Introduction to Statistical Issues for Phystat-ν

Louis Lyons
Imperial College, London
and
Oxford

Why bother ?

Experiments are expensive and time-consuming, so:

Worth investing effort in statistical analysis

→ better information from data

Statistical Procedures

Parameter Determination

Central value and range

e.g. $\sin^2 2\Theta$

Comparing data with Hypotheses → Discoveries, Upper Limits,…

Just one Hypothesis

Goodness of Fit

e.g. Current 3-neutrino mixing params

Comparing 2 Hypotheses

Hypothesis Testing

e.g. Standard 3ν or 3ν plus sterile ν

ν mass hierarchies

TOPICS

Combining results

A brief reminder about Likelihoods

A few comments on Bayes, Frequentism and Feldman-Cousins

MVA: How NNs work

Some issues related to Discovery claims

Upper Limits

Summary

COMBINING RESULTS

- Better to combine data than combine results

(Problems with non-Gaussian estimates

dealing with correlations

uncertainty estimates)

- Beware of uncertainty estimates that depend on parameter estimate

e.g. $n \pm \sqrt{n}$ 100 ± 10 and 80 ± 9

or $\tau \pm \tau/\sqrt{N}$ 1.00 ± 0.10 and 1.20 ± 0.12 (N=100)

Combining: oddities

- 1 variable :
  Best combination of 2 correlated measurements can be outside range of measurements

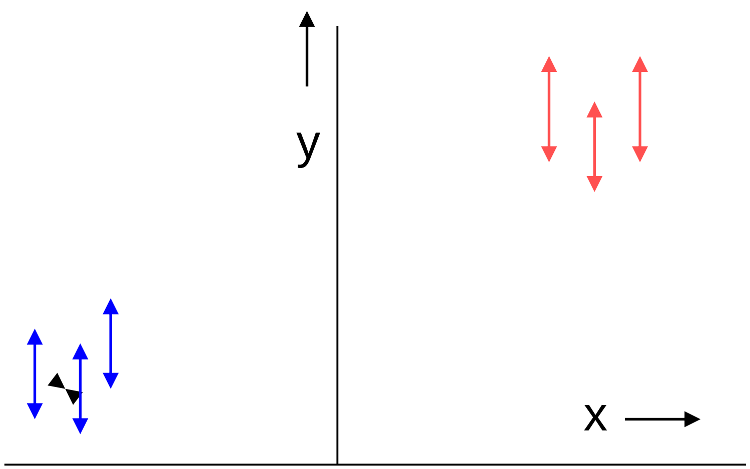
- 2 variables, α β
  Uncertainties on α_{best} and β_{best} much smaller than individual uncertainties.

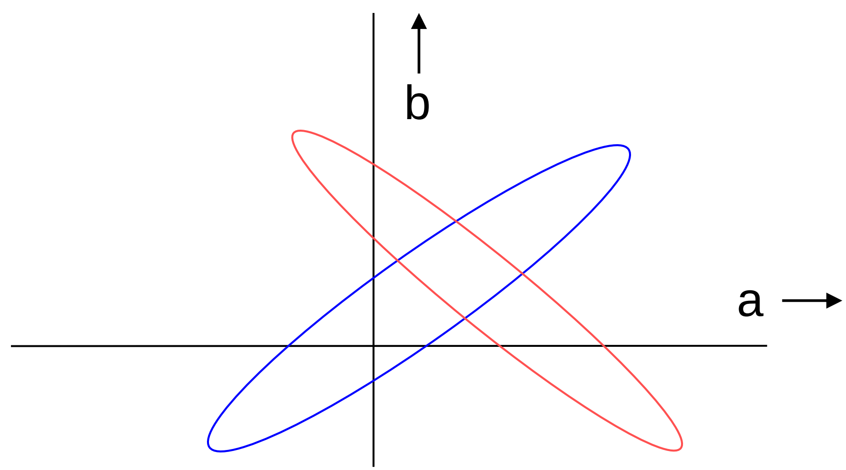
- 2 variables, α β
  $\alpha_{best} > \alpha_1$ and α_2 $\beta_{best} > \beta_1$ and β_2

Straight line fit to red points has large uncertainties on intercept and on gradient

Straight line fit to blue points has large uncertainties on intercept and on gradient

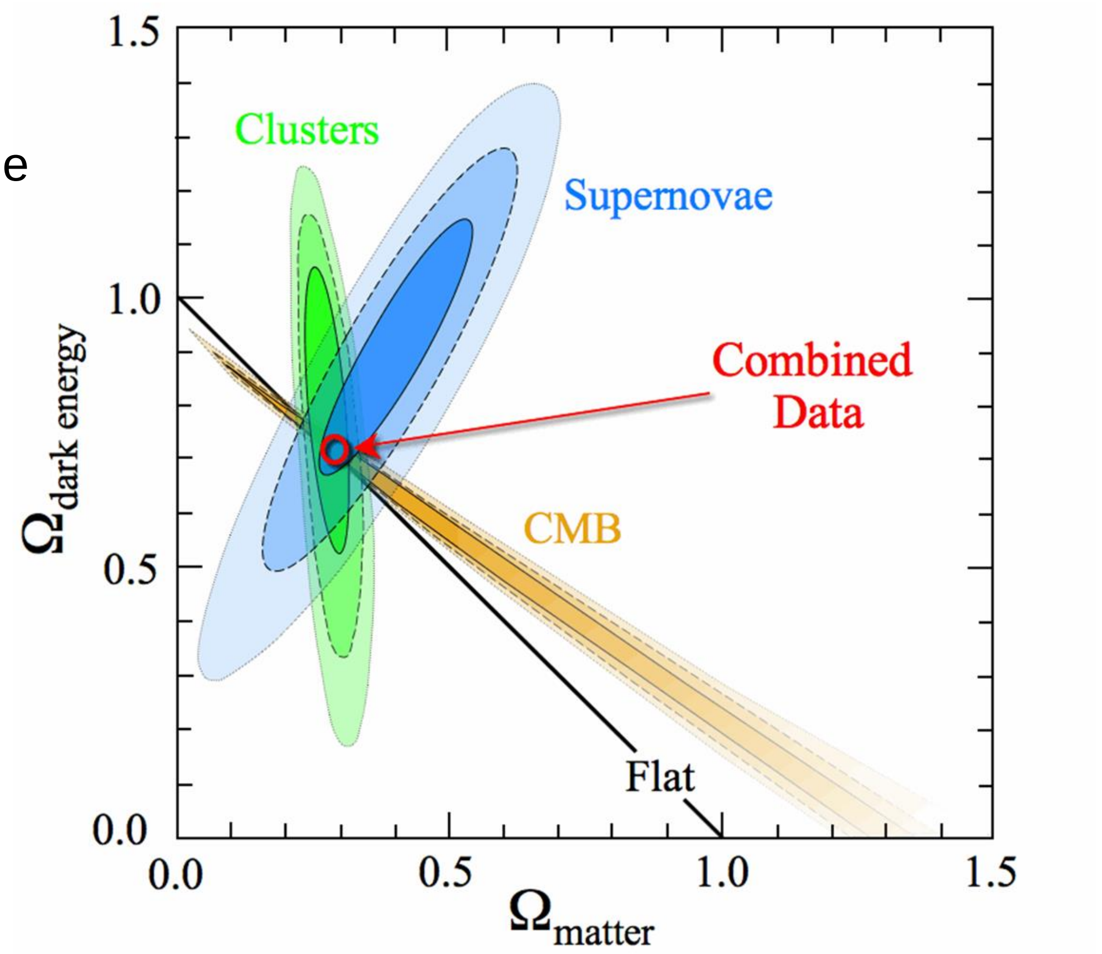
Combined straight line fit to red and blue points has much smaller uncertainties on intercept and on gradient





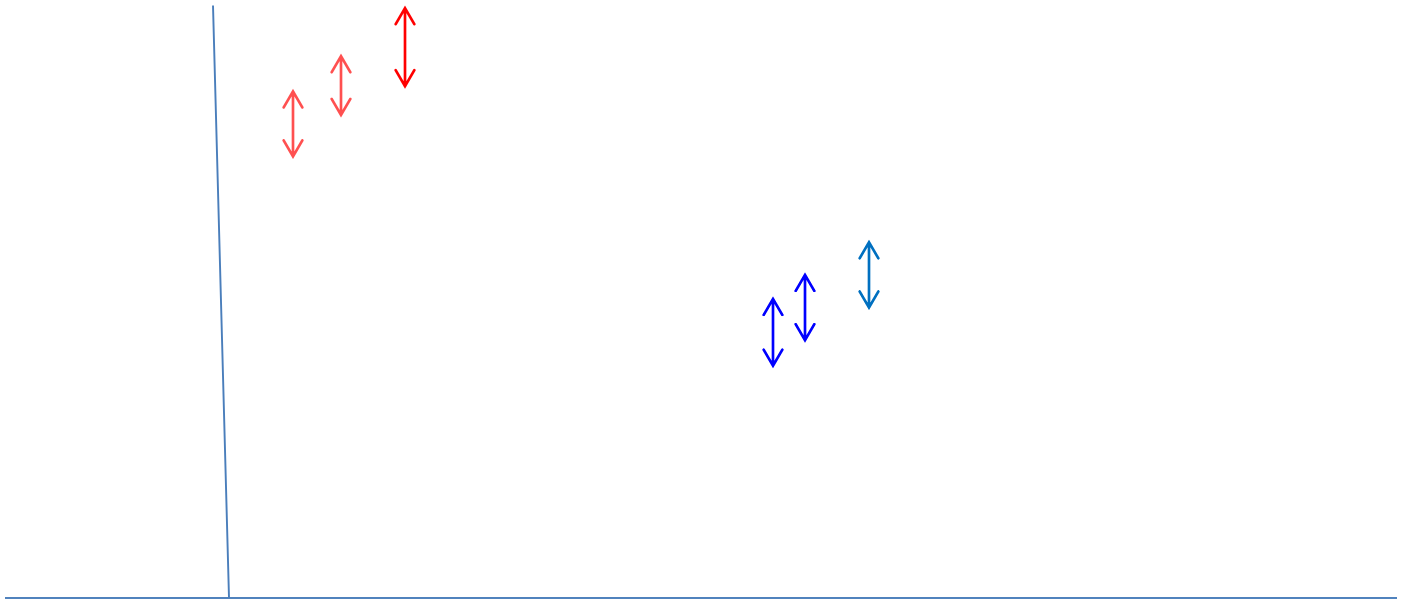
Uncertainty on $\Omega_{dark\ energy}$

When combining pairs of variables, the uncertainties on the combined parameters can be **much** smaller than any of the individual uncertainties
e.g. $\Omega_{dark\ energy}$

