Statistics for physics — part IV

Diego Tonelli — INFN Trieste
diego.tonelli@cern.ch

April 28, 2022

Future Flavours school 2022

Beyond point estimation

Issues with point estimates

Important details to check while quoting point estimates (that is: central value ± uncertainty).

- Is the estimator biased? How should I estimate/treat biases?
- Is the procedure able to guarantee the desired coverage?

E.g., Estimation of an efficiency. Use observed number n_{on} of successes out of n_{tot} trials to construct an estimator of the efficiency $\hat{\rho} = n_{on}/n_{tot}$

Observe 3 successes on 10 trials — what is our efficiency and its uncertainty?

Efficiency is binomial quantity ("either you succeed or not"). Usually replace $\hat{\rho} = 0.30$ into $\hat{\sigma} = \sqrt{n_{tot}\,\hat{\rho}(1-\hat{\rho})}$ and obtain the interval $[\rho_1, \rho_2] = \hat{\rho} \pm \hat{\sigma}$ This is not a proper confidence interval.

Flaw is manifest when $n_{on} = n_{tot}$ or $n_{on} = 0$, as the estimate would have zero uncertainty!

Move to a more general strategy for quoting results, which accounts for biases and ensure coverage — construction of confidence intervals.

Confidence intervals

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

To specify "extreme", need an ordering.

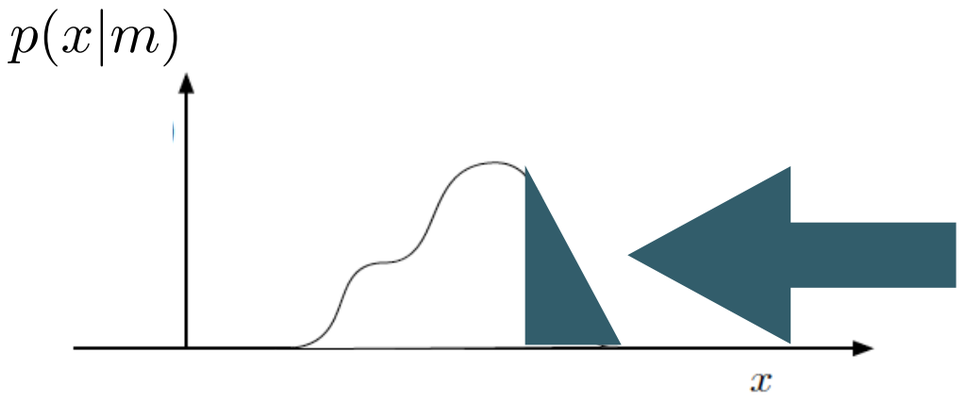
Rank values of x for each possible value of m. High rank means not extreme (likely to be included in the interval). Low rank means extreme (likely to be outside of the interval).

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%…

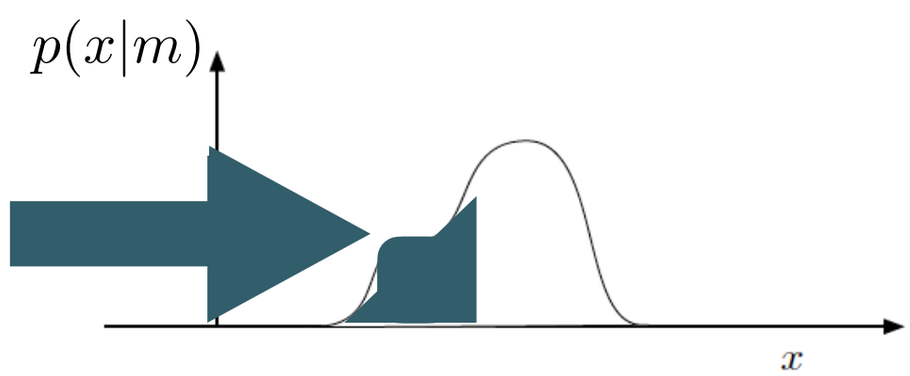
Given an ordering and a CL, the confidence interval $[m_1, m_2]$ includes those values of m for which x_0 are not "extreme" at the chosen CL

E.g, the 68% CL interval $[m_1, m_2]$ includes the values of m for which the observed data x_0 belongs to the least extreme 68% values of x

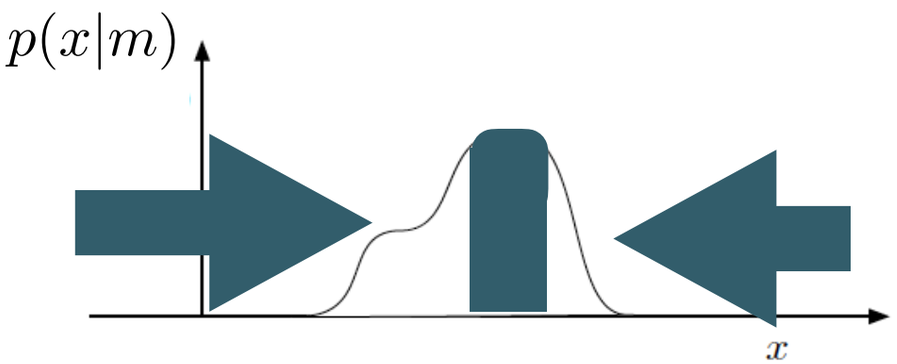
One-sided, two-sided.



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 (interval estimate of m)

(simplified interpretation that applies only when x is one-dimensional and p(x|m) is such that higher m imply higher average x). The confidence level CL is usually chosen to match the standard thresholds 68.3% (1σ) 95.5% (2σ) etc.

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

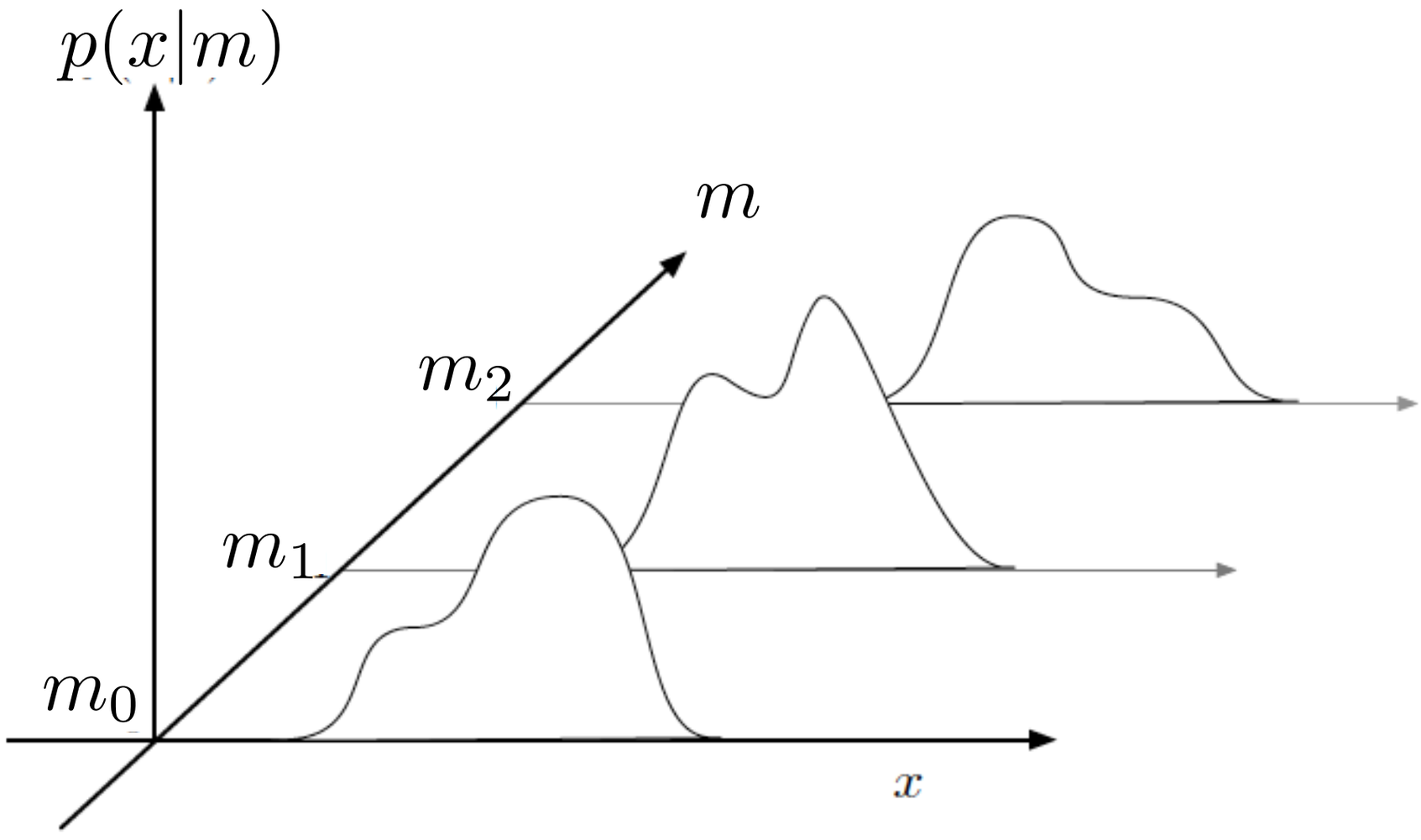
By J. NEYMAN

Reader in Statistics, University College, London

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

Neyman construction illustrated

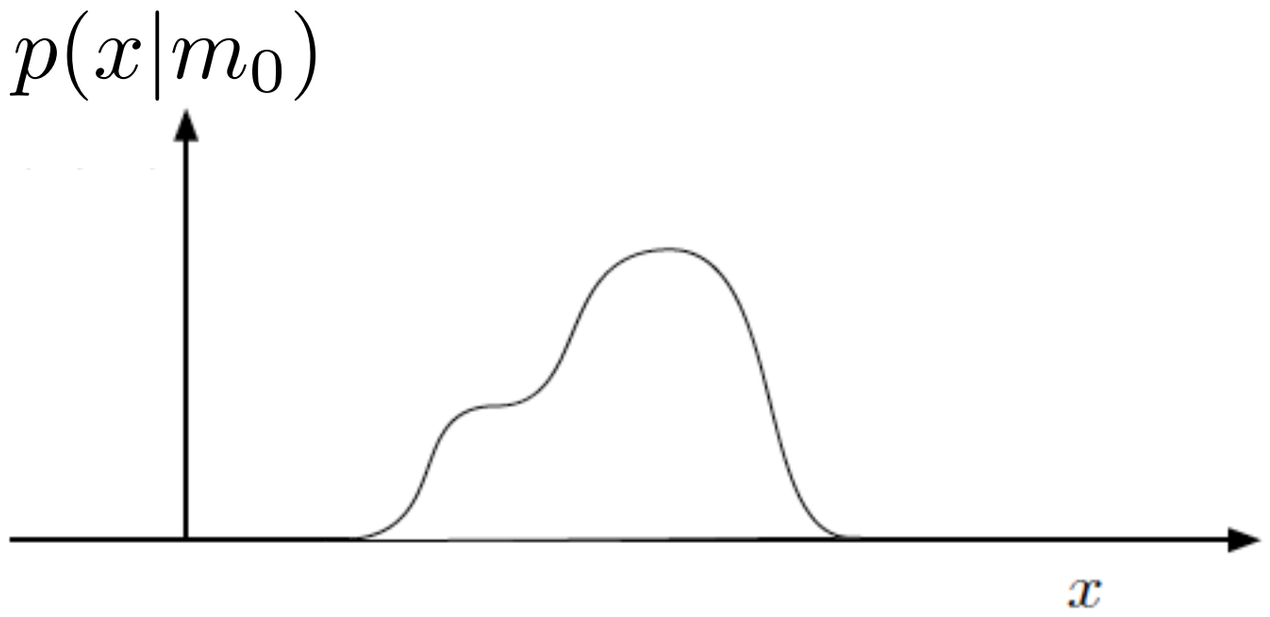
Prior to looking at data, for each possible true value of parameter m, consider p(x|m). Its shape can vary as a function of m.



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

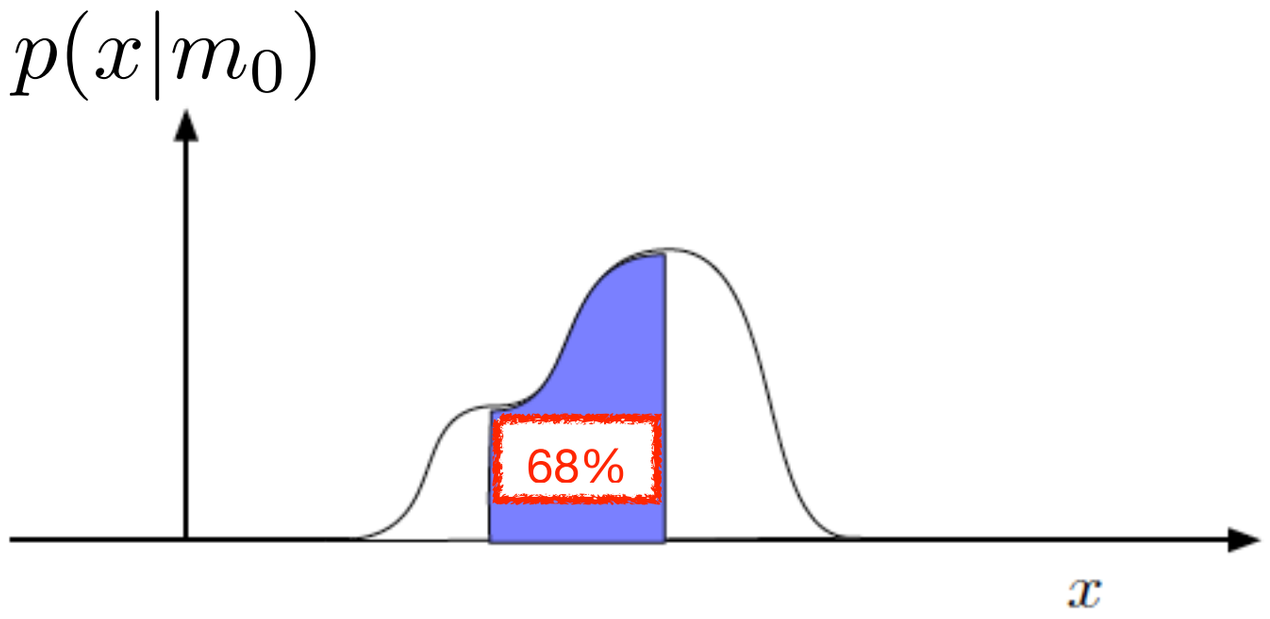
Neyman illustrated I

Take a specific value m_0 of the parameter



Neyman illustrated II

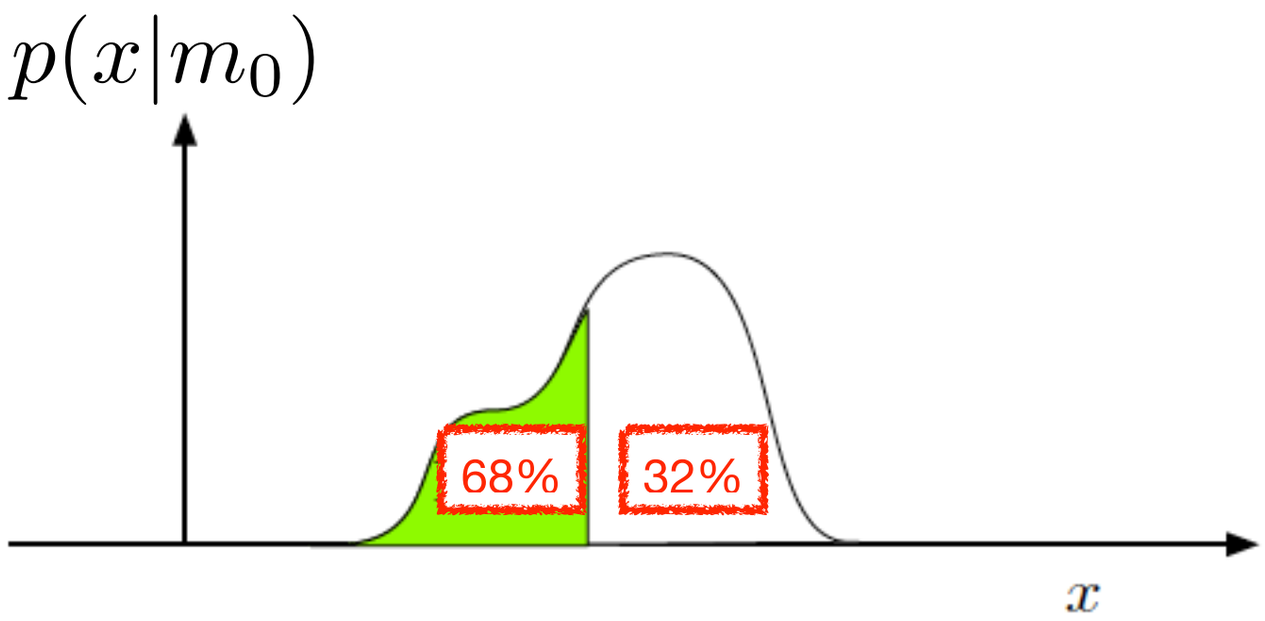
Use $p(x|m_0)$ to define an acceptance range in x, such that $p(x \in \text{range} \mid m_0) = 68\%$.



