BAYES and FREQUENTISM:
The Return of an Old Controversy

Louis Lyons and Lorenzo Moneta

Imperial College & Oxford University

CERN

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.

How can textbooks not even mention
Bayes / Frequentism?

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

with no constraint on μ_{true} , then

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

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 : 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→ 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 = I am in Spanish Pyrenees

B = I am giving a lecture

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

Completely uncontroversial, provided….

Bayesian

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

Bayes' Theorem

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

posterior likelihood prior

Problems: p(param) Has particular value

"Degree of belief"

Credible Intervals

Prior What functional form?

Coverage

Prior: What functional form?

Uninformative prior: Flat?

Cannot be normalised

Ranges 0-1 and 1089-1090 equally probable

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

$dp/dm = dp/d(\ln m) \times d(\ln m)/dm = (1/m) \times dp/d(\ln m)$

Even more problematic with more params

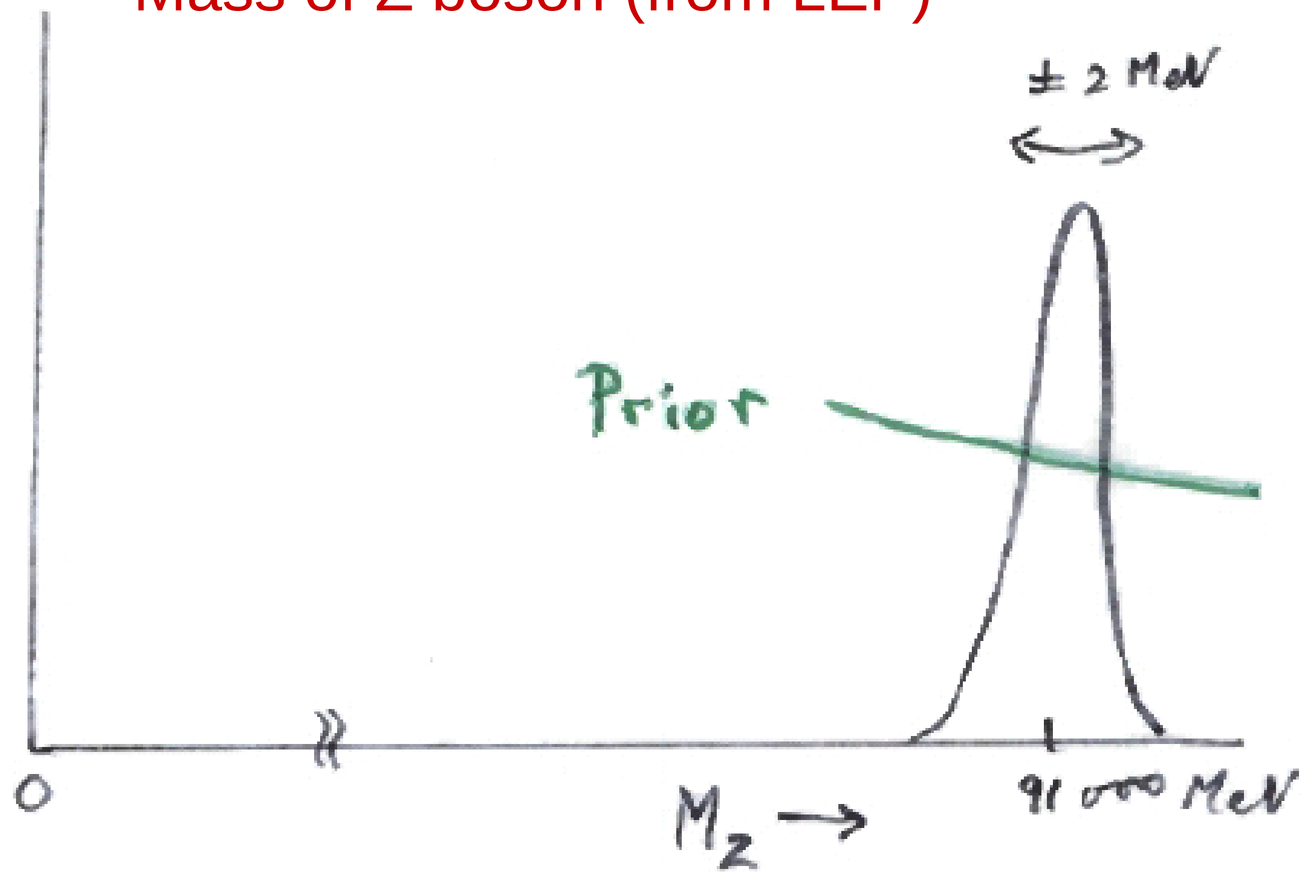
Unimportant if "data overshadows prior"

Important for limits

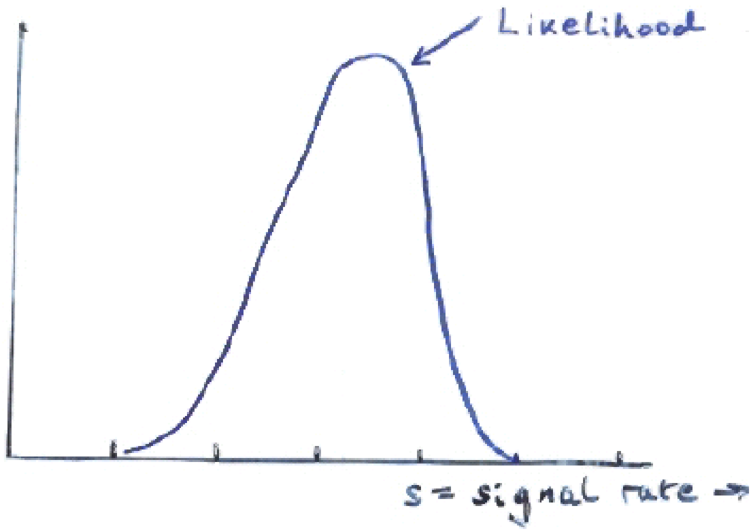
Subjective or Objective prior?

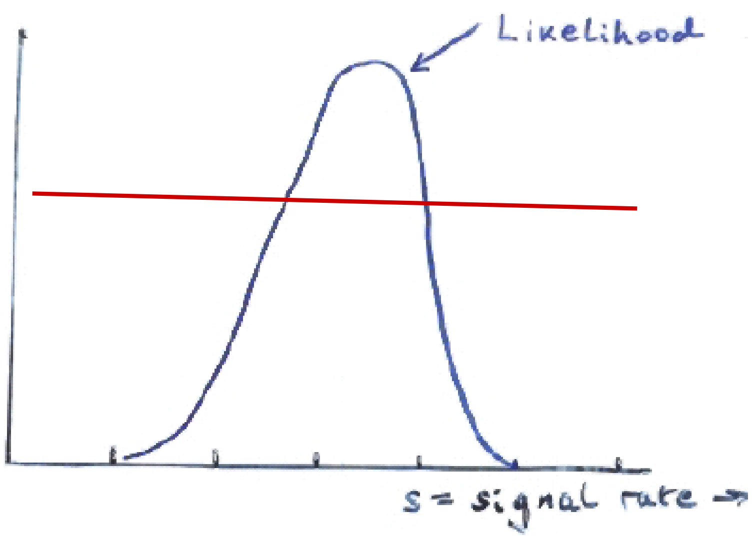
Priors might be OK for parametrising prior knowledge, but not so good for prior ignorance.

