L_{max} and Goodness of Fit?

Conclusion:

L has sensible properties with respect to parameters

NOT with respect to data

L_{max} within Monte Carlo peak is NECESSARY

not SUFFICIENT

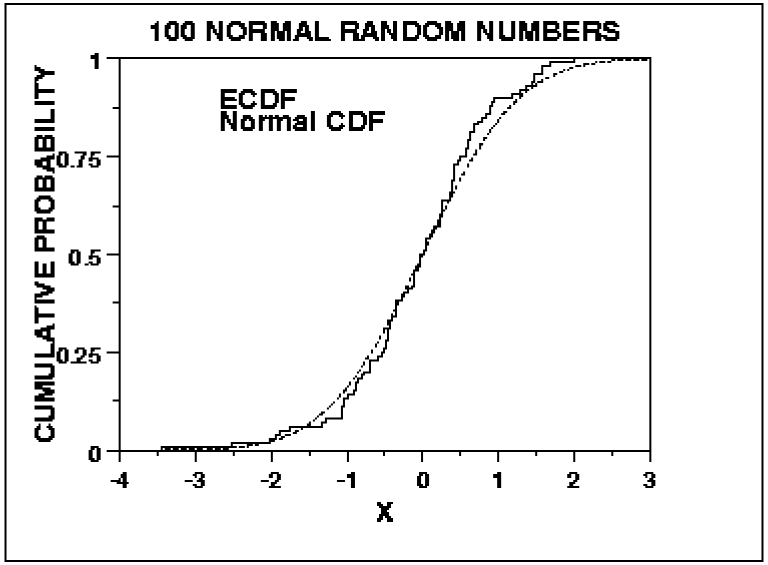
('Necessary' doesn't mean that you have to do it!)

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)

Goodness of fit: ‘Energy’ test

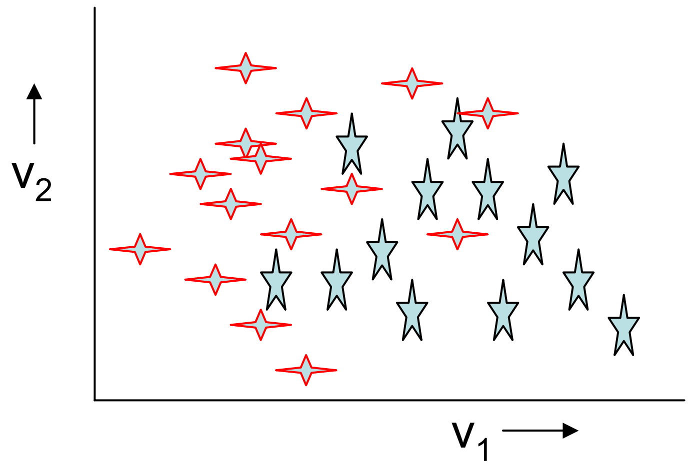
Assign +ve charge to data ; -ve charge to M.C.

Calculate ‘electrostatic energy E’ of charges

If distributions agree, E ~ 0

If distributions don’t overlap, E is positive

Assess significance of magnitude of E by MC



N.B.

1) Works in many dimensions
2) Needs metric for each variable (make variances similar?)
3) $E \sim \Sigma\, q_i q_j\, f(\Delta r = |r_i - r_j|)$, $f = 1/(\Delta r + \varepsilon)$ or $-\ln(\Delta r + \varepsilon)$

   Performance insensitive to choice of small ε

See Aslan and Zech’s paper at:
http://www.ippp.dur.ac.uk/Workshops/02/statistics/program.shtml

Combining different p-values

Several results quote 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

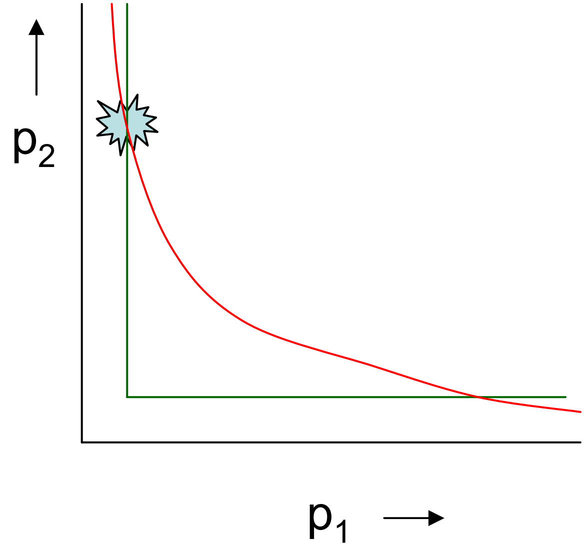
Conventional:

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	$2.0\ 10^{-2}$	$2.0\ 10^{-2}$	$2.0\ 10^{-4}$	$2.0\ 10^{-4}$

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:

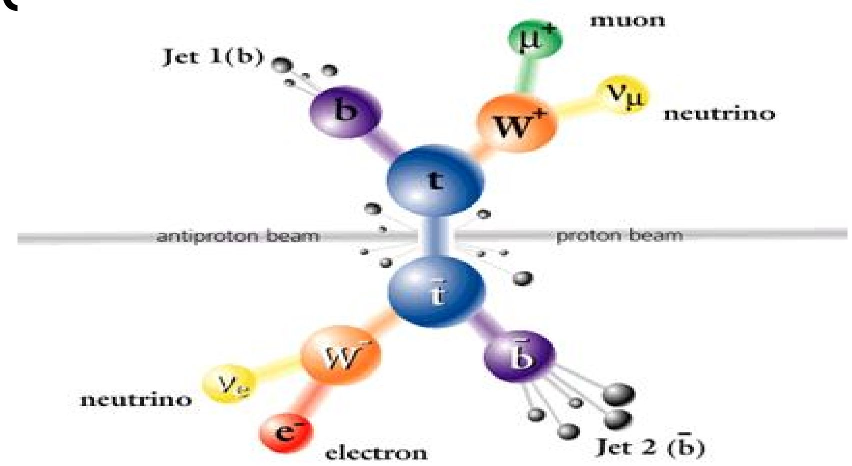
$$\underbrace{\frac{P(H0|data)}{P(H1|data)}}_{\text{Bayes posteriors}} = \frac{P(data|H0) * P(H0)}{P(data|H1) * P(H1)}$$

Bayes posteriors Likelihoods Priors

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

Sleuth

a quasi-model-independent search strategy for new physics



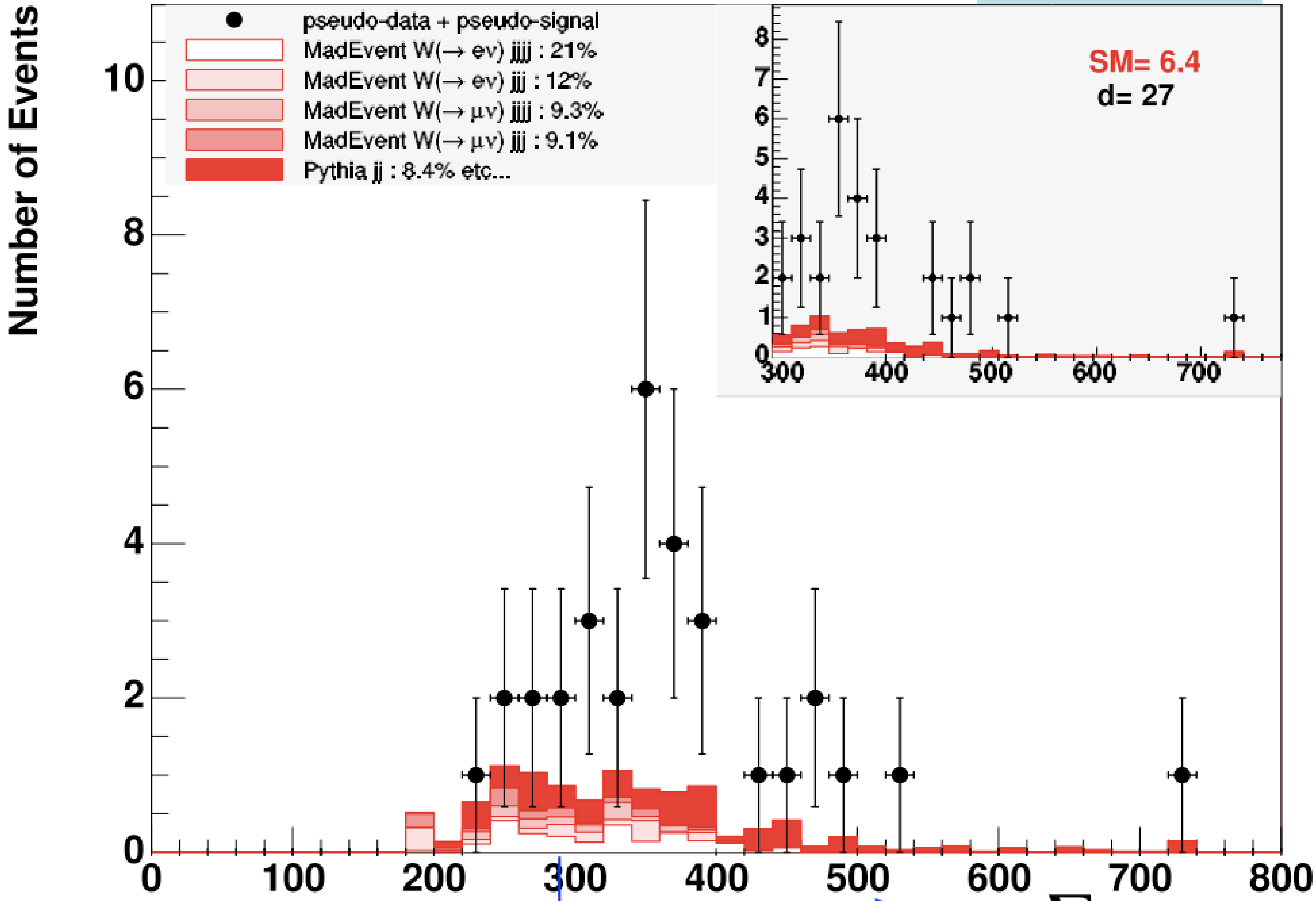
Assumptions:

1. Exclusive final state
2. Large $\sum p_T$
3. An excess

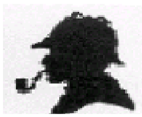
$$\int_{0001001}^{0608025} (\text{prediction})\, d(\text{hep-ph})$$

Rigorously compute the trials factor associated with looking everywhere

$W\,b\bar{b}\,jj$ pseudo discovery $P_{Wb\bar{b}jj} < 8e\text{-}08$ $\tilde{P} < 4e\text{-}05$



Sleuth



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

What is p good for?

Used to test whether data is consistent with H0

Reject H0 if p is small : $p \leq \alpha$ (How small?)

Sometimes make wrong decision:

Reject H0 when H0 is true: Error of 1st kind

Should happen at rate α

OR

Fail to reject H0 when something else (H1,H2,…) is true: Error of 2nd kind

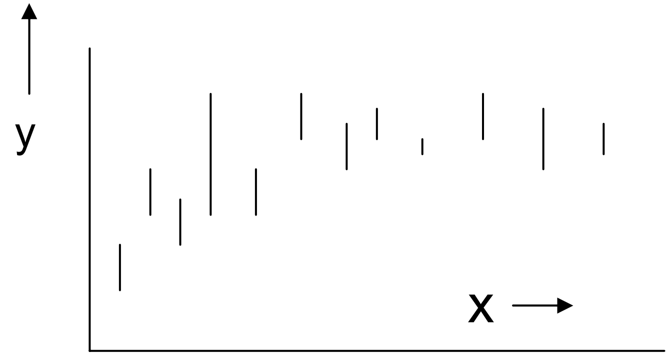
Rate at which this happens depends on……….

Errors of 2nd kind: How often?

e.g.1. Does data line on straight line?

Calculate χ^2

Reject if $\chi^2 \geq 20$



Error of 1st kind: $\chi^2 \geq 20$ Reject H0 when true

Error of 2nd kind: $\chi^2 \leq 20$ Accept H0 when in fact quadratic or..

How often depends on:

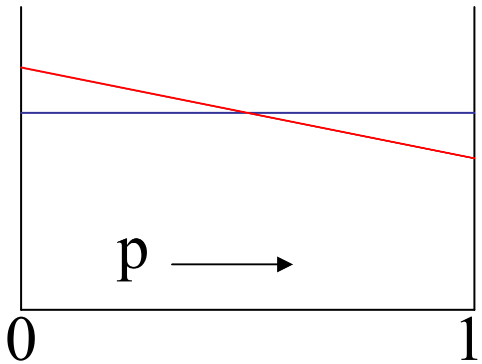
- Size of quadratic term
- Magnitude of errors on data, spread in x-values,…….
- How frequently quadratic term is present

Errors of 2^{nd} kind: How often?

e.g. 2. Particle identification (TOF, dE/dx, Čerenkov,…….)

Particles are π or μ

Extract p-value for H0 = π from PID information



π and μ have similar masses

Of particles that have p ~ 1% (‘reject H0’), fraction that are π is

a) ~ half, for equal mixture of π and μ

b) almost all, for “pure” π beam

c) very few, for “pure” μ beam

What is p good for?