Neyman illustrated III

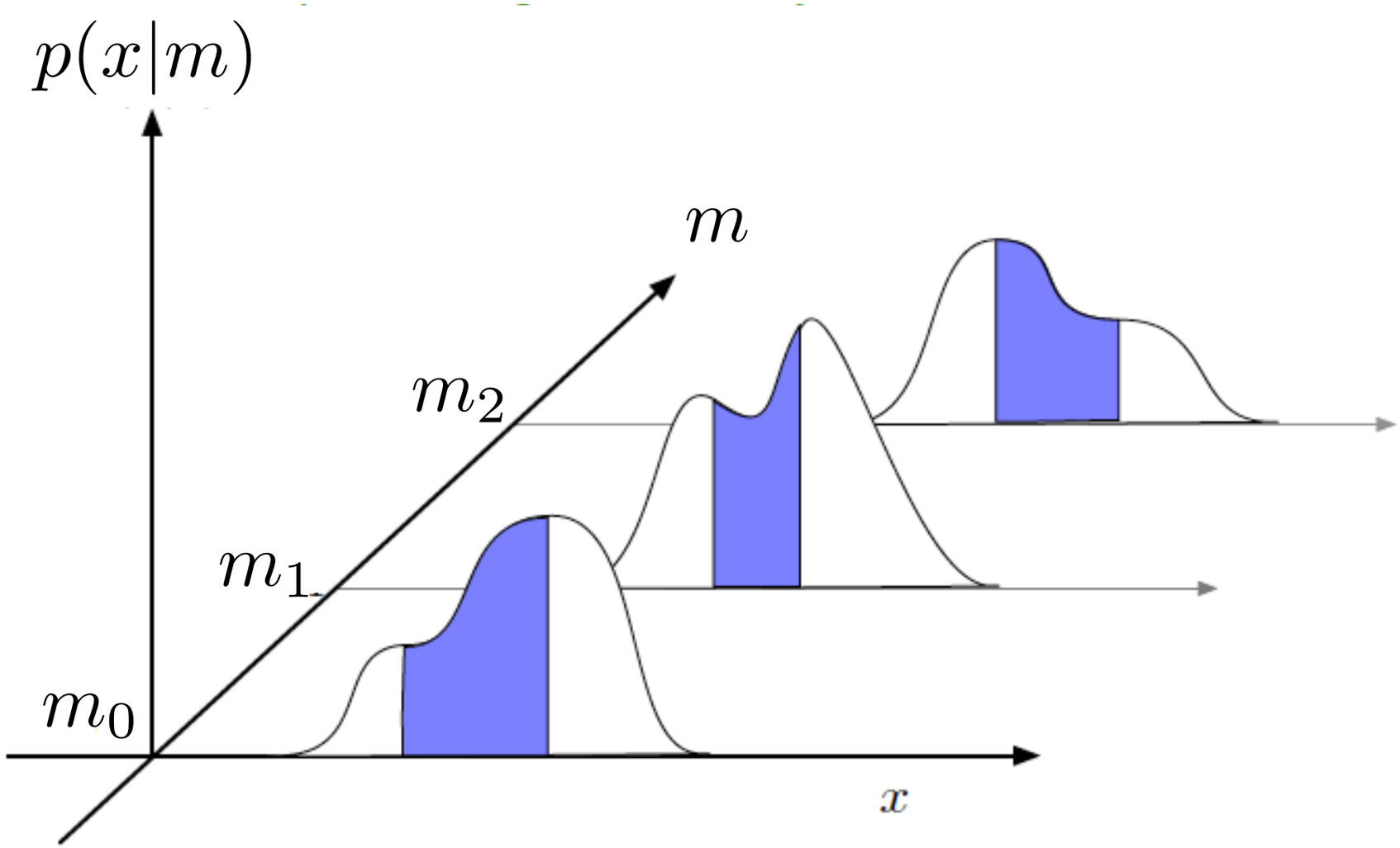
The definition of the acceptance range is not unique

The criterion to choose of the region is 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.



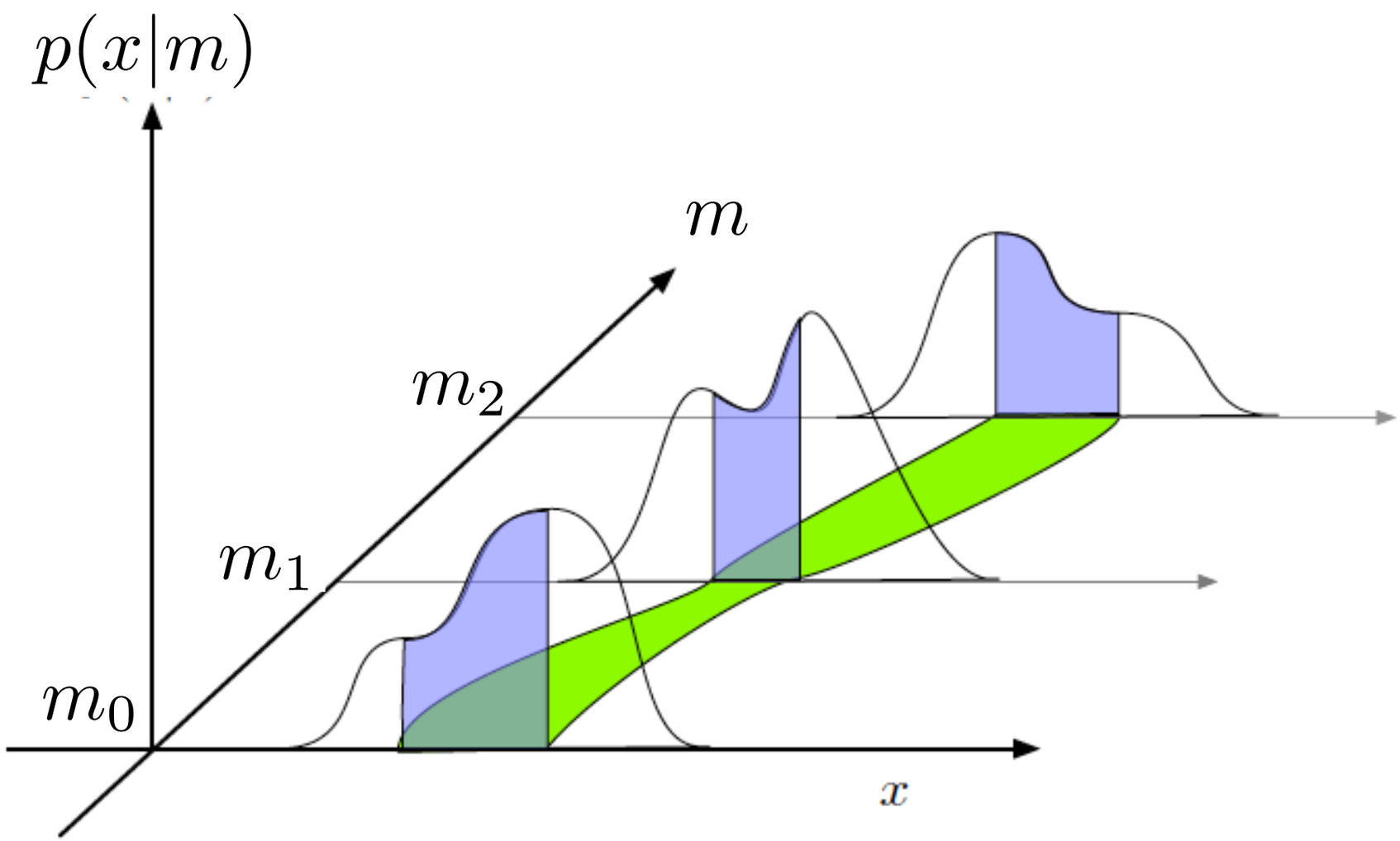
Neyman illustrated V

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



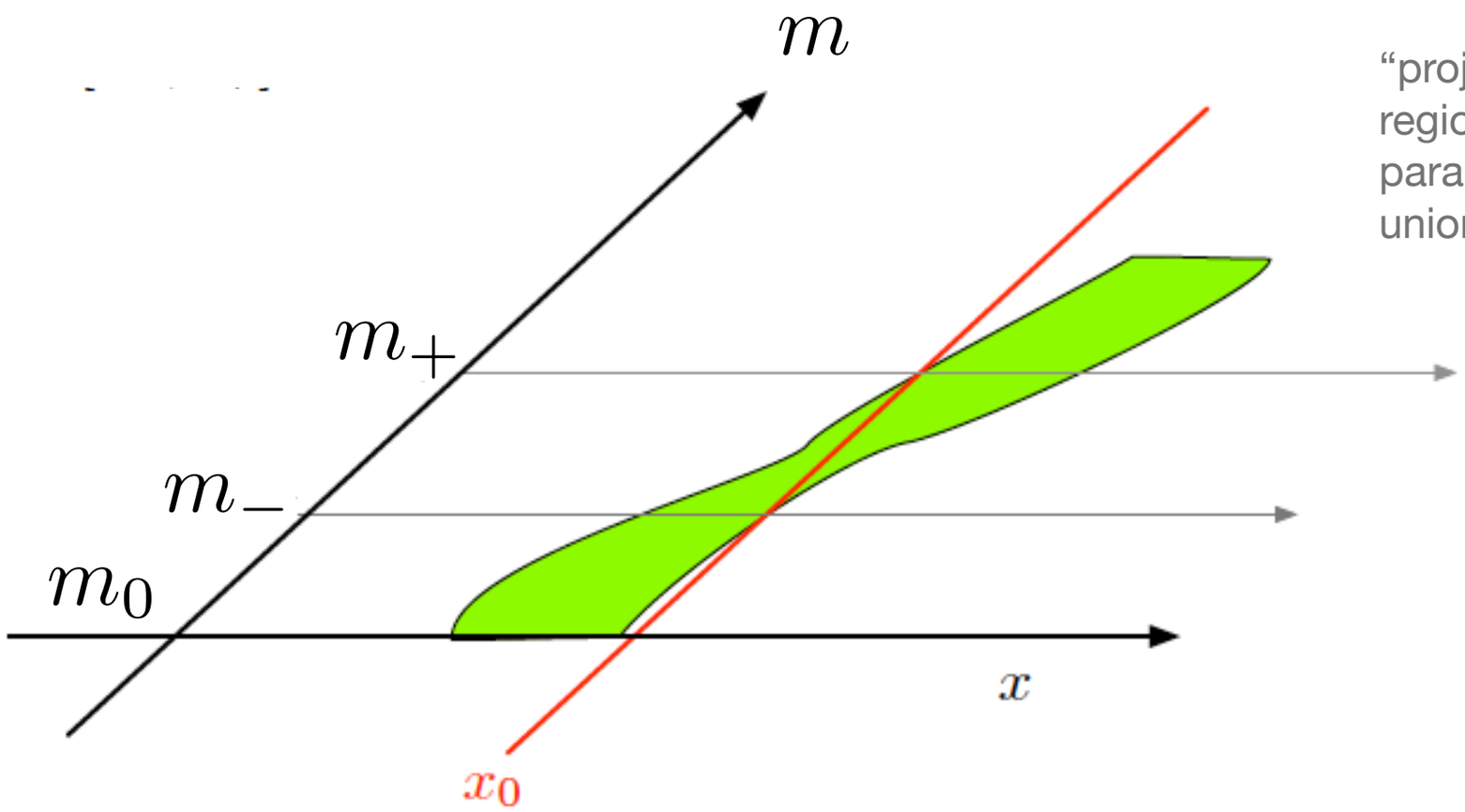
Neyman illustrated VI

This defines a confidence belt for m.



Neyman illustrated VII

Then you look at data and observe a value x_0. The observed value intersects the confidence belt. The *union* of all values of m for which the observed x_0 intercepts the confidence belt defines the confidence interval $[m_-(x_0)\ m_+(x_0)]$ at the 68% CL for the parameter. The extremes of the interval are random var. (functions of data x)



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

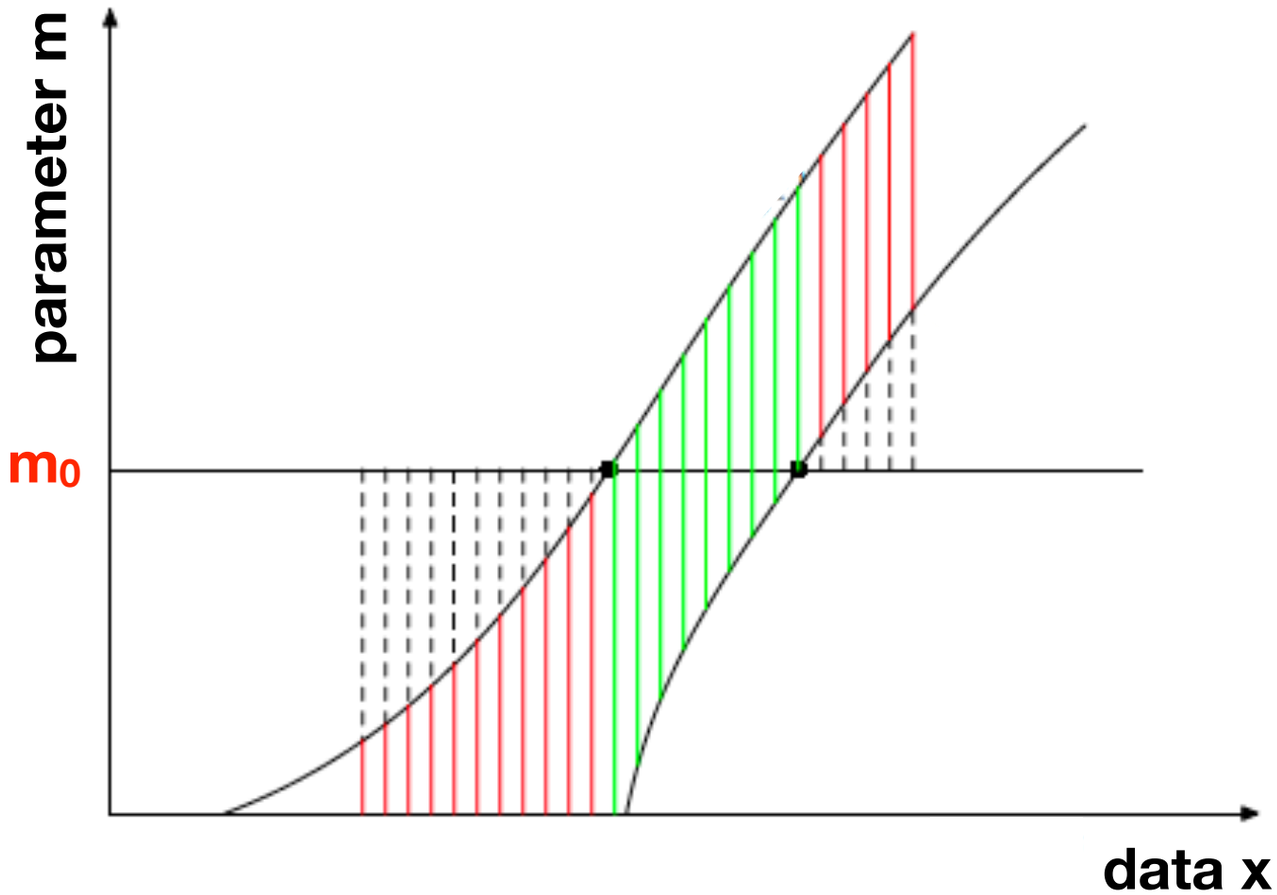
In repeated experiments, the boundaries of the confidence intervals $[m_-(x)\ m_+(x)]$ will differ, but 68% of them will contain the (unknown) true value of the parameter m

Neyman's "magic" explained

Suppose the true value is m_0

Depending on the data observed, could get either the red of the green intervals. Red intervals don't include m_0 — green intervals do.

Since the probability of observing data yielding 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



The procedure guarantees coverage. The result of a measurement is expressed as "m is in [a, b] at the 68% CL". It does not mean p(a < m < b) = 90%. It rather means that by repeating the procedure, 68% of the obtained intervals include the true value.

Toy example

I have externally identical bags of various classes. Each class contains a different fraction of white balls (class A = 1%, B= 5%, C = 50%, D= 95%, and D = 99%). Pick a bag, extract 5 balls, and infer whether the bag is class A, B etc, by setting a 95% CL upper limit on the true fraction of white balls.

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}

Start constructing one-sided confidence band…

For true value A, accumulate probability starting from high values of observations, which are “extreme” for an upper limit, until the probability accumulated is at least 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…

…continue band construction to true value B…

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

Green marks 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

Pick five balls from an unknown bag. Find only one white ball out of five.
==> D and E class are out of the confidence region: exclude class D and class E at the 95% CL.

