Upper limits

Belle II physics week — Oct 10, 2025 - KEK

Diego (Trieste)

The task

A little difficult to cover in 1 hr slot.

Executive decision (by myself) to only cover limits…

For fitting check these out

https://indico.belle2.org/login/?next=/event/1332/contributions/6424/attachments/3194/4876/Fitting.pdf

https://indico.belle2.org/event/3456/contributions/18541/attachments/10210/15687/fitting-belle2-academy.pdf

Inspirational premise

You (we…) all must strive to find signal

Even in very-rare-decay searches

Should be confident

Confidence is essential to avoid missing out on possible (big) discoveries

…unexpected and so not found…

APPARENT EVIDENCE OF POLARIZATION IN A BEAM OF β-RAYS

By R. T. Cox, C. G. McIlwraith and B. Kurrelmeyer*

New York University and Columbia University†

Communicated June 6, 1928

THE SCATTERING OF FAST ELECTRONS BY METALS. II. POLARIZATION BY DOUBLE SCATTERING AT RIGHT ANGLES

By Carl T. Chase

New York University, University Heights, N. Y.

(Received July 28, 1930)

We have made no attempt at a theoretical treatment of double scattering beyond a consideration of the question whether the results here reported are of an asymmetry of higher order than what might be expected of a

1930's "anomalous polarization" in β decays was early indication of parity violation

But Cox, Chase, collaborators and then HEP community were not ready just yet ;)

…unexpected and so not found…

CHAIRMAN'S SUMMARY

L M Lederman

Columbia University

1. $\left(\frac{e}{\gamma}\right)^+ \approx \left(\frac{e}{\pi}\right)^- \approx \left(\frac{\mu}{\gamma}\right)^+ \approx \left(\frac{\mu}{\pi}\right)^- \approx 10^{-4}$
2. This is independent of P_T from 1.5 to 5 GeV/c.
3. This is independent of nucleon target size.
4. This is independent of CM viewing angle.
5. This is independent of s from $\sqrt{s} = 7$ to $\sqrt{s} = 53$ (See Fig. 1).

All of these statements may be true to within a factor of 2 or so.

(A BNL point is taken from a comment by R Adair). The implications are that leptons and pions have a common origin. Statement 5 implies the source mass must be less than 3-4 GeV (no threshold effects) for

$$p + p \rightarrow X + \text{anything}$$
$$\hookrightarrow \text{leptons}$$

or less than 1.5-2 GeV for pion production e.g. Charmed particles. Statement (1) in its lack of charge asymmetry is discouraging for charmed meson sources analogous to K-mesons. The agreement of the ISR with NAL rules out low masses (M_X > few hundred MeV) because narrow angle leptons are vetoed in the ISR measurements.

The ISR muons and NAL electrons set limits on the production of single leptons e.g. from $W^{\pm}$ up to the kinematic limit. However, it is out of fashion to convert these limits to mass limits because the necessary models are currently discredited.

The lack of $P_\perp$ "bumps" means there are no significant heavy objects (M from 3 → 10 GeV) decaying into two leptons.

History is fickle. ff

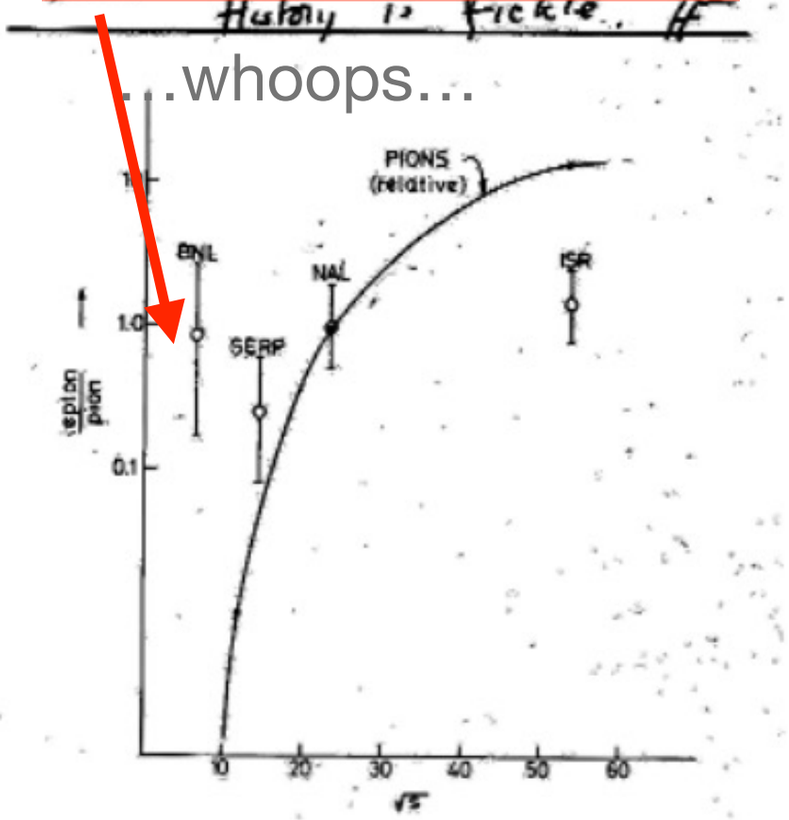


Fig. 1 lepton/pion ratio vs $\sqrt{s}$ compared to pion production ($P_\perp \sim 3$ GeV). Errors are estimated freely.

"the lack of pT 'bumps' means there are no significant heavy objects (M from 3 -> 10 GeV) decaying into two leptons"

L.M Lederman - 1971

…unexpected and so not found…

Experimental Limits on the Decays $K_L{}^0 \to \mu^+\mu^-$, e^+e^-, and $\mu^\pm e^\mp$†

Alan R. Clark, T. Elioff,* R. C. Field, H. J. Frisch, Rolland P. Johnson,
Leroy T. Kerth, and W. A. Wenzel
Lawrence Radiation Laboratory, University of California, Berkeley, California
(Received 9 April 1971)

We have performed a search at the Bevatron for the decays $K_L{}^0 \to \mu^+\mu^-$, e^+e^-, and $\mu^\pm e^\mp$ with a double magnetic spectrometer using wire spark chambers. Over 10^6 observed $K_L{}^0 \to \pi^+\pi^-$ decays determine the normalization for the di-lepton decay modes. No e^+e^- or $\mu^\pm e^\mp$ events were observed. For each of these decays the upper limit on the branching ratio relative to all modes is 1.57×10^{-9} (90% confidence level). For the decay $K_L{}^0 \to \mu^+\mu^-$, the limit is 1.82×10^{-9} (90% confidence level).

…whoops…

PDG today

$\mu^+\mu^-$	S1	$(6.84 \pm 0.11) \times 10^{-9}$

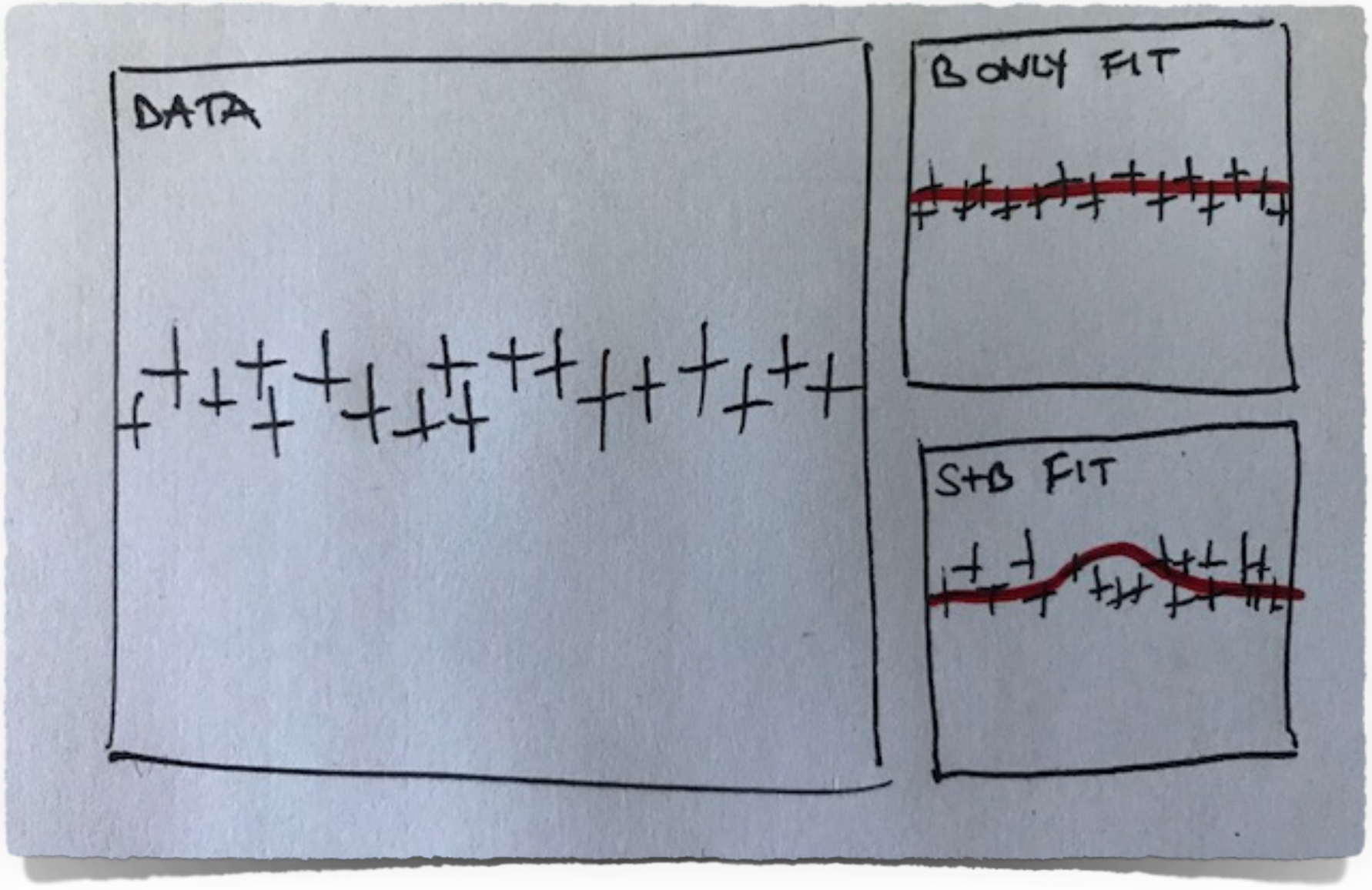
But not too confident

Sometimes the signal just isn’t there

Still, your results are important and informative, if *properly* rekindled as *exclusion limits*

The real talk

The simplest example…



Looking for distinctive signal structure over background

Fit data with a model that allows for signal and background, “the (S+B) model”

Data estimate for signal yield $\hat{N}_s$ consistent with zero….let’s set an exclusion limit

Fundamental ingredients

The model

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

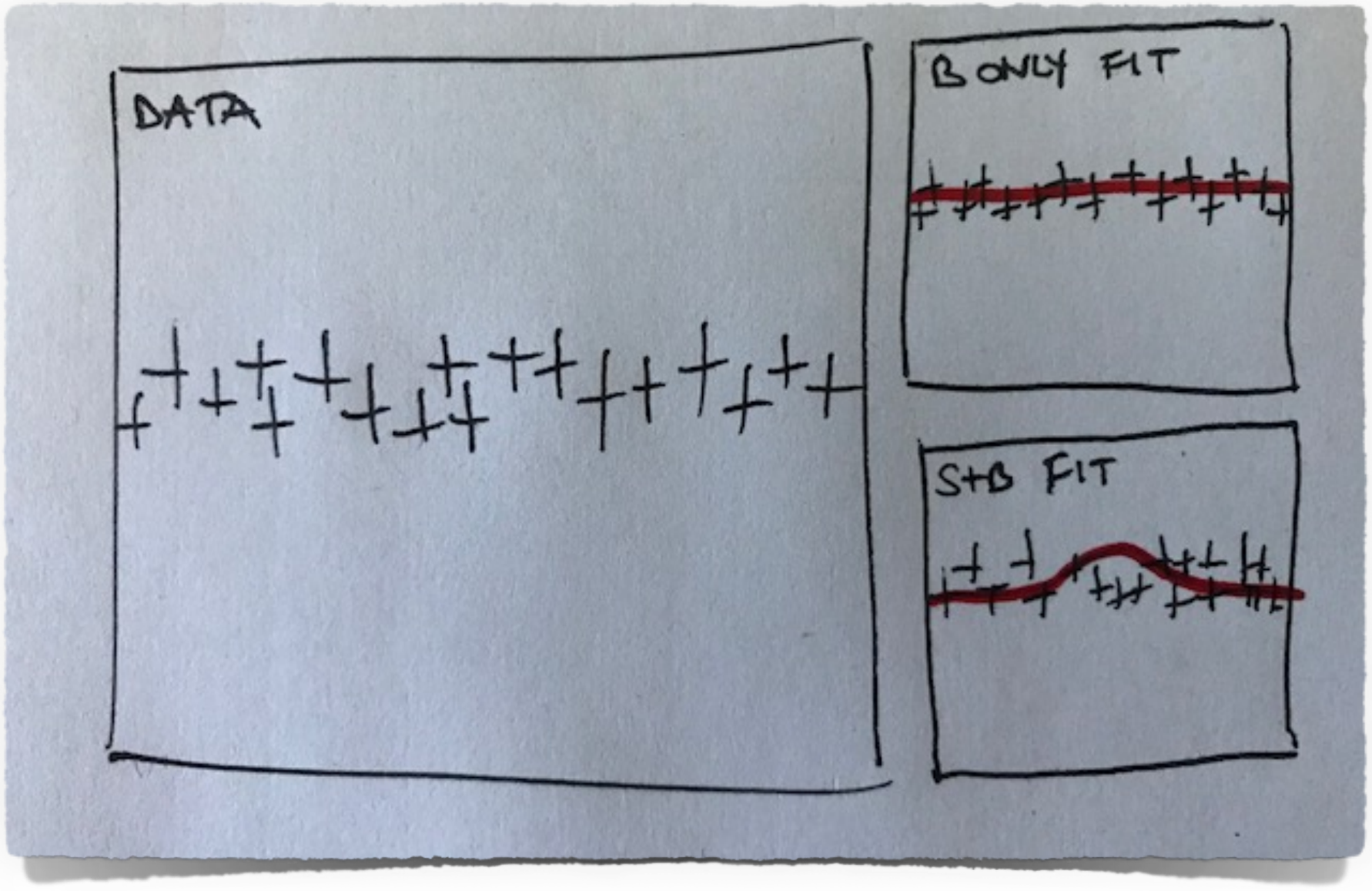
Probability of “data” given “parameters” — a mathematical construct.

Connects what you wanna measure with what you observe.

Interpreted as a function of data x (fixing $m = m_0$), it is the probability density function $p(x|m_0)$: *probability to observe each possible value of data x had the true value of m been m_0.*

Interpreted as a function of parameter m (fixing $x = x_0$) it is the likelihood $p(x_0|m) = L(m)$ to observe data x_0 for different choices of m

In our case



$$p(x \mid m) = p(\text{invariant mass} \mid \text{signal yield}) \propto \text{const} + \text{bump}$$

Accelerated recap: Bayesian/frequentist inference

Measuring physics parameter m consists in

- devising a model p(x|m) that approximately describes the phenomenon
- observing data x
- using the model and the data x to get information on m.

Bayesian — combine model with prior probabilities for m to determine the posterior probability p(m|x), which expresses the probability for each value of parameter m given the data. (“Prior” == known or chosen before observing x)

Frequentists — cannot define p(m|x), use model and probabilities for all other possible x outcomes to determine which values of m would produce the observed data x with highest probability

Bayesian limits

Probability for the parameter given the data

Use trivial property of conditional probabilities to answer the question: what's the probability that physics parameter has a certain value given the data I observed ?

