Statistical Issues in Searches for New Physics

Louis Lyons
Imperial College, London
and
Oxford

CERN
Jan, 2015

Theme:

Using data to make judgements about H1 (New Physics) versus
H0 (S.M. with nothing new)

Why?

Experiments are expensive and time-consuming
so
Worth investing effort in statistical analysis
→ better information from data

Topics:

- Blind Analysis
- Why 5σ for discovery?
- Significance
- P(A|B) ≠ P(B|A)
- Meaning of p-values
- Wilks' Theorem
- LEE = Look Elsewhere Effect
- Background Systematics
- Coverage
- Example of misleading inference
- p_0 v p_1 plots

(N.B. Several of these topics have no unique solutions from Statisticians)

Conclusions

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

Why 5σ for Discovery?

Statisticians ridicule our belief in extreme tails (esp. for systematics)

Our reasons:

1) Past history (Many 3σ and 4σ effects have gone away)

2) LEE (see later)

3) Worries about underestimated systematics

4) Subconscious Bayes calculation

$$\frac{p(H_1|x)}{p(H_0|x)} = \frac{p(x|H_1)}{p(x|H_0)} * \frac{\pi(H_1)}{\pi(H_0)}$$

Posterior prob   Likelihood ratio   Priors

"Extraordinary claims require extraordinary evidence"

N.B. Points 2), 3) and 4) are experiment-dependent

Alternative suggestion:

L.L. "Discovering the significance of 5σ"   http://arxiv.org/abs/1310.1284

How many σ's for discovery?

SEARCH	SURPRISE	IMPACT	LEE	SYSTEMATICS	No. σ
Higgs search	Medium	Very high	M	Medium	5
Single top	No	Low	No	No	3
SUSY	Yes	Very high	Very large	Yes	7
B_s oscillations	Medium/Low	Medium	Δm	No	4
Neutrino osc	Medium	High	$\sin^2 2\vartheta$, Δm^2	No	4
$B_s \rightarrow \mu\,\mu$	No	Low/Medium	No	Medium	3
Pentaquark	Yes	High/V. high	M, decay mode	Medium	7
$(g\text{-}2)_\mu$ anom	Yes	High	No	Yes	4
H spin ≠ 0	Yes	High	No	Medium	5
4^{th} gen q, l, ν	Yes	High	M, mode	No	6
Dark energy	Yes	Very high	Strength	Yes	5
Grav Waves	No	High	Enormous	Yes	8

Suggestions to provoke discussion, rather than `delivered on Mt. Sinai'

Bob Cousins: "2 independent expts each with 3.5σ better than one expt with 5σ"

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 [LEE]

•Choice of cuts (Blind analyses)

•Choice of bins (……………….)

For future experiments:

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

$P(A|B) \neq P(B|A)$

Remind Lab or University media contact person that:

Prob[data, given H0] is very small

does not imply that

Prob[H0, given data] is also very small.

e.g. Prob{data | speed of $\nu \leq c$}= very small

does not imply

Prob{speed of $\nu \leq c$ | data} = very small

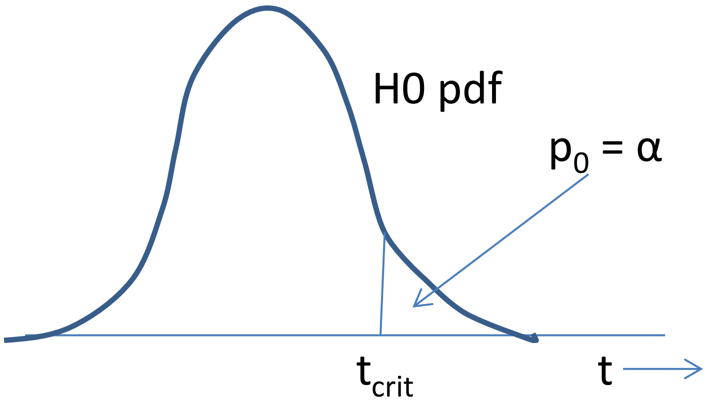
or Prob{speed of $\nu > c$ | data} ~ 1

Everyday example: pack of playing cards

p(spades|king) = 1/4

p(king|spades) = 1/13

What p-values are (and are not)



Reject H0 if $t > t_{crit}$ ($p < \alpha$)

p-value = prob that $t \geq t_{obs}$

Small p → data and theory have poor compatibility

Small p-value does **NOT** automatically imply that theory is unlikely

Bayes prob(Theory|data) related to prob(data|Theory) = Likelihood

by Bayes Th, including Bayesian prior

p-values are misunderstood. e.g. Anti-HEP jibe:

"Particle Physicists don't know what they are doing, because half their

$p < 0.05$ exclusions turn out to be wrong"

Demonstrates lack of understanding of p-values

[**All** results rejecting energy conservation with $p < \alpha = .05$ cut will turn out to be 'wrong']

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".

******* Better to combine data ************

Wilks' Theorem

Data = some distribution e.g. mass histogram

For H0 and H1, calculate best fit weighted sum of squares S_0 and S_1

Examples: 1) H0 = polynomial of degree 3

H1 = polynomial of degree 5

2) H0 = background only

H1 = bgd + peak with free M_0 and cross-section

3) H0 = normal neutrino hierarchy

H1 = inverted hierarchy

If H0 true, S_0 distributed as χ^2 with ndf = ν_0

If H1 true, S_1 distributed as χ^2 with ndf = ν_1

If H0 true, what is distribution of $\Delta S = S_0 - S_1$? Is it χ^2?

Wilks' Theorem: ΔS distributed as χ^2 with ndf = $\nu_1 - \nu_0$ provided:

a) H0 is true
b) H0 and H1 are nested
c) Params for H1→ H0 are well defined, and not on boundary
d) Data is asymptotic

Wilks' Theorem, contd

Examples: Does Wilks' Th apply?

1) H0 = polynomial of degree 3
   H1 = polynomial of degree 5

YES: ΔS distributed as χ^2 with ndf = (d-4) – (d-6) = 2

2) H0 = background only
   H1 = bgd + peak with free M_0 and cross-section

NO: H0 and H1 nested, but M_0 undefined when H1→ H0. $\Delta S \neq \chi^2$
(but not too serious for fixed M)

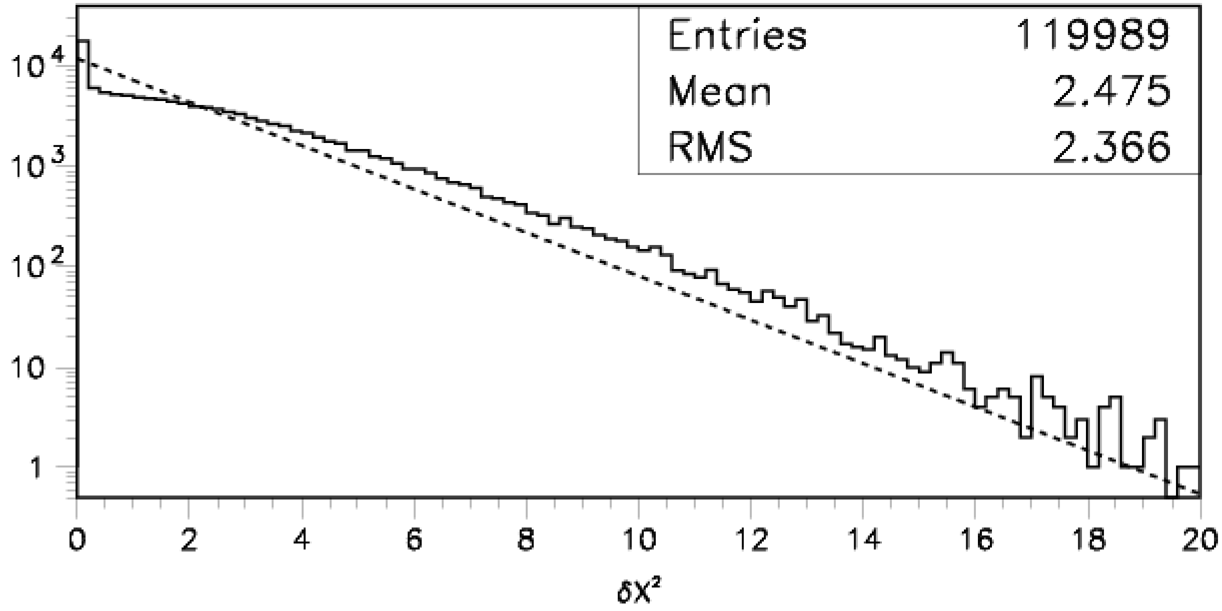
3) H0 = normal neutrino hierarchy
   H1 = inverted hierarchy

