p-values and Discovery

Louis Lyons

Oxford

l.lyons@physics.ox.ac.uk

Kyoto,

May 2007

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

Discoveries

H0 or H0 v H1

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)

THE paradox

Optimising for discovery and exclusion

Incorporating nuisance parameters

DISCOVERIES

"Recent" history:

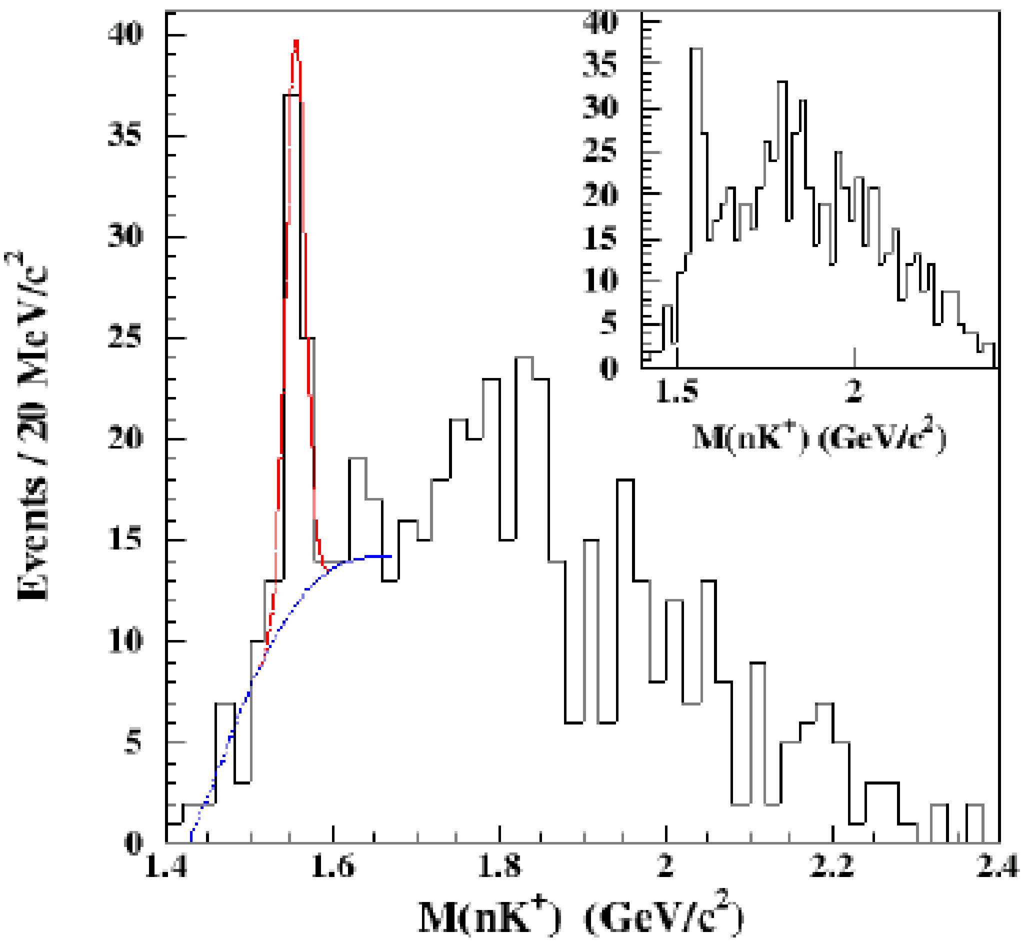
Charm	SLAC, BNL	1974
Tau lepton	SLAC	1977
Bottom	FNAL	1977
W,Z	CERN	1983
Top	FNAL	1995
{Pentaquarks	~Everywhere	2002 }
?	FNAL/CERN	2008?

? = Higgs, SUSY, q and l substructure, extra dimensions,
free q/monopoles, technicolour, 4th generation, black holes,…..

QUESTION: How to distinguish discoveries from fluctuations or goofs?

Penta-quarks?

Hypothesis testing: New particle or statistical fluctuation?



H0 or H0 versus H1 ?

H0 = null hypothesis

e.g. Standard Model, with nothing new

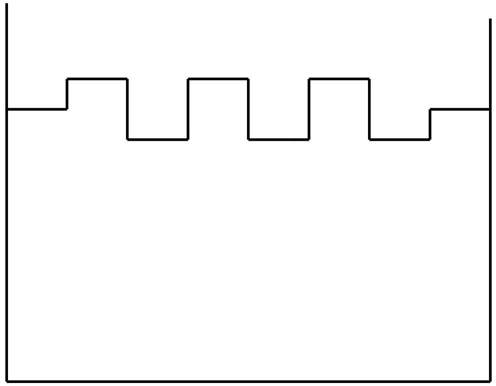
H1 = specific New Physics   e.g. Higgs with M_H = 120 GeV

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

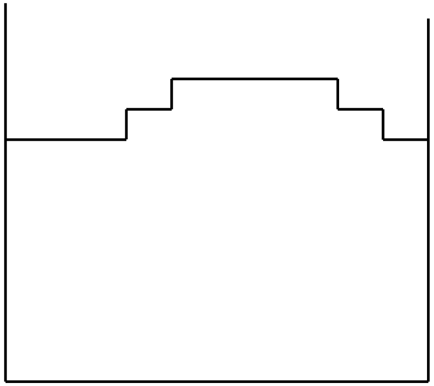
H0 v H1: “Hypothesis Testing” e.g. $\mathcal{L}$-ratio

Measures how much data favours one hypothesis wrt other

H0 v H1 likely to be more sensitive



or



Testing H0:
Do we have an alternative in mind?

1) Data is number (of observed events)

"H1" usually gives larger number

(smaller number of events if looking for oscillations)

2) Data = distribution. Calculate χ^2.

Agreement between data and theory gives χ^2 ~ndf

Any deviations give large χ^2

So test is independent of alternative?

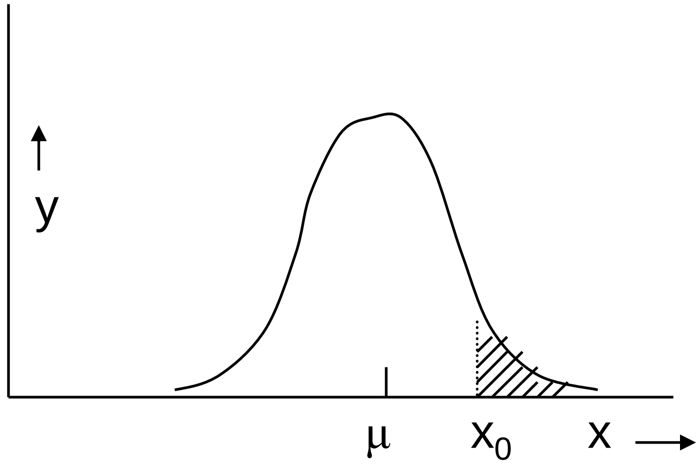
Counter-example: Cheating undergraduate

3) Data = number or distribution

Use $\mathcal{L}$-ratio as test statistic for calculating p-value

4) H0 = Standard Model

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

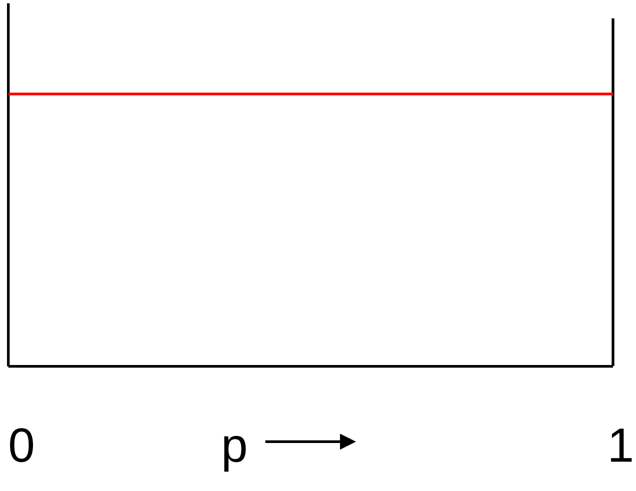
p-values, contd

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)