Likelihood of your data (model)

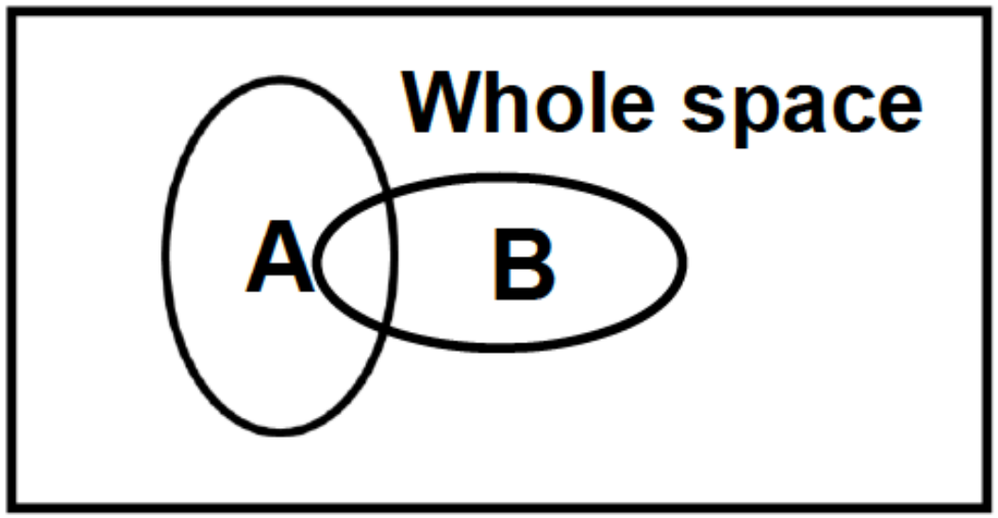
Prior probability (your assumption)

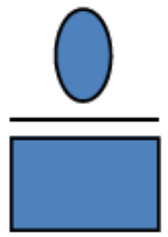
$$p(m|x) = \frac{p(x|m) \times p(m)}{p(x)}$$

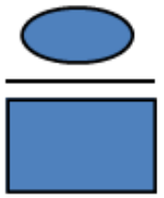
Posterior probability

Normalization

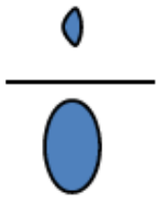
A “visual” demonstration of Bayes theorem

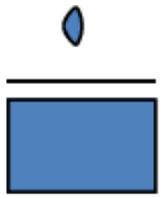


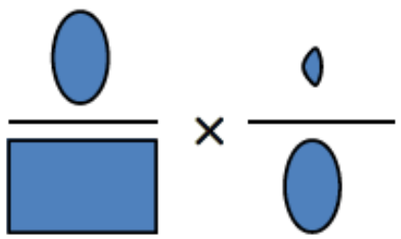
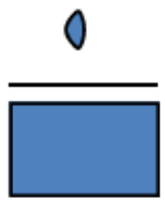
$P(A) =$ 

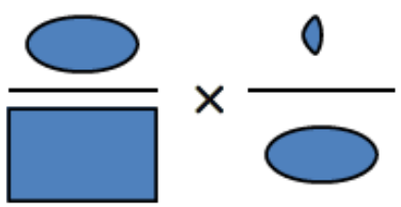
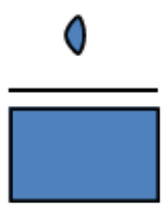
$P(B) =$ 

$P(A|B) =$ 

$P(B|A) =$ 

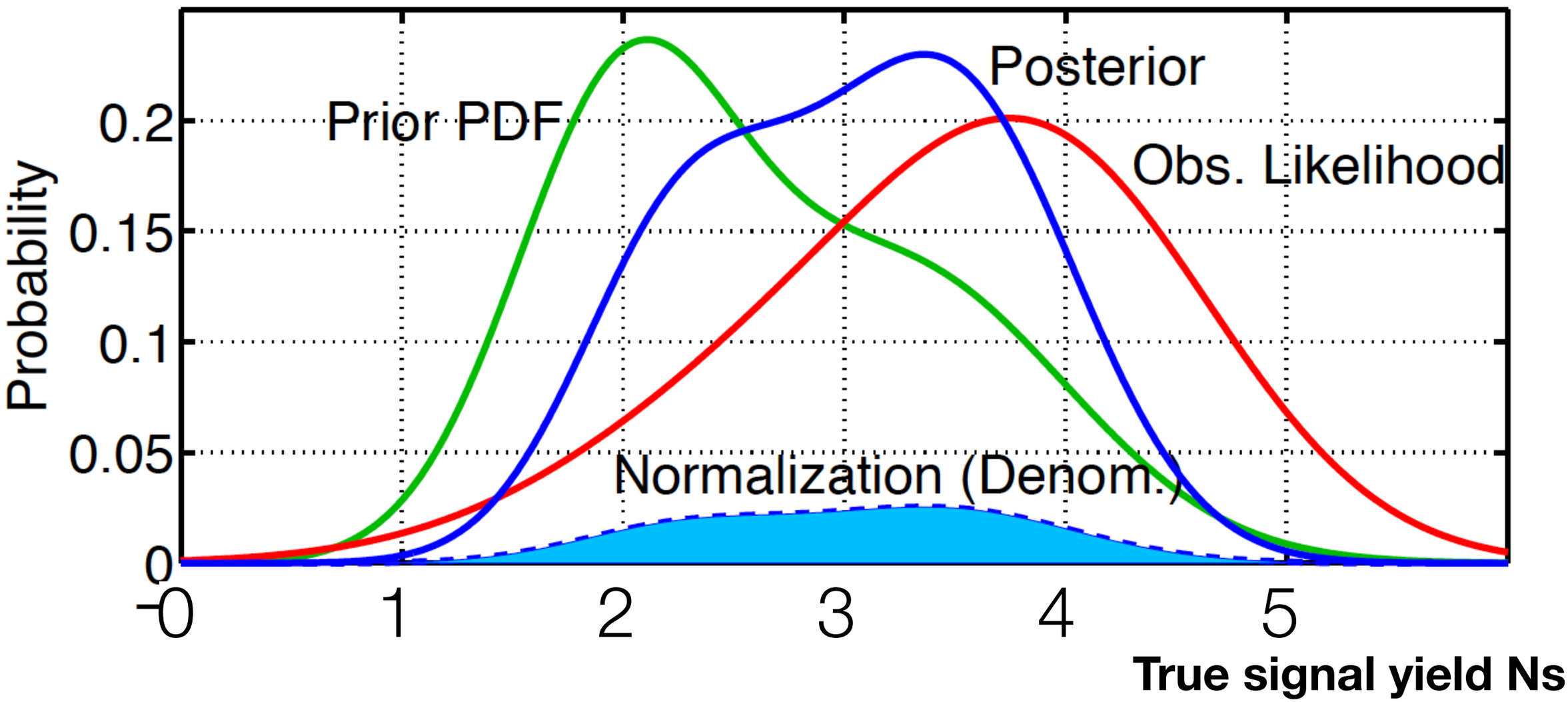
$P(A \cap B) =$ 

$P(A) \times P(B|A) =$  $=$  $= P(A \cap B)$

$P(B) \times P(A|B) =$  $=$  $= P(A \cap B)$

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

Those p(…|…) are all different things….



In case that's not yet clear….

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

Variable A: "pregnant", "not pregnant"

Variable B: "male", "female".

P(pregnant | female) ~ 3% but

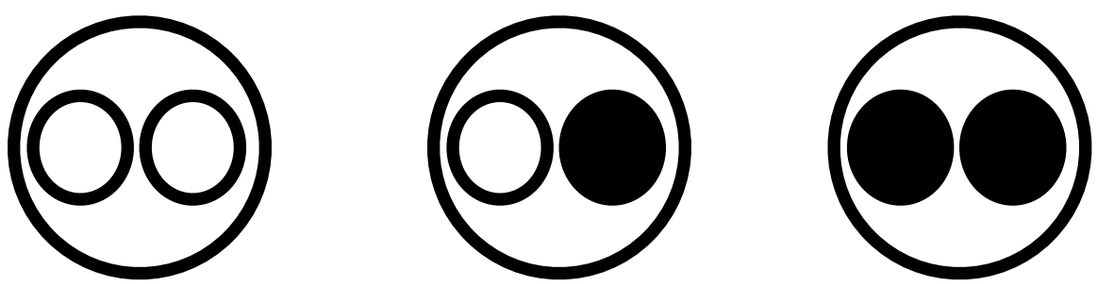
P(female | pregnant) >>> 3% !

[Lyons]

Bayesian inference — elementary example

homework

Three identical bags. Two balls in each. Balls can be black or white



bag A bag B bag C

Pick a random bag (of type **m — parameter, unobservable**) and a random ball inside it (of color **x — data, observable**)

Ball is white: x = w. What can one say about the chosen bag?

Wanna know probability p(m|w) for picking each type of bag, given the observed ball is white

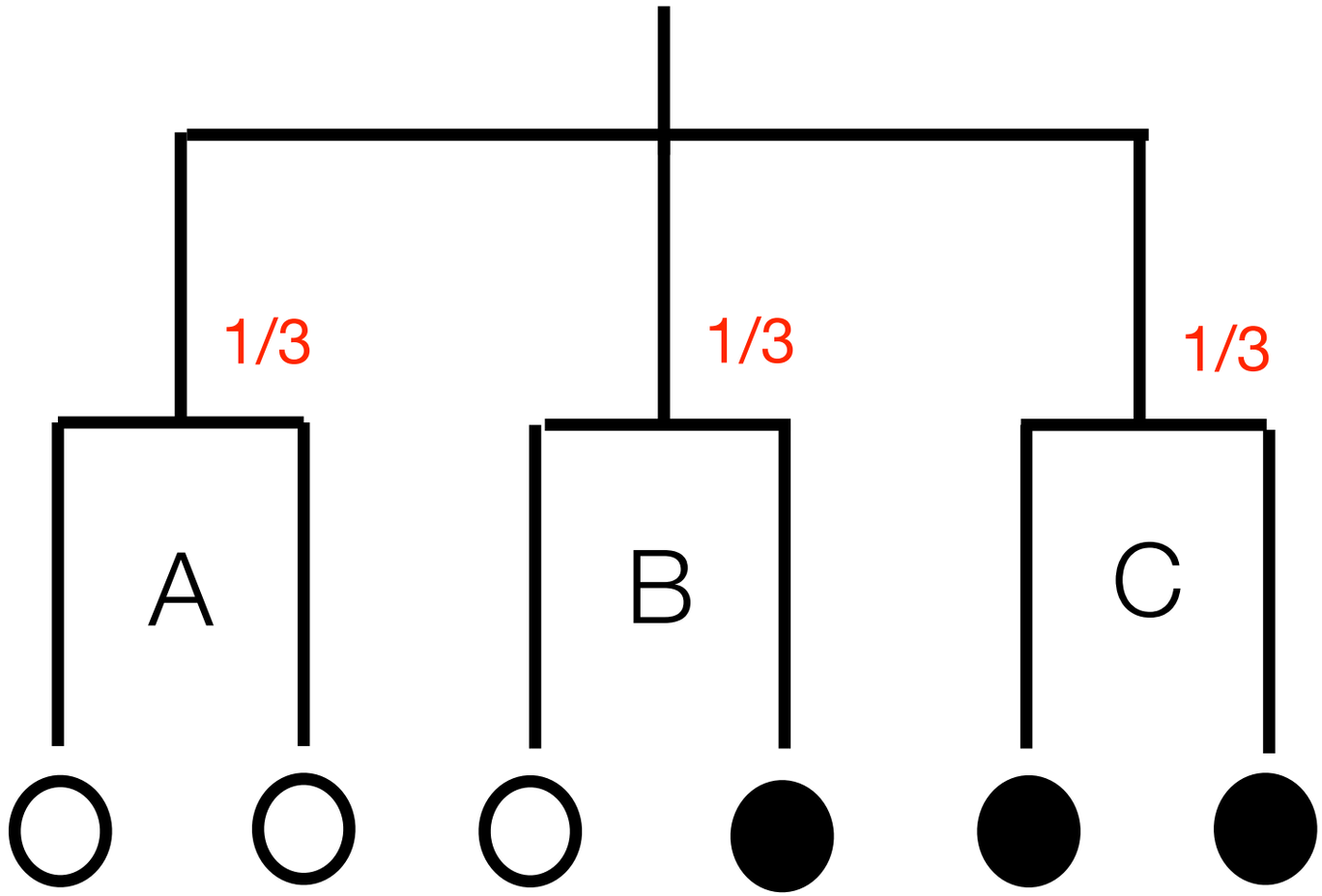
I know the priors p(m), which are 1/3 for each bag type

If priors are known - everyone should be Bayesian

Bayesian inference — elementary example

homework

$p(m)$



$p(w|m)$ — 1 | 1/2 | 0

$p(w,m) = p(w|m)\ p(m)$ — 1/3 | 1/6 | 0

$p(m|w) = p(w|m)\ p(m)/p(w)$ — 2/3 | 1/3 | 0

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

Once you got the posterior, limits are easy

Choose prior probablity $p(N_s)$: expresses knowledge on unknown parameter

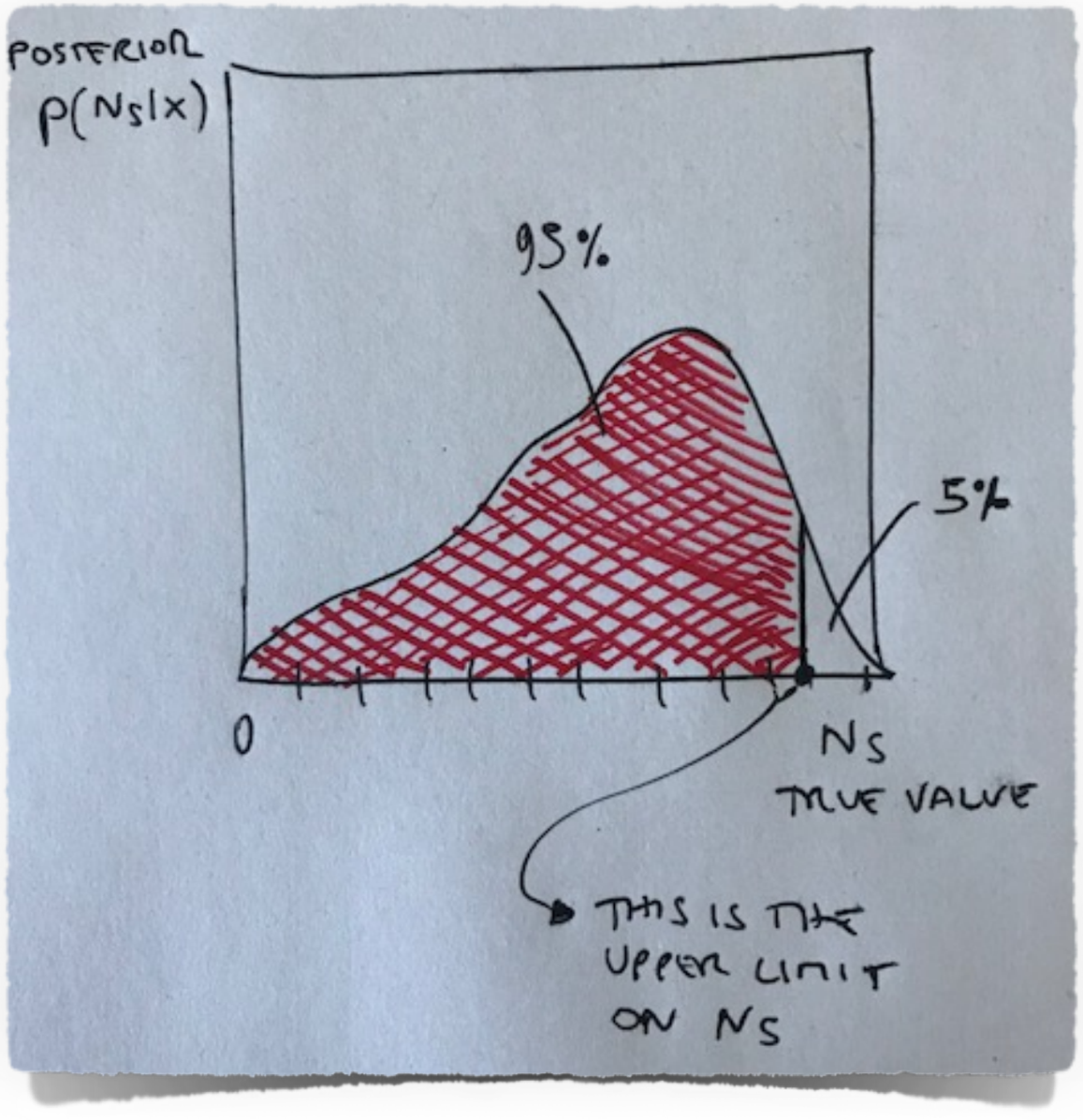
Integrate (marginalize) posterior

Likelihood Prior

$$p(N_s|x) = \frac{p(x|N_s)p(N_s)}{\int_{N_s} p(x|N_s)p(N_s)dN_s}$$

over true values of parameter N_s until reaching fractional area corresponding to desired *Bayesian credibility,* e.g. 95%.

