BAYES and FREQUENTISM:
The Return of an Old Controversy

Louis Lyons

Oxford & Imperial College

DESY July 2020

Topics

- Who cares?
- What is probability?
- Bayesian approach
- Examples
- Frequentist approach
- Summary

. Will discuss mainly in context of **PARAMETER ESTIMATION**. Also important for **GOODNESS of FIT** and **HYPOTHESIS TESTING**

It is possible to spend a lifetime analysing data without realising that there are two very different fundamental approaches to statistics:

Bayesianism and Frequentism.

For simplest case $(m \pm \sigma) \leftarrow Gaussian$

with no constraint on μ_{true} , then

$$m - k\sigma < \quad \mu_{true} \quad < m + k\sigma$$

at some probability, for both Bayes and Frequentist

(but different interpretations)

See Bob Cousins "Why isn't every physicist a Bayesian?" Amer Jrnl Phys 63(1995)398

We need to make a statement about

Parameters, Given Data

The basic difference between the two:

Bayesian : Prob(parameter, given data)
(an anathema to a Frequentist!)

Frequentist : Prob(data, given parameter)
(a likelihood function)

WHAT IS PROBABILITY?

MATHEMATICAL

Formal

Based on Axioms

FREQUENTIST

Ratio of frequencies as $n \rightarrow$ infinity

Repeated "identical" trials

Not applicable to **single event** or **physical constant**

BAYESIAN Degree of belief

Can be applied to single event or physical constant

(even though these have unique truth)

Varies from person to person ***

Quantified by "fair bet"

LEGAL PROBABILITY

Bayesian versus Classical

Bayesian

$P(A \text{ and } B) = P(A;B) \times P(B) = P(B;A) \times P(A)$

e.g. A = event contains t quark

B = event contains W boson

or A = a random day is in December

B = a random day is rainy

$P(A;B) = P(B;A) \times P(A) / P(B)$

Completely uncontroversial, provided….

Bayesian

$$P(A;B) = \frac{P(B;A) \times P(A)}{P(B)}$$

Bayes' Theorem

p(param | data) α p(data | param) * p(param)

↑ posterior ↑ likelihood ↑ prior

Problems: p(param) Has particular value

"Degree of belief"

Prior What functional form?

Coverage

P(parameter) Has specific value

"Degree of Belief"

Credible interval

Prior: What functional form?

Uninformative prior: flat?

In which variable? e.g. m, m^2, ln m,….?

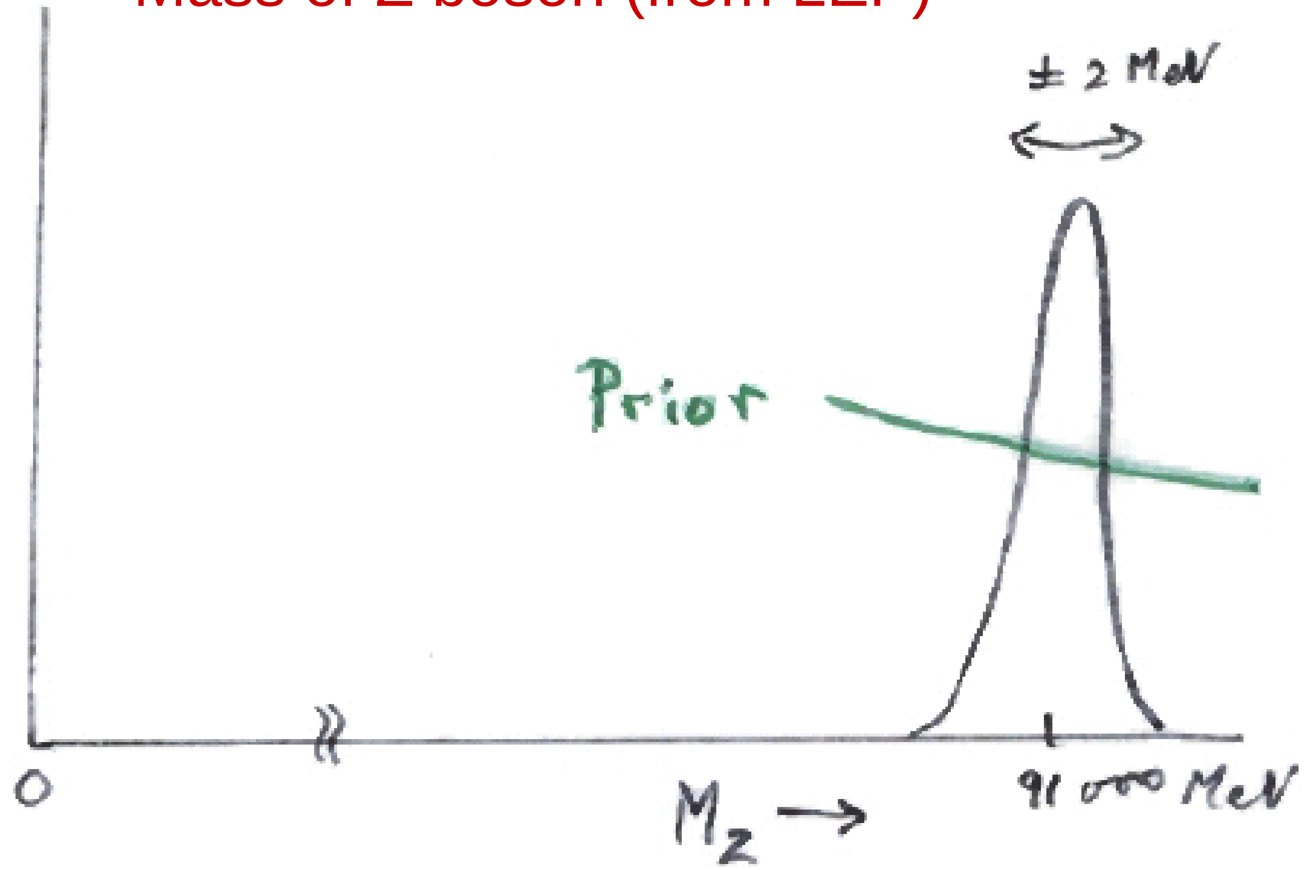
Even more problematic with more params

Unimportant if "data overshadows prior"

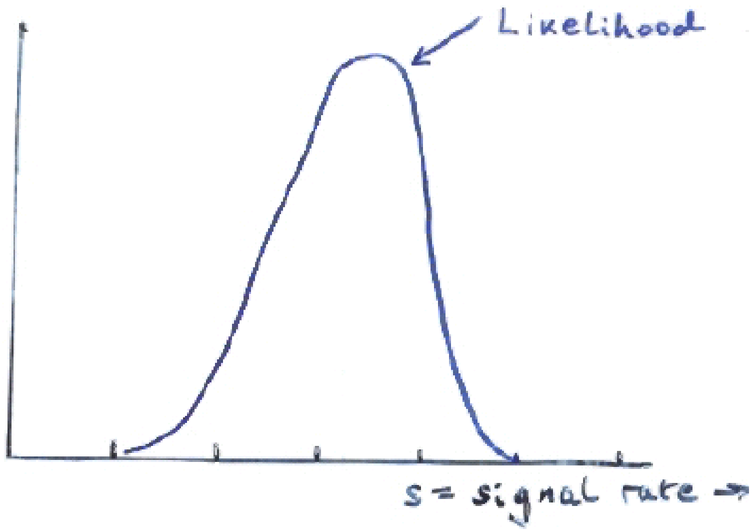
Important for limits

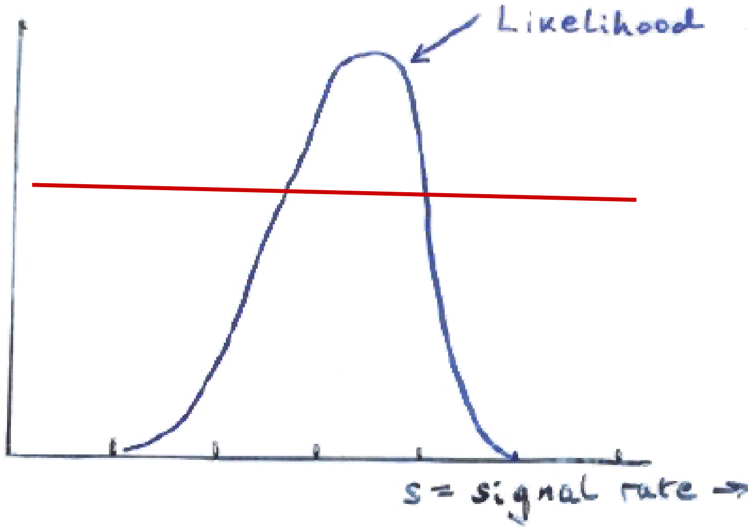
Subjective or Objective prior?

