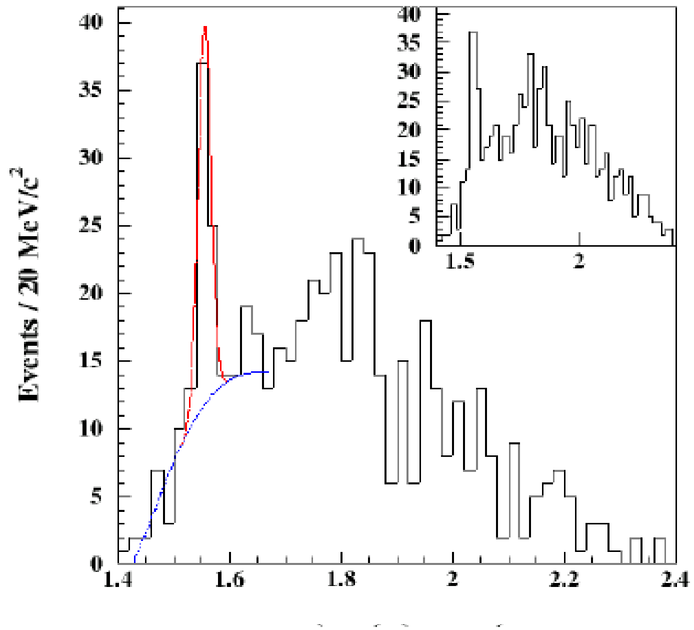
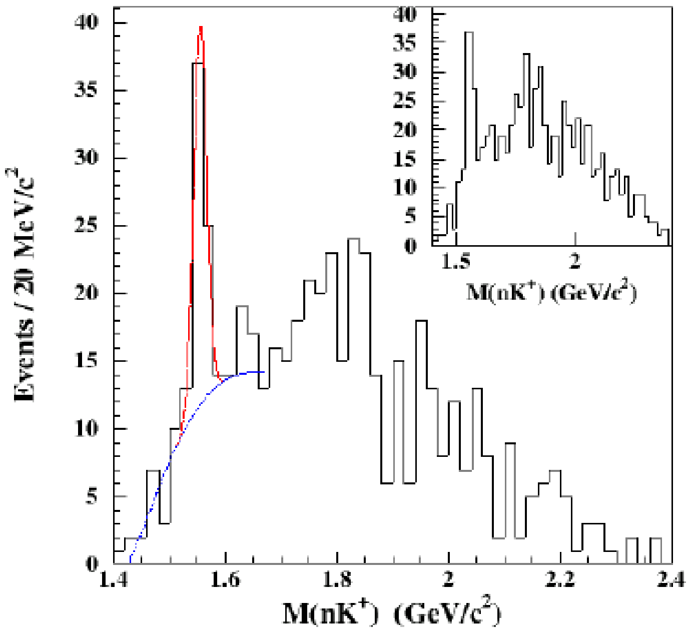
Is there evidence for a peak in this data?



Is there evidence for a peak in this data?



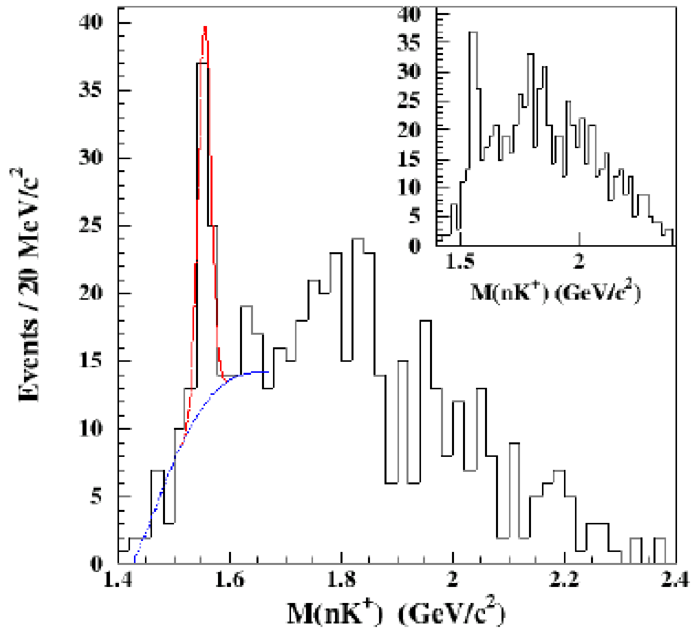
“Observation of an Exotic S=+1

Baryon in Exclusive Photoproduction from the Deuteron”

S. Stepanyan et al, CLAS Collab, Phys.Rev.Lett. 91 (2003) 252001

“The statistical significance of the peak is 5.2 ± 0.6 σ”

Is there evidence for a peak in this data?



"Observation of an Exotic S=+1 Baryon in Exclusive Photoproduction from the Deuteron"

S. Stepanyan et al, CLAS Collab, Phys.Rev.Lett. 91 (2003) 252001

"The statistical significance of the peak is 5.2 ± 0.6 σ"

"A Bayesian analysis of pentaquark signals from CLAS data"

D. G. Ireland et al, CLAS Collab, Phys. Rev. Lett. 100, 052001 (2008)

"The ln(RE) value for g2a (-0.408) indicates weak evidence in favour of the data model without a peak in the spectrum."

Comment on "Bayesian Analysis of Pentaquark Signals from CLAS Data" Bob Cousins, http://arxiv.org/abs/0807.1330

p-values and Discovery

Louis Lyons
IC and Oxford
l.lyons@physics.ox.ac.uk

RAL,

April 2010

PARADOX

Histogram with 100 bins

Fit 1 parameter

S_{min}: χ^2 with NDF = 99 (Expected $\chi^2 = 99 \pm 14$)

For our data, $S_{min}(p_0) = 90$

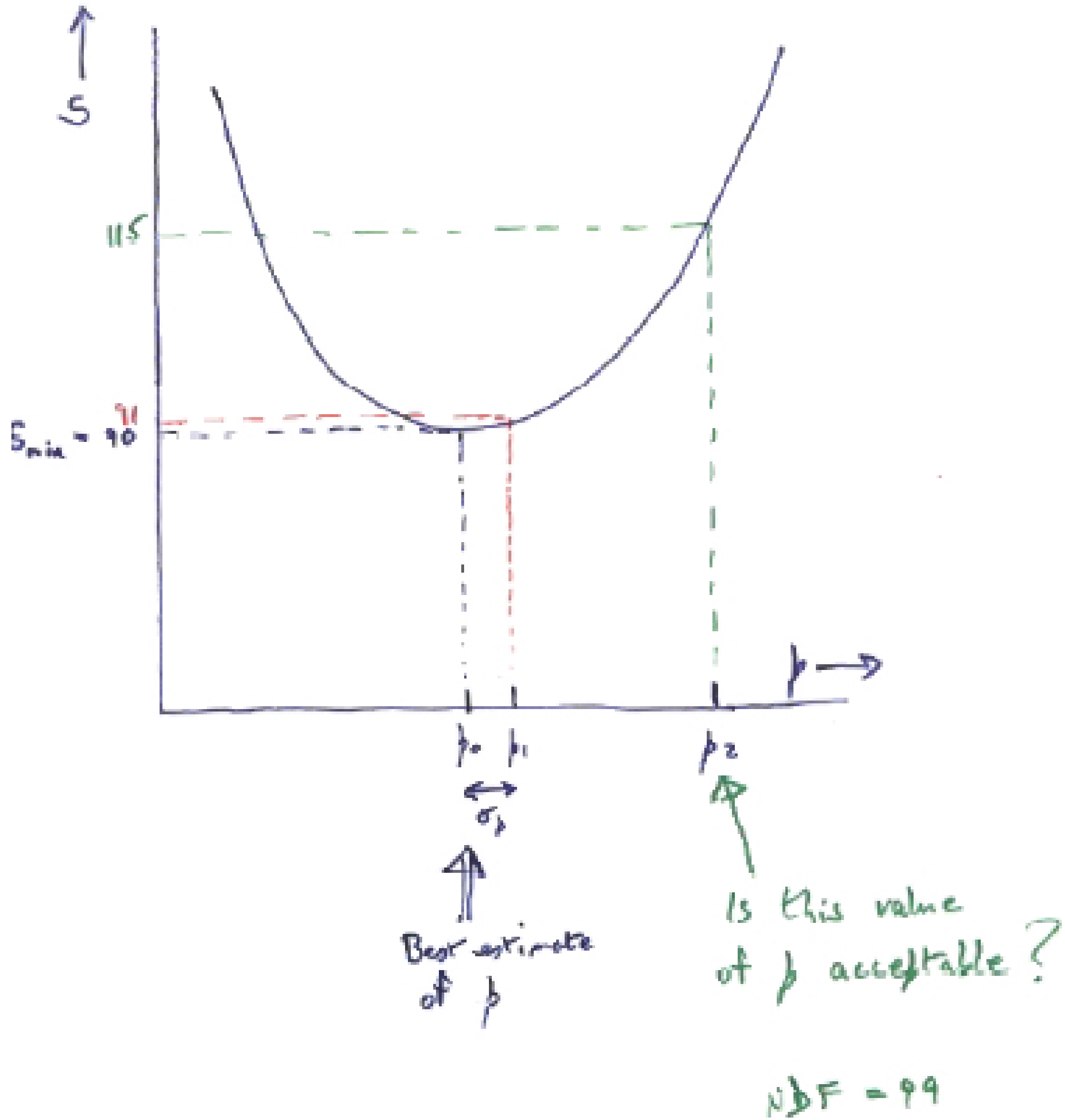
Is p_1 acceptable if $S(p_1) = 115$?

1) YES. Very acceptable χ^2 probability

2) NO. σ_p from $S(p_0 + \sigma_p) = S_{min} + 1 = 91$

But $S(p_1) - S(p_0) = 25$

So p_1 is 5σ away from best value


S
115
91
S min = 90
p
p0
p1
p2
σp
Best estimate of p
Is this value of p acceptable?
NDF = 99

ANOTHER EXAMPLE

$x = -1 \rightarrow +1$

100 bins
~100 events/bin
No FREE PARAMS

H1 : $1 + ax$
$a = 0.05$

H2 : $1 + b\cos(\pi x)$
$b = 0.05$

Generate events according to H1 (+ stat flucts)

Try fitting according to H1 or to H2
χ^2_1 χ^2_2

Look at dist of χ^2_1 As expected for NDF=100

χ^2_2 Bit bigger. Many* "satisfactory"

$\chi^2_2 - \chi^2_1$ Decision based on $\Delta\chi^2$ has much better power

Repeat for events generated according to H2

Look at dist of χ^2_1

χ^2_2

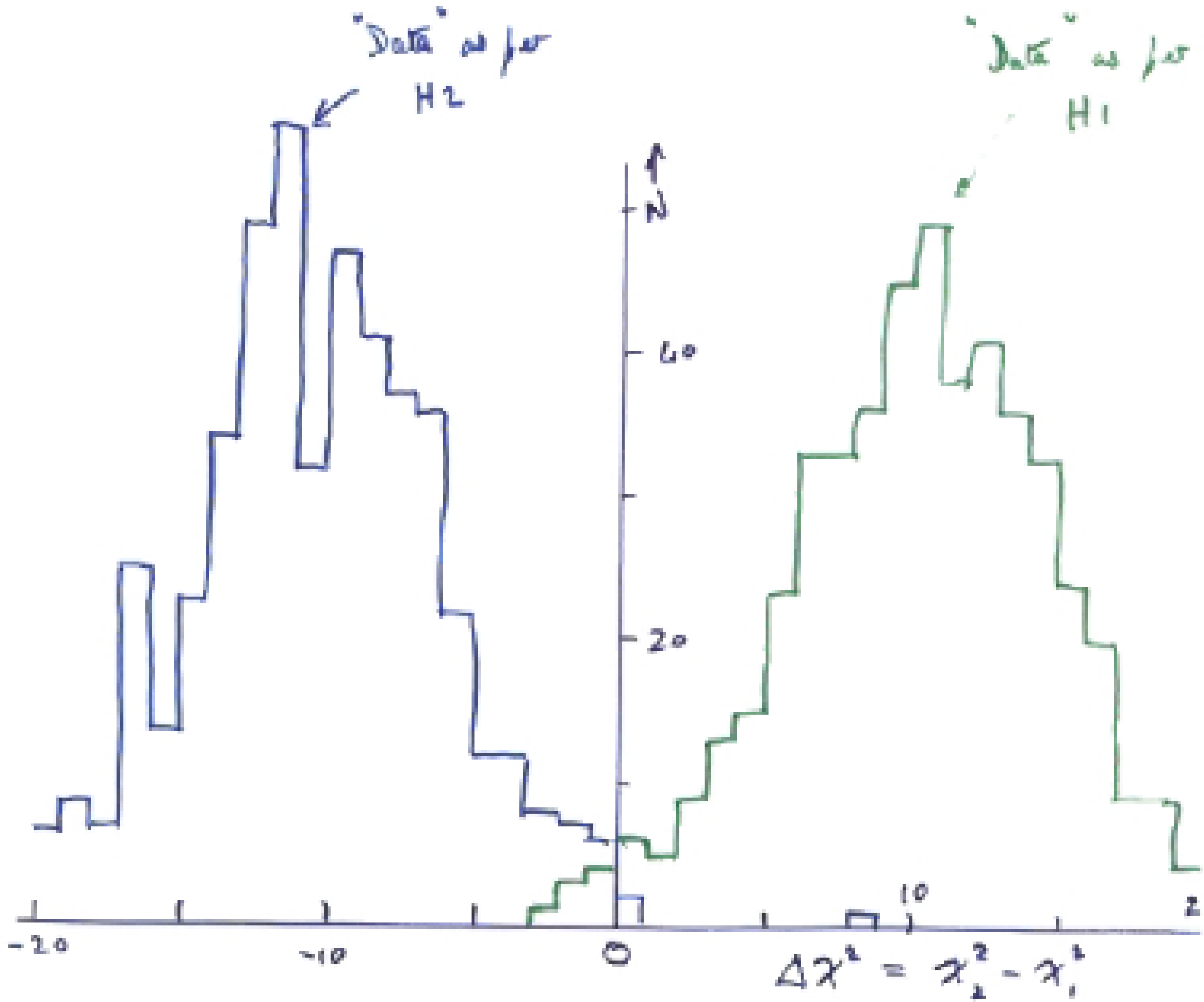
$\chi^2_2 - \chi^2_1$

* 69% have $\chi^2_2 < 130$

24 6

DISTINGUISHING 2 HYPOTHESES ON BASIS OF $\Delta\chi^2$

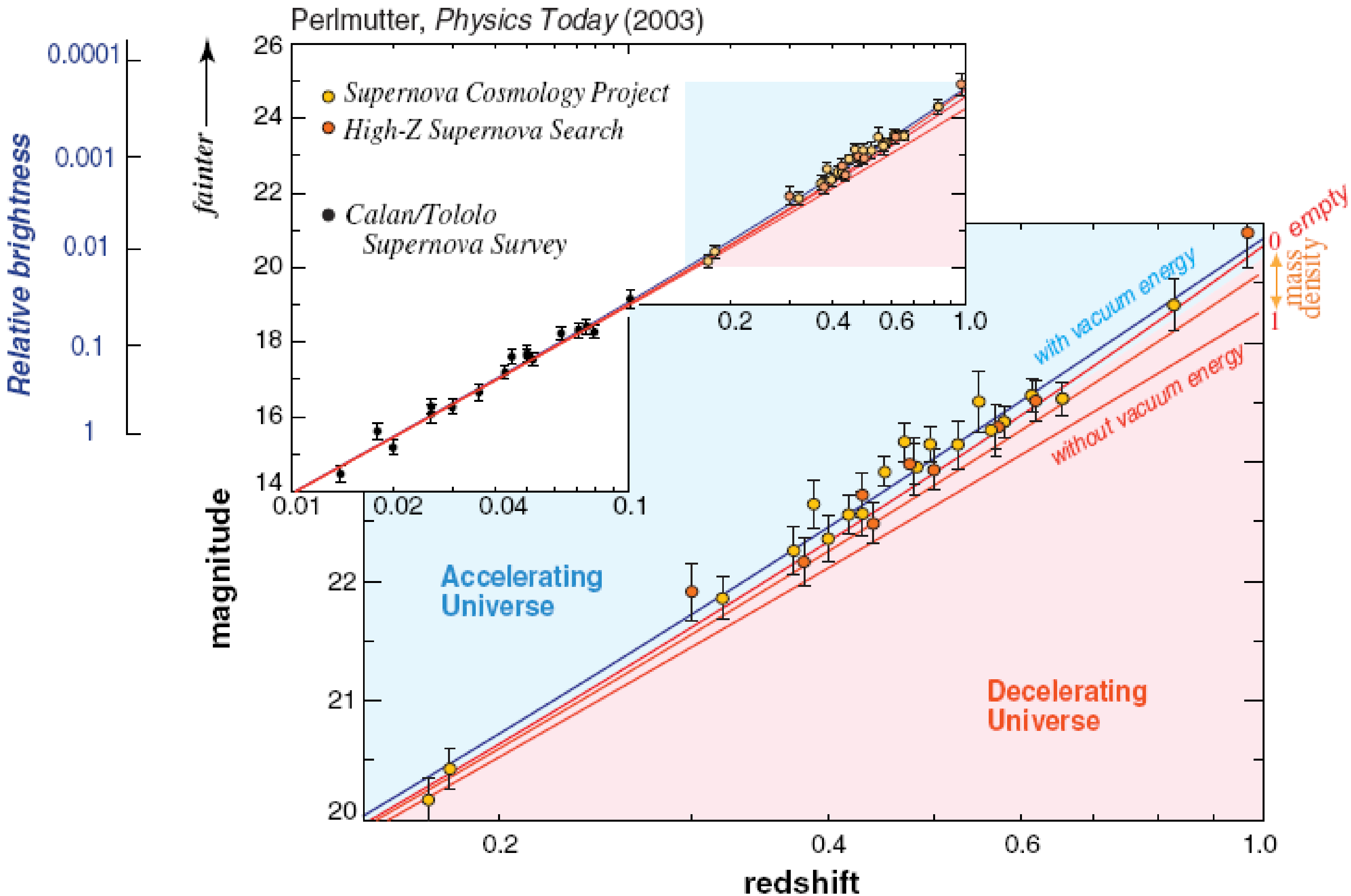
(500 SIMULATIONS)