(better to call it “Bayesian credibility” than confidence level if Bayesian inference is involved)



Obtained upper limit depends on the choice of the prior

Priors (long story short)

Priors carry *subjective information* that influences results.

Results dominated by *data information* (i.e., by the likelihood) rather than by prior information are preferable

While dealing with limits data typically scarce — so priors often relevant

Revert then to use “noninformative priors”

Noninformative priors do not exist.

“Flat priors” seem naively equanimous.

Plus, maximum of the posterior coincide with maximum of the likelihood when likelihood is only one-dimensional

But flat has no special role — it depends on the metric, can be as much informative as any other choice

Serious efforts toward priors (Jeffreys’ et al.) that inject statistically-motivated information in the inference — difficulties in high dimensions.

These days emphasis on prior-sensitivity studies

Assessing sensitivity to priors

How much is the final result driven by data and how much by the prior?

Change the prior and check variation in results

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]	

PRD 85, 072002 (2011)

Appendix B: Prior sensitivity study

To investigate the sensitivity of the results in Table I to the choice of priors, we derived the posterior mode and credible intervals for alternative sets of priors.

Firstly, we select truncated-normal priors, centered on the SM expectation (the only non-zero Wilson coefficient being $C_{\mathrm{VL}}^{\mathrm{SM}} = 6.6$), which disfavor deviations from the SM expectation,

$$p(\eta_i) = \begin{cases} \mathcal{N}(\eta_i|\mu = C_i^{\mathrm{SM}}, \sigma = 20) & \eta_i \geq 0 \\ 0 & \eta_i < 0 \end{cases}. \quad \text{(B1)}$$

Here $\eta_i \in [|C_{\mathrm{VL}}+C_{\mathrm{VR}}|, |C_{\mathrm{SL}}+C_{\mathrm{SR}}|, |C_{\mathrm{TL}}|]$ and C_i^{SM} correspond to the respective SM point $C_i^{\mathrm{SM}} \in [6.6, 0.0, 0.0]$.

Secondly, we select uniform priors in the squared Wilson coefficients, as these enter Eq. (7), which subse-

in Belle II too: arXiv: 2507.12393 (2025)

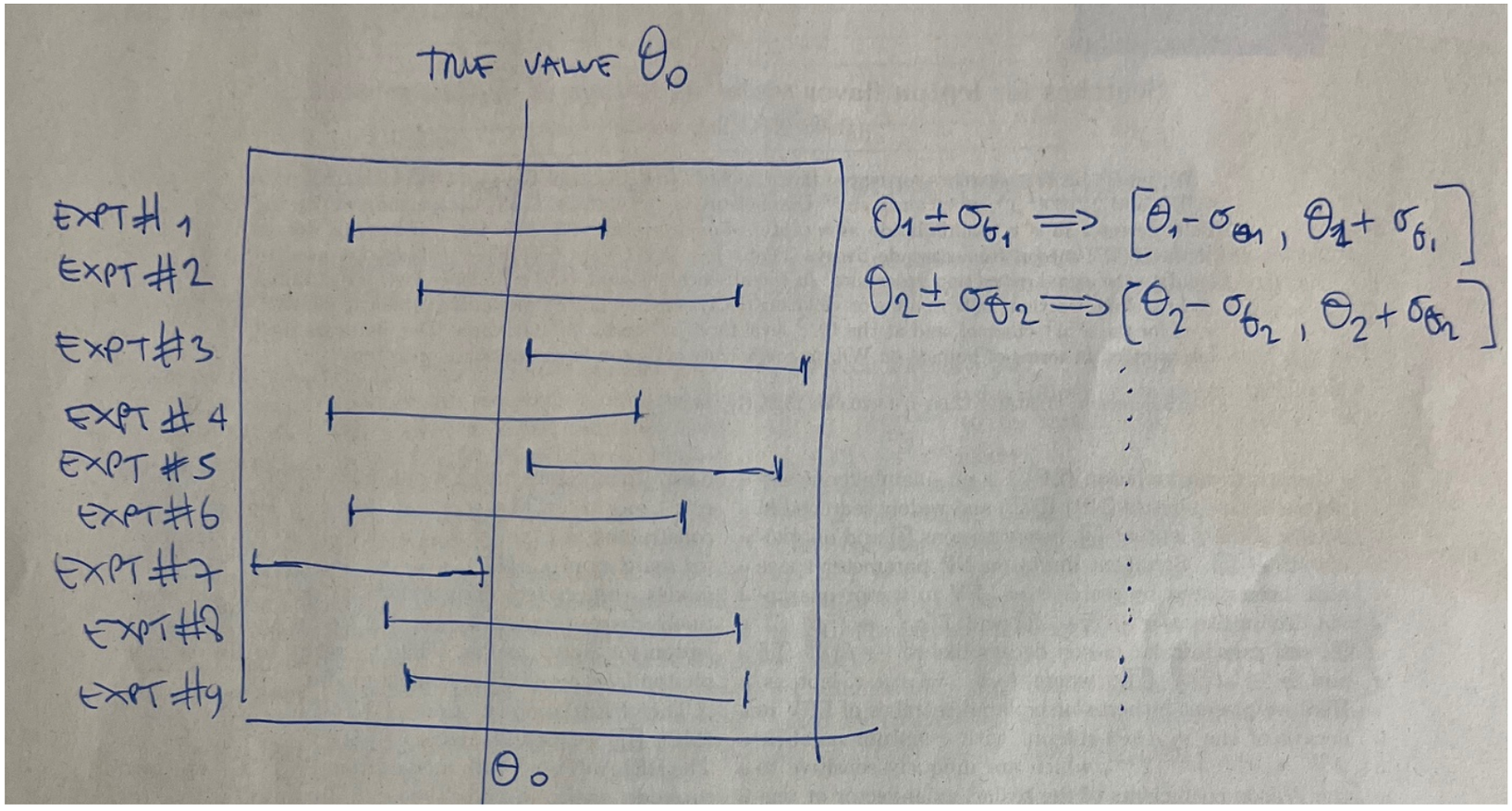
A desirable “calibration” of any Bayesian result.

Frequentist limits

Coverage (of the true value)

A property of an inference procedure: yield results that include the true value with the stated *confidence level*

For instance, 1/3 of the 1-sigma (68%) intervals should contain the true value



The true value isn't random — cannot move around or have a probability distribution

Data, that is, the interval extremes, are random and fluctuate

Coverage (of the true value)

Coverage is the central requirement of frequentist inference.

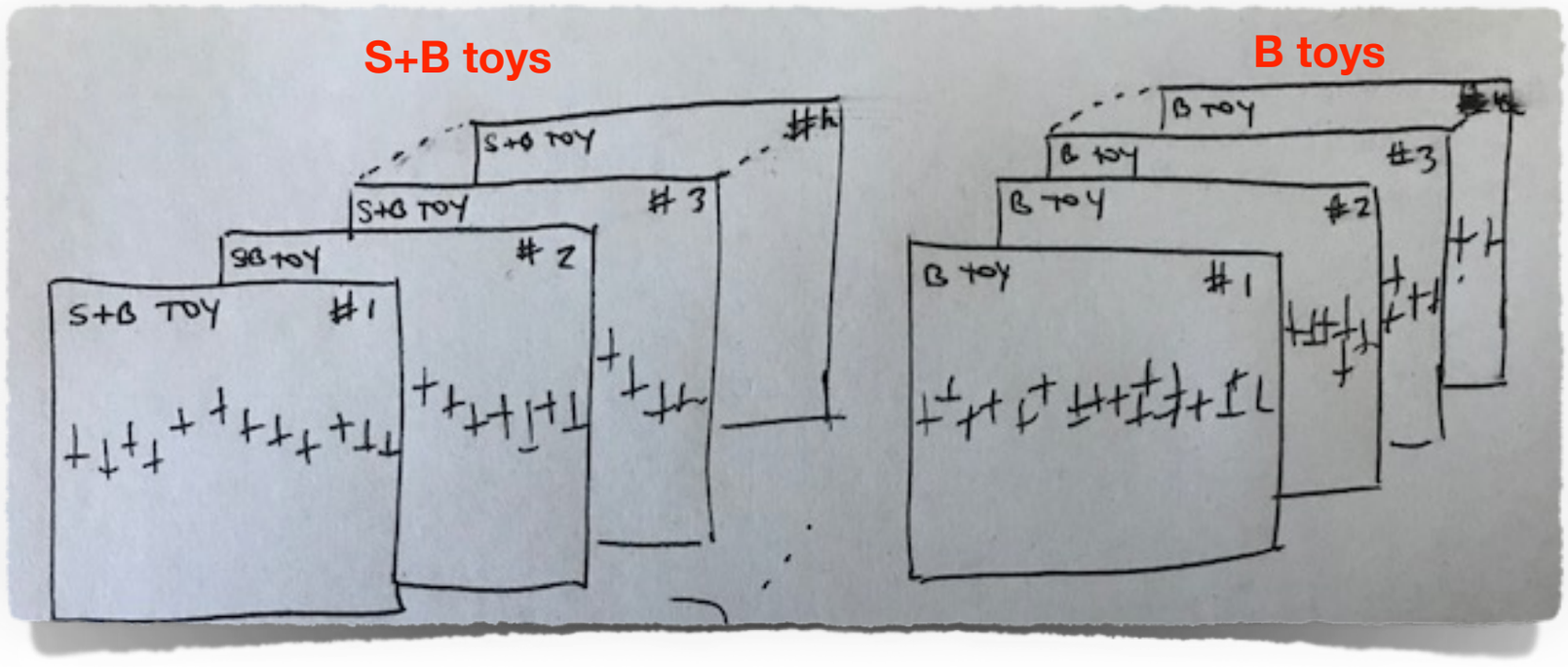
It provides a precise and objective meaning to the results of an inference

When someone reads your paper, she knows (or assumes) that the central values and uncertainties are obtained through a procedure that has coverage, therefore knowing where the true value is likely to be.

As coverage implies repeatability of the inference, toys simulation is commonplace in frequentist statistics.

Back to our example — testing a signal strength

Test compatibility of data with an assumed signal strength



Assume a signal strength, which determines expected signal yield in sample

Generate a set of toys by drawing simulated data from the signal+background model

Generate a set of toys by drawing simulated data from the background-only model

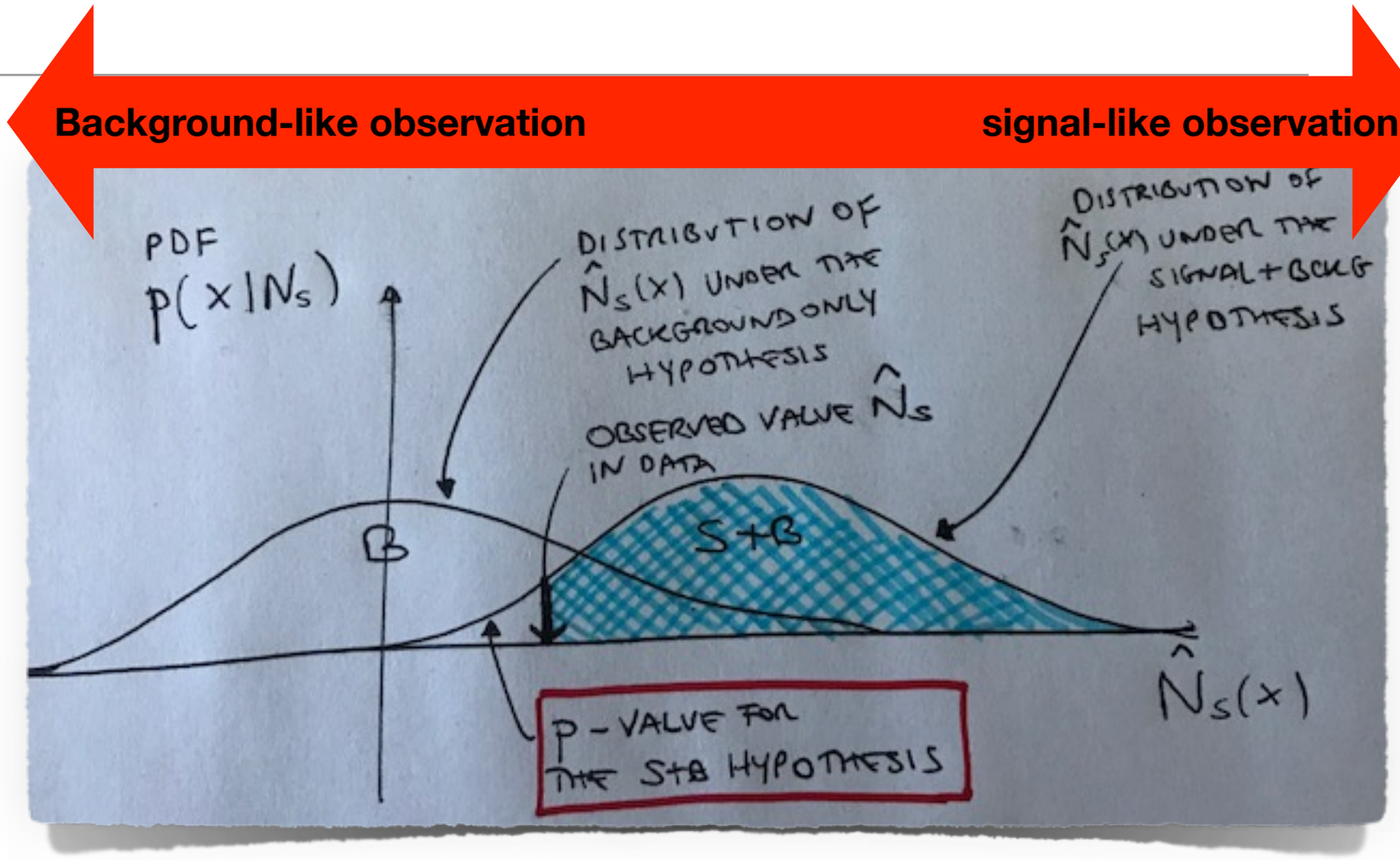
Each toy has same size as the real data (within total Poisson fluctuations of course)

Fit each toy of each set with S+B model: two sets of result for signal yield $\hat{N}_s$

Plot distributions of the fit results, separately for the two sets

p-value = 1- CL

Location of data observation relative to the two curves offers a measure of compatibility of the data with either.