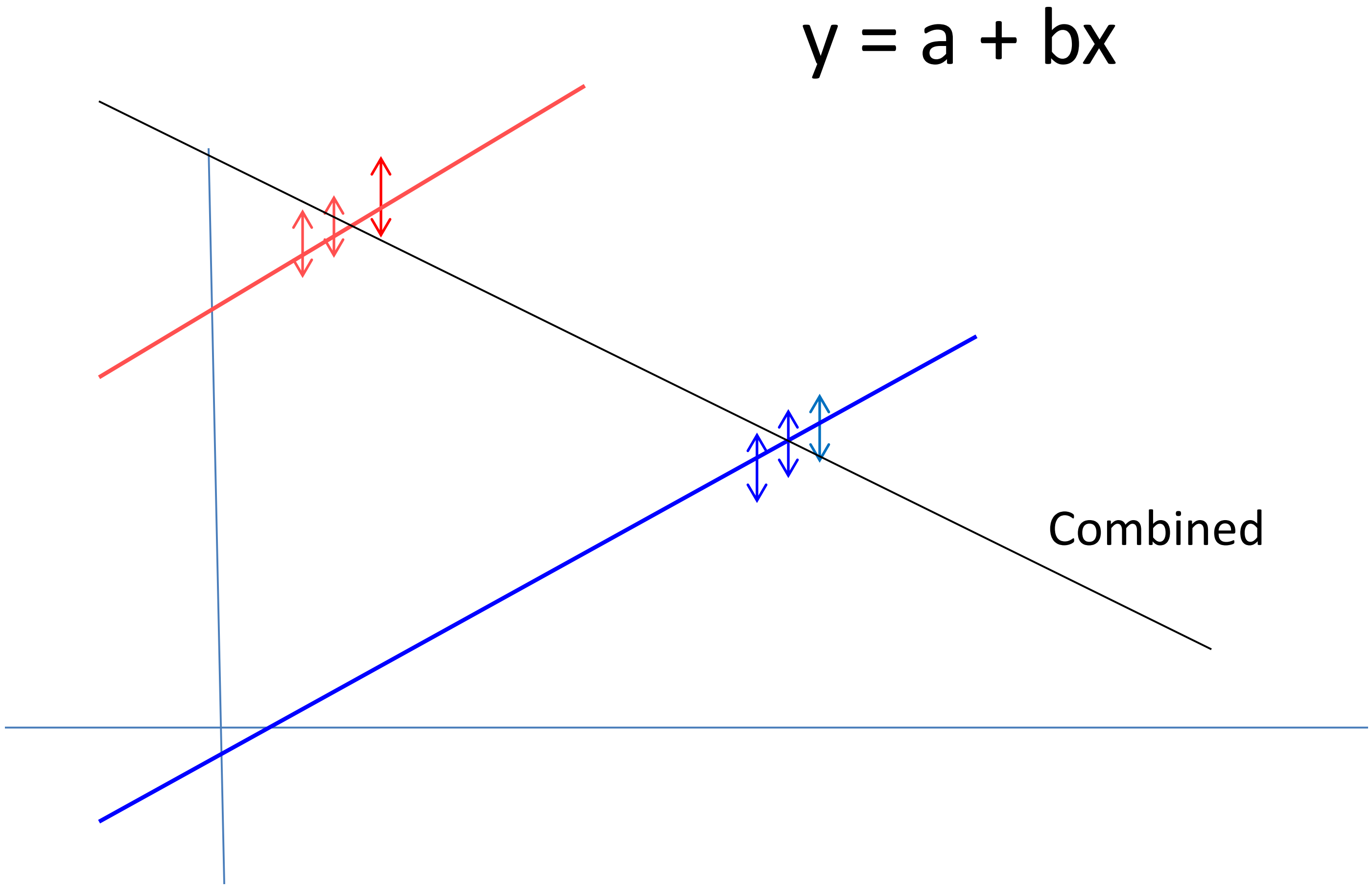


Best values of params a and b outside range of individual values

$y = a + bx$



Best values of params a and b outside range of individual values



Likelihoods

Here just for parameter determination
Also very important for Hypothesis Testing,
in Bayesian and Frequentist approaches

Procedure:
Write down P(data|hypothesis' param)

pdf: Regard this as fn of data, for fixed param values

Likelihood: Fn of parameter, for given data

e.g. Poisson $P(n|\mu) = e^{-\mu} \mu^n/n!$

Data:
Can be individual values. Does not have to be a histogram

Simple example of **Likelihood**: Angular distribution

$y = N\,(1 + \beta \cos^2\theta)$ $\qquad N = 1/\{2(1+\beta/3)\}$

$y_i = N\,(1 + \beta \cos^2\theta_i)$

= probability density of observing θ_i, given β

$\mathcal{L}(\beta) = \Pi\, y_i$

= probability density of observing the data set y_i, given β

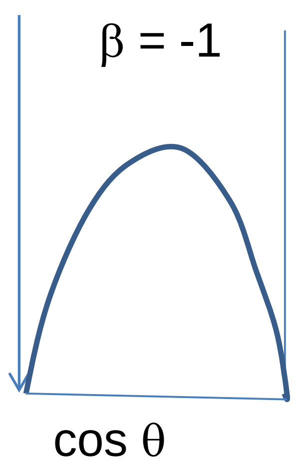
Best estimate of β is that which maximises L

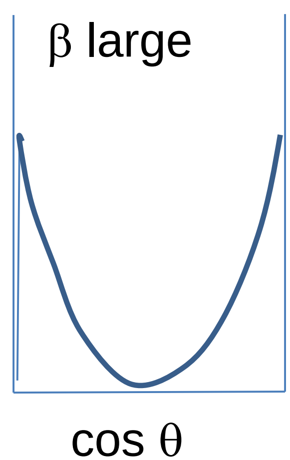
Values of β for which $\mathcal{L}$ is very small are ruled out

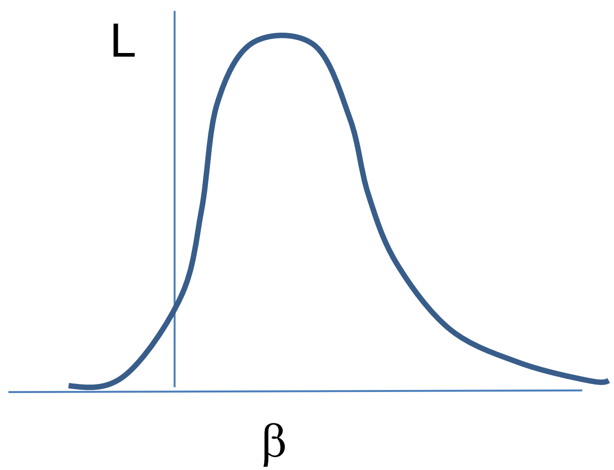
Precision of estimate for β comes from width of L distribution

****** **CRUCIAL** to normalise y $\qquad N = 1/\{2(1 + \beta/3)\}$

(Information about parameter β comes from **shape** of exptl distribution of $\cos\theta$)