Selecting sample of wanted events

e.g. kinematic fit to select $t\,\underline{t}$ events

$t \rightarrow bW$, $b \rightarrow jj$, $W \rightarrow \mu\nu$ $\quad \underline{t} \rightarrow \underline{b}W$, $\underline{b} \rightarrow jj$, $W \rightarrow jj$

Convert χ^2 from kinematic fit to p-value

Choose cut on χ^2 to select $t\,\underline{t}$ events

Error of 1st kind: Loss of efficiency for $t\,\underline{t}$ events

Error of 2nd kind: Background from other processes

Loose cut (large χ^2_{max}, small p_{min}): Good efficiency, larger bgd

Tight cut (small χ^2_{max}, larger p_{min}): Lower efficiency, small bgd

Choose cut to optimise analysis:

- More signal events: Reduced statistical error
- More background: Larger systematic error

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) $\neq$ P(data|H0)

P(theory|data) $\neq$ 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"

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

Theory = male or female

Data = pregnant or not pregnant

P (pregnant ; female) ~ 3%

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

Theory = male or female

Data = pregnant or not pregnant

$P\,(\text{pregnant ; female}) \sim 3\%$

but

$P\,(\text{female ; pregnant}) >>> 3\%$

Aside: Bayes' Theorem

$P(A \text{ and } B) = P(A|B) * P(B) = P(B|A) * P(A)$

$N(A \text{ and } B)/N_{tot} = N(A \text{ and } B)/N_B * N_B/N_{tot}$

If A and B are independent, $P(A|B) = P(A)$

Then $P(A \text{ and } B) = P(A) * P(B)$, but not otherwise

e.g. $P(\text{Rainy and Sunday}) = P(\text{Rainy})$

But $P(\text{Rainy and Dec}) = P(\text{Rainy}|\text{Dec}) * P(\text{Dec})$

$25/365 \quad = \quad 25/31 \quad * 31/365$

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

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.

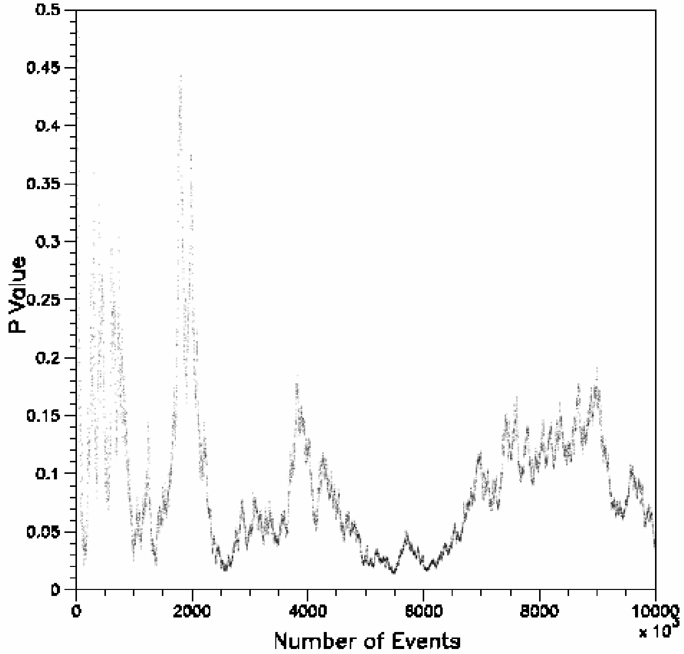
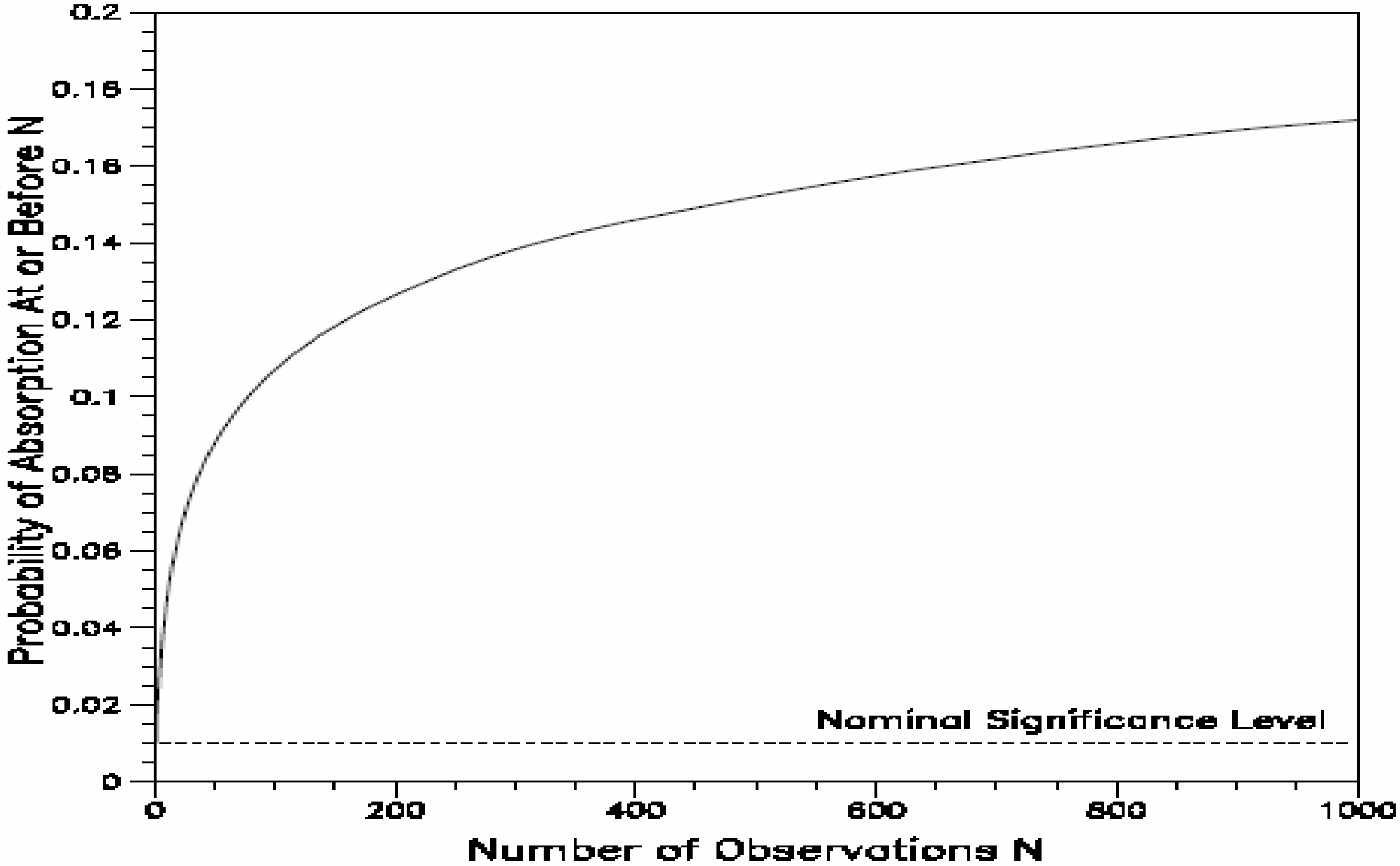