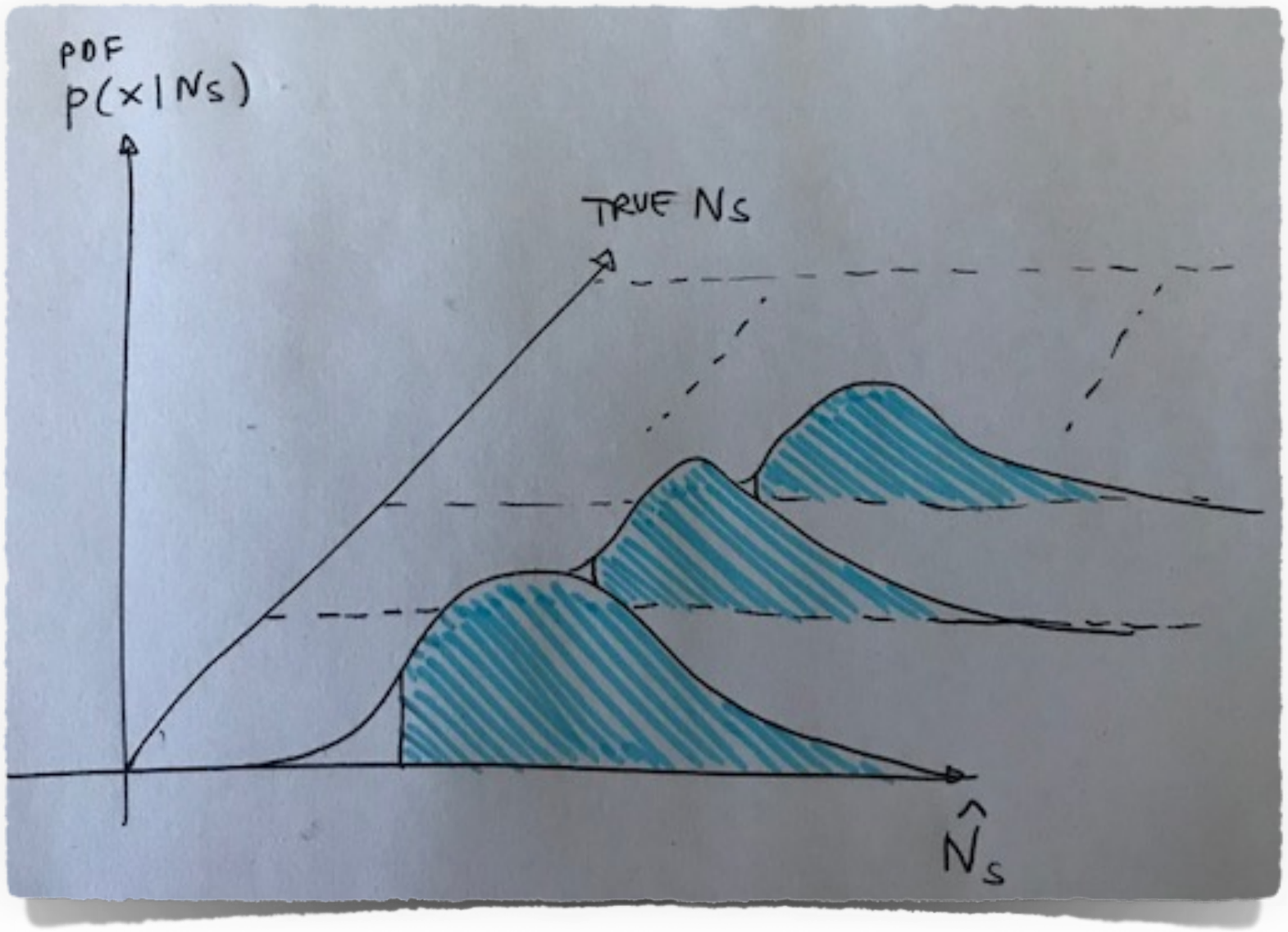


Fractional integral of the S+B curve over values as bgck-like as the one we observed, or more, is the p-value for the “S+B hypothesis”.

The smallest the p-value, the lower the compatibility of data with S+B hypothesis. Small p-value means it’s unlikely to observe our data if model S+B is realized

That is data “excludes the S+B model at a confidence level CL = 1- p”

Testing multiple signal strengths



We only have tested one signal strength.

Useful to test a whole range of signal stregths: repeat for different signal strengths

Comments

It seems straightforward.

It has issues

The principal issue comes from our sequence first look at data and then decide what to do

“Do you see a signal?” “Then measure its strength” “Do you see nothing?” “Then set a limit”)

This spoils *coverage*

Need a procedure that transitions consistently from limit-setting to signal-strength measurement, *prior to looking at data.*

Let’s look at this in detail from scratch

Confidence intervals

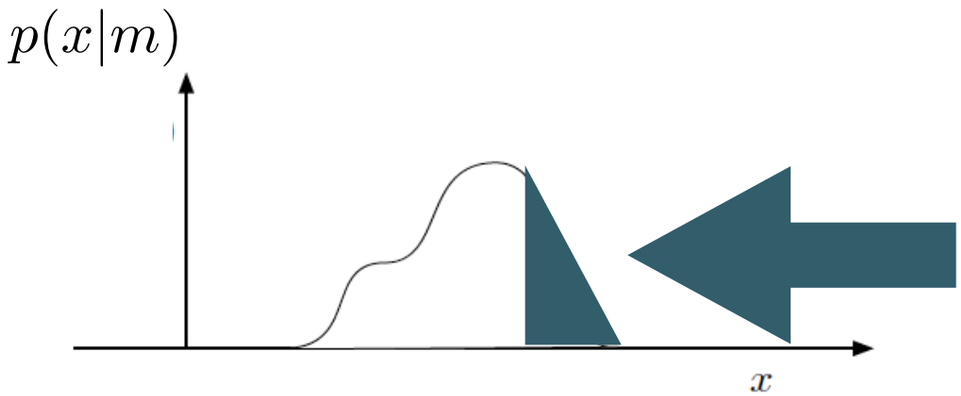
Given a model p(x|m), what values of the unknown parameter m make the observed data x_0 among the least extreme outcomes possible?

"Extreme" needs ordering: rank values of data observations x for each possible value of m from likely to extreme

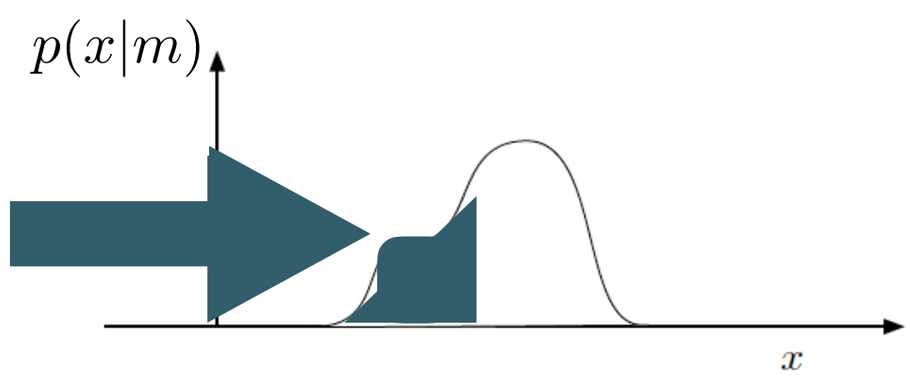
Accumulate the highest-ranked (i.e., less extreme) values of x and sum the corresponding p(x|m) until reaching a CL fraction of the x probability.

Given an ordering and chosen a CL, the confidence interval $[m_1(x), m_2(x)]$ includes values of m for which observed data x_0 are not "extreme" at the chosen CL

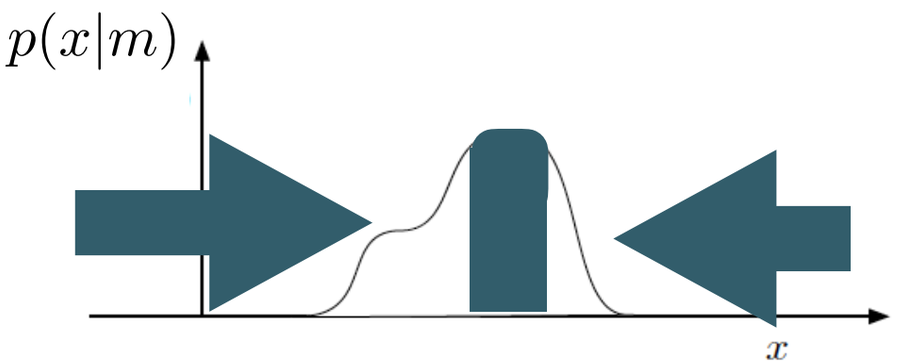
One-sided, two-sided.



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 (interval estimate of m)

(above is simplified: applies only for x one-dimensional and p(x|m) is such that higher m imply higher average x).

CL chosen to match the standard thresholds 68.3% (1σ) 95.5% (2σ) etc.

Neyman construction illustrated

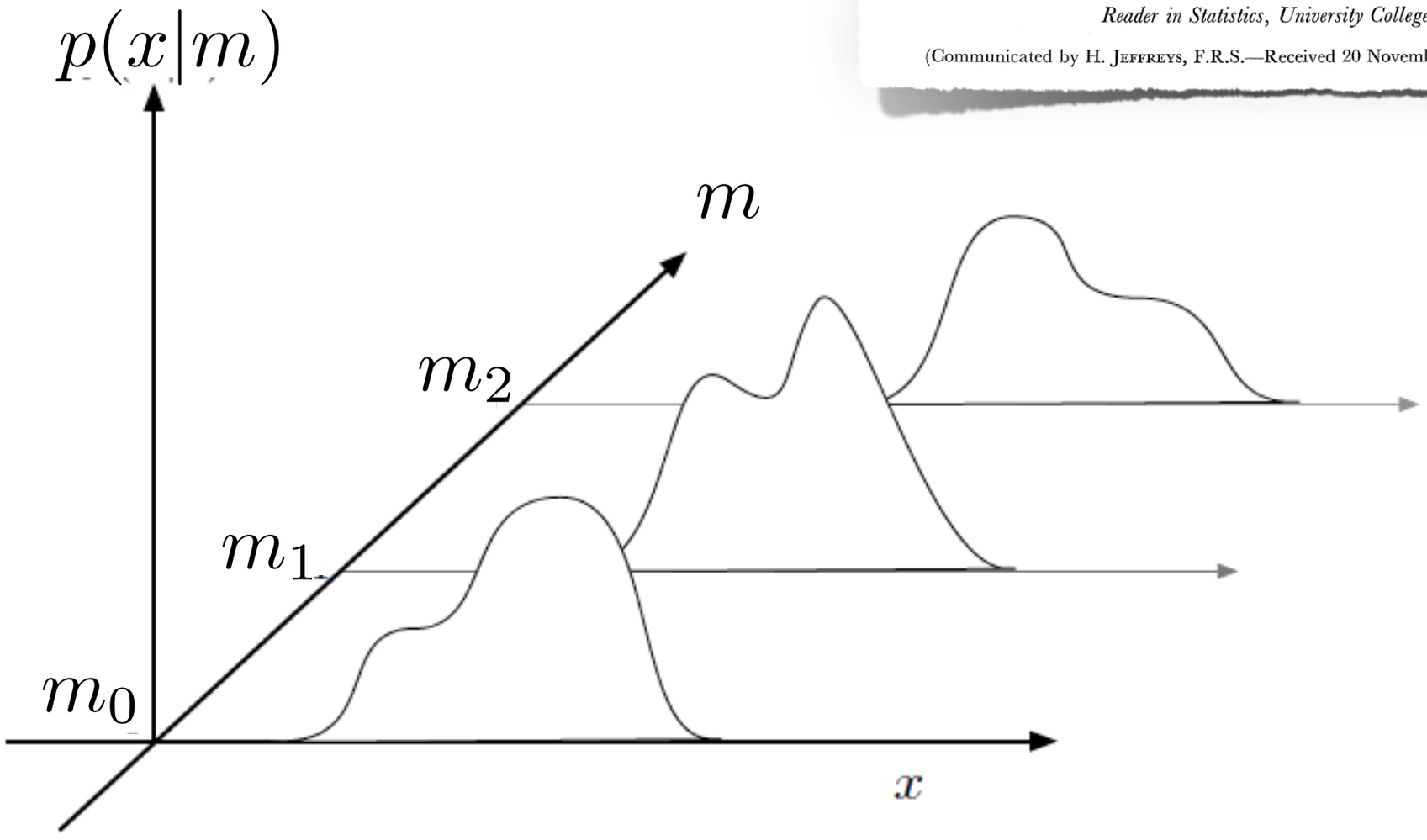
Prior to looking at data, consider p(x|m) for each possible true value of parameter m

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

By J. Neyman

Reader in Statistics, University College, London

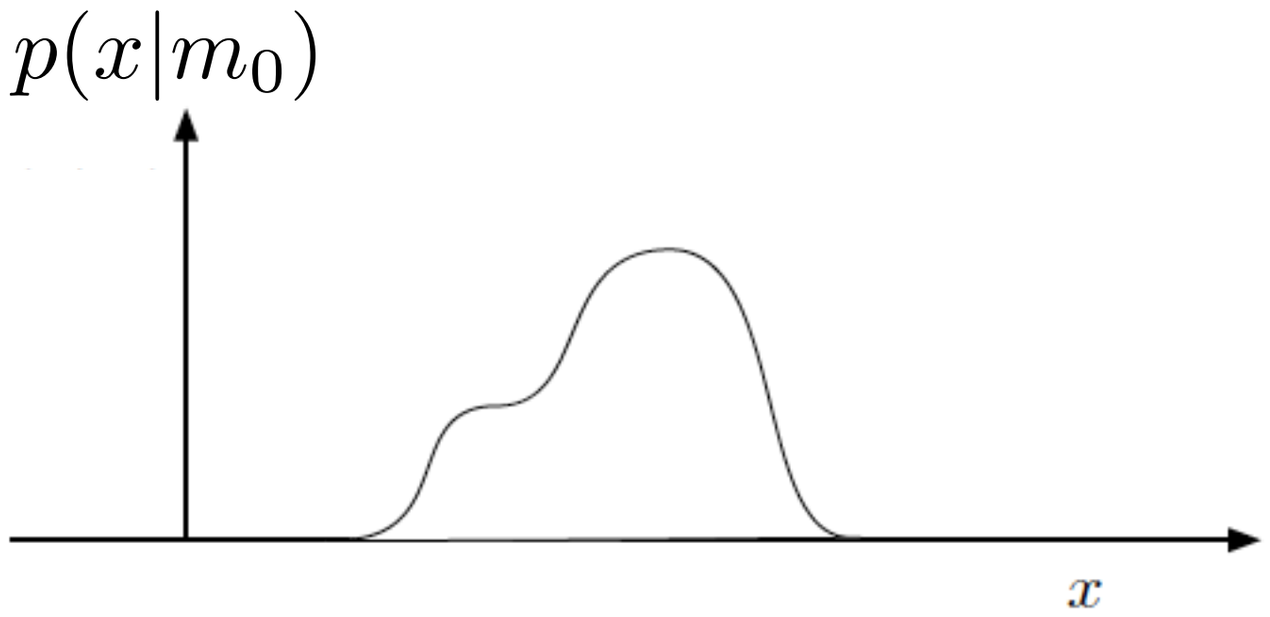
(Communicated by H. Jeffreys, F.R.S.—Received 20 November, 1936—Read 17 June, 1937)



(Tipically "x" is chosen to be the maximum likelihood estimator of m)