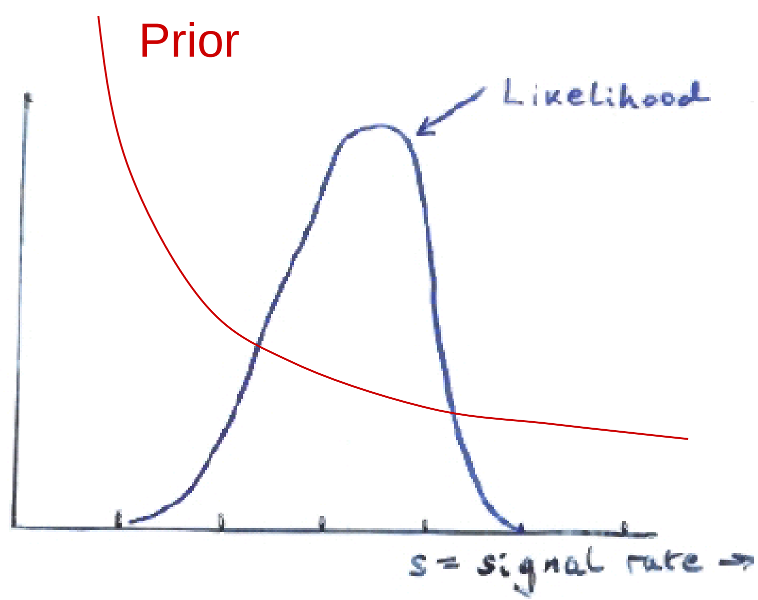
Mass of Z boson (from LEP)



Data overshadows prior

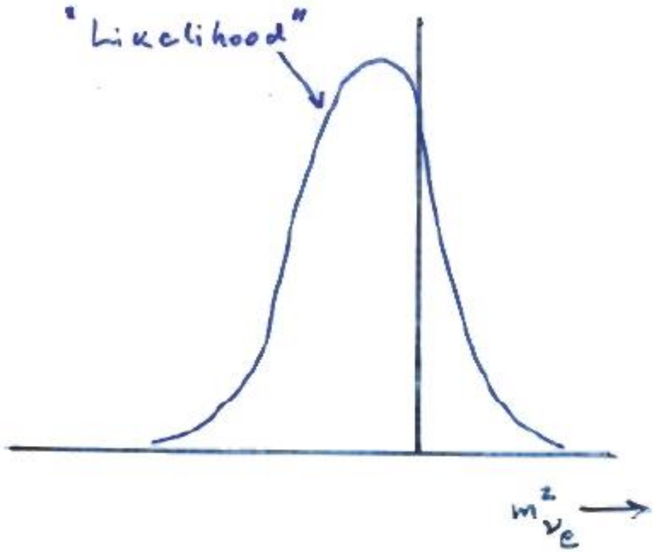


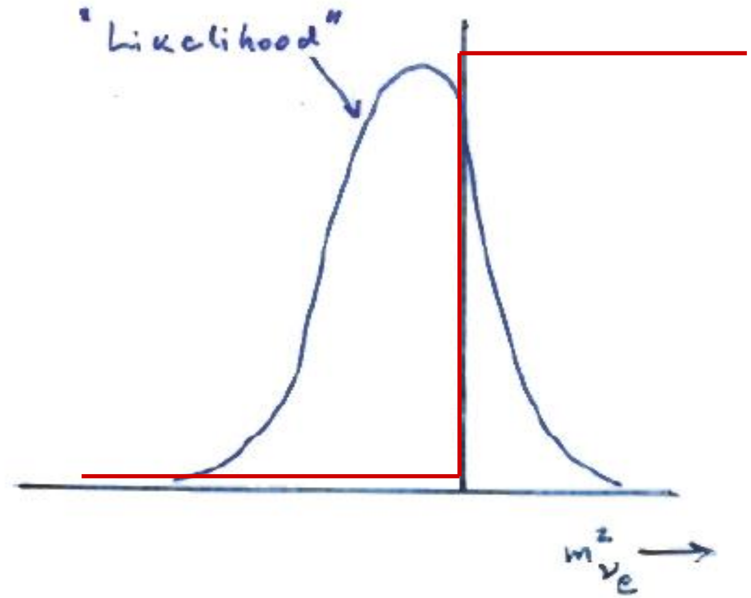




Even more important for UPPER LIMITS

Mass-squared of neutrino





Prior = zero in unphysical region

Fred James: “Is it a reindeer?”

