK

Distribution of the difference between ML estimate and the true value of the parameter, divided by the estimate of the std dev.



# What is the statistical uncertainty?

The square root of the variance of the estimator.

Usually, a result is quoted as  $\hat{m} \pm \sigma$

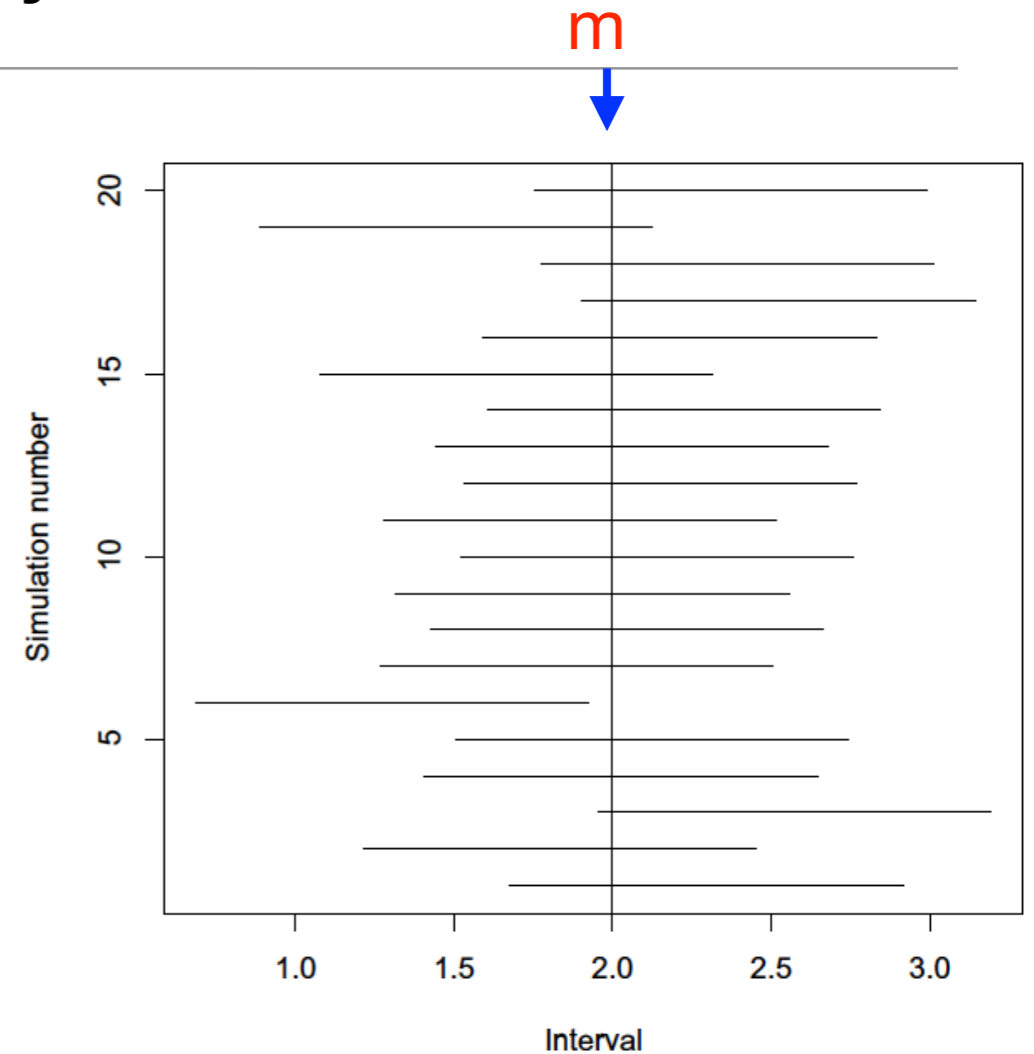
This means that in repeated experiments, 68.3% of the resulting  $[\hat{m} - \sigma, \hat{m} + \sigma]$  intervals include (“cover”) the true value of the parameter being estimated.

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 important.

The true value  $m$  is not random: cannot move around or have a probability.

Only data, that is, the interval extremes (which are functions of data), 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. Why?

Honest and consistent communication with our peers: when they read “the result is  $\hat{m} \pm \sigma$ ” they assume that the  $[\hat{m} - \sigma, \hat{m} + \sigma]$  interval has 68.3% probability to *include* the true value of the parameter being estimated.

