BAYES versus FREQUENTISM

The Return of an Old Controversy

- The ideologies, with examples
- Upper limits
- Systematics

Louis Lyons, Oxford University
and CERN

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

Bayesianism and Frequentism.

How can textbooks not even mention

Bayes/ Frequentism?

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

with no constraint on $m(true)$ then

$$m - k\sigma < m(true) < m + k\sigma$$

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

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

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"

Bayesian versus Classical

Bayesian

P(A and B) = P(A;B) x P(B) = P(B;A) x P(A)

e.g. A = event contains t quark

B = event contains W boson

or A = you are in CERN

B = you are at Workshop

Completely uncontroversial, provided….

P(A;B) = P(B;A) x P(A) /P(B)

Bayesian

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

Bayes Theorem

$$P(hyothesis; data) \; \alpha \; P(data; hypothesis) \text{ x } P(hypothesis)$$

↑ posterior ↑ likelihood ↑ prior

Problems: P(hyp..) true or false

"Degree of belief"

Prior What functional form?

Coverage

Goodness of fit

P(hypothesis.....)   True or False

"Degree of Belief"

credible interval

Prior: What functional form?

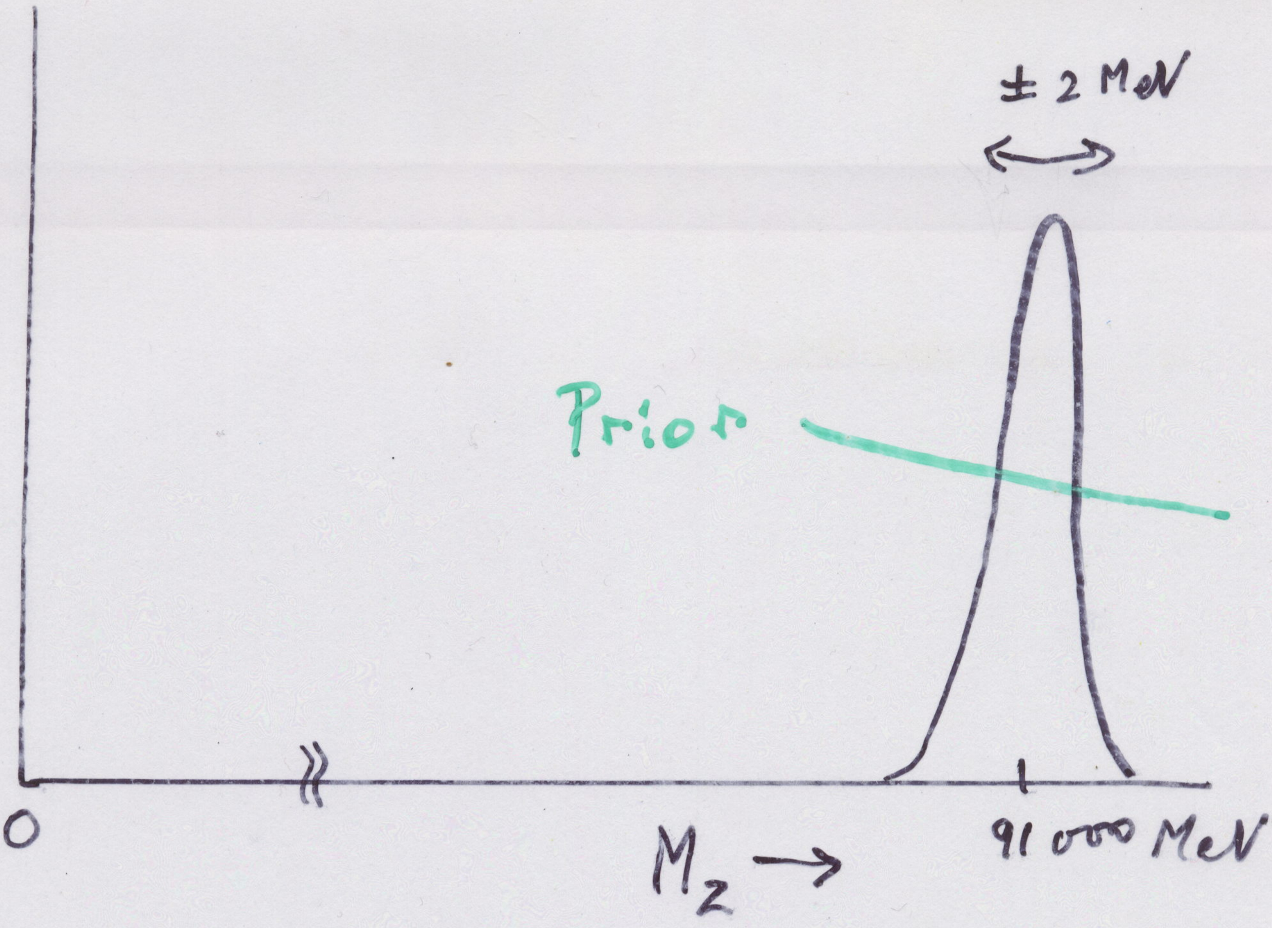
Uninformative prior:

flat? In which variable? e.g. $m, m^2, \ln m, ...?$

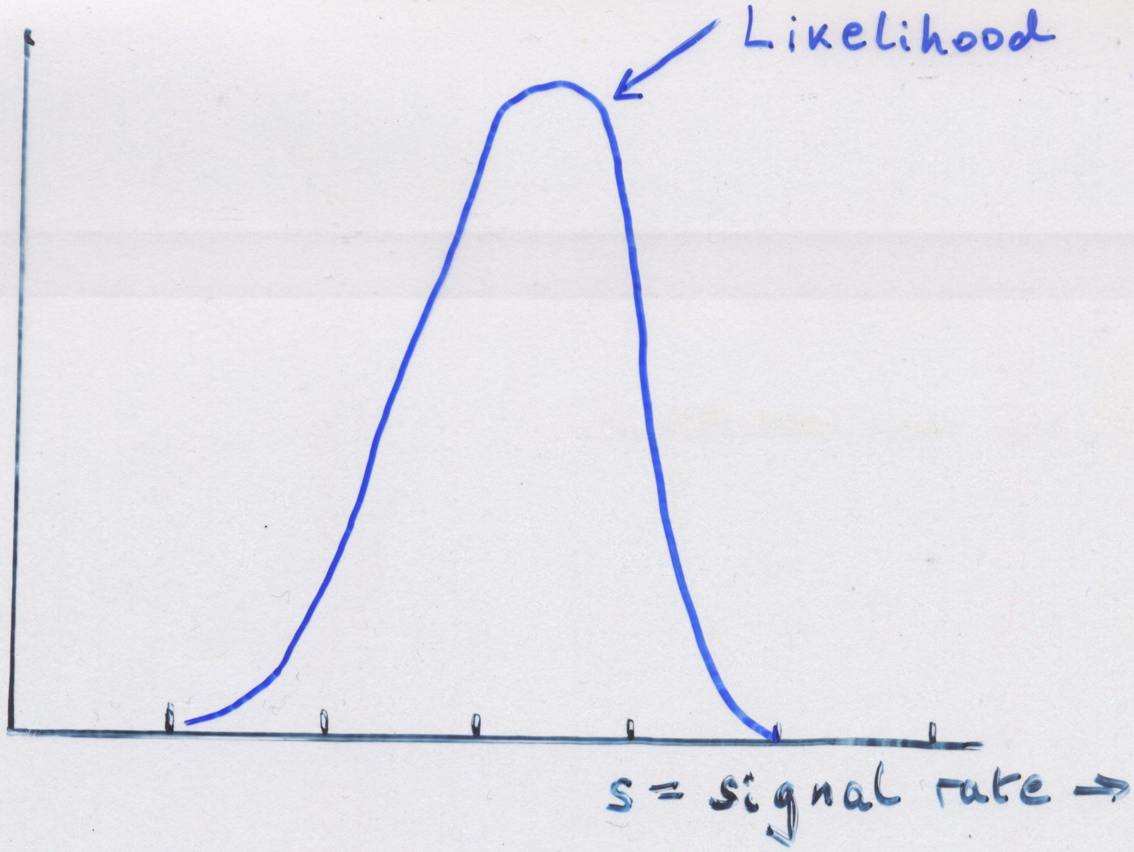
Unimportant if "data overshadows prior"