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

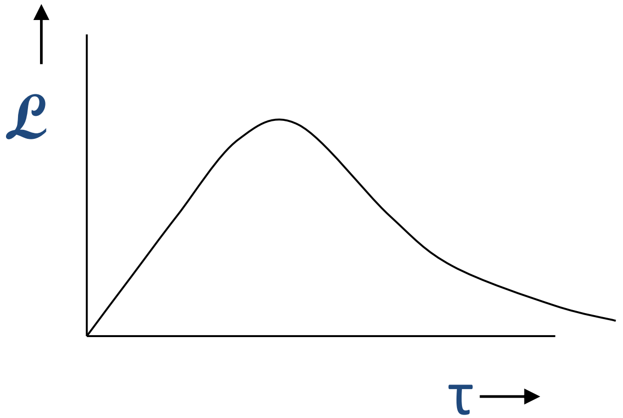
e.g. constant up to some large τ

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

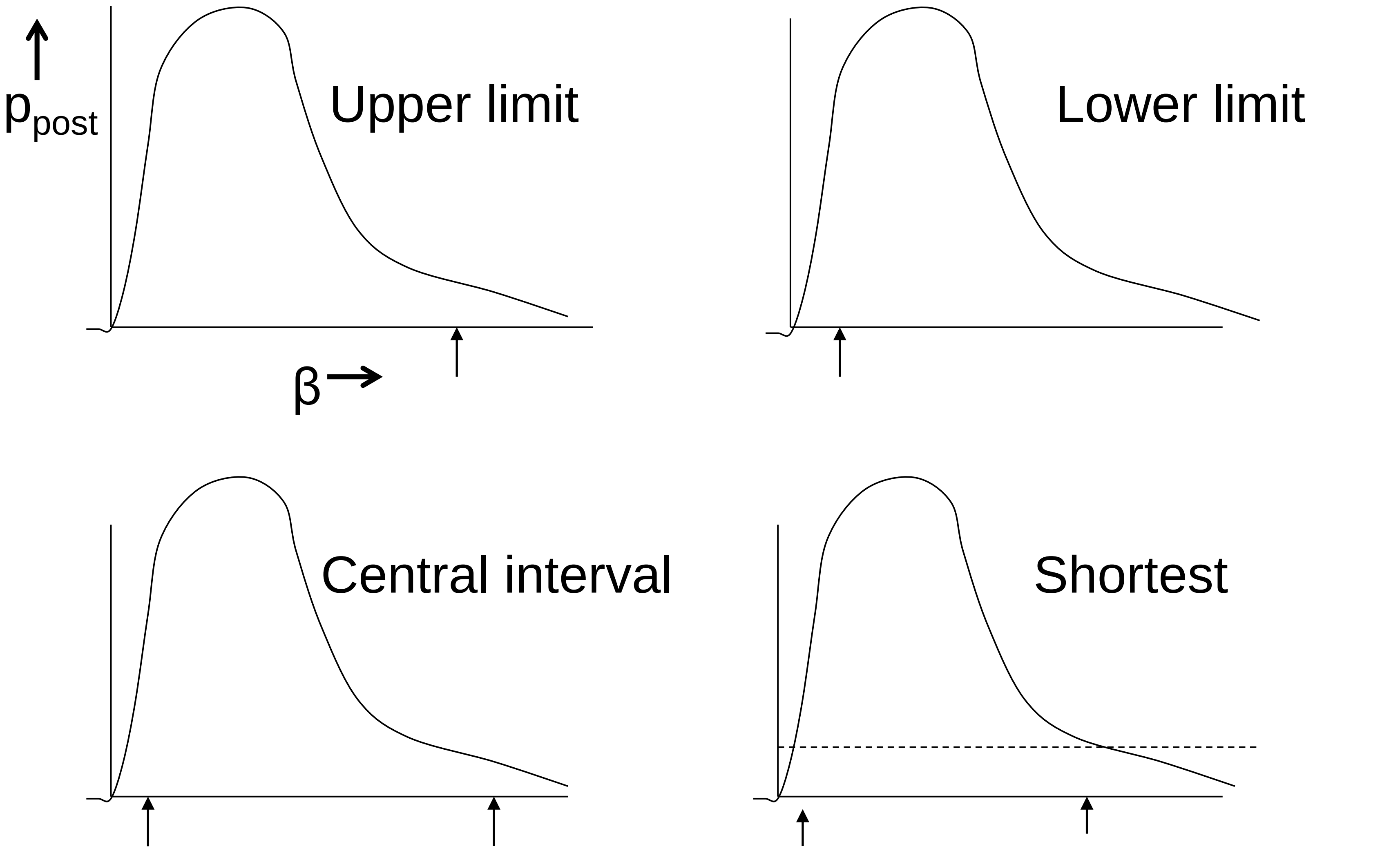


Sensitivity study: Compare with using different prior

e.g. Prior constant in decay rate $\lambda = 1/\tau \rightarrow$ different range

Contrast frequentist method for same situation later.

Bayesian posterior → intervals



UL→ includes 0; LL→ excludes 0; Central→ usually excludes 0; Shortest is metric dependent

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

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 = Murderer or not

Data = Eats bread for breakfast or not

P (eats bread ; murderer) ~ 99%

but

P(murderer; eats bread) ~ 10^{-6}

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

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

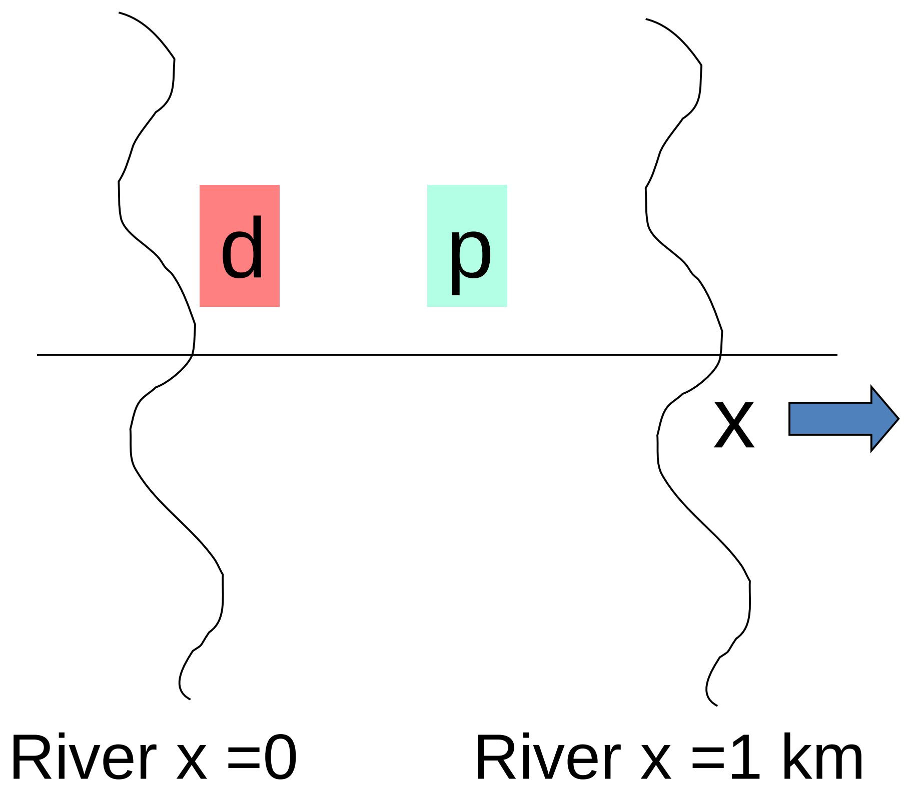
P (pregnant ; female) ~ 3%

but

P (female ; pregnant) >>>3%

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.

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

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

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

Statement b) untrue

Classical Approach

Neyman "confidence interval" avoids pdf for μ

Uses only P(x; μ)

Confidence interval $\mu_1 \to \mu_2$:

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

Varying intervals from ensemble of experiments

fixed

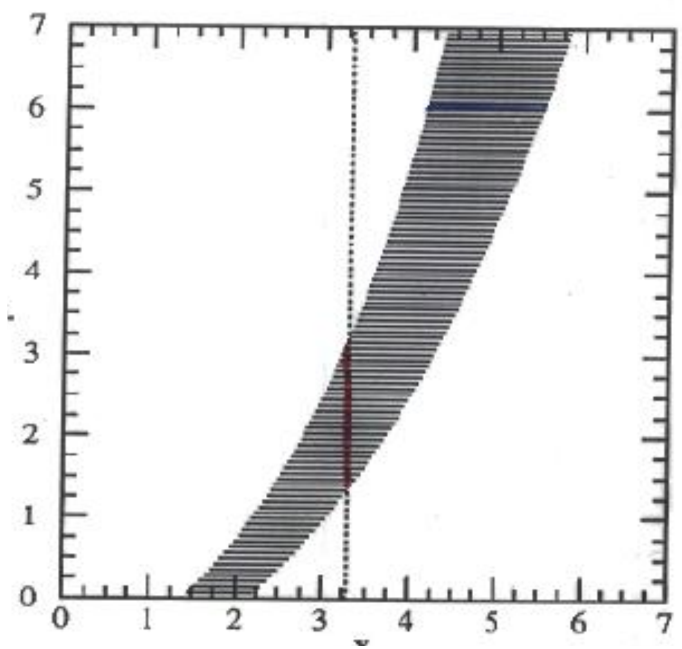
Gives range of μ for which observed value x_0 was "likely" (α)

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

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

90% Classical interval for Gaussian

$\sigma = 1 \qquad \mu \geq 0$

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

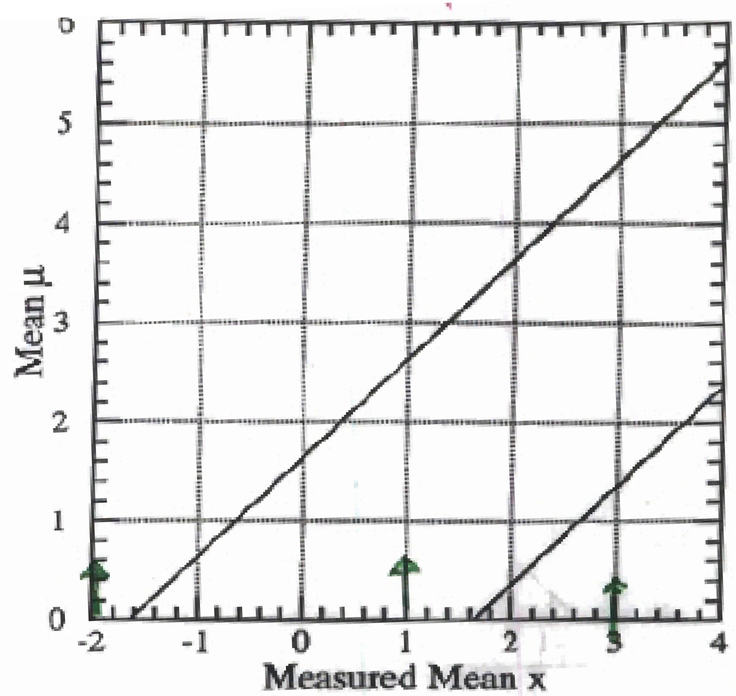


FIG. 3. Standard confidence belt for 90% C.L. central confidence intervals for the mean of a Gaussian, in units of the rms deviation.