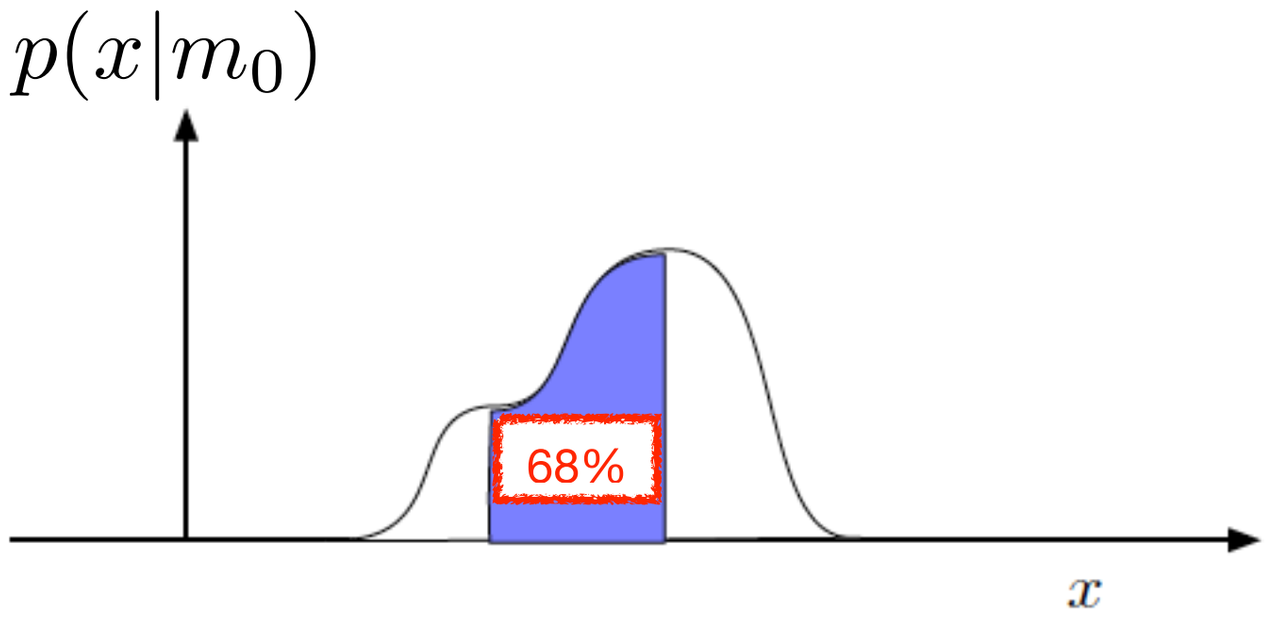
Neyman illustrated I

Take a specific value m_0 of the parameter



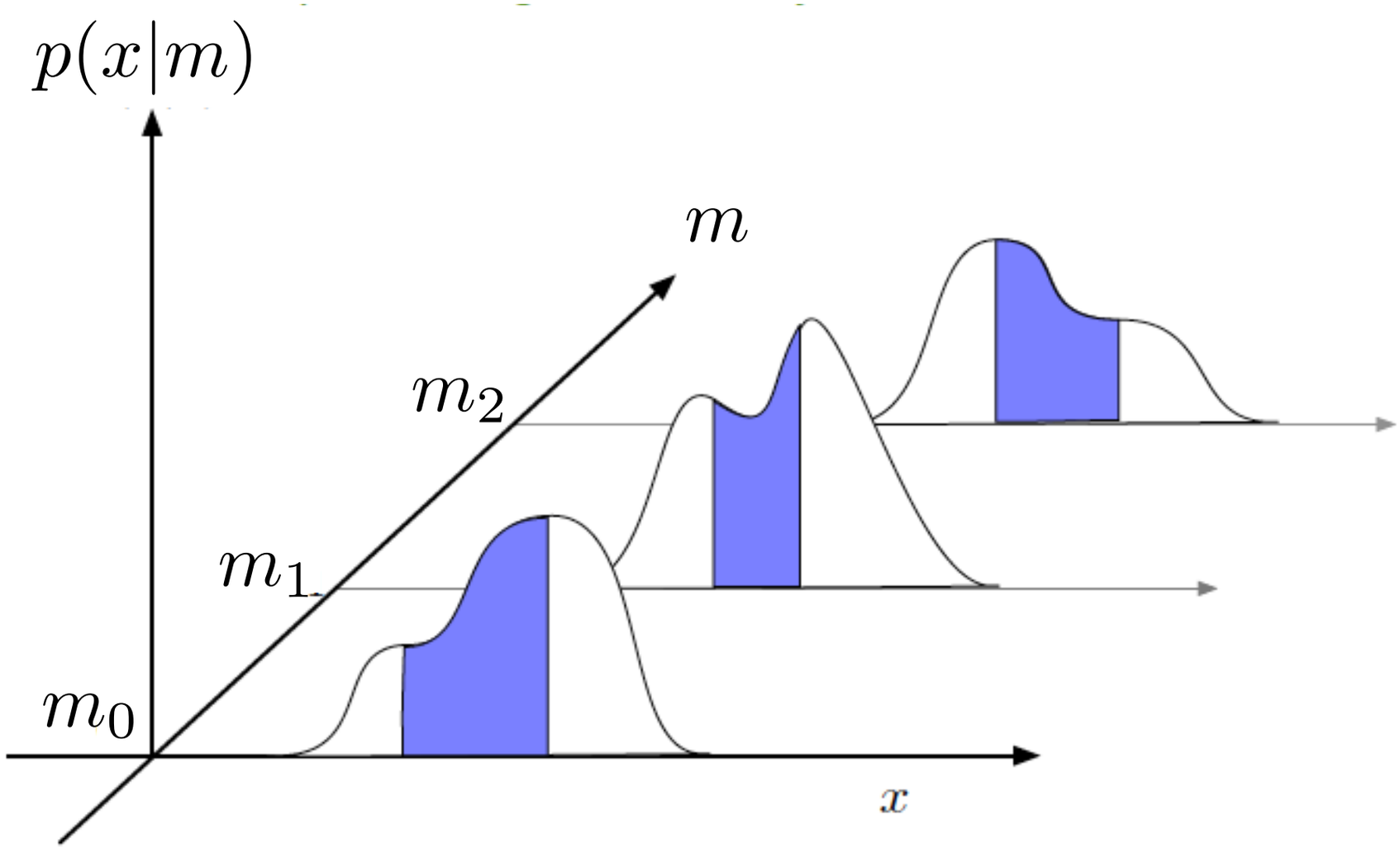
Neyman illustrated II

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



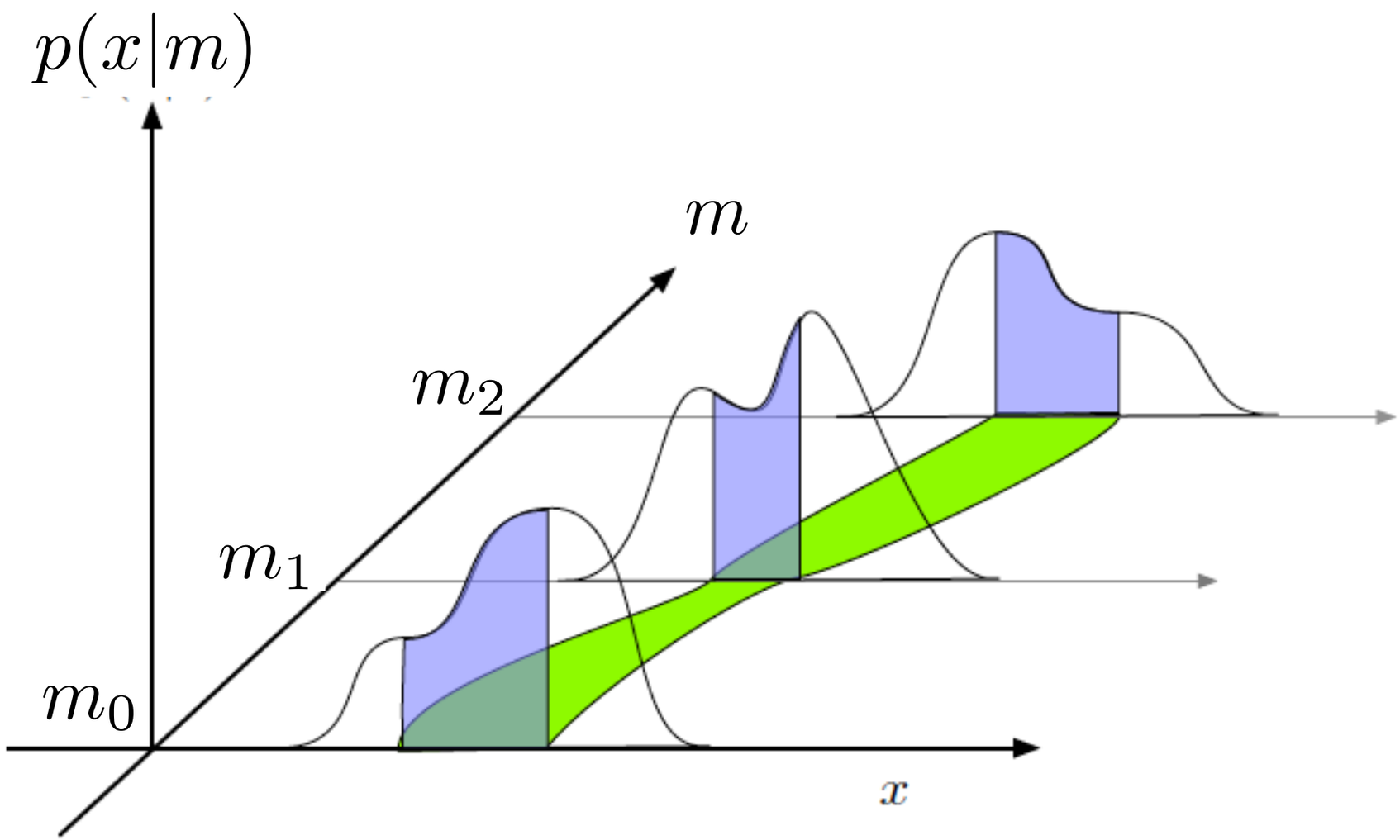
Neyman illustrated III

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



Neyman illustrated IV

This defines a confidence belt (aka acceptance region) for m.



The confidence belt consists of those values of parameter m for which the observed data values x are among the most probable to be observed.

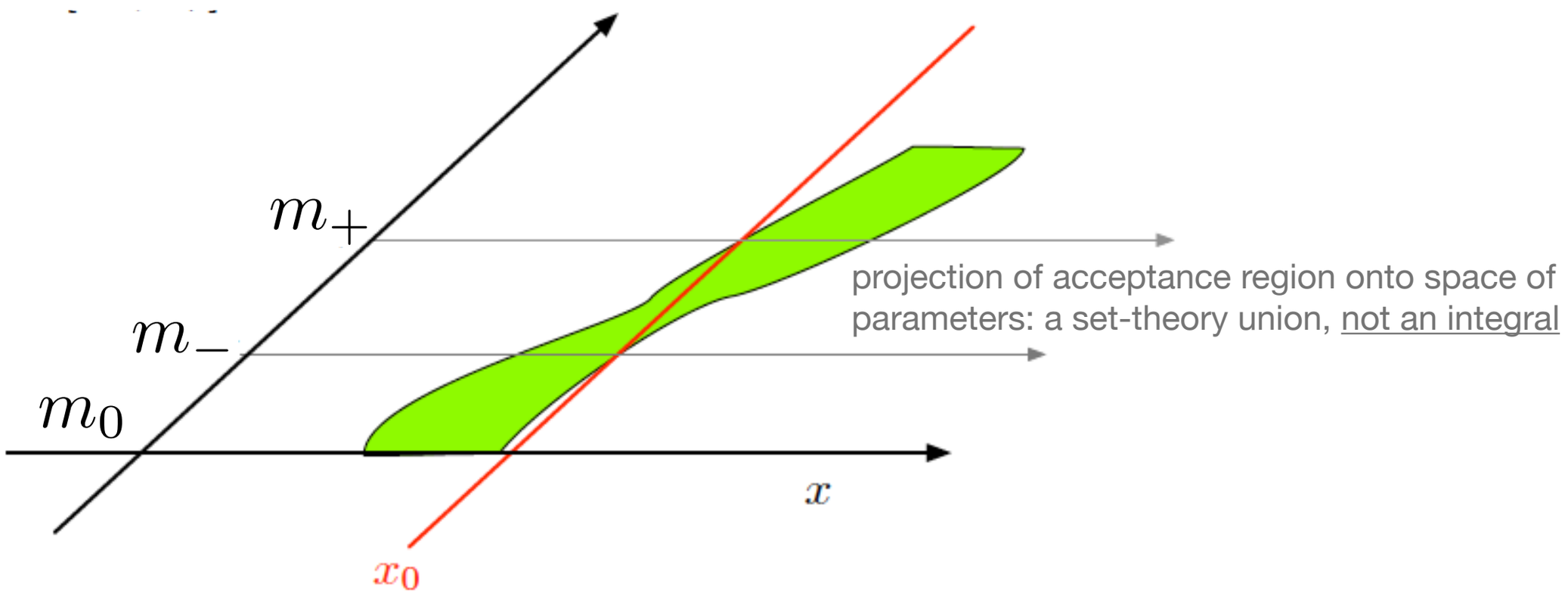
x and m don't need to have the same units, range, or dimensionality

Neyman illustrated V

Look at data and observe value x_0 —should intersect confidence belt.

Union of all m values for which x_0 intercepts the confidence belt defines the confidence interval $[m_-(x_0)\ m_+(x_0)]$ at the 68% CL for m.

The extremes of the interval are random variables (functions of data x)



In repeated experiments, the boundaries $[m_-(x)\ m_+(x)]$ will fluctuate, but 68% of them will contain the (unknown) true value of the parameter m

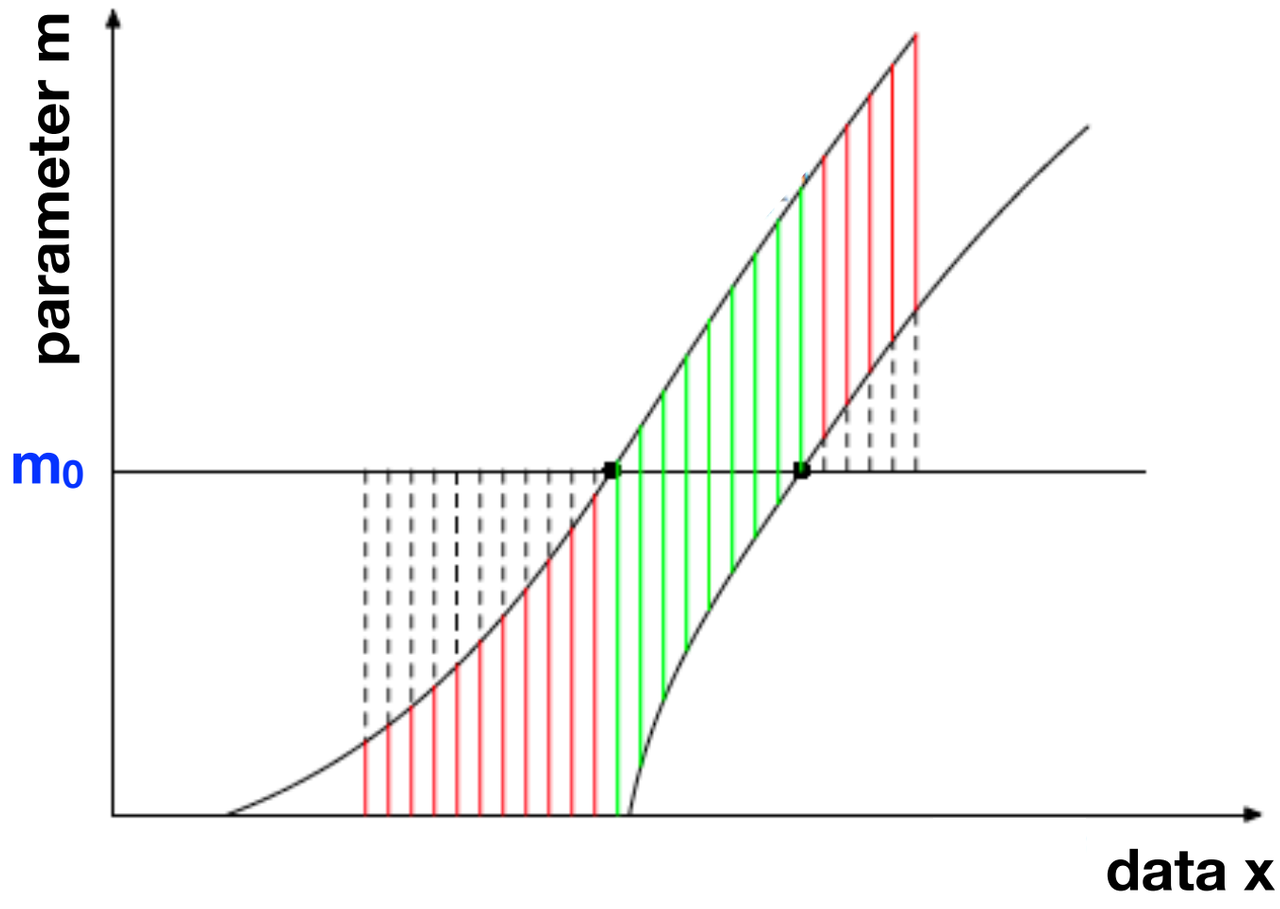
Neyman's "magic" explained

Suppose the true value is m_0

Depending on observation x, could pick either red of green intervals.