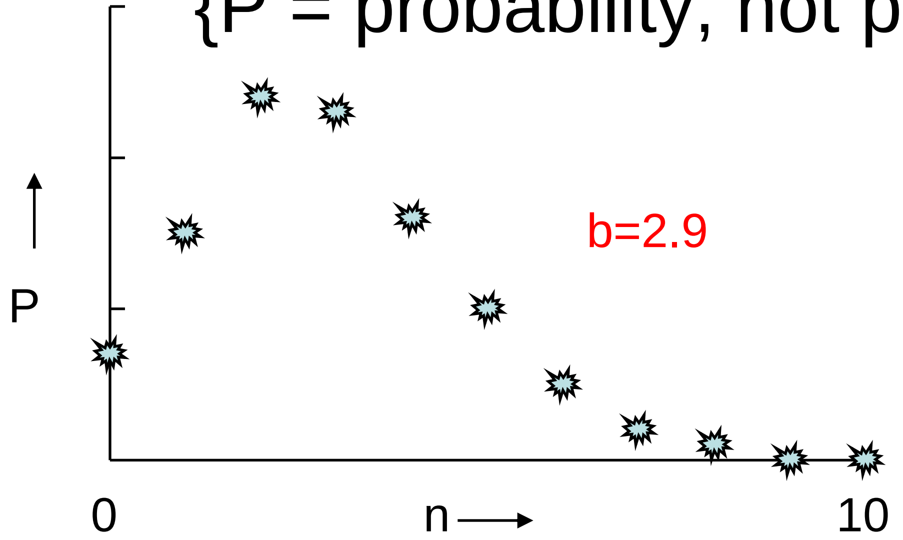


p-values for non-Gaussian distributions

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)

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^{-6}

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

$y = N\{1 + 10^{-6}\cos(x - x_0)\}$ x_0 = free param

b) Neutrino oscillations: almost degenerate parameters

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

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

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

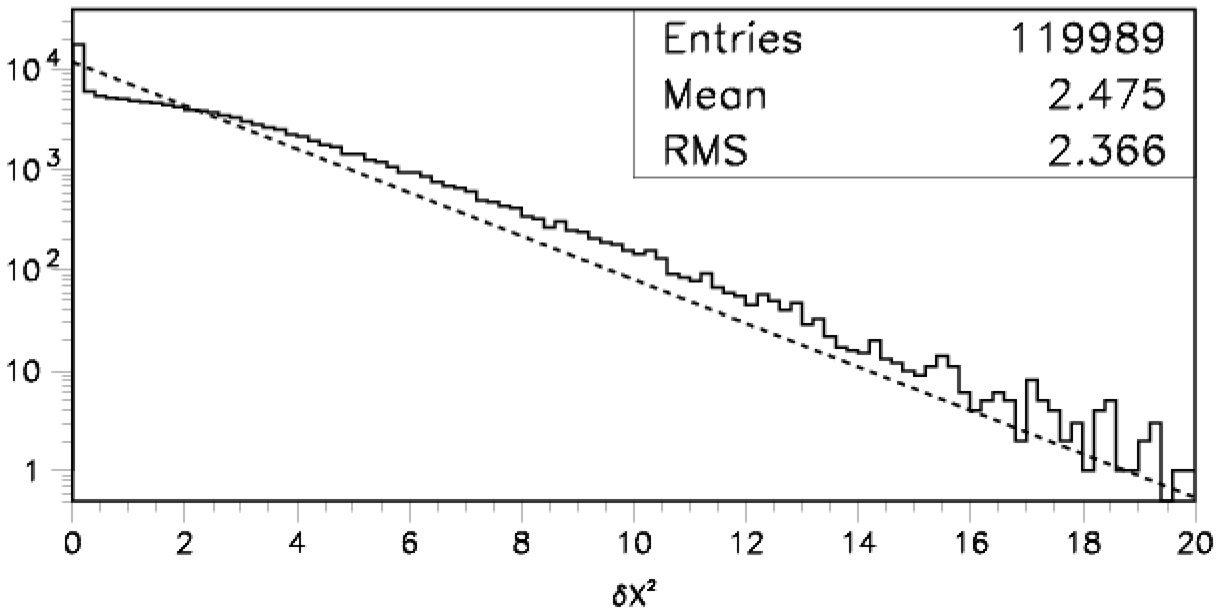
Is $\chi^2_{H0} - \chi^2_{H1}$ distributed as χ^2 with $\nu = k = 3$?

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

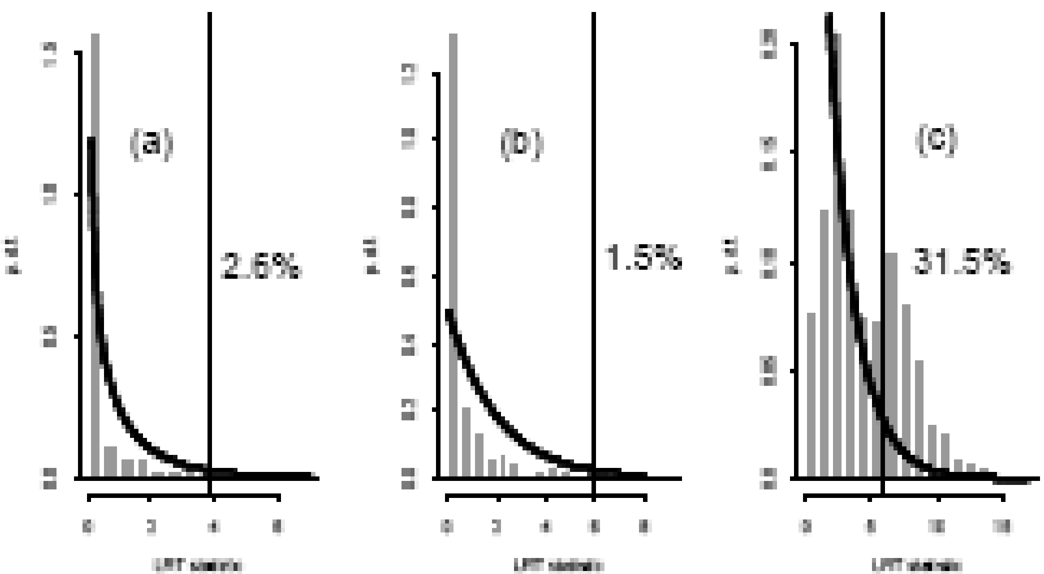
N.B. Under H0 ($y = C(x)$) : $A=0$ (boundary of physical region)

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)

Unbinned $\mathcal{L}_{max}$ and Goodness of Fit?