Important for limits

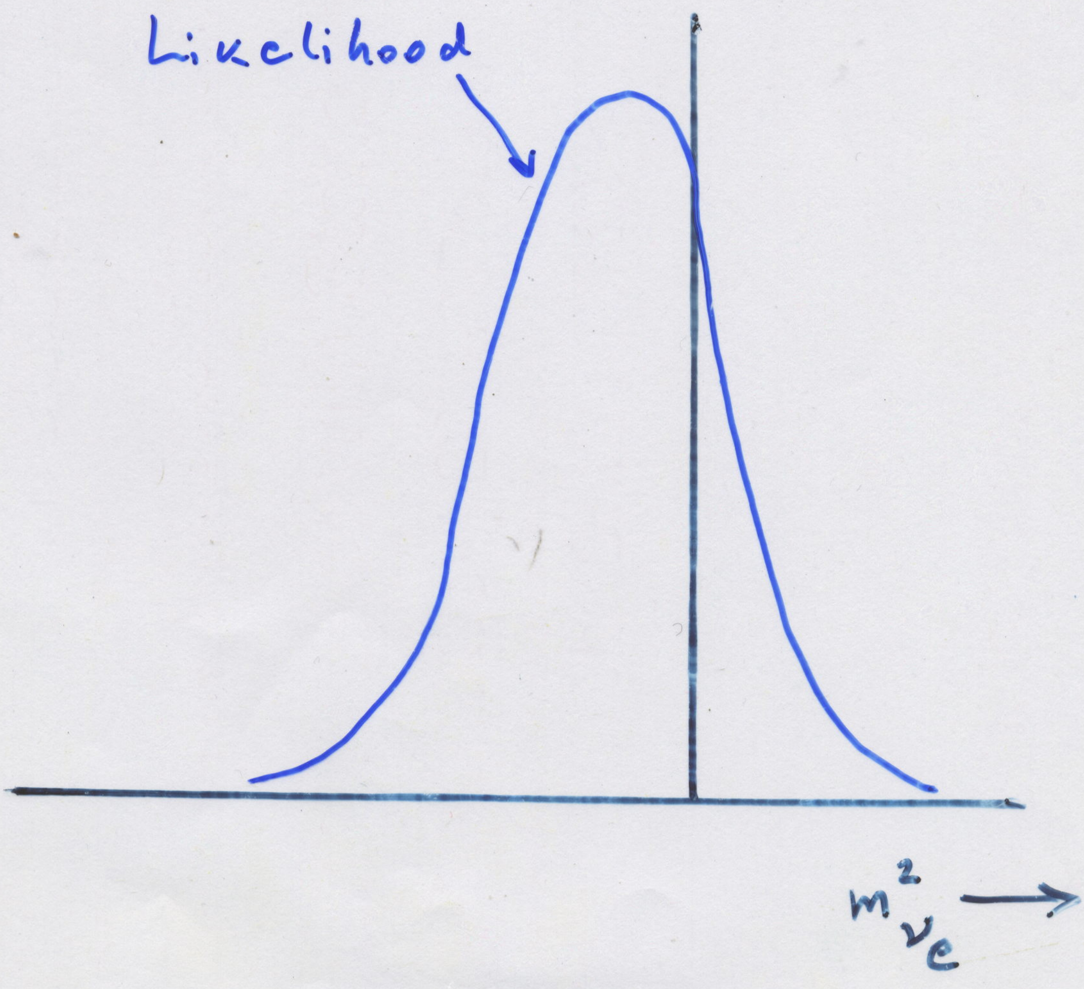
Subjective or Objective prior?