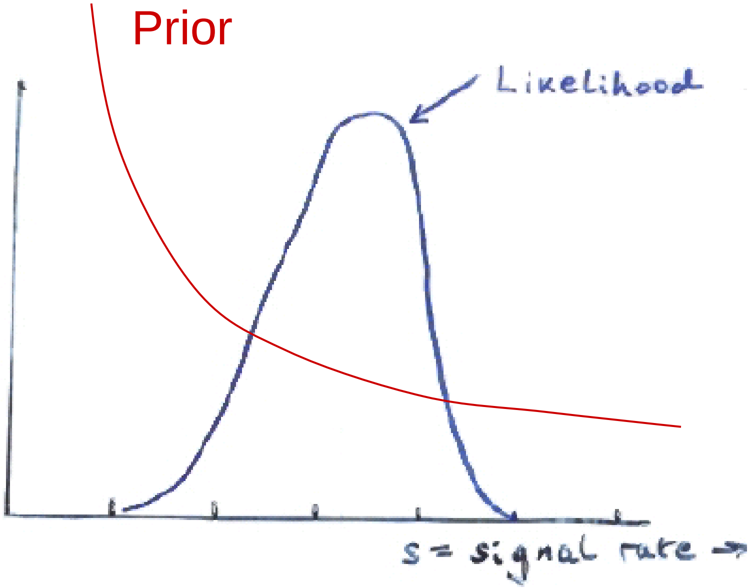
Mass of Z boson (from LEP)