$H2 = 1 + 0.05\cos(\pi x)$

$H1 = 1 + 0.05x$

Comparing data with different hypotheses



PHYSTAT-LHC Workshop

on

Statistical Issues for LHC Physics

CERN Geneva June 27-29, 2007

Further information and registration at http://cern.ch/phystat-lhc

TOPICS

H0 or H0 v H1

Upper limits

p-values: For Gaussian, Poisson and multi-variate data

Goodness of Fit tests

Why 5σ?

Blind analyses

What is p good for?

Errors of 1st and 2nd kind

What a p-value is not

P(theory|data) ≠ P(data|theory)

Optimising for discovery and exclusion

Incorporating nuisance parameters

H0 or H0 versus H1 ?

H0 = null hypothesis

e.g. Standard Model, with nothing new

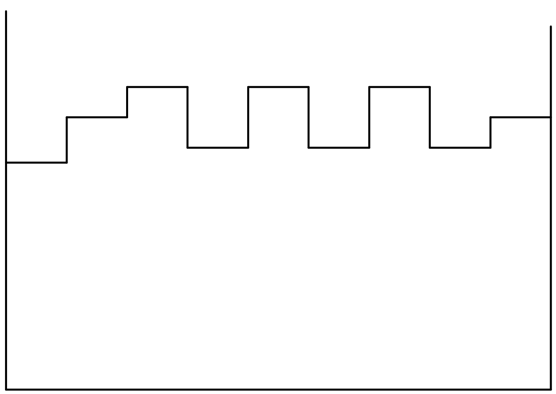
H1 = specific New Physics e.g. Higgs with M_H = 120 GeV

H0: "Goodness of Fit" e.g. χ^2, p-values

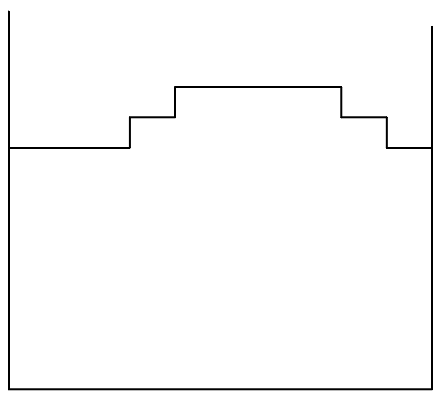
H0 v H1: "Hypothesis Testing" e.g. $\mathcal{L}$-ratio

Measures how much data favours one hypothesis wrt other

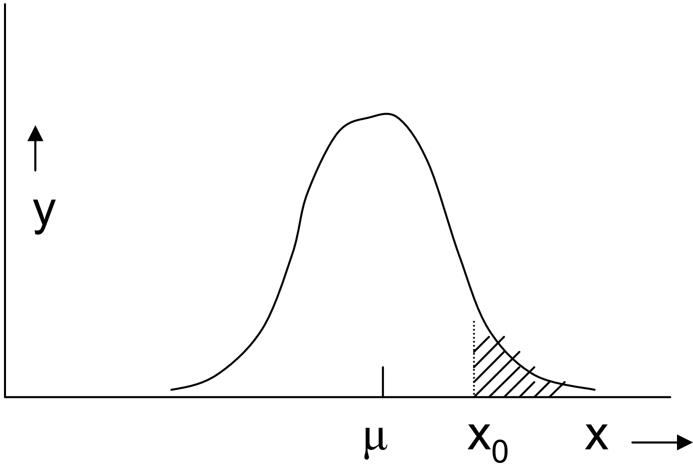
H0 v H1 likely to be more sensitive



or



p-values



Concept of pdf

Example: Gaussian

y = probability density for measurement x

$y = 1/(\sqrt{(2\pi)}\sigma) \exp\{-0.5*(x-\mu)^2/\sigma^2\}$

p-value: probablity that $x \geq x_0$

Gives probability of "extreme" values of data (in interesting direction)

$(x_0-\mu)/\sigma$	1	2	3	4	5
p	16%	2.3%	0.13%	0. 003%	$0.3*10^{-6}$

i.e. Small p = unexpected

p-values, contd

Assumes:

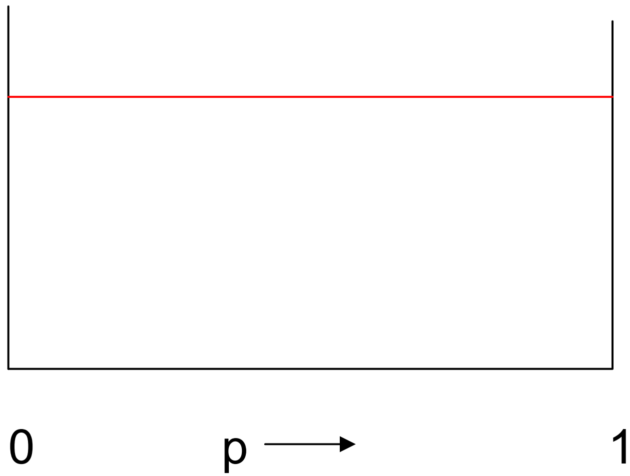
Gaussian pdf (no long tails)

Data is unbiassed

σ is correct

If so, Gaussian x ⟹ uniform p-distribution

(Events at large x give small p)

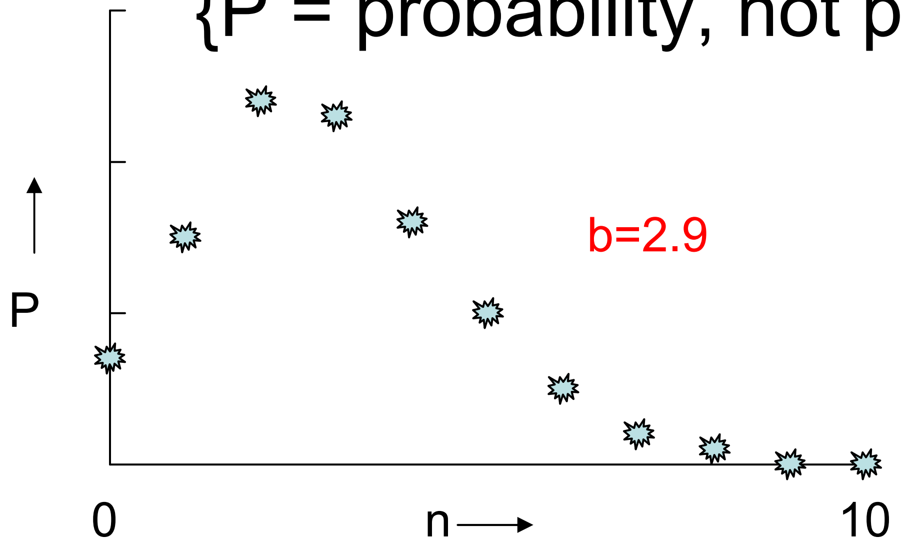


p-values for non-Gaussian distributions

e.g. Poisson counting experiment, bgd = b

$P(n) = e^{-b} * b^n/n!$

{P = probability, not prob density}



For n=7, p = Prob(at least 7 events) = P(7) + P(8) + P(9) +…….. = 0.03

Poisson p-values

n = integer, so p has discrete values

So p distribution cannot be uniform

Replace Prob{$p \le p_0$} = p_0, for continuous p

by Prob{$p \le p_0$} $\le p_0$, for discrete p

(equality for possible p_0)

p-values often converted into equivalent Gaussian σ

e.g. $3*10^{-7}$ is "5σ" (one-sided Gaussian tail)

Does NOT imply that pdf = Gaussian

LIMITS

- Why limits?
- Methods for upper limits
- Desirable properties
- Dealing with systematics
- Feldman-Cousins
- Recommendations

WHY LIMITS?

Michelson-Morley experiment → death of aether

HEP experiments

CERN CLW (Jan 2000)

FNAL CLW (March 2000)

Heinrich, PHYSTAT-LHC, “Review of Banff Challenge”

SIMPLE PROBLEM?

Gaussian

$\sim \exp\{-0.5*(x-\mu)^2/\sigma^2\}$

No restriction on μ, σ known exactly

$\mu \geq x_0 + k\,\sigma$

BUT Poisson {$\mu = s\varepsilon + b$}

$s \geq 0$

ε and b with uncertainties

Not like : 2 + 3 = ?

N.B. Actual limit from experiment ≠ Expected (median) limit

Methods (no systematics)

Bayes (needs priors e.g. const, $1/\mu$, $1/\sqrt{\mu}$, μ, …..)

Frequentist (needs ordering rule,

possible empty intervals, F-C)

Likelihood (DON'T integrate your L)

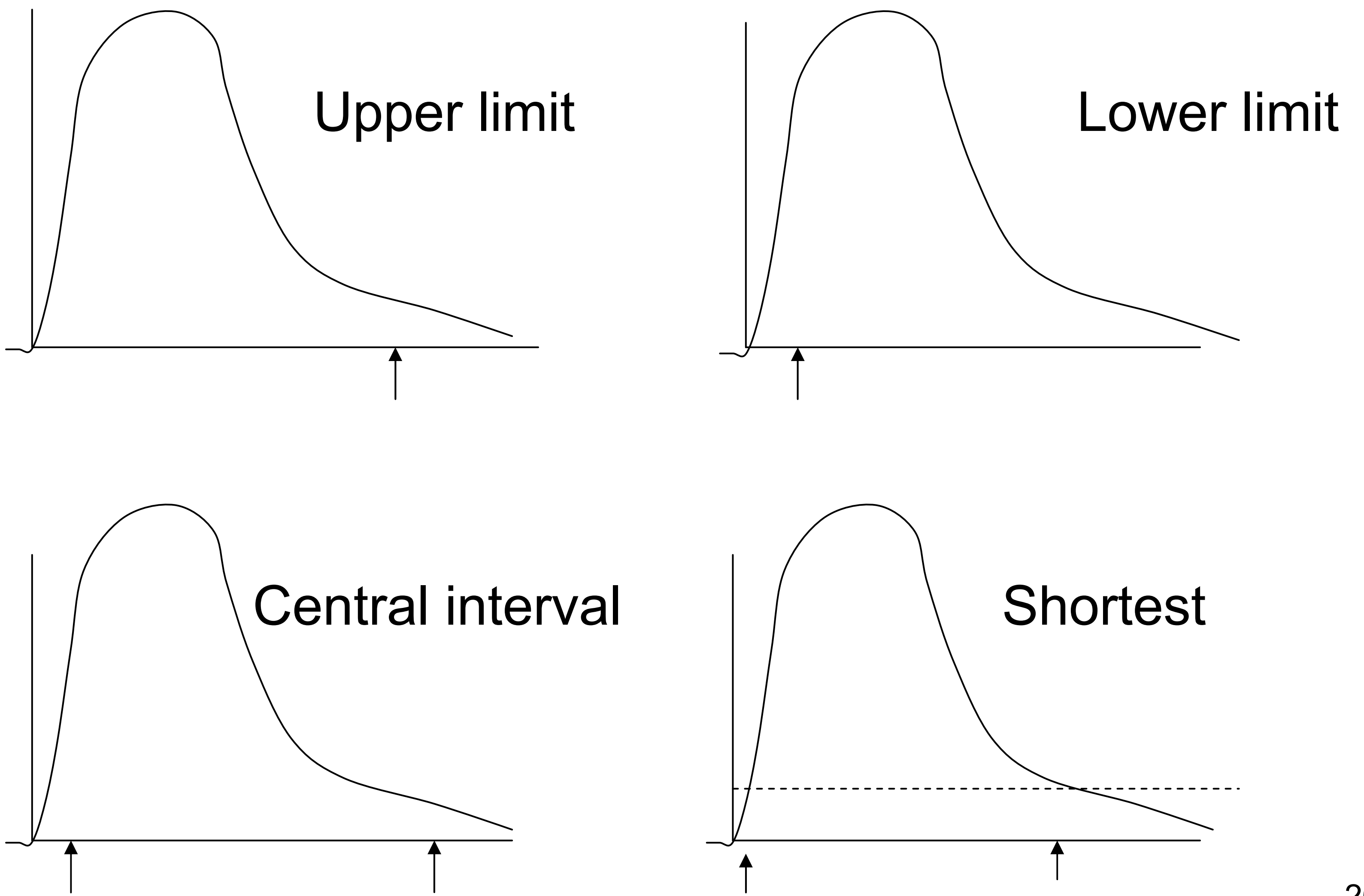
χ^2 ($\sigma^2 = \mu$)

χ^2(σ^2 = n)