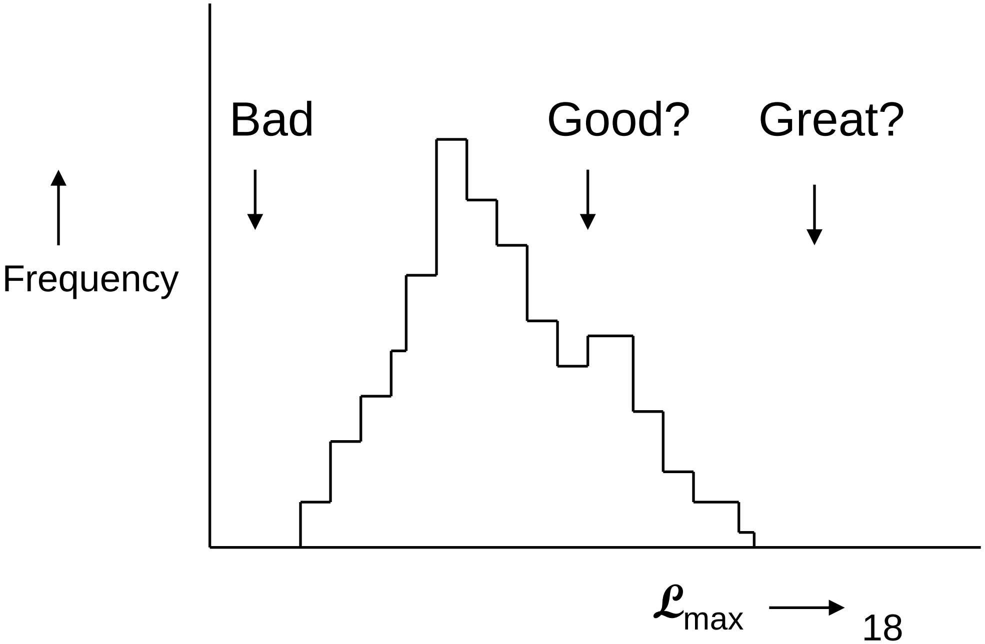
Find params by maximising $\mathcal{L}$

So larger $\mathcal{L}$ better than smaller $\mathcal{L}$

So $\mathcal{L}_{max}$ gives Goodness of Fit ??

Monte Carlo distribution

of unbinned $\mathcal{L}_{max}$



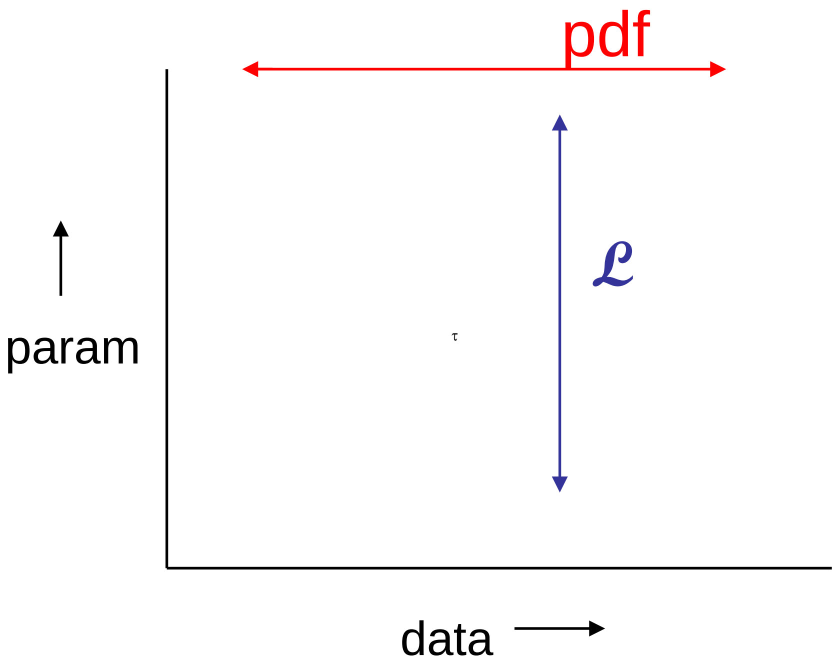
Not necessarily:

$\mathcal{L}$(data,params)

fixed vary

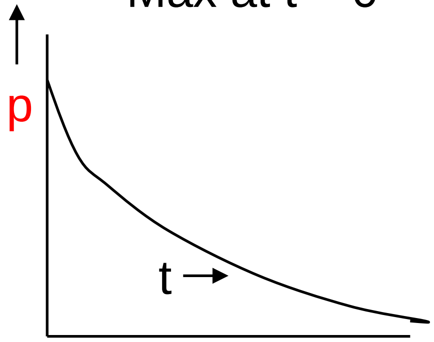
Contrast pdf(data,params)

vary fixed

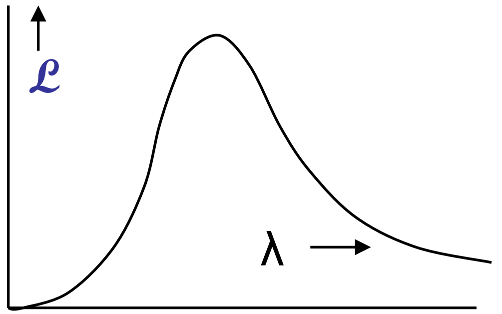


e.g. $p(t,\lambda) = \lambda * \exp(-\lambda t)$

Max at t = 0



Max at λ=1/t



Example 1: Exponential distribution

Fit exponential λ to times t_1, t_2, t_3 ……. [Joel Heinrich, CDF 5639]

$$\mathcal{L} = \prod_i \lambda e^{-\lambda t}$$

$$\ln\mathcal{L}_{max} = -N(1 + \ln t_{av})$$

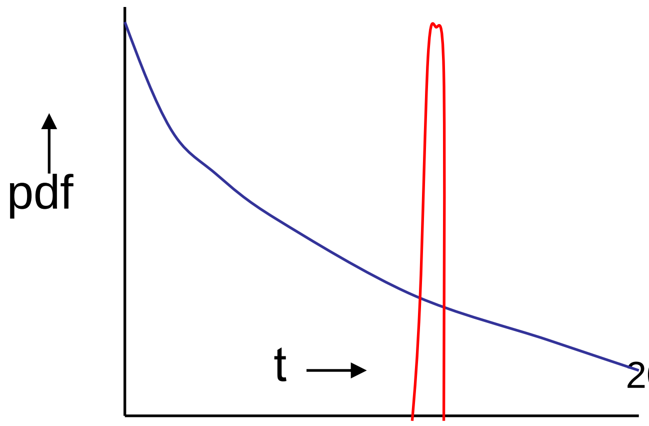
i.e. $\ln\mathcal{L}_{max}$ depends only on AVERAGE t, but is

INDEPENDENT OF DISTRIBUTION OF t (except for……..)