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

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

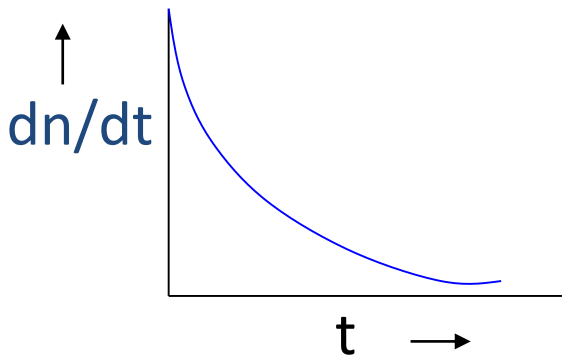
Probability/credible statement about μ

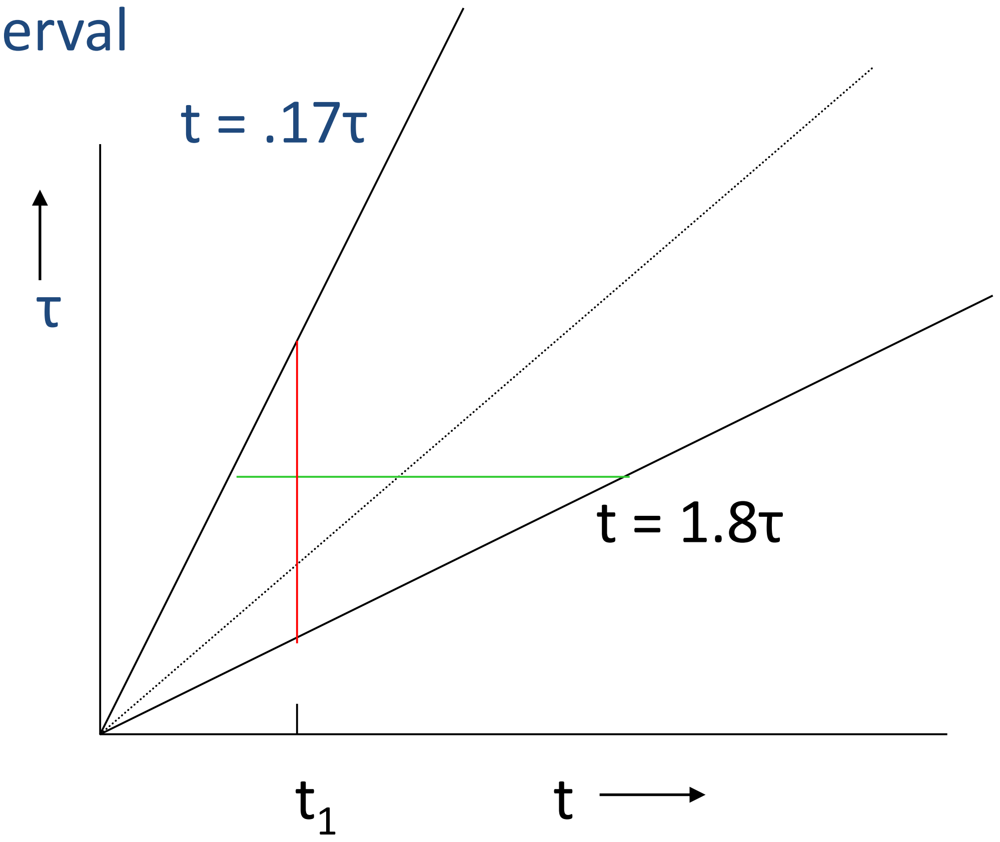
Frequentism: Specific example

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

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

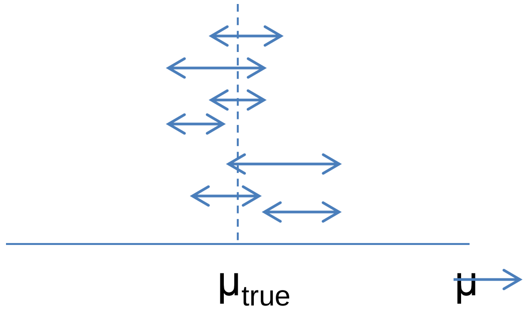
Construct 68% central interval





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

Coverage



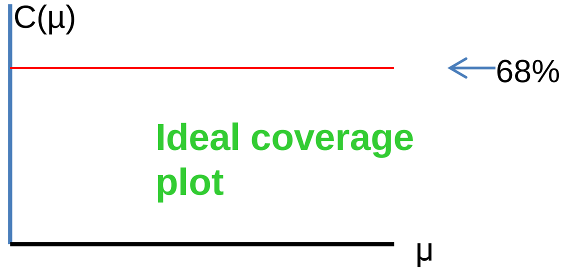
* What it is:

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

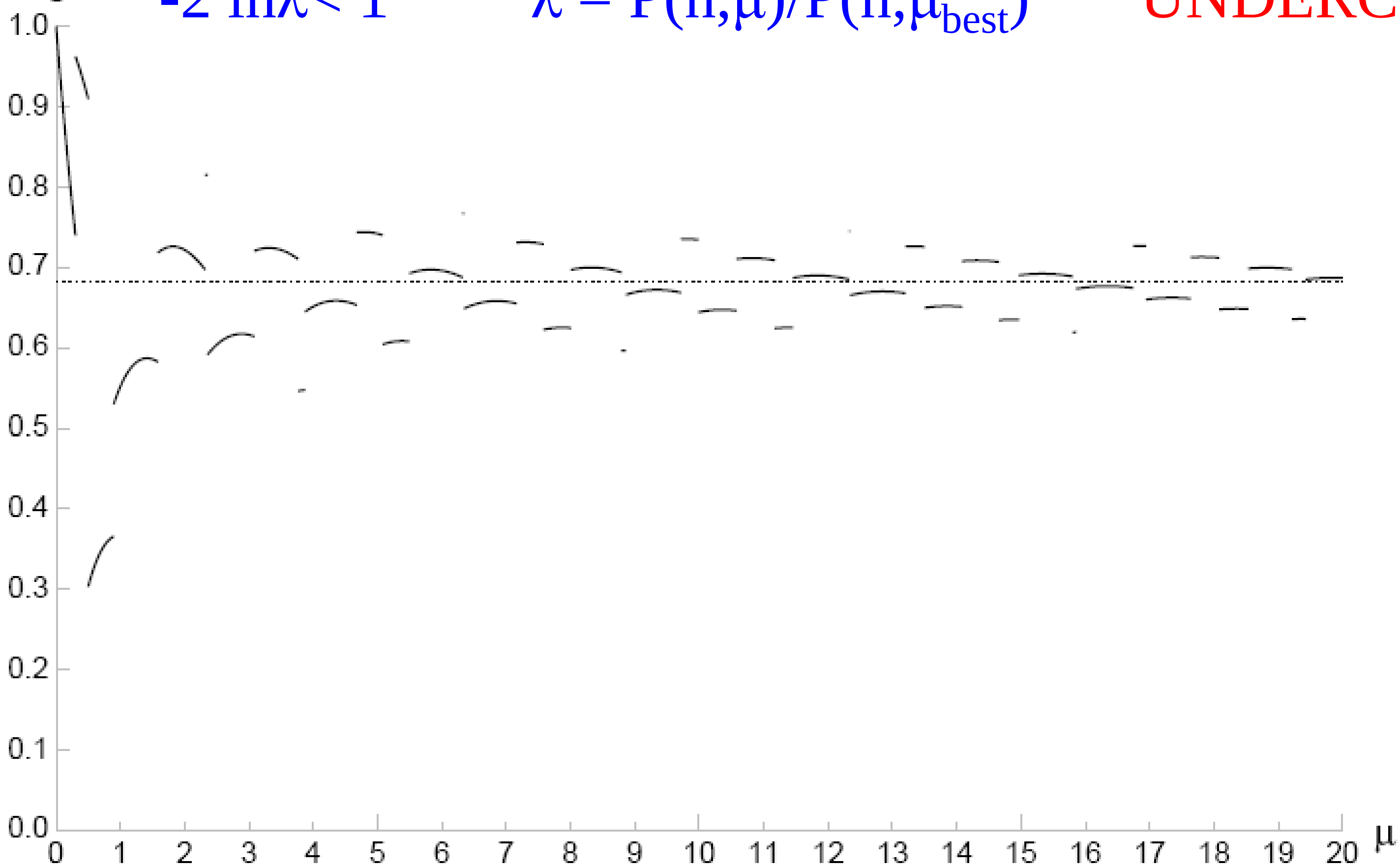
$\mathcal{L}(\mu) = e^{-\mu}\mu^n/n!$ and range of μ from $\ln\{\mathcal{L}(\mu_{best})/\mathcal{L}(\mu)\} < 0.5$