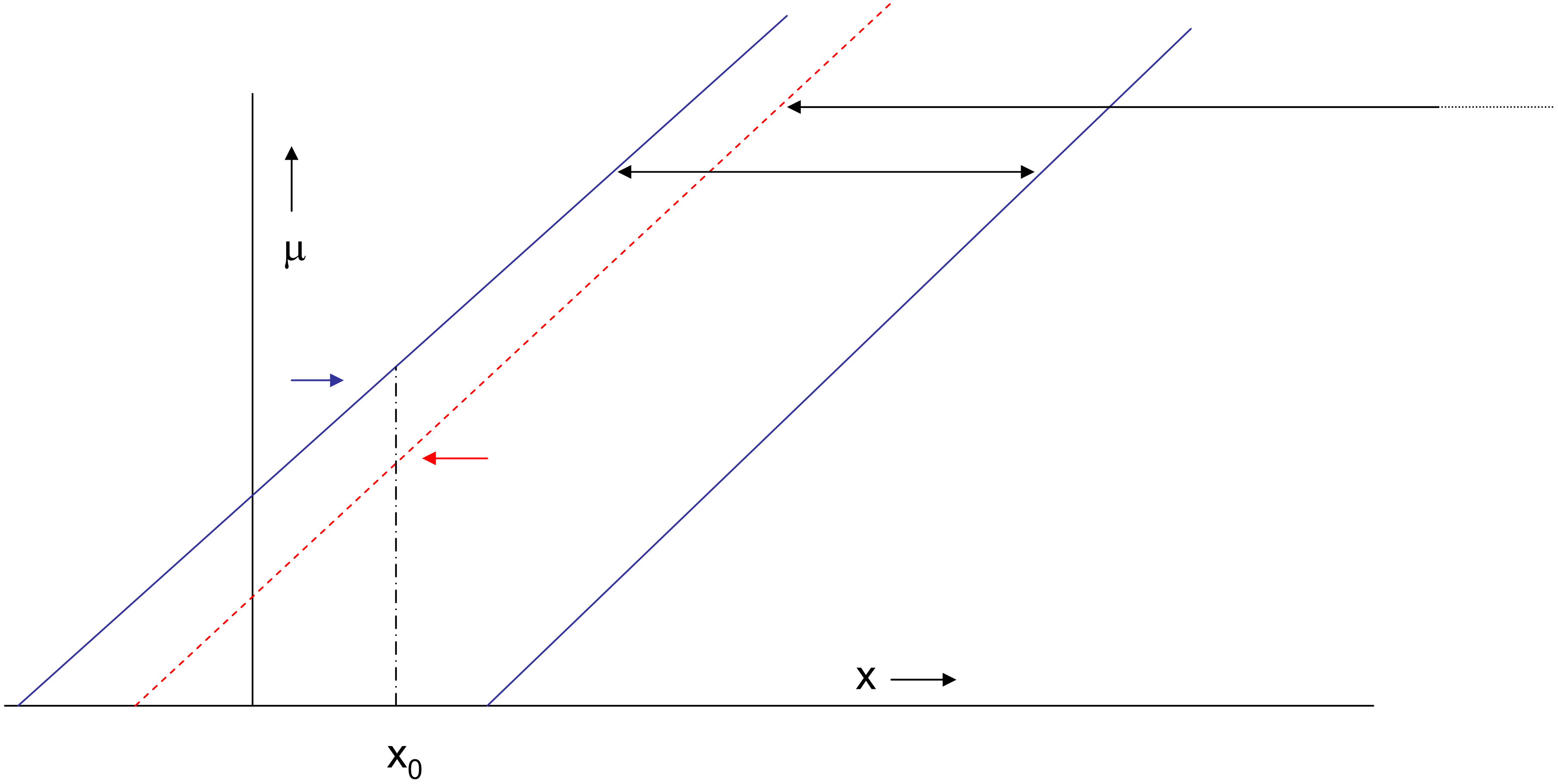
Recommendation 7 from CERN CLW: "Show your **L**"

1) Not always practical
2) Not sufficient for frequentist methods

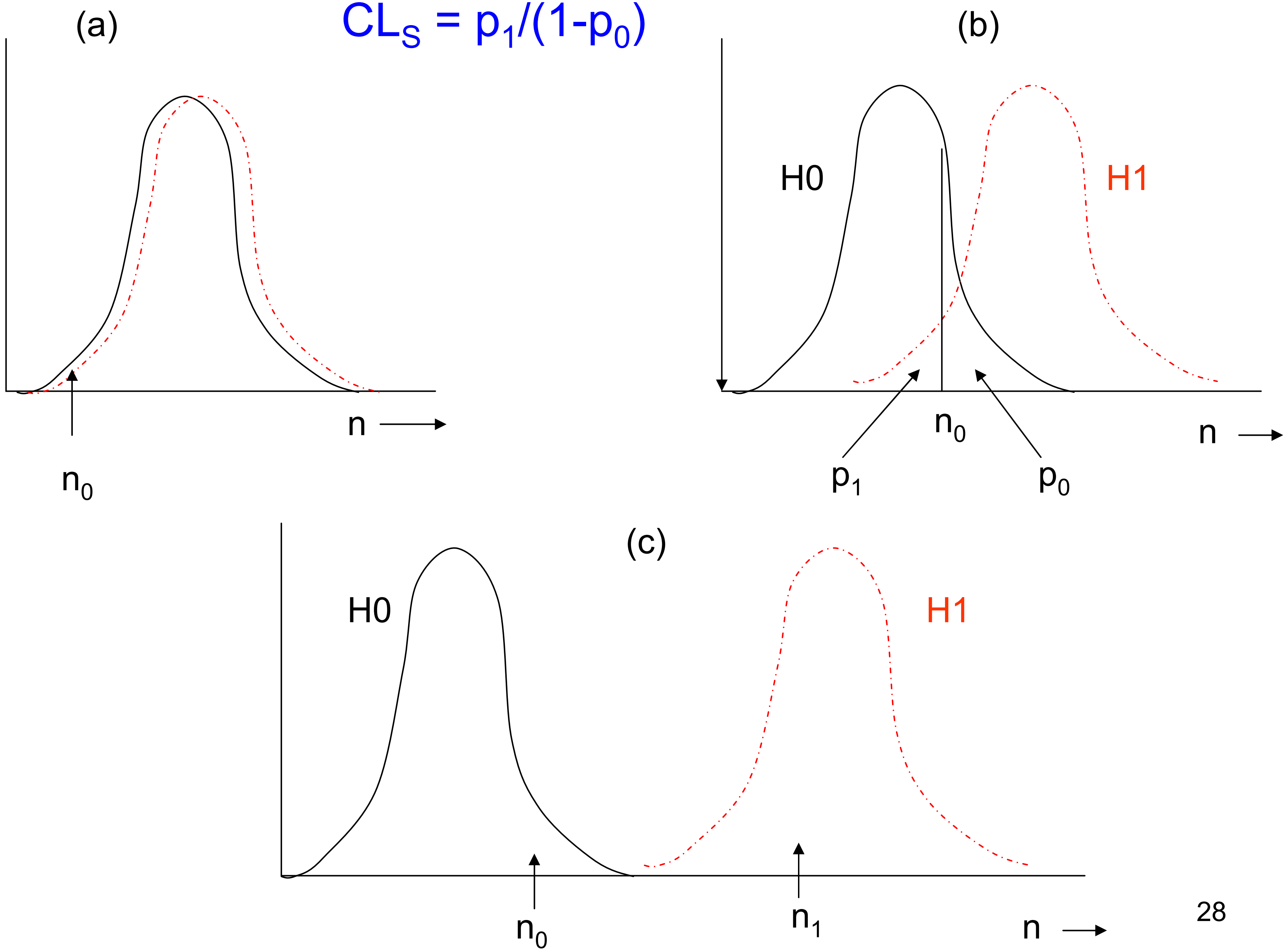
Bayesian posterior → intervals



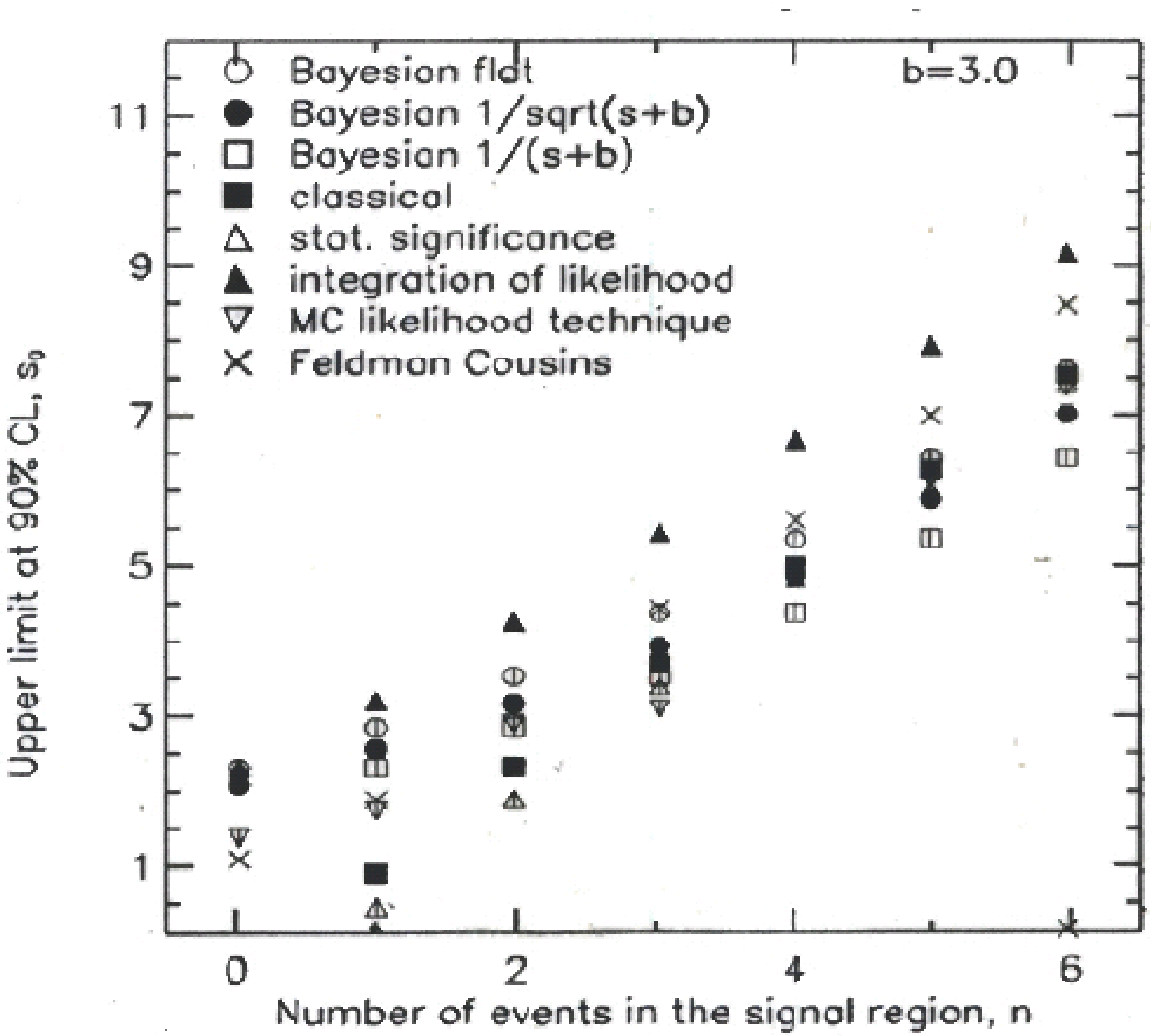
90% C.L. Upper Limits



$CL_S = p_1/(1-p_0)$



Ilya Narsky, FNAL CLW 2000



DESIRABLE PROPERTIES

- Coverage
- Interval length
- Behaviour when $n < b$
- Limit increases as σ_b increases

$\Delta \ln \mathcal{L} = -1/2$ rule

If $\mathcal{L}(\mu)$ is Gaussian, following definitions of σ are equivalent:

1) RMS of $\mathcal{L}(\mu)$

2) $1/\sqrt{(-\ell n\mathcal{L}/d\mu^2)}$

3) $\ln(\mathcal{L}(\mu\pm\sigma) = \ln(\mathcal{L}(\mu_0)) -1/2$

If $\mathcal{L}(\mu)$ is non-Gaussian, these are no longer the same

~~"Procedure 3) above still gives interval that contains the true value of parameter μ with 68% probability"~~

Heinrich: CDF note 6438 (see CDF Statistics Committee Web-page)

Barlow: Phystat05

COVERAGE

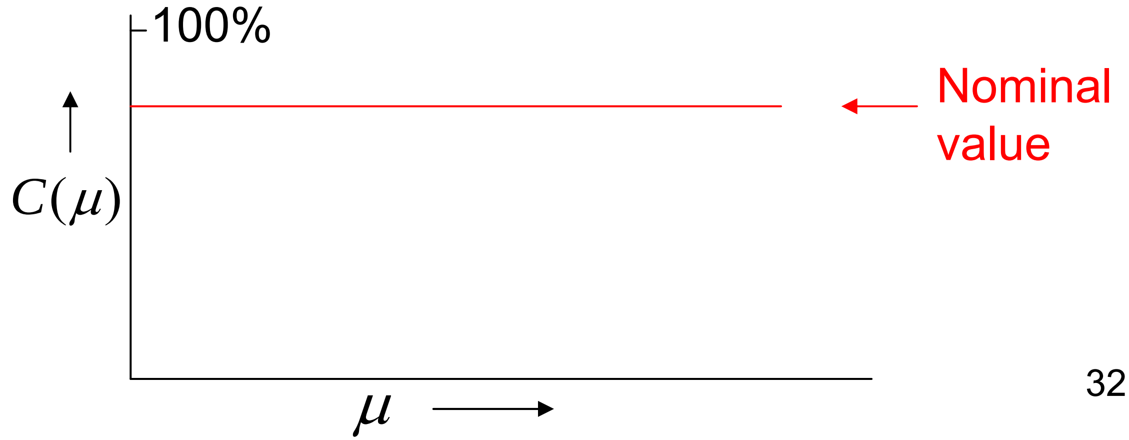
How often does quoted range for parameter include param's true value?

N.B. Coverage is a property of METHOD, not of a particular exptl result

Coverage can vary with μ

Study coverage of different methods of Poisson parameter μ, from observation of number of events n

Hope for:



COVERAGE

If true for all μ : "correct coverage"

P< α for some μ "undercoverage"
(this is serious !)

P> α for some μ "overcoverage"