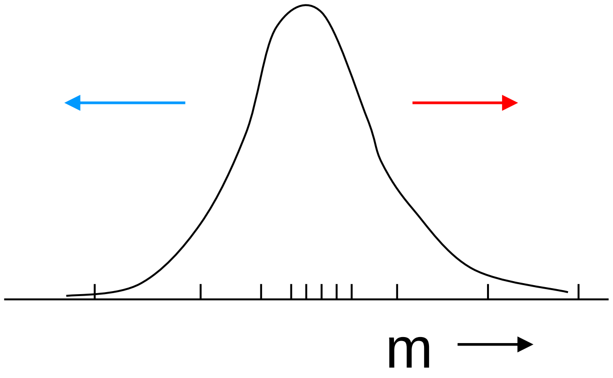


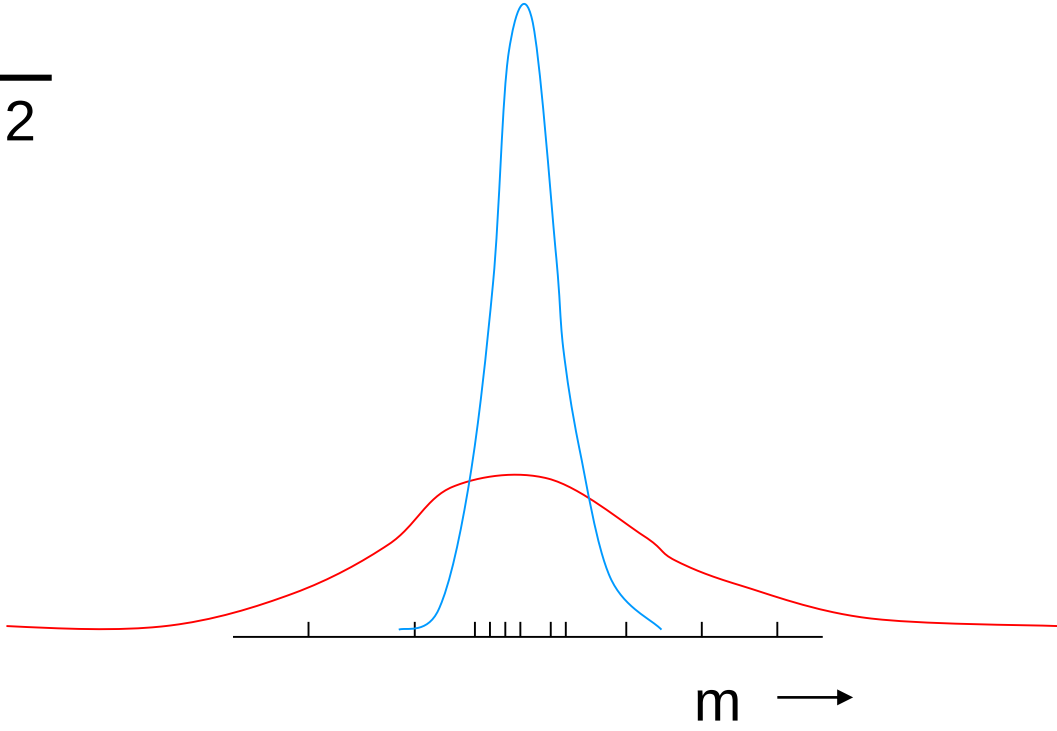




How it works: Resonance

$$y \sim \frac{\Gamma/2}{(m-M_0)^2 + (\Gamma/2)^2}$$





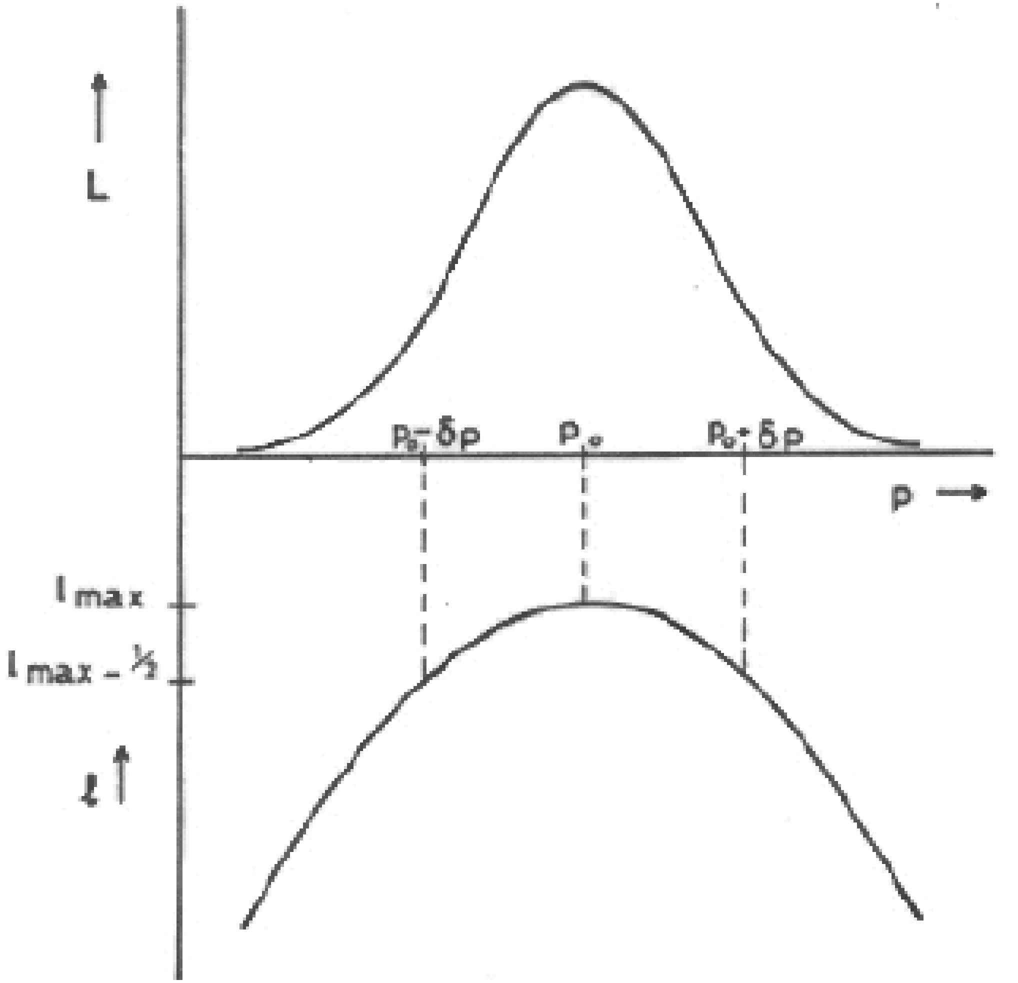
Vary M_0

Vary Γ

Conventional to consider

$$\ell = \ln(\mathcal{L}) = \Sigma \ln(y_i)$$

For large N, $\mathcal{L} \rightarrow$ Gaussian



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

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

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

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

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

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

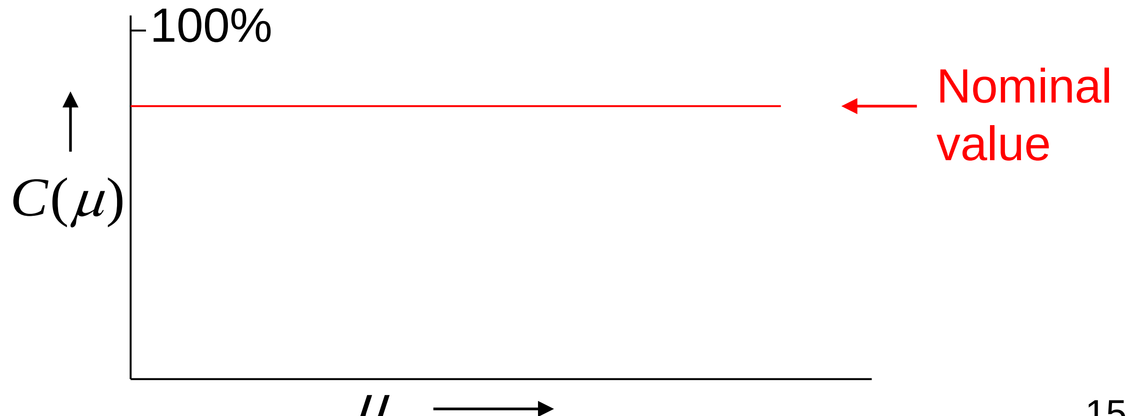
How often does quoted range for parameter include param's true value?

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

Coverage can vary with μ

Study coverage of different methods for 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"

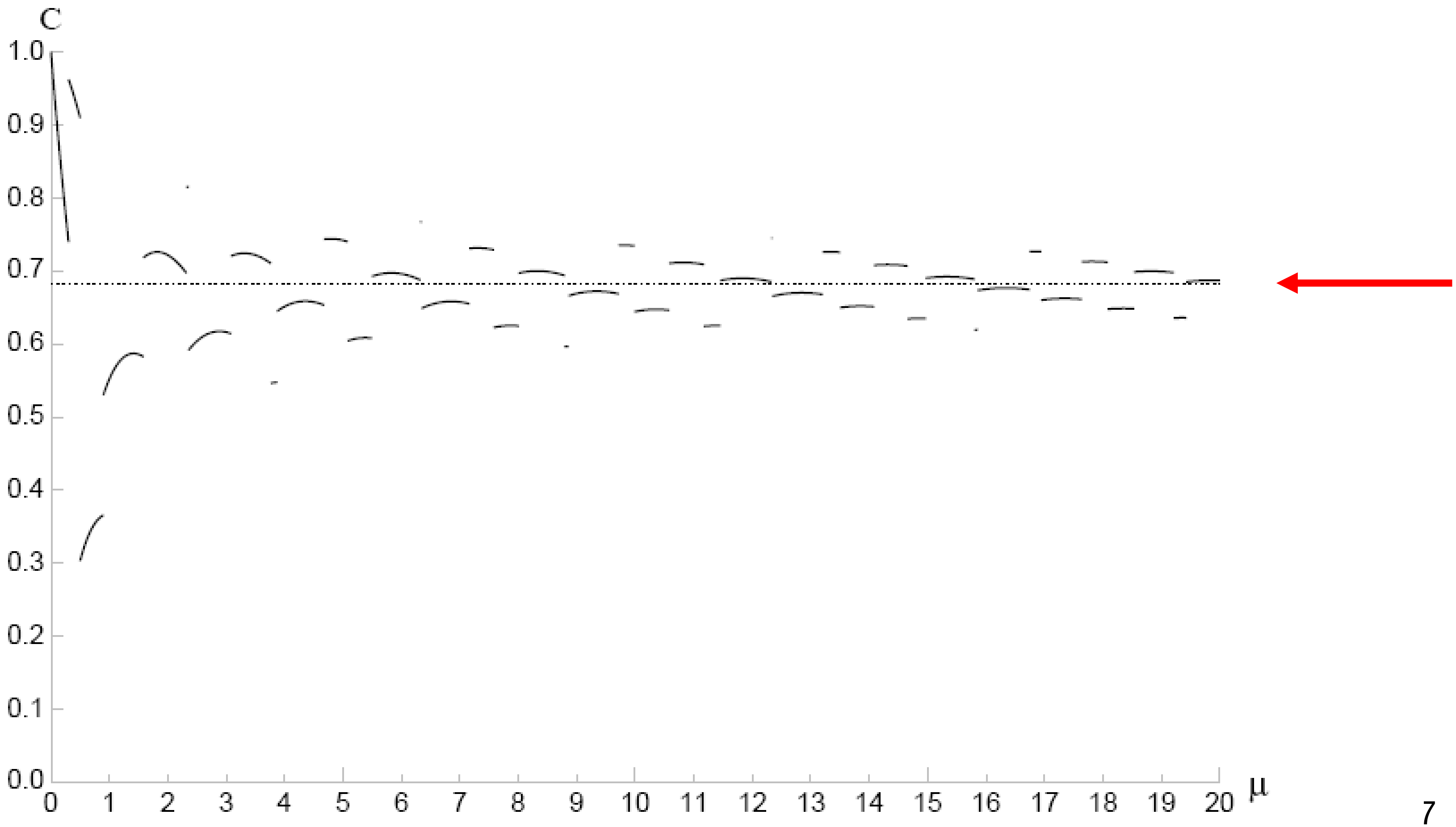
Conservative

Loss of rejection power

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

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

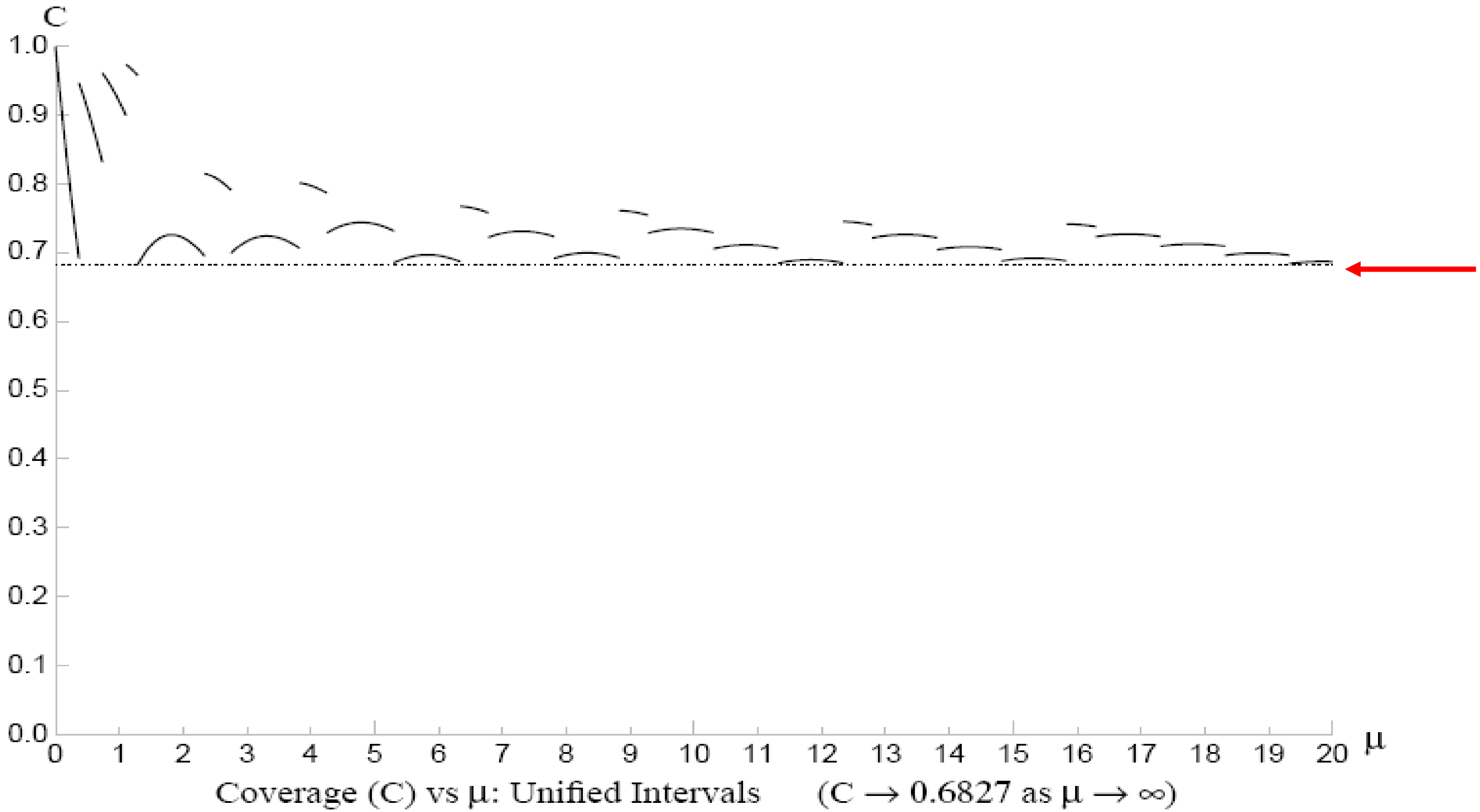
$-2 \ln\lambda < 1$ $\lambda = P(n,\mu)/P(n,\mu_{best})$ UNDERCOVERS



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

Feldman-Cousins Unified intervals

Neyman construction so NEVER undercovers



	Moments	Max Like	Least squares
Easy?	**Yes, if…**	Normalisation, maximisation messy	Minimisation
Efficient?	Not very	Usually best	Sometimes = Max Like
Input	Separate events	**Separate events**	Histogram
Goodness of fit	Messy	No (unbinned)	Easy
Constraints	No	Yes	Yes
N dimensions	Easy if ….	Norm, max messier	Easy
Weighted events	Easy	Errors difficult	Easy
Bgd subtraction	Easy	Troublesome	Easy
Uncertainty estimates	Observed spread, or analytic	$\left\{ \frac{-\partial^2 l}{\partial p_i \partial p_j} \right\}$	$\left\{ \frac{\partial^2 S}{2 \partial p_i \partial p_j} \right\}$
Main feature	Easy	Best for params	**Goodness of Fit**

BAYES and FREQUENTISM: Different views of probability

We need to make a statement about

Parameters, Given Data

The basic difference between the two:

Bayesian : Probability (parameter, given data)
(an anathema to a Frequentist!)

Frequentist : Probability (data, given parameter)
(a likelihood function)

PROBABILITY

MATHEMATICAL

Formal

Based on Axioms

FREQUENTIST

Ratio of frequencies as n → infinity

Repeated "identical" trials

Not applicable to **single event** or **physical constant**

BAYESIAN Degree of belief

Can be applied to single event or physical constant

(even though these have unique truth)

Varies from person to person ***

Quantified by "fair bet"

Bayesian versus Classical

Bayesian

$$P(A \text{ and } B) = P(A;B) \times P(B) = P(B;A) \times P(A)$$

{ If A and B independent, $P(A;B) = P(A)$ → $P(A \text{ and } B) = P(A)\, P(B)$ }

e.g. A = event contains t quark

B = event contains W boson

or A = I am in CERN

B = I am giving a lecture

$P(A;B) = P(B;A) \times P(A) / P(B)$ Bayes' Theorem

Completely uncontroversial, provided….