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)

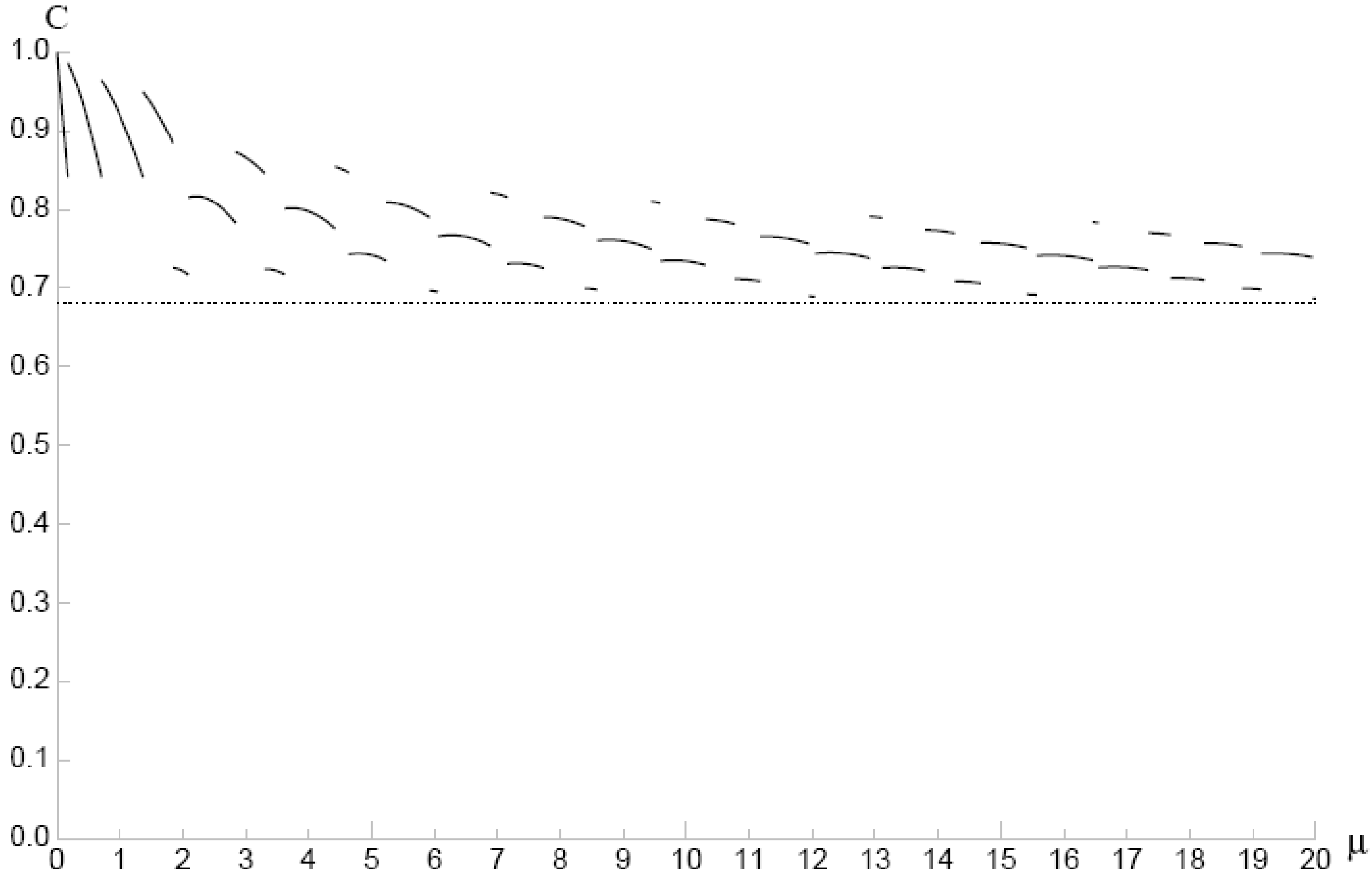
$-2\ \ln\lambda < 1$ $\lambda = P(n,\mu)/P(n,\mu_{best})$ UNDERCOVERS



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

Frequentist central intervals, NEVER undercovers

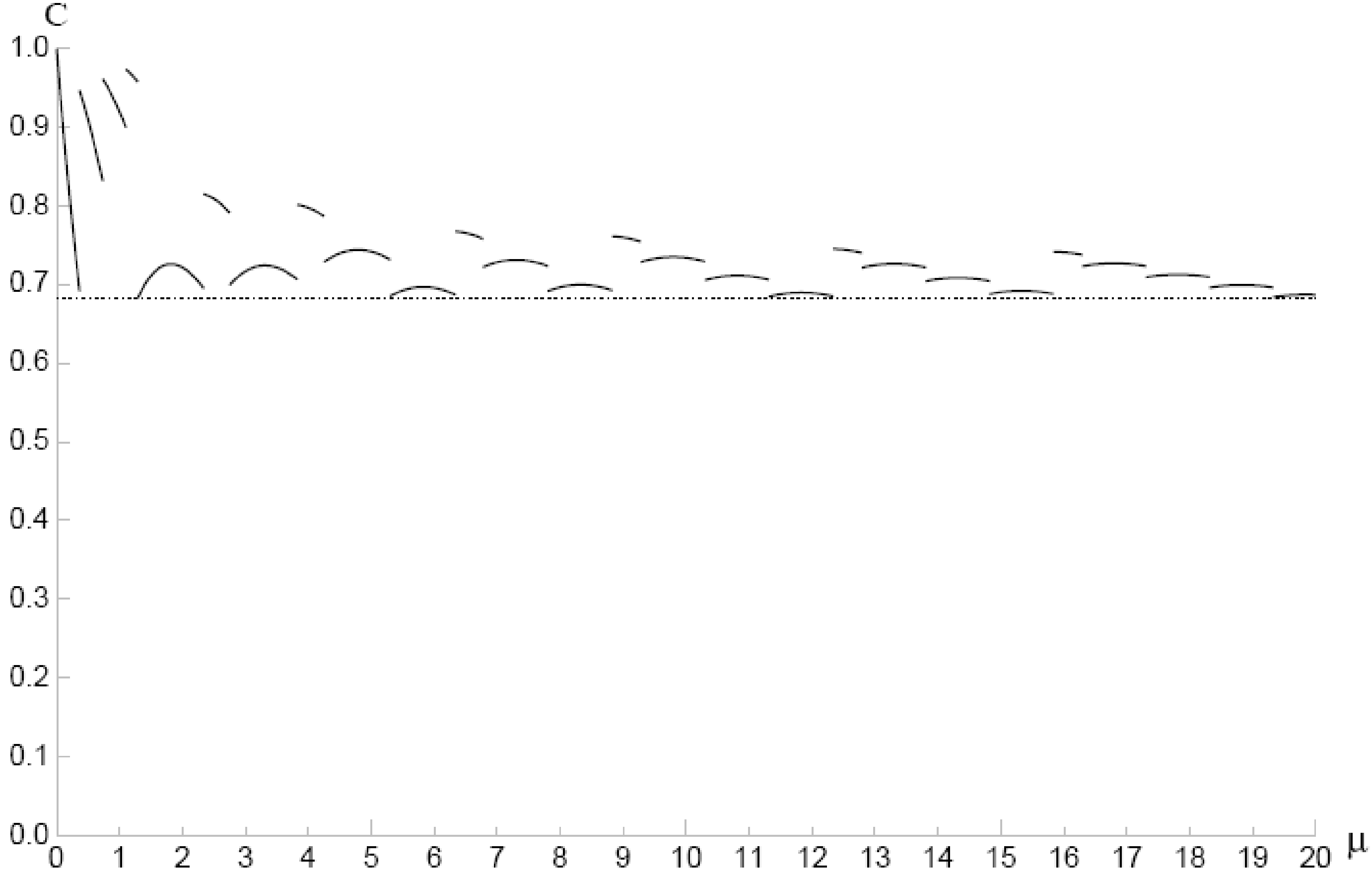
(Conservative at both ends)