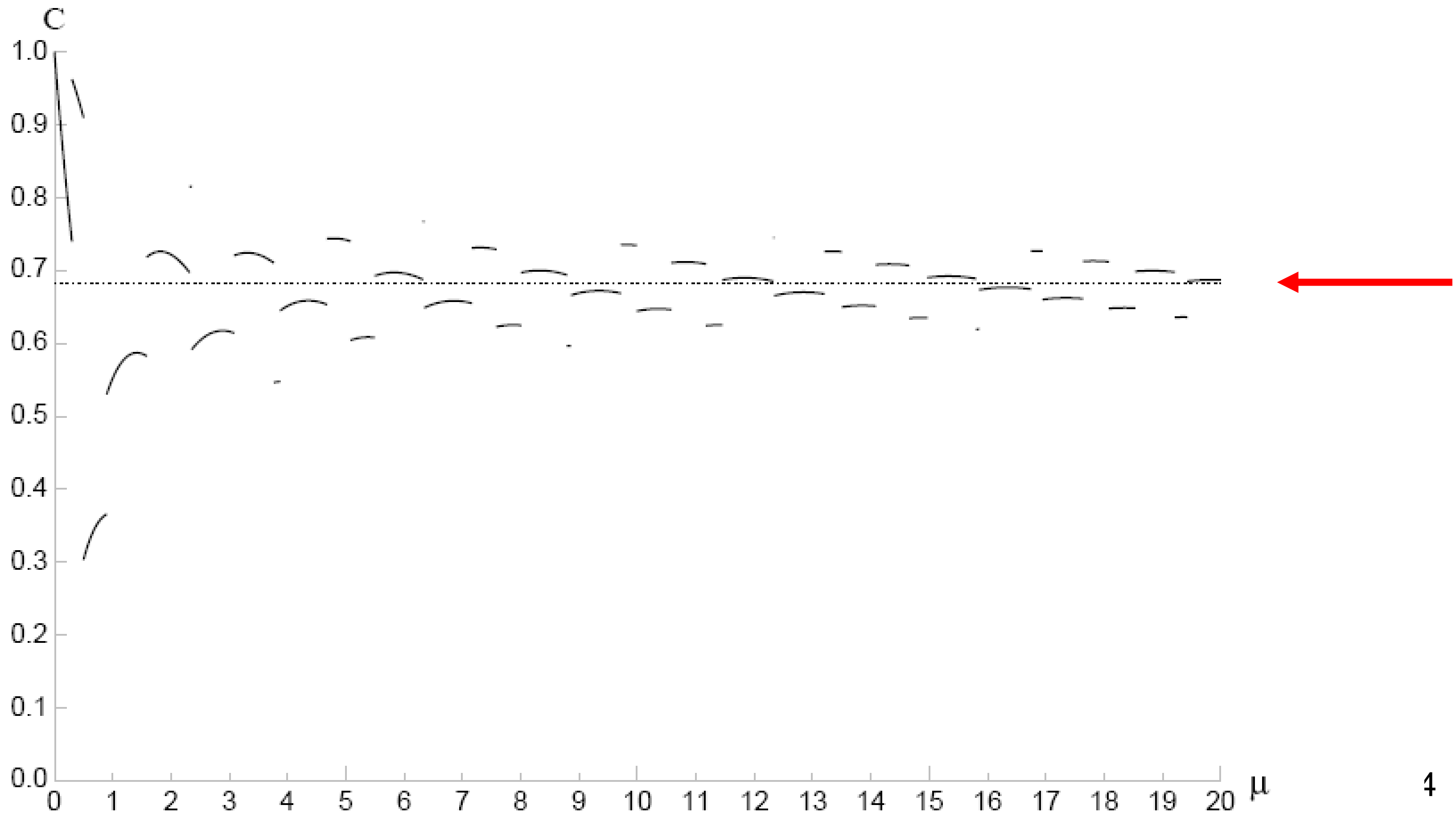
Conservative

Loss of rejection power

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

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

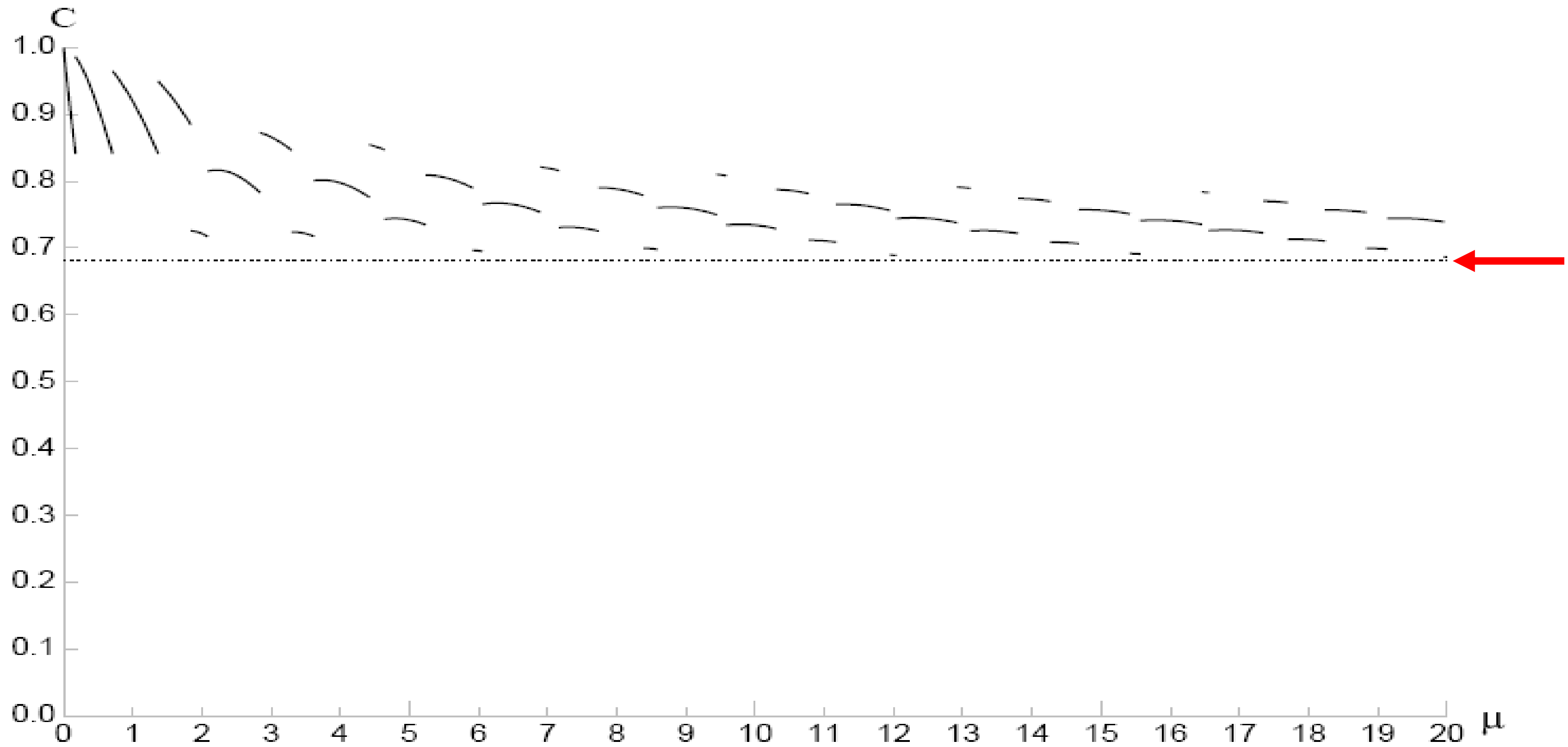
$-2\ \ln\lambda < 1$ $\lambda = P(n,\mu)/P(n,\mu_{best})$ UNDERCOVERS



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

Frequentist central intervals, NEVER undercovers

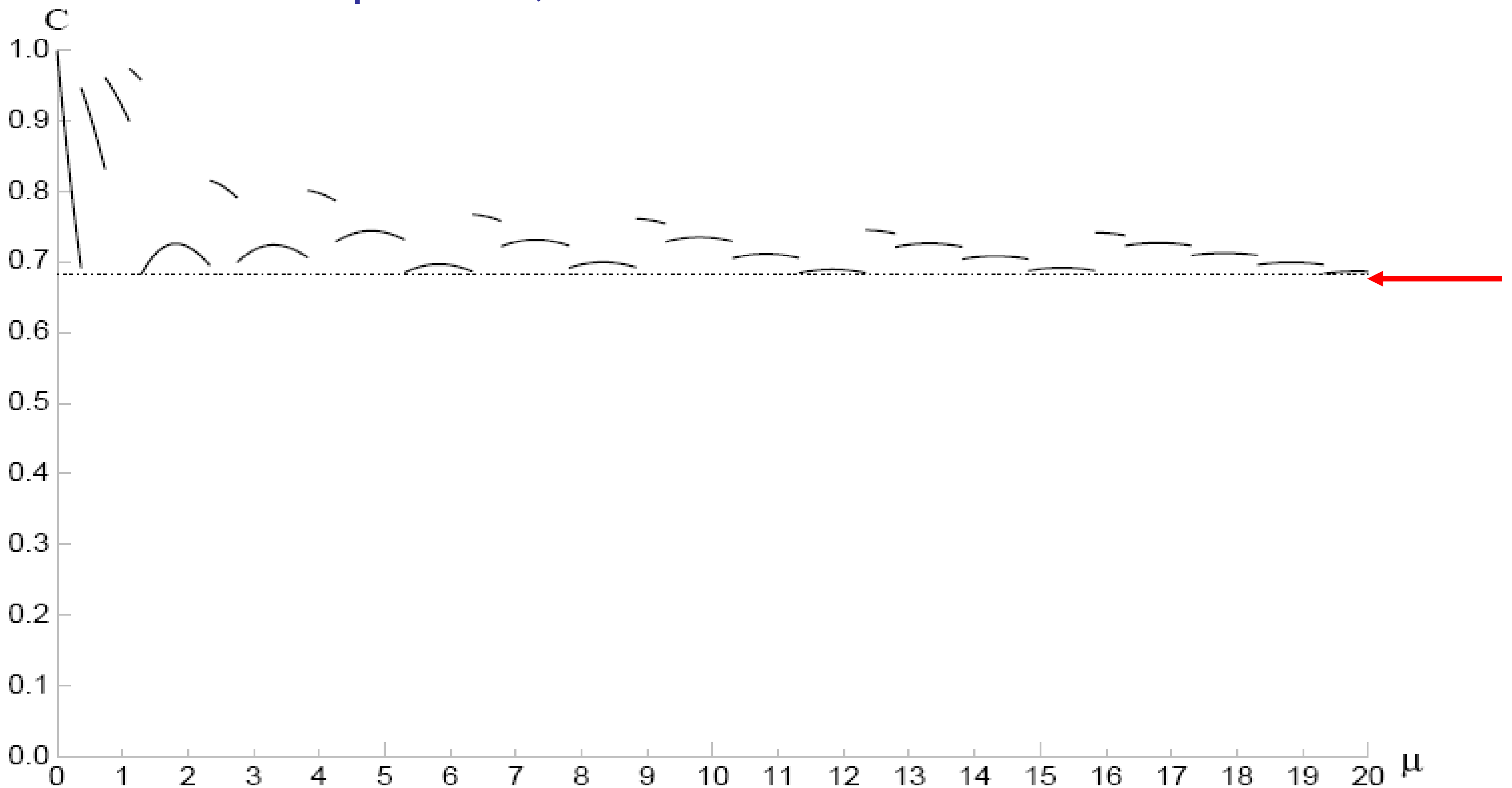
(Conservative at both ends)



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

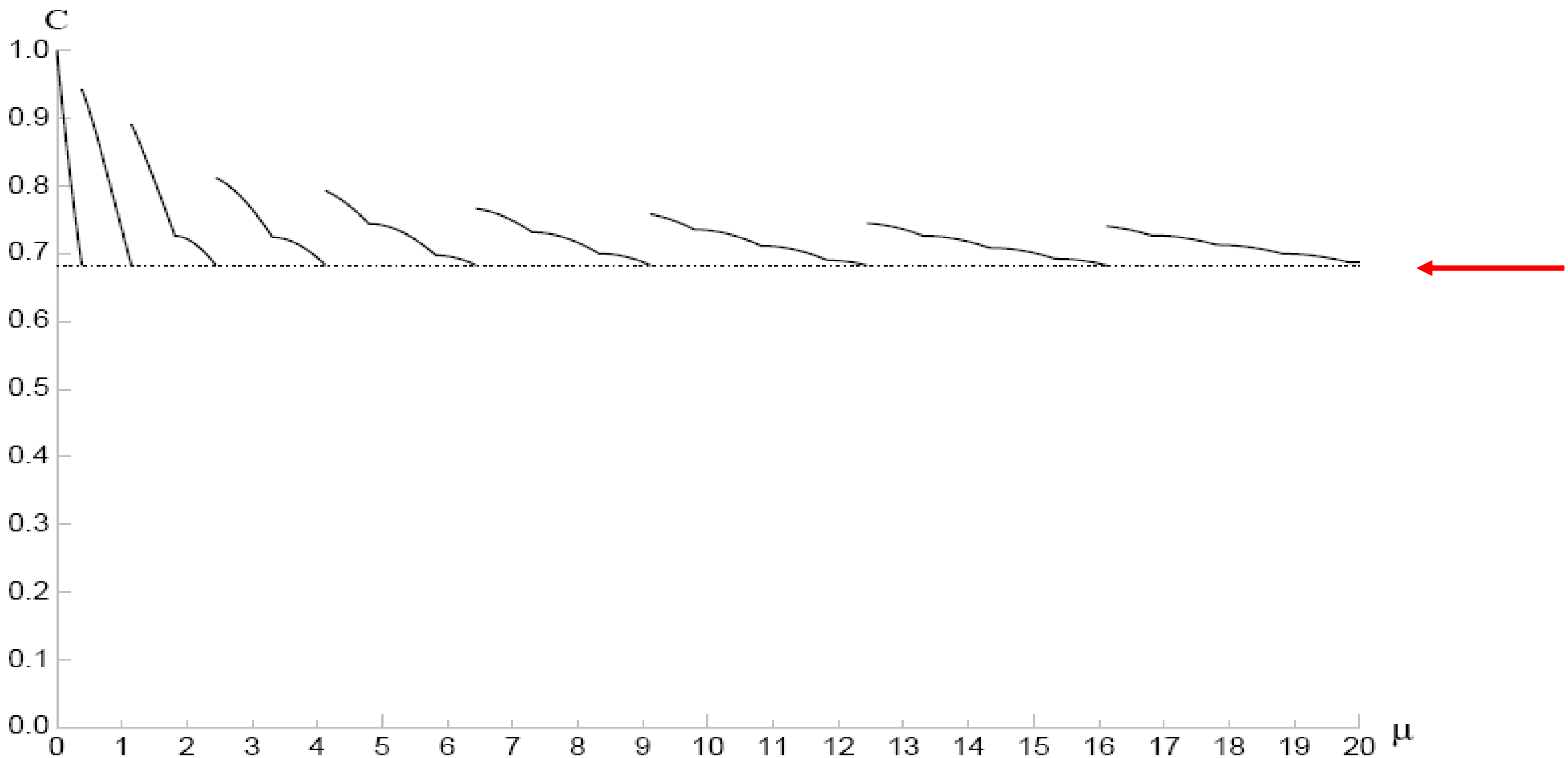
Feldman-Cousins Unified intervals

Frequentist, so NEVER undercovers



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

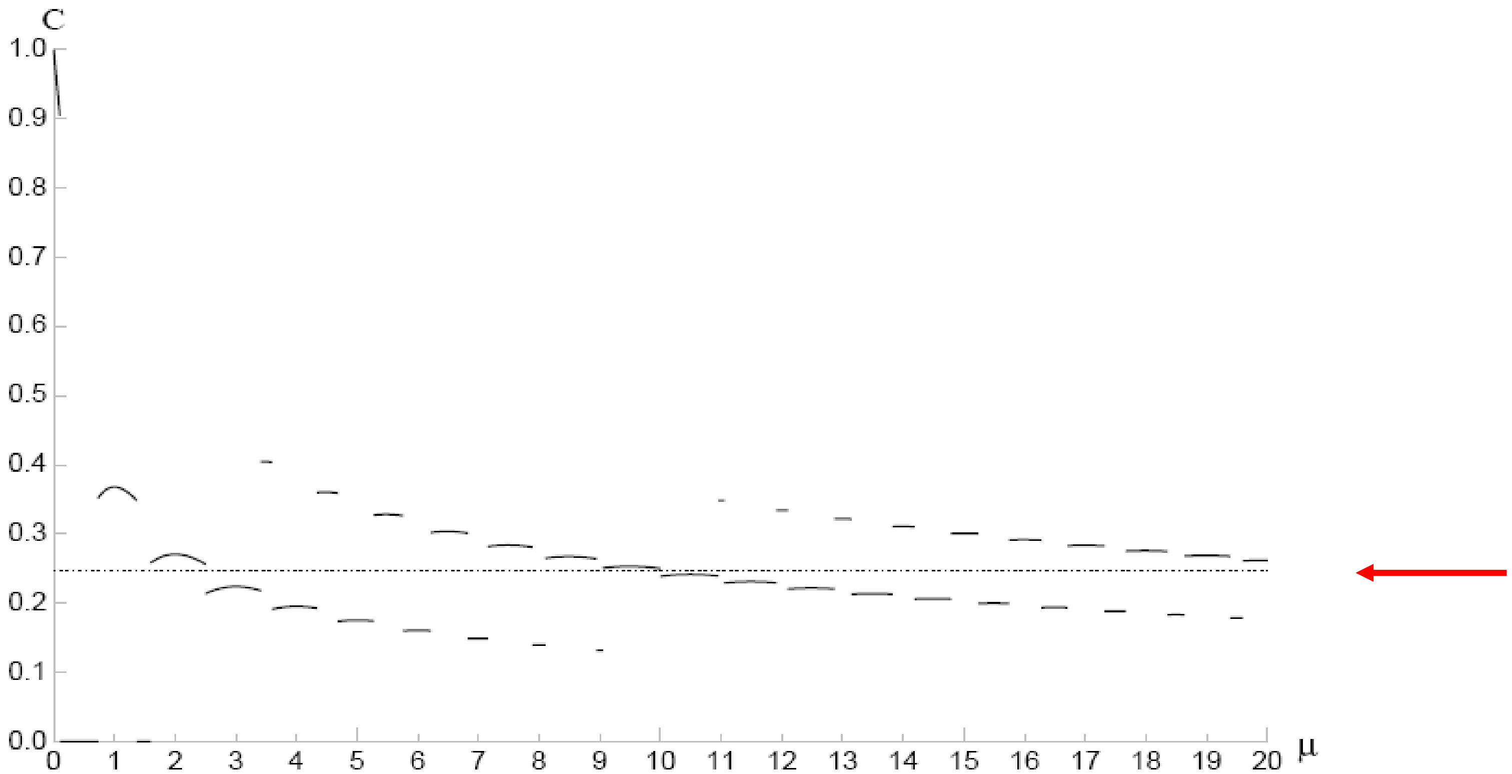
Probability ordering



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

$\chi^2 = (n-\mu)^2/\mu \qquad \Delta\chi^2 = 0.1 \longrightarrow$ 24.8% coverage?

NOT frequentist : Coverage = 0% → 100%



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

COVERAGE

N.B. Coverage alone is not sufficient

e.g. Clifford (CERN CLW, 2000)

"Friend thinks of number

Procedure for providing interval that includes number 90% of time."

COVERAGE

N.B. Coverage alone is not sufficient

e.g. Clifford (CERN CLW, 2000)

Friend thinks of number

Procedure for providing interval that includes number 90% of time.

90%: Interval = $-\infty$ to $+\infty$

10%: number = 102.84590135…..