Bayesian

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

Bayes' Theorem

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

↑ posterior ↑ likelihood ↑ prior

Problems: p(param) Has particular value

"Degree of belief"

Prior What functional form?

Coverage

P(parameter)   Has specific value

"Degree of Belief"

Credible interval

Prior: What functional form?

Uninformative prior: flat?

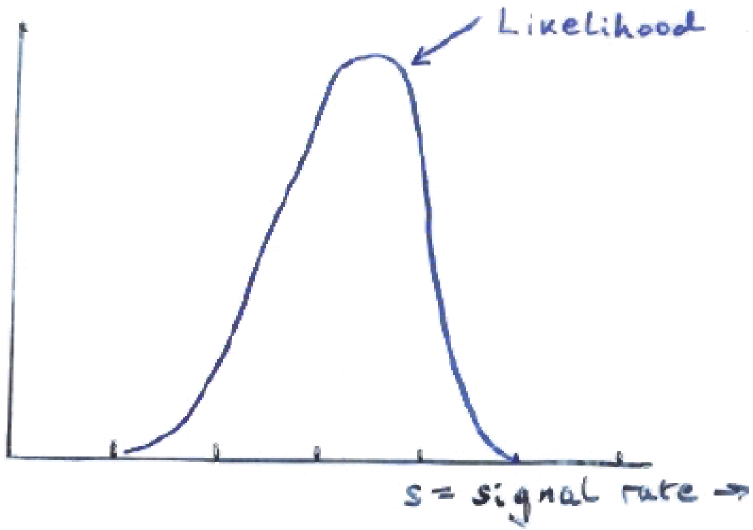
In which variable?   e.g. m, m^2, ln m,….?

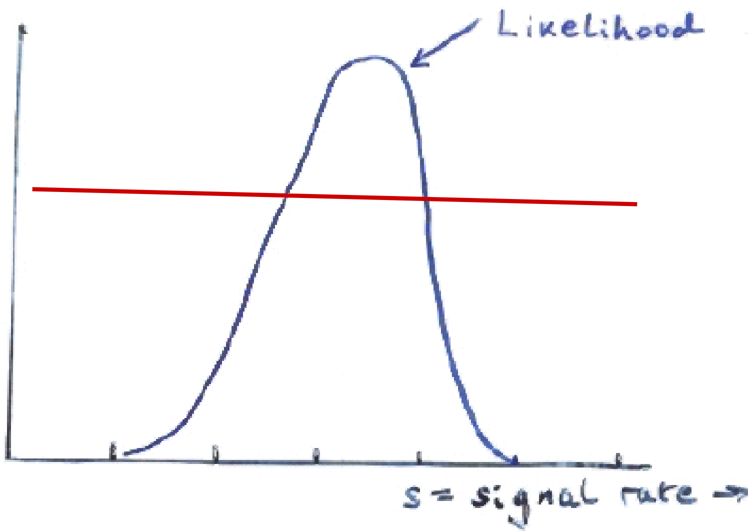
Even more problematic with more params

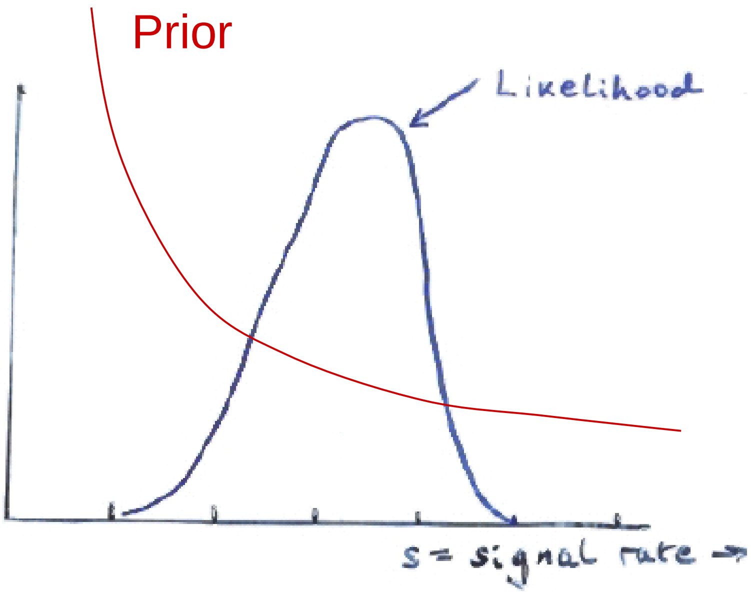
Unimportant if "data overshadows prior"

Important for limits

Subjective or Objective prior?

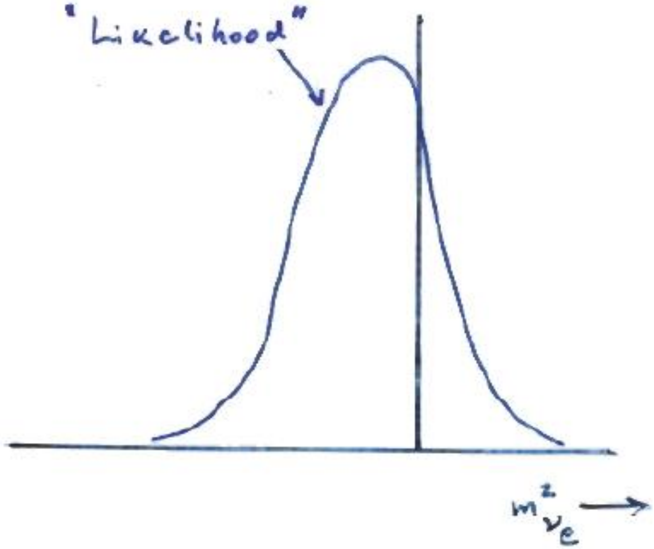


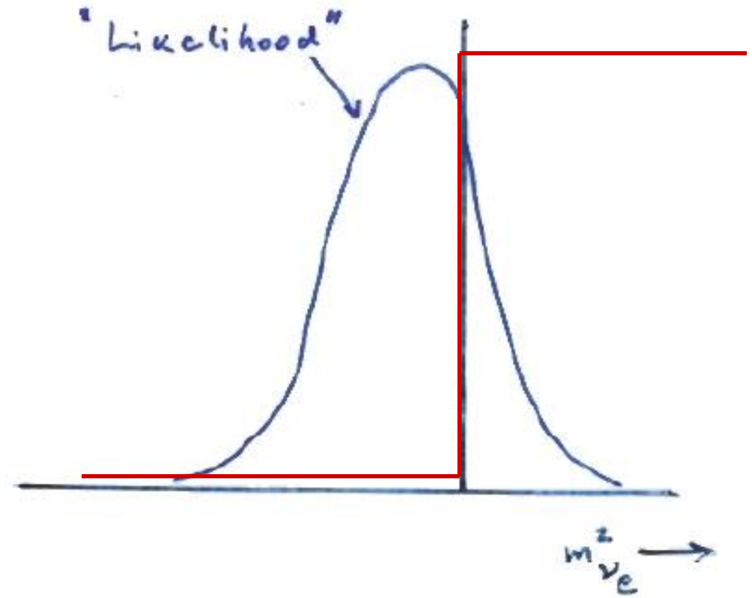




Even more important for UPPER LIMITS

Mass-squared of neutrino





Prior = zero in unphysical region

Bayes: Specific example

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

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

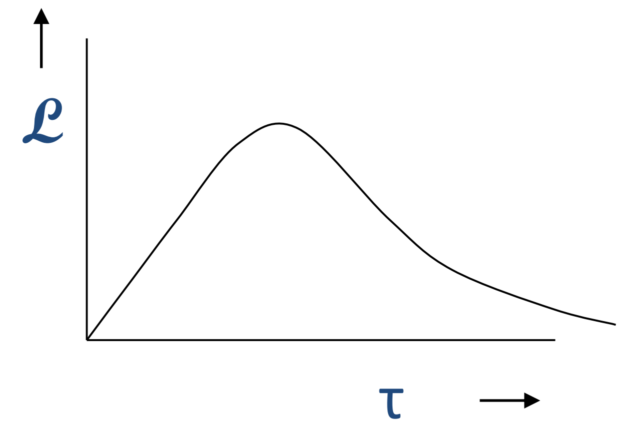
Choose prior $\pi(\tau)$ for τ

e.g. constant up to some large τ

Then posterior $p(\tau) = \mathcal{L}(\tau) * \pi(\tau)$

has almost same shape as $\mathcal{L}(\tau)$

Use $p(\tau)$ to choose interval for τ in usual way



Contrast frequentist method for same situation later.

Classical Approach

Neyman "confidence interval" avoids pdf for μ

Uses only P(x; μ)

Confidence interval $\mu_1 \rightarrow \mu_2$:

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

Varying intervals from ensemble of experiments

fixed

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

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

Classical (Neyman) Confidence Intervals

Uses only P(data|theory)

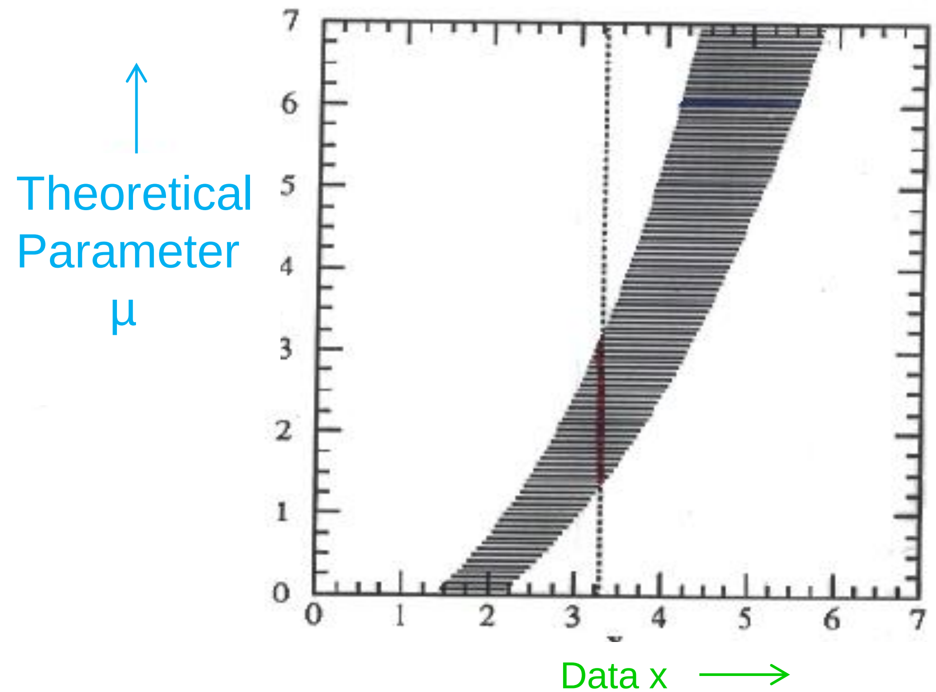


FIG. 1. A generic confidence belt construction and its use. For each value of μ, one draws a horizontal acceptance interval $[x_1, x_2]$ such that $P(x \in [x_1, x_2] | \mu) = \alpha$. Upon performing an experiment to measure x and obtaining the value x_0, one draws the dashed vertical line through x_0. The confidence interval $[\mu_1, \mu_2]$ is the union of all values of μ for which the corresponding acceptance interval is intercepted by the vertical line.