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

Feldman-Cousins Unified intervals

Neyman construction, so NEVER undercovers



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

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% Confidence Interval for Gaussian

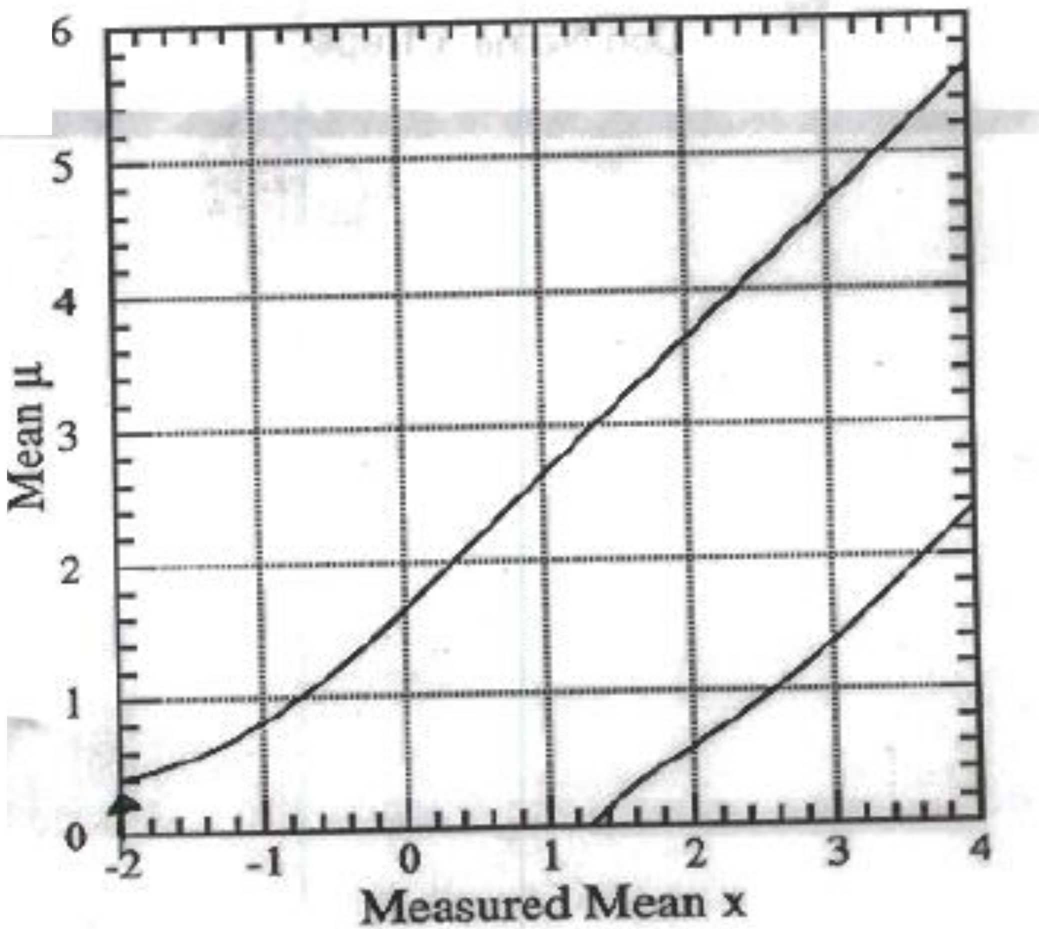
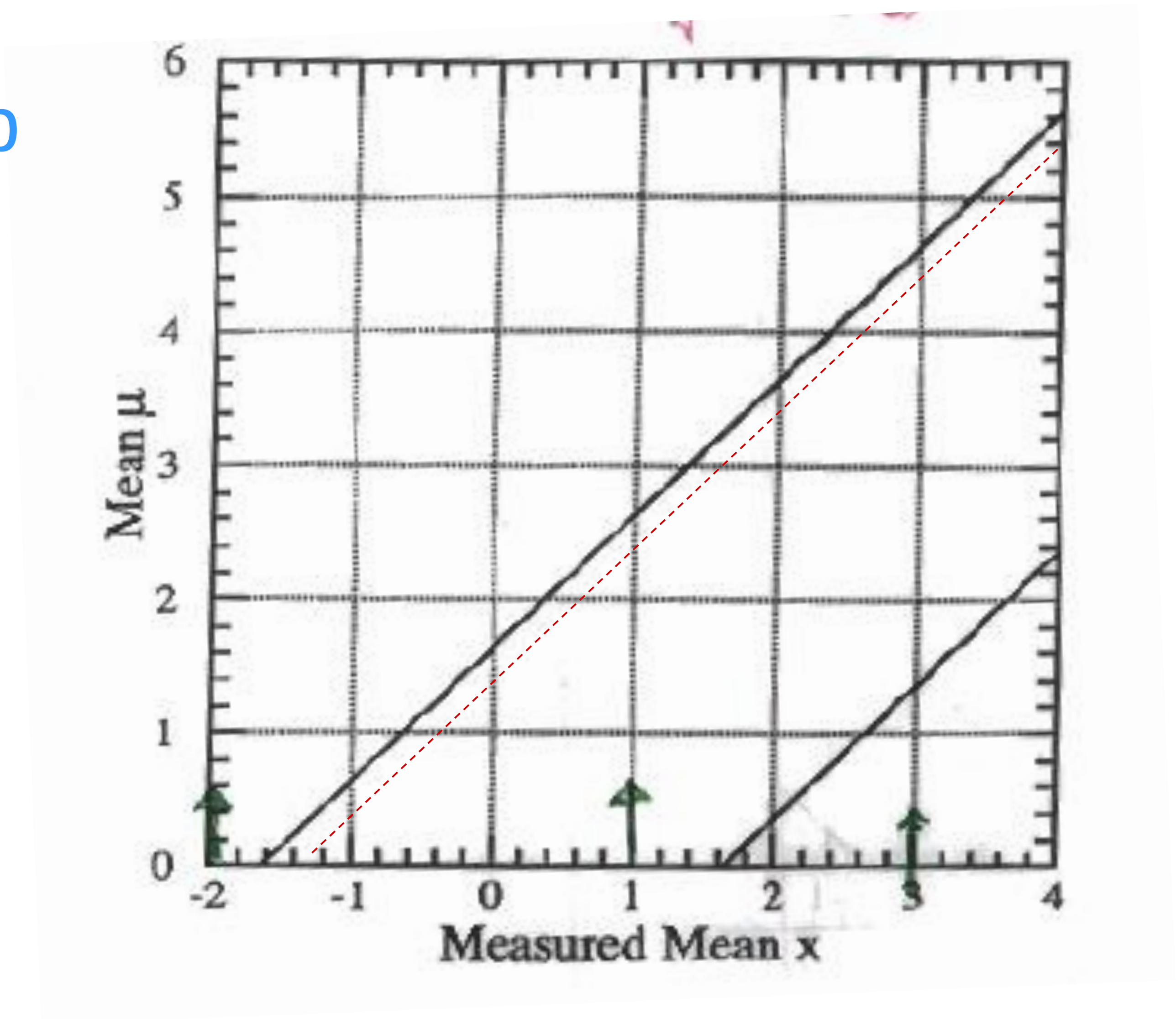


FIG. 10. Plot of our 90% confidence intervals for mean of a Gaussian, constrained to be non-negative, described in the text.