INTERVAL LENGTH

Empty → Unhappy physicists

Very short→ False impression of sensitivity

Too long→ loss of power

(2-sided intervals are more complicated because 'shorter' is not metric-independent: e.g. 0→4 or 4 →9)

90% Classical interval for Gaussian

$\sigma = 1 \qquad \mu \geq 0 \qquad$ e.g. $m^2(\nu_e)$

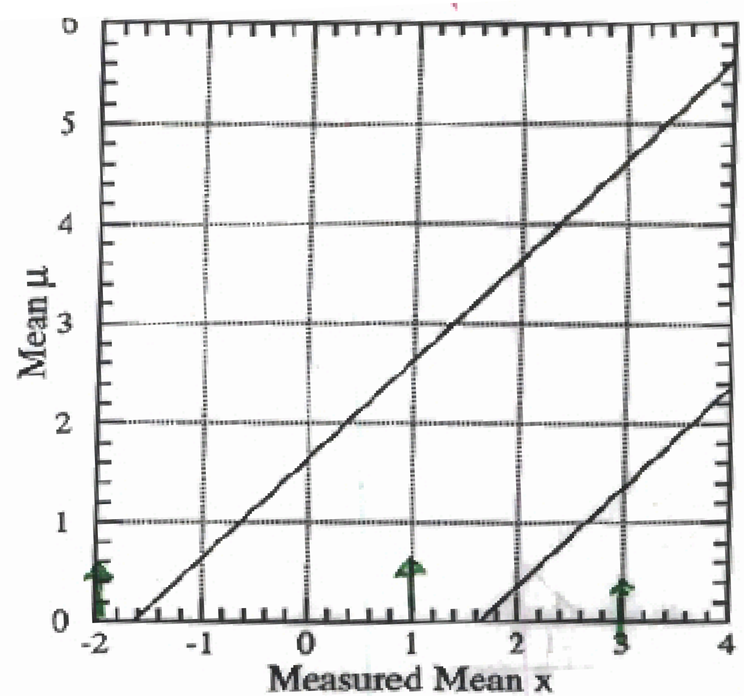


FIG. 3. Standard confidence belt for 90% C.L. central confidence intervals for the mean of a Gaussian, in units of the rms deviation.

$x_{obs} = 3$ Two sided limit

$x_{obs} = 1$ Upper limit

$x_{obs} = -2$ No region for μ

Behaviour when $n < b$

Frequentist: Empty for $n < < b$

Frequentist: Decreases as n decreases below b

Bayes: For $n = 0$, limit independent of b

Sen and Woodroofe: Limit increases as data decreases below expectation

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]

Unified → No 'Flip-Flop' problem

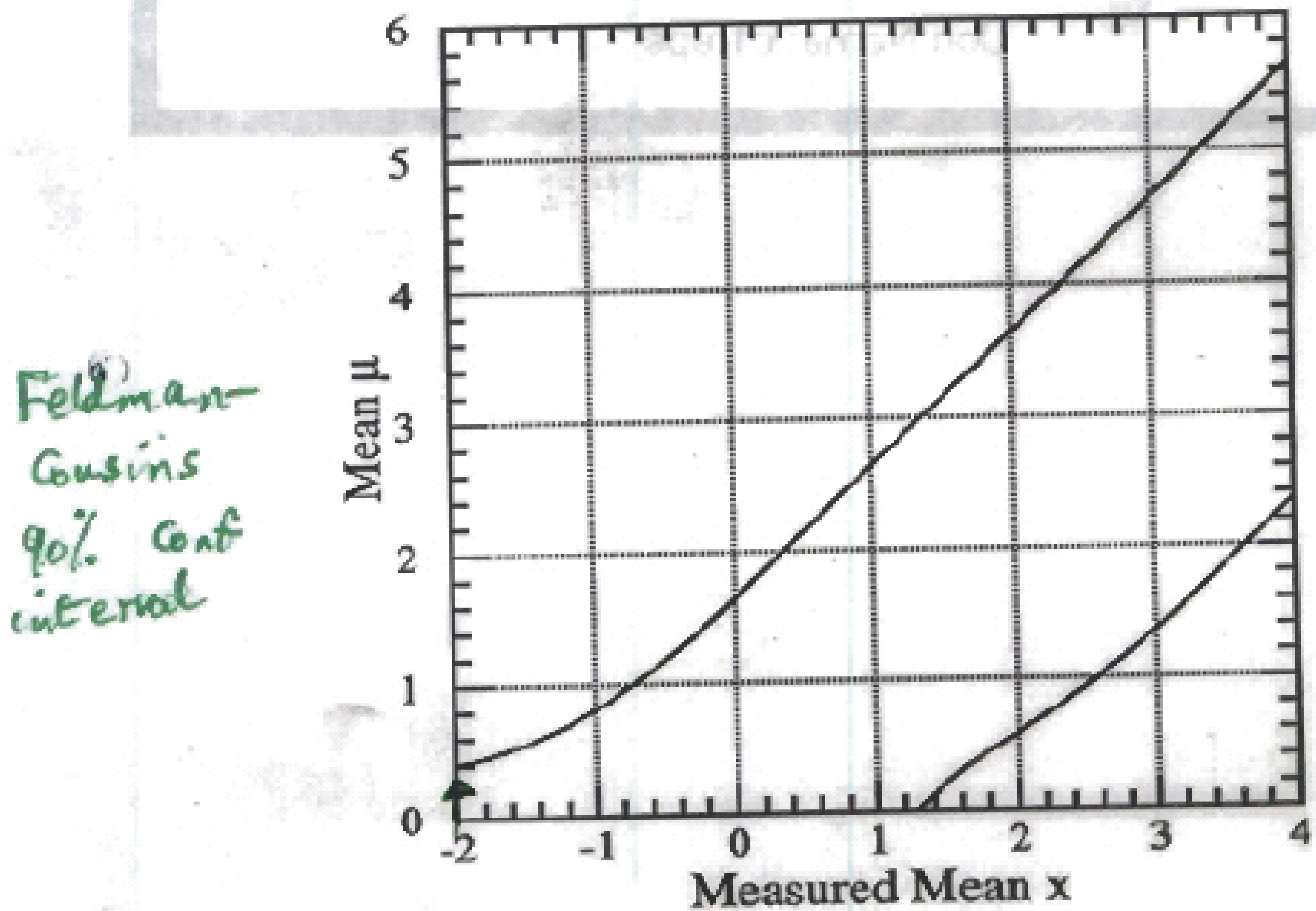


FIG. 10. Plot of our 90% confidence intervals for mean of a Gaussian, constrained to be non-negative, described in the text.

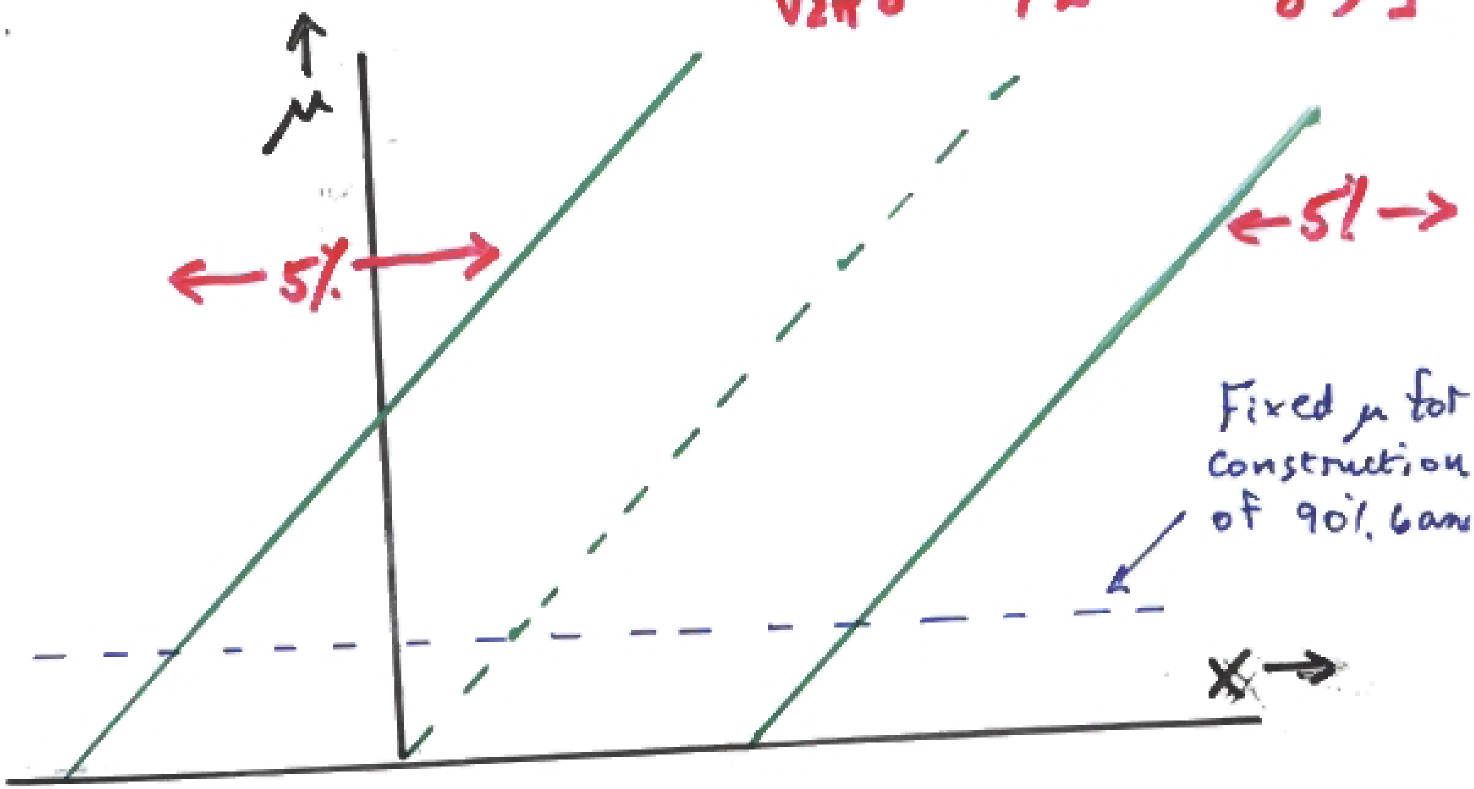
X_{obs} = -2 now gives upper limit

FELDMAN-COUSINS ORDERING RULE

$R = p(x,\mu)/p(x,\mu_{best})$ [Likelihood ratio ordering]

Gaussian example $p(x,\mu) = G(x,\mu,\sigma)$

$$= \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

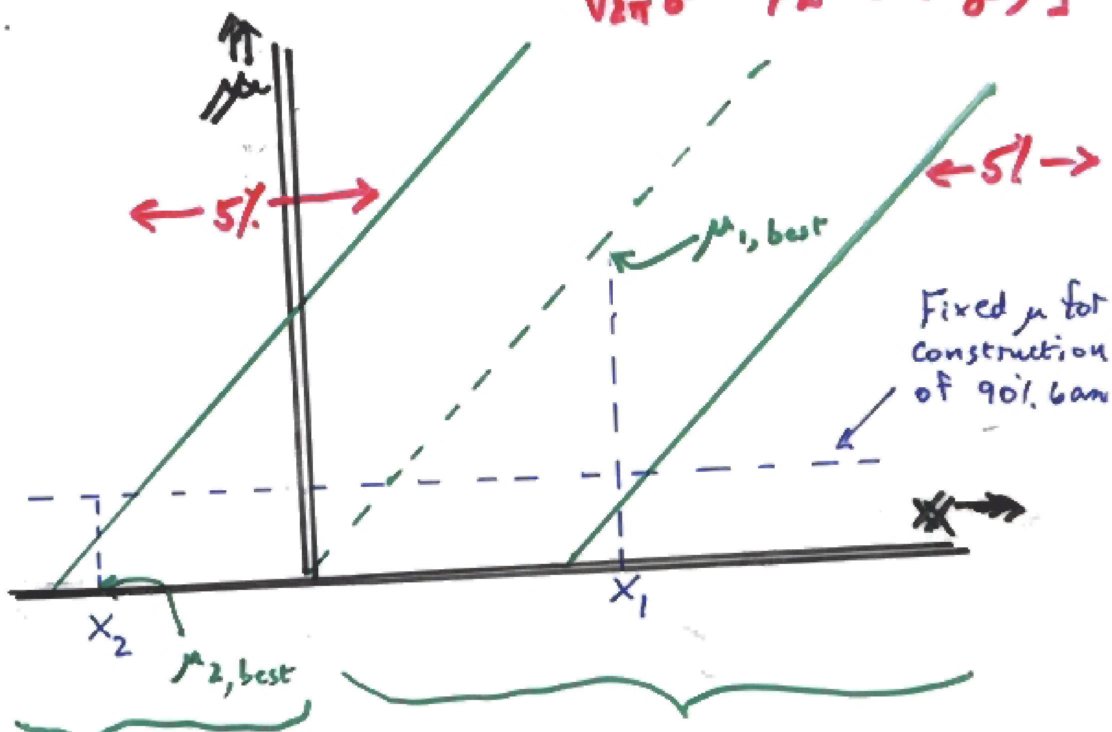


FELDMAN-COUSINS ORDERING RULE

$R = p(x,\mu)/p(x,\mu_{best})$ [Likelihood ratio ordering]