Figure 1: P value versus sample size.

More "More and more data"



PARADOX

Histogram with 100 bins

Fit 1 parameter

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

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

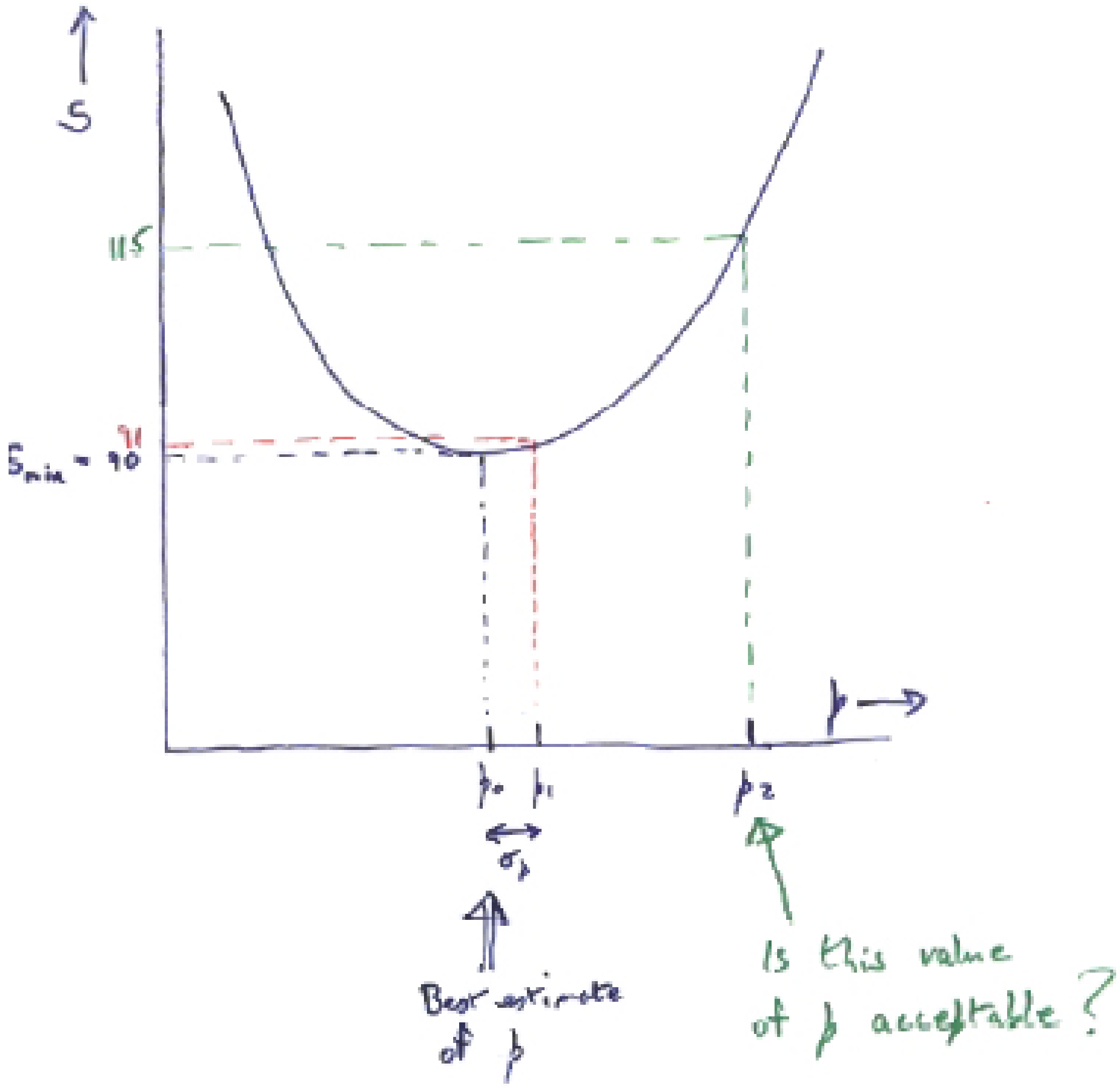
Is p_1 acceptable if $S(p_1) = 115$?