(Average t is a sufficient statistic)

Variation of $\mathcal{L}_{max}$ in Monte Carlo is due to variations in samples' average t , but

NOT TO BETTER OR WORSE FIT

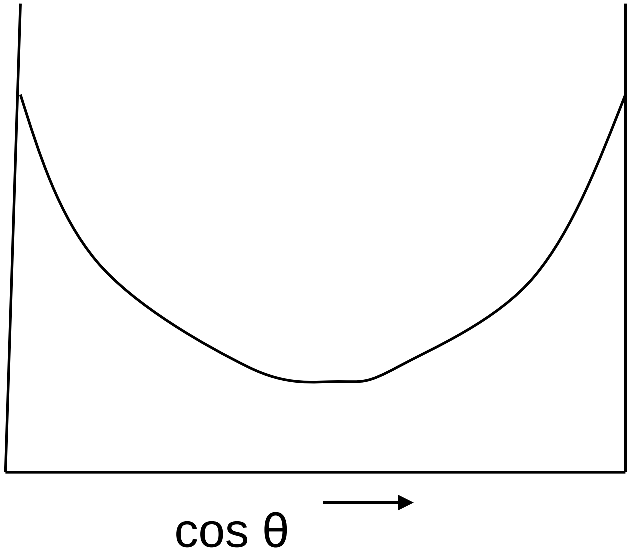


Same average t $\Longrightarrow$ same $\mathcal{L}_{max}$

Example 2

$$\frac{dN}{d\cos\theta} = \frac{1+\alpha\cos^2\theta}{1+\alpha/3}$$

$$\mathcal{L} = \prod_i \frac{1+\alpha\cos^2\vartheta_i}{1+\alpha/3}$$



pdf (and likelihood) depends only on $\cos^2\theta_i$

Insensitive to sign of $\cos\theta_i$

So data can be in very bad agreement with expected distribution

e.g. all data with $\cos\theta < 0$, but $\mathcal{L}_{max}$ does not know about it.

Example of general principle

Example 3

Fit to Gaussian with variable μ, fixed σ

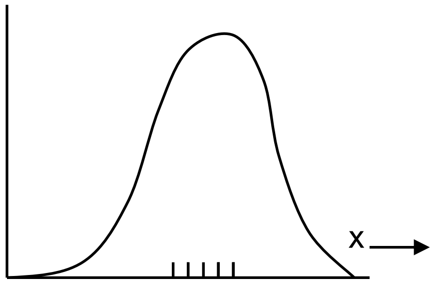
$$pdf = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

$$\underbrace{\ln \mathcal{L}_{max} = N(-0.5 \ln 2\pi}_{\text{constant}} - \ln\sigma) - 0.5\, \underbrace{\Sigma(x_i - x_{av})^2}_{\text{\textasciitilde variance(x)}} / \sigma^2$$

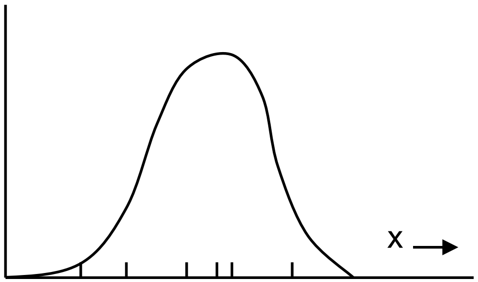
i.e. $\mathcal{L}_{max}$ depends only on variance(x),

which is not relevant for fitting μ ($\mu_{est} = x_{av}$)

Smaller than expected variance(x) results in larger $\mathcal{L}_{max}$



Worse fit, larger $\mathcal{L}_{max}$



Better fit, lower $\mathcal{L}_{max}$

$\mathcal{L}_{max}$ and Goodness of Fit?

Conclusion:

$\mathcal{L}$ has sensible properties with respect to parameters

NOT with respect to data

$\mathcal{L}_{max}$ within Monte Carlo peak is NECESSARY

not SUFFICIENT

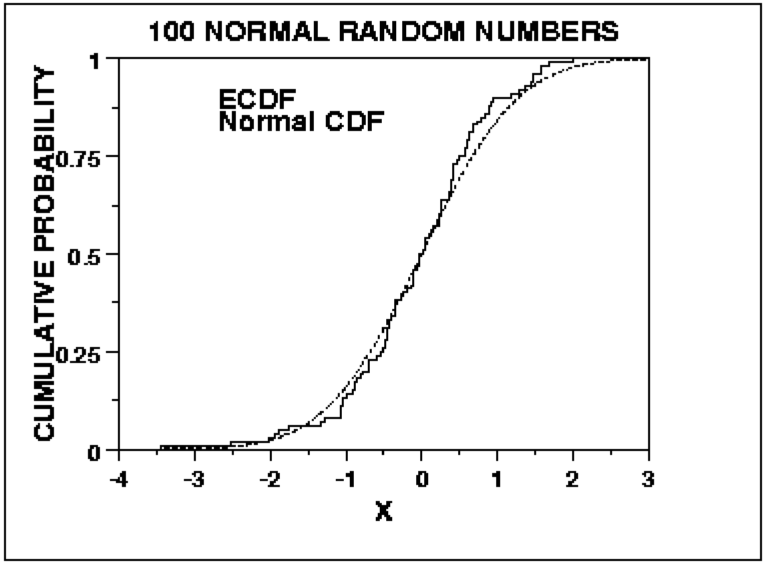
('Necessary' doesn't mean that you have to do it!)

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

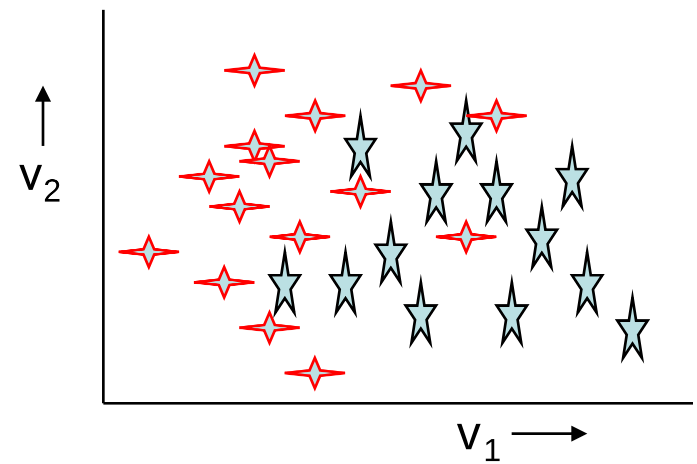
Assign +ve charge to data ; -ve charge to M.C.