Example:

Param = Temp at centre of Sun

Data = Est. flux of solar neutrinos

$\text{Prob}(\mu_l < \mu < \mu_u) = \alpha$

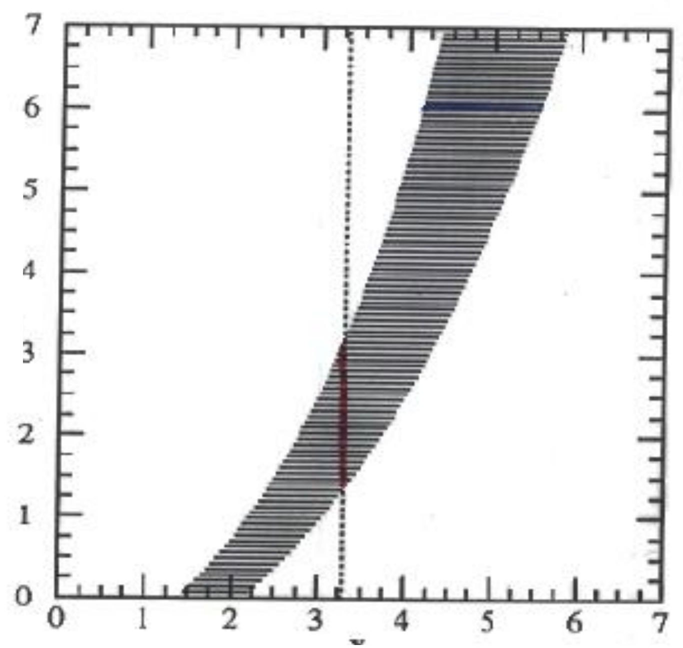
$\mu \geq 0$

No prior for μ

Classical (Neyman) Confidence Intervals

Uses only P(data|theory)

Theoretical Parameter µ ↑



Data x ⟶

FIG. 1. A generic confidence belt construction and its use. For each value of μ, one draws a horizontal acceptance interval $[x_1, x_2]$ such that $P(x \in [x_1, x_2] | \mu) = \alpha$. Upon performing an experiment to measure x and obtaining the value x_0, one draws the dashed vertical line through x_0. The confidence interval $[\mu_1, \mu_2]$ is the union of all values of μ for which the corresponding acceptance interval is intercepted by the vertical line.

Data x	**µ range**
<1.5	Empty
1.5 – 2.2	Upper limit
>2.2	2-sided

Example:

Param = Temp at centre of Sun

Data = est. flux of solar neutrinos

µ≥0 No prior for µ

90% Classical interval for Gaussian

σ = 1 $\mu \geq 0$

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

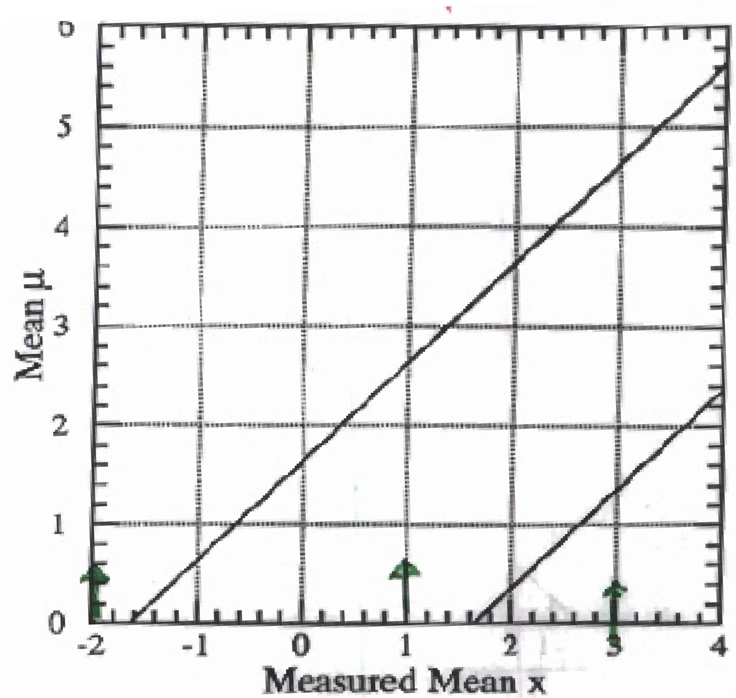


FIG. 3. Standard confidence belt for 90% C.L. central confidence intervals for the mean of a Gaussian, in units of the rms deviation.

x_{obs}=3 Two-sided range
x_{obs}=1 Upper limit
x_{obs}=-1 No region for μ

Other methods have different behaviour at negative x

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

Feldman-Cousins 90% conf intervals

Uses different ordering rule

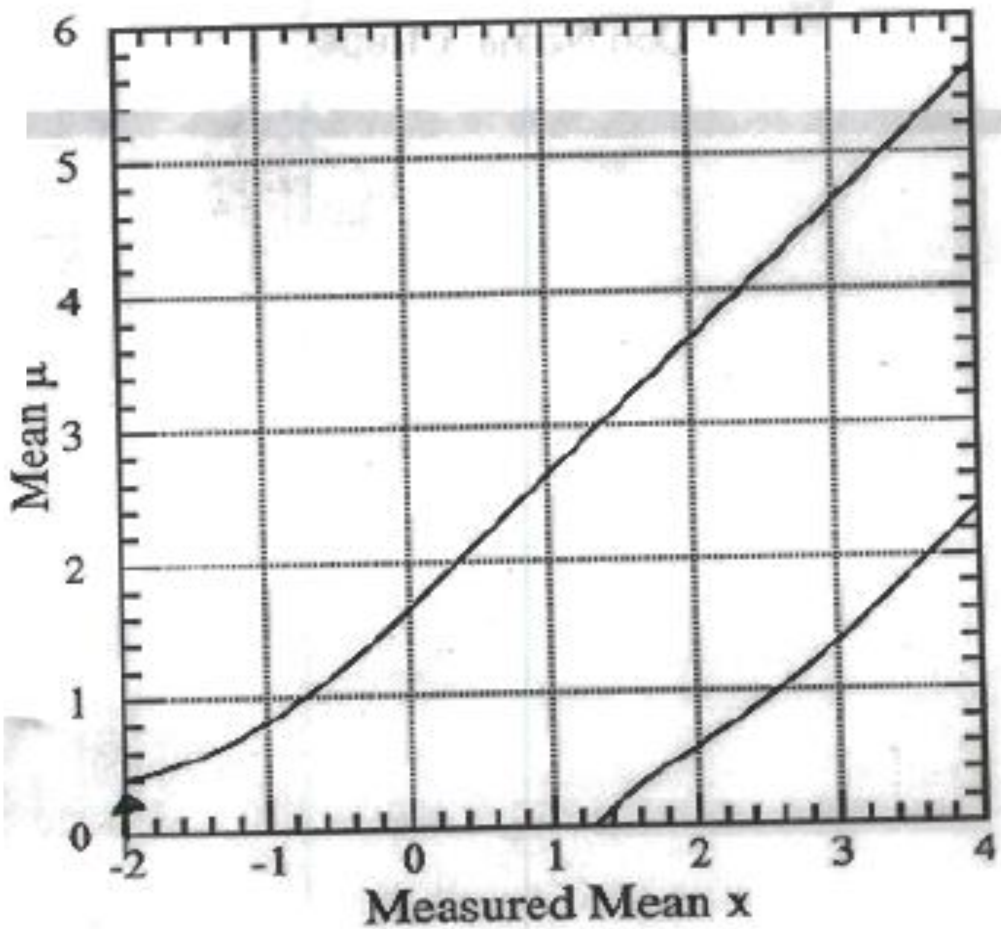


FIG. 10. Plot of our 90% confidence intervals for mean of a Gaussian, constrained to be non-negative, described in the text.

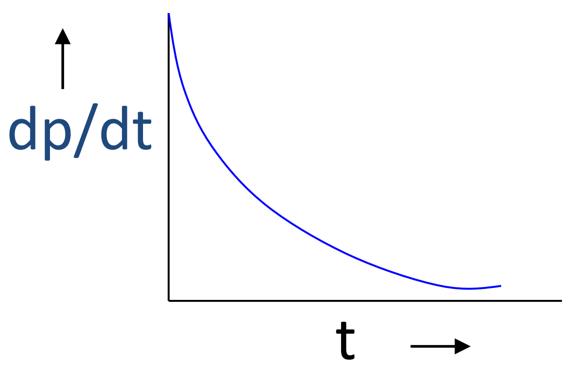
X_{obs} = -2 now gives upper limit

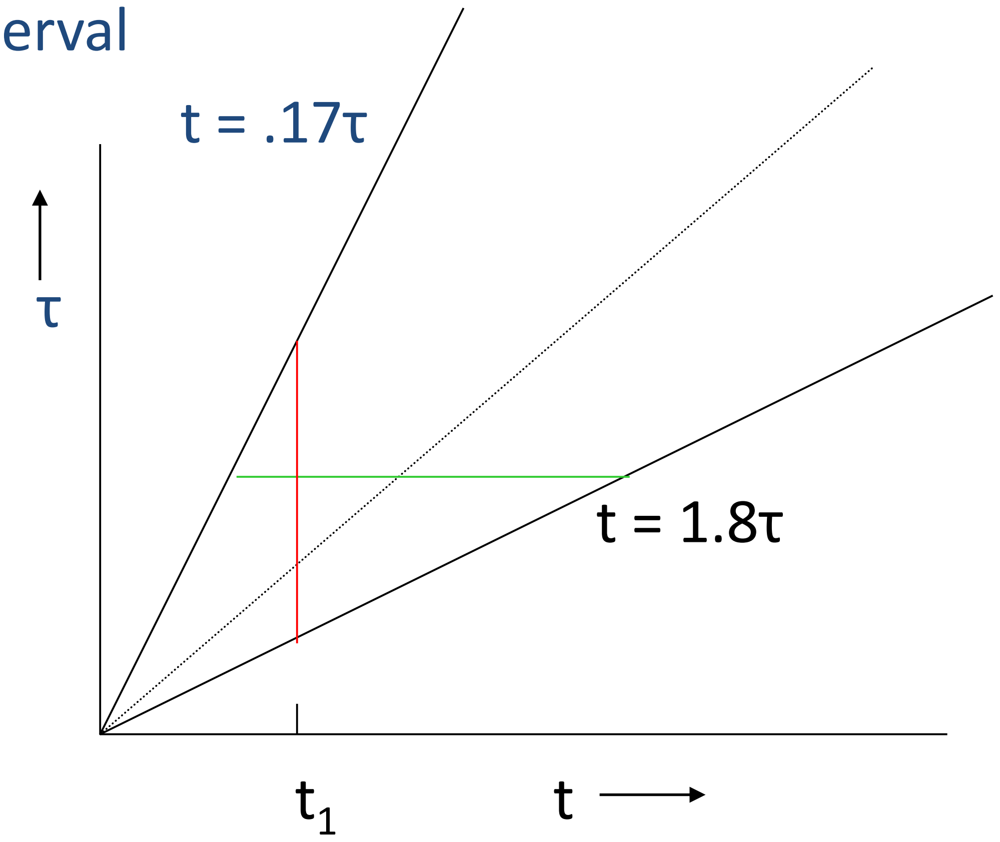
Frequentism: Specific example

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

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

Construct 68% central interval





68% conf. int. for τ from

$t_1/1.8 \rightarrow t_1/0.17$

$\mu_l \le \mu \le \mu_u$ at 90% confidence

Frequentist

μ_l and μ_u known, but random

μ unknown, but fixed

Probability statement about μ_l and μ_u

Bayesian

μ_l and μ_u known, and fixed

μ unknown, and random

Probability/credible statement about μ

MULTIVARIATE ANALYSIS

Example: Aim to separate signal from background

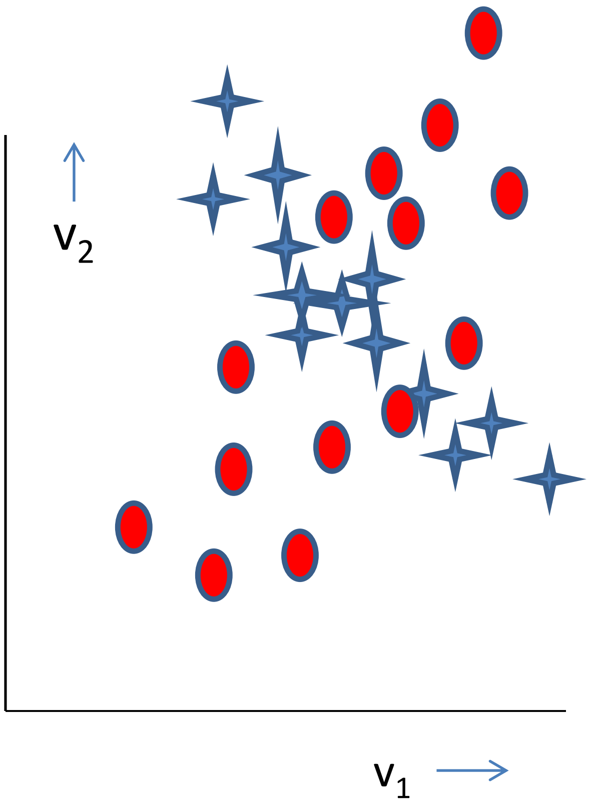
Neyman-Pearson Lemma:
Imagine all possible contours that select signal with efficiency ε (Loss = Error of 1st Kind)
Best is one containing minimal amount of background (Contamination = Error of 2nd Kind)