NO: Not nested. $\Delta S \neq \chi^2$ (e.g. can have $\Delta\chi^2$ negative)

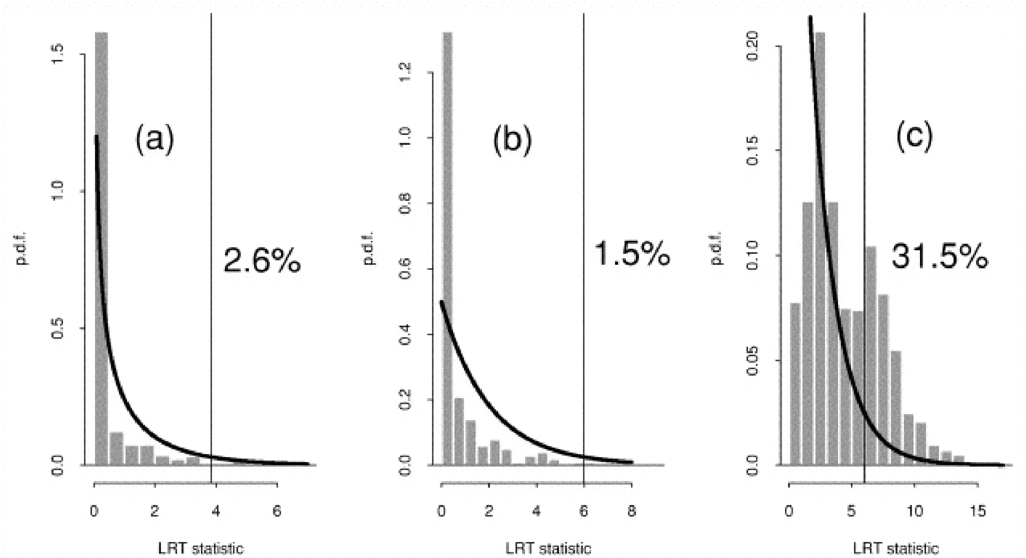
N.B. 1: Even when W. Th. does not apply, it does not mean that ΔS is irrelevant, but you cannot use W. Th. for its expected distribution.

N.B. 2: For large ndf, better to use ΔS, rather than S_1 and S_0 separately

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

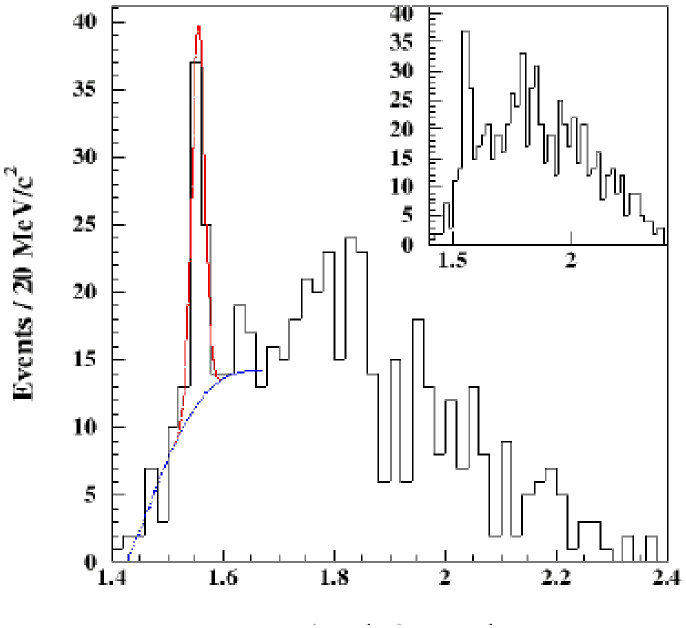
N.B.

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)

3) Asymptotic formulae may be useful (see K. Cranmer, G. Cowan, E. Gross and O. Vitells, 'Asymptotic formulae for likelihood-based tests of new physics', http://link.springer.com/article/10.1140%2Fepjc%2Fs10052-011-1554-0)

Look Elsewhere Effect (LEE)



Prob of bgd fluctuation at that place = local p-value
Prob of bgd fluctuation 'anywhere' = global p-value
Global p > Local p
Where is `anywhere'?

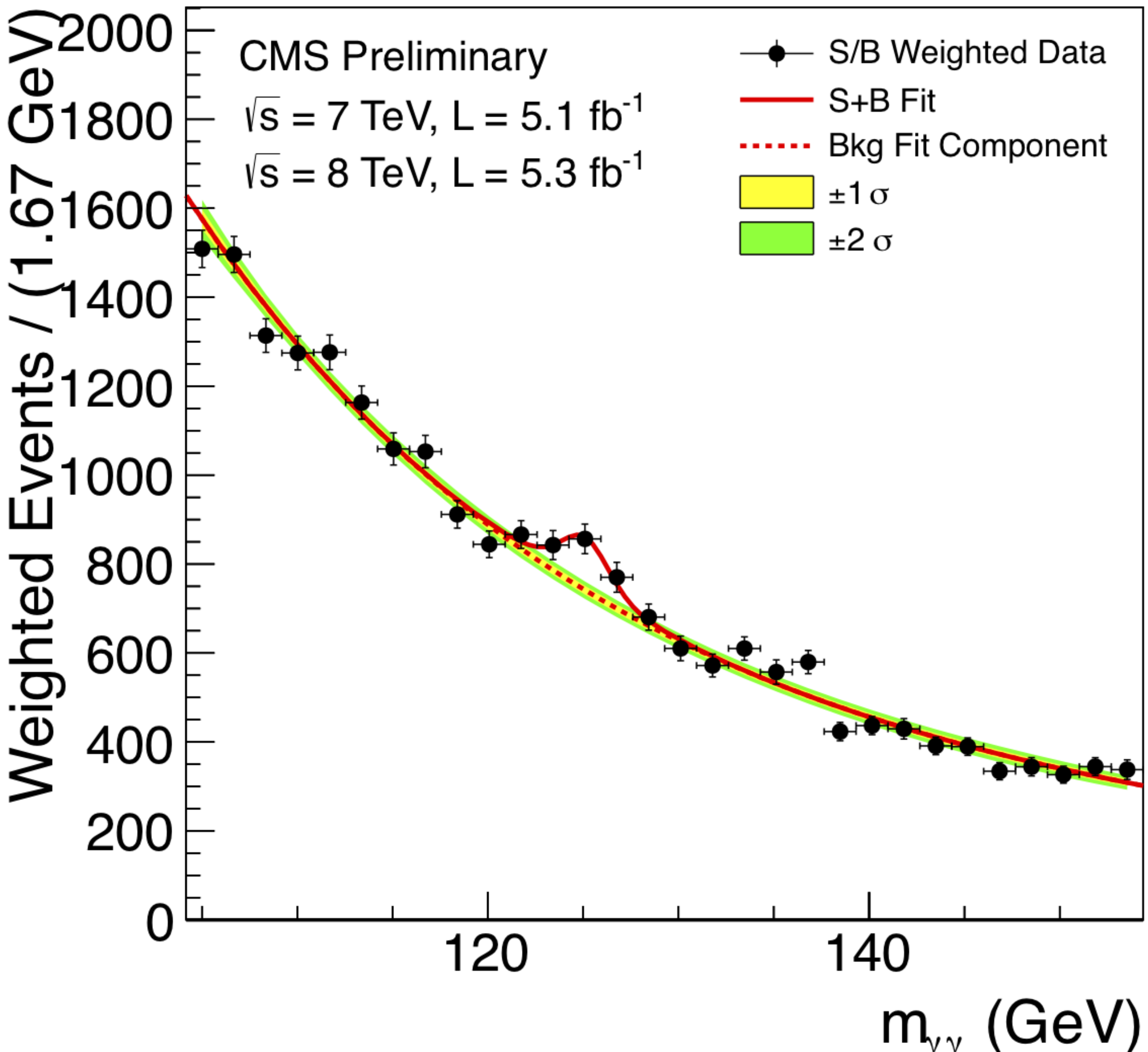
a) Any location in this histogram in sensible range
b) Any location in this histogram
c) Also in histogram produced with different cuts, binning, etc.
d) Also in other plausible histograms for this analysis
e) Also in other searches in this PHYSICS group (e.g. SUSY at CMS)
f) In any search in this experiment (e.g. CMS)
g) In all CERN expts (e.g. LHC expts + NA62 + OPERA + ASACUSA + ….)
h) In all HEP expts

etc.