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

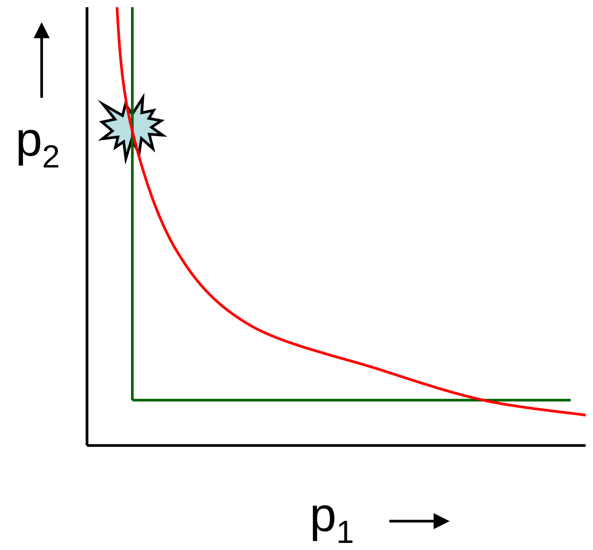
Conventional:

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:

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

Bayes posteriors Likelihoods Priors

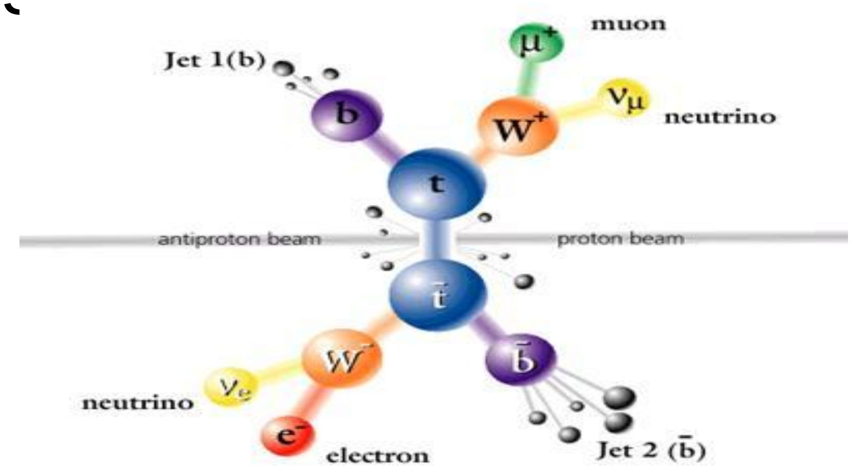
Prior for {H0 = S.M.} >>> Prior for {H1 = New Physics}

Sleuth

a quasi-model-independent search strategy for new physics

Assumptions:

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

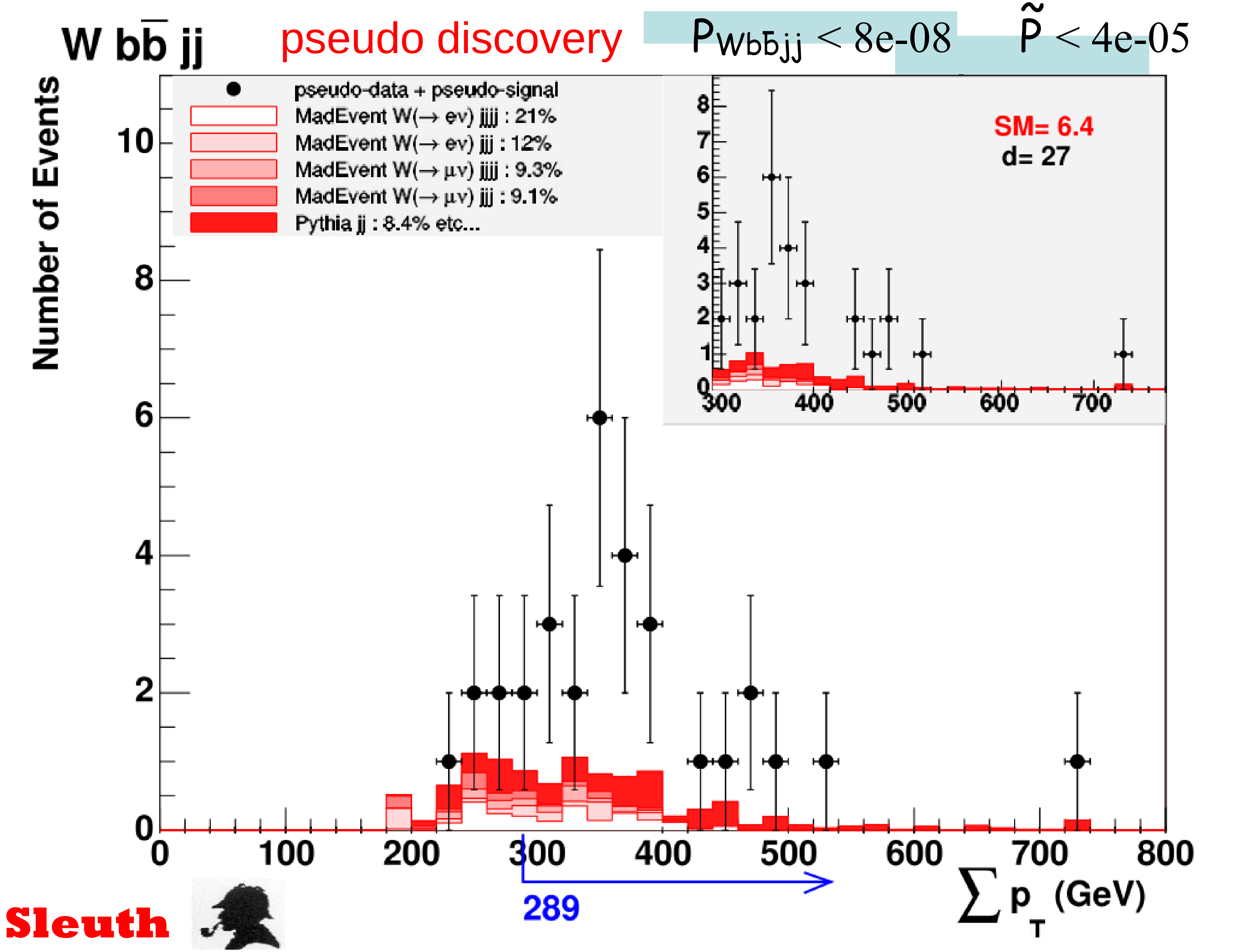


0608025

(prediction) d(hep-ph)

0001001

Rigorously compute the trials factor associated with looking everywhere



W bb̄ jj
pseudo discovery
$P_{Wb\bar{b}jj} < 8e\text{-}08$
$\tilde{P} < 4e\text{-}05$
Number of Events
pseudo-data + pseudo-signal
MadEvent W(→ eν) jjjj : 21%
MadEvent W(→ eν) jjj : 12%
MadEvent W(→ μν) jjjj : 9.3%
MadEvent W(→ μν) jjj : 9.1%
Pythia jj : 8.4% etc...
SM= 6.4
d= 27
289
$\sum p_T$ (GeV)
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

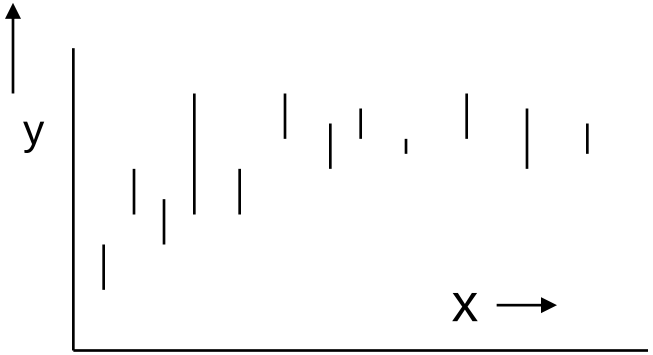
Rate at which this happens depends on……….

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,…….