Equivalent to ordering data by

$$\mathcal{L}\text{-ratio} = \mathcal{L}_s(v_1, v_2, \ldots\ldots.) / \mathcal{L}_b(v_1, v_2, \ldots)$$

IF variables are independent

$$\mathcal{L}\text{-ratio} = \{\mathcal{L}_s(v_1)/\mathcal{L}_b\{v_1)\} \times \{\mathcal{L}_s(v_2)/\mathcal{L}_b(v_2)\} \times \ldots..$$



PROBLEM:

Don’t know $\mathcal{L}$-ratio exactly because:

1) Signal & bdg generated by M.C. with finite statistics
2) Nuisance params (systematics) and signal params
3) Neglected sources of bgd
4) Hard to implement in many dimensions

METHODS TO DEAL WITH THIS

Cuts

Kernel Density Estimation

Fisher Discriminant

Principal Component Analysis

Boosted Decision Trees

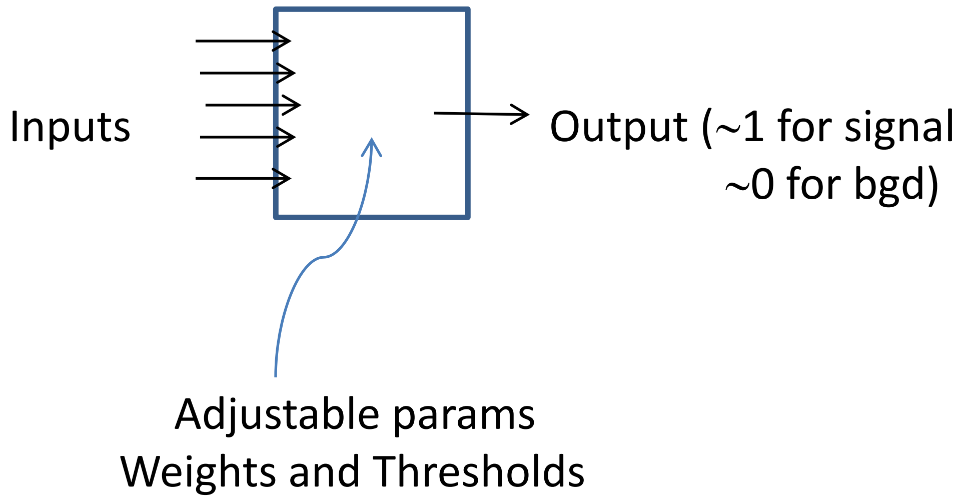
Support Vector Machines

Neural Nets

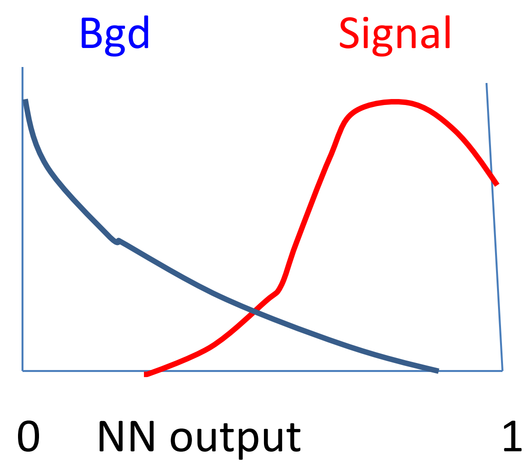
Deep Nets

NEURAL NETWORKS

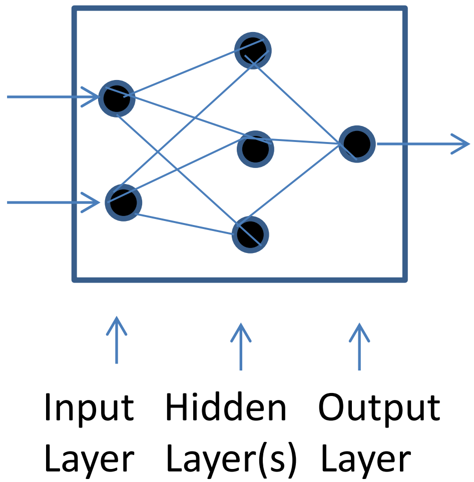
Typical application: Classify events as signal or bgd



- Learning process:
  Input = Known signal & bgd (e.g M.C.)
  Adjust params → 'Best' output
- Testing process
  Make sure not 'overtraining'
- Use trained network on actual data
  Classify events as signal if output > cut



HOW DOES IT WORK?



For each hidden or output node j

$\text{Output}_j = F\ [\Sigma\ \text{Input}_i * W_{ij} + T_j]$

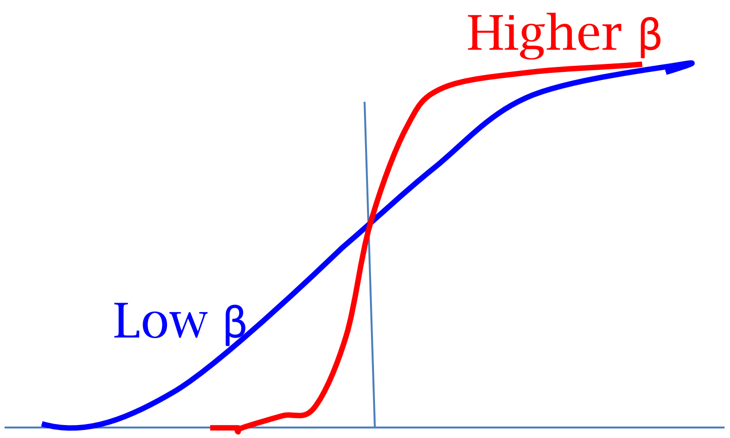
(W and T = network params)

Typical $F(x) = 1/(1+ e^{-\beta x})$ Sigmoid

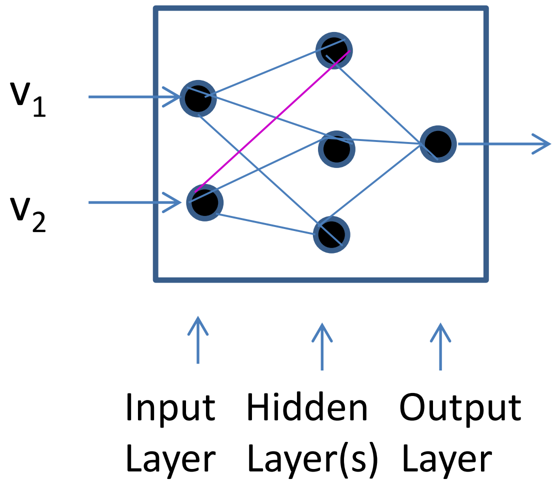
For large β, output of node j is 'ON' if

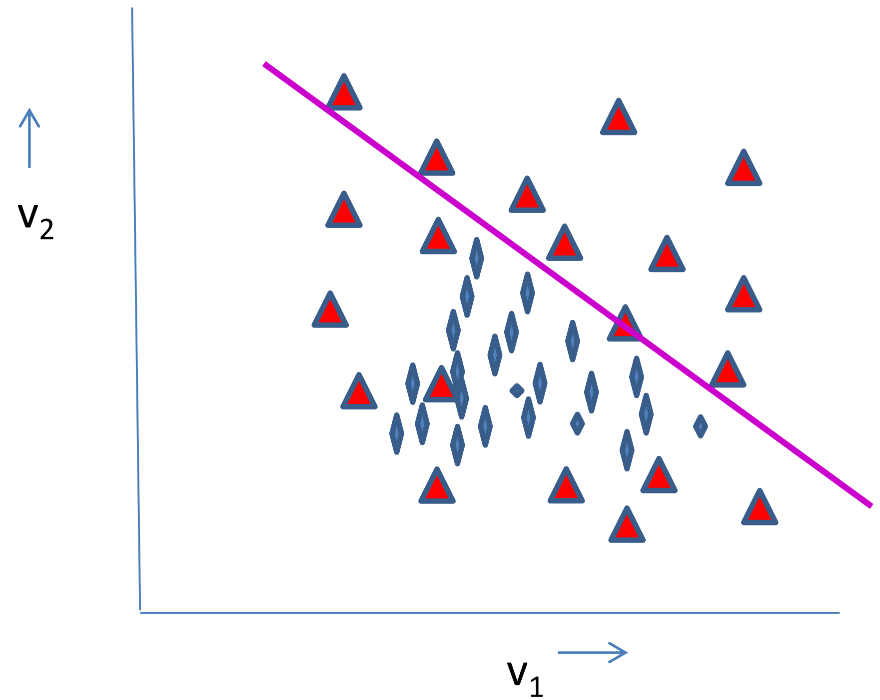
$\Sigma\ I_i\, w_{ij} + T_j > 0$

This is 'hyper-plane' in I space ●



HOW DOES IT WORK?





For First hidden node

Straight line is

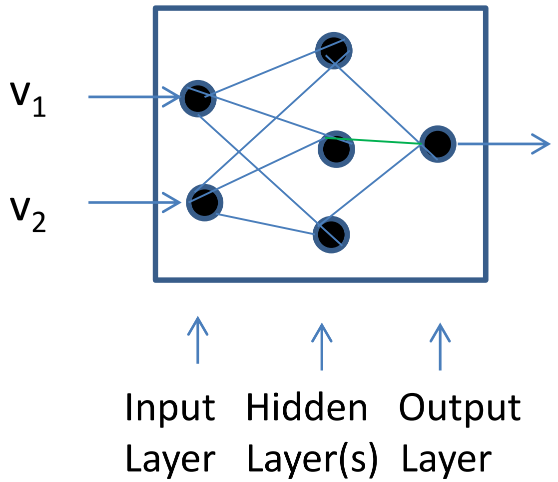
$$w_{\mathbf{11}} * v_1 + w_{21} * v_2 + T_{10} = 0$$