d) relevant for graduate student doing analysis
f) relevant for experiment's Spokesperson

INFORMAL CONSENSUS:
Quote local p, and global p according to a) above.
Explain which global p

Background systematics



Background systematics, contd

Signif from comparing χ^2's for H0 (bgd only) and for H1 (bgd + signal)
Typically, bgd = functional form f_a with free params
e.g. 4^{th} order polynomial
Uncertainties in params included in signif calculation
But what if functional form is different ? e.g. f_b
Typical approach:
If f_b best fit is bad, not relevant for systematics
If f_b best fit is ~comparable to f_a fit, include contribution to systematics
But what is '~comparable'?
Other approaches:
Profile likelihood over different bgd parametric forms
http://arxiv.org/pdf/1408.6865v1.pdf?
Background subtraction
sPlots
Non-parametric background
Bayes
etc

No common consensus yet among experiments on best approach
{Spectra with multiple peaks are more difficult}

"Handling uncertainties in background shapes: the discrete profiling method"

Dauncey, Kenzie, Wardle and Davies (Imperial College, CMS)

arXiv:1408.6865v1 [physics.data-an]

Has been used in CMS analysis of H→$\gamma\gamma$

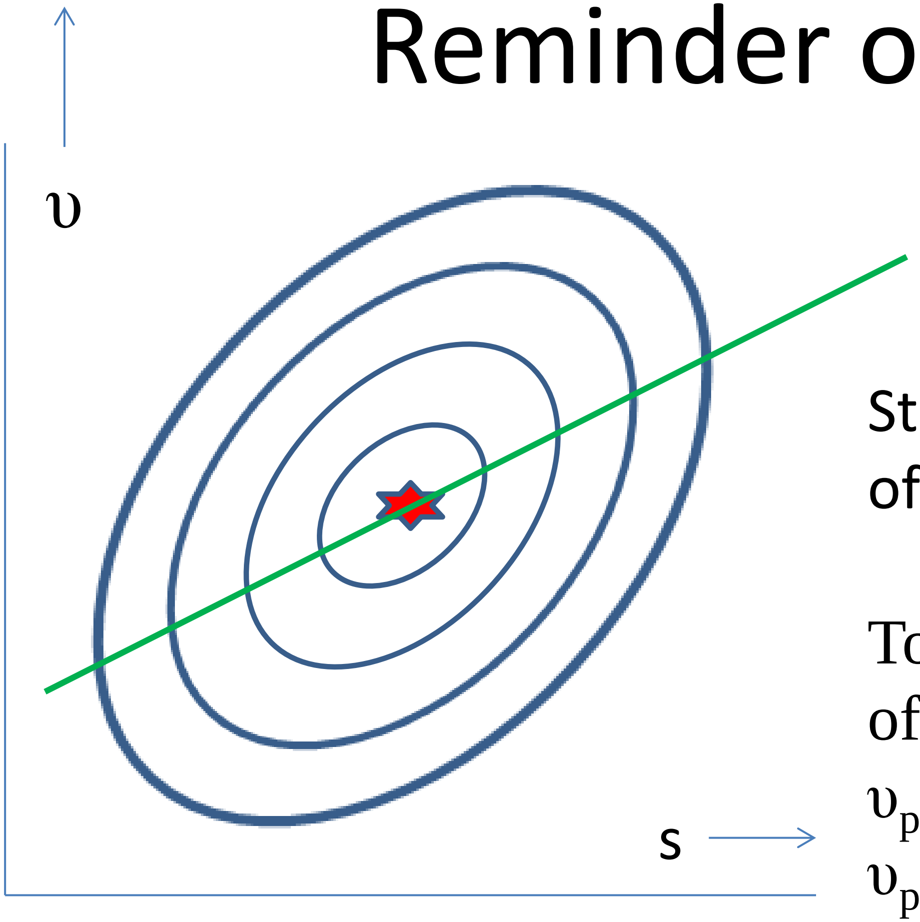
EPJC doi:10.1140/epjc/s10052-014-3076-z

Problem with 'Typical approach': Alternative functional forms do or don't contribute to systematics by hard cut, so systematics can change discontinuously wrt $\Delta\chi^2$

Method is like profile $\mathcal{L}$ for continuous nuisance params.

Here 'profile' over discrete functional forms

Reminder of Profile $\mathcal{L}$

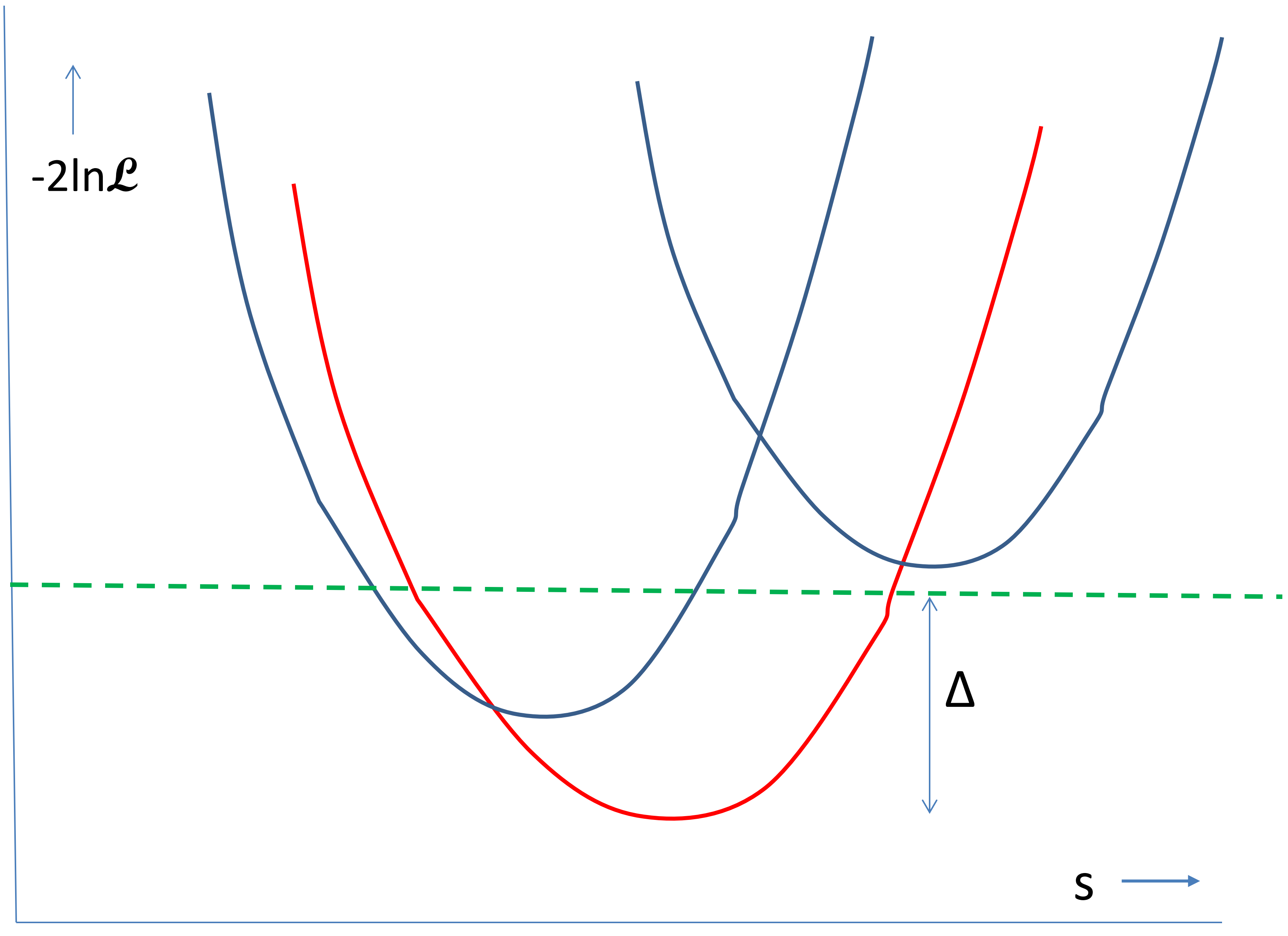


Contours of ln$\mathcal{L}$(s,υ)
s = physics param
υ = nuisance param

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 ≥ stat uncertainty



-2ln$\mathcal{L}$
Δ
s

Red curve: Best value of nuisance param υ

Blue curves: Other values of υ

Horizontal line: Intersection with red curve → statistical uncertainty

'Typical approach': Decide which blue curves have small enough Δ

Systematic is largest change in minima wrt red curves'.