Data overshadows prior

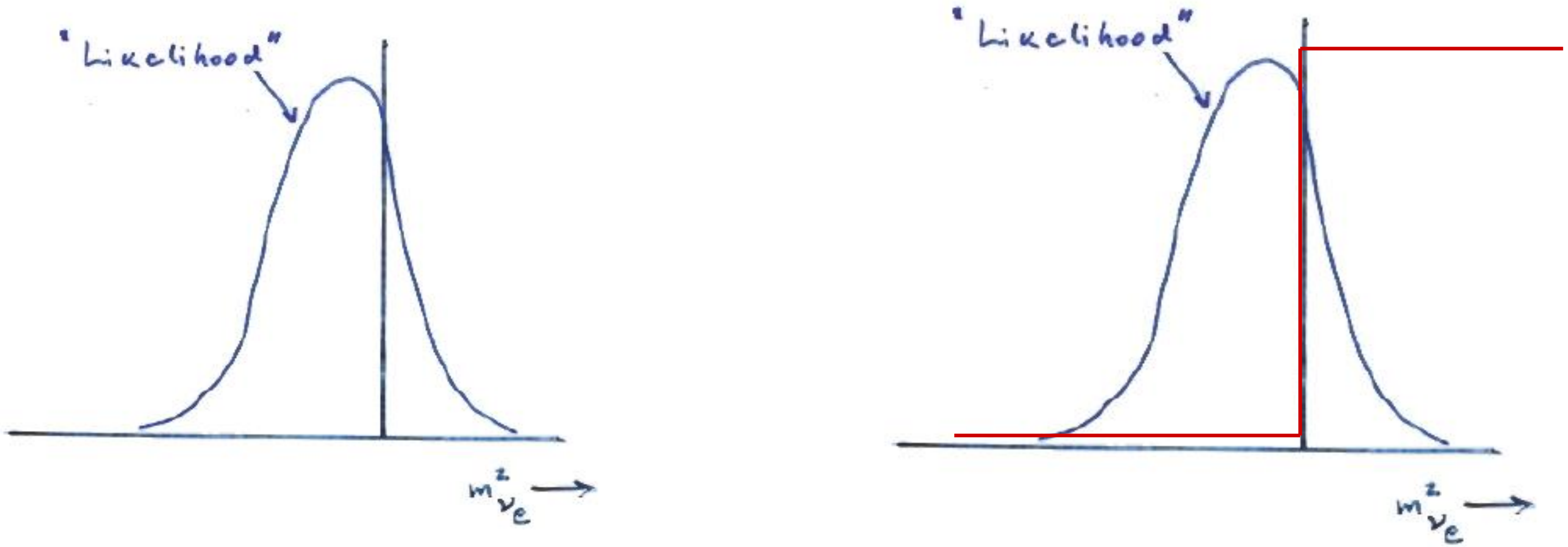






Even more important for UPPER LIMITS

Mass-squared of neutrino



Prior = zero in unphysical region

Bayes: Specific example

Particle decays exponentially: $dn/dt = (1/\tau) \exp(-t/\tau)$

Observe 1 decay at time t_1: $\mathcal{L}(\tau) = (1/\tau) \exp(-t_1/\tau)$

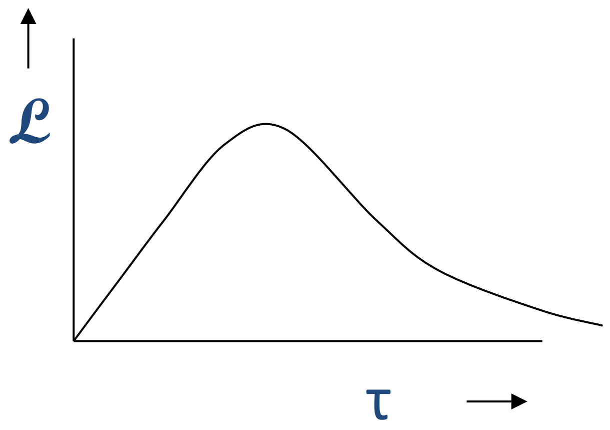
Choose prior $\pi(\tau)$ for τ

e.g. constant up to some large τ

Then posterior $p(\tau) = \mathcal{L}(\tau) * \pi(\tau)$

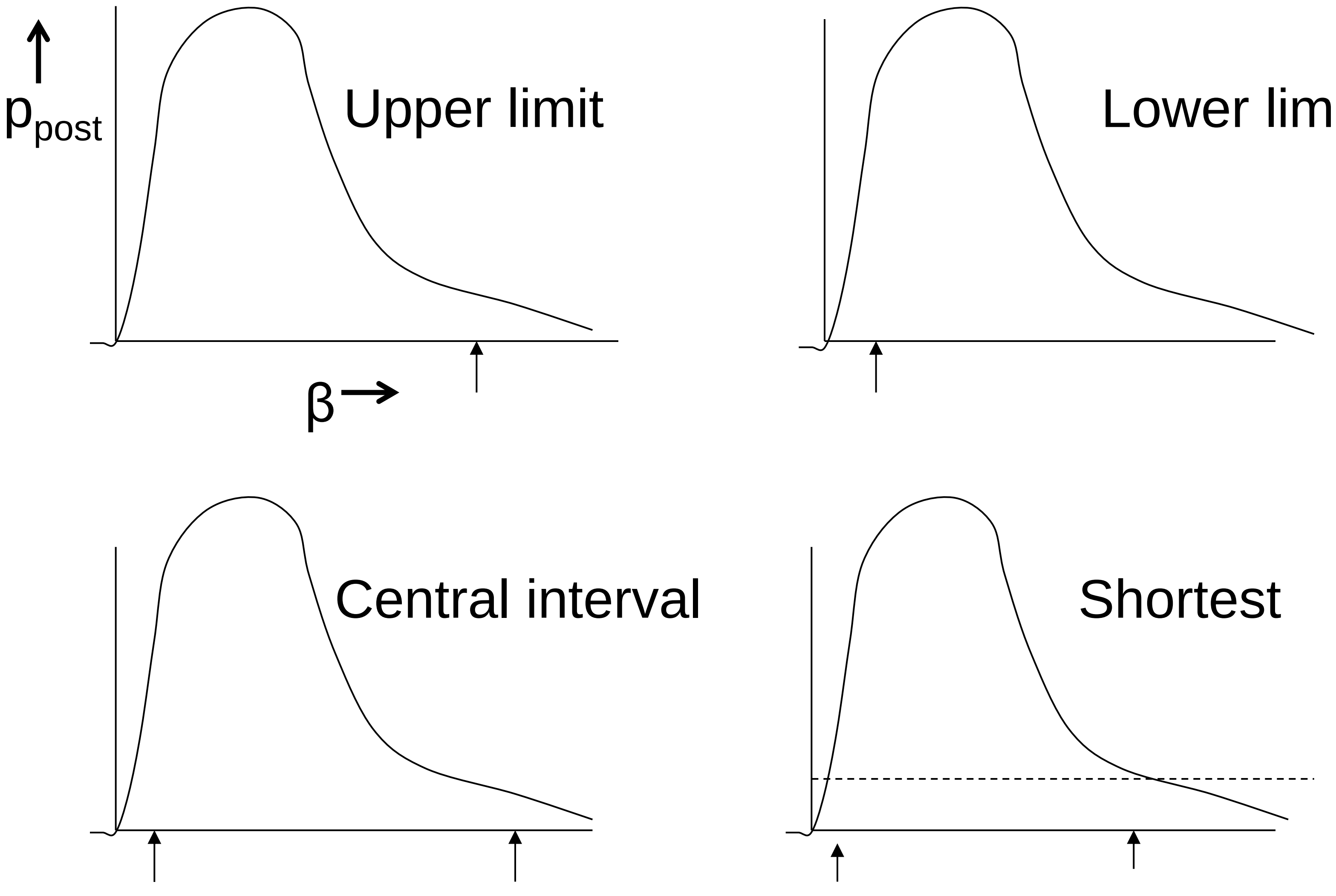
has almost same shape as $\mathcal{L}(\tau)$

Use $p(\tau)$ to choose interval for τ in usual way



Contrast frequentist method for same situation later.

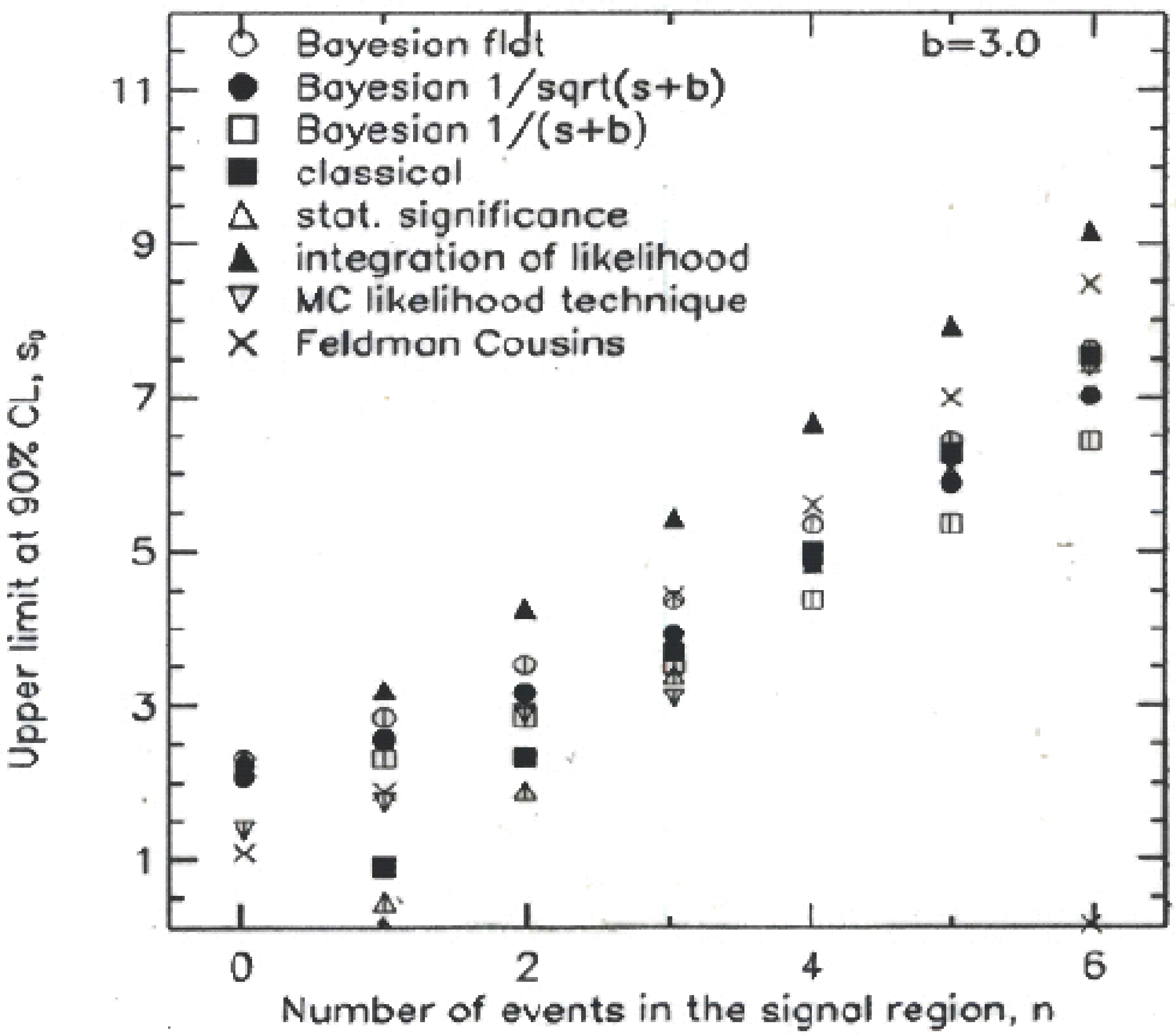
Bayesian posterior → intervals



Ilya Narsky, FNAL CLW 2000

Upper Limits from Poisson data

Expect b = 3.0, observe n events



Upper Limits important for excluding models

P (Data;Theory) $\neq$ P (Theory;Data)

HIGGS SEARCH at CERN in 2000

Is data consistent with:

Null Hyp: Standard Model, but no Higgs; or

Alternative Hyp: Standard Model + Higgs?

End of Sept 2000: Data not very consistent with 'No H'

Prob (Data ; No H) < 1% valid frequentist statement

Turned by the press into: Prob (No Higgs ; Data) < 1%
and therefore Prob (Higgs ; Data) > 99%

i.e. "It is almost certain that the Higgs has been seen"

$P\,(\text{Data};\text{Theory}) \neq P\,(\text{Theory};\text{Data})$ *

* Certainly true for Frequentists, as

P(Theory) is not allowed

$$P\,(\text{Data};\text{Theory}) \neq P\,(\text{Theory};\text{Data})$$

Theory = male or female

Data = pregnant or not pregnant

P (pregnant ; female) ~ 3%

P (Data;Theory) $\neq$ P (Theory;Data)

Theory = male or female

Data = pregnant or not pregnant

P (pregnant ; female) ~ 3%

but

P (female ; pregnant) >>>3%

Example 1 : Is coin fair ?