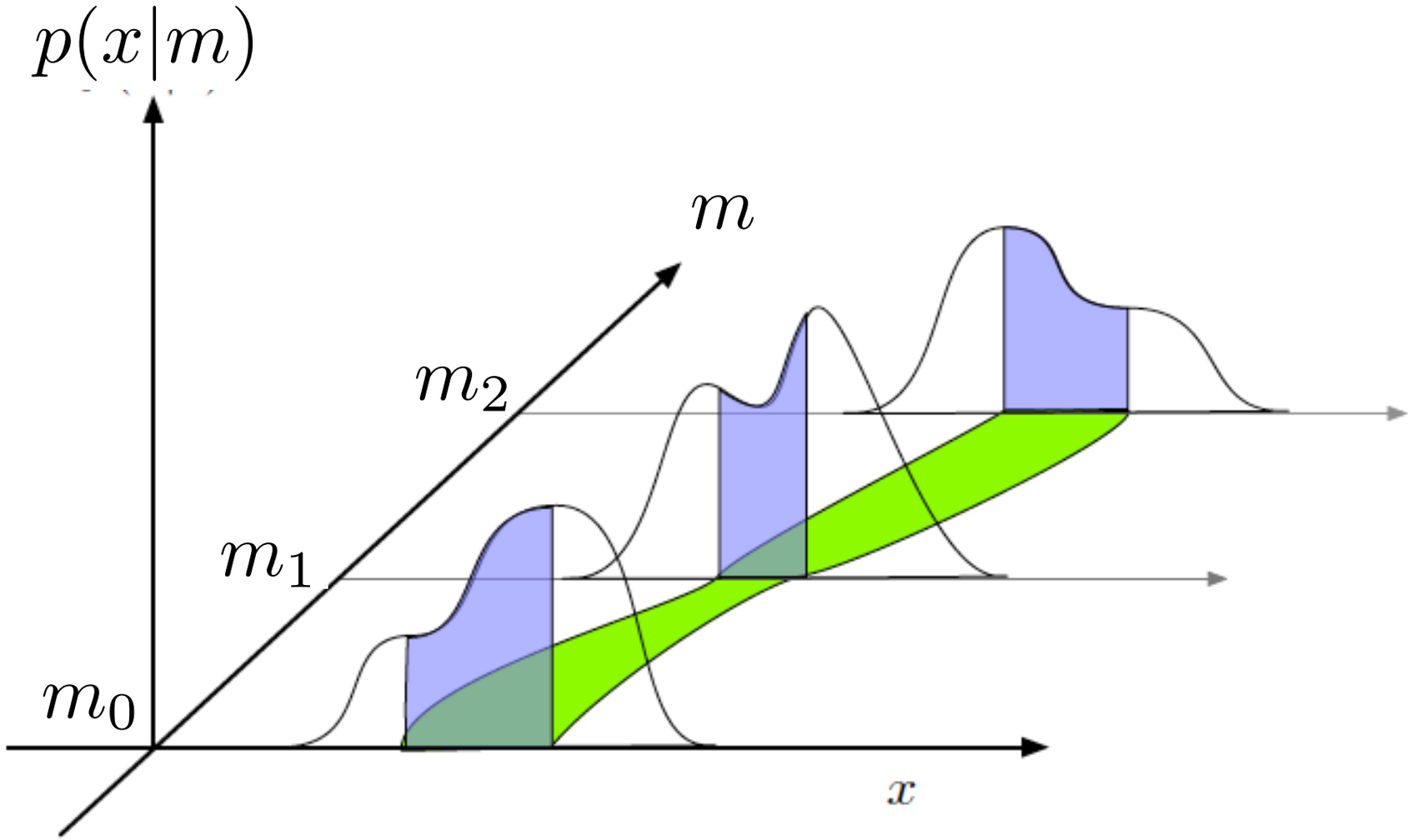
True fraction of white balls (this is “m”)

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}

EXCLUDED at 95% CL

Again on Neyman construction

It is true (in precisely the sense defined by the chosen ordering) that the confidence interval 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

Back to our efficiency

$\hat{\rho} = n_{on}/n_{tot}$ Observe 3 successes on 10 trials — what is our efficiency and its uncertainty? Let's find exact 68% central confidence intervals $[\rho_1, \rho_1]$

Convenient practical trick: endpoints of a central interval at given CL can be found from one-sided confidence intervals (lower and upper limits) at 1-(1-CL)/2:

Find lower limit ρ_1 with C.L. = 1 – (1 – 68%)/2. = 84%
I.e., Find ρ_1 such that $Bi(n_{on} < 3 \mid n_{tot}=10, \rho_1) = 84\%$

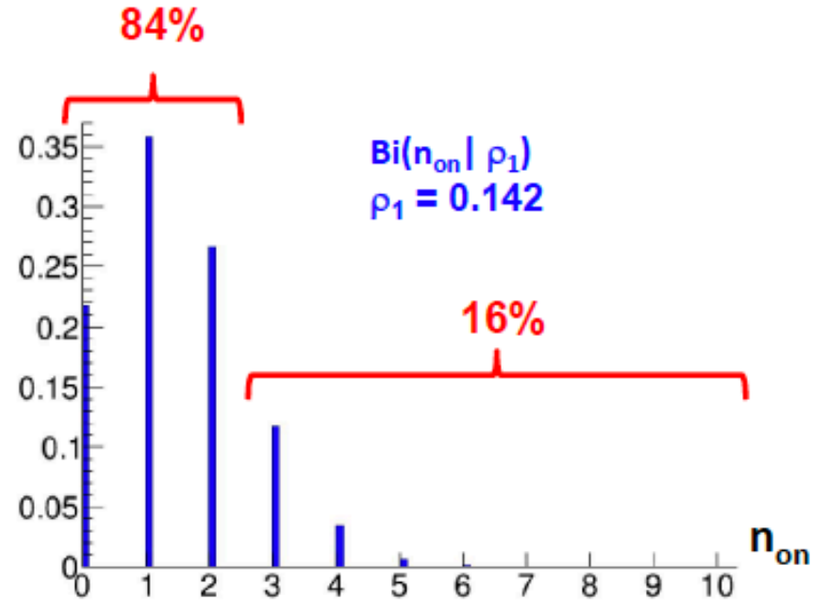
Find upper limit ρ_2 with C.L. = 84%
I.e., Find ρ_2 such that $Bi(n_{on} > 3 \mid n_{tot}=10, \rho_2) = 84\%$

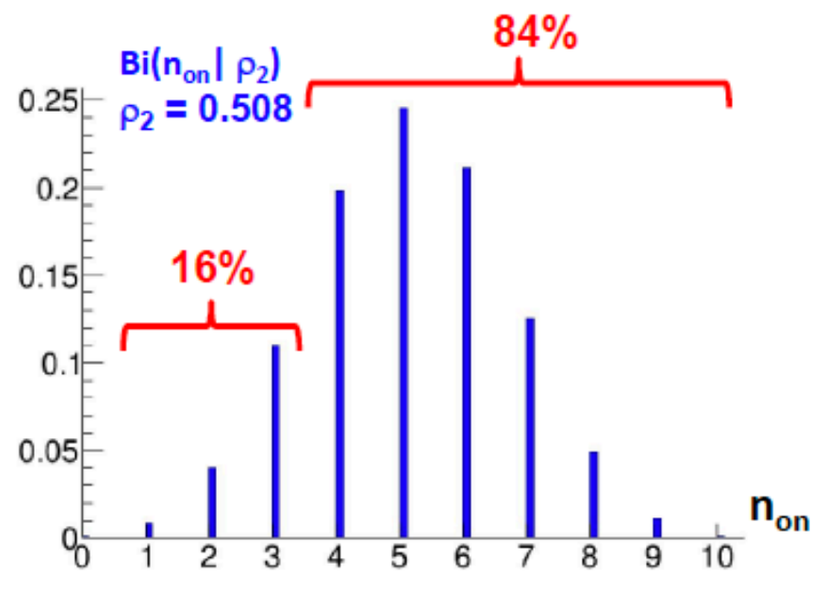
Confidence intervals for binomial

n_{on} = 3 , n_{tot}=10.
Find ρ_1 such that
Bi(n_{on} < 3 | ρ_1) = 84%
Bi($n_{on} \geq 3$ | ρ_1) = 16%
(lower limit at 84% C.L.)
Solve: ρ_1 = 0.142

And find ρ_2 such that
Bi(n_{on} > 3 | ρ_2) = 84%
Bi($n_{on} \leq 3$ | ρ_2) = 16%
(upper limit at 84% C.L.)
Solve: ρ_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)



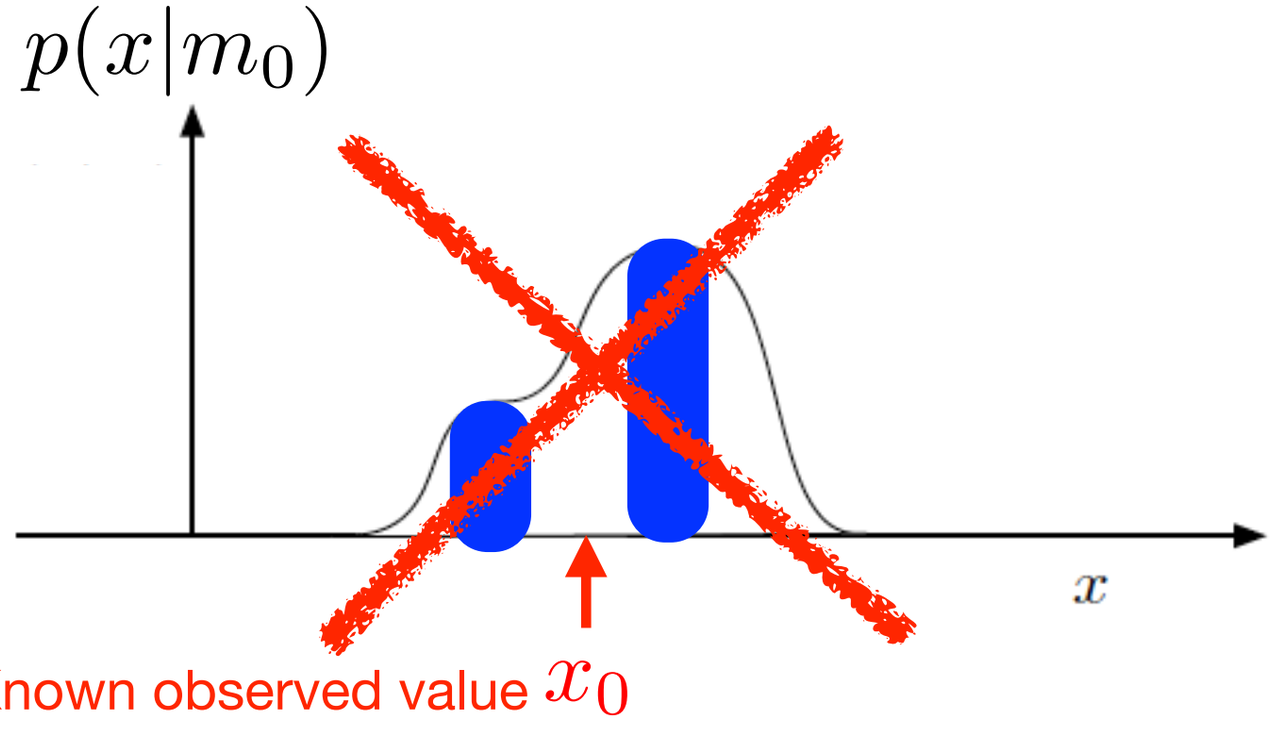


Poisson example: Fig. 3a,b; R. Cousins, Am. J. Phys. 63 398 (1995) DOI: 10.1119/1.17901

Ordering guidelines

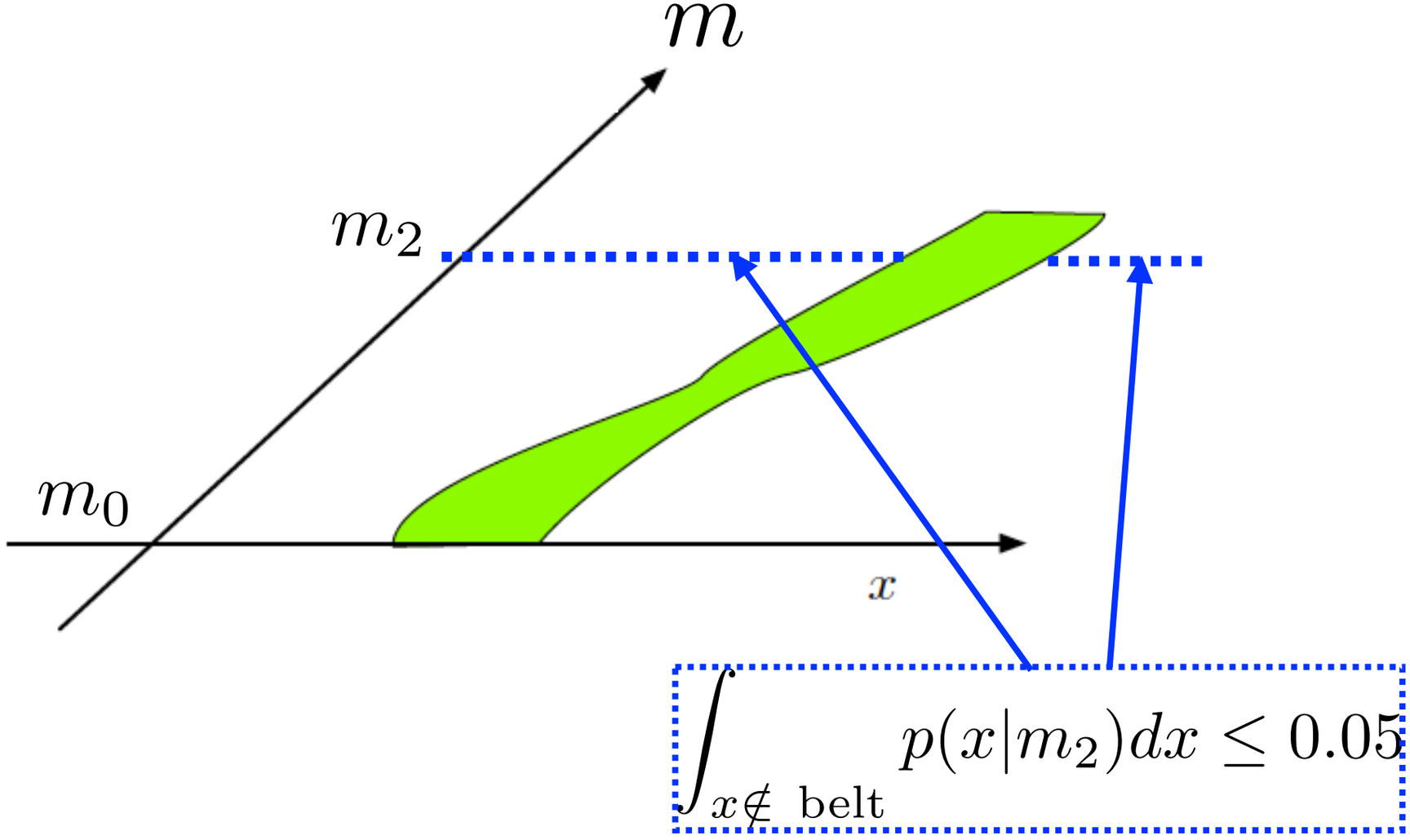
Despite arbitrariness, standards/conventions that are usually followed in the construction of the region.