Profile $\mathcal{L}$: Envelope of lots of blue curves

Wider than red curve, because of systematics (υ)

For $\mathcal{L}$ = multi-D Gaussian, agrees with 'Typical approach'

Dauncey et al use envelope of finite number of functional forms

Point of controversy!

Two types of 'other functions':

a) Different function types e.g.

$\Sigma a_i x_i$ versus $\Sigma a_i/x_i$

b) Given fn form but different number of terms

DDKW deal with b) by $-2\ln L \rightarrow -2\ln L + kn$

n = number of extra free params wrt best

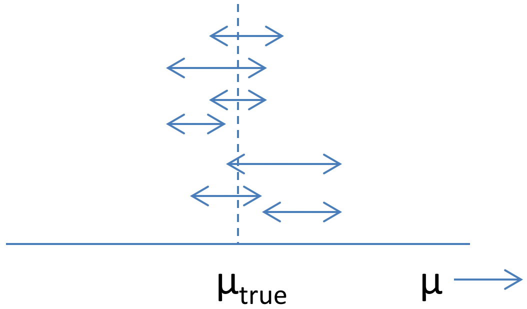
k = 1 {cf AIC = Akaike Information Criterion}

Opposition claim choice k=1 is arbitrary.

DDKW agree but have studied different values, and say k =1 is optimal for them.

Also, any parametric method needs to make such a choice

Coverage



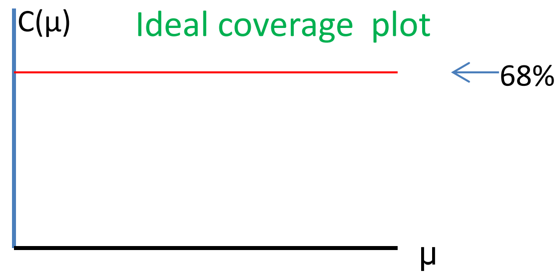
* What it is:

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

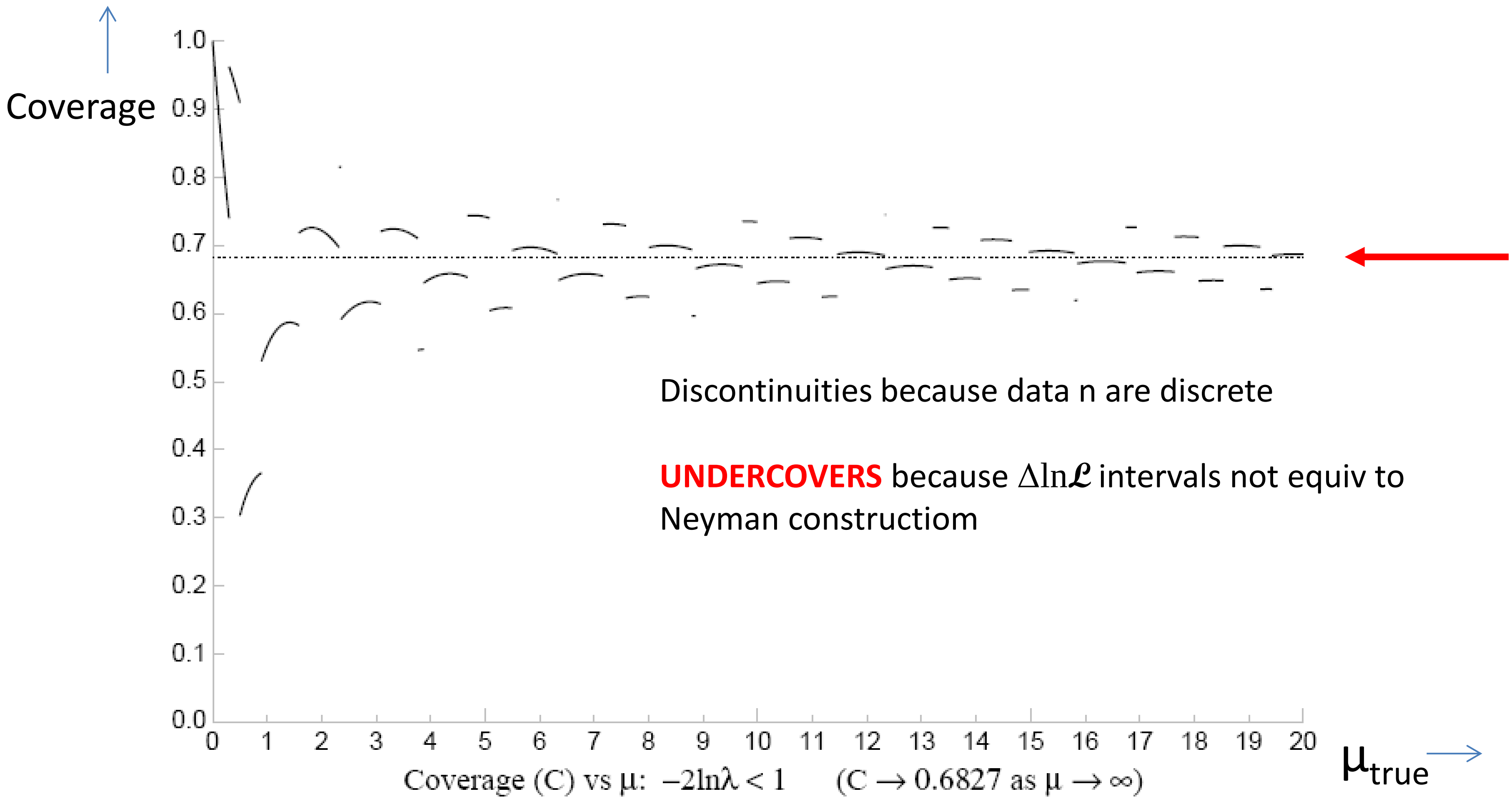
$L(\mu) = e^{-\mu} \mu^n/n!$ and range of μ from $\ln\{L(\mu_{best})/L(\mu)\} < 0.5$