Toss coin: 5 consecutive tails

What is P(unbiased; data) ? i.e. $p = ½$

Depends on Prior(p)

If village priest: prior ~ $\delta(p = 1/2)$

If stranger in pub: prior ~ 1 for $0 < p < 1$

(also needs cost function)

Example 2 : Particle Identification

Try to separate π's and protons

probability (p tag; real p) = 0.95

probability (π tag; real p) = 0.05

probability (p tag; real π) = 0.10

probability (π tag; real π) = 0.90

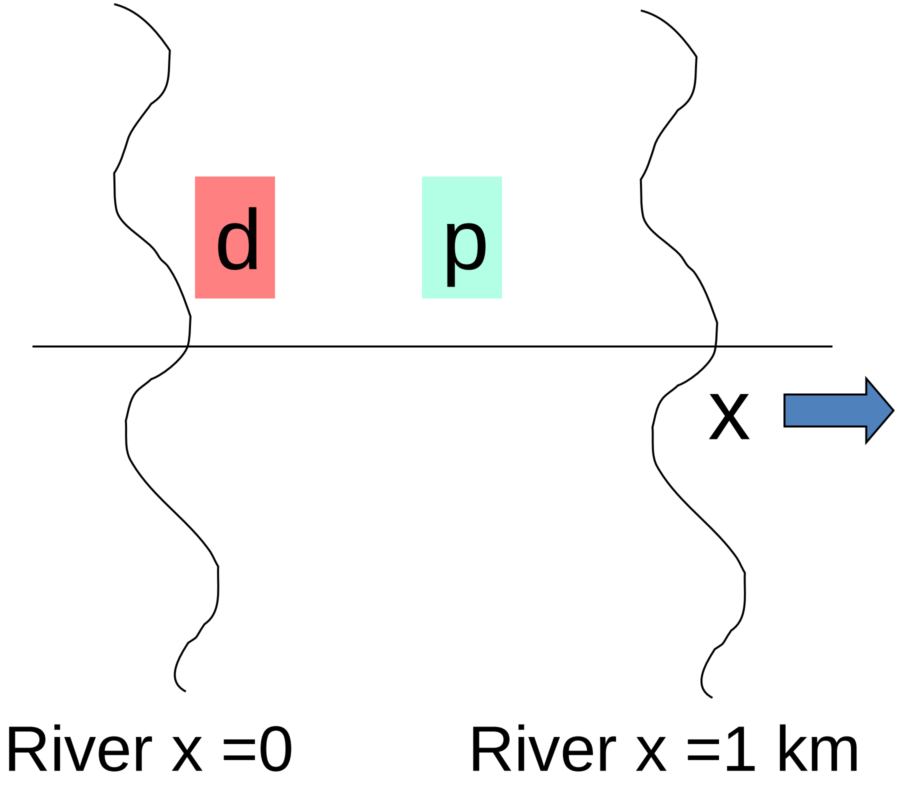
Particle gives proton tag. What is it?

Depends on prior = fraction of protons

If proton beam,	very likely
If general secondary particles,	more even
If pure π beam,	~ 0

Peasant and Dog

1) Dog d has 50% probability of being 100 m. of Peasant p
2) Peasant p has 50% probability of being within 100m of Dog d ?



Given that: a) Dog d has 50% probability of being 100 m. of Peasant,

is it true that: b) Peasant p has 50% probability of being within 100m of Dog d ?

Additional information

- Rivers at zero & 1 km. Peasant cannot cross them.

$$0 \le h \le 1\,\text{km}$$

- Dog can swim across river - Statement a) still true

If dog at –101 m, Peasant cannot be within 100m of dog

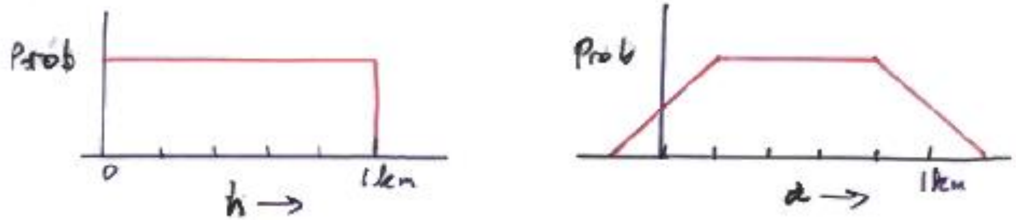
Statement b) untrue

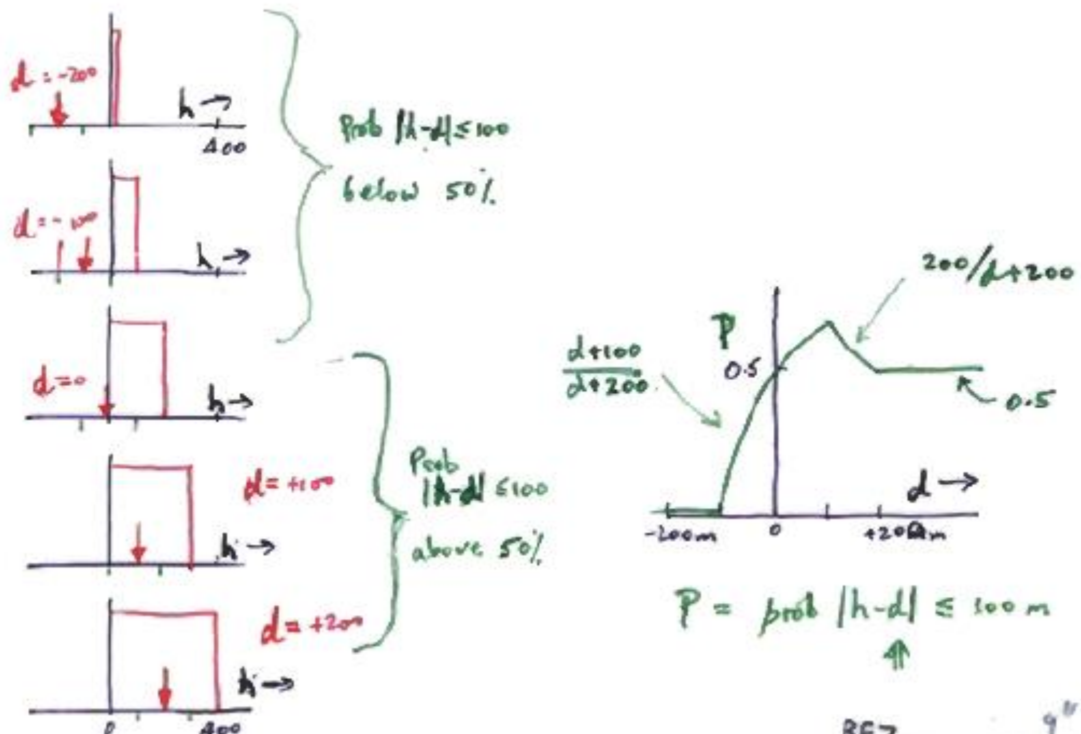
1) More specific on statement ①:

$$\text{Prob}(d - h) = \begin{cases} \text{const} & \text{for } |d-h| < 200\text{m} \\ 0 & \text{for } |d-h| > 200\text{m} \end{cases}$$

[L'Hoo]

2) Hunter h uniform in 0 → 1 km [PRIOR]





BF7 9''

Classical Approach

Neyman "confidence interval" avoids pdf for μ

Uses only P(x; μ)

Confidence interval $\mu_1 \rightarrow \mu_2$:

P($\mu_1 \rightarrow \mu_2$ contains μ_t) = α True for any μ_t

Varying intervals from ensemble of experiments

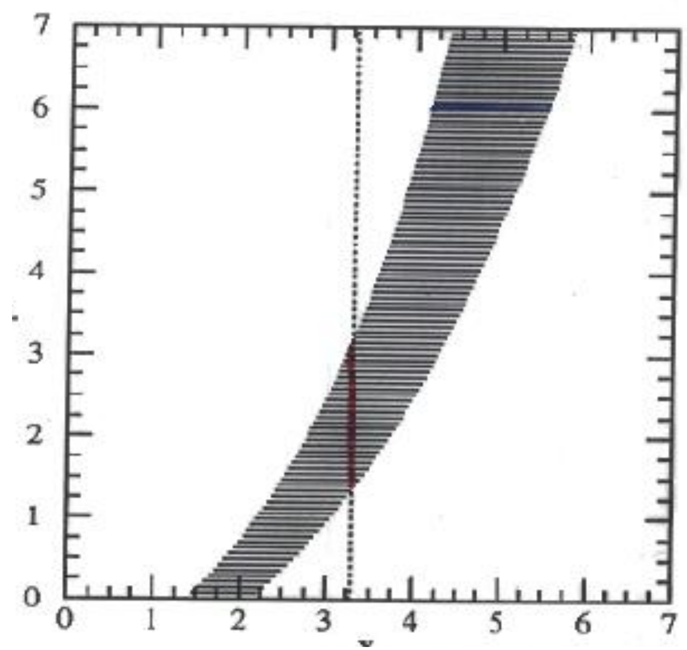
fixed

Gives range of μ for which observed value x_0 was "likely" (α)

Contrast Bayes : Degree of belief = α that μ_t is in $\mu_1 \rightarrow \mu_2$

Classical (Neyman) Confidence Intervals

Uses only P(data|theory)



Theoretical Parameter μ

Observation x →

FIG. 1. A generic confidence belt construction and its use. For each value of μ, one draws a horizontal acceptance interval $[x_1, x_2]$ such that $P(x \in [x_1, x_2] | \mu) = \alpha$. Upon performing an experiment to measure x and obtaining the value x_0, one draws the dashed vertical line through x_0. The confidence interval $[\mu_1, \mu_2]$ is the union of all values of μ for which the corresponding acceptance interval is intercepted by the vertical line.