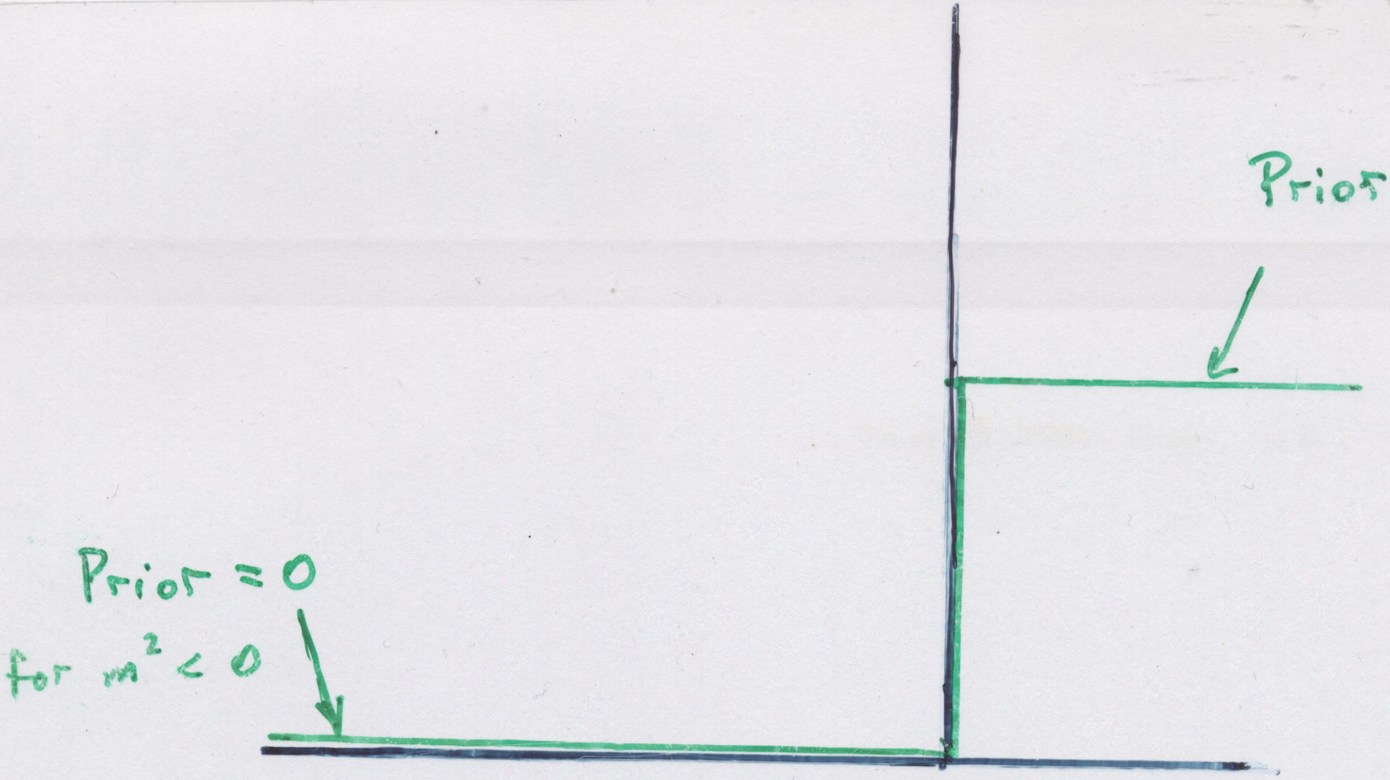
Data overshadows the Prior

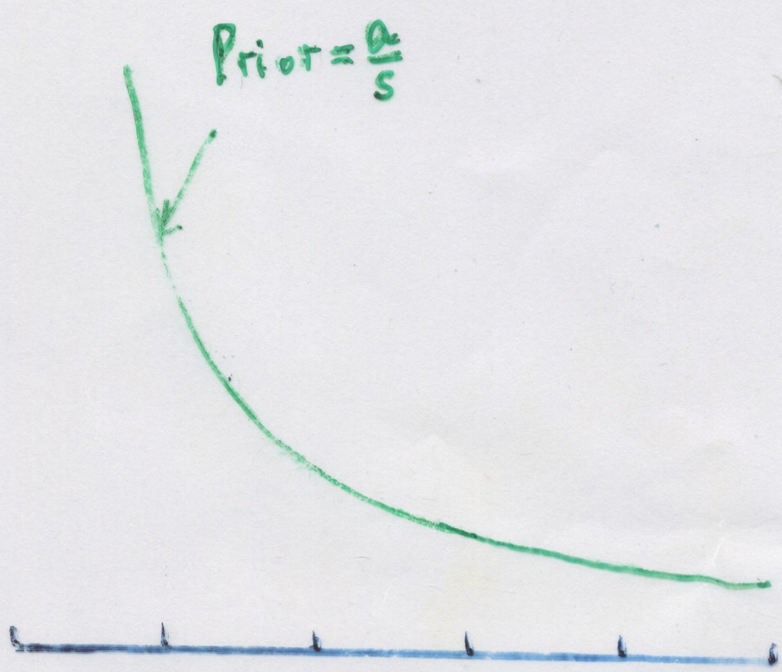


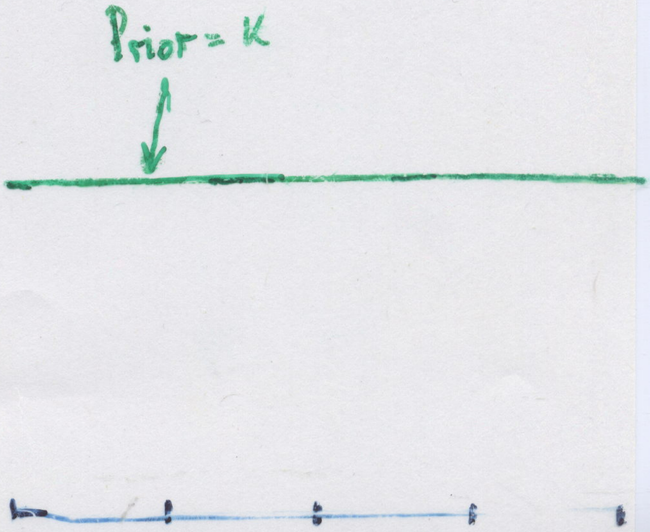
Likelihood
s = signal rate →



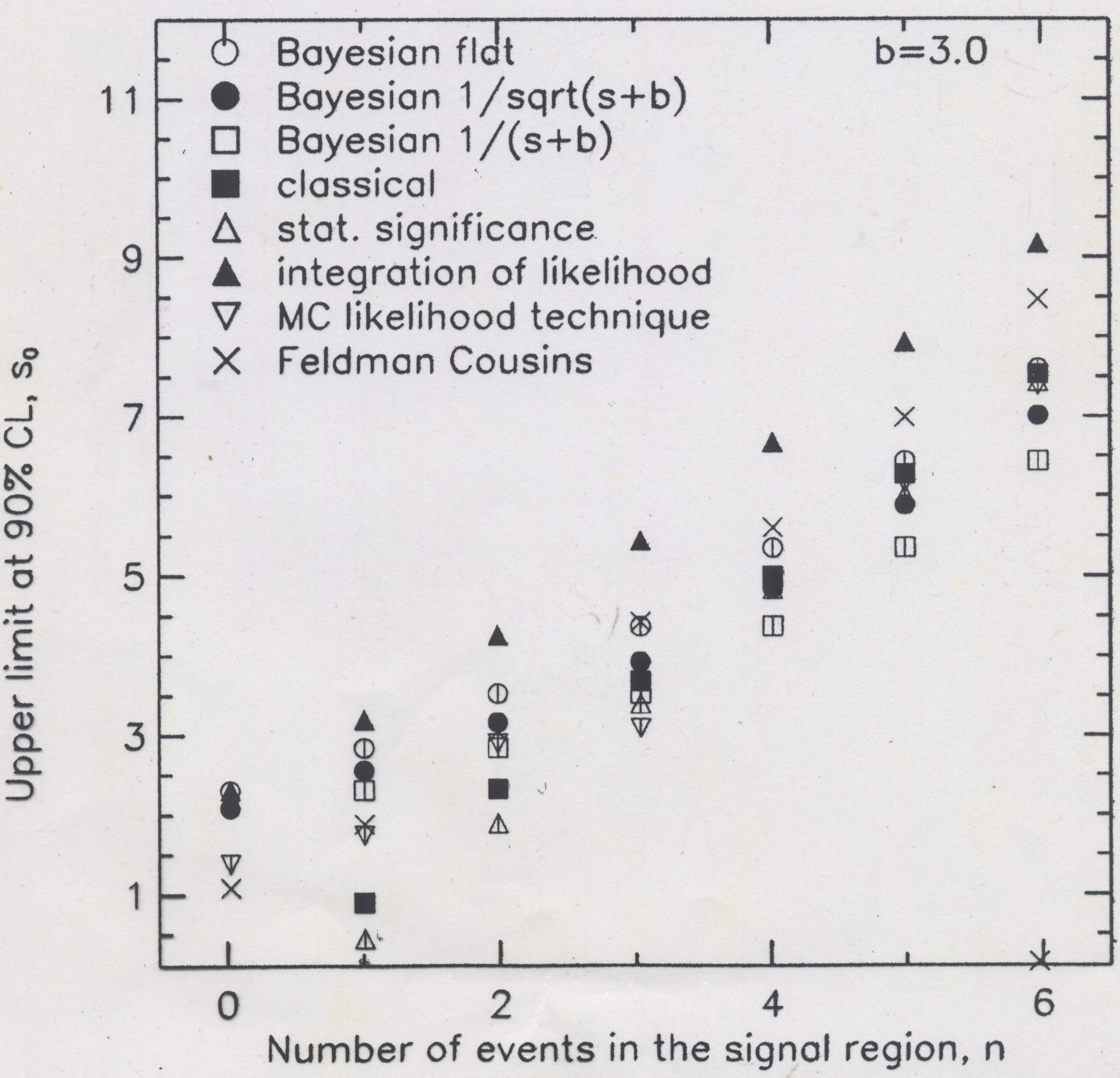
Likelihood
$m^2_{\nu_e}$ →


Prior
Prior = 0
for m² < 0


Prior = a/s


Prior = K

ILYA NARSKY
FNAL CL Workshop



P (Data;Theory) ≠ P (Theory;Data)

HIGGS SEARCH at CERN

Is data consistent with Standard Model?

or with Standard Model + Higgs?

End of Sept 2000 Data not very consistent with S.M.

Prob (Data ; S.M.) < 1% valid frequentist statement

Turned by the press into: Prob (S.M. ; Data) < 1%

and therefore Prob (Higgs ; Data) > 99%

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

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

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 ~ $\delta(1/2)$

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

(also needs cost function)

Example 2 : Particle Identification

Try to separate π 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

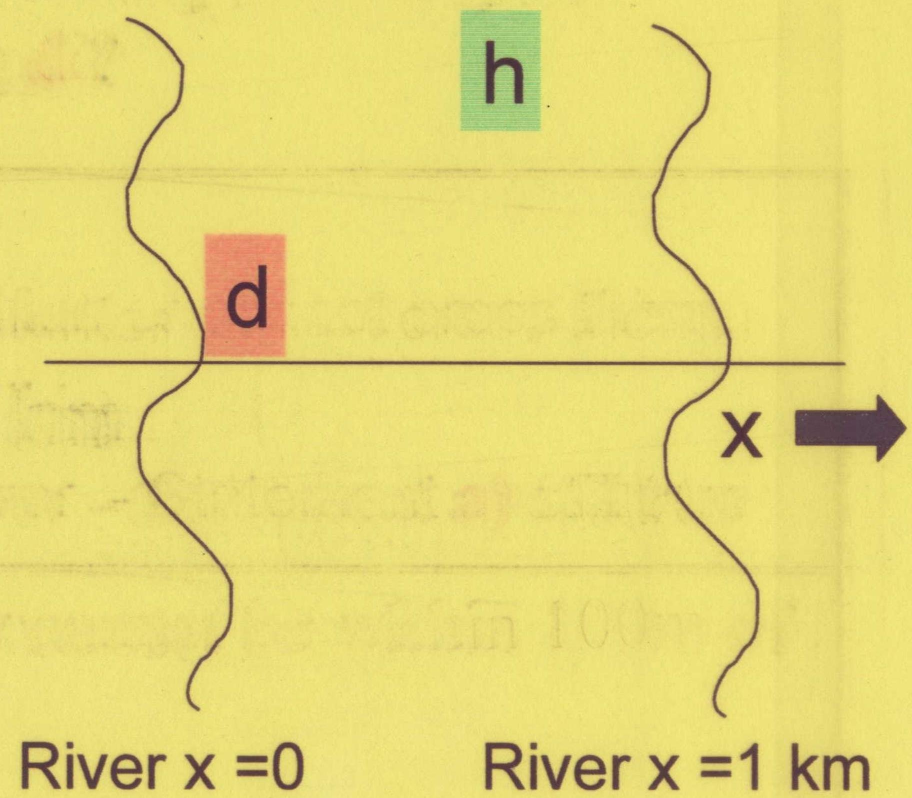
If proton beam, very likely

If general secondary particles, more even

If pure π beam, ~ 0

Hunter and Dog

1) Dog d has 50% probability of being 100 m. of Hunter h
2) Hunter h has 50% probability of being within 100m of Dog d



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

Is it true that b) Hunter h has 50% probability of being within 100m of Dog d ?

Additional information • Rivers at zero & 1 km. Hunter cannot cross them. $0 \le h \le 1\,\text{km}$ • Dog can swim across river - Statement a) still true
If dog at –101 m, hunter cannot be within 100m of dog Statement b) untrue

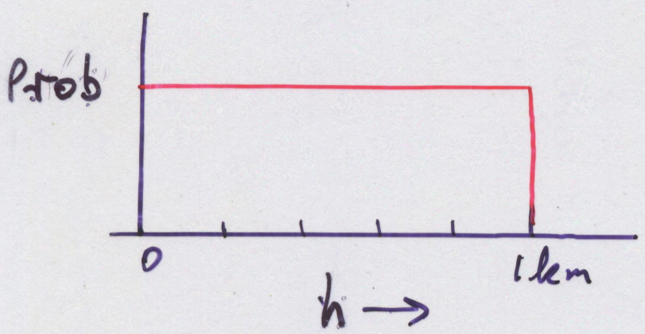
Example:

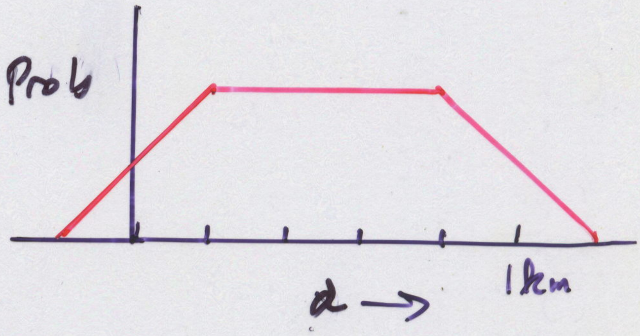
1) More specific on statement (1):