Red intervals don't include m_0 — green intervals do.

Since probability of observing data that yields a green interval is CL by construction, and green intervals contain m_0, then any observation yields an interval that include true value with probability CL



Coverage enforced by construction.

Result is expressed as "m is contained in the interval [a, b] at the 68% CL".

Not assigning a probability to true value m, which is fixed and unknown, but to the integral extremes

Toy example

homework

Identical bags of various classes. Each class contains a different fraction of white balls (class A has 1%, B has 5%, C has 50%, D has 95%, D has 99%).

Pick a bag, extract 5 balls, and find the bag class by setting a 95% CL upper limit on the true fraction of white balls.

white balls observed (this is "x")	True fraction of white balls (this is "m")				
	Class A = 1%	Class B = 5%	Class C = 50%	Class D = 95%	Class E = 99%
5	10^{-10}	$3*10^{-7}$	3.1%	77.4%	95.1%
4	$5*10^{-8}$	$3*10^{-5}$	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%	$3*10^{-5}$	$5*10^{-8}$
0	95.1%	77.4%	3.1%	$3*10^{-7}$	10^{-10}

Start constructing one-sided confidence band…

homework

For true value A, accumulate probability p(x|m) starting from high values of observations x, which are “extreme” for an upper limit, until accumulated probability is at least CL (chosen to be 95%)

True fraction of white balls (this is “m”)

white balls observed (this is “x”)

	Class A = 1%	Class B = 5%	Class C = 50%	Class D = 95%	Class E = 99%
5	10^{-10}	$3*10^{-7}$	3.1%	77.4%	95.1%
4	$5*10^{-8}$	$3*10^{-5}$	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%	$3*10^{-5}$	$5*10^{-8}$
0	95.1%	77.4%	3.1%	$3*10^{-7}$	10^{-10}

…keep constructing the confidence band…

homework

white balls observed (this is "x")	True fraction of white balls (this is "m")				
	Class A = 1%	Class B = 5%	Class C = 50%	Class D = 95%	Class E = 99%
5	10^{-10}	$3*10^{-7}$	3.1%	77.4%	95.1%
4	$5*10^{-8}$	$3*10^{-5}$	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%	$3*10^{-5}$	$5*10^{-8}$
0	95.1%	77.4%	3.1%	$3*10^{-7}$	10^{-10}

Confidence band is complete

homework

Green is the acceptance region, white the exclusion region

True fraction of white balls (this is “m”)

white balls observed (this is “x”)	Class A = 1%	Class B = 5%	Class C = 50%	Class D = 95%	Class E = 99%
5	10^{-10}	$3*10^{-7}$	3.1%	77.4%	95.1%
4	$5*10^{-8}$	$3*10^{-5}$	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%	$3*10^{-5}$	$5*10^{-8}$
0	95.1%	77.4%	3.1%	$3*10^{-7}$	10^{-10}

Now look at data

homework

Pick five balls from an unknown bag. Find only one white ball out of five. .

True fraction of white balls (this is “m”)

white balls observed (this is “x”)	Class A = 1%	Class B = 5%	Class C = 50%	Class D = 95%	Class E = 99%
5	10^{-10}	$3*10^{-7}$	3.1%	77.4%	95.1%
4	$5*10^{-8}$	$3*10^{-5}$	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%	$3*10^{-5}$	$5*10^{-8}$
0	95.1%	77.4%	3.1%	$3*10^{-7}$	10^{-10}

==> D and E class out of the confidence region.

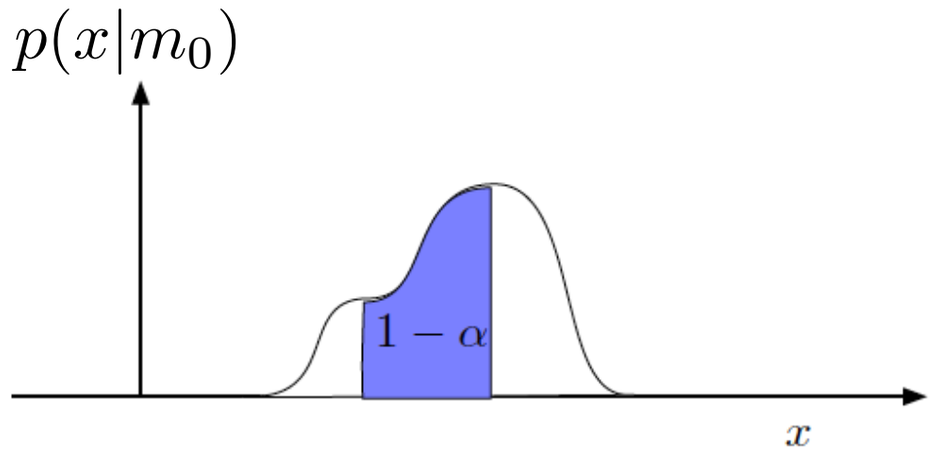
EXCLUDED at 95% CL

Which ordering?

Definition of acceptance range not unique.

Only constraint: p(x|m) outside it is <= 1-CL

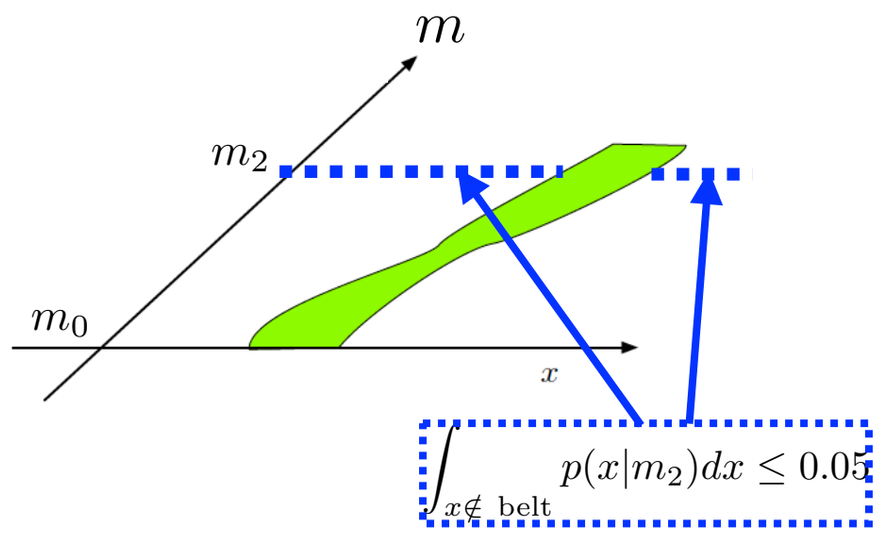
Criterion to choose acceptance region is *ordering rule* — *order* of accumulation of possible observations x until a CL amount of probability is accumulated.



Ordering may determines the precision of your inference

Decide prior to look at data otherwise could artificially exclude the result of the experiment

(Also, usually one wants a connected region and no “zebra” bands)

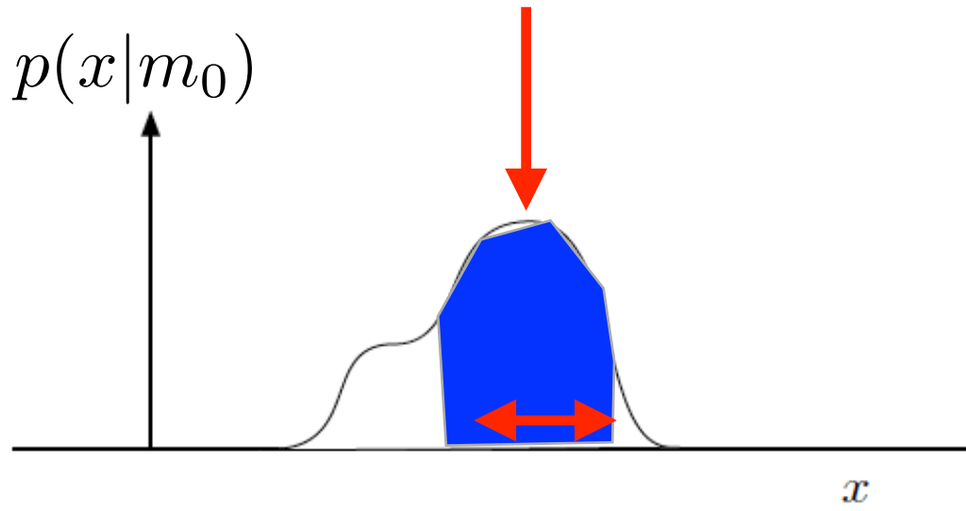


Probability ordering

Try to get the shortest possible interval, so that resulting confidence intervals are narrower (more precise results).

Probability ordering or Crow-Gardner ordering”

1. Choose one value for m, $m=m_0$, 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



Ill-defined: probability depends on the metric for observable x

Shortest interval in one metric isn’t shortest in others.

Issues — empty intervals

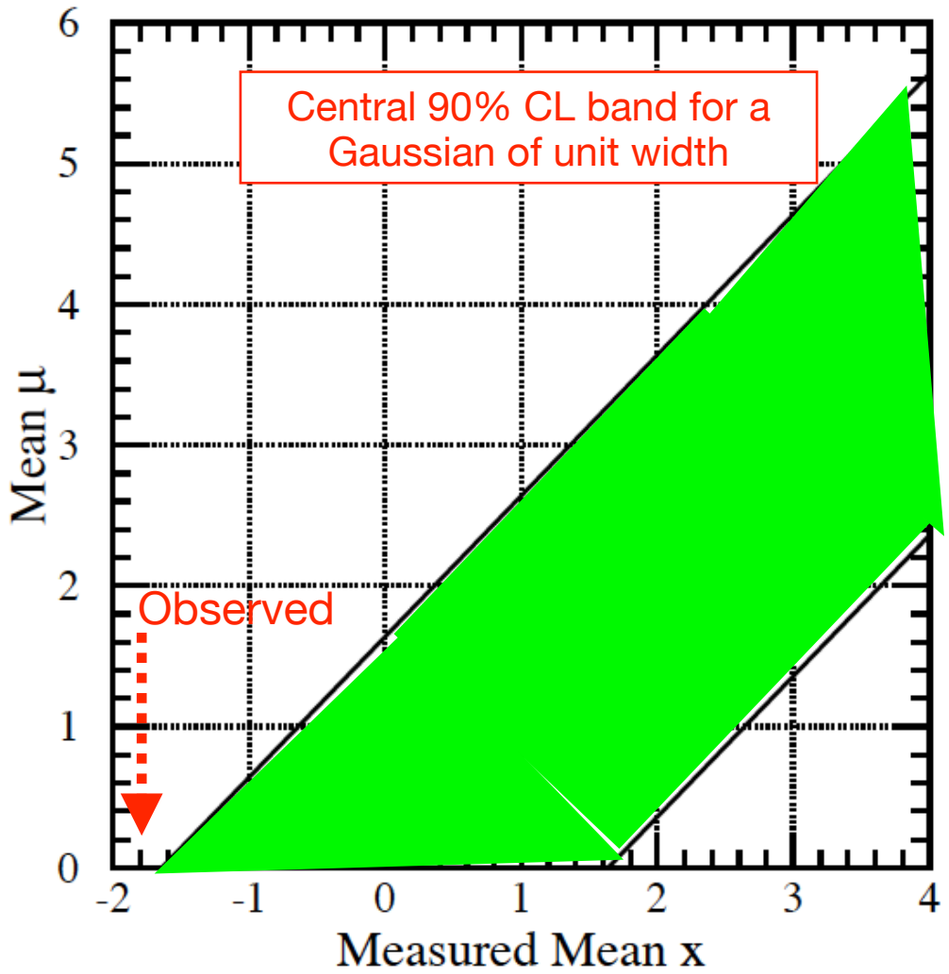
Long-standing inconsistencies found in simplistic ordering criteria.

Measurement with Gaussian resolution near physical boundary (e.g., neutrino mass square close to zero)

Observe a negative fluctuation

Resulting confidence regions is empty - that is no true value of parameter m could have generated the data I see….

Clearly a problem.



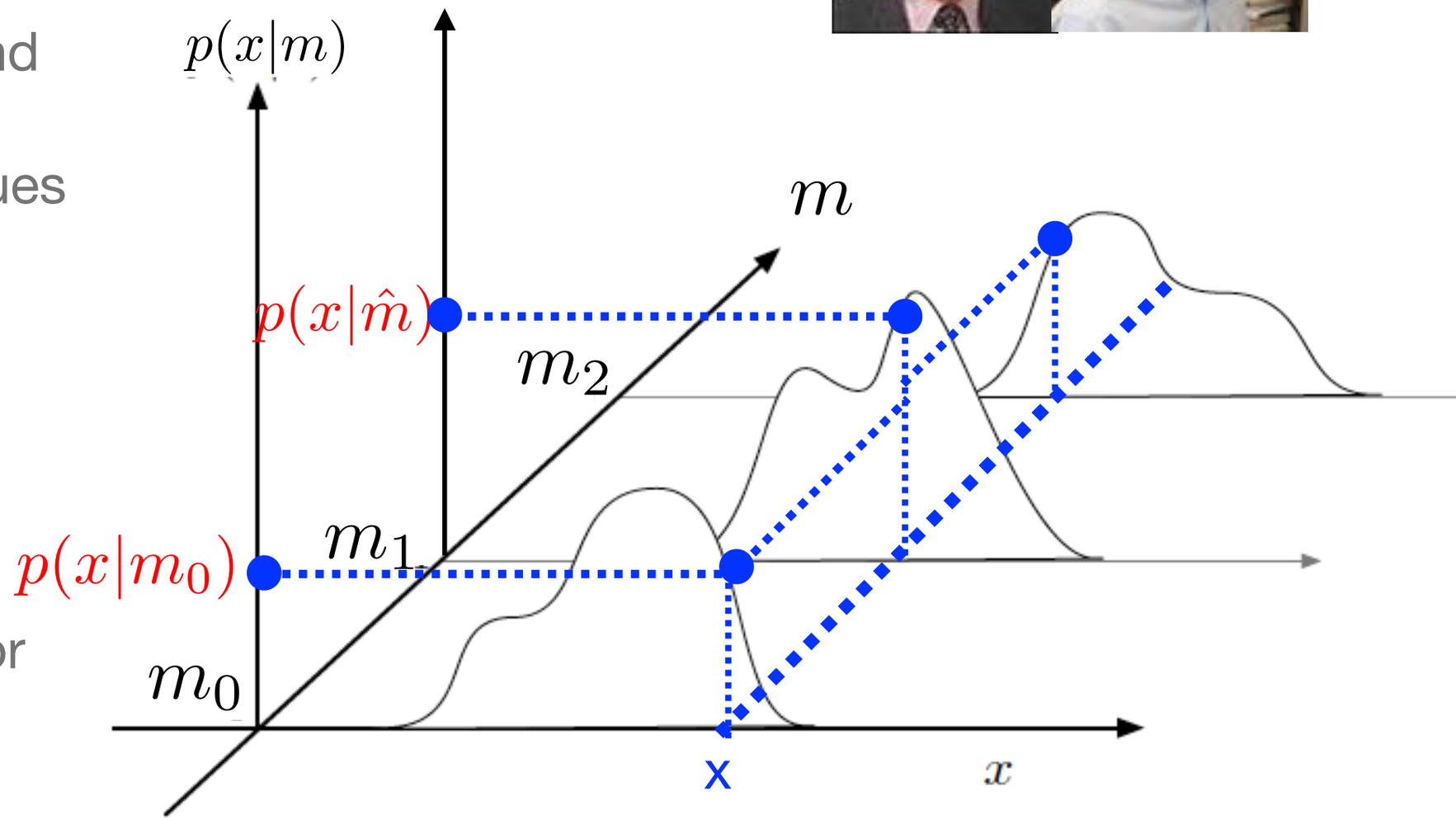
Likelihood-ratio ordering ("Feldman and Cousins")

Issues solved by adopting the likelihood-ratio ordering

When constructing the band for each value m_0 of the parameter accumulate values of x in decreasing order of

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

where $\hat{m}$ is the value that maximizes the likelihood for that x



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

Got your brain tangled? Try with Poisson.

homework

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 μ in the presence of known mean background $b = 3.0$. Here we find the acceptance interval for $\mu = 0.5$.