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

Coverage : $\Delta \ln \mathcal{L}$ intervals for μ

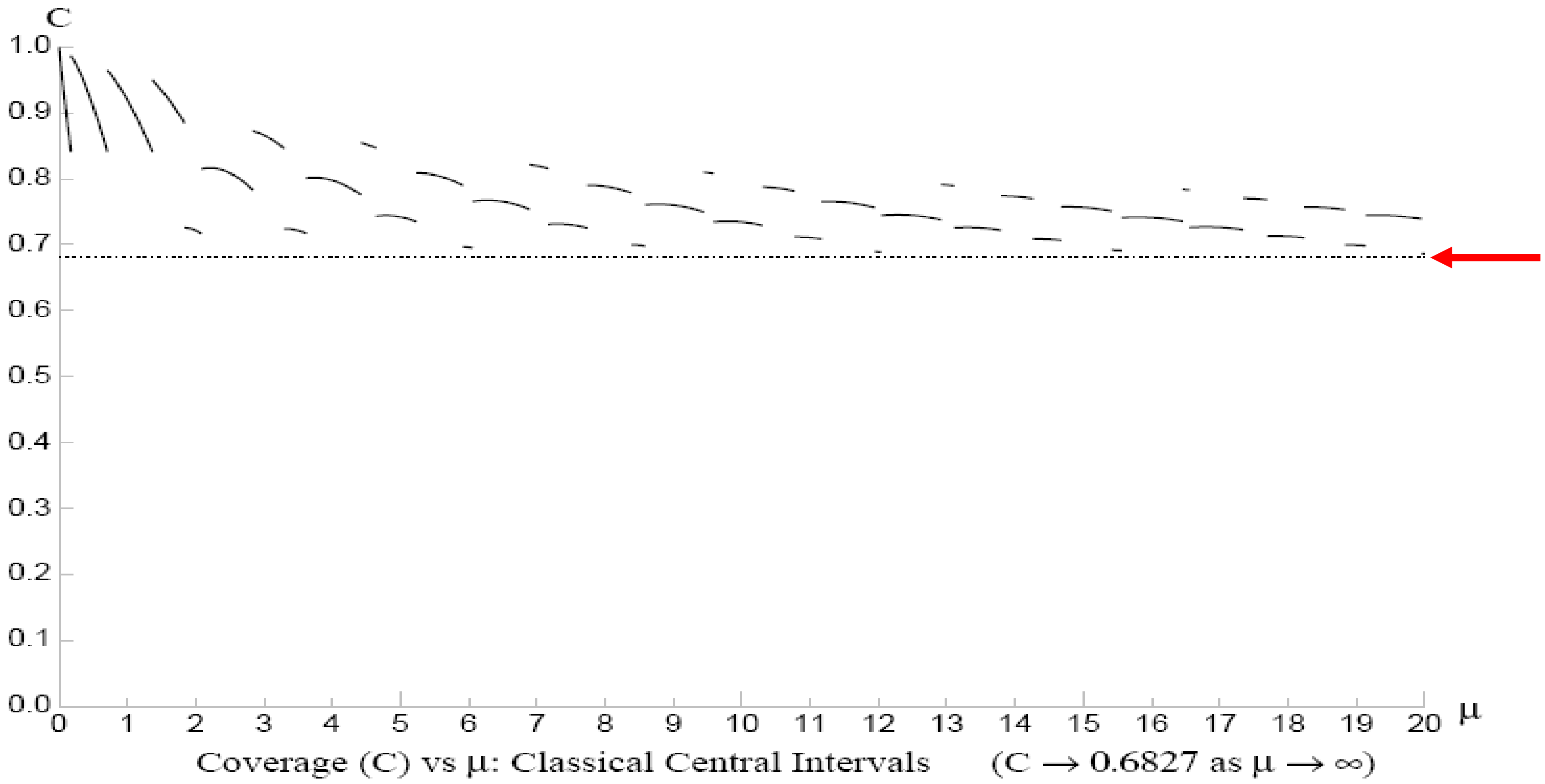
$P(n,\mu) = e^{-\mu}\mu^n/n!$ (Joel Heinrich CDF note 6438)

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



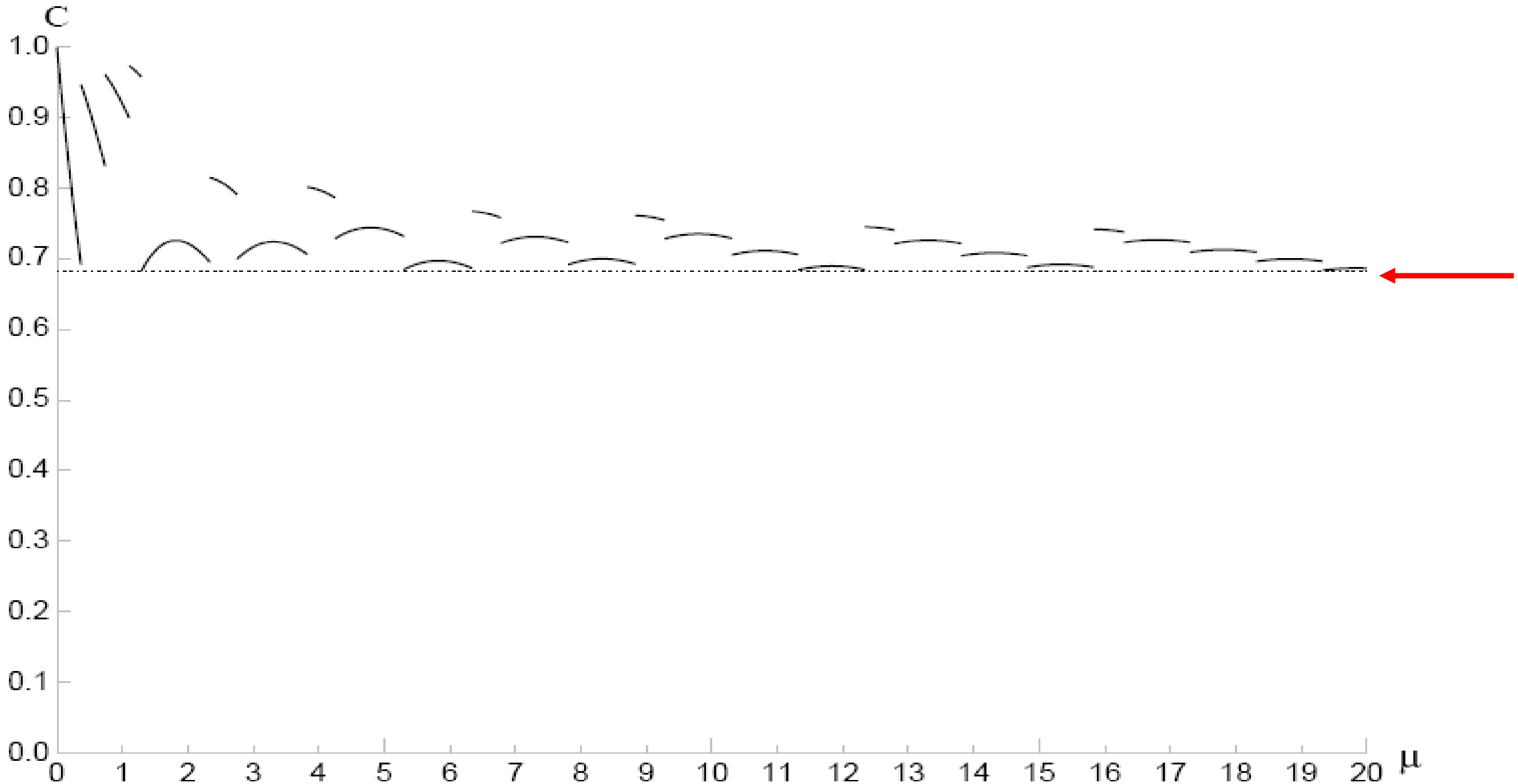
Frequentist central intervals, NEVER undercover

(Conservative at both ends)



Feldman-Cousins Unified intervals

Frequentist, so NEVER undercovers



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

Example of misleading inference

Ofer Vitells, Weizmann Institute PhD thesis (2014)

On-off problem (signal + bgd, bgd only)
e.g. $n_{on} = 10$, $m_{off} = 0$
i.e. convincing evidence for signal

Now, to improve analysis, look at spectra of events (e.g. in mass) in "on" and "off" regions
e.g. Use 100 narrow bins → $n_i = 1$ for 10 bins, $m_i = 0$ for all bins

Assume bins are chosen so that signal s_i is uniform in all bins
but bgd b_i is unknown

$\mathcal{L}$ikelihood: $\mathcal{L}(s,b_i) = e^{-Ks}\, e^{-(1+\tau)\Sigma bi}\, \Pi_j(s+b_j)$

K = number of bins (e.g. 100)

τ = scale factor for bgd (e.g. 1)

j = "on" bins with event (e.g. 1….. 10)

Profile over background nuisance params b_i

$\mathcal{L}_{prof}(s)$ maximises at

$s=0$ if $n_{on} < K/(1+\tau)$

$s=n_{on}/K$ if $n_{on} \geq K/(1+\tau)$

{Similar result for Bayesian marginalisation of $\mathcal{L}(s,b_i)$ over backgrounds b_i}

i.e. With many bins, profile (or marginalised) $\mathcal{L}$ maximises at s=0, even though n_{on} = 10 and m_{off}=0

BUT when mass distribution ignored (i.e. just counting experiment), signal+bgd is favoured over just bgd

WHY?