Gaussian example $p(x,\mu) = G(x,\mu,\sigma)$

$$= \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$



$\mu_{best} = 0$

$p(x,\mu_b) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-x^2}$

$p(x_1,\mu) > p(x_2,\mu)$

BUT $R(x_2,\mu) > R(x_1,\mu)$

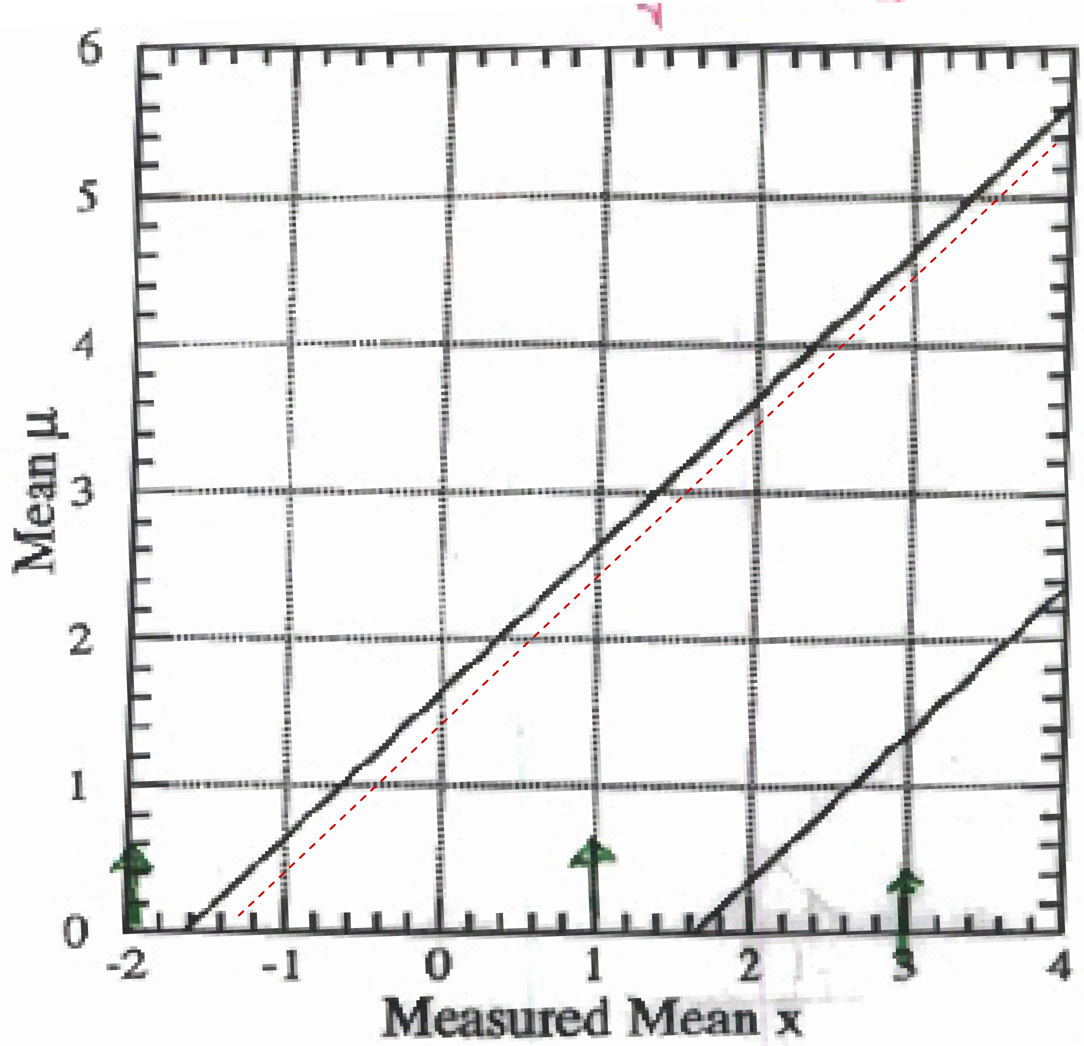
$\mu_{best} = x$

$p(x,\mu_b) = \frac{1}{\sqrt{2\pi}\,\sigma} = \text{const}$

Standard : Select x_1 before x_2

F-C : Select x_2 before x_1

Flip-flop



Black lines Classical 90% central interval

Red dashed: Classical 90% upper limit

FLIP – FLOP

90% upper limit for $x_{obs} \leq 3$

90% 2-sided interval for $x_{obs} > 3$

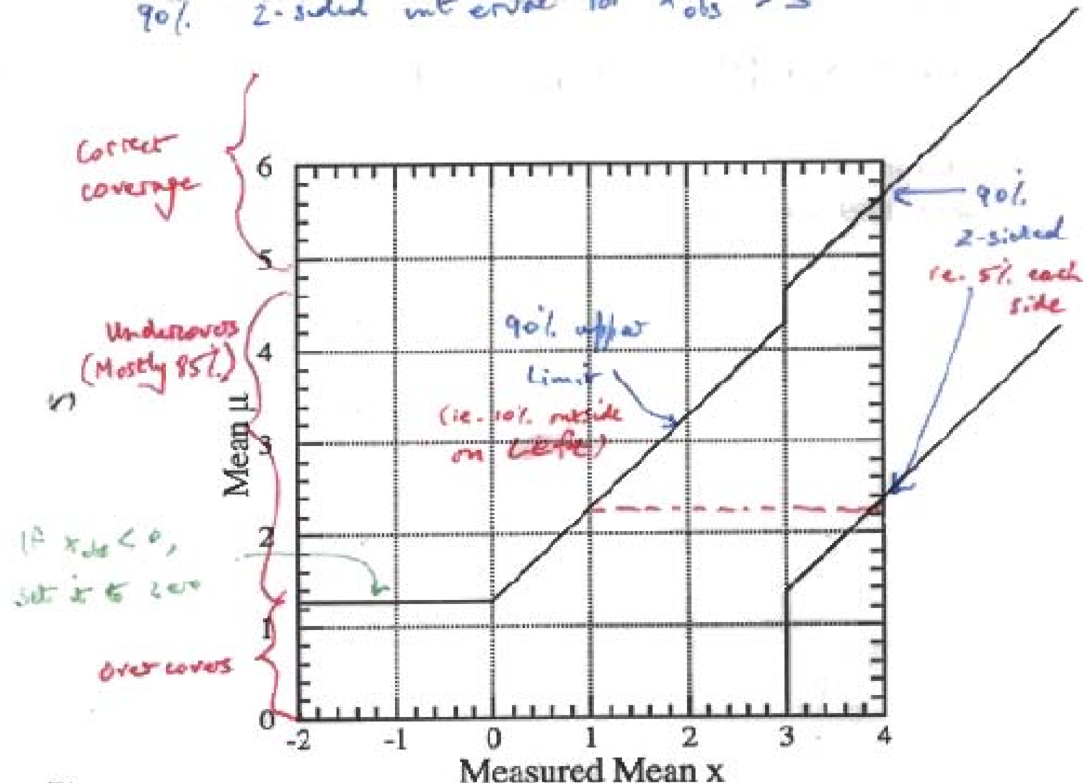


FIG. 4. Plot of confidence belts implicitly used for 90% C.L. confidence intervals (vertical intervals between the belts) quoted by flip-flopping Physicist X, described in the text. They are not valid confidence belts, since they can cover the true value at a frequency less than the stated confidence 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 goes smoothly from 1-sided → 2-sided

Poisson confidence intervals.   Background = 3

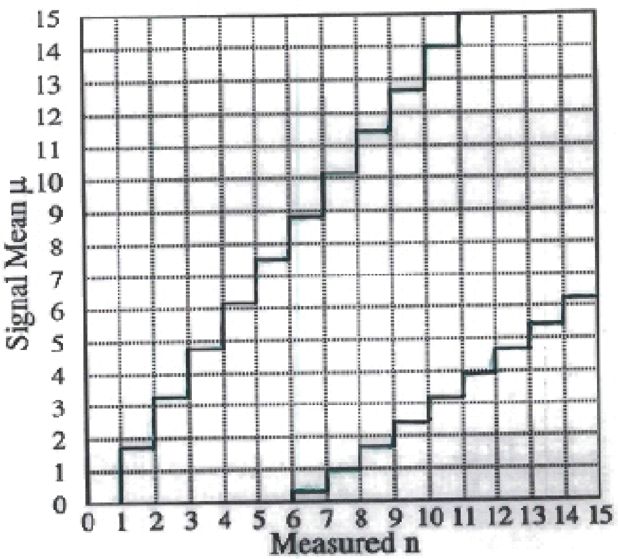


FIG. 6. Standard confidence belt for 90% C.L. central confidence intervals, for unknown Poisson signal mean μ in the presence of Poisson background with known mean $b = 3.0$.

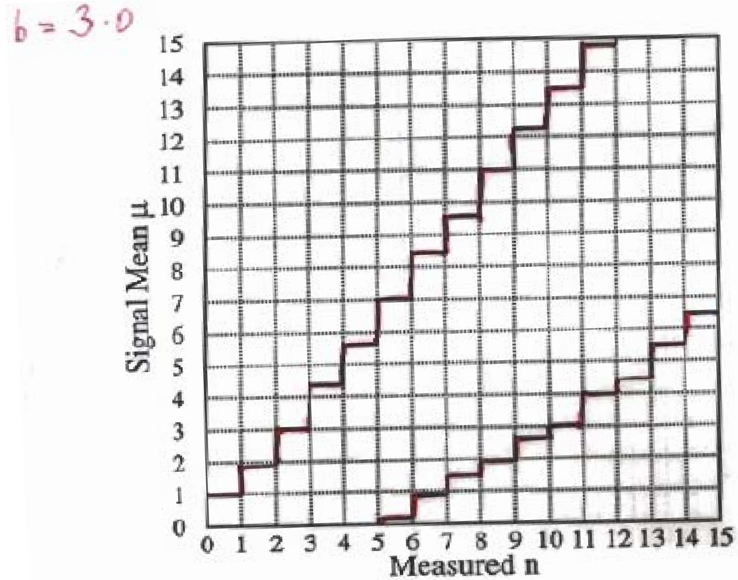


FIG. 7. Confidence belt based on our ordering principle, for 90% C.L. confidence intervals for unknown Poisson signal mean μ in the presence of Poisson background with known mean $b = 3.0$.

Standard Frequentist

Feldman - Cousins

FREQUENTIST POISSON C.B. CONSTRN.

< 10%

< 5%

Prob ordering

TABLES

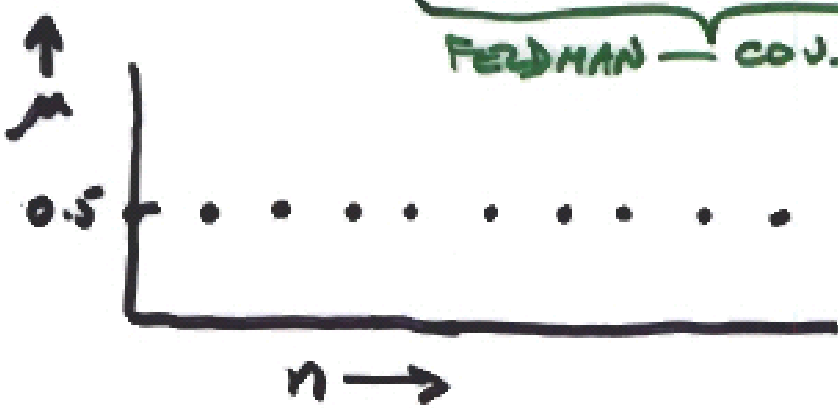
TABLE I. Illustrative calculations in the confidence belt construction for signal mean μ in the presence of known mean background $b = 3.0$. Here we find the acceptance interval for $\mu = 0.5$.

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%

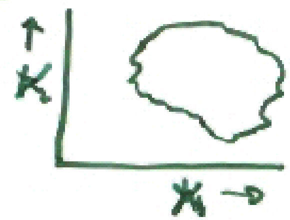


FEATURES OF F+C

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"

MAY PROB DENSITY
5% AT ENDS ?

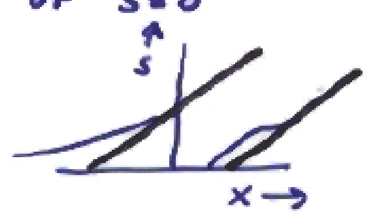
NEYMAN CONSTRUCTION ⇒ CPU-INTENSIVE
(ESP IN SEVERAL DIMENSIONS)

MINOR PATHOLOGIES : DISTANT INTERVALS
WRONG BEHAVIOUR WRT BGD

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

n_{obs}	b_{gd}	90% Limit
0	3.0	1.08
0	0	2.44

UNIFIED ⇒ QUICKER EXCLUSION OF s=0



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

Profile Likelihood

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

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

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

$$\uparrow$$

prior

With systematics

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

$$\uparrow$$

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

Then integrate over LA and b

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

31

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

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 $\sigma_{\max}$

BAYES 90% UPPER LIMITS

Bgd	$\epsilon = 1.0 \pm 0.1$: 0	$\epsilon = 1.0 \pm 0.1$: 3	$\epsilon = 1$ exactly: 0	$\epsilon = 1$ exactly: 3
n_{obs}				
0	2.35 indep of b		2.30 indep of b	
1	3.99	2.90	3.89	2.84
2	5.47	3.60	5.32	3.52
3	6.87	4.46	6.68	4.36
4	8.24	5.48	7.99	5.34
⋮	⋮	⋮	⋮	⋮
20	28.3	25.04	27.05	24.04

Less than 10% bigger than for $\epsilon = 1$ exactly

$\Delta = 0$ for $b = 0$

$\Delta = 3$ for large b

$\sim n + \kappa\sqrt{n}$

Joel Heinrich et al (CDF St. Cttee)

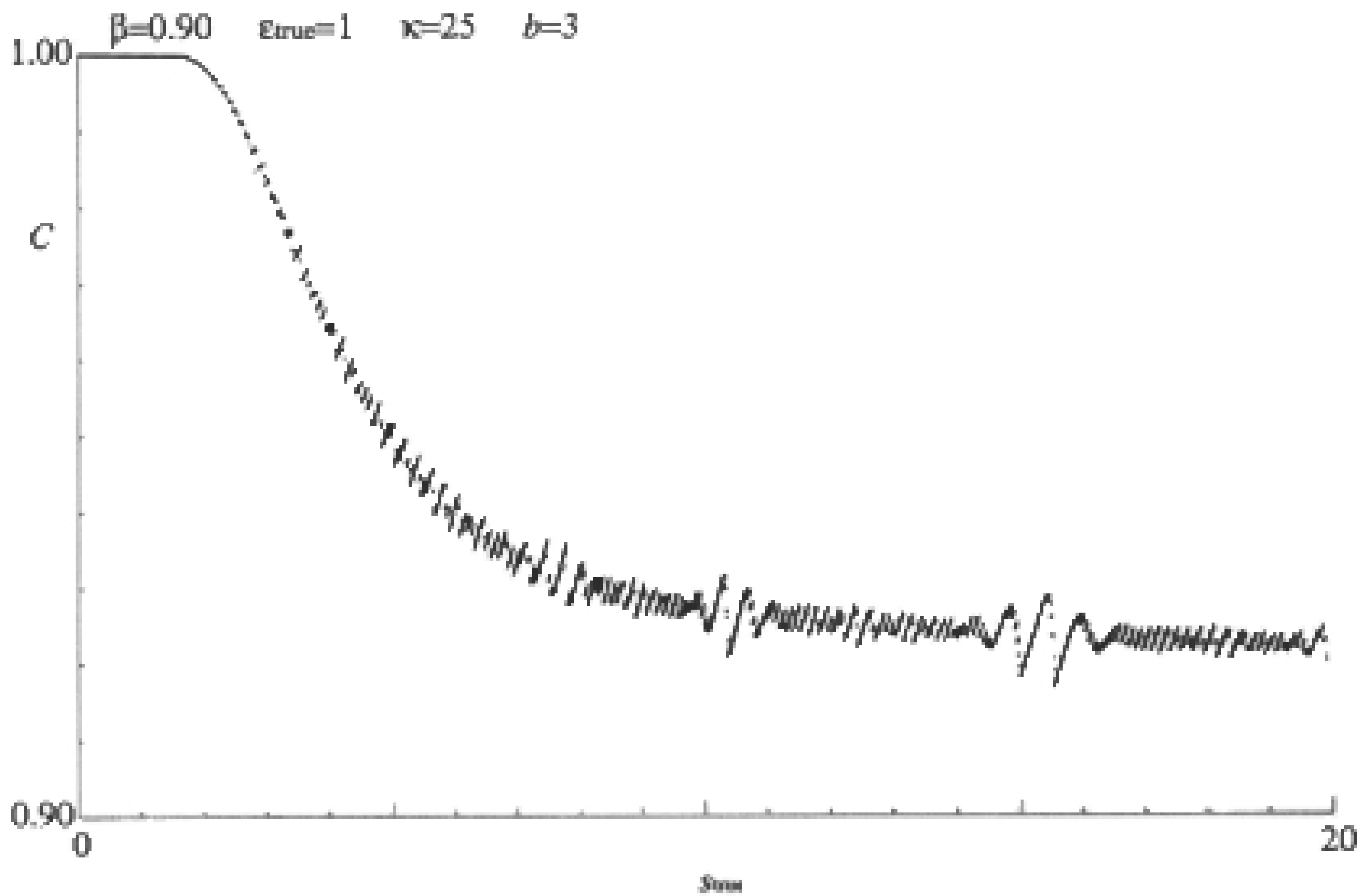


Figure 6: Coverage of 90% upper limits as a function of s_{true} for $\epsilon_{\text{true}} = 1$, nominal 20% uncertainty of the subsidiary measurement of ϵ, and $b = 3$ background expected.

Frequentist

Full Method

Imagine just 2 parameters σ and LA

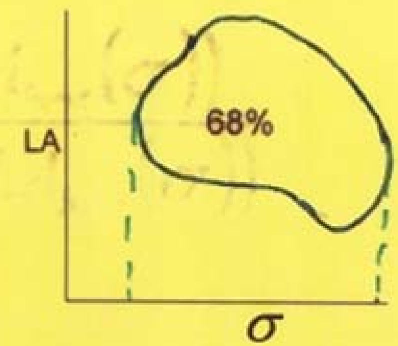
and 2 measurements N and M

N ↑ Physics, M ↑ Nuisance

Do Neyman construction in 4-D

Use observed **N** and **M**, to give

Confidence Region



33

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

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

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

Contrast Bayes marginalisation

Distinguish "profile ordering"

35

Properties being studied by Giovanni Punzi

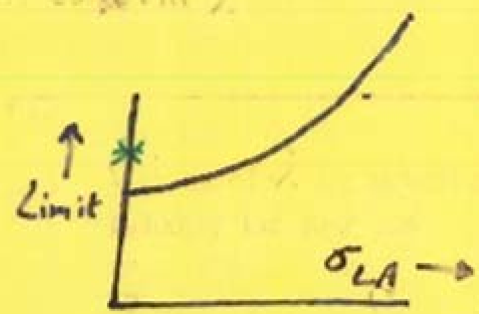
Method: Mixed Frequentist - Bayesian

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)

Coverage studied by Tegenfeldt + Conrad

PROFILE $\mathcal{L}$

Rolke, Lopez, Conrad + James

"Limits + Confidence Intervals in the presence of Nuisance Parameters"

$$p\mathcal{L}(\mu|\text{data}) = \mathcal{L}(\mu, b_{best}|\text{data})$$

$$\Delta \ln p\mathcal{L} = 0.5$$

Coverage much smoother (as fn of μ) than for standard Bayesian without nuisance parameters

Recommendations?