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



Ordering

The ordering algorithm is arbitrarily chosen, provided that (i) has been defined prior to look at the data (ii) for each value m of the parameter, the integral of the pdf along the x region outside of the belt does not exceed 1-CL, e.g, 5% in a 95% CL confidence region construction

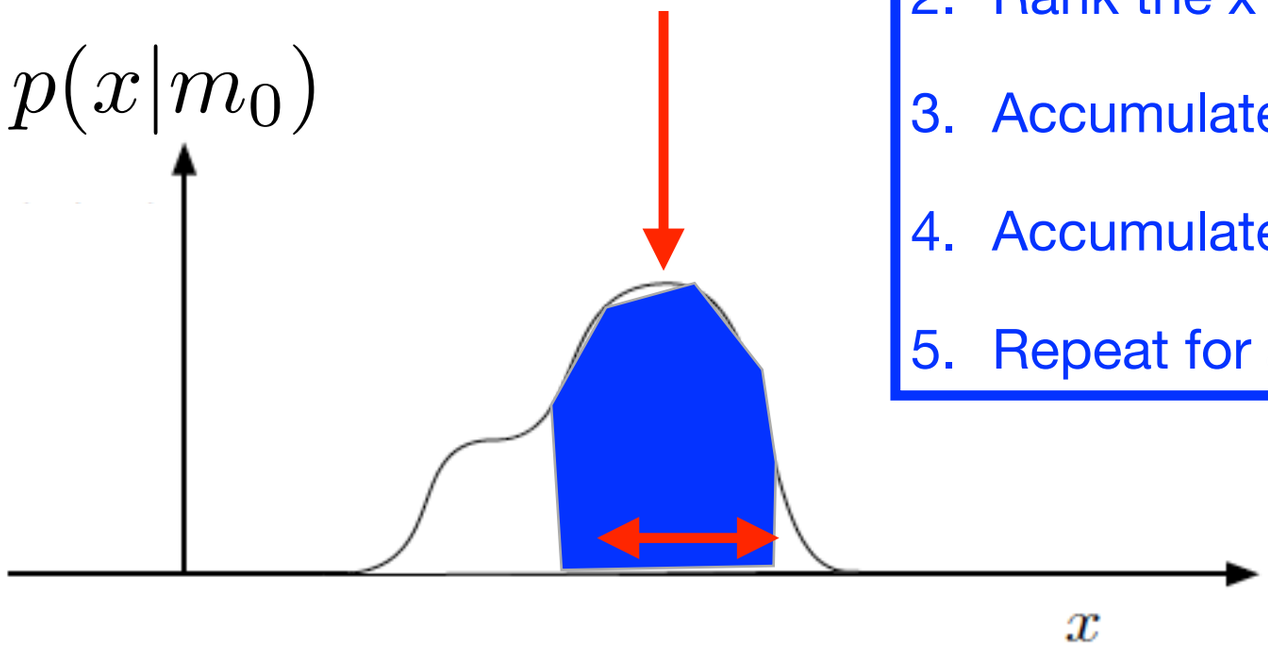


[Cranmer]

Probability ordering

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

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

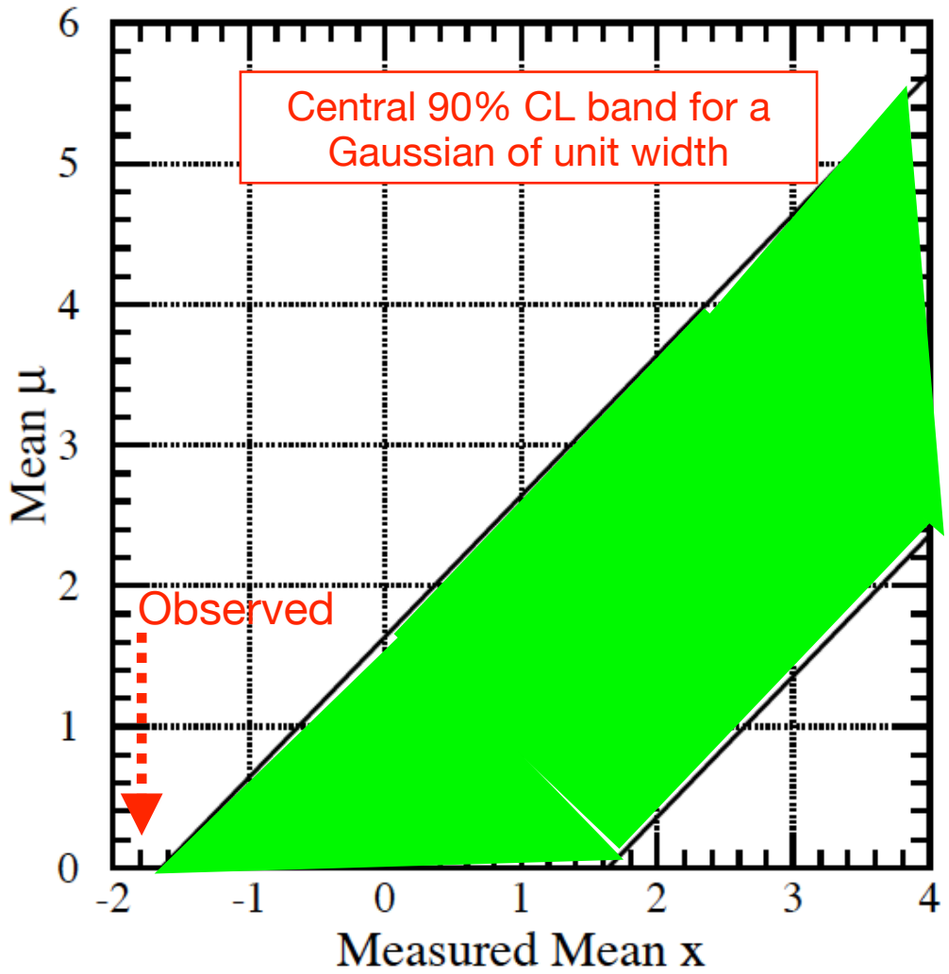


This is ill-defined: as probability depends on the metric for the observable x, the shortest interval in one metric isn’t shortest in others.

Issues — empty intervals

Long-standing inconsistencies found in simplistic ordering criteria.

For instance Gaussian measurement resolution near a physical boundary (e.g., like a measurement of neutrino mass square close to zero).



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

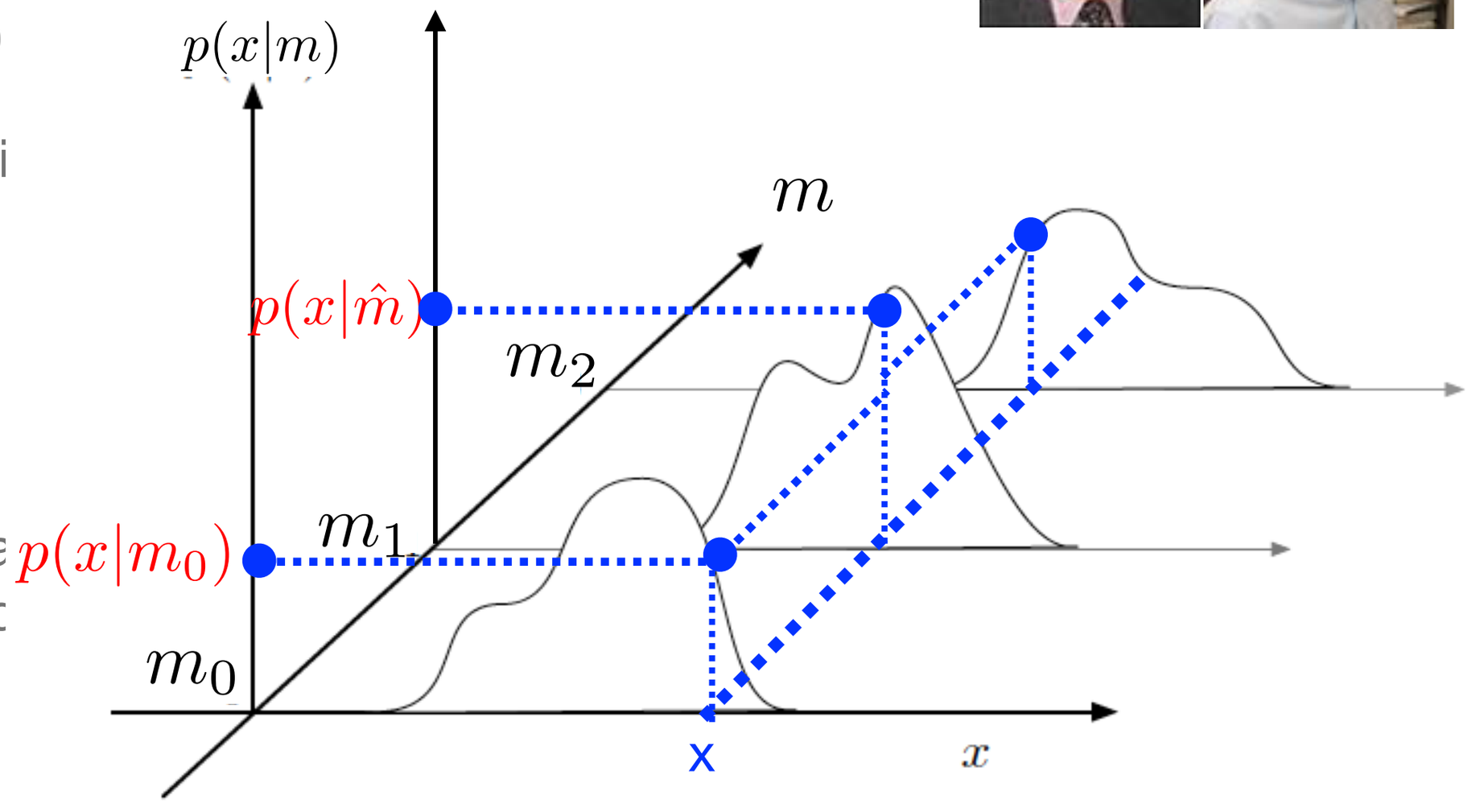
Likelihood-ratio ordering ("Feldman and Cousins")


Issues solved by adopting the likelihood-ratio ordering

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

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

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

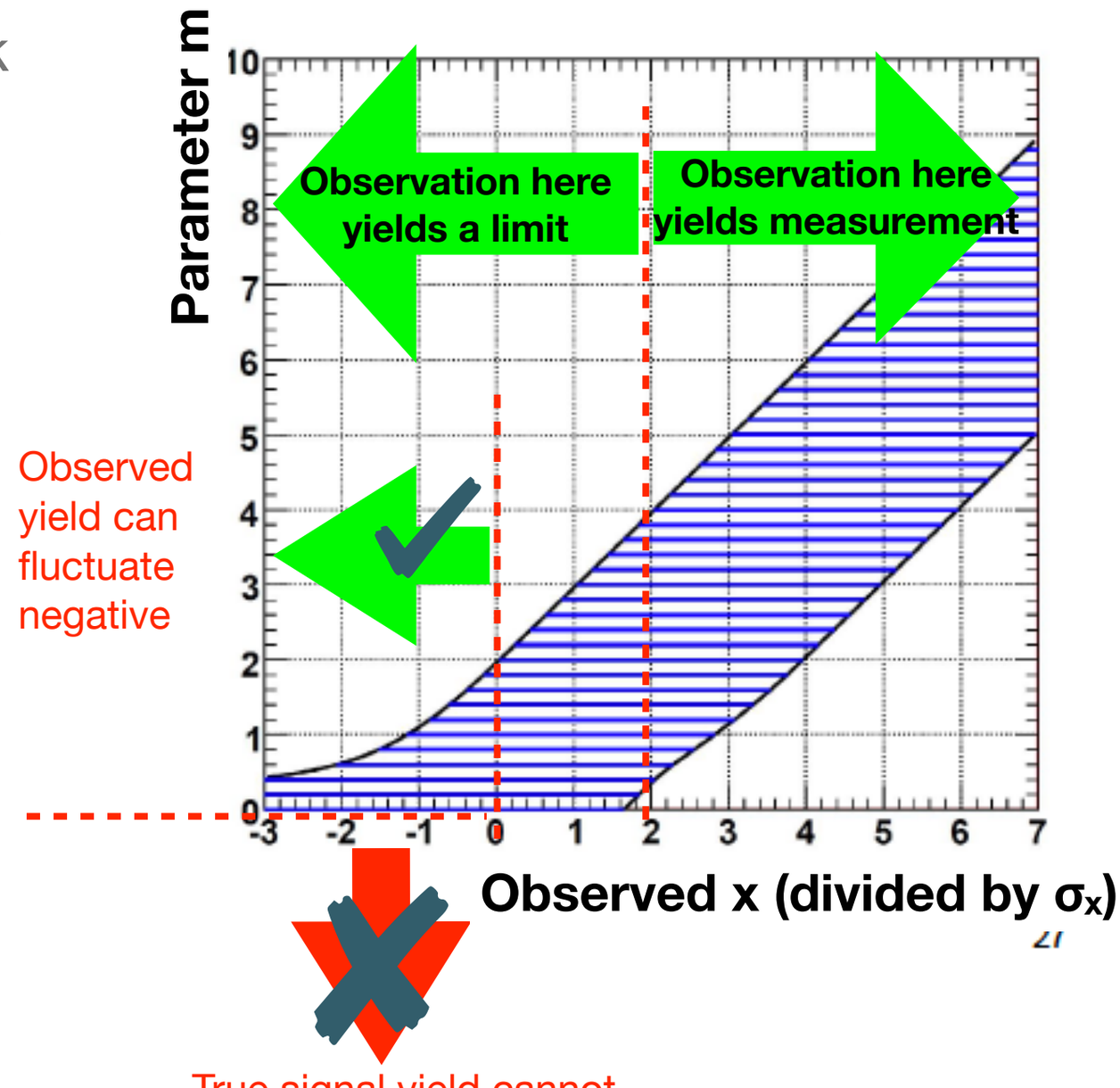


The "accumulation score" of each element in x, no longer depends only on $p(x|m_0)$ but also on $p(x|m)$ at other m values

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 observation x calculate (i) the value of the likelihood at m_0, $p(x|m_0)=L(m_0)$ and (ii) the maximum likelihood $L(\hat{m})$ over the space of m values (for that observation)
3. Rank all x in decreasing order of likelihood ratio $LR=L_x(m_0)/L_x(\hat{m})$.
4. Accumulate probability starting from the x with higher LR until the desired CL is reached.
5. Repeat for all m

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

In your work, take LR-ordering as default option unless there are strong motivations against it.

Got your brain tangled? Try with Poisson.

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

TABLE I. Illustrative calculations in the confidence belt construction for signal mean 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(μ =0.5)/L(μ̂) (ordering score)

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

In many cases, can use tabulated values

Poisson limits (from the original Feldman-Cousins paper).

Handy in a counting experiment (e.g,, a search for an excess in a bin) where I observe n0 and know ~precisely the expected background yield b

TABLE IV. 90% C.L. intervals for the Poisson signal mean μ, for total events observed n_0, for known mean background b ranging from 0 to 5.

$n_0\backslash b$	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	5.0
0	0.00, 2.44	0.00, 1.94	0.00, 1.61	0.00, 1.33	0.00, 1.26	0.00, 1.18	0.00, 1.08	0.00, 1.06	0.00, 1.01	*0.00, 0.98*
1	0.11, 4.36	0.00, 3.86	0.00, 3.36	0.00, 2.91	0.00, 2.53	0.00, 2.19	0.00, 1.88	0.00, 1.59	0.00, 1.39	0.00, 1.22
2	0.53, 5.91	0.03, 5.41	0.00, 4.91	0.00, 4.41	0.00, 3.91	0.00, 3.45	0.00, 3.04	0.00, 2.67	0.00, 2.33	0.00, 1.73
3	1.10, 7.42	0.60, 6.92	0.10, 6.42	0.00, 5.92	0.00, 5.42	0.00, 4.92	0.00, 4.42	0.00, 3.95	0.00, 3.53	0.00, 2.78
4	1.47, 8.60	1.17, 8.10	0.74, 7.60	0.24, 7.10	0.00, 6.60	0.00, 6.10	0.00, 5.60	0.00, 5.10	0.00, 4.60	0.00, 3.60
5	1.84, 9.99	1.53, 9.49	1.25, 8.99	0.93, 8.49	0.43, 7.99	0.00, 7.49	0.00, 6.99	0.00, 6.49	0.00, 5.99	0.00, 4.99
6	2.21,11.47	1.90,10.97	1.61,10.47	1.33, 9.97	1.08, 9.47	0.65, 8.97	0.15, 8.47	0.00, 7.97	0.00, 7.47	0.00, 6.47
7	3.56,12.53	3.06,12.03	2.56,11.53	2.09,11.03	1.59,10.53	1.18,10.03	0.89, 9.53	0.39, 9.03	0.00, 8.53	0.00, 7.53
8	3.96,13.99	3.46,13.49	2.96,12.99	2.51,12.49	2.14,11.99	1.81,11.49	1.51,10.99	1.06,10.49	0.66, 9.99	0.00, 8.99
9	4.36,15.30	3.86,14.80	3.36,14.30	2.91,13.80	2.53,13.30	2.19,12.80	1.88,12.30	1.59,11.80	1.33,11.30	0.43,10.30
10	5.50,16.50	5.00,16.00	4.50,15.50	4.00,15.00	3.50,14.50	3.04,14.00	2.63,13.50	2.27,13.00	1.94,12.50	1.19,11.50
11	5.91,17.81	5.41,17.31	4.91,16.81	4.41,16.31	3.91,15.81	3.45,15.31	3.04,14.81	2.67,14.31	2.33,13.81	1.73,12.81
12	7.01,19.00	6.51,18.50	6.01,18.00	5.51,17.50	5.01,17.00	4.51,16.50	4.01,16.00	3.54,15.50	3.12,15.00	2.38,14.00
13	7.42,20.05	6.92,19.55	6.42,19.05	5.92,18.55	5.42,18.05	4.92,17.55	4.42,17.05	3.95,16.55	3.53,16.05	2.78,15.05
14	8.50,21.50	8.00,21.00	7.50,20.50	7.00,20.00	6.50,19.50	6.00,19.00	5.50,18.50	5.00,18.00	4.50,17.50	3.59,16.50
15	9.48,22.52	8.98,22.02	8.48,21.52	7.98,21.02	7.48,20.52	6.98,20.02	6.48,19.52	5.98,19.02	5.48,18.52	4.48,17.52
16	9.99,23.99	9.49,23.49	8.99,22.99	8.49,22.49	7.99,21.99	7.49,21.49	6.99,20.99	6.49,20.49	5.99,19.99	4.99,18.99
17	11.04,25.02	10.54,24.52	10.04,24.02	9.54,23.52	9.04,23.02	8.54,22.52	8.04,22.02	7.54,21.52	7.04,21.02	6.04,20.02
18	11.47,26.16	10.97,25.66	10.47,25.16	9.97,24.66	9.47,24.16	8.97,23.66	8.47,23.16	7.97,22.66	7.47,22.16	6.47,21.16
19	12.51,27.51	12.01,27.01	11.51,26.51	11.01,26.01	10.51,25.51	10.01,25.01	9.51,24.51	9.01,24.01	8.51,23.51	7.51,22.51
20	13.55,28.52	13.05,28.02	12.55,27.52	12.05,27.02	11.55,26.52	11.05,26.02	10.55,25.52	10.05,25.02	9.55,24.52	8.55,23.52

TABLE V. 90% C.L. intervals for the Poisson signal mean μ, for total events observed n_0, for known mean background b ranging from 6 to 15.