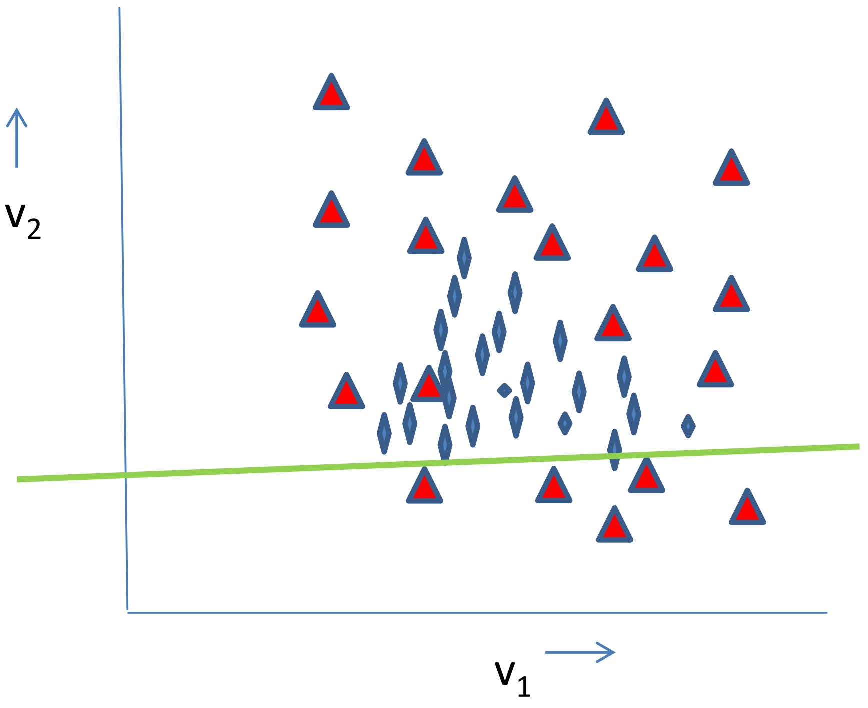
where

w_{ij} is weight from i^{th} input node to j^{th} hidden node

T_{k0} is threshold for k^{th} hidden node

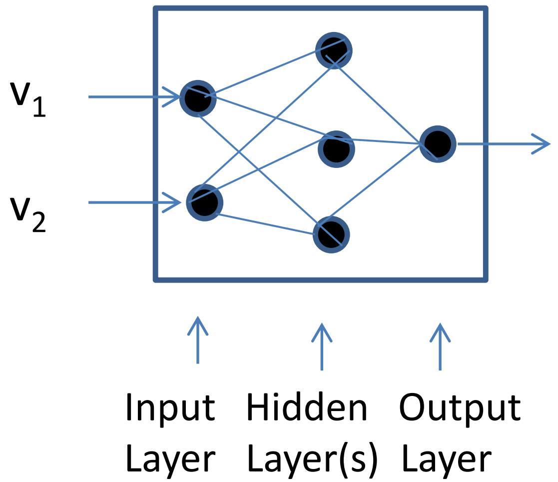
HOW DOES IT WORK?



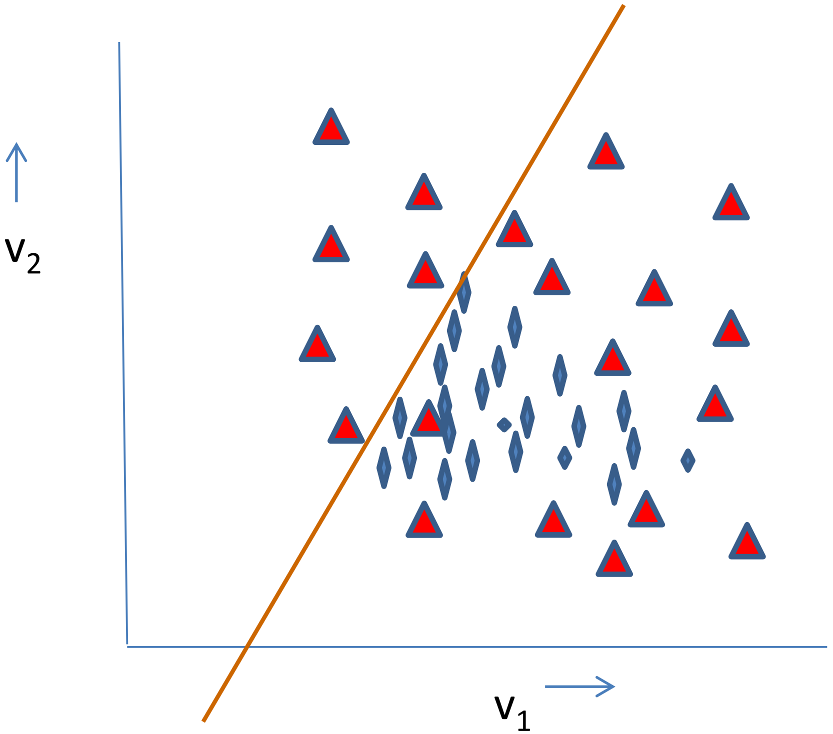


For second hidden node

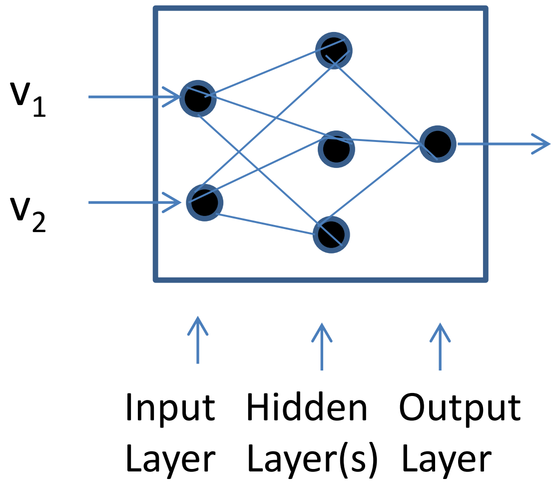
HOW DOES IT WORK?

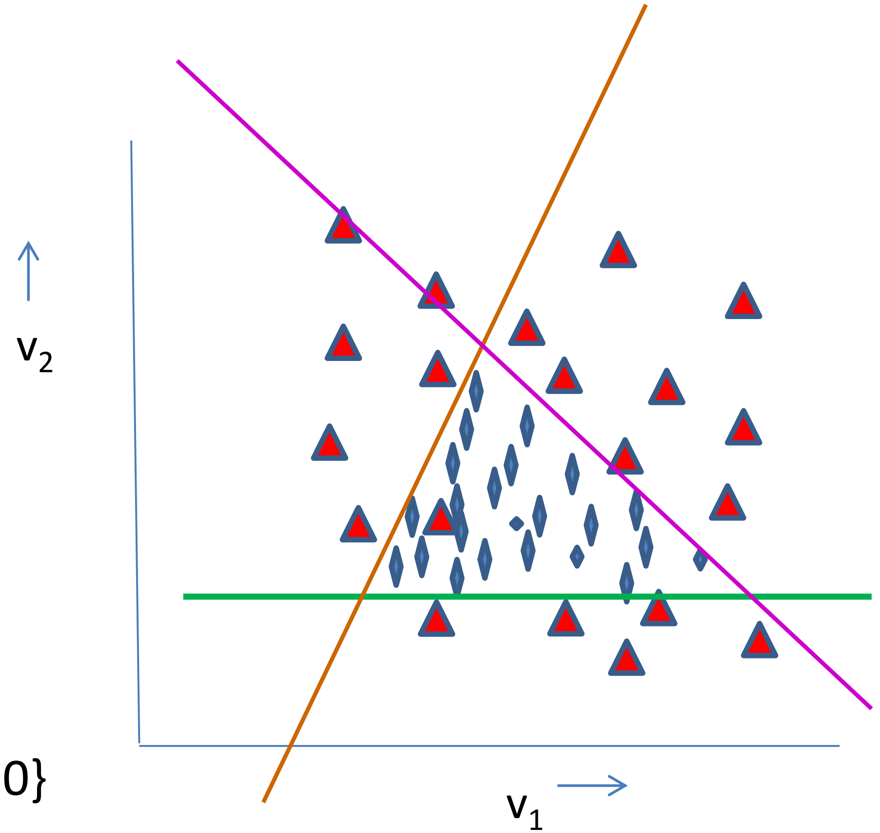


For third hidden node



HOW DOES IT WORK?





Output = Sigmoid{$0.4H_1 + 0.4H_2 + 0.4H_3 - 1.0$}

Output is 'On' only if H_1 H_2 H_3 all are 'On'

N.B.

* Complexity of final region depends on number of hidden nodes.
* Finite β → rounded edges for selected region; and contours of constant output in (v_1, v_2) plane.

BEWARE

- Training sets are reliable
- Don’t train with variable you want to measure
- Data does not extend outside range of training samples (in multi-dimensions)
- Don’t overtrain
- Approx equal numbers of signal and bgd

Is NN better* than simple cuts?

In principle, NO

Can cut on complicated variable e.g. NN output

In practice: YES

But:

Better NN performance → more work by 'Cuts' analysis to improve performance

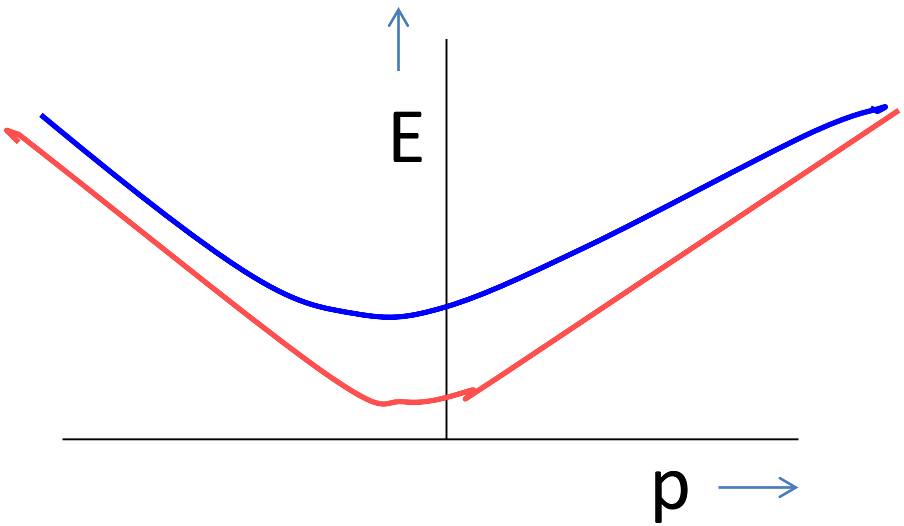
* Better = improved efficiency v mistag rate

SIMPLE EXAMPLE

Try to separate π and proton using E and p

$\pi: E^2 = p^2 + m_\pi^2$

$P: E^2 = p^2 + m_p^2$



Easy: $p = 0 \rightarrow 2$ GeV

Harder: $p = -4 \rightarrow 4$ GeV

Hardest: $p_x, p_y, p_z = -4 \rightarrow 4$ GeV

More realistic: Add expt scatter of data wrt curves

PHYSICS EXAMPLE

Separate b-jets from light flavour, gluons, W, Z:

Input variables: Track IPs, SV mass, distance, quality, etc.

Output: 0 → 1

Issues:

Pre NN cuts

Training and testing samples (Where from? How many events? Ratios of different bgds,….)

How many inputs?

Network structure

How many networks?

Single output or several

Systematics (use different sets of testing events}

Stability wrt NN cut

NN Summary

- ADVANTAGES:
  Very flexible
  Correlations OK
  Tunable cut
- DISADVANTAGES
  Training takes time
  Tendency to include too many variables
  Treat as black box

* Past attitude: Need to convince colleagues NN is sensible
More recently: Why aren’t you using NN?
Now/future: Why aren’t you using a Deep Network?