CDF note 7739 (May 2005)

Decide method in advance

No valid method is ruled out

Bayes is simplest for incorporating nuisance params

- Check robustness
- Quote coverage

Quote sensitivity

Use same method as other similar expts

Explain method used

Significance

Significance = $S/\sqrt{B}$?

Potential Problems:

•Uncertainty in B

•Non-Gaussian behaviour of Poisson, especially in tail

•Number of bins in histogram, no. of other histograms [FDR]

•Choice of cuts (Blind analyses)

•Choice of bins (………………….)

For future experiments:

• Optimising $S/\sqrt{B}$ could give S =0.1, B = 10^{-4}

Look Elsewhere Effect

See ‘peak’ in bin of histogram

p-value is chance of fluctuation at least as significant as observed under null hypothesis

1) at the position observed in the data; or
2) anywhere in that histogram; or
3) including other relevant histograms for your analysis; or
4) including other analyses in Collaboration; or
5) anywhere in HEP.

Goodness of Fit Tests

Data = individual points, histogram, multi-dimensional,
multi-channel

χ^2 and number of degrees of freedom

$\Delta\chi^2$ (or *ln*$\mathcal{L}$-ratio): Looking for a peak

Unbinned $\mathcal{L}_{max}$?

Kolmogorov-Smirnov

Zech energy test

Combining p-values

Lots of different methods. Software available from:

http://www.ge.infn.it/statisticaltoolkit

χ^2 with ν degrees of freedom?

1) ν = data – free parameters ?

Why **asymptotic** (apart from Poisson → Gaussian) ?

a) Fit flatish histogram with

$y = N\ \{1 + 10^{-6} \exp\{-0.5(x-x_0)^2\}$ $\quad x_0$ = free param

b) Neutrino oscillations: almost **degenerate parameters**

$y \sim 1 - A \sin^2(1.27\ \Delta m^2\ L/E)$ $\qquad$ 2 parameters

$\xrightarrow[\text{Small }\Delta m^2]{} 1 - A\ (1.27\ \Delta m^2\ L/E)^2$ $\qquad$ 1 parameter

χ^2 with ν degrees of freedom?

2) Is difference in χ^2 distributed as χ^2 ?

H0 is true.

Also fit with H1 with k extra params

e. g. Look for Gaussian peak on top of smooth background

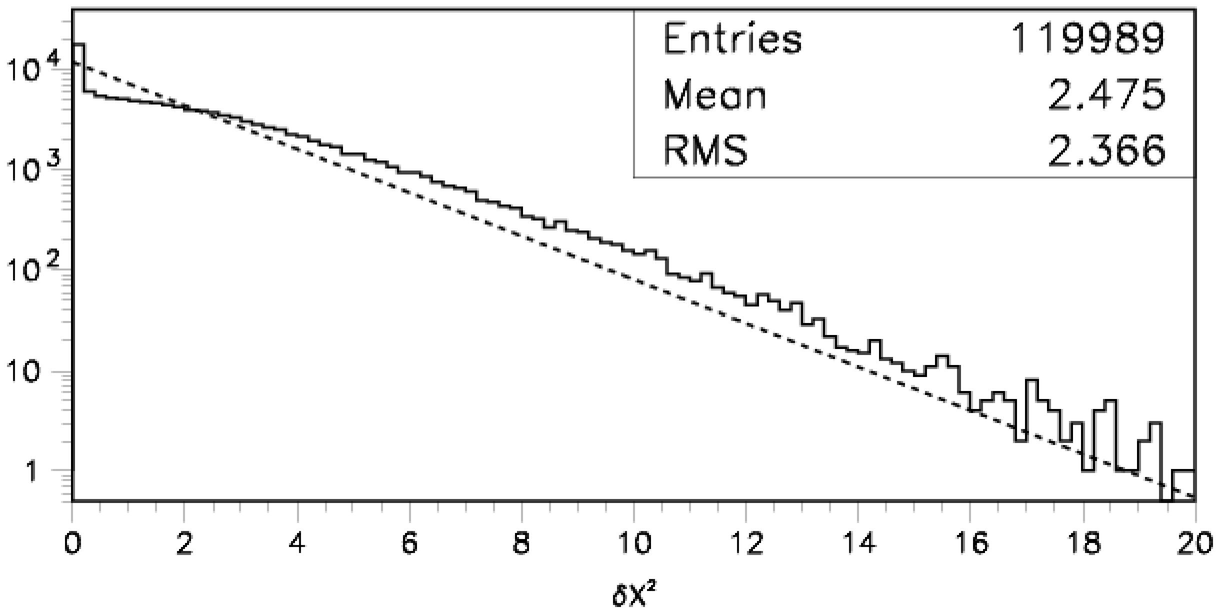
$$y = C(x) + A \exp\{-0.5 \, ((x-x_0)/\sigma)^2\}$$

Is $\chi^2_{H0} - \chi^2_{H1}$ distributed as χ^2 with $\nu = k = 3$?

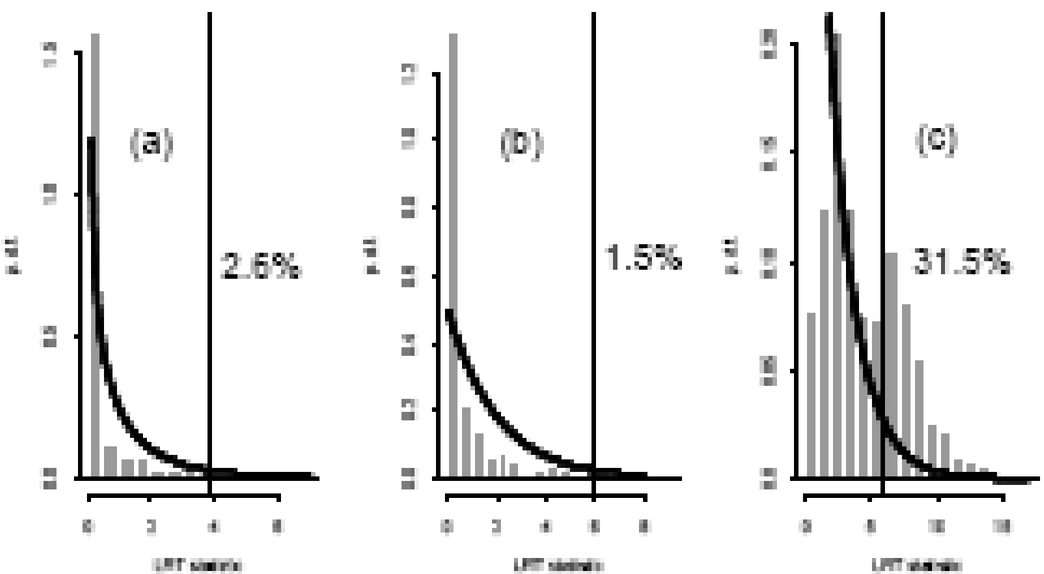
Relevant for assessing whether enhancement in data is just a statistical fluctuation, or something more interesting

N.B. Under H0 ($y = C(x)$) : $A=0$ (boundary of physical region)
x_0 and σ undefined

Is difference in χ^2 distributed as χ^2 ?



Demortier:
H0 = quadratic bgd
H1 = ……………… +
Gaussian of fixed width,
variable location & ampl



Protassov, van Dyk, Connors, ….
H0 = continuum
(a) H1 = narrow emission line
(b) H1 = wider emission line
(c) H1 = absorption line

Nominal significance level = 5%

Is difference in χ^2 distributed as χ^2 ?, contd.

So need to determine the $\Delta\chi^2$ distribution by Monte Carlo

N.B.

1) Determining $\Delta\chi^2$ for hypothesis H1 when data is generated according to H0 is not trivial, because there will be lots of local minima

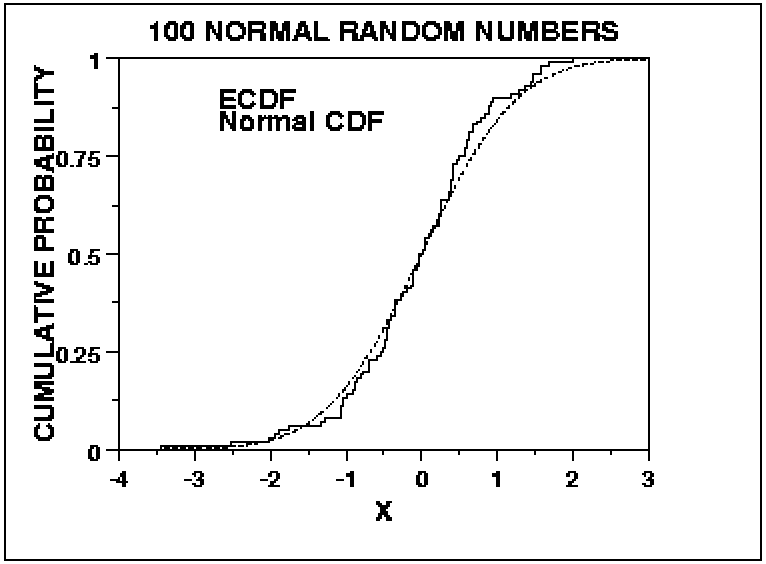
2) If we are interested in 5σ significance level, needs lots of MC simulations (or intelligent MC generation)

Goodness of Fit: Kolmogorov-Smirnov

Compares data and model cumulative plots

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)

Combining different p-values

Several results quote independent p-values for same effect:

p_1, p_2, p_3….. e.g. 0.9, 0.001, 0.3 ……..

What is combined significance? Not just $p_1 {*} p_2 {*} p_3$…..

If 10 expts each have p ~ 0.5, product ~ 0.001 and is clearly **NOT** correct combined p

$$S = z \; * \sum_{j=0}^{n-1} (-\ln z)^j / j! \; , \qquad z = p_1 p_2 p_3 \ldots\ldots$$

(e.g. For 2 measurements, $S = z * (1 - \ln z) \geq z$)

Slight problem: Formula is not associative

Combining {{p_1 and p_2}, and then p_3} gives different answer from {{p_3 and p_2}, and then p_1} , or all together

Due to different options for "more extreme than x_1, x_2, x_3".

Combining different p-values

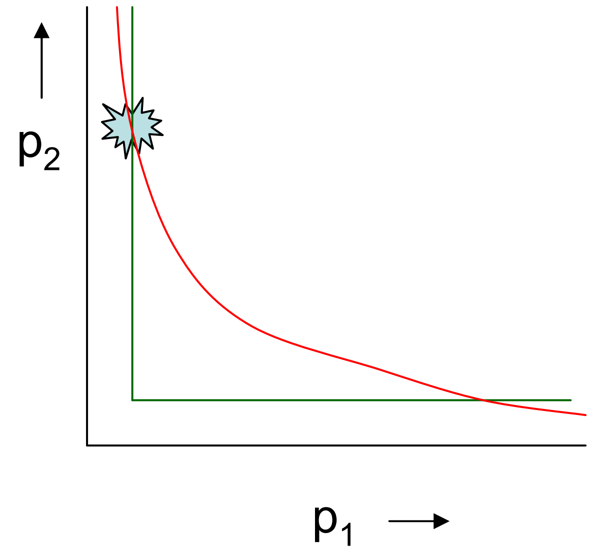
Conventional:

Are set of p-values consistent with H0?

SLEUTH:

How significant is smallest p?

$$1\text{-}S = (1-p_{smallest})^n$$



	$p_1 = 0.01$		$p_1 = 10^{-4}$	
	$p_2 = 0.01$	$p_2 = 1$	$p_2 = 10^{-4}$	$p_2 = 1$
Combined S				
Conventional	$1.0\ 10^{-3}$	$5.6\ 10^{-2}$	$1.9\ 10^{-7}$	$1.0\ 10^{-3}$
SLEUTH	$2.0\ 10^{-2}$	$2.0\ 10^{-2}$	$2.0\ 10^{-4}$	$2.0\ 10^{-4}$

Example of ambiguity

Combine two tests:

a) $\chi^2 = 80$ for $\boldsymbol{\nu} = 100$

b) $\chi^2 = 20$ for $\boldsymbol{\nu} = 1$

1) b) is just another similar test:

$\chi^2 = 100$ for $\boldsymbol{\nu} = 101$

ACCEPT

2) b) is very different test

p_1 is OK, but p_2 is very small. Combine p's

REJECT

Basic reason for ambiguity

Trying to transform uniform distribution in unit hypercube to uniform one dimensional distribution ($p_{comb} = 0 \rightarrow 1$)

Why 5σ?

- Past experience with 3σ, 4σ,… signals
- Look elsewhere effect:

  Different cuts to produce data

  Different bins (and binning) of this histogram

  Different distributions Collaboration did/could look at

  Defined in SLEUTH
- Bayesian priors:

$$\frac{P(H0|data)}{P(H1|data)} = \frac{P(data|H0) * P(H0)}{P(data|H1) * P(H1)}$$

Bayes posteriors   Likelihoods   Priors

Prior for {H0 = S.M.} >>> Prior for {H1 = New Physics}

Why 5σ?

BEWARE of tails,

especially for nuisance parameters

Same criterion for all searches?

Single top production

Higgs

Highly speculative particle

Energy non-conservation

BLIND ANALYSES

Why blind analysis?
Selections, corrections, method

Methods of blinding

Add random number to result *
Study procedure with simulation only
Look at only first fraction of data
Keep the signal box closed
Keep MC parameters hidden
Keep unknown fraction visible for each bin

After analysis is unblinded, ……..

* Luis Alvarez suggestion re "discovery" of free quarks

p-value is not ……..

Does **NOT** measure Prob(H0 is true)

i.e. It is **NOT** P(H0|data)

It is P(data|H0)

N.B. P(H0|data) ≠ P(data|H0)

P(theory|data) ≠ P(data|theory)

"Of all results with p ≤ 5%, half will turn out to be wrong"

N.B. Nothing wrong with this statement

e.g. 1000 tests of energy conservation

~50 should have p ≤ 5%, and so reject H0 = energy conservation

Of these 50 results, all are likely to be "wrong"

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

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%

More and more data

1) Eventually p(data|H0) will be small, even if data and H0 are very similar.

p-value does not tell you how different they are.

2) Also, beware of multiple (yearly?) looks at data.

"Repeated tests eventually sure to reject H0, independent of value of α"

Probably not too serious – < ~10 times per experiment.

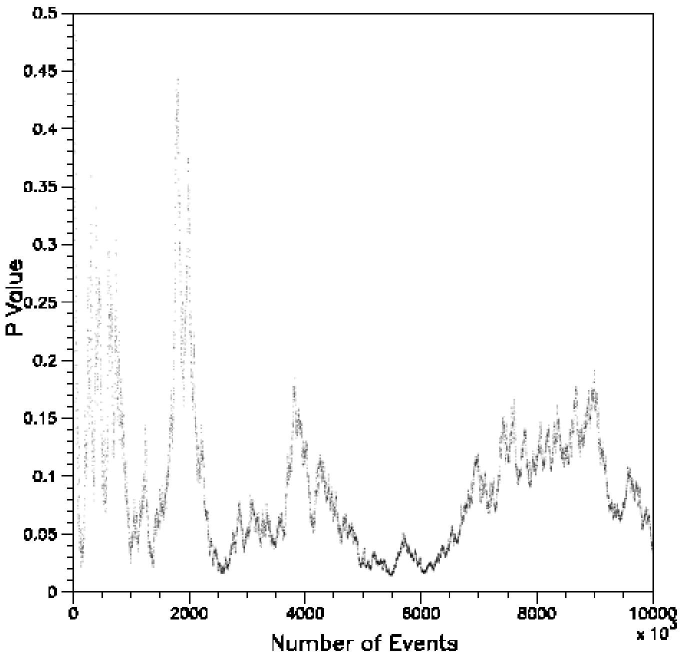


Figure 1: P value versus sample size.