μ≥0

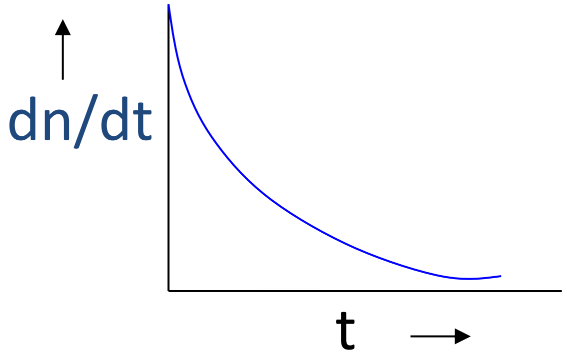
No prior for μ

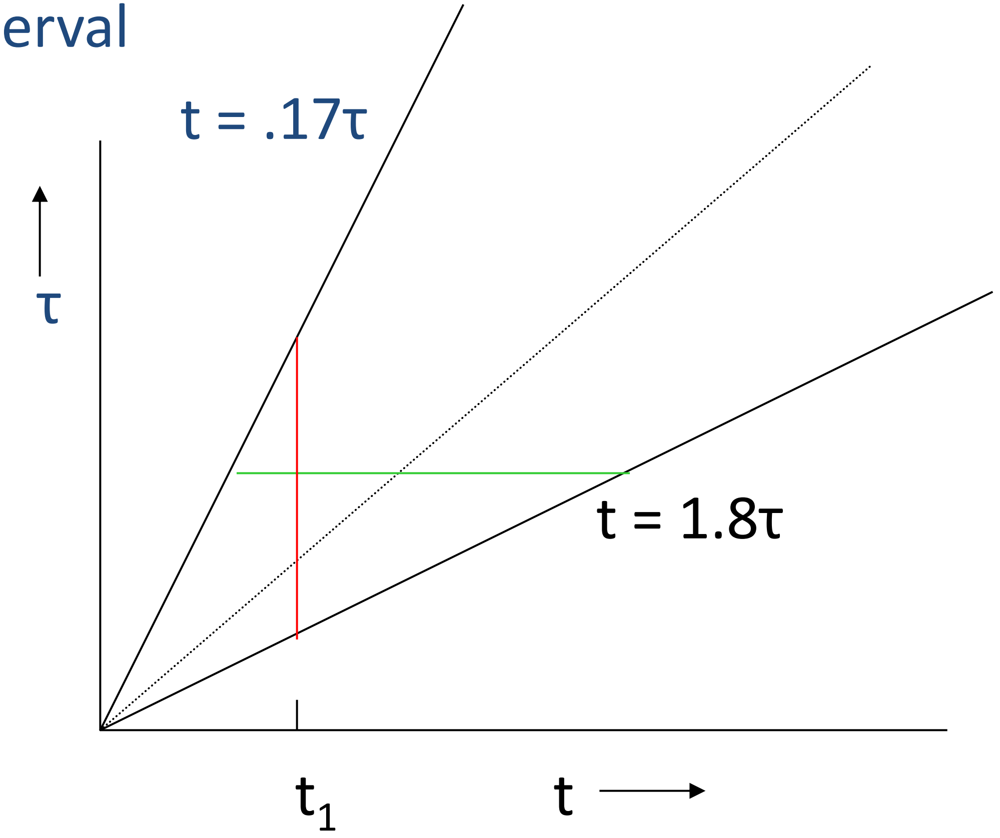
Frequentism: Specific example

Particle decays exponentially: $dn/dt = (1/\tau)\exp(-t/\tau)$

Observe 1 decay at time t_1: $\mathcal{L}(\tau) = (1/\tau)\exp(-t_1/\tau)$

Construct 68% central interval





68% conf. int. for τ from $t_1/1.8 \rightarrow t_1/0.17$

90% Classical interval for Gaussian

Measurement x   σ = 1   True value $\mu \geq 0$

e.g. $m^2(\nu_e)$,   length of small object

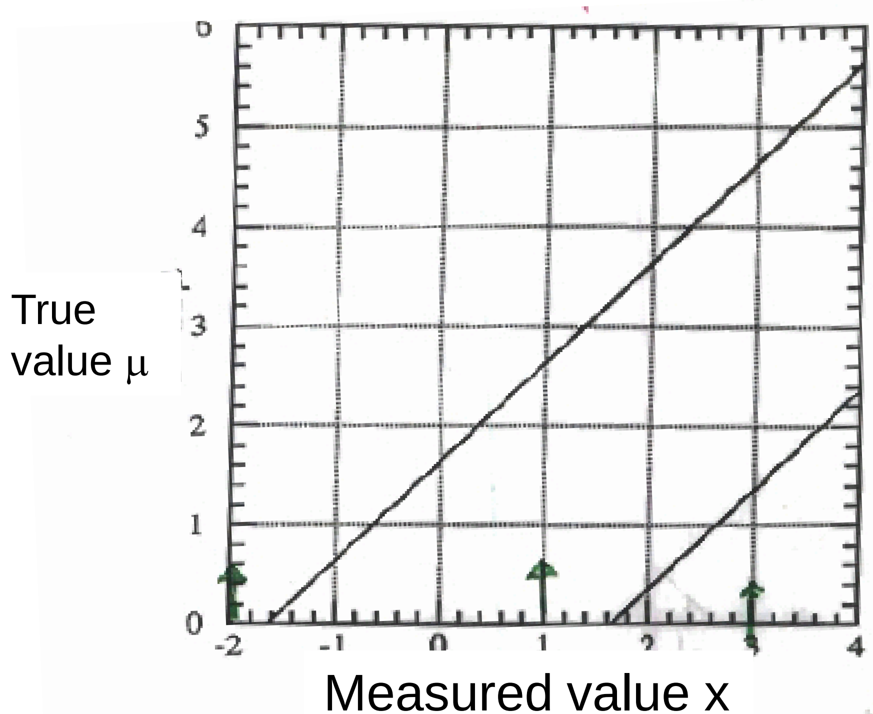


FIG. 3. Standard confidence belt for 90% C.L. central confidence intervals for the mean of a Gaussian, in units of the rms deviation.

x_{obs}=3 Two-sided range
x_{obs}=1 Upper limit
x_{obs}=-1 No region for μ

Other methods have different behaviour at negative x

$\mu_l \le \mu \le \mu_u$ at 90% confidence

Frequentist

μ_l and μ_u known, but random

μ unknown, but fixed

Probability statement about μ_l and μ_u

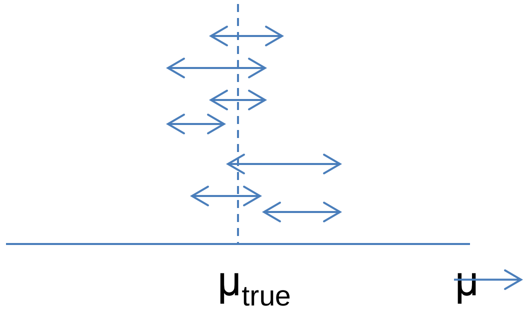
Bayesian

μ_l and μ_u known, and fixed

μ unknown, and random

Probability/credible statement about μ

Coverage



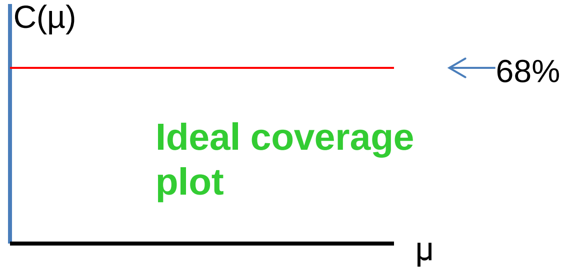
* What it is:

For given statistical method applied to many sets of data to extract confidence intervals for param μ, coverage C is fraction of ranges that contain true value of param.   Can vary with μ

* Does not apply to **your** data:

It is a property of the **statistical method** used

It is **NOT** a probability statement about whether μ_{true} lies in your confidence range for μ



* Coverage plot for Poisson counting expt

Observe n counts

Estimate μ_{best} from maximum of likelihood

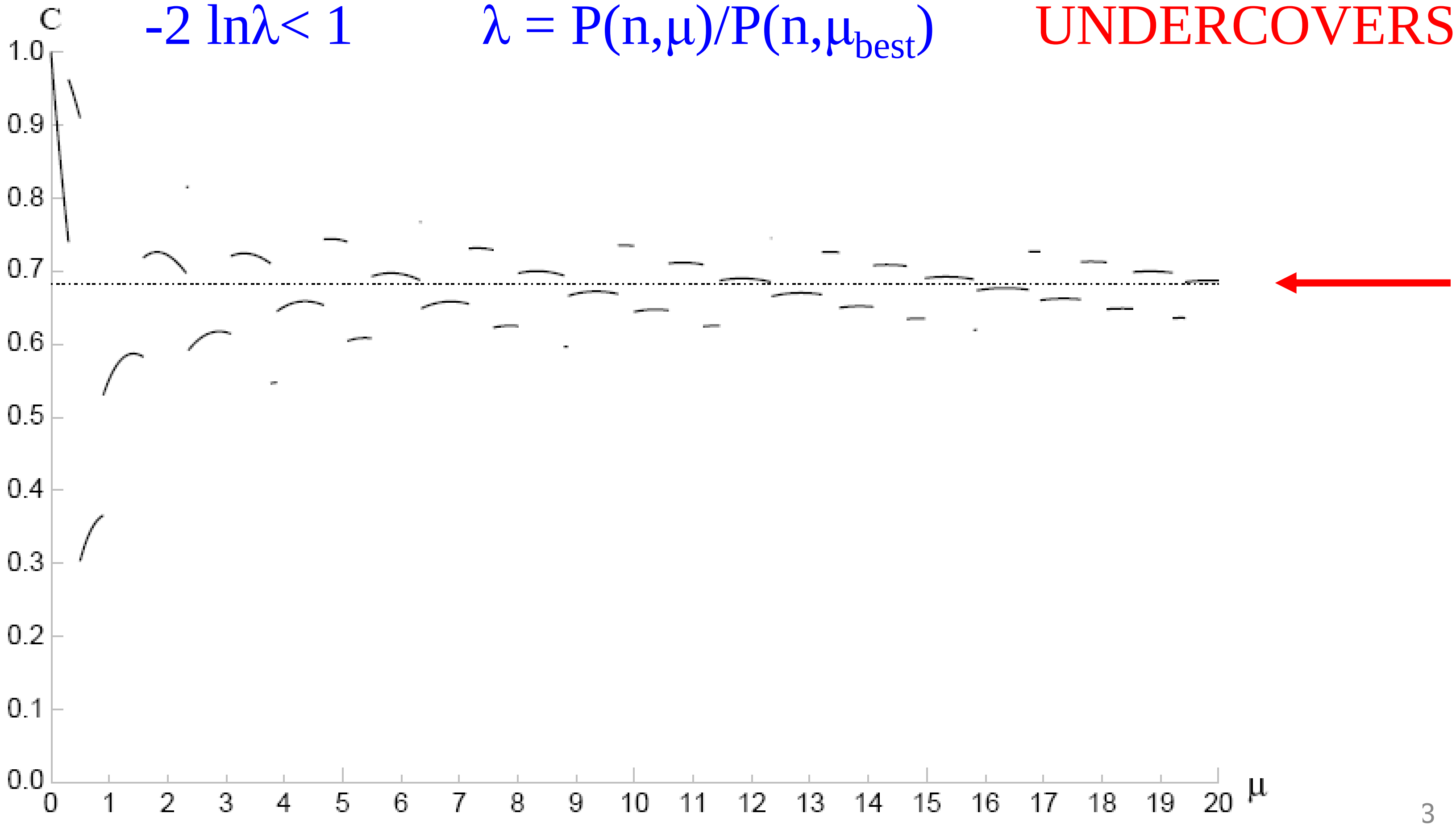
$$\mathcal{L}(\mu) = e^{-\mu}\,\mu^n/n! \quad \text{and range of } \mu \text{ from} \quad \ln\{\mathcal{L}(\mu_{best})/\mathcal{L}(\mu)\} < 0.5$$

For each μ_{true} calculate coverage $C(\mu_{true})$, and compare with nominal 68%

Coverage : $\mathcal{L}$ approach
(Not Neyman construction)

$P(n,\mu) = e^{-\mu}\mu^n/n!$ (Joel Heinrich CDF note 6438)

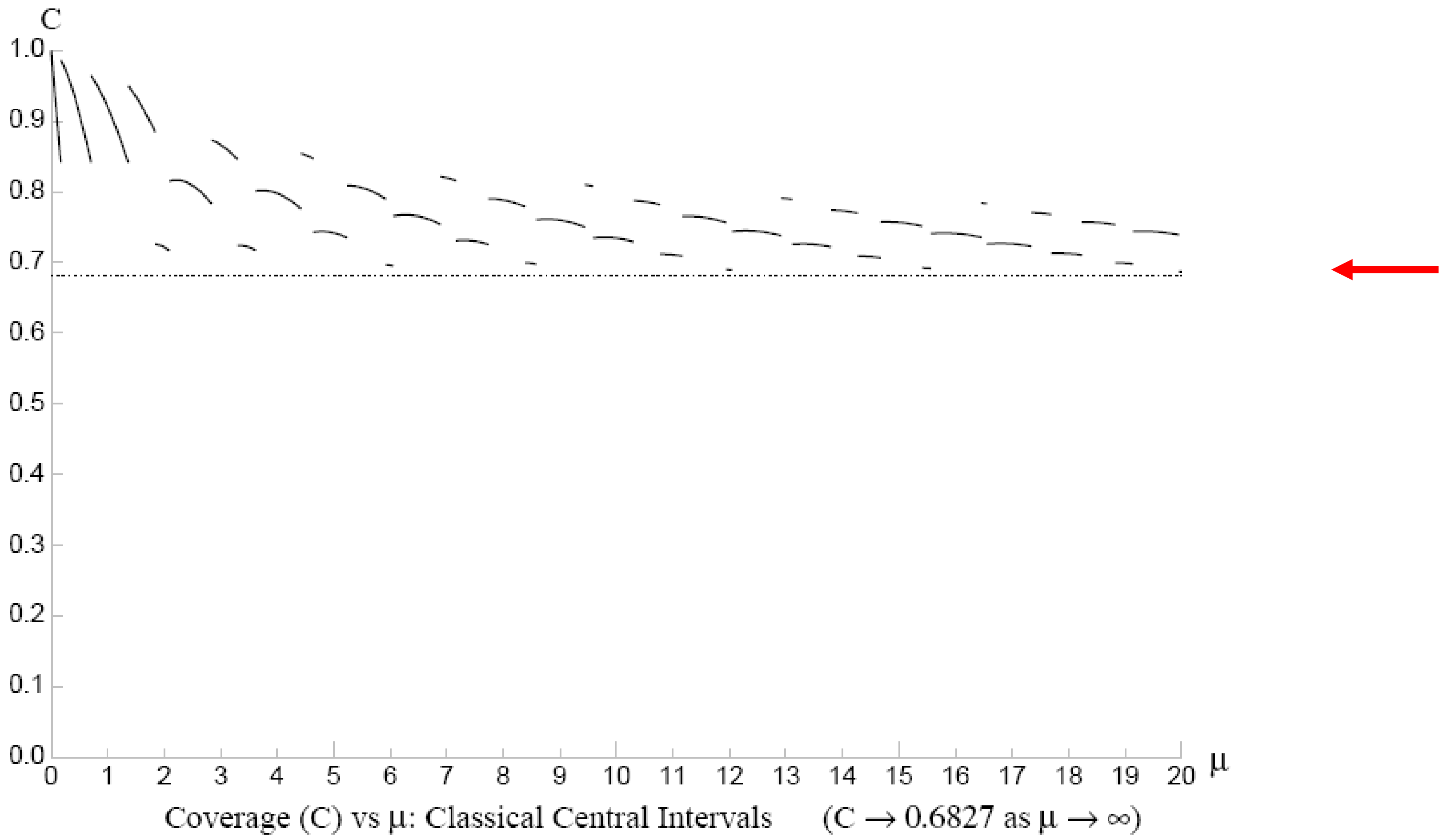
$-2\ln\lambda < 1$ $\lambda = P(n,\mu)/P(n,\mu_{best})$ UNDERCOVERS