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
p_0
p_1
p_2
σ_p
Best estimate of p
Is this value of p acceptable?

NDF = 99

SELECTING BETWEEN TWO HYPOTHESES

Louis Lyons

OUNP-99-12

MATHEMATICAL FORMULATION

$$S(x) = \sum \frac{(x_i - x)^2}{\sigma^2} \equiv \sum \frac{(x_i - \bar{x})^2}{\sigma^2} + N\frac{(\bar{x} - x)^2}{\sigma^2}$$

SCATTER OF POINTS WRT THEIR MEAN.

INDEP OF x

THIS IS TERM WHICH HAS EXPECTED VALUE $(N-1) \pm \sqrt{2(N-1)}$

χ^2_{N-1}

HOW WELL x AGREES WITH $\bar{x}$

VARIES WITH x

BEST VALUE IS $x = \bar{x}$

INCREASES BY 1 FOR $x = \bar{x} \pm \frac{\sigma}{\sqrt{N}}$

χ^2_1

CONCLUSION FOR THIS CASE

COMPARING H1 : $p = p_1$

& H2 : $p = p_2$

DECISION DEPENDS ON $\Delta\chi^2$

ANOTHER EXAMPLE

$x = -1 \to +1$

100 bins
~100 events/bin
No free params

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

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

Generate events according to H1 (+ stat fluctn)