X_{obs} = -2 now gives upper limit

Flip-flop



Black lines Classical 90% central interval

Red dashed: Classical 90% upper limit

FLIP-FLOP

If $x_{obs} < 3$, Upper Limit

If $x_{obs} > 3$, 2-sided interval

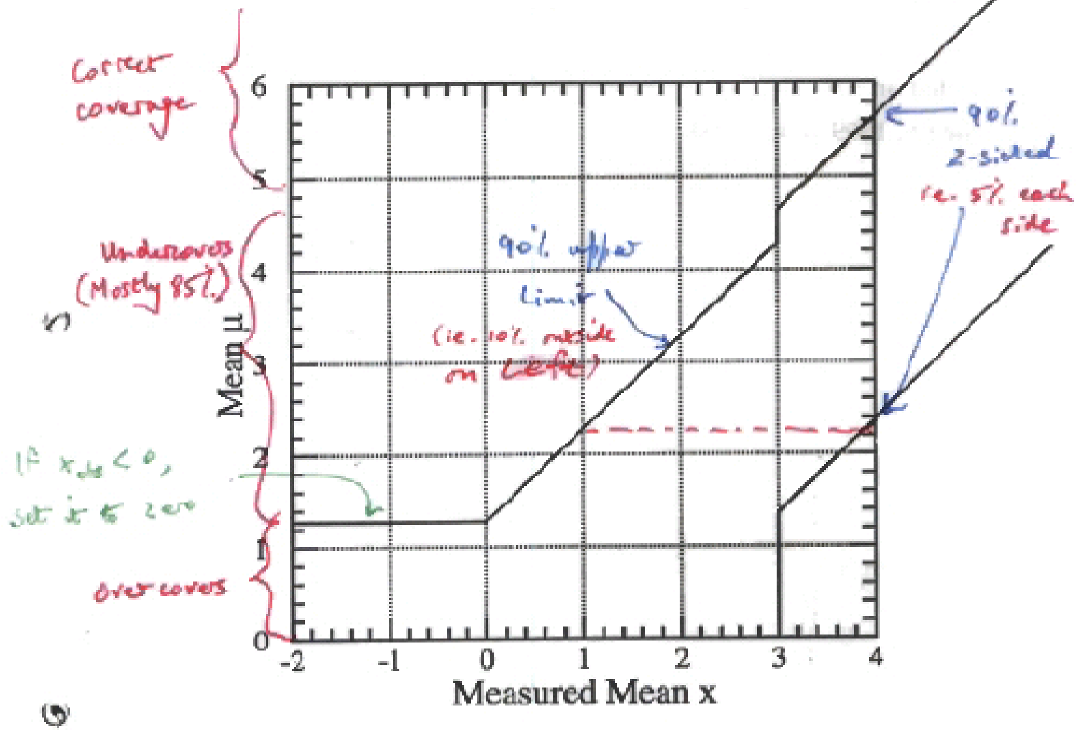


FIG. 4. Plot of confidence belts implicitly used for 90% C.L. confidence intervals (vertical intervals between the belts) quoted by flip-flopping Physicist X, described in the text. They are not valid confidence belts, since they can cover the true value at a frequency less than the stated confidence level. For $1.36 < \mu < 4.28$, the coverage (probability contained in the horizontal acceptance interval) is 85%.

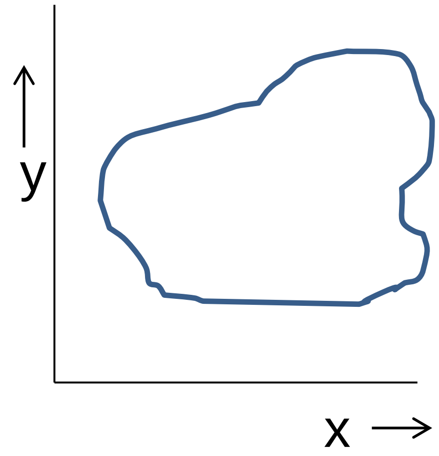
Not good to let x_{obs} determine how result will be presented.

F-C: Move smoothly from 1-sided to 2-sided interval

Features of Feldman-Cousins

Almost no empty intervals
Unified 2-sided and 1-sided intervals
Eliminates flip-flop
No arbitrariness of interval
Less over-coverage than 'x% at both ends'
'Readily' extends to several dimensions

'x% at each end' or 'Max prob density' problematic



Neyman construction time-consuming (esp in n-dimensions)
Minor pathologies: Occasional disjoint intervals
Wrong behaviour wrt background
Tight limits when $b>n_{obs}$ e.g.

n_{obs}	bgd	90% UL
0	3.0	1.08
0	0.0	2.44

Exclusion of s=0 at lower x

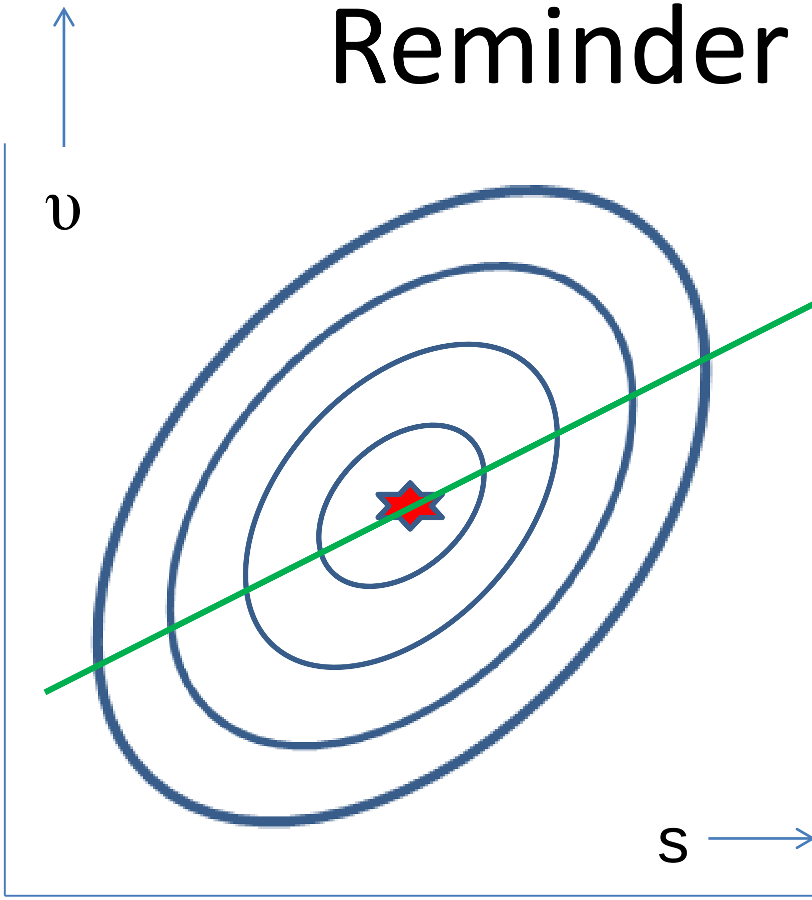
Taking Systematics into account

Result for physics param s depends on systematic param ν
e.g. Mass of $H \rightarrow \gamma\gamma$ depends on energy scale for γ
Subsidiary measurement/info about ν