Coverage (C) vs μ: $-2\ln\lambda < 1$ ($C \to 0.6827$ as $\mu \to \infty$)

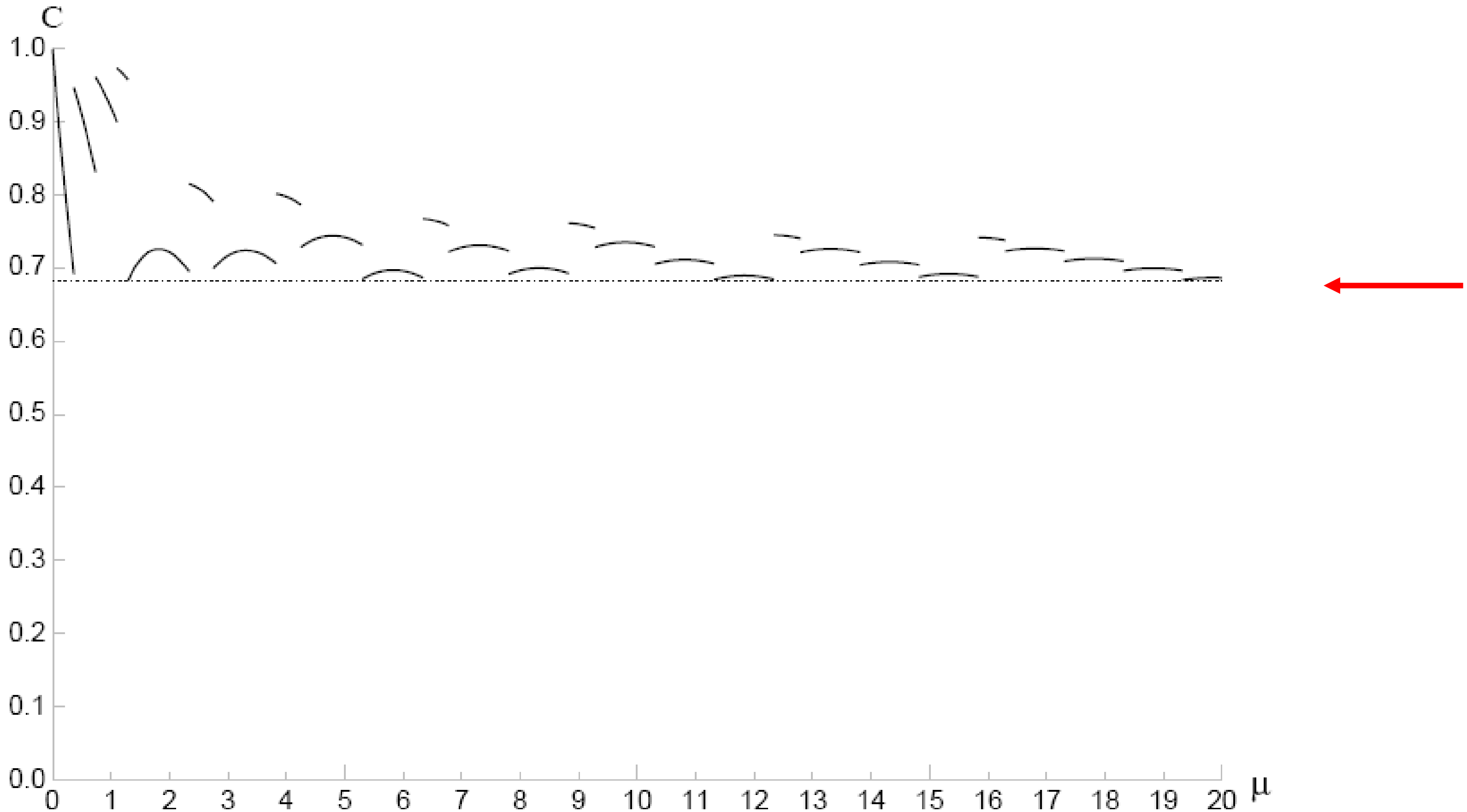
Frequentist central intervals, NEVER undercovers

(Conservative at both ends)



Feldman-Cousins Unified intervals

Neyman construction, so NEVER undercovers



Coverage (C) vs μ: Unified Intervals ($C \to 0.6827$ as $\mu \to \infty$)

FELDMAN - COUSINS

Wants to avoid empty classical intervals →

Uses "$\mathcal{L}$-ratio ordering principle" to resolve ambiguity about "which 90% region?" →

[Neyman + Pearson say $\mathcal{L}$-ratio is best for hypothesis testing]

No 'Flip-Flop' problem

Feldman-Cousins 90% Conf Int for Gaussian

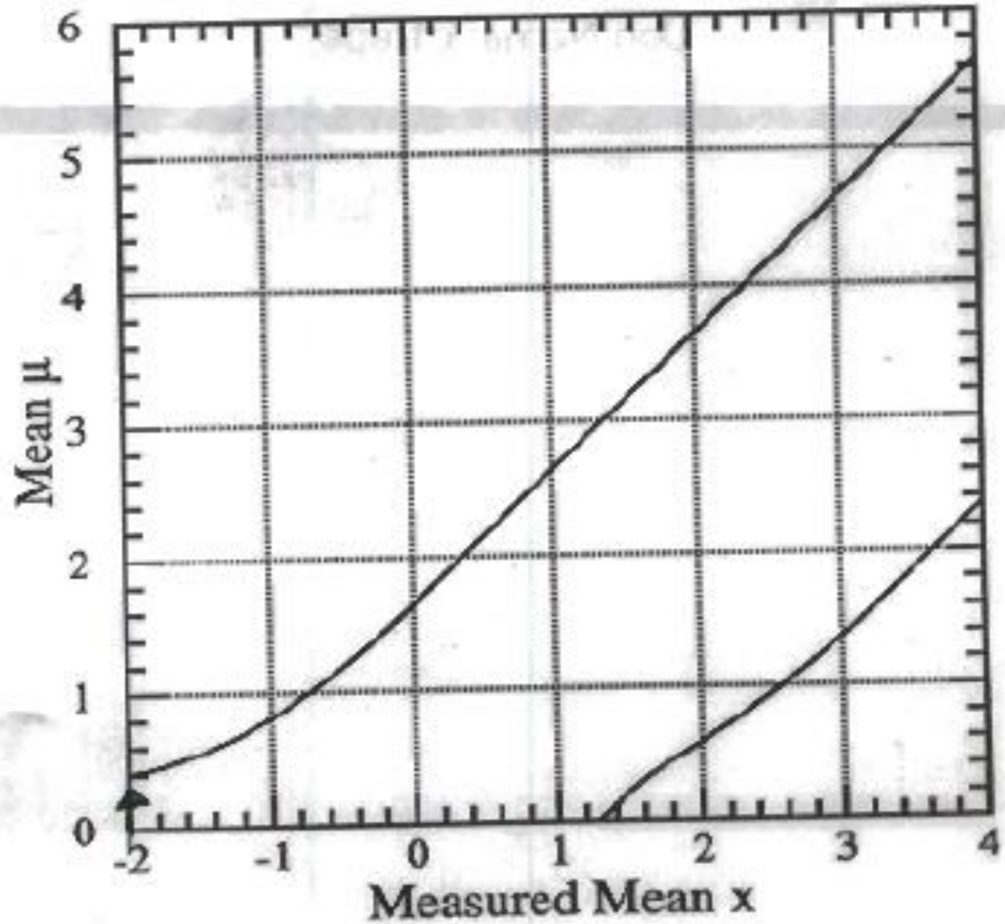


FIG. 10. Plot of our 90% confidence intervals for mean of a Gaussian, constrained to be non-negative, described in the text.