Choosing between 2 hypotheses

Possible methods:

$\Delta\chi^2$

p-value of statistic $\rightarrow$

$ln\mathcal{L}$–ratio

Bayesian:

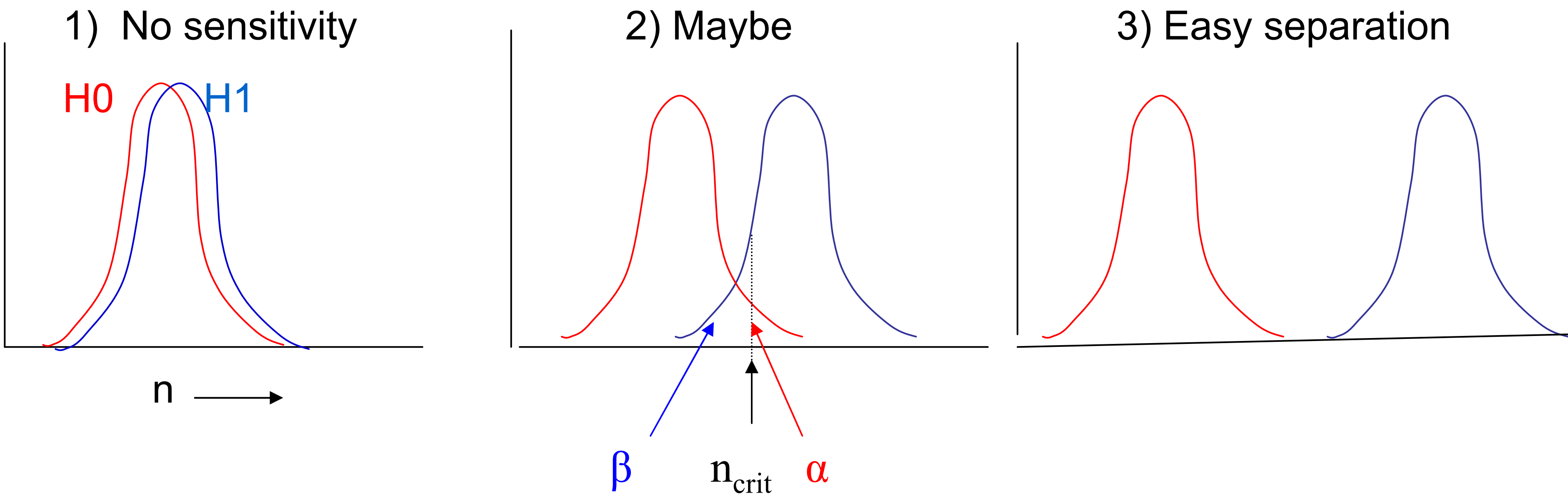
Posterior odds

Bayes factor

Bayes information criterion (BIC)

Akaike …….. (AIC)

Minimise "cost"



Procedure: Choose α (e.g. 95%, 3σ, 5σ ?) and CL for β (e.g. 95%)

Given b, α determines n_{crit}

s defines β. For $s > s_{min}$, separation of curves → discovery or excln

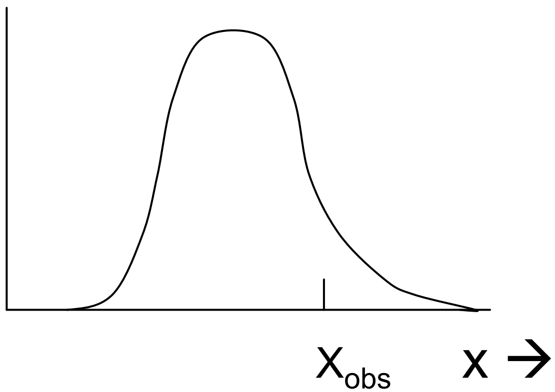
s_{min} = Punzi measure of sensitivity For $s \geq s_{min}$, 95% chance of 5σ discovery

Optimise cuts for smallest s_{min}

Now data: If $n_{obs} \geq n_{crit}$, discovery at level α

If $n_{obs} < n_{crit}$, no discovery. If $\beta_{obs} < 1 - CL$, exclude H1

p-values or $\mathcal{L}$ikelihood ratio?



$\mathcal{L}$ = height of curve

p = tail area

Different for distributions that

a) have dip in middle

b) are flat over range

Likelihood ratio favoured by Neyman-Pearson lemma (for simple H0, H1)

Use $\mathcal{L}$-ratio as statistic, and use p-values for its distributions for H0 and H1

Think of this as either

i) p-value method, with $\mathcal{L}$-ratio as statistic; or

ii) $\mathcal{L}$-ratio method, with p-values as method to assess value of $\mathcal{L}$-ratio

Bayes' methods for H0 versus H1

Bayes' Th: $P(A|B) = P(B|A) * P(A) / P(B)$

$$\frac{P(H0|data)}{P(H1|data)} = \frac{P(data|H0) * Prior(H0)}{P(data|H1) * Prior(H1)}$$

Posterior odds ratio

Likelihood ratio

Priors

N.B. Frequentists object to this

(and some Bayesians object to p-values)

Bayes' methods for H0 versus H1

$$\frac{P(H0|data)}{P(H1|data)} = \frac{P(data|H0) * Prior(H0)}{P(data|H1) * Prior(H1)}$$

Posterior odds   Likelihood ratio   Priors

e.g. data is mass histogram
H0 = smooth background
H1 = ……………………… + peak

1) Profile likelihood ratio also used but not quite Bayesian
(Profile = **maximise** wrt parameters.
Contrast Bayes which **integrates** wrt parameters)
2) Posterior odds
3) Bayes factor = Posterior odds/Prior ratio
(= Likelihood ratio in simple case)
4) In presence of parameters, need to integrate them out, using priors.
e.g. peak's mass, width, amplitude
Result becomes dependent on prior, and more so than in parameter determination.
5) Bayes information criterion (BIC) tries to avoid priors by
$BIC = -2 * \ln\{\mathcal{L} \text{ ratio}\} + k * \ln\{n\}$ k= free params; n=no. of obs
6) Akaike information criterion (AIC) tries to avoid priors by
$AIC = -2 * \ln\{\mathbf{L} \text{ ratio}\} + 2k$

etc etc etc

Why p ≠ Bayes factor

Measure different things:

p_0 refers just to H0; B_{01} compares H0 and H1

Depends on amount of data:

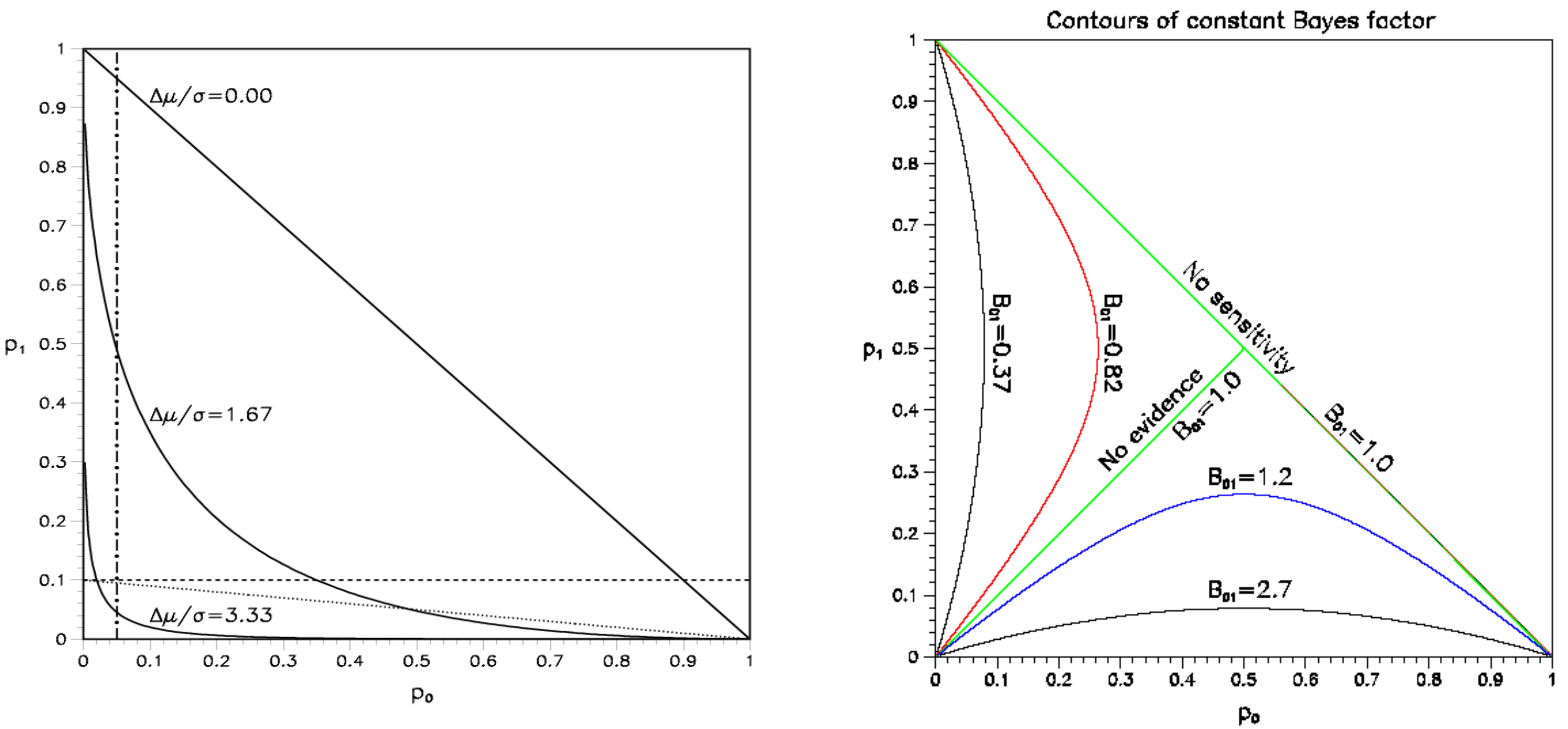
e.g. Poisson counting expt little data:

For H0, $\mu_0 = 1.0$. For H1, $\mu_1 = 10.0$

Observe $n = 10$ $p_0 \sim 10^{-7}$ $B_{01} \sim 10^{-5}$

Now with 100 times as much data, $\mu_0 = 100.0$ $\mu_1 = 1000.0$

Observe $n = 160$ $p_0 \sim 10^{-7}$ $B_{01} \sim 10^{+14}$



p_0 versus p_1 plots

Optimisation for Discovery and Exclusion

Giovanni Punzi, PHYSTAT2003:

"Sensitivity for searches for new signals and its optimisation"

http://www.slac.stanford.edu/econf/C030908/proceedings.html

Simplest situation: Poisson counting experiment,

Bgd = b, Possible signal = s, n_{obs} counts

(More complex: Multivariate data, $ln\mathcal{L}$-ratio)

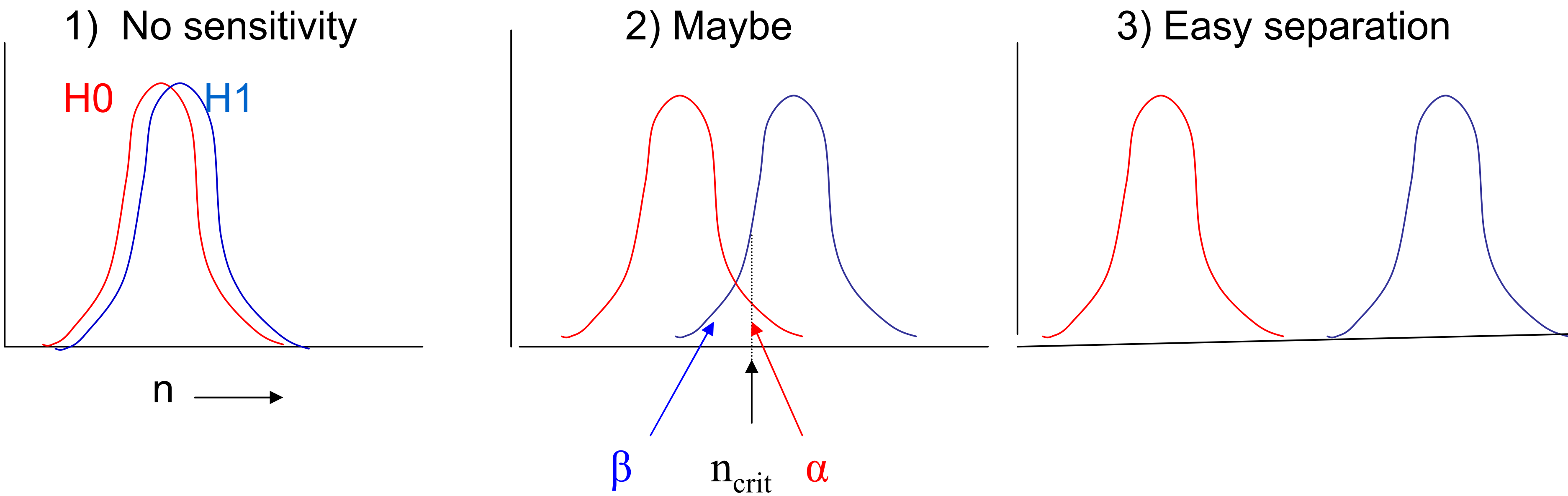
Traditional sensitivity:

Median limit when s=0

Median σ when $s \neq 0$ (averaged over s?)

Punzi criticism: Not most useful criteria

Separate optimisations



Procedure: Choose α (e.g. 95%, 3σ, 5σ ?) and CL for β (e.g. 95%)

Given b, α determines n_{crit}

s defines β. For $s > s_{min}$, separation of curves → discovery or excln

s_{min} = Punzi measure of sensitivity For $s \geq s_{min}$, 95% chance of 5σ discovery

Optimise cuts for smallest s_{min}

Now data: If $n_{obs} \geq n_{crit}$, discovery at level α

If $n_{obs} < n_{crit}$, no discovery. If $\beta_{obs} < 1 - CL$, exclude H1

1) No sensitivity

Data almost always falls in peak

β as large as 5%, so 5% chance of H1 exclusion even when no sensitivity. (CL_s)

2) Maybe

If data fall above n_{crit}, discovery

Otherwise, and $n_{obs} \rightarrow \beta_{obs}$ small, exclude H1

(95% exclusion is easier than 5σ discovery)

But these may not happen → no decision

3) Easy separation

Always gives discovery or exclusion (or both!)

Disc	Excl	1)	2)	3)
No	No	□	□	
No	Yes		□	□
Yes	No		(□)	□
Yes	Yes			□!

Incorporating systematics in p-values

Simplest version:

Observe n events

Poisson expectation for background only is $b \pm \sigma_b$

σ_b may come from:

acceptance problems

jet energy scale

detector alignment

limited MC or data statistics for backgrounds

theoretical uncertainties

Luc Demortier,“p-values: What they are and how we use them”, CDF memo June 2006

http://www-cdfd.fnal.gov/~luc/statistics/cdf0000.ps

Includes discussion of several ways of incorporating nuisance parameters

Desiderata:

- Uniformity of p-value (averaged over ν, or for each ν?)
- p-value increases as σ_ν increases
- Generality
- Maintains power for discovery

Ways to incorporate nuisance params in p-values

- Supremum — Maximise p over all ν. Very conservative
- Conditioning — Good, if applicable
- Prior Predictive — Box. Most common in HEP

  $p = \int p(\nu)\, \pi(\nu)\, d\nu$
- Posterior predictive — Averages p over posterior
- Plug-in — Uses best estimate of ν, without error
- $\mathcal{L}$-ratio
- Confidence interval — Berger and Boos.

  $p = \mathrm{Sup}\{p(\nu)\} + \beta$, where $1-\beta$ Conf Int for ν
- Generalised frequentist — Generalised test statistic

Performances compared by Demortier

Summary

- P(H0|data) ≠ P(data|H0)
- p-value is NOT probability of hypothesis, given data
- Many different Goodness of Fit tests

  Most need MC for statistic → p-value
- For comparing hypotheses, $\Delta\chi^2$ is better than χ^2_1 and χ^2_2
- Blind analysis avoids personal choice issues
- Different definitions of sensitivity
- Worry about systematics

PHYSTAT-LHC Workshop at CERN, June 2007

“Statistical issues for LHC Physics Analyses”

Proceedings at http://phystat-lhc.web.cern.ch/phystat-lhc/2008-001.pdf