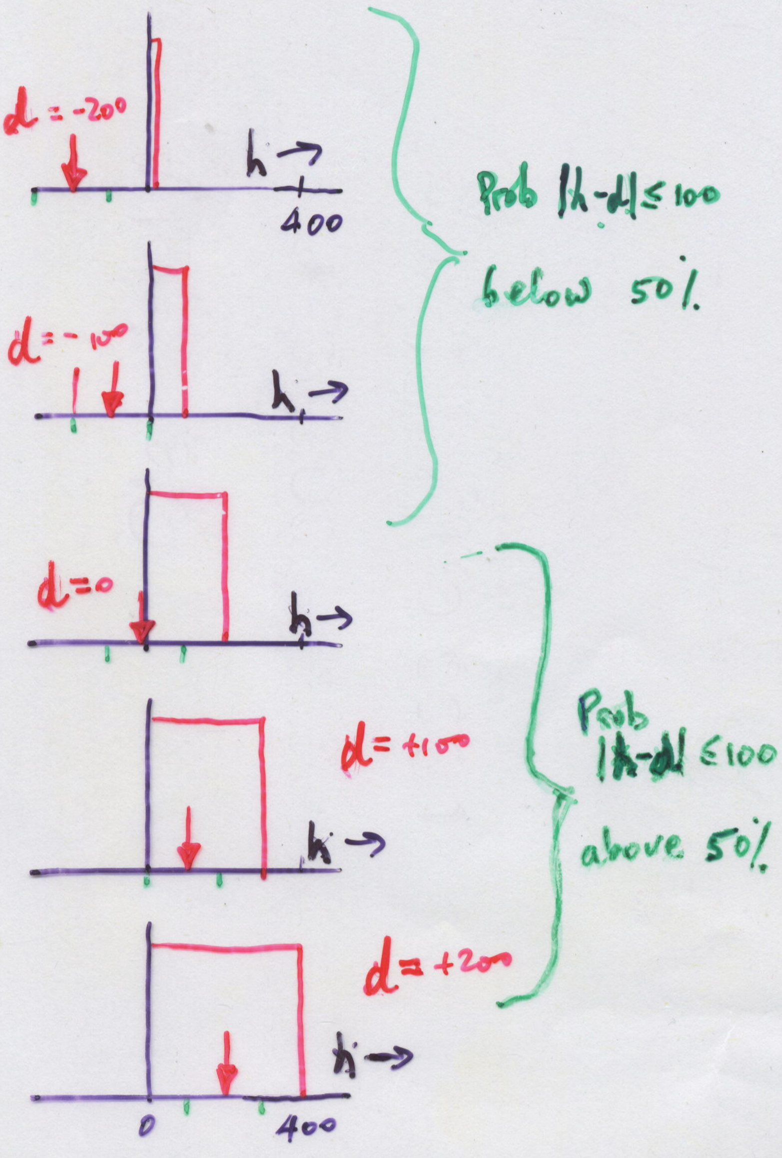
$$\text{Prob}(d - h) = \begin{cases} \text{const.} & \text{for } |d-h| < 200\,\text{m} \\ 0 & \text{for } |d-h| > 200\,\text{m} \end{cases}$$

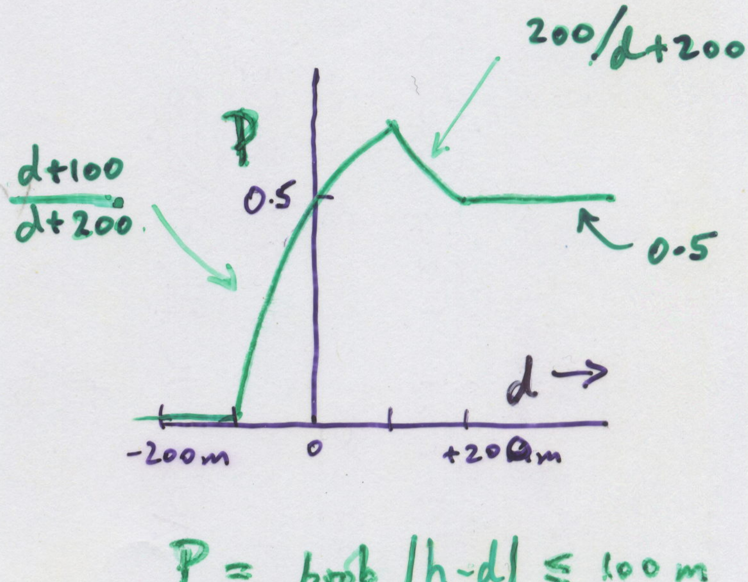
⇑ [L'HOOD]

2) Hunter h uniform in $0 \rightarrow 1\,\text{km}$ [PRIOR]









P = prob $|h-d| \leq 100\,\text{m}$

⇑

Classical Approach

Neyman "confidence interval" avoids pdf for μ

uses only $P(x; \mu)$

Confidence interval $\mu_1 \rightarrow \mu_2$:

$P(\mu_1 \rightarrow \mu_2$ contains $\mu) = \alpha$ True for any μ

Varying intervals from ensemble of experiments

fixed

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

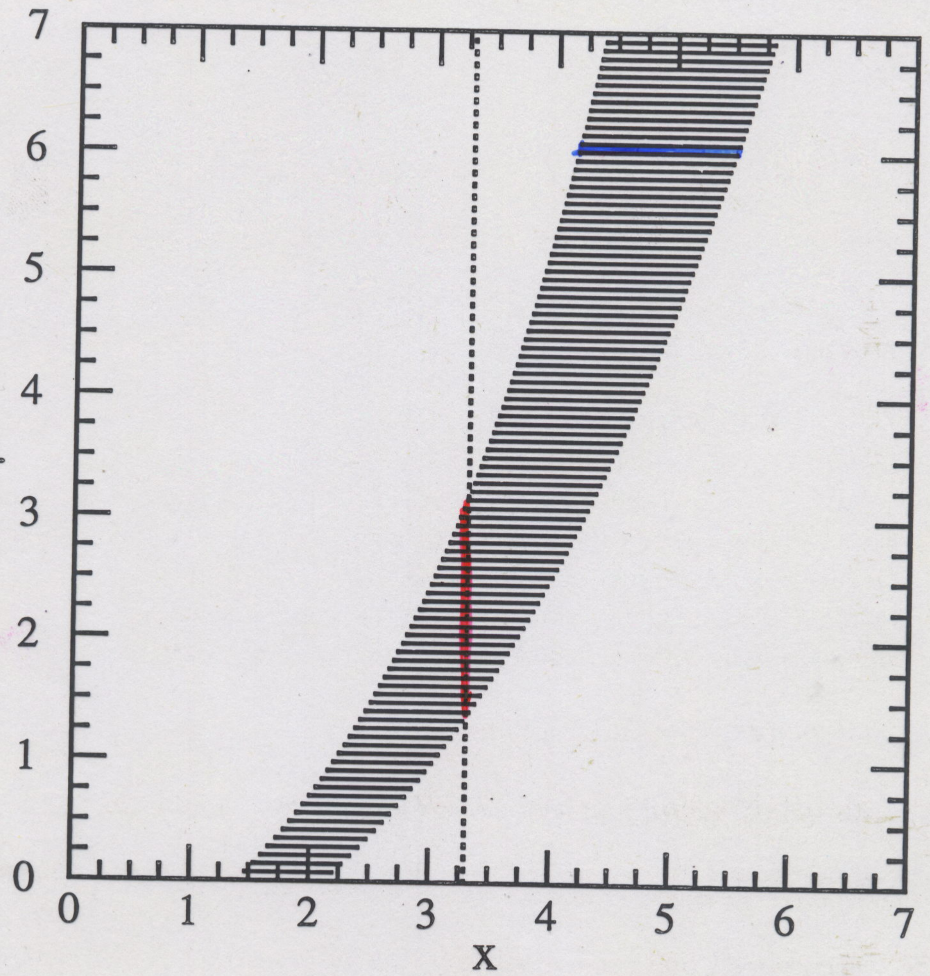
Contrast Bayes : Degree of belief = α that μ_t is in $\mu_1 \rightarrow \mu_2$

CLASSICAL (NEYMAN) CONFIDENCE INTERVALS

Uses only P(data | theory)

FIGURES

Theoretical parameter ↑



Observation ⟶

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.

NO PRIOR INVOLVED

COVERAGE

If true for all μ : "correct coverage"

$P< \alpha$ for some μ: "undercoverage"
(this is serious !)

$P> \alpha$ for some μ: "overcoverage"

Conservative

Loss of rejection power

$\mu_l \leq \mu \leq \mu_u$ 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

Classical Intervals

- Problems

Hard to understand e.g. d'Agostini e-mail
Arbitrary choice of interval
Possibility of empty range
Over-coverage for integer observation
e.g. # of events
Nuisance parameters (systematic errors)