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

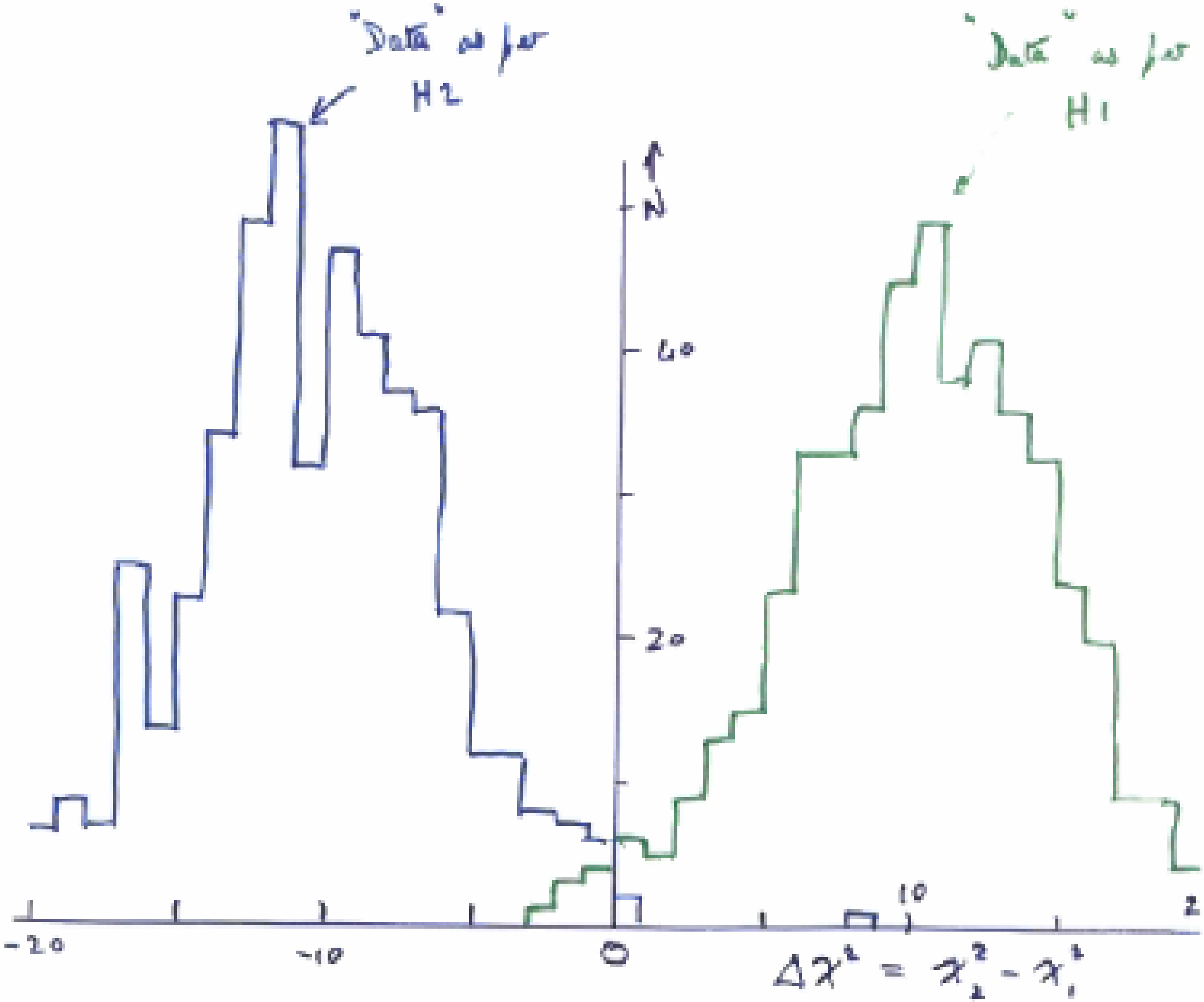
Look at dist of χ^2_1
χ^2_2
$\chi^2_2 - \chi^2_1$

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

2H 6

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

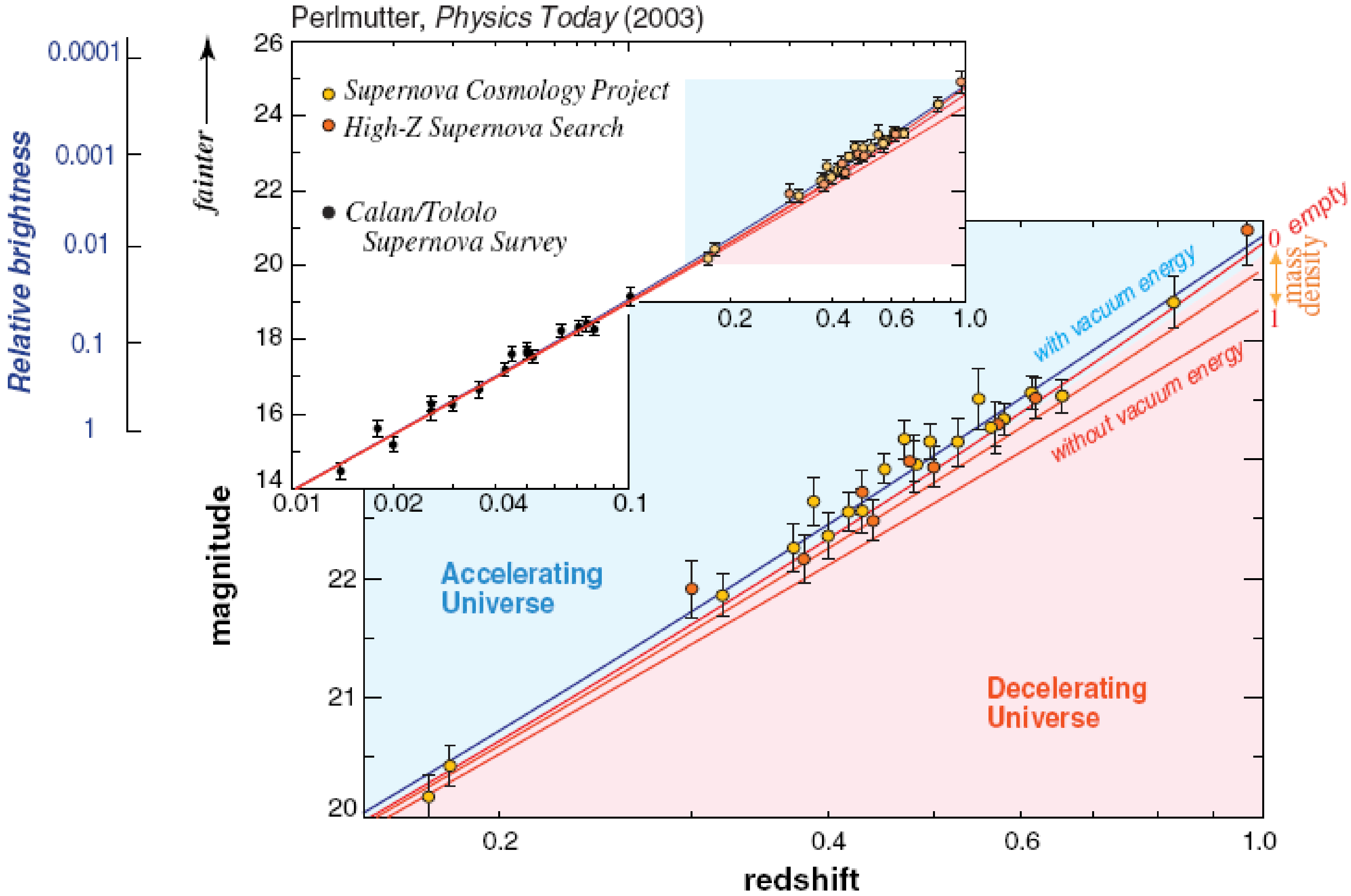
(500 SIMULATIONS)