$n_0\backslash b$	6.0	7.0	8.0	9.0	10.0	11.0	12.0	13.0	14.0	15.0
0	*0.00, 0.97*	*0.00, 0.95*	*0.00, 0.94*	*0.00, 0.94*	*0.00, 0.93*	*0.00, 0.93*	*0.00, 0.92*	*0.00, 0.92*	*0.00, 0.92*	*0.00, 0.92*
1	0.00, 1.14	*0.00, 1.10*	*0.00, 1.07*	*0.00, 1.05*	*0.00, 1.03*	*0.00, 1.01*	*0.00, 1.00*	*0.00, 0.99*	*0.00, 0.99*	*0.00, 0.98*
2	0.00, 1.57	0.00, 1.38	0.00, 1.27	*0.00, 1.21*	*0.00, 1.15*	*0.00, 1.11*	*0.00, 1.09*	*0.00, 1.08*	*0.00, 1.06*	*0.00, 1.05*
3	0.00, 2.14	0.00, 1.75	0.00, 1.49	0.00, 1.37	0.00, 1.29	*0.00, 1.24*	*0.00, 1.21*	*0.00, 1.18*	*0.00, 1.15*	*0.00, 1.14*
4	0.00, 2.83	0.00, 2.56	0.00, 1.98	0.00, 1.82	0.00, 1.57	0.00, 1.45	*0.00, 1.37*	*0.00, 1.31*	*0.00, 1.27*	*0.00, 1.24*
5	0.00, 4.07	0.00, 3.28	0.00, 2.60	0.00, 2.38	0.00, 1.85	0.00, 1.70	0.00, 1.58	0.00, 1.48	*0.00, 1.39*	*0.00, 1.32*
6	0.00, 5.47	0.00, 4.54	0.00, 3.73	0.00, 3.02	0.00, 2.40	0.00, 2.21	0.00, 1.86	0.00, 1.67	0.00, 1.55	*0.00, 1.47*
7	0.00, 6.53	0.00, 5.53	0.00, 4.58	0.00, 3.77	0.00, 3.26	0.00, 2.81	0.00, 2.23	0.00, 2.07	0.00, 1.86	0.00, 1.69
8	0.00, 7.99	0.00, 6.99	0.00, 5.99	0.00, 5.05	0.00, 4.22	0.00, 3.49	0.00, 2.83	0.00, 2.62	0.00, 2.11	0.00, 1.95
9	0.00, 9.30	0.00, 8.30	0.00, 7.30	0.00, 6.30	0.00, 5.30	0.00, 4.30	0.00, 3.93	0.00, 3.25	0.00, 2.64	0.00, 2.45
10	0.22,10.50	0.00, 9.50	0.00, 8.50	0.00, 7.50	0.00, 6.50	0.00, 5.56	0.00, 4.71	0.00, 3.95	0.00, 3.27	0.00, 3.00
11	1.01,11.81	0.02,10.81	0.00, 9.81	0.00, 8.81	0.00, 7.81	0.00, 6.81	0.00, 5.81	0.00, 4.81	0.00, 4.39	0.00, 3.69
12	1.57,13.00	0.83,12.00	0.00,11.00	0.00,10.00	0.00, 9.00	0.00, 8.00	0.00, 7.00	0.00, 6.05	0.00, 5.19	0.00, 4.42
13	2.14,14.05	1.50,13.05	0.65,12.05	0.00,11.05	0.00,10.05	0.00, 9.05	0.00, 8.05	0.00, 7.05	0.00, 6.08	0.00, 5.22
14	2.83,15.50	2.13,14.50	1.39,13.50	0.47,12.50	0.00,11.50	0.00,10.50	0.00, 9.50	0.00, 8.50	0.00, 7.50	0.00, 6.55
15	3.48,16.52	2.56,15.52	1.98,14.52	1.26,13.52	0.30,12.52	0.00,11.52	0.00,10.52	0.00, 9.52	0.00, 8.52	0.00, 7.52
16	4.07,17.99	3.28,16.99	2.60,15.99	1.82,14.99	1.13,13.99	0.14,12.99	0.00,11.99	0.00,10.99	0.00, 9.99	0.00, 8.99
17	5.04,19.02	4.11,18.02	3.32,17.02	2.38,16.02	1.81,15.02	0.98,14.02	0.00,13.02	0.00,12.02	0.00,11.02	0.00,10.02
18	5.47,20.16	4.54,19.16	3.73,18.16	3.02,17.16	2.40,16.16	1.70,15.16	0.82,14.16	0.00,13.16	0.00,12.16	0.00,11.16
19	6.51,21.51	5.51,20.51	4.58,19.51	3.77,18.51	3.05,17.51	2.21,16.51	1.58,15.51	0.67,14.51	0.00,13.51	0.00,12.51
20	7.55,22.52	6.55,21.52	5.55,20.52	4.55,19.52	3.55,18.52	2.81,17.52	2.23,16.52	1.48,15.52	0.53,14.52	0.00,13.52

In many cases, can use tabulated values

Gaussian limits (from the original Feldman-Cousins paper)

Handy in a fit to a signal where pulls for the signal yield are Gaussian

TABLE X. Our confidence intervals for the mean μ of a Gaussian, constrained to be non-negative, as a function of the measured mean x_0, for commonly used confidence levels. Italicized intervals corresponds to cases where the goodness-of-fit probability (Sec. IV C) is less than 1%. All numbers are in units of σ.

x_0	68.27% C.L.	90% C.L.	95% C.L.	99% C.L.
-3.0	*0.00, 0.04*	*0.00, 0.26*	*0.00, 0.42*	*0.00, 0.80*
-2.9	*0.00, 0.04*	*0.00, 0.27*	*0.00, 0.44*	*0.00, 0.82*
-2.8	*0.00, 0.04*	*0.00, 0.28*	*0.00, 0.45*	*0.00, 0.84*
-2.7	*0.00, 0.04*	*0.00, 0.29*	*0.00, 0.47*	*0.00, 0.87*
-2.6	*0.00, 0.05*	*0.00, 0.30*	*0.00, 0.48*	*0.00, 0.89*
-2.5	*0.00, 0.05*	*0.00, 0.32*	*0.00, 0.50*	*0.00, 0.92*
-2.4	*0.00, 0.05*	*0.00, 0.33*	*0.00, 0.52*	*0.00, 0.95*
-2.3	0.00, 0.05	0.00, 0.34	0.00, 0.54	0.00, 0.99
-2.2	0.00, 0.06	0.00, 0.36	0.00, 0.56	0.00, 1.02
-2.1	0.00, 0.06	0.00, 0.38	0.00, 0.59	0.00, 1.06
-2.0	0.00, 0.07	0.00, 0.40	0.00, 0.62	0.00, 1.10
-1.9	0.00, 0.08	0.00, 0.43	0.00, 0.65	0.00, 1.14
-1.8	0.00, 0.09	0.00, 0.45	0.00, 0.68	0.00, 1.19
-1.7	0.00, 0.10	0.00, 0.48	0.00, 0.72	0.00, 1.24
-1.6	0.00, 0.11	0.00, 0.52	0.00, 0.76	0.00, 1.29
-1.5	0.00, 0.13	0.00, 0.56	0.00, 0.81	0.00, 1.35
-1.4	0.00, 0.15	0.00, 0.60	0.00, 0.86	0.00, 1.41
-1.3	0.00, 0.17	0.00, 0.64	0.00, 0.91	0.00, 1.47
-1.2	0.00, 0.20	0.00, 0.70	0.00, 0.97	0.00, 1.54
-1.1	0.00, 0.23	0.00, 0.75	0.00, 1.04	0.00, 1.61
-1.0	0.00, 0.27	0.00, 0.81	0.00, 1.10	0.00, 1.68
-0.9	0.00, 0.32	0.00, 0.88	0.00, 1.17	0.00, 1.76
-0.8	0.00, 0.37	0.00, 0.95	0.00, 1.25	0.00, 1.84
-0.7	0.00, 0.43	0.00, 1.02	0.00, 1.33	0.00, 1.93
-0.6	0.00, 0.49	0.00, 1.10	0.00, 1.41	0.00, 2.01
-0.5	0.00, 0.56	0.00, 1.18	0.00, 1.49	0.00, 2.10
-0.4	0.00, 0.64	0.00, 1.27	0.00, 1.58	0.00, 2.19
-0.3	0.00, 0.72	0.00, 1.36	0.00, 1.67	0.00, 2.28
-0.2	0.00, 0.81	0.00, 1.45	0.00, 1.77	0.00, 2.38
-0.1	0.00, 0.90	0.00, 1.55	0.00, 1.86	0.00, 2.48
0.0	0.00, 1.00	0.00, 1.64	0.00, 1.96	0.00, 2.58
0.1	0.00, 1.10	0.00, 1.74	0.00, 2.06	0.00, 2.68
0.2	0.00, 1.20	0.00, 1.84	0.00, 2.16	0.00, 2.78
0.3	0.00, 1.30	0.00, 1.94	0.00, 2.26	0.00, 2.88
0.4	0.00, 1.40	0.00, 2.04	0.00, 2.36	0.00, 2.98
0.5	0.02, 1.50	0.00, 2.14	0.00, 2.46	0.00, 3.08
0.6	0.07, 1.60	0.00, 2.24	0.00, 2.56	0.00, 3.18
0.7	0.11, 1.70	0.00, 2.34	0.00, 2.66	0.00, 3.28
0.8	0.15, 1.80	0.00, 2.44	0.00, 2.76	0.00, 3.38
0.9	0.19, 1.90	0.00, 2.54	0.00, 2.86	0.00, 3.48
1.0	0.24, 2.00	0.00, 2.64	0.00, 2.96	0.00, 3.58
1.1	0.30, 2.10	0.00, 2.74	0.00, 3.06	0.00, 3.68
1.2	0.35, 2.20	0.00, 2.84	0.00, 3.16	0.00, 3.78
1.3	0.42, 2.30	0.02, 2.94	0.00, 3.26	0.00, 3.88
1.4	0.49, 2.40	0.12, 3.04	0.00, 3.36	0.00, 3.98
1.5	0.56, 2.50	0.22, 3.14	0.00, 3.46	0.00, 4.08
1.6	0.64, 2.60	0.31, 3.24	0.00, 3.56	0.00, 4.18
1.7	0.72, 2.70	0.38, 3.34	0.06, 3.66	0.00, 4.28
1.8	0.81, 2.80	0.45, 3.44	0.16, 3.76	0.00, 4.38
1.9	0.90, 2.90	0.51, 3.54	0.26, 3.86	0.00, 4.48
2.0	1.00, 3.00	0.58, 3.64	0.35, 3.96	0.00, 4.58
2.1	1.10, 3.10	0.65, 3.74	0.45, 4.06	0.00, 4.68
2.2	1.20, 3.20	0.72, 3.84	0.53, 4.16	0.00, 4.78
2.3	1.30, 3.30	0.79, 3.94	0.61, 4.26	0.00, 4.88
2.4	1.40, 3.40	0.87, 4.04	0.69, 4.36	0.07, 4.98
2.5	1.50, 3.50	0.95, 4.14	0.76, 4.46	0.17, 5.08
2.6	1.60, 3.60	1.02, 4.24	0.84, 4.56	0.27, 5.18
2.7	1.70, 3.70	1.11, 4.34	0.91, 4.66	0.37, 5.28
2.8	1.80, 3.80	1.19, 4.44	0.99, 4.76	0.47, 5.38
2.9	1.90, 3.90	1.28, 4.54	1.06, 4.86	0.57, 5.48
3.0	2.00, 4.00	1.37, 4.64	1.14, 4.96	0.67, 5.58

A shortcut to LR — Wilks' theorem

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

$$-2\log \mathrm{LR}(m) = -2\log \frac{p(x\,|\,m)}{p(x\,|\,\hat{m})}$$

Samuel S. Wilks (1906-1964)

approaches a $\chi^2(n)$ distribution with # of degrees of freedom n equal to the number of additional free parameters the numerator has wrt the denominator.

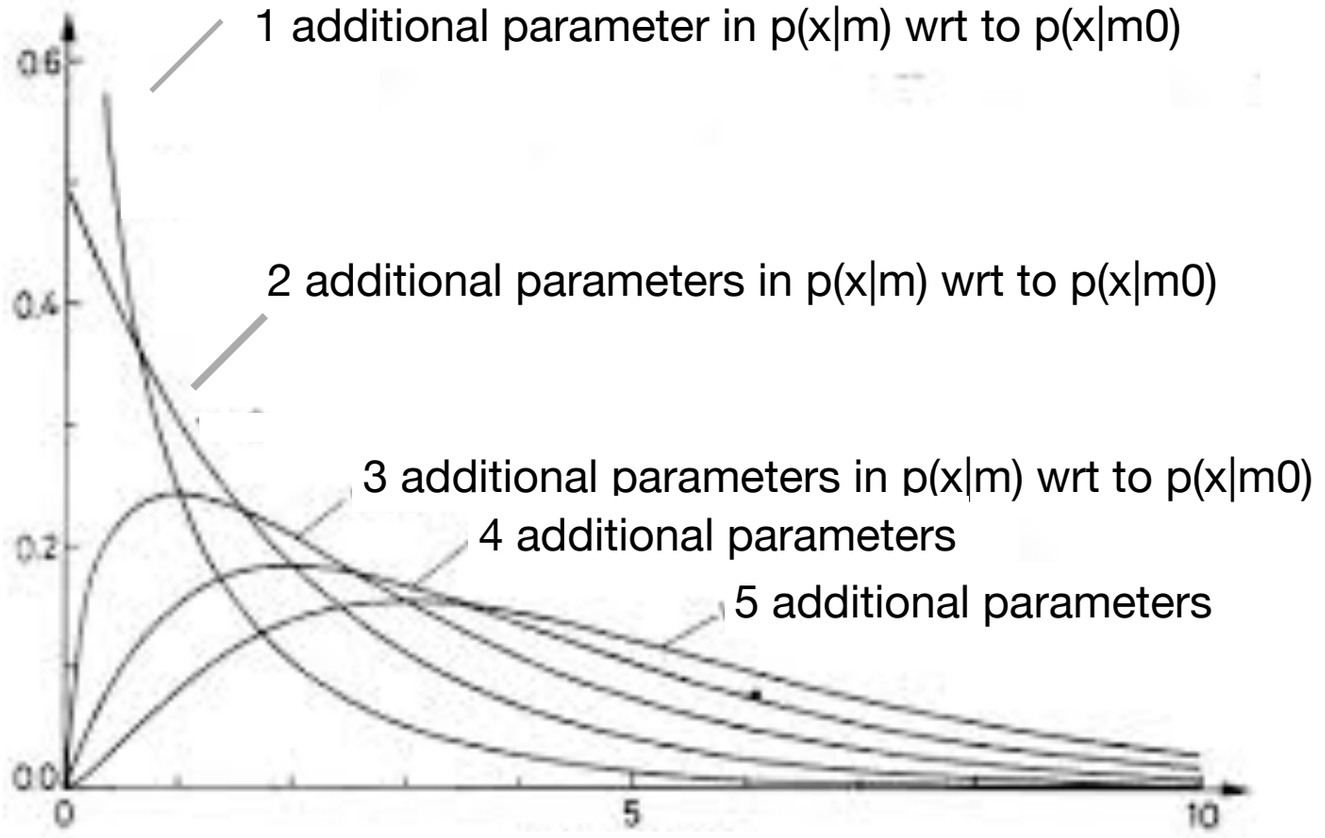
This holds independently of the shape of p(x|m) (does not need to be Gaussian!) and on the value of m (so-called "distribution-free")

It works due to an asymptotic limit based on the central-limit theorem and applies with some restrictions on the family of models: basically a shortcut of the FC in the asymptotic limit. If the likelihood is regular enough to be in asymptotic regime with the sample at hand, save you massive production of simulated experiments.