- Advantages

Widely applicable
Well defined coverage

90% classical interval for Gaussian

$\sigma = 1$

$\mu \geq 0$

e.g. $m^2(\nu_e)$

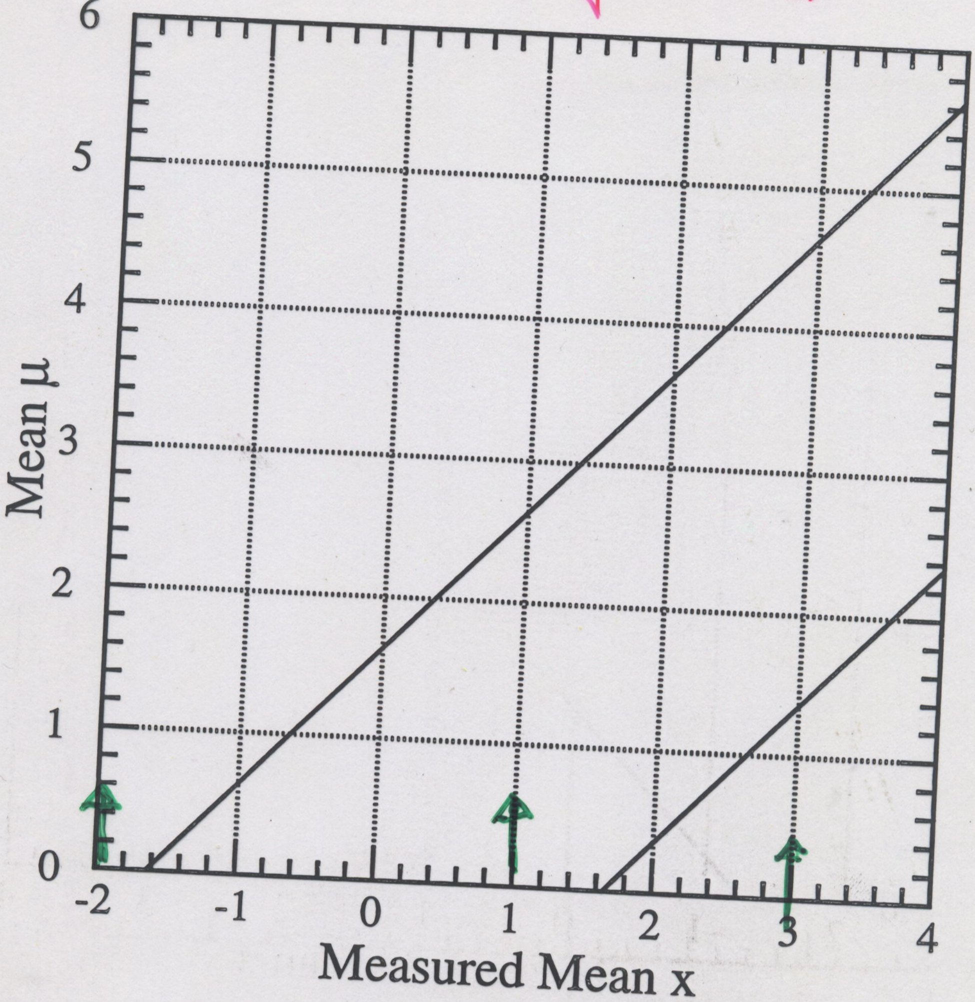


FIG. 3. Standard confidence belt for 90% C.L. central confidence intervals for the mean of a Gaussian, in units of the rms deviation.

$x_{obs} = 3$ Two sided limit

$x_{obs} = 1$ Upper limit

$x_{obs} = -2$ No region for μ

BF 10 17

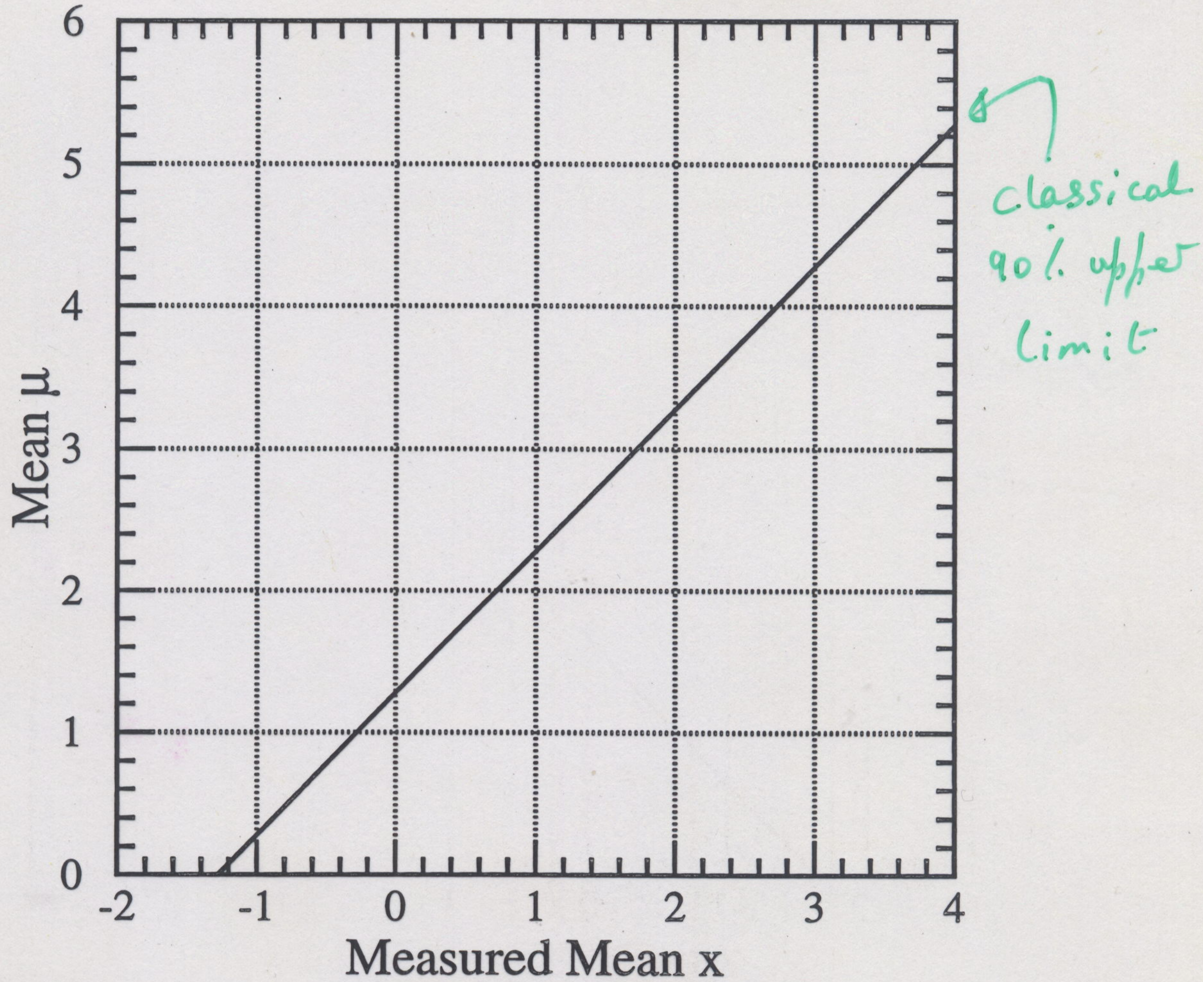


FIG. 2. Standard confidence belt for 90% C.L. upper limits for the mean of a Gaussian, in units of the rms deviation. The second line in the belt is at $x = +\infty$.

Feldman-Cousins 90% conf interval

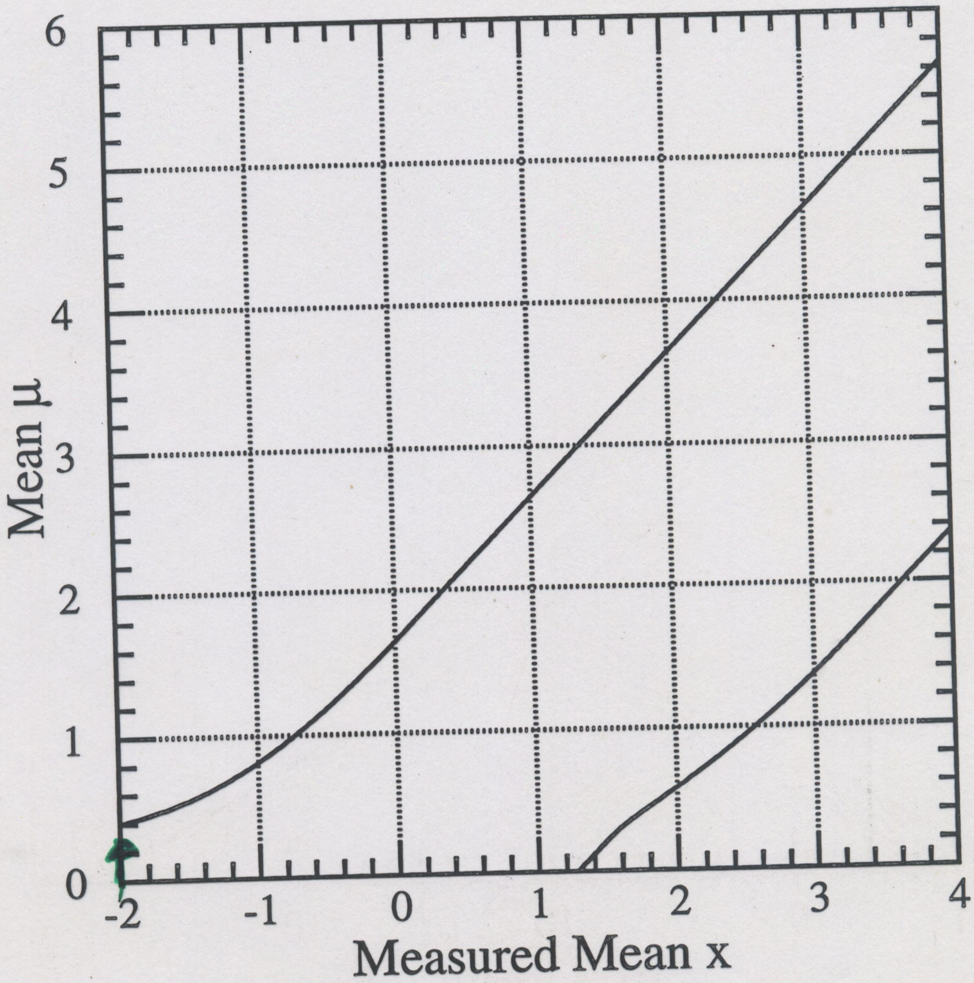


FIG. 10. Plot of our 90% confidence intervals for mean of a Gaussian, constrained to be non-negative, described in the text.