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

$H1 = 1 + 0.05x$

Comparing data with different hypotheses



Choosing between 2 hypotheses

Possible methods:

$\Delta\chi^2$

$ln\mathbf{L}$–ratio

Bayesian evidence

Minimise “cost”

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*L-ratio)

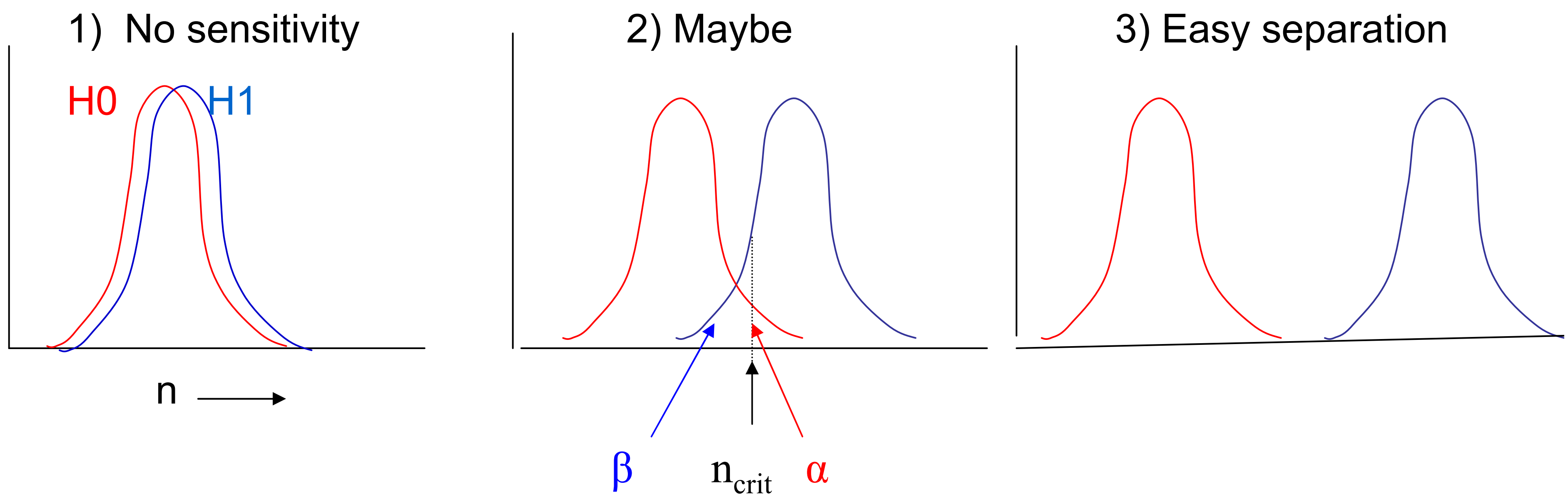
Traditional sensitivity:

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

スライド 50

N1 NPL, 2005/11/06

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
- 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) $\neq$ 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
- Worry about systematics

PHYSTAT Workshop at CERN, June 27 → 29 2007

"Statistical issues for LHC Physics Analyses"

Final message

Send interesting statistical issues to

l.lyons@physics.ox.ac.uk