Background given greater freedom with large number K of nuisance parameters

Compare:
Neyman and Scott, "Consistent estimates based on partially consistent observations", Econometrica 16: 1-32 (1948)

Data = n pairs $X_{1i} = G(\mu_i, \sigma^2)$
$X_{2i} = G(\mu_i, \sigma^2)$
Param of interest = σ^2
Nuisance params = μ_i. Number increases with n
Profile L estimate of σ^2 are biassed $E = \sigma^2/2$
and inconsistent (bias does not tend to 0 as $n \rightarrow \infty$)

MORAL: Beware!

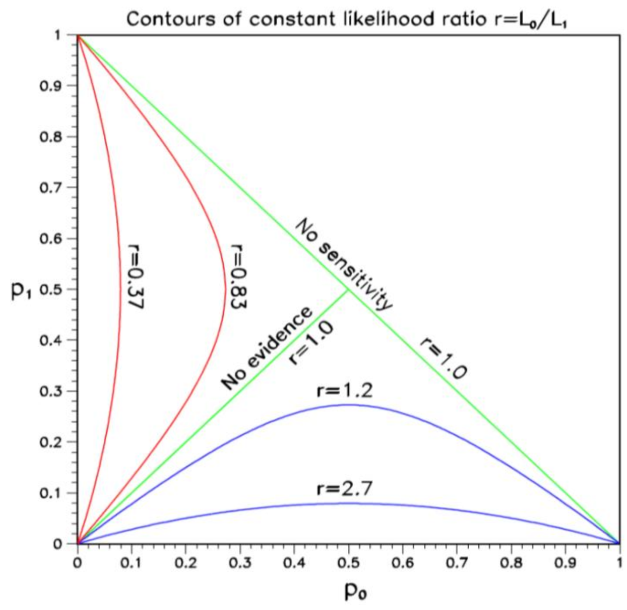
p_0 v p_1 plots

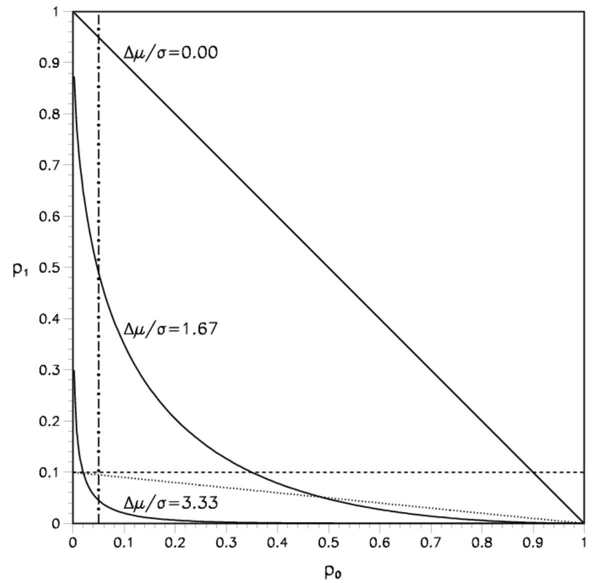
Preprint by Luc Demortier and LL,
"Testing Hypotheses in Particle Physics: Plots of p_0 versus p_1"
http://arxiv.org/abs/1408.6123

For hypotheses H0 and H1, p_0 and p_1 are the tail probabilities for data statistic t

Provide insights on:

- CLs for exclusion
- Punzi definition of sensitivity
- Relation of p-values and Likelihoods
- Probability of misleading evidence
- Sampling to foregone conclusion
- Jeffreys-Lindley paradox



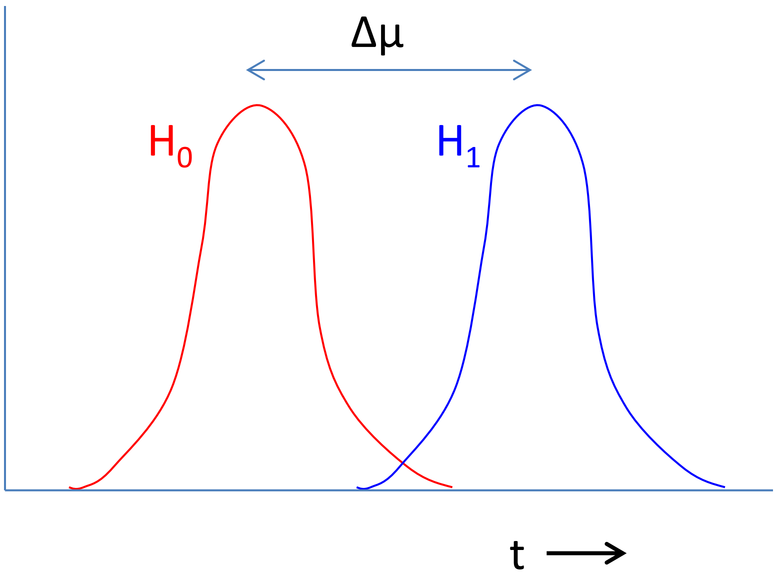


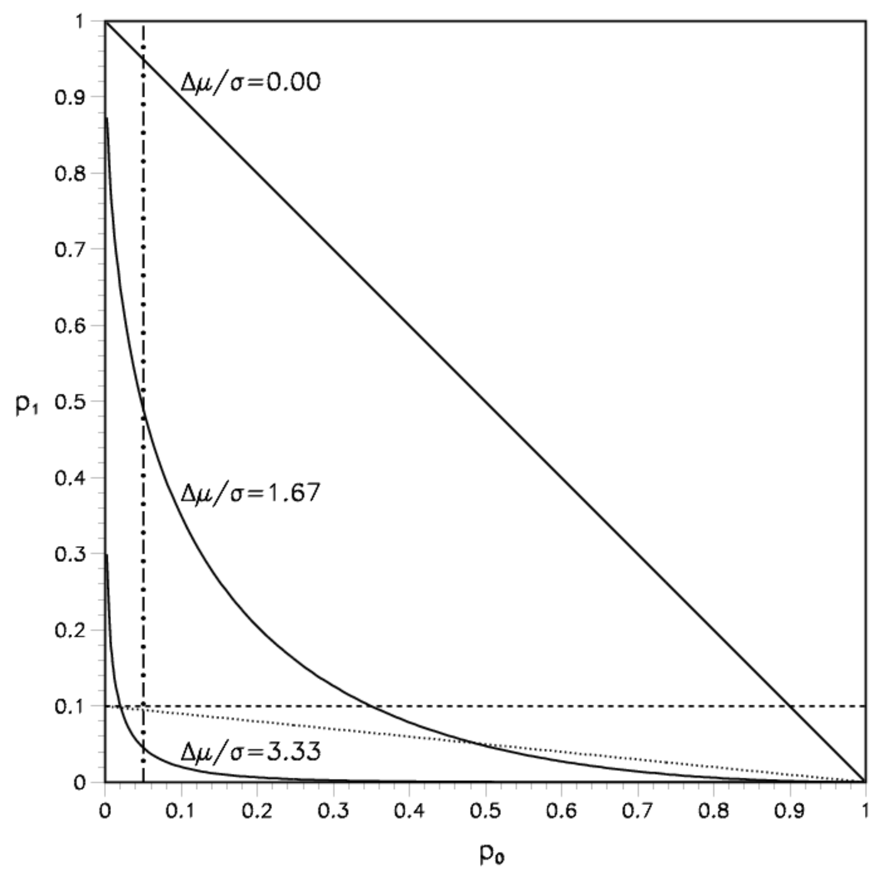
CLs = $p_1/(1-p_0)$ → diagonal line

Provides protection against excluding H_1 when little or no sensitivity

Punzi definition of sensitivity:

Enough separation of pdf's for no chance of ambiguity





Can read off power of test

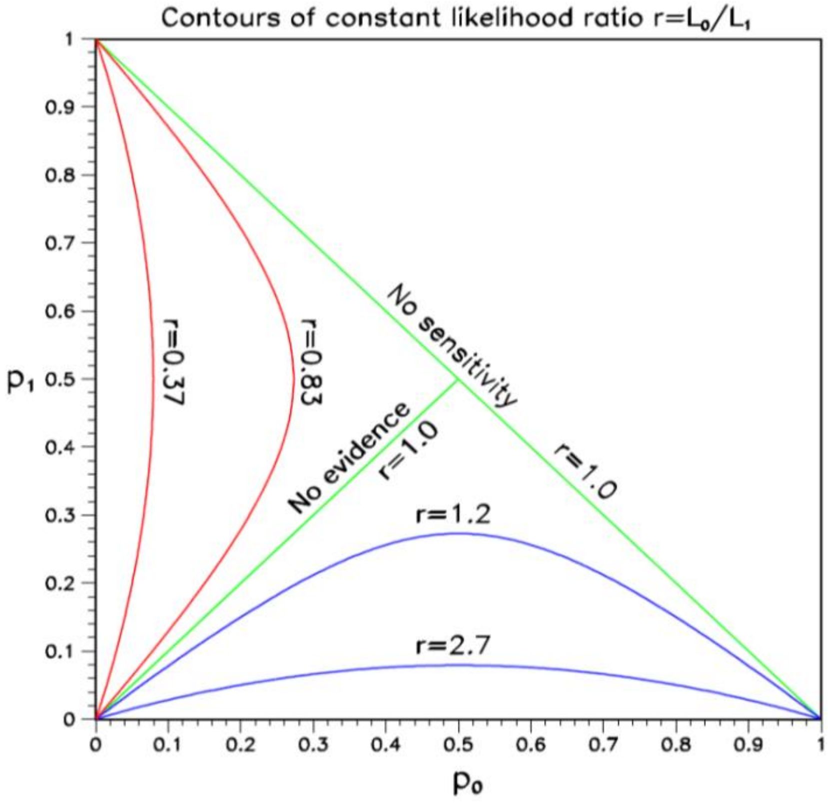
e.g. If H_0 is true, what is

prob of rejecting H_1?

N.B. p_0 = tail towards H_1

p_1 = tail towards H_0

Why p ≠ Likelihood ratio



Measure different things:

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

Depends on amount of data:

e.g. Poisson counting expt little data:

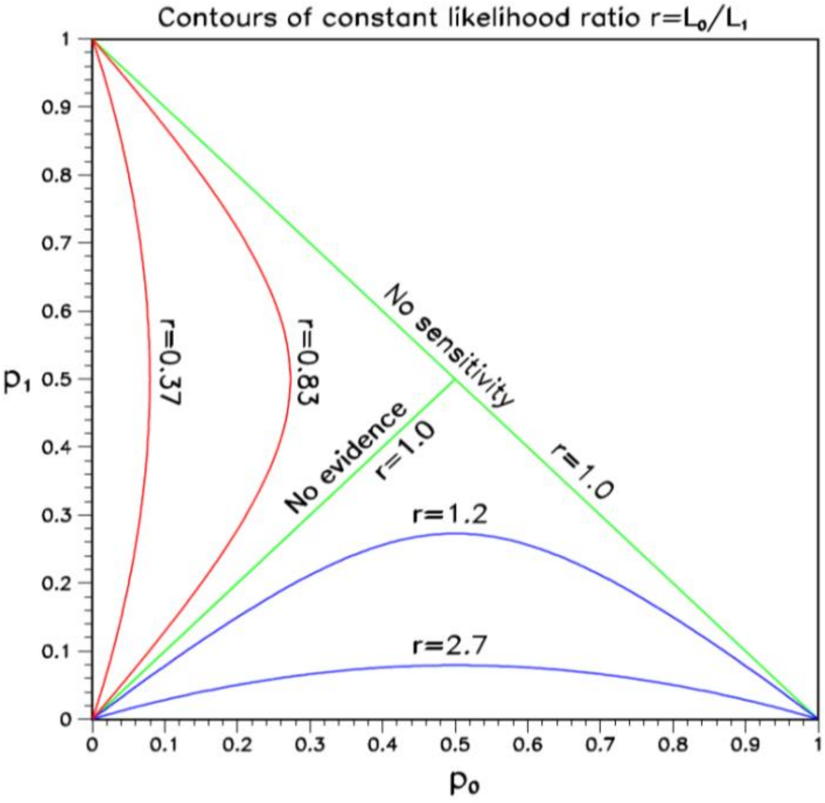
For H0, $\mu_0 = 1.0$. For H1, $\mu_1 = 10.0$

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

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

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

Jeffreys-Lindley Paradox



H_0 = simple, H_1 has μ free

p_0 can favour H_1, while B_{01} can favour H_0

$B_{01} = L_0 / \int L_1(s)\, \pi(s)\, ds$

Likelihood ratio depends on signal :

e.g. Poisson counting expt small signal s:

For H_0, $\mu_0 = 1.0$. For H_1, $\mu_1 = 10.0$

Observe n = 10 $p_0 \sim 10^{-7}$ $L_{01} \sim 10^{-5}$ and favours H_1

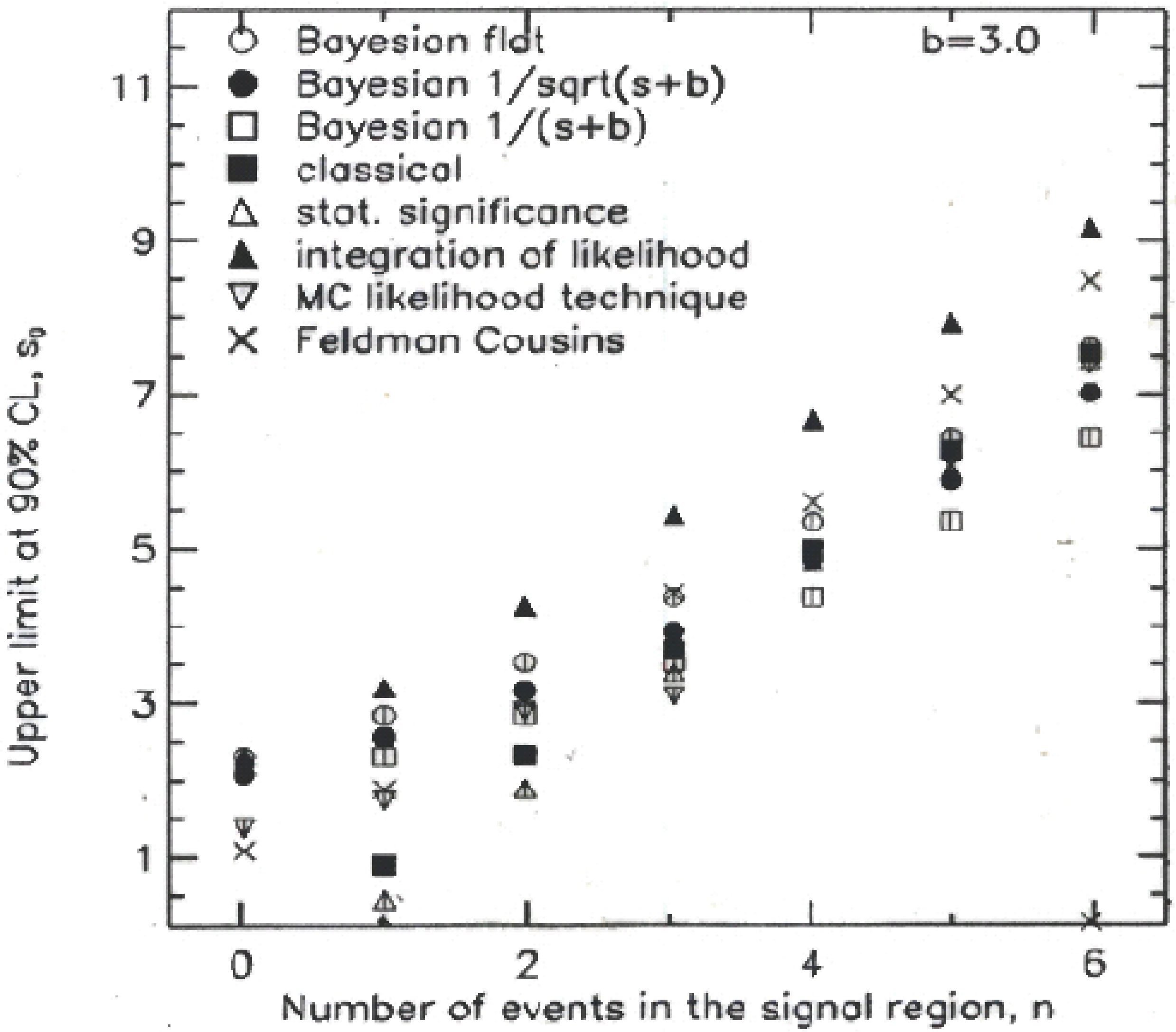
Now with 100 times as much signal s, $\mu_0 = 100.0$ $\mu_1 = 1000.0$

Observe n = 160 $p_0 \sim 10^{-7}$ $L_{01} \sim 10^{+14}$ and favours H_0

B_{01} involves intergration over s in denominator, so a wide enough range will result in favouring H_0

However, for B_{01} to favour H_0 when p_0 is equivalent to 5σ, integration range for s has to be $O(10^6)$ times Gaussian widths

Ilya Narsky, FNAL CLW 2000



Conclusions

Resources:

Software exists: e.g. RooStats

Books exist: Barlow, Cowan, James, Lyons, Roe,…..