Wilks' theorem tells us how LR distribute

No need to generate the distributions of likelihood ratio statistic (no need for toys — saves lots of work!

Look at where the value of likelihood ratio observed in data falls along the appropriate curve (determined by the number of degrees of freedom)



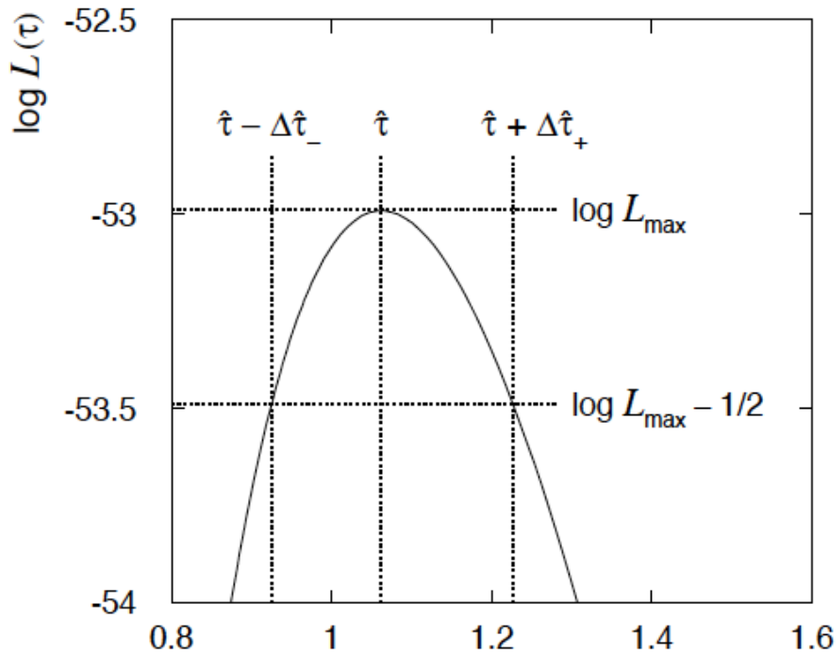
$$-2\log \mathrm{LR}(m) = -2\log \frac{p(x\,|\,m)}{p(x\,|\,\hat{m})}$$

Varying “thresholds” on the value of LR correspond to intervals in the space of parameters. Every locus of iso-LR values relative to the minimum “projects down” into an interval of the space of parameters with a known CL.

This is what the MINOS algorithm in MINUIT does

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

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



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

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

"projection" onto the space of parameters of a 1(2)-dimensional likelihood at the point where -2lnLR varies by 1.0 units identifies a 1(2)-dimensional 68(39)% CL central interval

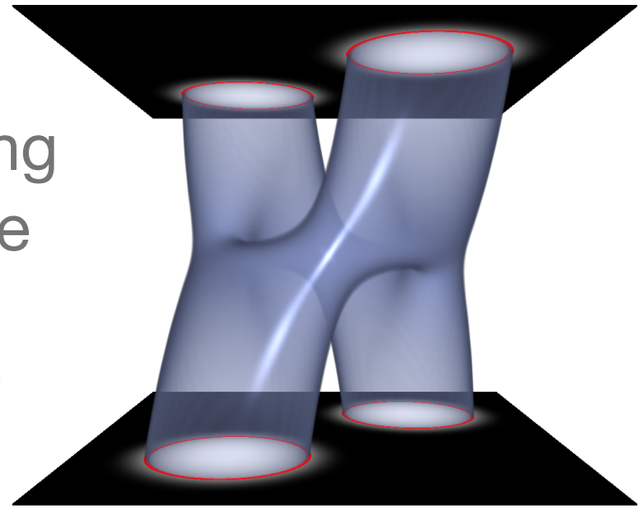
CL	Δ				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

"projection" onto the space of parameters of a 3-dimensional likelihood at the point where -2lnLR varies by 6.25 units identifies a 3-dimensional 90%CL central interval

Real life — high dimensions

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

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



With highly-dimensional likelihoods the “set-theory projection” of the full-dimensional confidence band into the lower-dimensional subspace of the parameter of interest leads to information loss: a lot of information on structure in the full dimensional space is lost when projected.

The resulting confidence interval is bigger (less precise results).

Systematic uncertainties

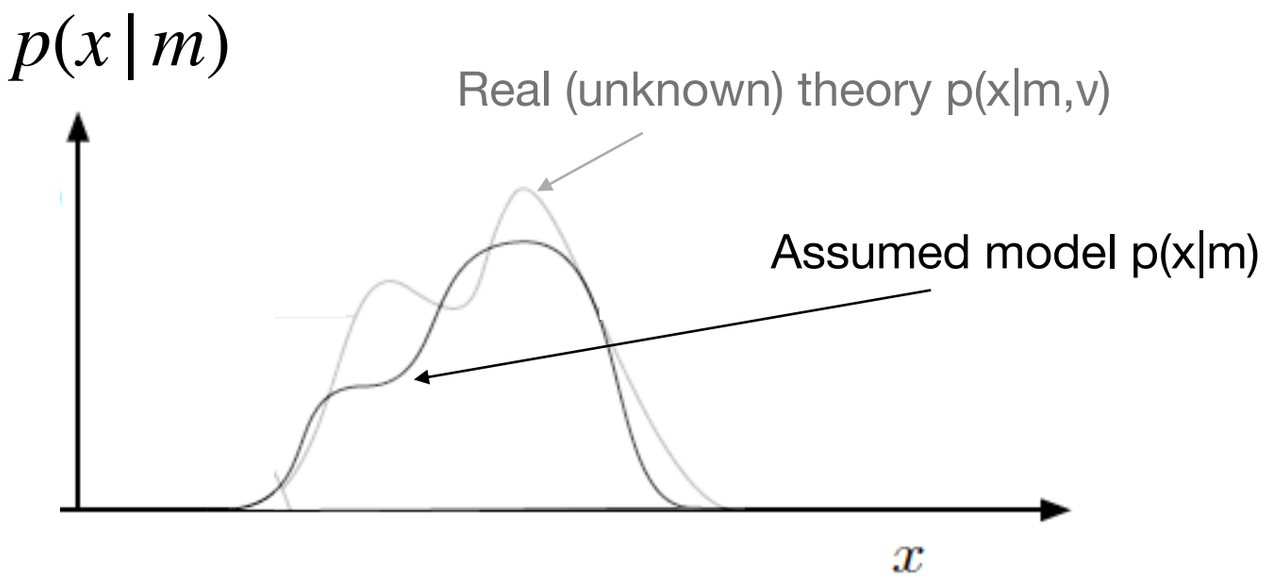
What systematic uncertainty is

The systematic uncertainty accounts for the differences between our model and reality

p(x|m) is an approximation of the real (and unknown) theory p(x|m,v).

Parametrize the difference with dpendence on additional unknown parameters v.

The better the approximation, the smaller the systematic uncertainty



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

Here we **do not know** the value of v *and* we **do not know** the distribution p(x|v).

If we knew p(x|v), we'd include it in our model to improve it and v would no longer contribute a systematic uncertainty. This is often done for v's of experimental origin (e.g., external experimental inputs), which are often Gaussian

How systematic uncertainty is interpreted

When a HEP physicist reads '***result is $\hat{m}$ = 10 ± 4 (stat) ± 3 (syst)' ,*** or the compact form that combines independent uncertainties $\hat{m}$ = 10 ± 5 (tot), no universal probabilistic interpretation exists.

Typically assume (consciously or not) that the range 5 < $\hat{m}$ < 15 contains the true value of m with >= 68.3% probability.

For this to happen, need systematic uncertainty sized to ensure coverage for any of the possible alternate models we could have picked (among those consistent with our data).

Reporting details on how systematic effects are considered is important. It allows for readers to form their own scientific opinion on the robustness of our work

Many assume that the results including systematics follow a Gaussian distribution. This is in general **NOT TRUE**, for the very conceptual definition of systematic uncertainty (if I knew it was Gaussian, then it would no longer be a systematic)

Nuisance parameters

Our likelihood L(m)=p(x|m) is based on model p(x|m) which is an approximation of the real (and unknown) theory p(x|m,ν) that depends on nuisance parameters ν

How does the "distance" between our model and reality impacts results? How can we make our results robust against that distance?

It is important to distinguish between

- nuisance parameters that contribute model refinements (ν has **unknown value** but p(x|ν) is **known**).
- nuisance parameters that contribute genuine systematics (both the value and p(x|ν) of ν are **unknown**).

Model refinements

In HEP, likelihoods often depend on parameters v that have **unknown** true value but **known** distribution p(x|v)**,** e.g.,

- reconstruction efficiencies, which may be determined from MC or control samples, and thus have known p(x|v)
- BFs of reference modes, which may be known by other experiments and thus have known p(x|v)

Inclusion of this information in the likelihood is straightforward:

Multiply your default p(x|m) by the likelihood for the nuisance parameters p(x|v) and fit the data.

You'll get a determination of v too (may not be relevant) but the **statistical uncertainty on m will include a contribution due to the uncertainty on v.**

Proper systematics in likelihood fits

The likelihood may also depend on parameters v where **both true value and distribution p(x|v) are unknown** e.g.,

- parameters regulating the shapes of components (e.g, backgrounds) of your sample that cannot be inferred precisely from MC or control samples.
- theoretical parameters that are needed to convert what you measure into a higher level physical quantity (decay constants from lattice QCD, ISR/FSR etc.)

Treatment of these cases is NOT straightforward.

Various approaches exist.

It is important to fully understand the assumptions and limitations associated with each so that the implications on the final results and their interpretation is known and can be documented.

Bayesian approach

A straightforward generalization of the standard Bayesian treatment.

Even if **p(x|v)** is unknown, *assume* a prior **p(v)** for the nuisance parameters (typically flat or Gaussian) and integrate ("marginalize") the product of that prior by the likelihood over **v.** Obtain a posterior 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. What you should know/do?

- Priors introduce subjective input. Not a problem in Bayesian view. Prone to inconsistencies (e.g., inconsistent results obtained by different groups using the same data). Concerns worsen for high-dimensional likelihood, since impact of priors on final results explodes with dimensionality. Test and report the robustness of your results under various choices of priors

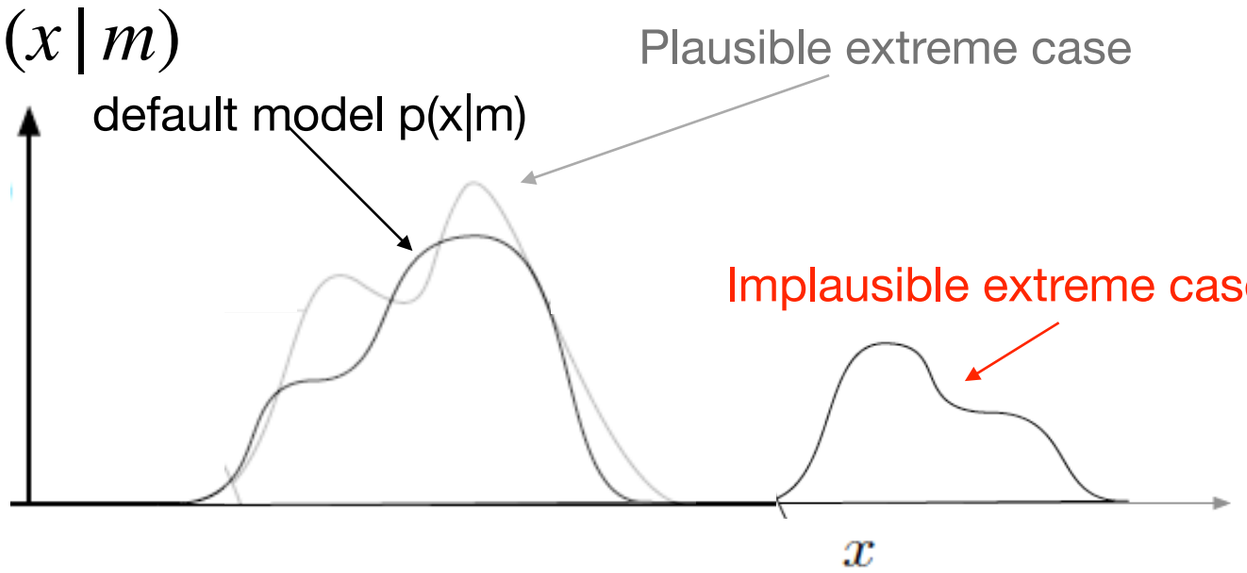
- No guarantee that final results contain the true value with the desired probability (68% or else). Unimportant in Bayesian statistics, but can lead to misinterpreting one's results. Important to state what is done.

Frequentist: step 1, bracketing the model

p(x|m,v) is unknown, but often data and physics allow for identifying a few `**extreme configurations**' for v, say v', and v", that bracket all **possible** configurations of the unknown parameter.

Extreme and possible do not mean to include *any* configuration for v.

Only the subset of configurations that are consistent with our data and objective knowledge are meaningful.



E.g., I might be uncertain about the specific mass shape of a background component, but I might have data sidebands that provide *some* information about that shape. Hence, my extreme shape configurations will be chosen among those that are still capable of fitting satisfactorily my sideband data.

Frequentist: step 2 replicating possible Universes

Once extreme cases for v are chosen, need to know how results would be if nature had chosen v' or v" as true values for v

- Draw from p(x|m,v) to generate ensembles of toys for each of the possible configurations — 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.

- Fit each toy and plot the distributions of results $\hat{m}$ for each ensemble.

The shape and spread of the $\hat{m}$ distributions under the various configurations determines the proper interval that contains the true m value with the desired probability, *regardless of the true value of v.*

[Occasionally, the number of nuisance parameters and their correlations prevents from identifying intuitively suited extreme cases. Then, sample randomly (uniformly in all dimensions) the space of nuisance parameters sufficient times and repeat the two points above]

Frequentist approach - What ya see is what ya get

What you should know/do if you go this way?

- You get coverage: results will contain the true value with at least the desired probability (68% or else). Reduced risk of misinterpreting one's results.
- Results might be less exciting at face value. The condition that intervals should contain the true value of m *for whatever configuration of nuisance parameters* is pessimistic for most possible values of *v.* That is, for most v, intervals will contain m with higher than 68% probability (i.e., you get larger uncertainties)
- It takes a lot of work. Occasionally, people skip the search for extreme values v', v" and just generate toys under the single configuration p(x|m,$\hat{v}$), where $\hat{v}$ is the ML estimate of v in data (so-called *plug-in* method). This saves work but spoils coverage if the value $\hat{v}$ differs from the unknown real value of v.
- Highly dimensional likelihoods might spoil the final precision, due to the topologically inherent information-loss that occurs when projecting from the higher-dimensional (m, v) space to the m subspace (see next slides)

Frequentist approach - profiling the likelihood

To obviate the loss of information (and computational complexity) in problems where likelihoods are functions of many variables, replace the likelihood with a lower-dimensional structure, the profile-likelihood, and base inference on that.

The profile likelihood is obtained by maximizing the likelihood with respect to a subset of its variables (usually the nuisance parameters *v*) and replacing their maximized values v̂ inside it:

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

The original likelihood is function of n + m variables and the profile likelihood is function of only n variables.

The profile likelihood is not a likelihood nor it has its mathematical properties.

However, profile-likelihood properties approach sufficiently well the likelihood properties in many problems thus offering a solid lower-dimensional instrument to perform inference.

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`)

While statistical reliability of these methods may vary, it is rarely desirable to mix Bayesian and frequentist techniques within the same inference, as that obfuscates a proper interpretation of the final results.

Profile-likelihood ratio ordering

Reduce the dimensionality of the problem. Not by integration as Bayesian do, but by derivation: replacing the likelihood with a lower-dimensional function obtained by maximizing the likelihood wrt the nuisance parameters.

FC ratio-ordering applied to likelihoods profiled (i.e., maximized) with respect to the uninteresting parameters. The profile-likelihood *is not a likelihood.* It is a lower-dimensional derivation of it that preserves some of the attractive features of the likelihood.

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

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

In practice

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

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

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

Key: how to treat nuisance paramt. in generation

Step 6 is essential to ensure coverage for all values of the nuisance parameters: this is the *"supremum p-value method"* Can be expensive.

Often circumvented using the *"plugin method"*: only generate pseudodata at the s values estimated on data. Equivalent to assume that the true values of the nuisance parameters are exactly those measured in data. Likely to be an optimistic assumption that spoils coverage.

Midway between plugin and supremum: generate pseudata at $\vec{s}$ values sampled in a plausible subvolume centered on their estimates in data. Berger and Boos: sample along each dimension s_i a range around the estimated value $\hat{s}$ with CL much larger than the target CL of the profile-likelihood interval. (e.g, when constructing a 68% CL band in m, sample a 99.7% CL range in each dimension in s space).
Applied in JASA, 89, 427 (1994) https://arxiv.org/pdf/0810.3229.pdf Phys. Rev. Lett 100 161802,

Comprehensive review of treatment of nuisance parameters: Sec 4 in www-cdf.fnal.gov/~luc/statistics/cdf8662.pdf

Interval estimation roadmap

I want to report my results as a confidence interval

Assume model p(x|m)

I am feeling Bayesian…

Assume priors for physics and nuisance parameters

Integrate (marginalize) priors on nuisance parameters to get the posterior on physics parameters

Check for prior dependence

Tough integrals

I don't stomach priors..