$x_{obs} = -2$. Now gives upper limit

18

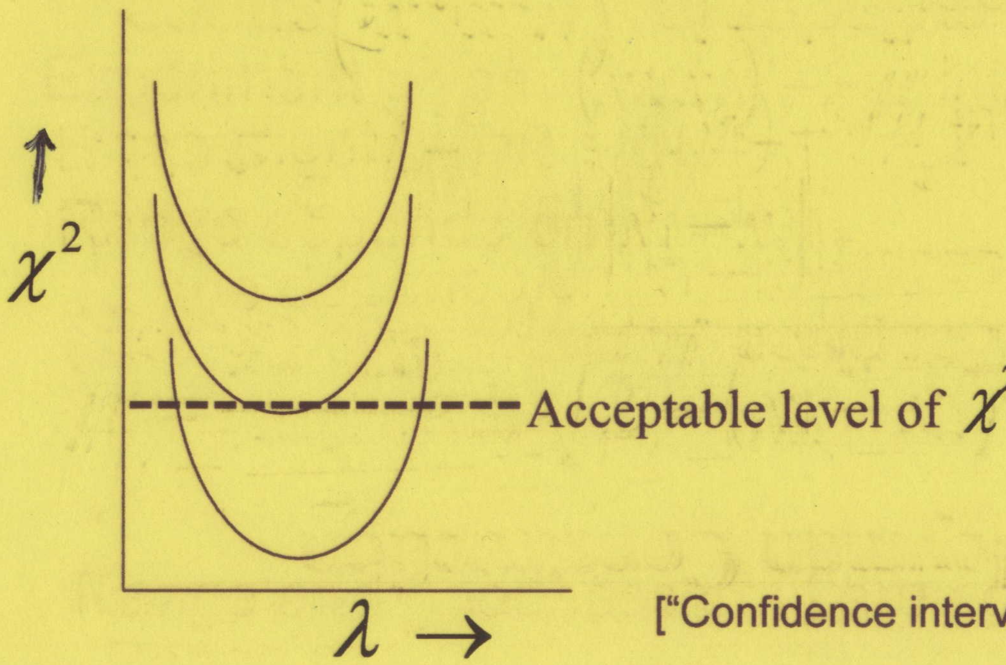
Importance of Ordering Rule

Neyman construction in 1 parameter μ

2 measurements $\mathbf{X}_1$ $\mathbf{X}_2$

$$p(x;\mu) = G(x-\mu, 1)$$

An aside: Determination of single parameter λ via χ^2



Range of parameters given by

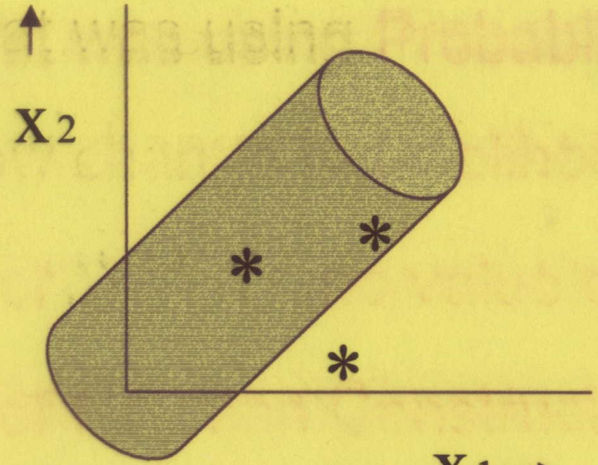
1) Values of λ for which data is likely i.e. p(χ^2) is acceptable or

2) $\chi^2(\lambda) < \chi^2_{min}(\lambda) + 1$

2) is good

1) Range depends on χ^2_{min}

["Confidence interval coupled to goodness of fit"]



Neyman Construction

For given μ, acceptable (x_1, x_2) satisfy

$$\chi^2 = (x_1 - \mu)^2 + (x_2 - \mu)^2 \leq Ccut$$

Defines cylinder in (μ, x_1, x_2) space

Experiment gives $(x_1, x_2) \rightarrow \mu$ interval

Range depends on $|x_1 - x_2|$

$$\mu = \frac{x_1 + x_2}{2} \pm \sqrt{2 - (x_1 - x_2)^2} / 2$$

Range and goodness of fit are coupled

That was using Probability Ordering

Now change to Likelihood Ratio Ordering

For $x_1 \neq x_2$,no value of μ gives very good fit

For Neyman Construction at fixed μ , compare:

$(x_1 - \mu)^2 + (x_2 - \mu)^2$ with $(x_1 - \mu_{best})^2 + (x_2 - \mu_{best})^2$

where $\mu_{best} = (x_1 + x_2)/2$

giving $2\left[\mu^2 - \mu(x_1 + x_2) + \frac{1}{4}(x_1 + x_2)^2\right] = 2\left[\mu - \frac{1}{2}(x_1 + x_2)\right]^2$

Cutting on Likelihood Ratio Ordering gives:

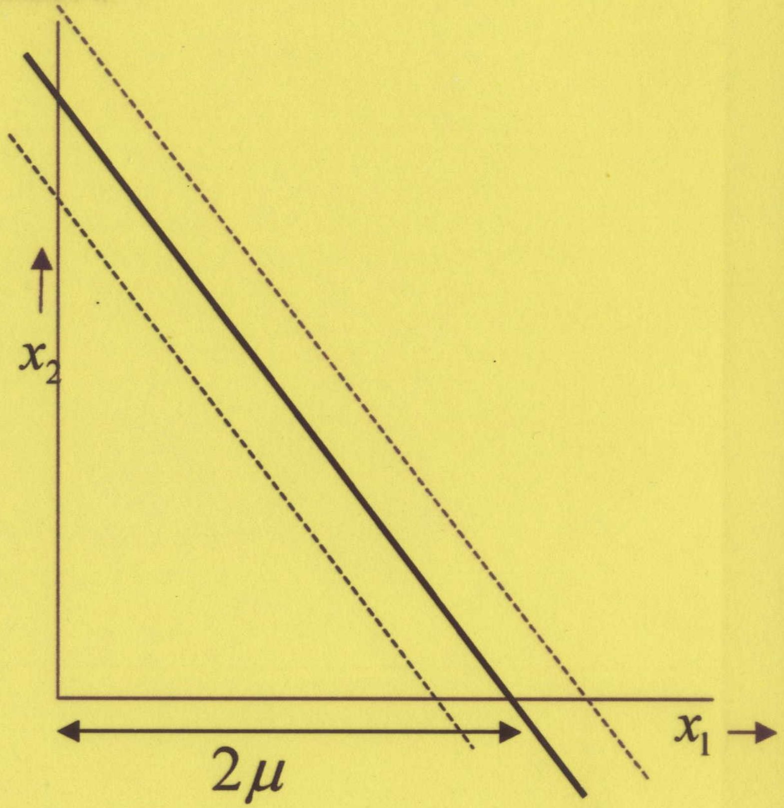
$$\mu = \frac{x_1 + x_2}{2} \pm \sqrt{\frac{C}{2}}$$

$$\mu = \frac{x_1 + x_2}{2} \pm \sqrt{\frac{C}{2}}$$

Therefore, range of μ is

Constant Width

Independent of $x_1 - x_2$



Confidence Range and Goodness of Fit are completely decoupled

Bayesian

Pros:

Easy to understand

Physical Interval

Cons:

Needs prior

Hard to combine

Coverage

Standard Frequentist

Pros:

Coverage

Cons:

Hard to understand

Small or Empty Intervals

Different Upper Limits

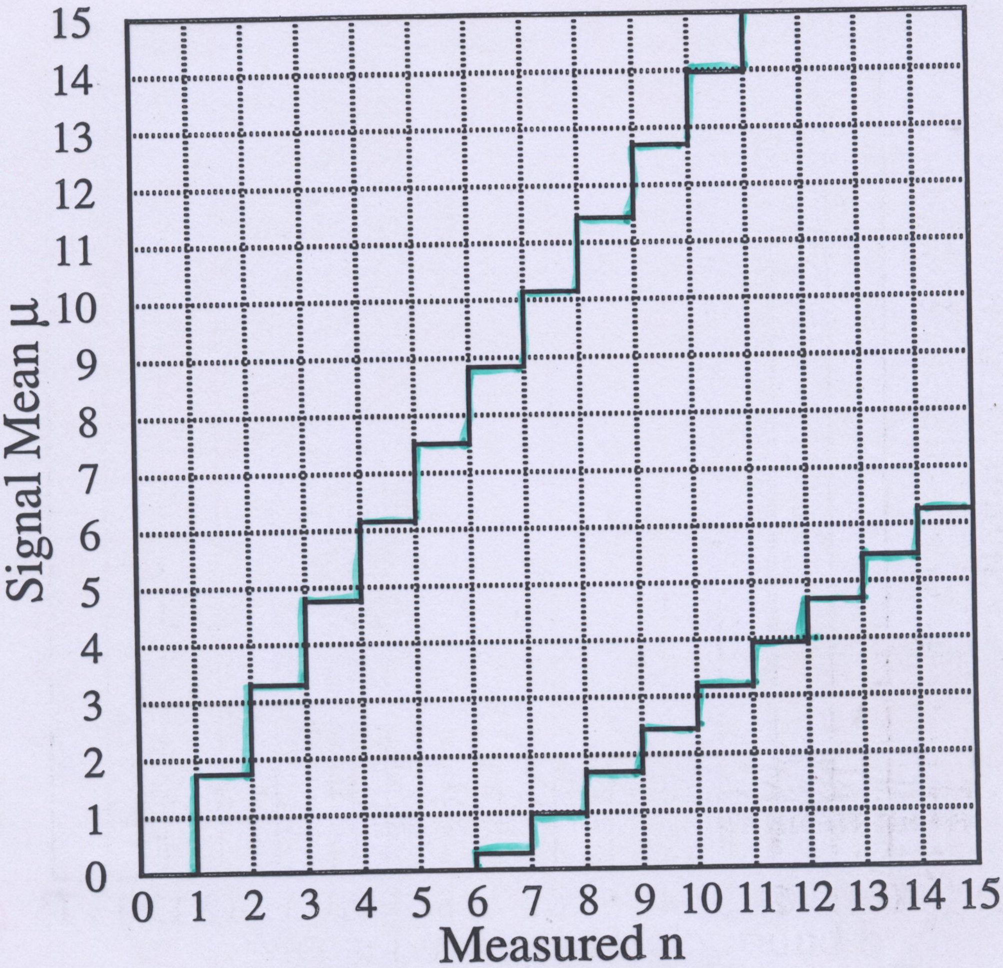
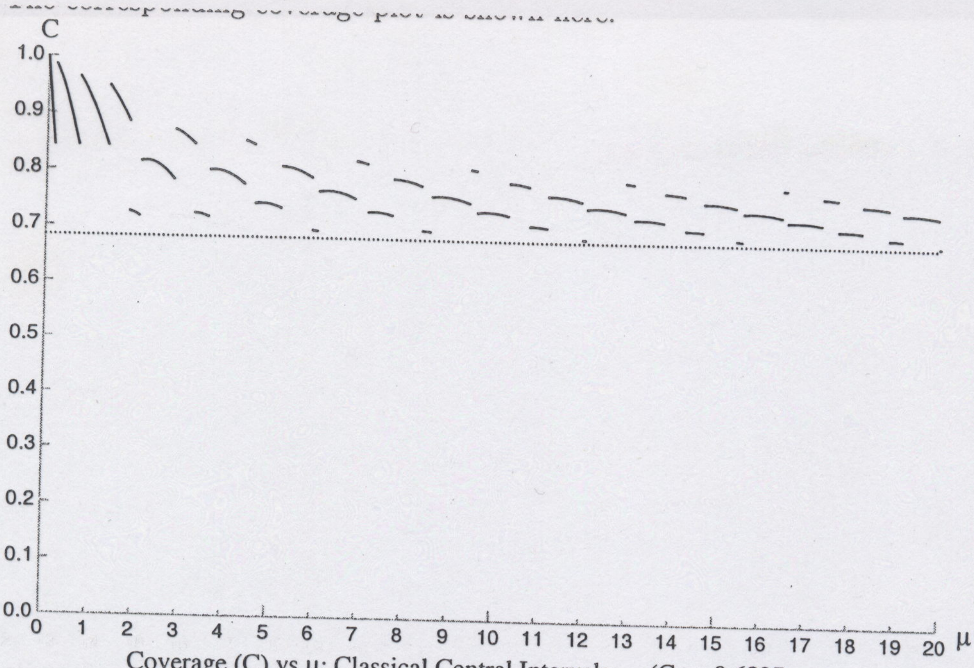


FIG. 6. Standard confidence belt for 90% C.L. central confidence intervals, for unknown Poisson signal mean μ in the presence of Poisson background with known mean $b = 3.0$.

Standard Frequentist
for Poisson mean μ

COVERAGE OF ERROR BARS FOR POISSON DATA

JOEL HEINRICH
CDF 6438



Coverage (C) vs μ: Classical Central Intervals (C → 0.6827 as μ → ∞)

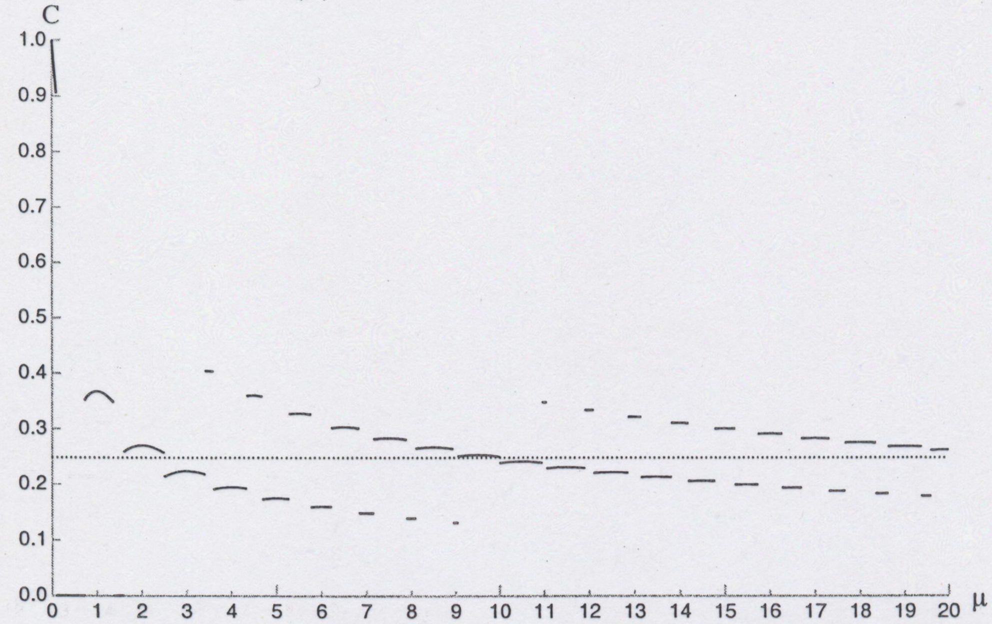
$$P(n, \mu) = e^{-\mu} \mu^n / n!$$

Classical central intervals at 68.3% coverage

COVERAGE OF ERROR BARS FOR POISSON DATA — JOEL HEINRICH

CDF 6438

plot of the coverage $C(\mu)$ for the case $\Delta = 0.1$.



Coverage (C) vs μ: $\chi^2 < 0.1$ $\quad$ ($C \to 0.2482$ as $\mu \to \infty$)

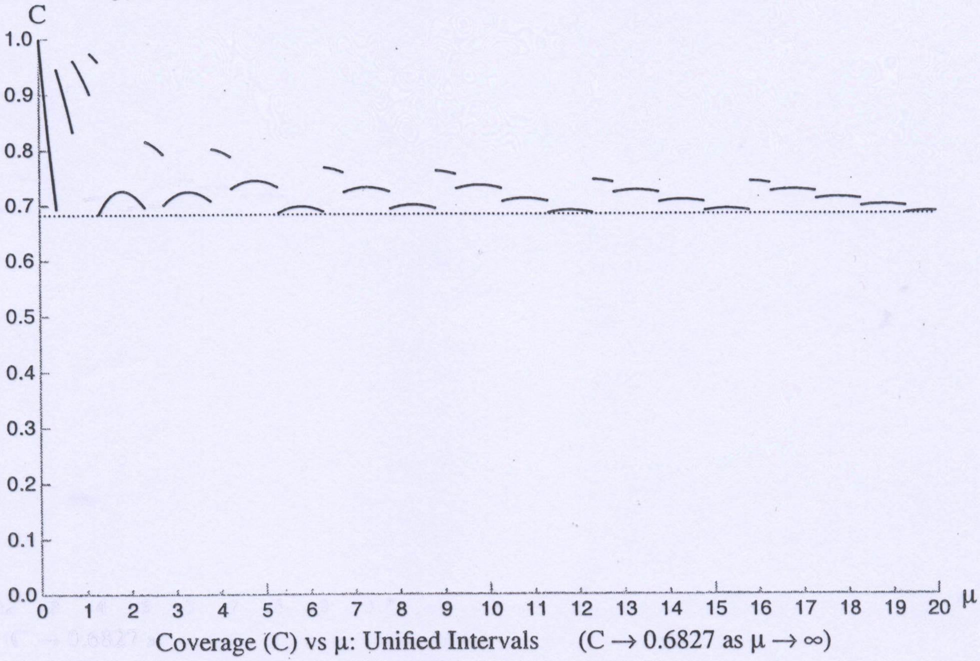
$$P(n, \mu) = e^{-\mu} \mu^n / n!$$

$$\chi^2 = \left(\frac{n - \mu}{\sqrt{\mu}} \right)^2$$

$\Delta \chi^2 = 0.1 \implies 24.8\%$ coverage ✓

NOT FREQUENTIST

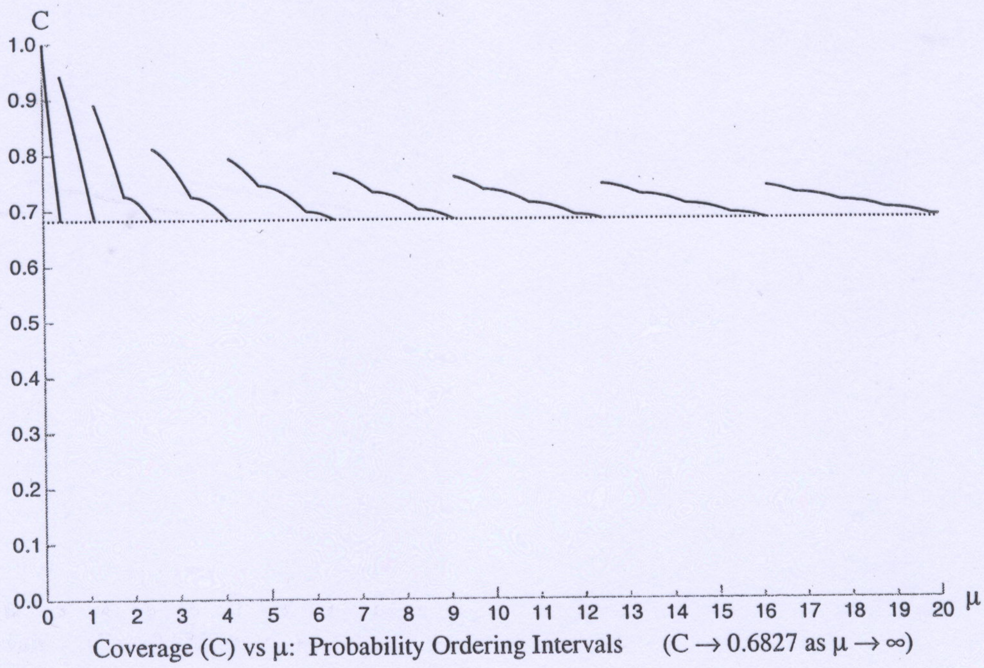
Finally, we show the coverage of the 1σ unified intervals:



Coverage (C) vs μ: Unified Intervals (C → 0.6827 as μ → ∞)

$P(n, \mu) = e^{-\mu} \mu^{n} / n!$

Unified intervals at 68.3% coverage

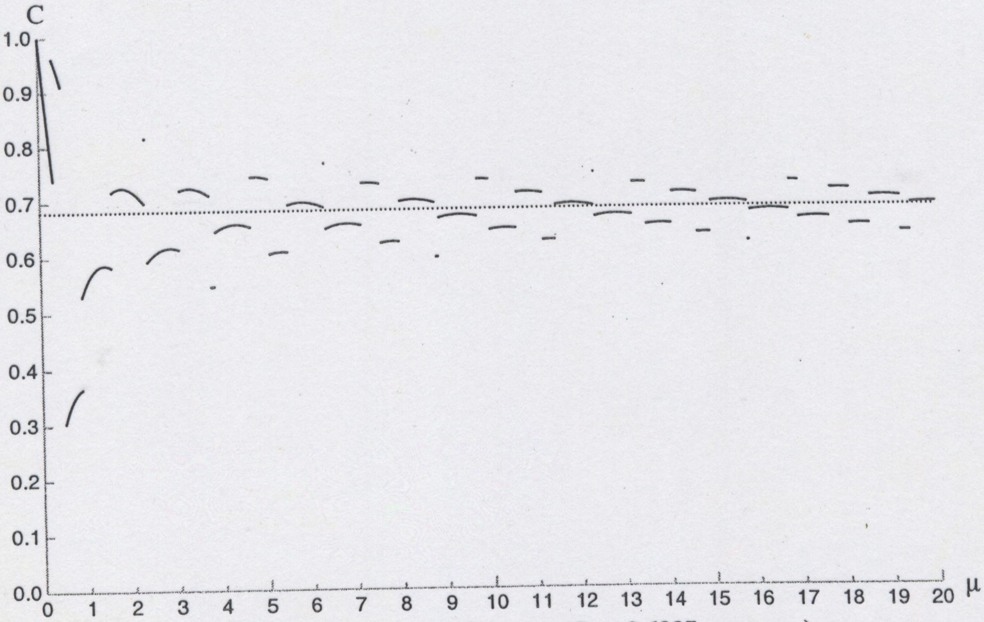


Coverage (C) vs μ: Probability Ordering Intervals (C → 0.6827 as μ → ∞)

$$P(n, \mu) = e^{-\mu} \mu^n / n!$$

Probability ordering intervals

at 68.3% coverage



Coverage (C) vs μ: $-2\ln\lambda < 1$ (C → 0.6827 as μ → ∞)

$$P(n, \mu) = e^{-\mu} \mu^{n} / n!$$

$$-2 \ln \lambda < 1$$

$$\left[\lambda = P(n, \mu) / P(n, \mu_{best}) \right]$$

SYSTEMATICS

For example

$$N_{events} = \sigma LA + b$$

↑ Observed ↑ $N \pm \sqrt{N}$ for statistical errors

↑ Physics parameter

↑ ↑ we need to know these, probably from other measurements (and/or theory)

Uncertainties → error in σ

Some are arguably statistical errors

$$LA = LA_o \pm \sigma_{LA}$$

$$b = b_o \pm \sigma_b$$

Shift Central Value

Bayesian

Frequentist

Mixed

$$N_{events} = \sigma \, LA + b$$

Simplest Method

Evaluate σ_0 using LA_0 and b_0

Move nuisance parameters (one at a time) by their errors → $\delta\sigma_{LA}$ & $\delta\sigma_{b}$

If nuisance parameters are uncorrelated

Combine these contributions in quadrature

→ total systematic

Bayesian

Without systematics

$$p(\sigma;N) \propto p(N;\sigma)\Pi(\sigma)$$

↑ prior

With systematics

$$p(\sigma, LA, b; N) \propto p(N;\sigma, LA, b)\Pi(\sigma, LA, b)$$

↑

$$\sim \Pi_1(\sigma)\Pi_2(LA)\Pi_3(b)$$

Then integrate over LA and b

$$p(\sigma;N) = \iint p(\sigma, LA, b; N)\, dLA\, db$$

$$p(\sigma ; N) = \iint p(\sigma , LA , b; N)\, dLA \; db$$

If $\Pi_1(\sigma) =$ constant and $\Pi_2(\mathrm{LA})$ = truncated Gaussian TROUBLE!

Upper limit on σ from $\int p(\sigma;N)\,\mathrm{d}\sigma$

Significance from likelihood ratio for $\sigma = 0$ and $\sigma_{\max}$

Frequentist

Full Method

Imagine just 2 parameters σ and LA

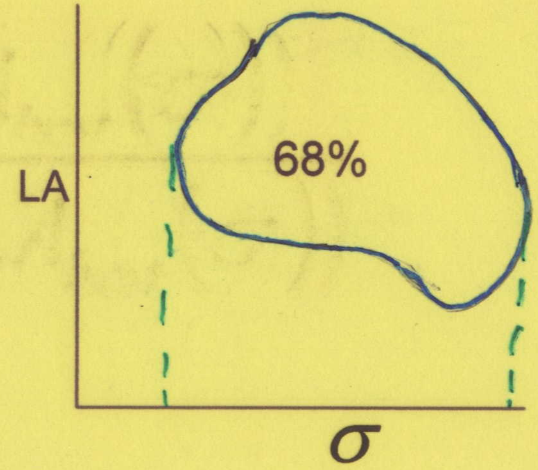
and 2 measurements N and M

↑ Physics ↑ Nuisance

Do Neyman construction in 4-D

Use observed N and M, to give

Confidence Region



Then project onto σ axis

This results in OVERCOVERAGE

Aim to get better shaped region, by suitable choice of ordering rule

Example: Profile likelihood ordering

$$\frac{L(N_0 M_0; \sigma, LA_{best}(\sigma))}{L(N_0 M_0; \sigma_{best}, LA_{best}(\sigma))}$$

Full frequentist method hard to apply in several dimensions

Used in ≤ 3 parameters

For example: Neutrino oscillations (CHOOZ)

$$\sin^2 2\theta \, , \Delta m^2$$

Normalisation of data

Use approximate frequentist methods that reduce dimensions to just physics parameters

e.g. Profile pdf

i.e. $pdf_{profile}\left(N;\sigma\right) = pdf\left(N, M_0;\sigma, LA_{best}\right)$

Contrast Bayes marginalisation

Distinguish "profile ordering"

Properties being studied by Giovanni Punzi

Talks at FNAL CONFIDENCE LIMITS WORKSHOP

(March 2000) by:

Gary Feldman

Wolfgang Rolke hep-ph/0005187 version 2

Acceptance uncertainty worse than Background uncertainty

Limit of C.L. as $\sigma \rightarrow 0$

$\neq$ C.L. for $\sigma = 0$

Need to check Coverage

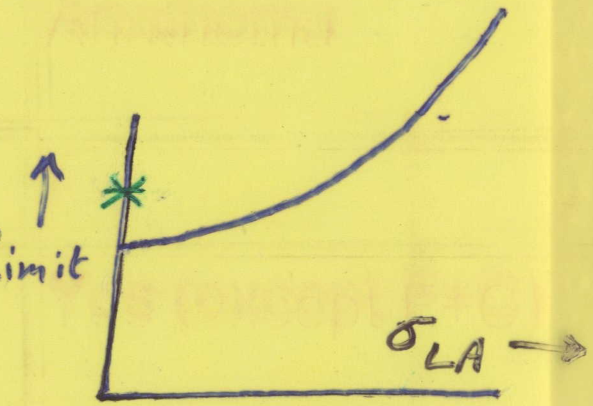
Method: Mixed Frequentist - Bayesian

Bayesian for nuisance parameters and

Frequentist to extract range

Philosophical/aesthetic problems?

Highland and Cousins



(Motivation was paradoxical behavior of Poisson limit when LA not known exactly)

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
Problem of parameters?	Yes	Anathema
Needs prior?	Yes	No
Choice of interval?	Yes	Yes (except F+C)
Data considered	Only data you have	….+ more extreme
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"