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

Observed count

L(μ =0.5) of observed count

μ̂ that maximizes L of observed count

L(μ̂) of observed count

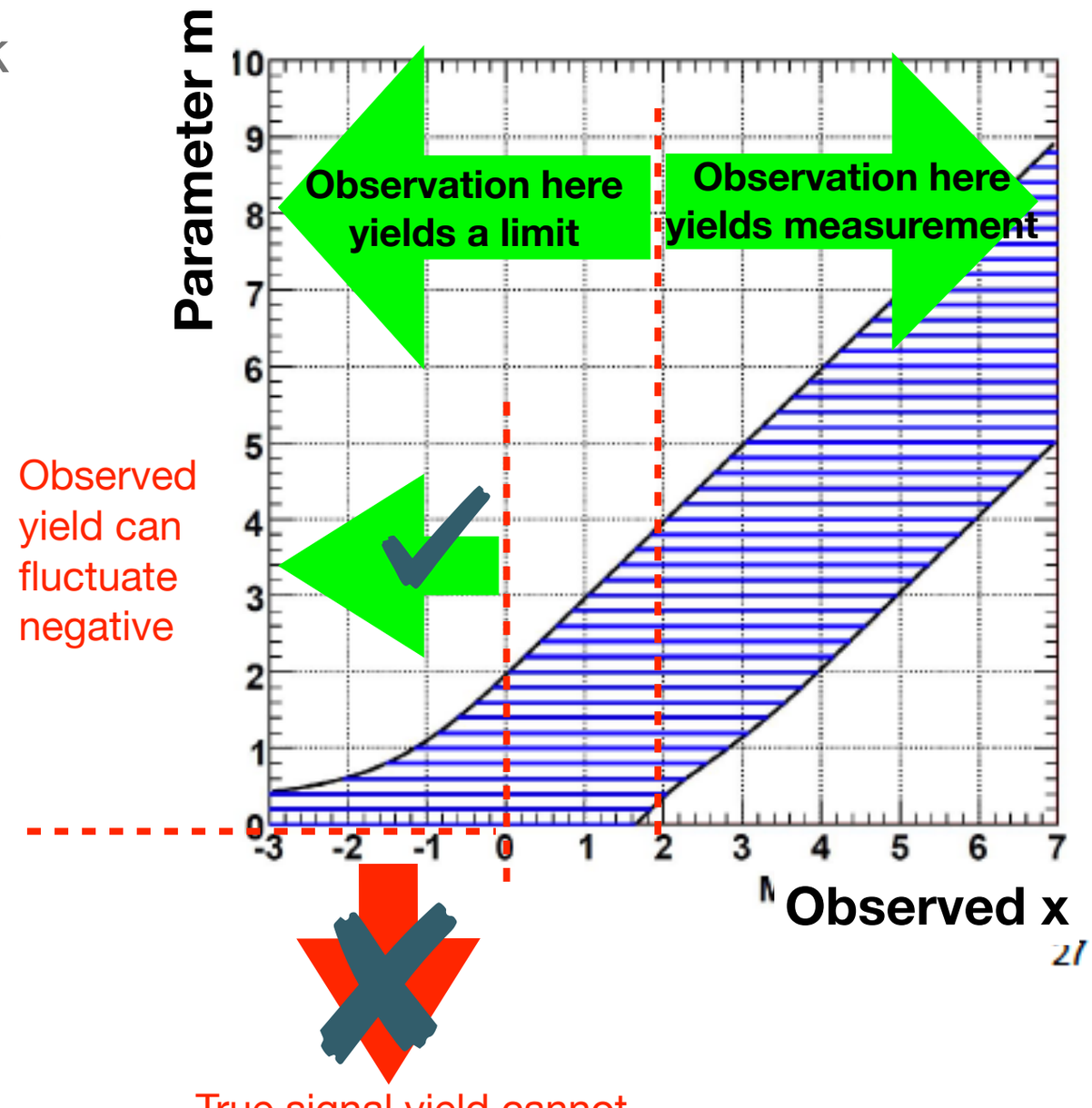
Likelihood ratio L(μtest = 0.5)/L(μ̂) (ordering score)

Can use this for Poisson http://stats.areppim.com/calc/calc_poisson.php

Natural transition from limit to point estimate

Important to keep distinct

- data x, which, due to resolution and bck fluctuations could fluctuate negative.
- parameter N_s, for which negative values do not exist in the model



Likelihood-ratio ordering

1. Choose one value for m, m_0 and generate simulated pseudodata accordingly.
2. For each simulated data set 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 (for that observation)
3. Rank all x values in decreasing order of likelihood ratio $LR=L_x(m_0)/L_x(\hat{m})$.
4. Accumulate probability p(x|m) starting from the x with higher LR until desired CL is reached.
5. Repeat for all m — now acceptance band is constructed

As the likelihood is metric-invariant so is the ratio of likelihoods.

Therefore LR-ordering preserves the metric and avoids empty regions (mostly, see sec. B.3 in https://arxiv.org/abs/hep-ex/9912048)

By far the most popular ordering in HEP.

Recommendation #1: FC is your go-to default option

Take likelihood-ratio ordering as default option in your work unless there are strong motivations against it.

Among standard frequentists inference procedures, FC has the most convenient, and statistically supported, properties

Recommendation #2: provide expected limits too

All limit setting procedures incur in counter-intuive results under certain conditions due to degraded statistical properties when samples are small.

For instance, for zero observations FC limits for Poisson plus background improve with background yield

To identify those cases, do report expected upper limits in addition to the observed ones

Expected upper limits are usually medians of limits obtained in many simulated background-only samples

Reliable assessment of the actual sensitivity of measurements procedure, regardless of the specific fluctuations happening in your data sample

Real life (i.e., your analysis)

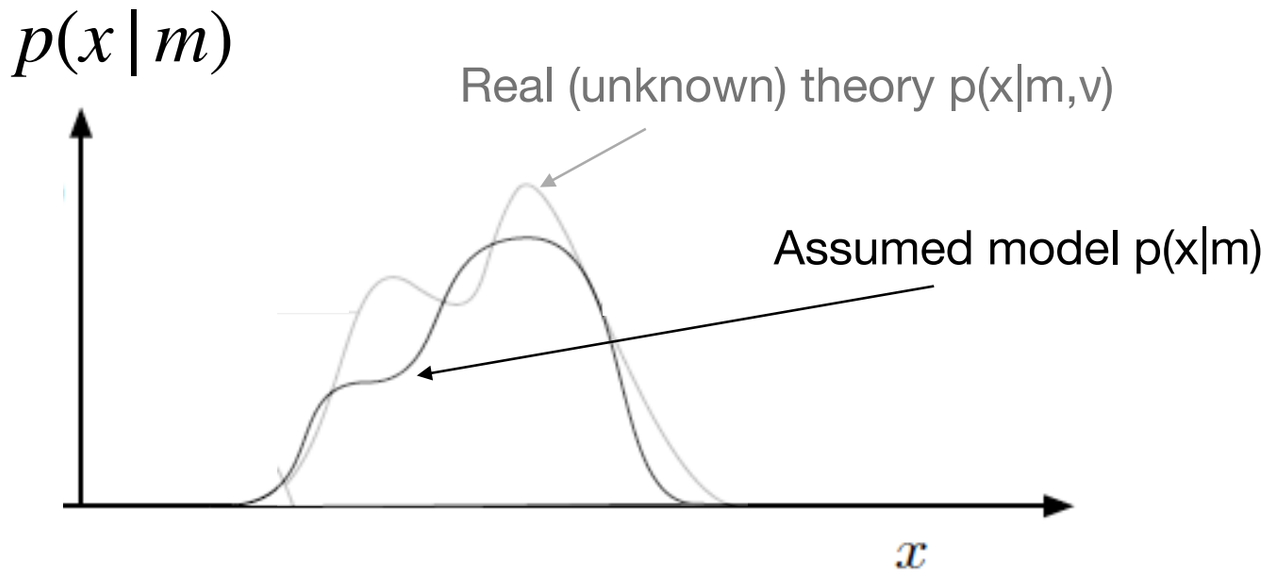
Systematic uncertainties

What systematic uncertainty is

Parametrization of the differences between our model and reality

p(x|m) is an approximation ofreal unknown theory p(x|m,v).

Parametrize difference with dependence on unknown parameters v



Not only one does not know which data x will be observed for a true value m.
One does not even know precisely the probability for each possible x.

In general both the value of v and its functional dependence p(x|v) are unknown

That's why it's generally wrong to assume systematic uncertainties Gaussian

Bayesian approach

Assume prior **p(v)** for the nuisance parameters and integrate ("marginalize") the product of that prior by the likelihood over **v.**

Revert to p(x|m) that no longer depends on the nuisance parameters

$$p(x\,|\,m) = \int_{\nu} p(x\,|\,m,\nu)p(\nu)d\nu$$

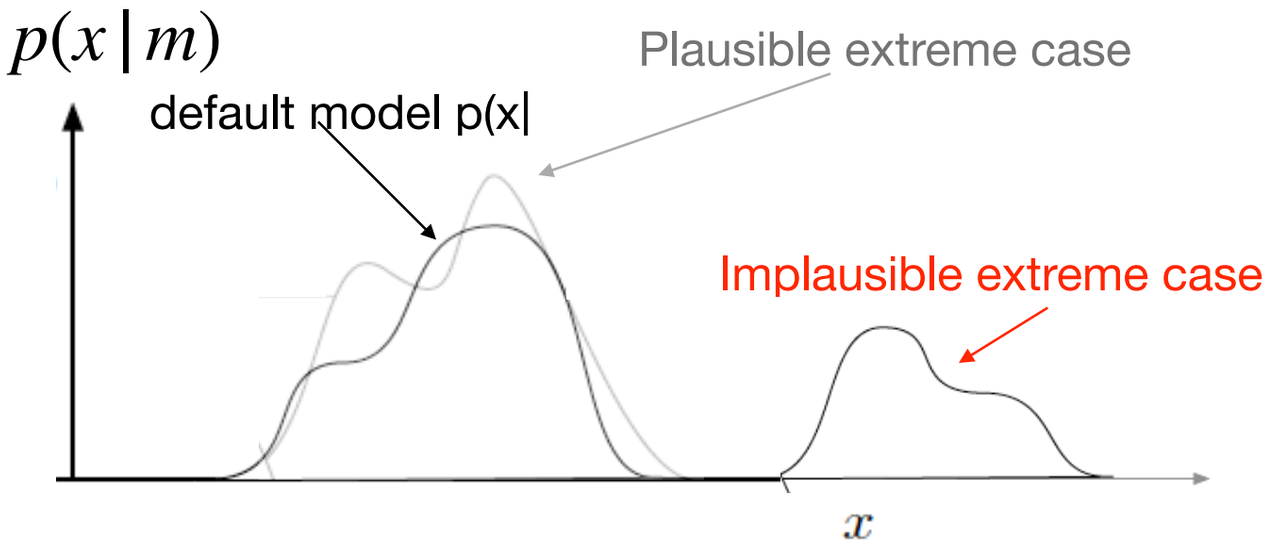
and then proceed with Bayesian inference as shown before.

Results may depend on priors and won't guarantee coverage but are valid Bayesian results - especially if systenatic sources are few

For many systematic sources, the impact of priors in driving the results explodes with dimensionality rendering the inference problematic

Frequentist step 1: simulating alternative models

Identify a few `extreme configurations' for v, (v', v",v"",…) that bracket all plausible configurations of the unknown parameter.



If identifiying extreme cases is not obvious, sample uniformly the nuisance parameter space to define alternative "Universes" v',v", v"", etc…

Simulate sets of data, each from an alternative Universe using v', v",… as true values for v say

1000 toys from $p(x|m_0, v')$, 1000 from $p(x|m_0, v")$, 1000 from $p(x|m_1, v')$, 1000 from $p(x|m_1, v")$…. — where m_i are possible true values for the parameters of interest.

Frequentist step 2: constructing acceptance band

Replace likelihood with lower-dimensional structure, the profile-likelihood, and base inference on that.

Profile likelihood: likelihood maximized with respect to a subset of its variables (usually the nuisance parameters *ν*) and replacing their maximized values ν̂ inside it:

$$L(m_1, m_2, \ldots, m_n, \nu_1, \nu_2, \ldots, \nu_m \,|\, x) \Rightarrow L_p(m_1, m_2, \ldots, m_n \,|\, \hat{\nu}_1, \hat{\nu}_2, \ldots, \hat{\nu}_m, x)$$

Gain convenience due to reduced dimensionality. But lose rigor: profile likelihood is not a likelihood nor it has its mathematical properties.

However, it approximates sufficiently likelihood properties in many problems thus offering a reliable inference instrument

Fit each toy with profiled likelihood and plot the distributions of results m̂ for each ensemble.

Construct the confidence band by using the Universe that yields the lowest CL. This will make the results to have coverage *regardless of the true value of ν.*

Example from Poisson plus uncertain background

In the signal region, we observe n_{on} events from a Poisson process with mean $\mu_s + \mu_b$

In the sideband region, we observe n_{off} events

The unknown mean number of signal events μ_s is the parameter of interest

We only consider the systematic uncertainty due to the uncertainty in background expectation (treatment is straightforwardly extensible to multiple nuisance pars.)

The mean number of background events μ_b is the sole nuisance parameter, estimated to be $\mu^*_b \pm \sigma^*_b$ from a (scaled) fit to our sideband events.

We know that the likelihood for μ_s is Poisson

We need to know what the likelihood for μ^*_b *is.* Since that results from a fit, it might be ~Gaussian, but should look at the distribution of the estimator μ^*_b using toys or bootstrap

Let's suppose it is Gaussian (any other shape could be replaced in what follows).

Then the likelihood for our problem is

$$P(n_{on}\,|\,\mu_s,\mu_b) = L_G(\mu_s,\mu_b;n_{on}) = \frac{(\mu_s+\mu_b)^{n_{on}}}{n_{on}!}e^{-(\mu_s+\mu_b)}\frac{1}{\sigma_b\sqrt{2\pi}}\exp\left(-\frac{(\mu^*_b-\mu_b)^2}{2\sigma_b{}^{*2}}\right)$$

(assume the estimate σ^*_b to be a good approximation of its unknown true value σ_b) 1

Example from Poisson plus uncertain background

We now need to construct confidence region based on the Feldman Cousins procedure applied to the above likelihood - *profiled* against the nuisance parameter μ_b.