Coverage is 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 set of intervals that have coverage under repeated sampling).

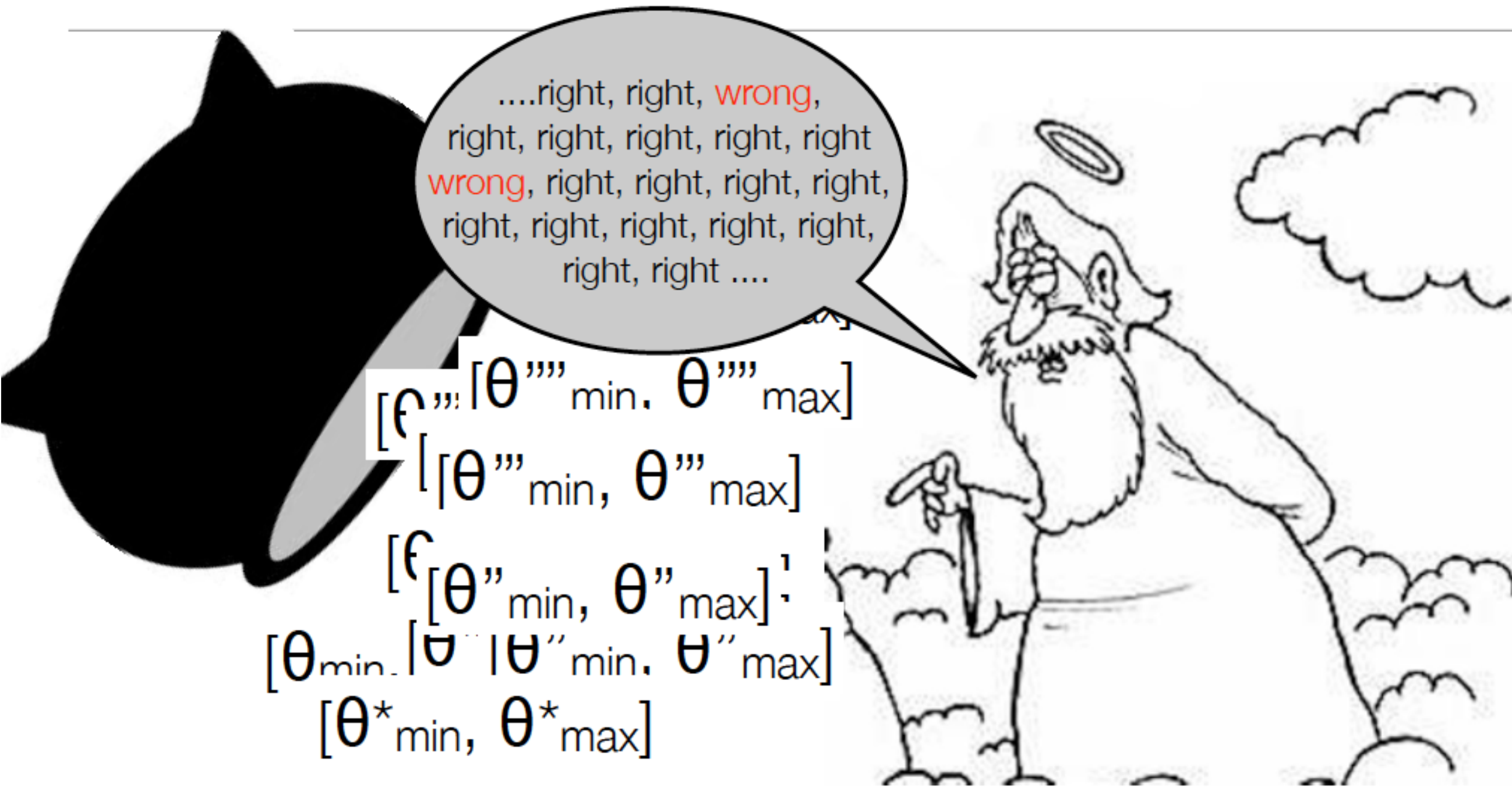
Explicitly checking coverage against various choices of true values for the unknown parameter  $m$  is a key (and oft-overlooked) check for any analysis, especially complicated global fits used to make exciting claims.