Check if Wilks theorem conditions apply

yes

Use differences in PLR to determine final interval

no/not sure

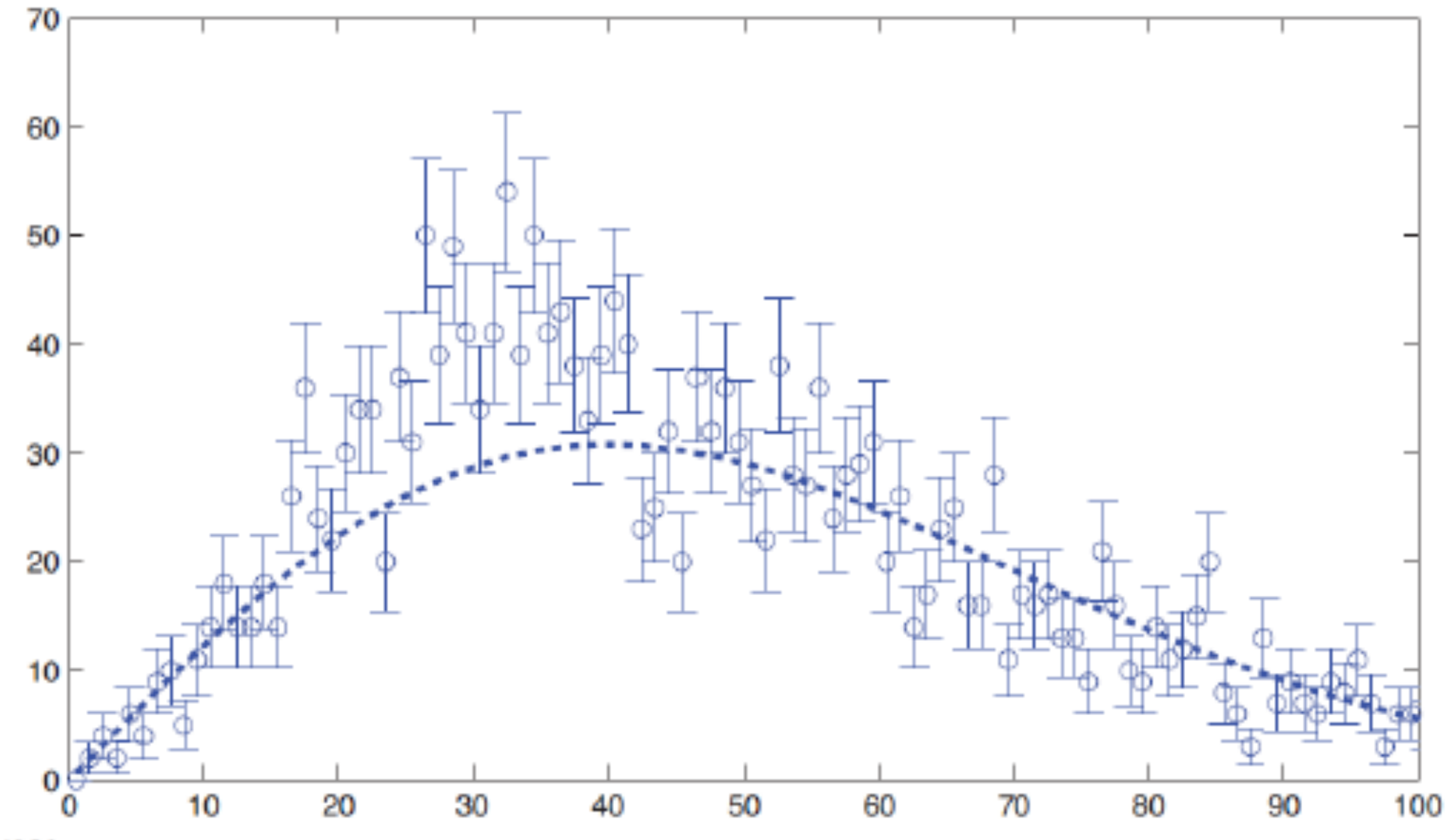
Construct confidence region using PLR-ordering

Check for undercoverage

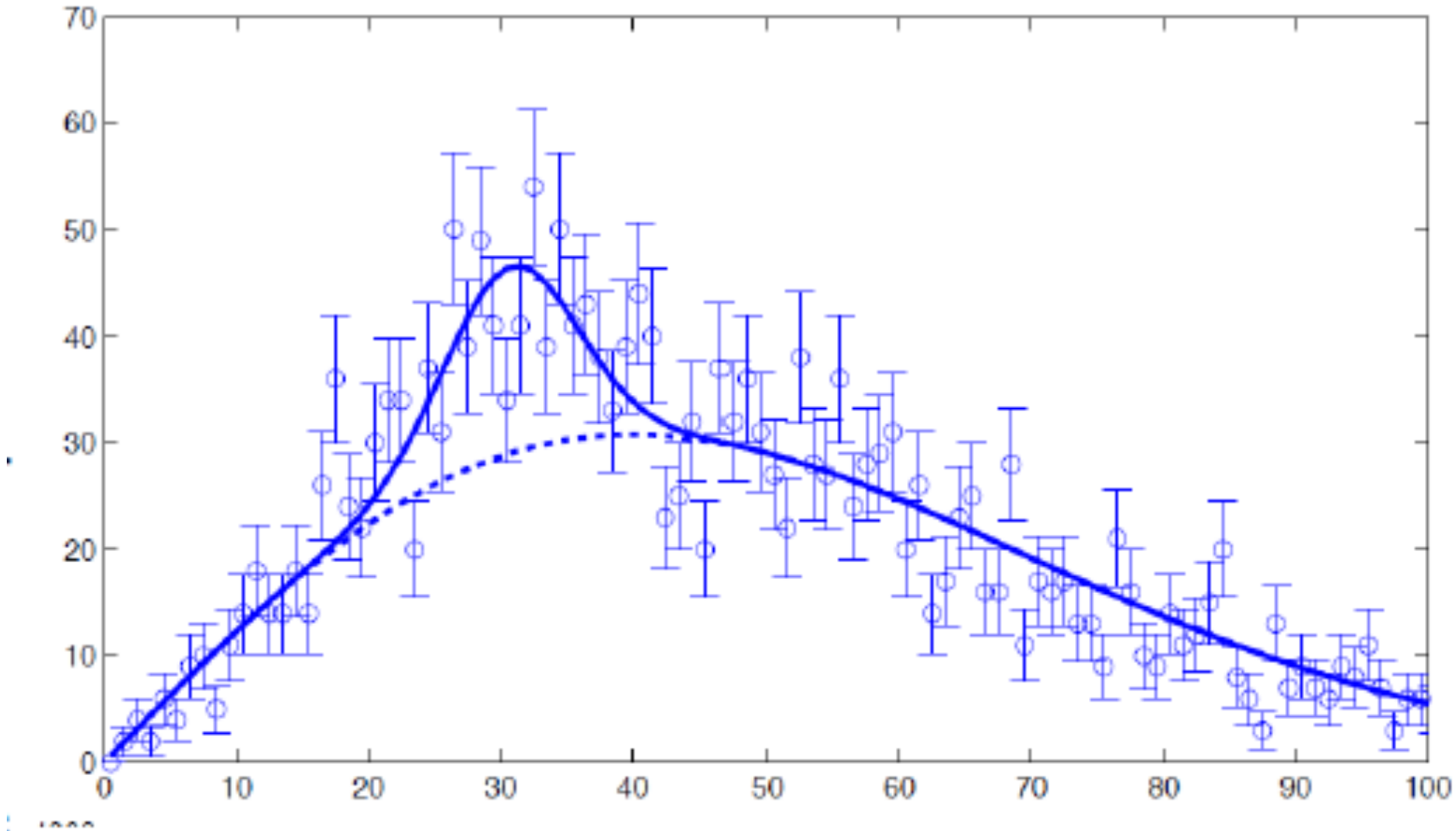
Lots of simulation

Hypothesis testing

Are my data compatible with background?



Or they suggest the presence of an anomaly?

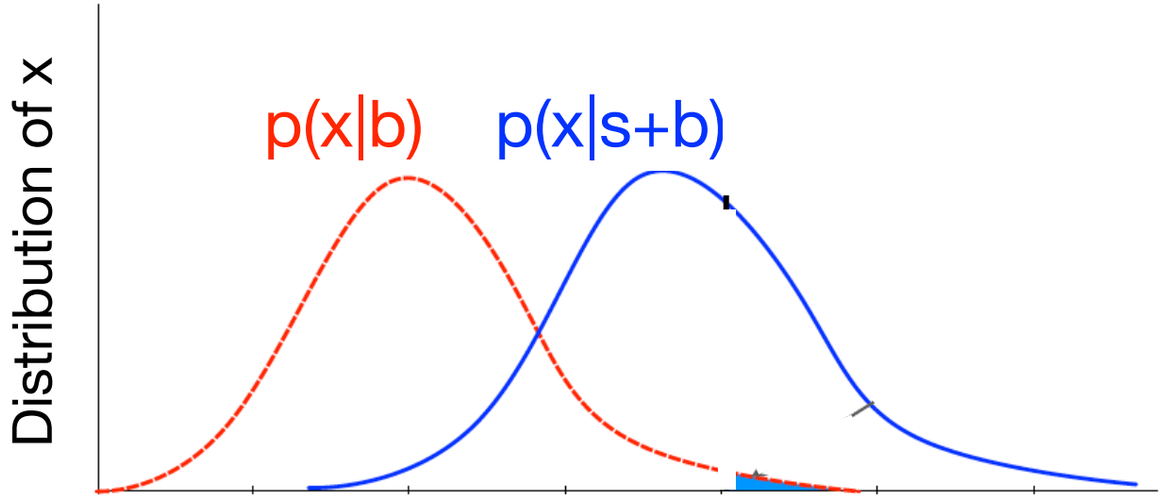


The p-value is a random variable that helps answering this question

http://priceonomics.com/the-guinness-brewer-who-revolutionized-statistics/

Ingredients (prepare prior to any observation)

1. Need two hypotheses. For instance: only known phenomena contribute"null" or "background") new phenomena contribute too ("alternate" or "signal")



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

2. Need a function x of the data (e.g., signal-event count), whose distribution under the null p(x|b) "differs" from that under the signal hypothesis p(x|s+b).

3. Generate these two distributions (typically done using simulation)

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

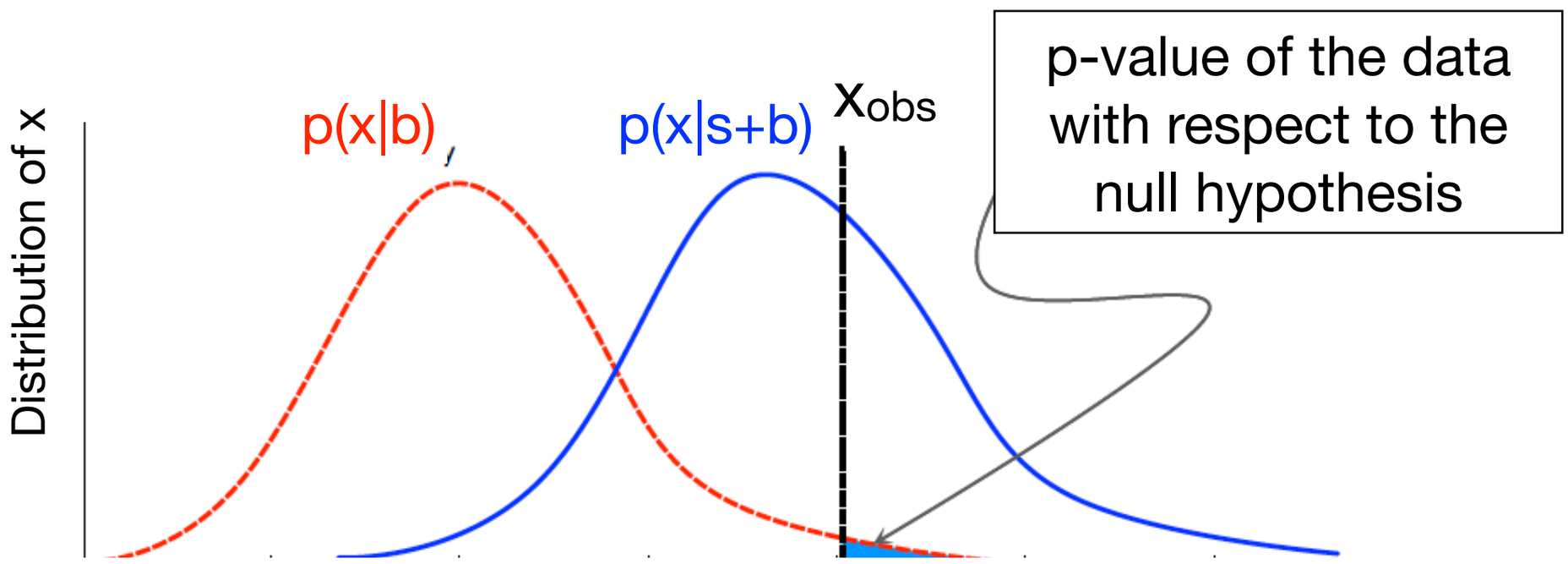
Step 2: look at the data

That is, look at what particular value x_{obs} the quantity x takes up in your data

p-values for discovering a new effect

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

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

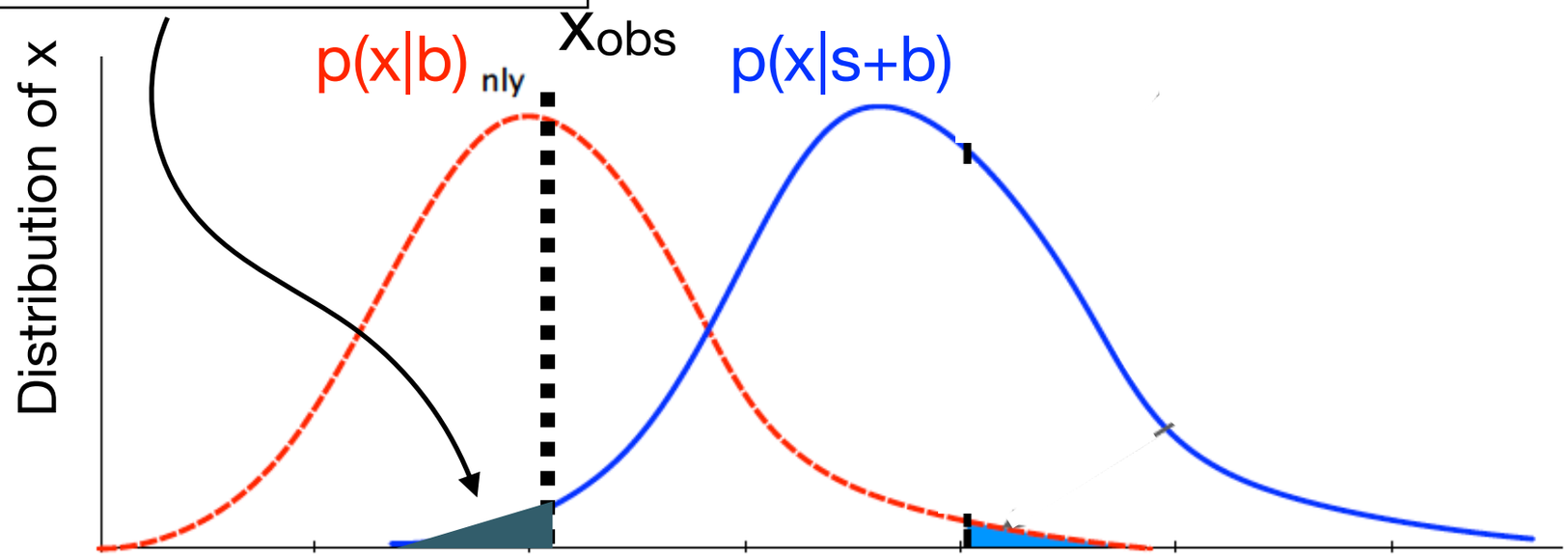


p-values for excluding a new effect

If the purpose is to exclude a new effect, then one tests the signal hypothesis, and quotes the p-value with respect to that.