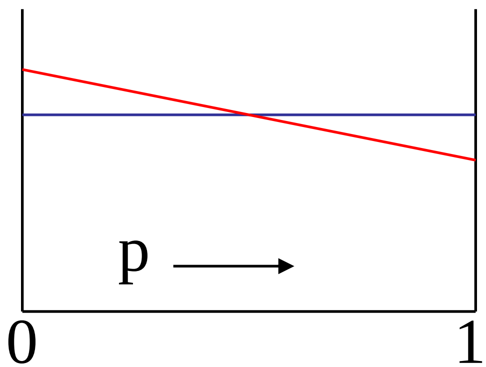
How frequently quadratic term is present

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

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

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) ≠ 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"

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

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%

Aside: Bayes' Theorem

P(A and B) = P(A|B) * P(B) = P(B|A) * P(A)

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

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

Then P(A and B) = P(A) * P(B), but not otherwise

e.g. P(Rainy and Sunday) = P(Rainy)

But P(Rainy and Dec) = P(Rainy|Dec) * P(Dec)

25/365 = 25/31 * 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.

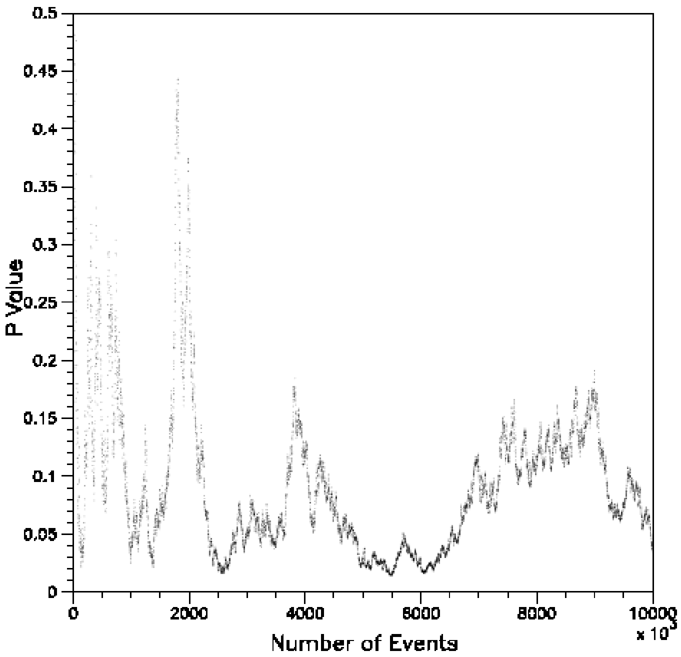
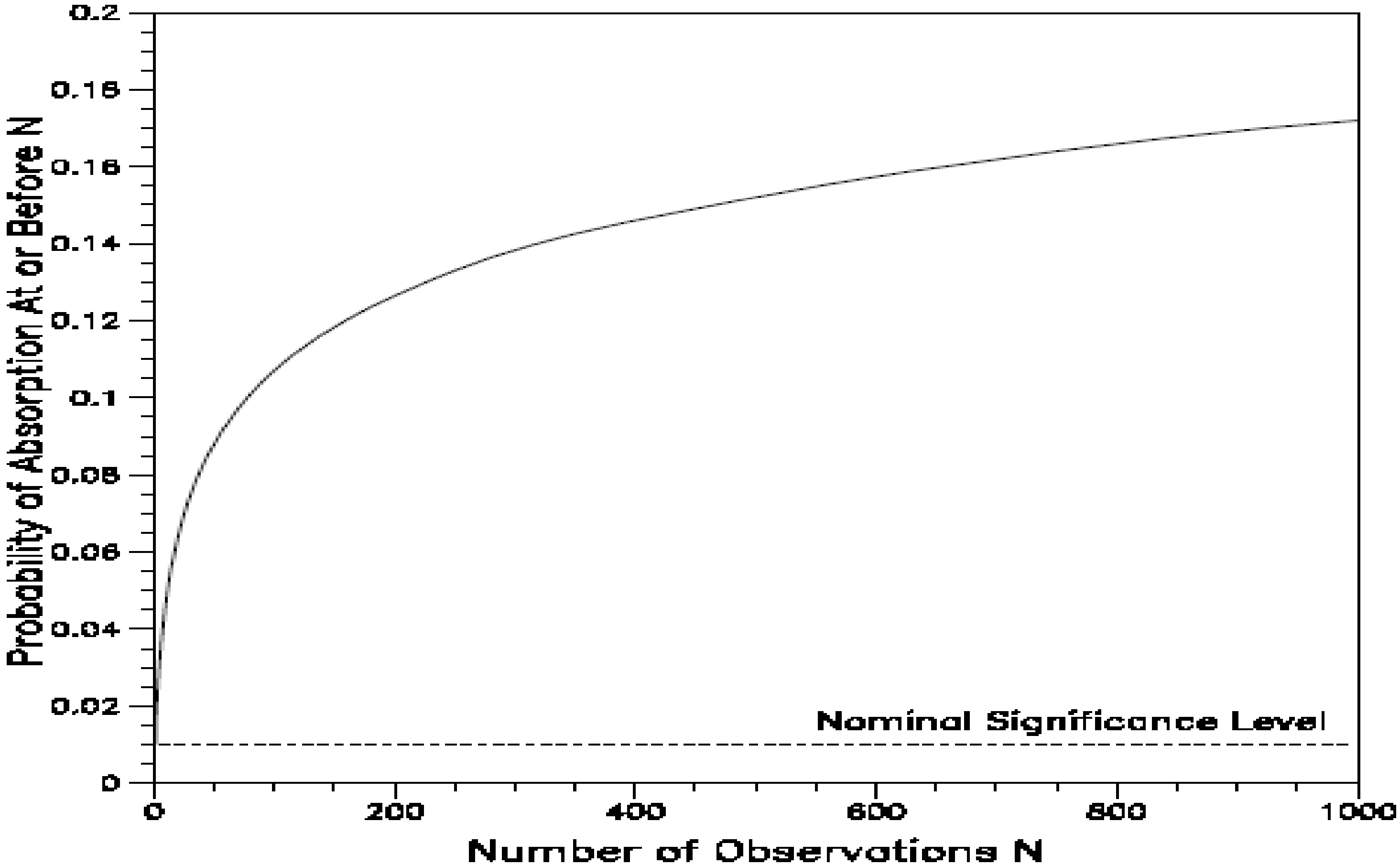


Figure 1: P value versus sample size.