# Coverage — reminder

---

A property of the procedure, not of the single measurement.

---



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

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

$$[\theta''', \min, \theta''', \max]$$

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

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

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

# When the MLE fails

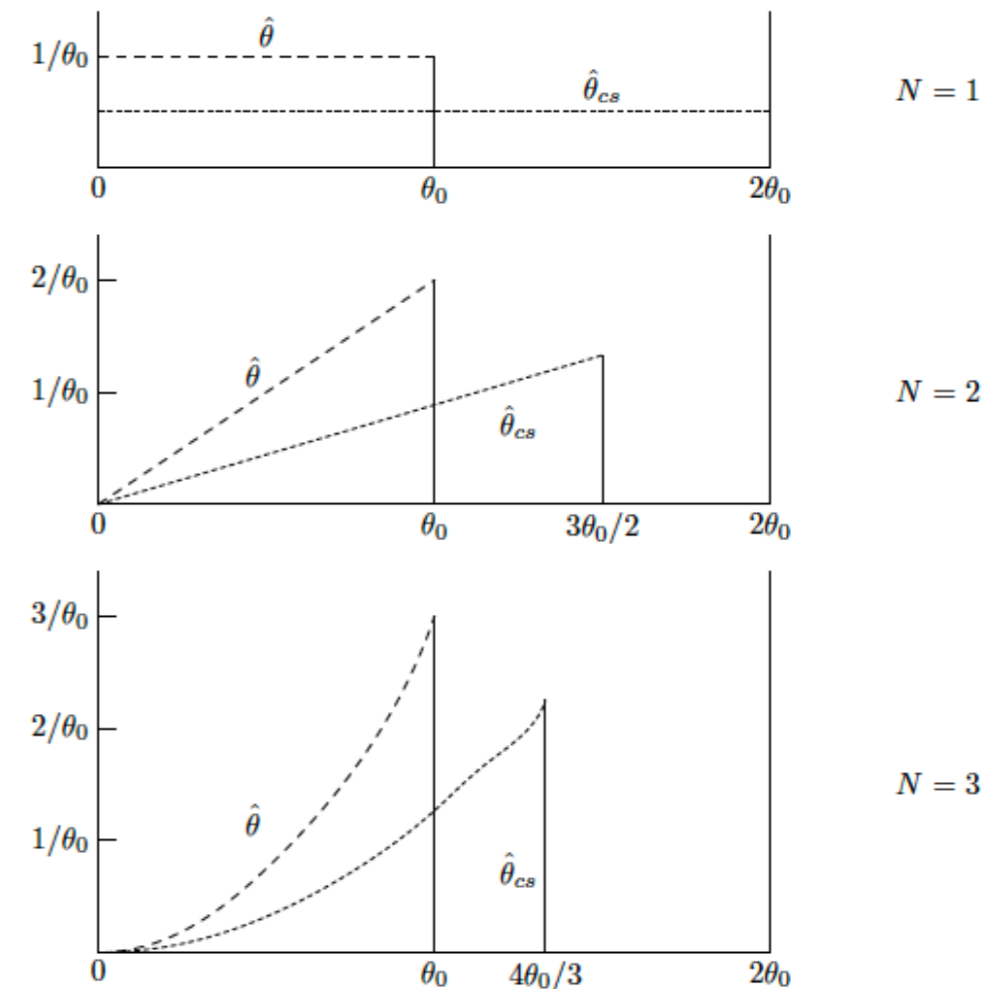
Suppose that one observes  $N$  events  $X_i$  chosen randomly from a uniform distribution between 0 and  $\theta$ , where the upper bound  $\theta$  is the unknown parameter.

This is a case where **the range of the data depends on the value of the parameter  $\theta$** .

Since  $\theta \geq X_i$  for all  $i$ , the likelihood function  $L = \theta^{-N}$  will have its maximum at  $\hat{\theta} = X_{\max}$ , where  $X_{\max}$  is the largest observed value of  $X$ . It is clear that this estimator (almost) always gives a result which is too small, and the obvious correction is to use the common-sense estimate:

$$\hat{\theta}_{cs} = X_{\max} + \frac{X_{\max}}{N}.$$

This estimate in fact turns out to be unbiased, as can easily be verified.