New: `Data Analysis in HEP: A Practical Guide to Statistical Methods' , Behnke et al.

PDG sections on Prob, Statistics, Monte Carlo

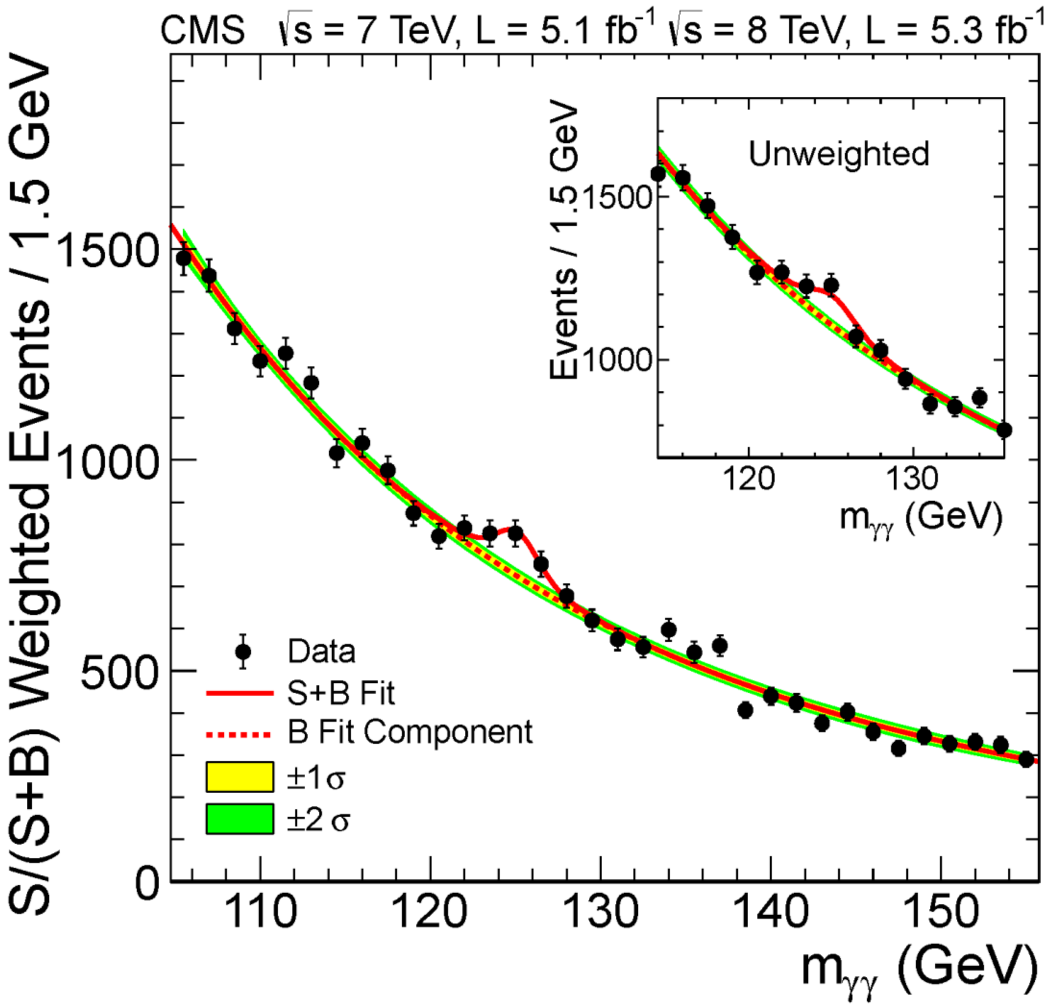
CMS and ATLAS have Statistics Committees (and BaBar and CDF earlier) – see their websites

Before re-inventing the wheel, try to see if Statisticians have already found a solution to your statistics analysis problem.

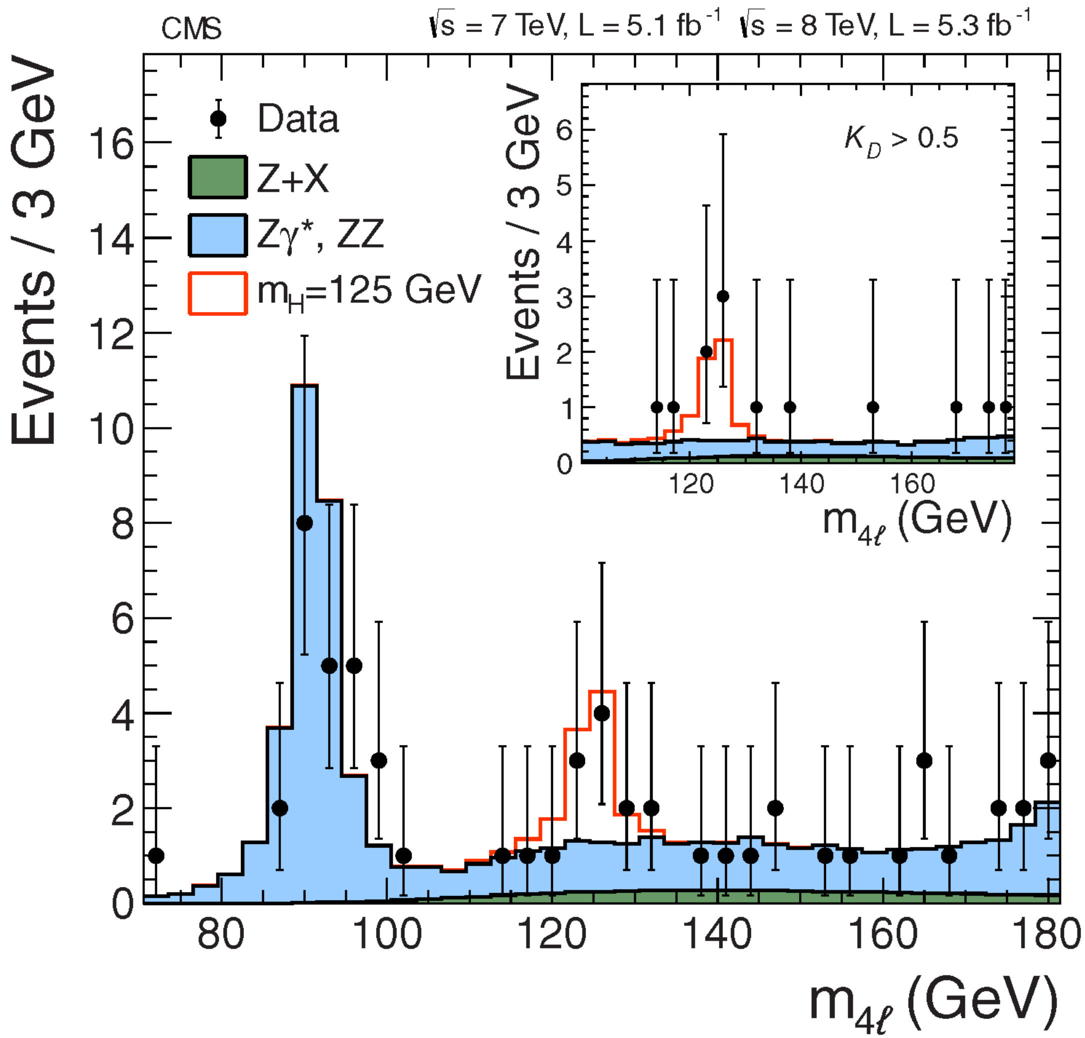
Don't use a square wheel if a circular one already exists.

"Good luck"

H→ γ γ: low S/B, high statistics



H→Z Z → 4 l: high S/B, low statistics



p-value for ‘No Higgs’ versus m_H