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, $S_{min}(p_0) = 90$

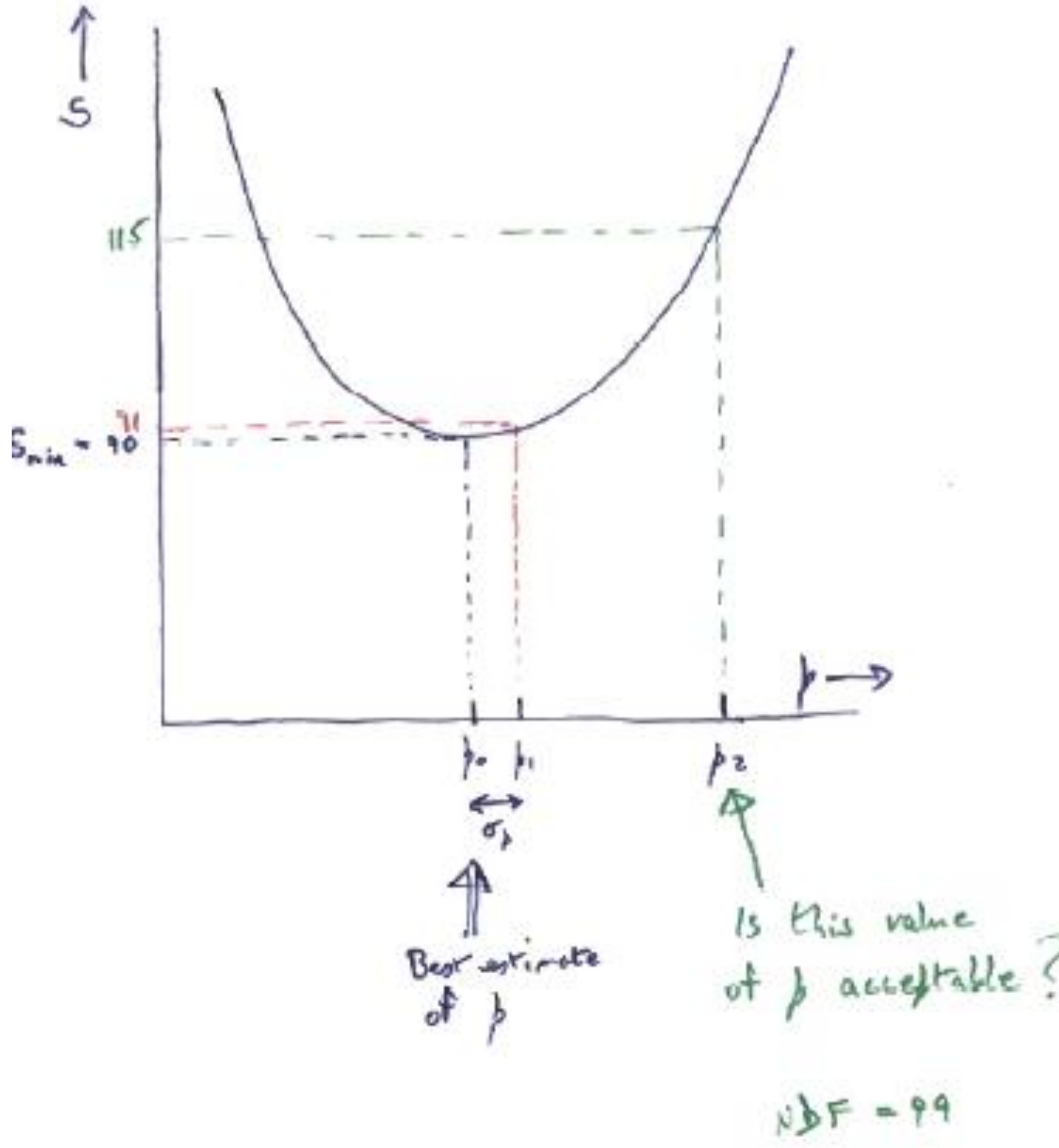
Is p_1 acceptable if $S(p_1) = 115$?

1) YES. Very acceptable χ^2 probability

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

But $S(p_1) – S(p_0) = 25$

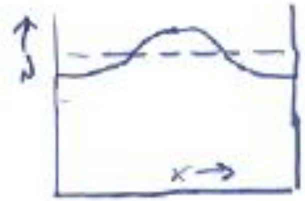
So p_1 is 5σ away from best value


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

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 flucts)

Try fitting according to H1 or to H2
χ_1^2 χ_2^2

Look at dist of χ_1^2 As expected for NDF = 100

χ_2^2 Bit bigger Many* "satisfactory"

$\chi_2^2 - \chi_1^2$ Decision based on $\Delta\chi^2$ has much better power

Repeat for events generated according to H2

Look at dist of χ_1^2

χ_2^2

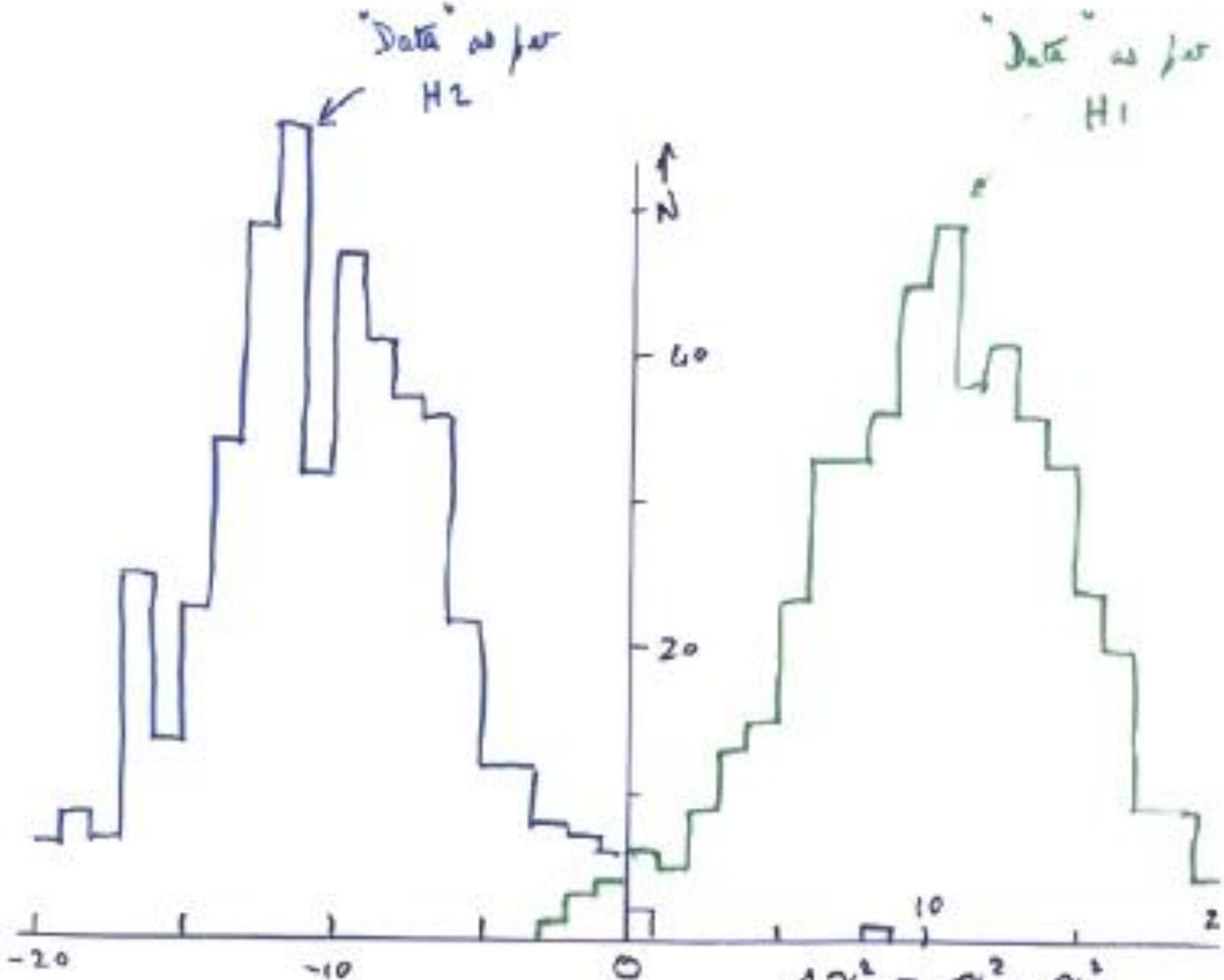
$\chi_2^2 - \chi_1^2$

* 69% have $\chi_2^2 < 130$

2H 6

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

(500 SIMULATIONS)