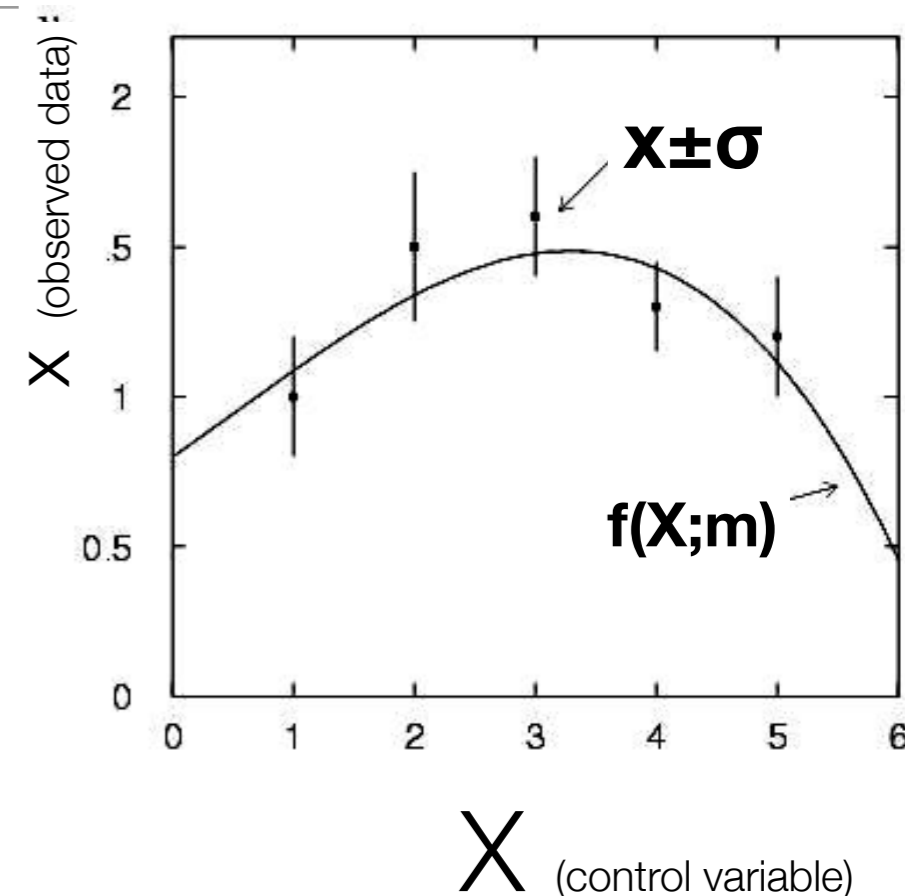
# Other estimators: least-squares (in one slide)

N independent observations  $x_1, \dots, x_N$  that **fluctuate following Gaussian distributions of known\* variance**  $\sigma_i^2$  around **their known\* expected values  $f(X_i; m)$**  that are functions of **a known\* control variable  $X_i$**  and unknown parameter  $m$ .

Maximizing the (Gaussian) likelihood corresponds to minimizing the least-squares estimator of  $m$

$$\chi(m) = \sum_{i=1}^N \frac{(x_i - f(X_i; m))^2}{\sigma_i^2}$$

LS properties are inferior or equivalent to the MLE. But LS offers an advantage: its value at the minimum  $\chi^2_{\min}$  offers **a measure of the agreement of the model to the data (goodness-of-fit)** since the distribution of  $\chi^2_{\min}$  is known and can be compared with the observed  $\chi^2_{\min}$  value. Holds rigorously only if the red conditions above are true (happens rarely in real analysis)



\*known means known — not estimated. That is, known means there is no uncertainty, something that ~never happens in real life. One usually approximates the known variance with the estimated one

# Can we get goodness-of-fit for MLE?

---

**Not in fits to unbinned data.**

Usage of the MLE distribution to derive goodness-of-fit as suggested by some books is flawed (see <https://arxiv.org/abs/physics/0310167>)

Heuristic methods discussed in <https://arxiv.org/abs/1006.3019> are shown to “work” in simple toy MC examples, but no general demonstration of their success and properties is given — no guarantee exists that they’ll work in any general problem.

**To date, no widely accepted method for evaluating goodness-of-fit in unbinned fits exists.**