Means constructing region by ordering observations according to the profile-likelihood ratio

$$\Lambda(\mu_s; n_{on}) = \frac{L_G(\mu_s, \hat{\hat{\mu}}_b(\mu_s); n_{on})}{L_G(\hat{\mu}_s, \hat{\mu}_b; n_{on})}$$

where

- $\hat{\mu}_s$ and $\hat{\mu}_b$ are the maximum-likelihood estimates of the numbers of signal and background events (μ_s and μ_b are floating)
- $\hat{\hat{\mu}}_b$ is the result of maximizing the likelihood with respect to the number of background events only, as a function of μ_s. (μ_b only is floating)
- n_{on} is the observation in the given sample

Example from Poisson plus uncertain background

Choose a plausible test true value for the signal mean μ_s^0

Choose the worse-case true value for the background μ_b^w

Need to "horizontally"construct the band: accumulate values of observed total yield n_{on}^i in decreasing order of $\Lambda(\mu_s; n_{on})$ until the sum of the $P(n_{on}^i \mid \mu_s^0, \mu_b)$ pieces corresponding to the included n_{on}^i values is $\sum_i P(n_{on}^i \mid \mu_s^0, \hat{\hat{\mu}}_b)$ >= 90%

In practice, for each n_{on}^i need to

- calculate the pdf $P(n_{on}^i \mid \mu_s^0, \hat{\hat{\mu}}_b)$
- calculate $L_G(\hat{\mu}_s, \hat{\mu}_b; n_{on}^i)$ - likelihood, at given n_{on}^i, maximized with respect to both pars
- calculate $L_G(\mu_s^0, \hat{\hat{\mu}}_b(\mu_s^0); n_{on}^i)$ - as above but maximized only wrt μ_b with μ_s fixed at μ_s^0
- calculate their ratio $\Lambda(\mu_s^0; n_{on}^i)$
- Rank n_{on}^i values in decreasing order of $\Lambda(\mu_s^0; n_{on}^i)$ until the sum of $P(n_{on}^i \mid \mu_s^0, \hat{\hat{\mu}}_b)$ reaches 90%

Repeat the whole procedure for several relevant values of μ_s^i

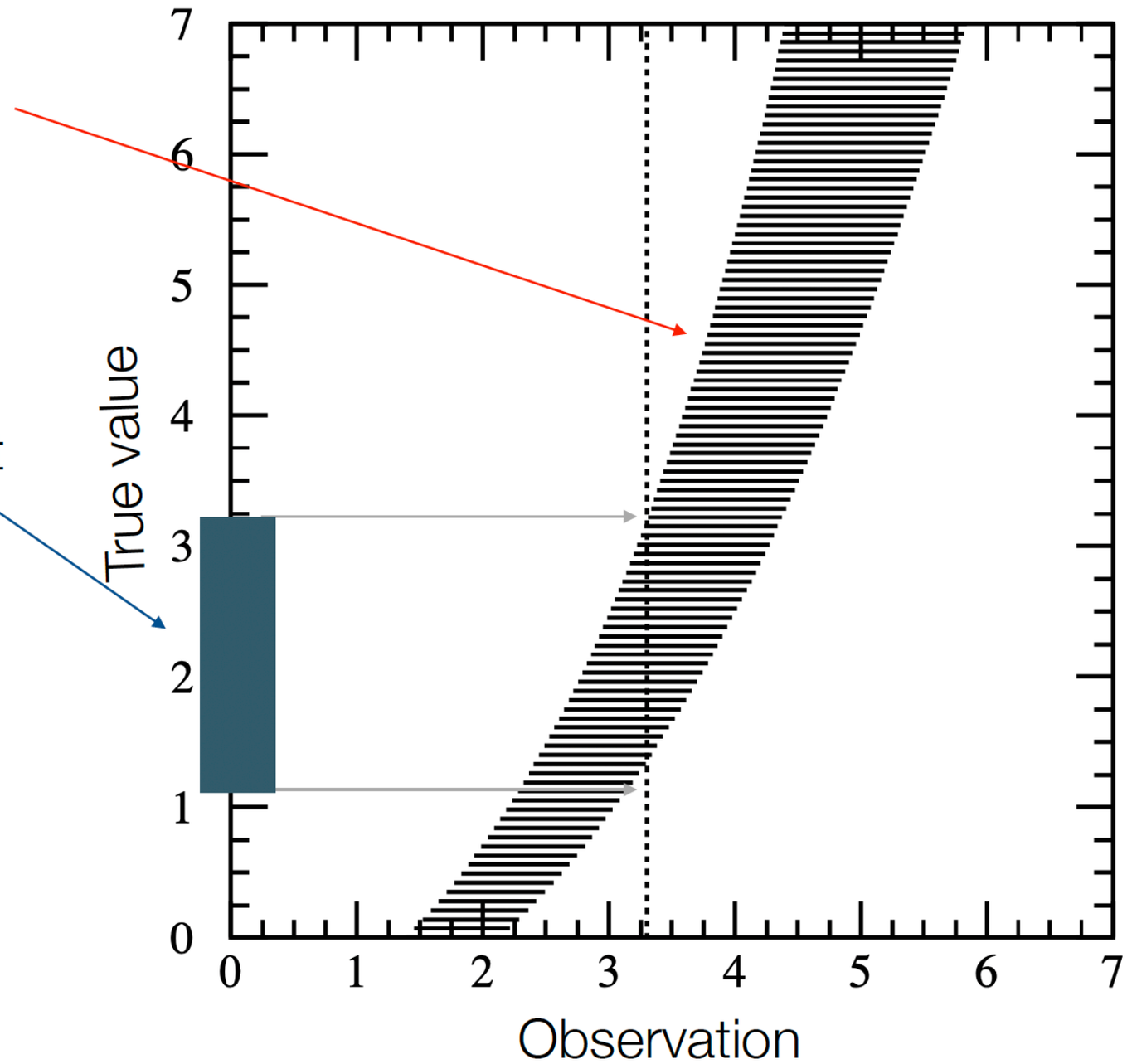
NB: "calculate likelihood" above means either do it analytically (when possible) or numerically (i.e., fit) otherwise

Example from Poisson plus uncertain background

Each of the bulleted sequences above determines one tiny horizontal line here

Repeating for multiple true values μ_s^i determines the full acceptance band

Then, projecting onto the true value the intersection of the observation in data with acceptance band, will give the result



Hybrid approaches

Some mix Bayesian and frequentist approaches, especially when trying to include systematic uncertainties in exclusion limits.

These approaches usually involve “folding in” the systematic uncertainty along with the statistical one first, and then determining limits using the total uncertainty

The folding can happen by either

- convolving the likelihood with a Gaussian of width equal to systematic uncert.;
- summing in quadrature the statistical and systematic uncertainty;
- marginalizing the likelihood **only** with respect to the nuisance parameters (as in slide 18) and then treat the resulting posterior as a proper likelihood for usage in standard frequentist inference (Cousins-Highland, NIM A320, 331 (1992), `RooStats::HybridCalculator`)

I recommend against these as mixed treatment obfuscates a proper interpretation of the final results - which are likely to be improper both from a frequentist and a Bayesian standpoints

Combining limits

Can quickly become a mess.

No consensus on statistically proper procedure to combine limits.

Much better to combine point-estimates (central values with uncertainties) and then extract limit from combined result.

That's why whenever quoting limits is good practice to report also central value with uncertainties, so that combination will be straightforward.

Final pontification

Try to discover signals, not to exclude them

Try to discover signals not to exclude them

Try to discover signals not to exclude them

…but if you really don't have signal, then make sure your "exclusion" is proper

Refrain from pre-cooked sw packages, invest time in coding your model, producing your toys, and construct your inference from scratch — instructive, transparent, flexible.

Be Bayesian all the way. Or frequentist all the way. Do not mix approaches.

If you like frequentist, Feldman Cousins is the best go-to option by default

If you like Bayesian, consider *temporary apostasy* if dimensions are many

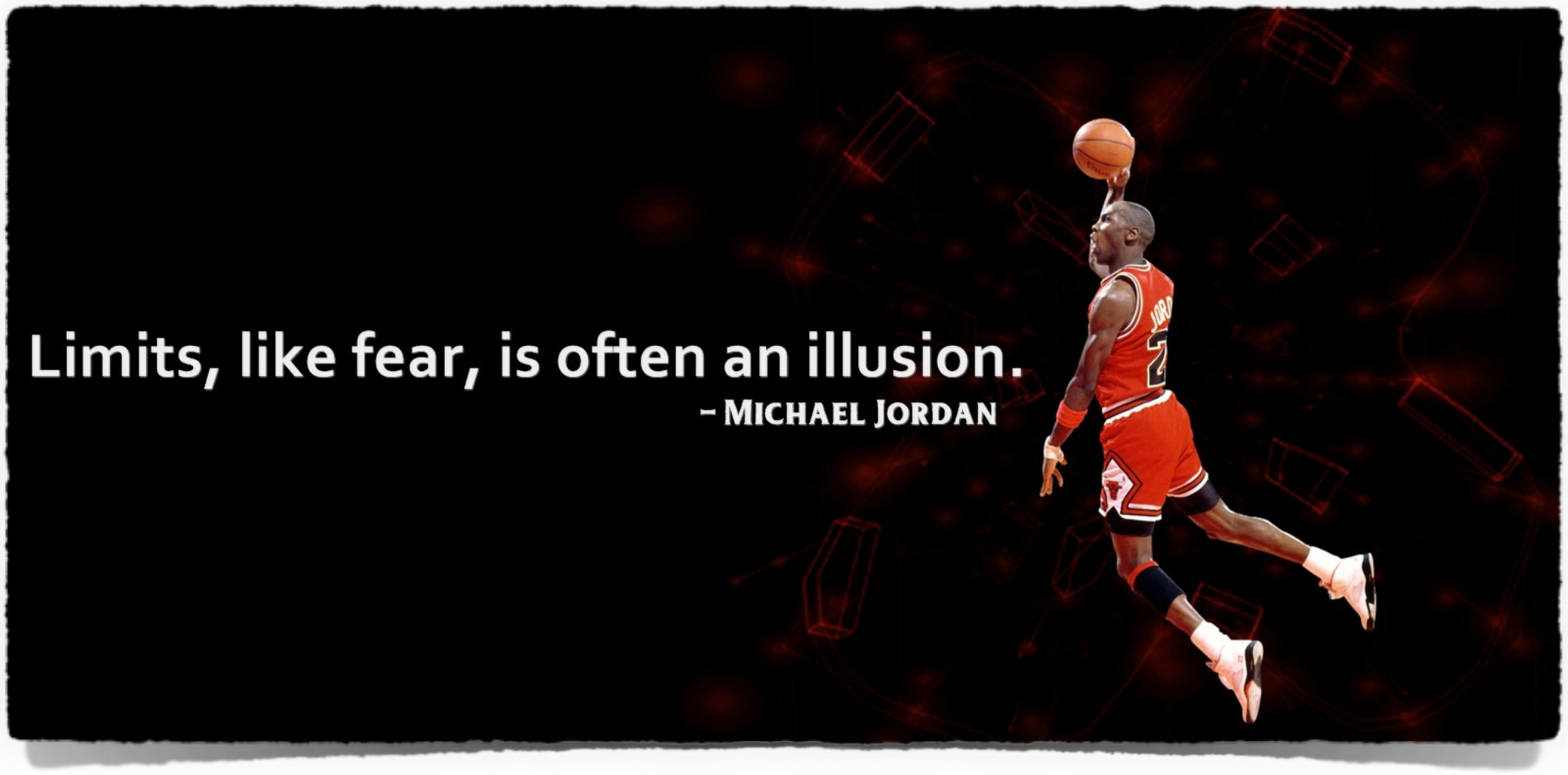
In any case check for coverage (frequentist) or for prior sensitivity (Bayesian)

Report expected limits for unbiased sensitivity assessments

Report central values with uncertainties to facilitate combination

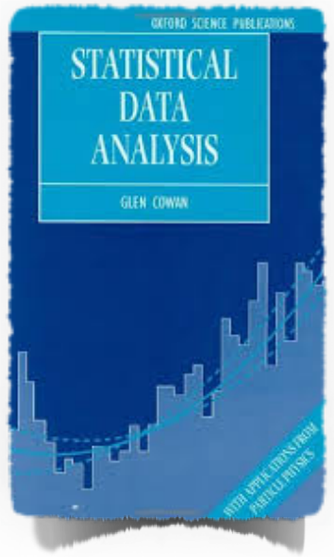
Exclusion is a poor word: suggests boolean (yes/no) logical condition, while here it is associated with probability. For ~1000 limits in the history of PDG there have to be ~100 cases in which the true value was found to lie in the "excluded" range - is this the case?

Thanks



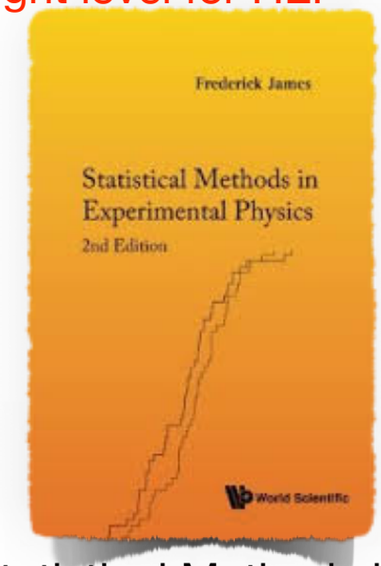
Further readings

- Good starting point



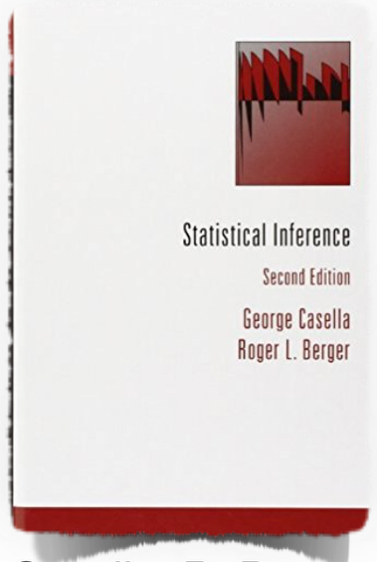
G. Cowan,"Statistical data analysis"

- Very good book at the right level for HEP



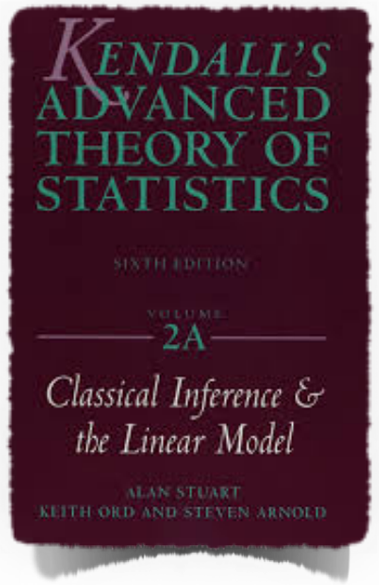
F. James, "Statistical Methods in Experimental Physics, data analysis"

- Advanced book



G. Casella, R. Berger, "Statistical Inference

- Ultimate bible



A. Stuart, et al "Kendall's Advanced Theory of Statistics Vol 2A"

- Statistics@ http://hcpss.web.cern.ch/hcpss/ (Excellent lectures by K. Cranmer, G. Cowan, B. Cousins et al. Some are video-recorded). Similar expertise level to the lectures given here.
- Lectures from Glen Cowan's page https://www.pp.rhul.ac.uk/~cowan/ Similar or more basic expertise level
- Terascale Stat School (especially 2015 F. James' lectures) https://indico.desy.de/conferenceDisplay.py?confId=11244 More advanced level.
- T. Junk's lectures from www-cdf.fnal.gov/~trj/ Similar expertise level
- L. Lyons lectures: https://indico.cern.ch/event/431038/ Similar or more basic expertise level
- Notes from CDF's Statistics Committee public page https://www-cdf.fnal.gov/physics/statistics/ Basic to advanced
- B. Cousins: find his (CMS-restricted) "Statistics in Theory - prelude to Statistics in Practice" lectures. Look at his statistics papers on inspire and the references he reccommends. Advanced
- Proceedings/docs from the PHYSTAT conferences and workshops, linked from phystat.org Advanced