Is the relative fraction of the integral of the signal model over values of x as background-like as that observed and more. The smaller the p-value, the stronger the evidence against the signal hypothesis.

p-value of the data with respect to the signal hypothesis

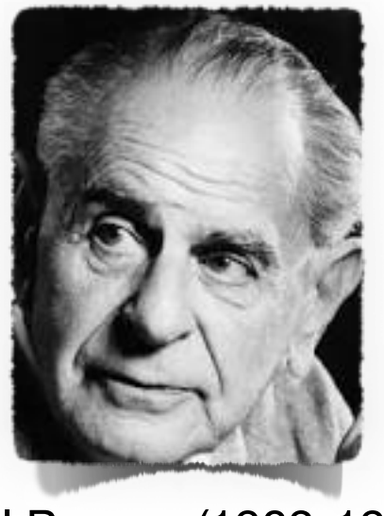


Arbitrary function x of the data that allows for separation between the two hypotheses

This is Popperian testing

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

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



Karl Popper (1902-1994)

A p-value is not a probability. It is a random variable (function of the data) that is distributed uniformly if the tested hypothesis is true.

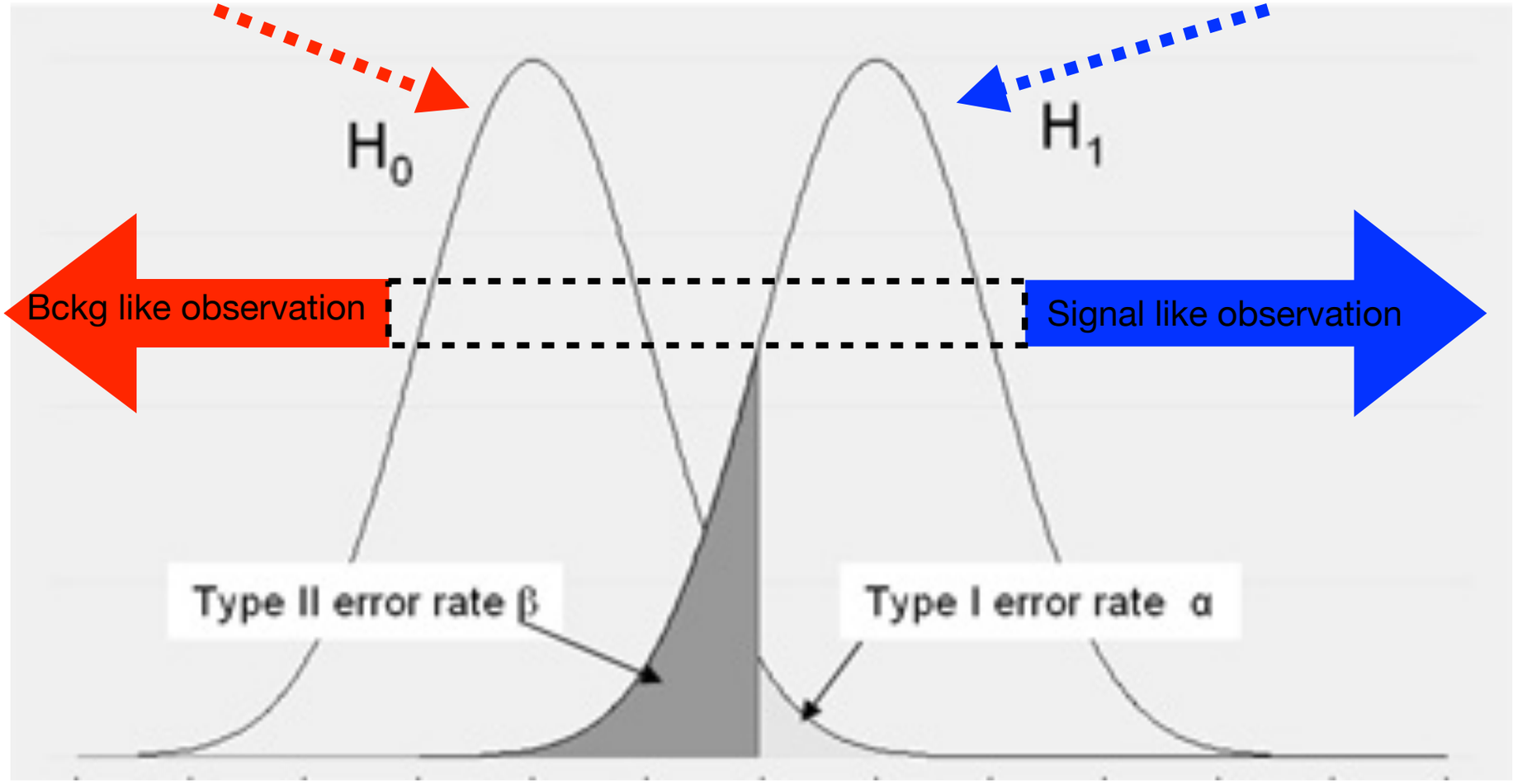
It does not express the probability that an hypothesis is true or false!
Wrong claim "The measurement shows that the probability for hypothesis blah is .."

P-values connect to the probability to observe x_{obs} or a more extreme value *if a specific hypothesis were true*. Proper claim: "Assuming that the hypothesis blah holds, the probability to observe a fluctuation as extreme as that observed in our data or more is…"

One-slide recap

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

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



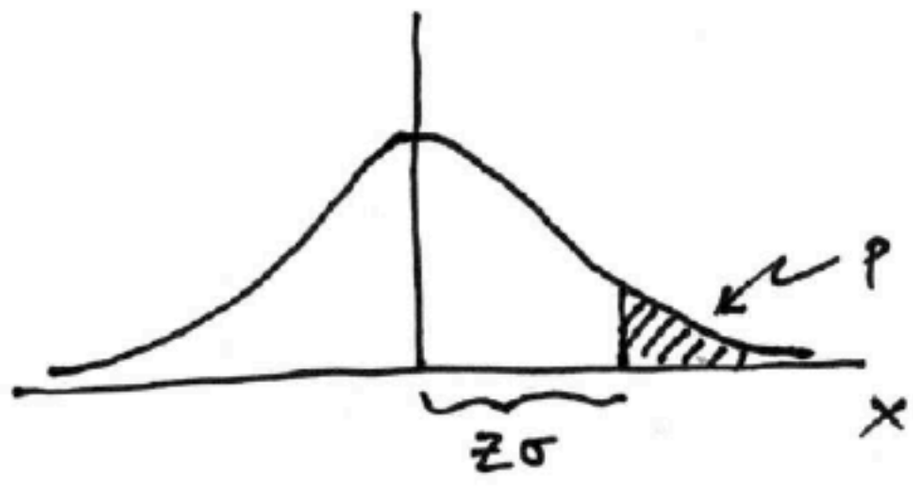
This is x, whatever function of data whose distribution is sensitive to separate H0 from H1

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

“Significance”

“At how many sigma such and such result is significant?”

The “number of sigma” (or z-value) is just a remapping of p-values into integrals of one tail of a Gaussian. It expresses by how many sigma away of mean of my observation would be if the test statistic x would be distributed as Gaussian



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

$$Z = \Phi^{-1}(1-p)$$ `TMath::NormQuantile`

Examples: p-values in coin tossing

Check if a coin is fair. The probability to observe j heads in n trials is binomial

$$f(j;n,p) = \binom{n}{j} p^j (1-p)^{n-j} = \frac{n!}{(n-j)!j!} p^j (1-p)^{n-j}$$

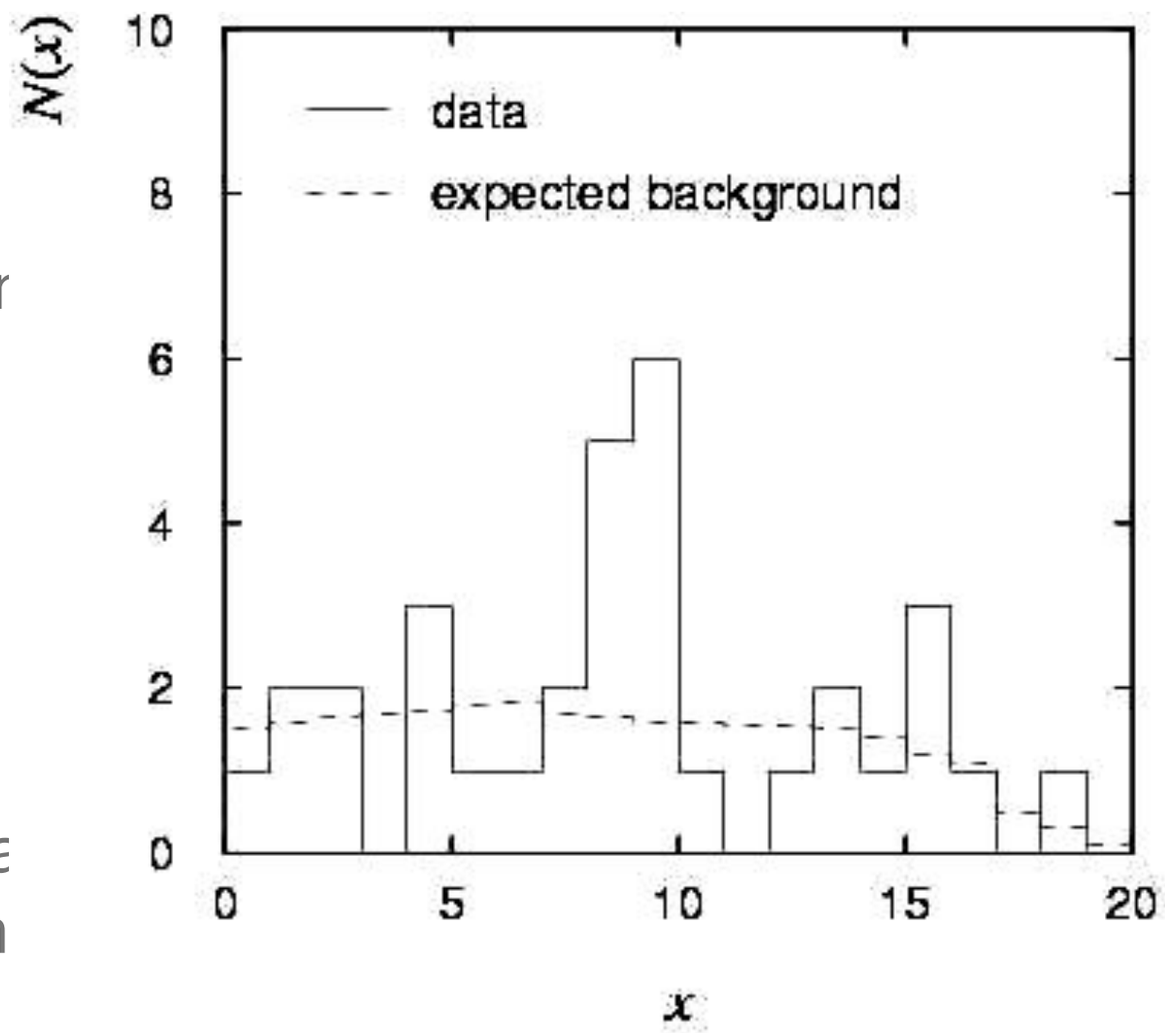
Null hypothesis: the coin is fair (p=0.5). Get 17 heads out of 20 trials. Regions of data space with equal or lesser compatibility with null, relative to j=17 include n=17, 18, 19, 20, *0, 1, 2, 3.*

P(n=0,1,2,3,17,18,19,or 20) = 0.26%.

Hence, if the null were true (coin is fair) and we would repeat the experiment many times, only 0.26% of the times we would obtain a result as extreme or more than that observed.

p-values in mass peak

Suppose you measure a value x for each event and bin the resulting distribution.



The count in each bin is a Poisson randor variable, whose mean in the bck-only hypothesis is given by the dashed line

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

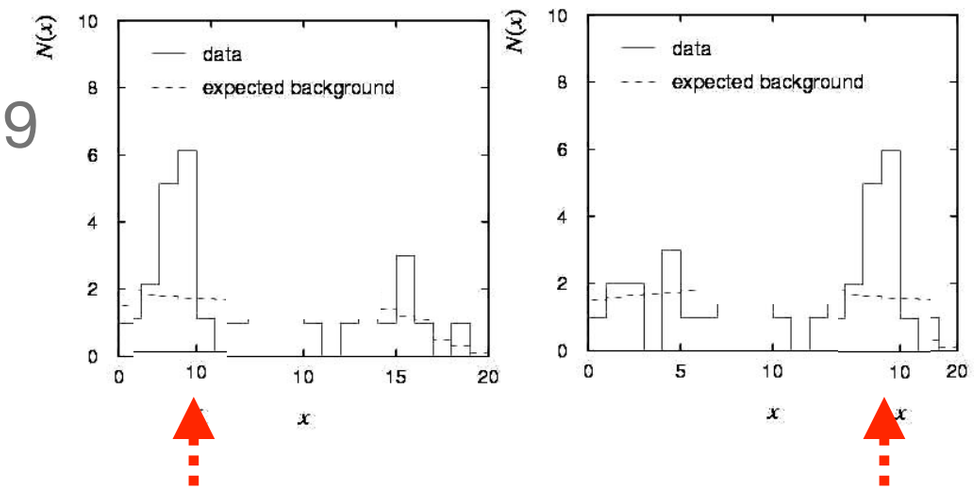
Observe a peak of 11 events in the centra bins, with expected background 3.2 even

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

Is this evaluation fair or biased?

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

That evaluation accounts for the chances of an upward fluctuation only in that very position at x~9 where I observed it. That’s the “local p-value”.



“global p-value” need to account for the chances that an excess could have arisen in any pair of adjacent bins. With 20 bins (10 pairs of adjacent bins) the local p-value gets multiplied by ≈10.

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

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

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

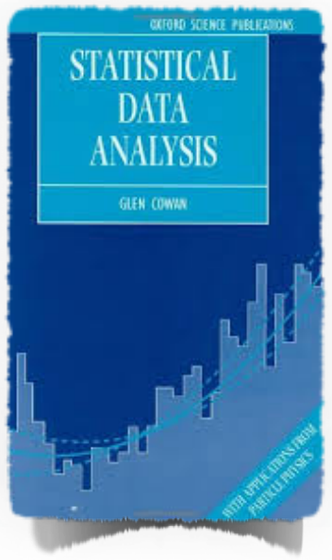
Look elsewhere, monkeys, and Shakespeare

A trillion monkeys typing on a trillion typewriters will, sooner or later, reproduce the works of William Shakespeare.

Don't be a monkey.

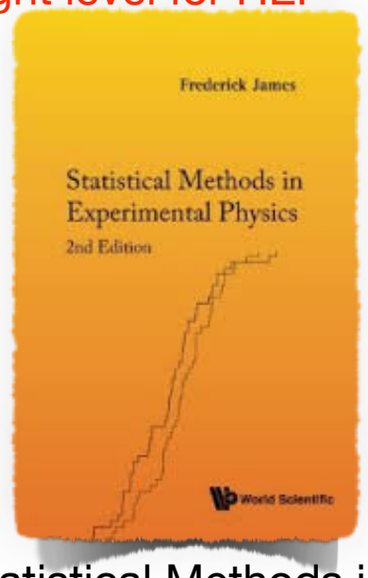
Sources - books

- Good starting point



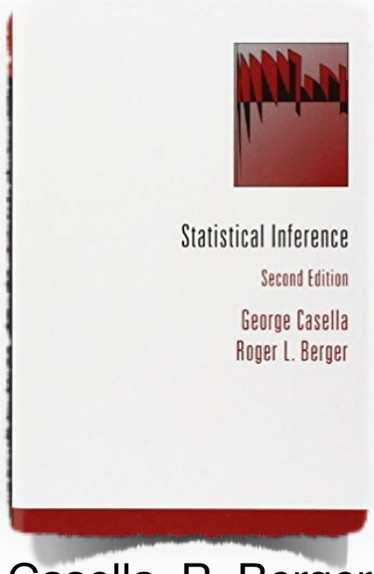
G. Cowan,“Statistical data analysis”

- Very good book at the right level for HEP



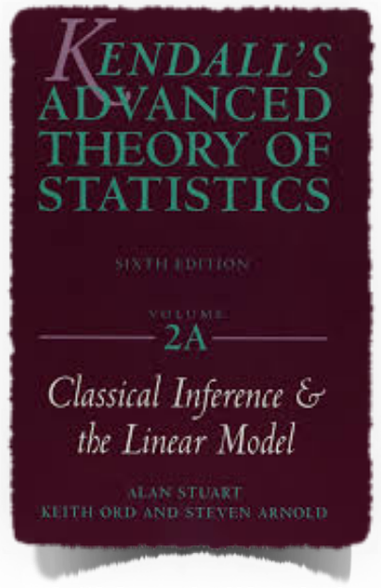
F. James, “Statistical Methods in Experimental Physics, data analysis”

- Advanced book



G. Casella, R. Berger, “Statistical Inference”

- Ultimate bible



A. Stuart, et al “Kendall’s Advanced Theory of Statistics Vol 2A”

If you are serious about statistics, get the Cowan and James, and consult the other two in the library when needed.

Further readings — lecture slides/docs

- 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.
- 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: https://arxiv.org/abs/1807.05996. Look at his statistics papers on http://inspirehep.net/help/easy-search and at the references he reccommends. Advanced
- Proceedings/docs from the PHYSTAT conferences and workshops, linked from phystat.org Advanced

Thanks for your attention