X_{obs} = -2 now gives upper limit

Poisson confidence intervals. Background = 3

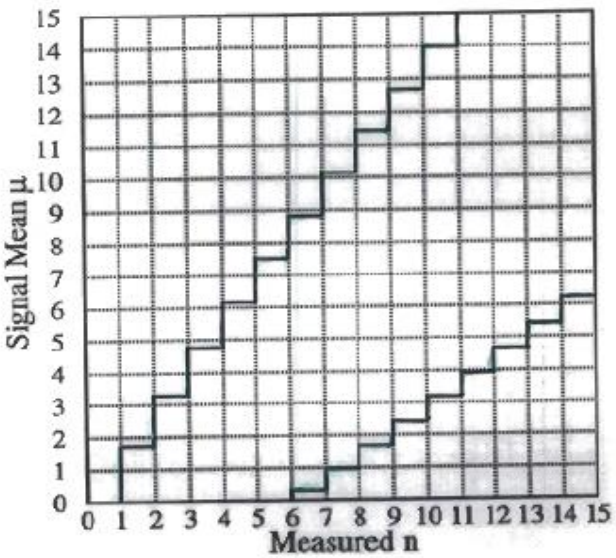


FIG. 6. Standard confidence belt for 90% C.L. central confidence intervals, for unknown Poisson signal mean μ in the presence of Poisson background with known mean $b = 3.0$.

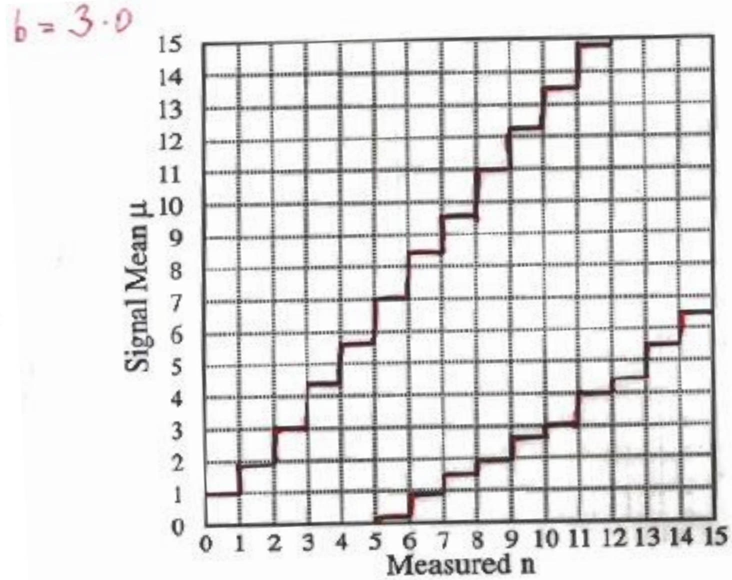


FIG. 7. Confidence belt based on our ordering principle, for 90% C.L. confidence intervals for unknown Poisson signal mean μ in the presence of Poisson background with known mean $b = 3.0$.

Standard Frequentist

Feldman - Cousins

FREQUENTIST POISSON C.B. CONSTRN.

<10% <5% Prob ordering

TABLES

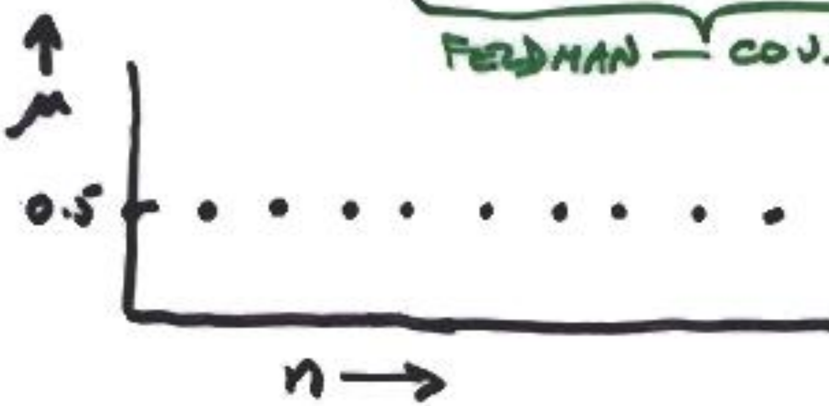
TABLE I. Illustrative calculations in the confidence belt construction for signal mean μ in the presence of known mean background $b = 3.0$. Here we find the acceptance interval for $\mu = 0.5$.

n	$P(n\|\mu)$	μ_{best}	$P(n\|\mu_{best})$	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		✓	

5 3 1 2 4/6 7

FELDMAN — COUSINS

<5%

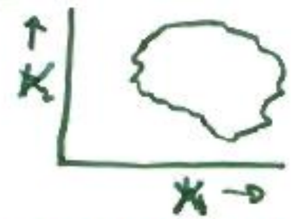


FEATURES OF F+C

REDUCES EMPTY INTERVALS

- UNIFIED 1-SIDED + 2-SIDED INTERVALS
- ELIMINATES FLIP-FLOP
- NO ARBITRARINESS OF INTERVAL

"READILY" EXTENDS TO SEVERAL DIMENSIONS



LESS OVERCOVERAGE THAN "5% AT ENDS"

MAX PROB DENSITY

5% AT ENDS ?

NEYMAN CONSTRUCTION ⇒ CPU-INTENSIVE (ESP IN SEVERAL DIMENSIONS)

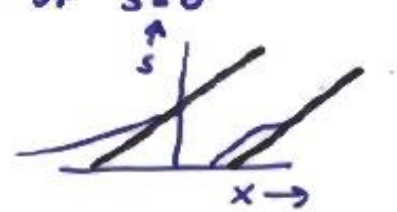
MINOR PATHOLOGIES : DISTANT INTERVALS

WRONG BEHAVIOUR WRT BGD

TIGHT LIMITS FOR $b > n_{obs}$ e.g.

n_{obs}	bgd	90% Limit
0	3.0	1.08
0	0	2.44

UNIFIED ⇒ QUICKER EXCLUSION OF s = 0



Standard Frequentist

Pros:

Coverage

Widely applicable

Cons:

Hard to understand

Small or empty intervals

Difficult in many variables (e.g. systematics)

Needs ensemble

Bayesian

Pros:

Easy to understand

Physical interval

Cons:

Needs prior

Coverage not guaranteed

Hard to combine

Bayesian versus Frequentism

	Bayesian	Frequentist
Basis of method	Bayes Theorem → Posterior probability distribution	Uses pdf for data, for fixed parameters
Meaning of probability	Degree of belief	Frequentist definition
Prob of parameters?	Yes	Anathema
Needs prior?	Yes	No
Choice of interval?	Yes	Yes (except F+C)
Data considered	Only data you have	….+ other possible data
Likelihood principle?	Yes	No

Bayesian versus Frequentism

	Bayesian	Frequentist
Ensemble of experiment	No	Yes (but often not explicit)
Final statement	Posterior probability distribution	Parameter values → Data is likely
Unphysical/ empty ranges	Excluded by prior	Can occur
Systematics	Integrate over prior	Extend dimensionality of frequentist construction
Coverage	Unimportant	Built-in
Decision making	Yes (uses cost function)	Not useful

Bayesianism versus Frequentism

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

"Frequentists use impeccable logic to deal with an issue of no interest to anyone"

Approach used at LHC

Recommended to use both Frequentist and Bayesian approaches

If agree, that's good

If disagree, see whether it is just because of different approaches