1) Bayesian:
Post(s:n) = $\int$Post(s,ν;n) dν MARGINALISE
where Post(s,ν;n) = $\mathcal{L}$(n;s,ν) π(s) π(ν) / $\int\mathcal{L}$(n;s,ν) π(s) π(ν) ds dν
$\pi(\nu)$ from subsidiary expt. Maybe Gaussian $N(\nu_0, \sigma_\nu)$
2) $\mathcal{L}_{prof}(s) = \mathcal{L}(s,\nu_{best}(s))$ PROFILE $\rightarrow$
Then use $\mathcal{L}_{prof}(s)$ in likelihood, Frequentist, Bayesian approach
3) Frequentist:
Region in (s,ν) space for which measured values in main and subsid expts were likely. Problematic computationally
4) Mixed (Highland-Cousins):
Frequentist for main expt, but Bayesian smearing over ν

Usually (many) more than one nuisance parameter

Reminder of PROFILE $\mathcal{L}$



Contours of $\ln\mathcal{L}(s,\upsilon)$

s = physics param

υ = nuisance param

Stat uncertainty on s from width of $\mathcal{L}$ fixed at υ_{best}

Total uncertainty on s from width of $\mathcal{L}(s,\upsilon_{prof(s)}) = \mathcal{L}_{prof}$

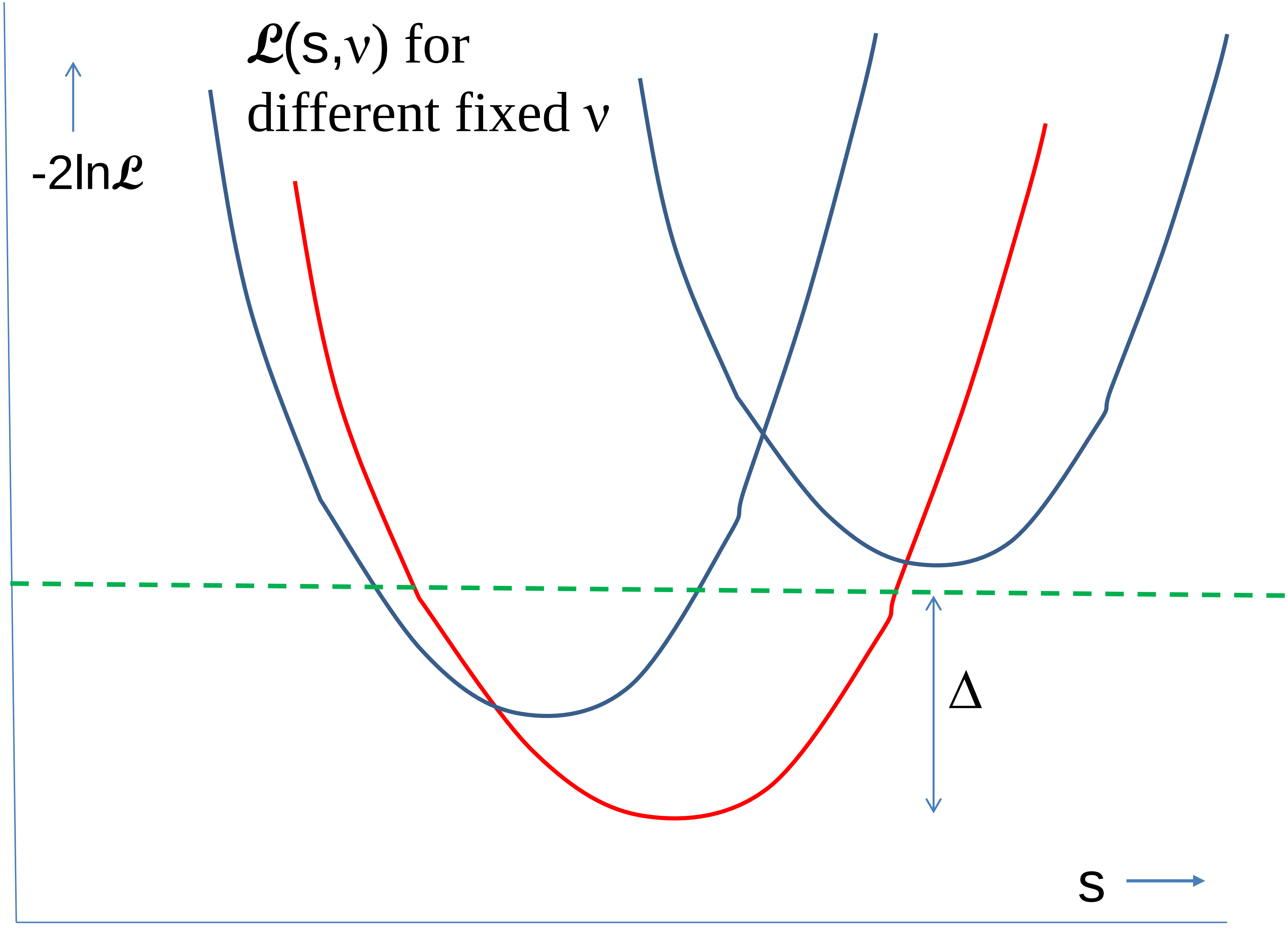
$\upsilon_{prof(s)}$ is best value of υ at that s

$\upsilon_{prof(s)}$ as fn of s lies on **green line**

Total uncert $\geq$ stat uncertainty

Contrast with MARGINALISE

Integrate over ν



$\mathcal{L}(s,\nu)$ for
different fixed ν
$-2\ln\mathcal{L}$
Δ
s

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 for parameter determination

If agree, that's good

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

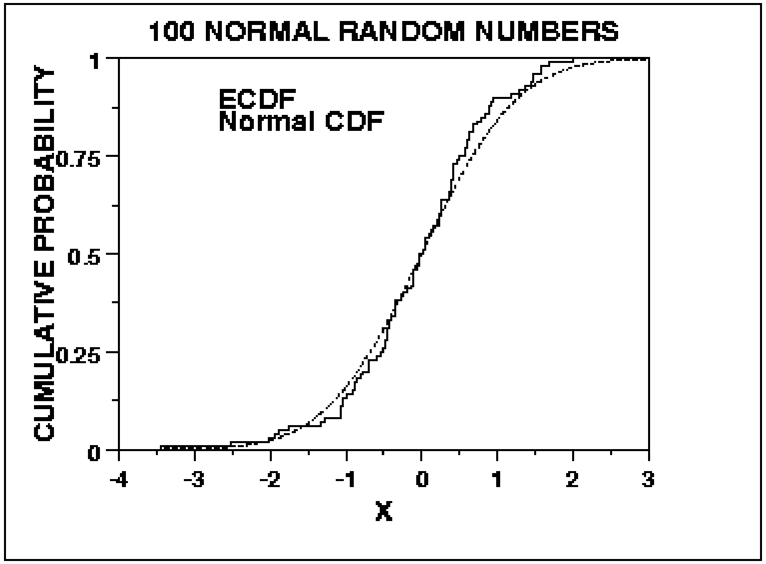
Goodness of Fit: Kolmogorov-Smirnov

Compares data and model cumulative plots

(or 2 sets of data)

Uses largest discrepancy between dists.

Model can be analytic or MC sample



Uses individual data points

Not so sensitive to deviations in tails

(so variants of K-S exist)

Not readily extendible to more dimensions

Distribution-free conversion to p; depends on n

(but not when free parameters involved – needs MC)