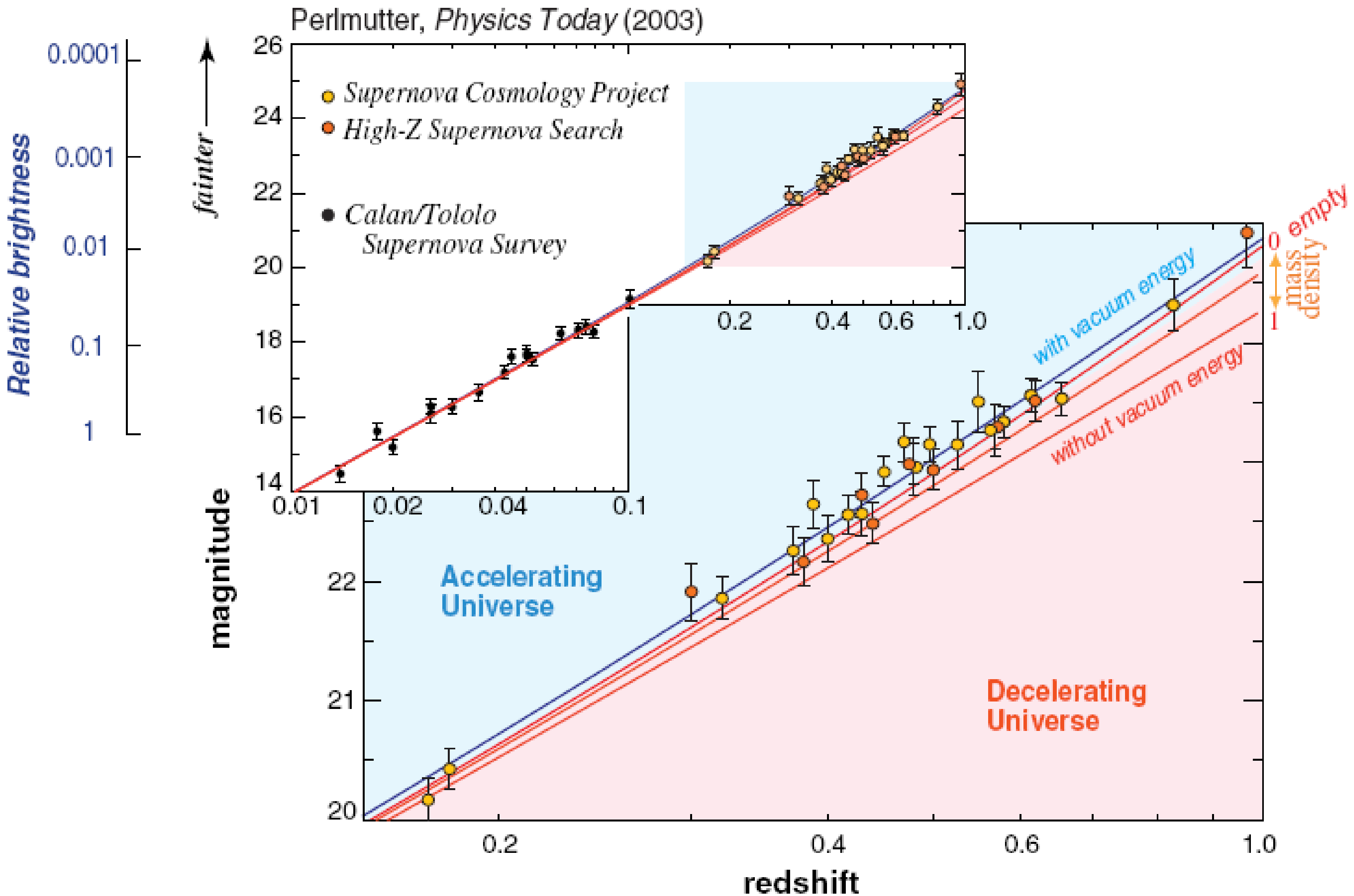


$\Delta\chi^2 = \chi_2^2 - \chi_1^2$

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

$H1 = 1 + 0.05x$

Comparing data with different hypotheses



Choosing between 2 hypotheses

Possible methods:

$\Delta\chi^2$

$ln\mathcal{L}$–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, $ln\mathcal{L}$-ratio)

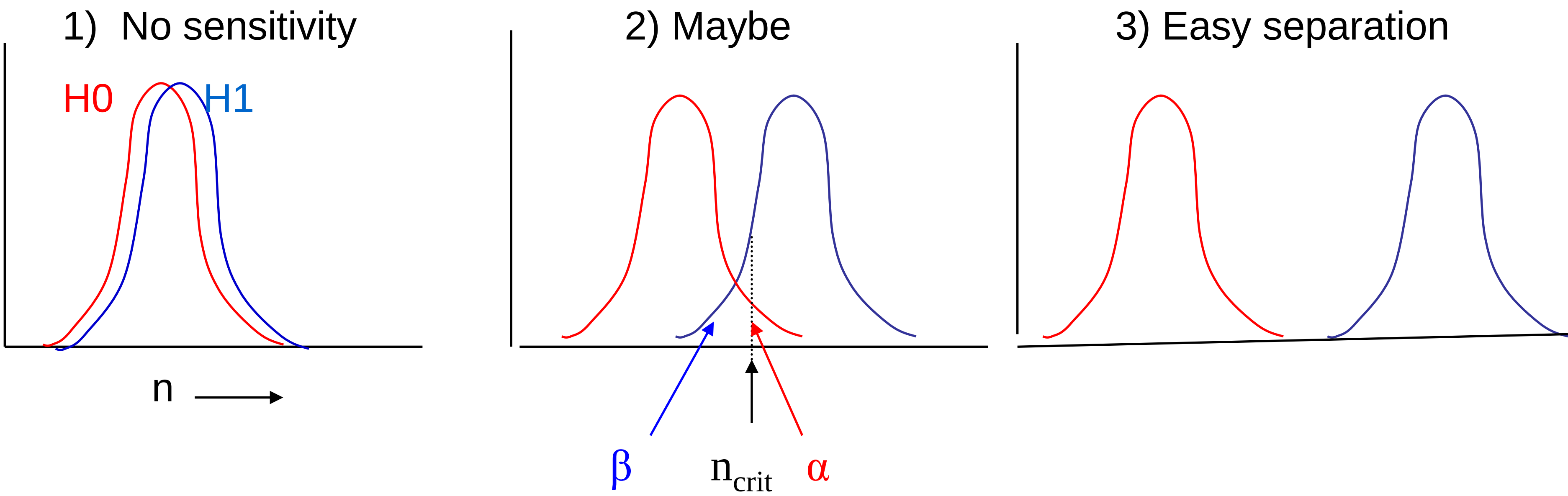
Traditional sensitivity:

Median limit when s=0

Median σ when s ≠ 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 $\beta_{obs} < 1 - \mathrm{CL}$, 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 ± σ_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
- 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