Choosing between 2 hypotheses

Possible methods:

$\Delta\chi^2$

p-value of statistic →

ln$\mathcal{L}$–ratio

Bayesian:

- Posterior odds
- Bayes factor
- Bayes information criterion (BIC)
- Akaike …….. (AIC)

Minimise "cost"

See 'Comparing two hypotheses'
http://www-cdf.fnal.gov/physics/statistics/notes/H0H1.pdf

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

Topics:

- Example of Hypotheses
- H0 or H0 v H1?
- 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
- Upper Limits
- Higgs search: Discovery and spin

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

Examples of Hypotheses

1) Event selector

Selection of event sample based on required features

e.g. H0: Cerenkov ring produced by electron   H1: Produced by other particle

Possible outcomes: Events assigned as H0 or H1

2) Result of experiment

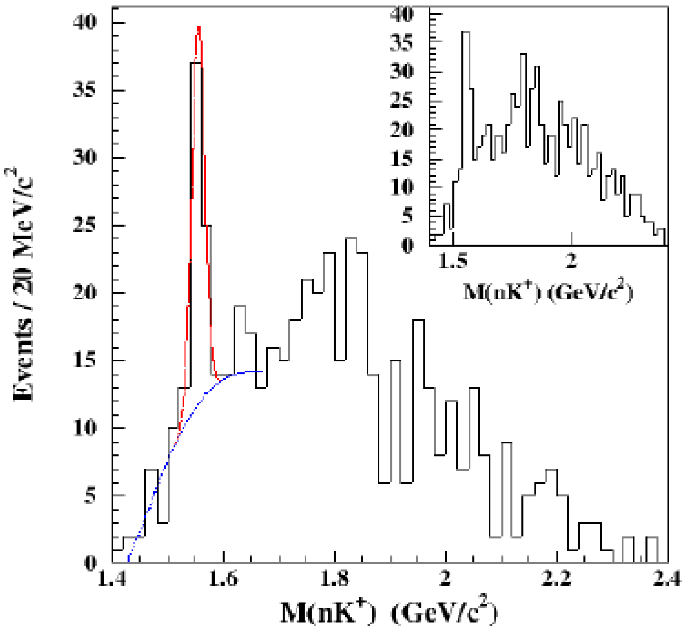
e.g. H0 = nothing new

H1 = new particle produced as well

(Sterile neutrino,…..)

Possible outcomes	H0	H1	
	✓	X	Exclude H1
	X	✓	Discovery
	✓	✓	No decision
	X	X	?

.



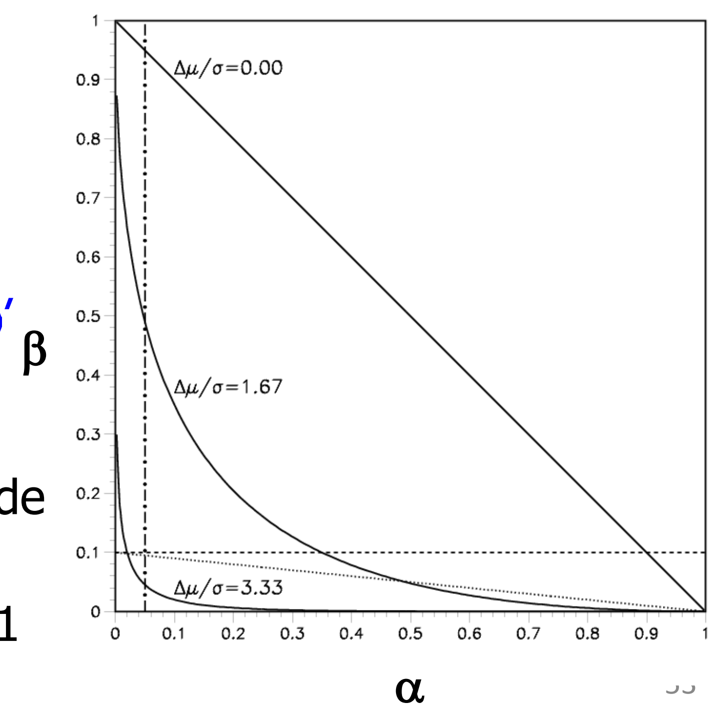
Errors of 1st and 2nd Kind

- 1st Kind: Reject H0 when H0 true
  Should happen at rate α
- 2nd Kind: Fail to reject H0 when H0 is false
  Rate depends on:
  How similar H0 and H1 are
  Relative rates of H0 and H1 (for event selector)

For event selector: E1st = Loss of efficiency
E2nd = Contamination
As $\alpha \downarrow$, efficiency $\uparrow$ and contamination $\downarrow$

For result of expt , E1st gives incorrect result
E2nd fails to make discovery

α = E1st
β = Prob of failing to exclude H0, if H1 = true
1- β = power of test for H1



H0 or H0 versus H1 ?

H0 = null hypothesis

e.g. Standard Model, with nothing new

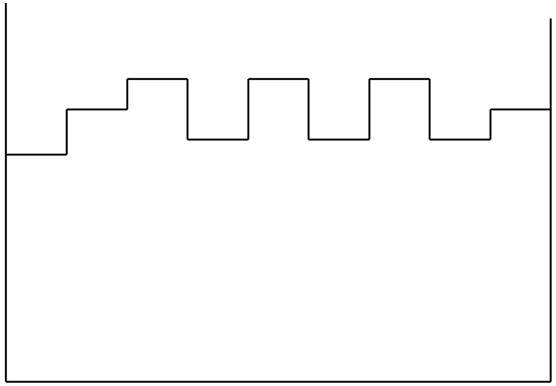
H1 = specific New Physics e.g. Higgs with M_H = 125 GeV

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

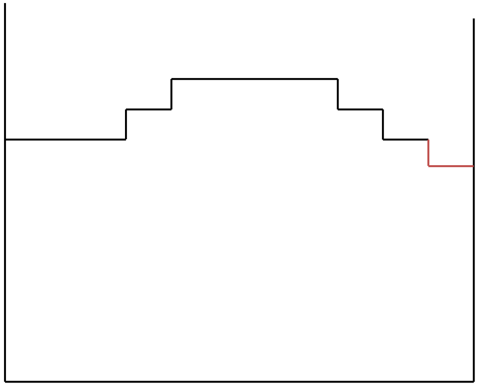
H0 v H1: “Hypothesis Testing” e.g. $\mathcal{L}$-ratio

Measures how much data favours one hypothesis wrt other

H0 v H1 likely to be more sensitive for H1



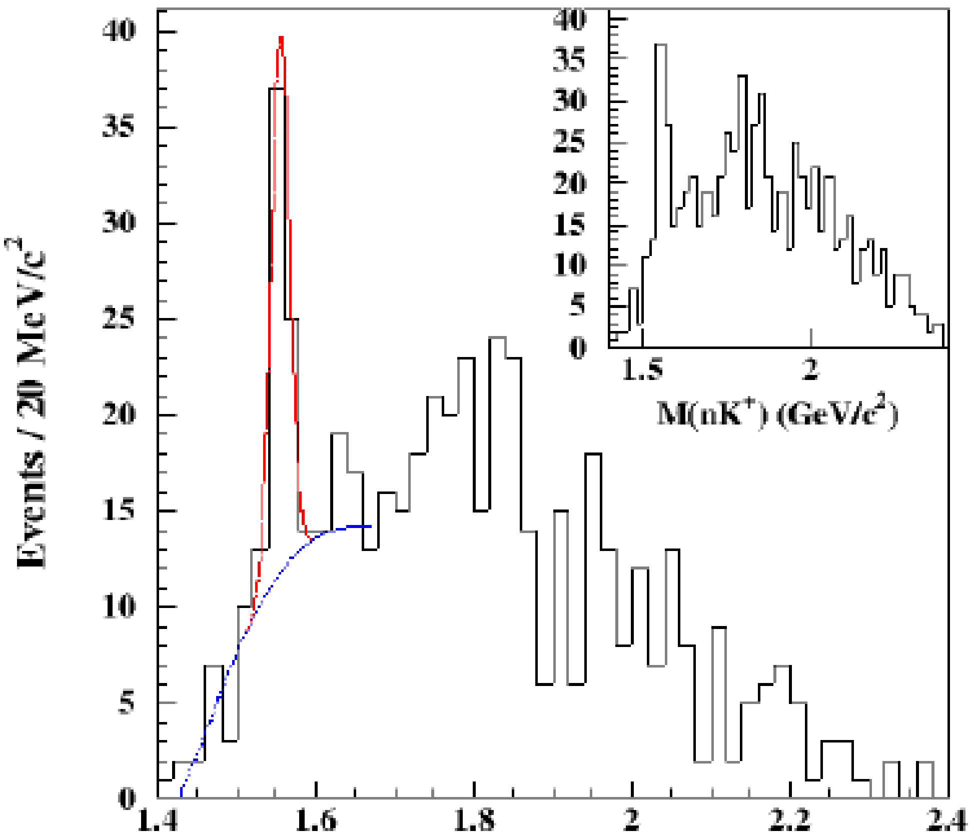
or



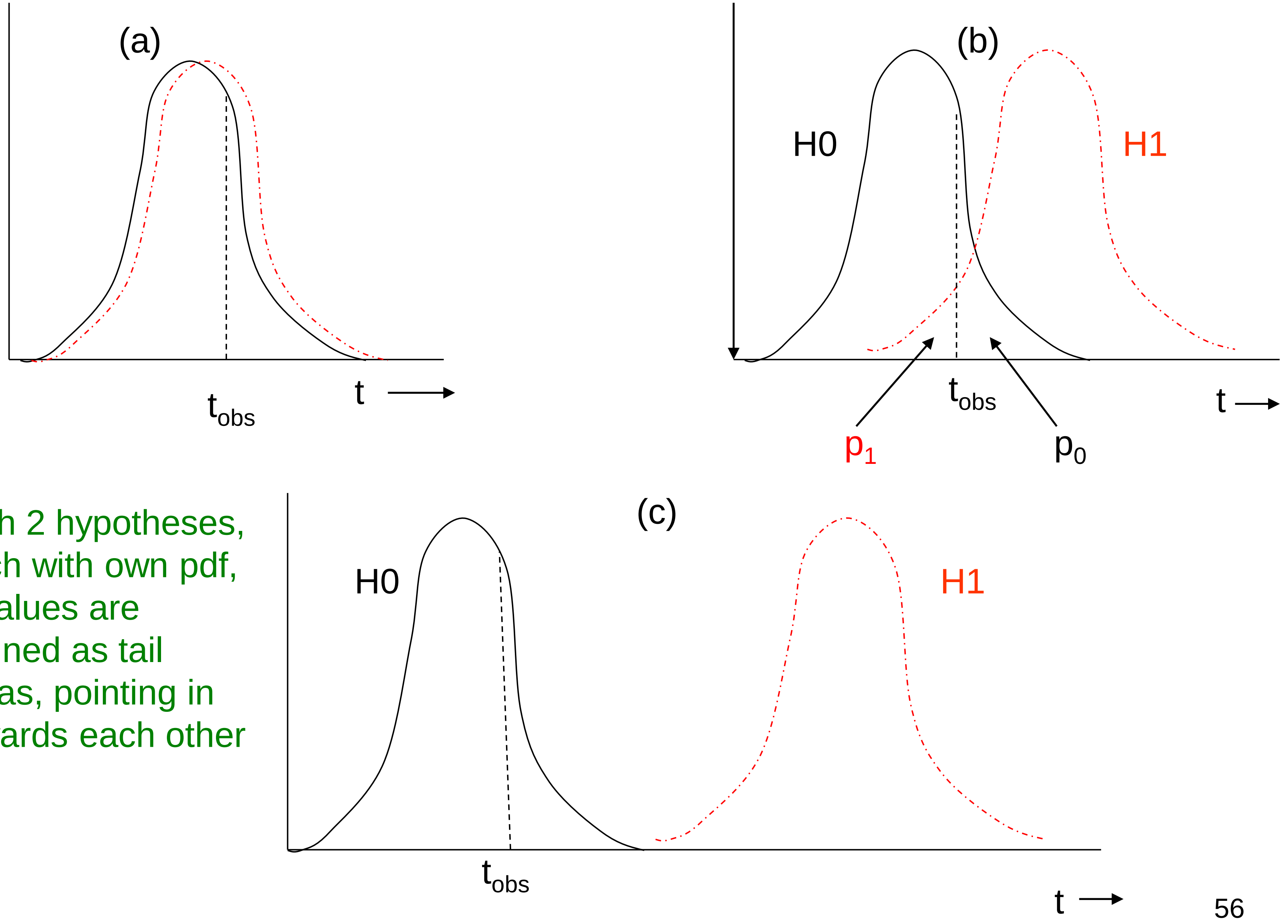
Choosing between 2 hypotheses

Hypothesis testing: New particle or statistical fluctuation?

H0 = b   H1 = b + s

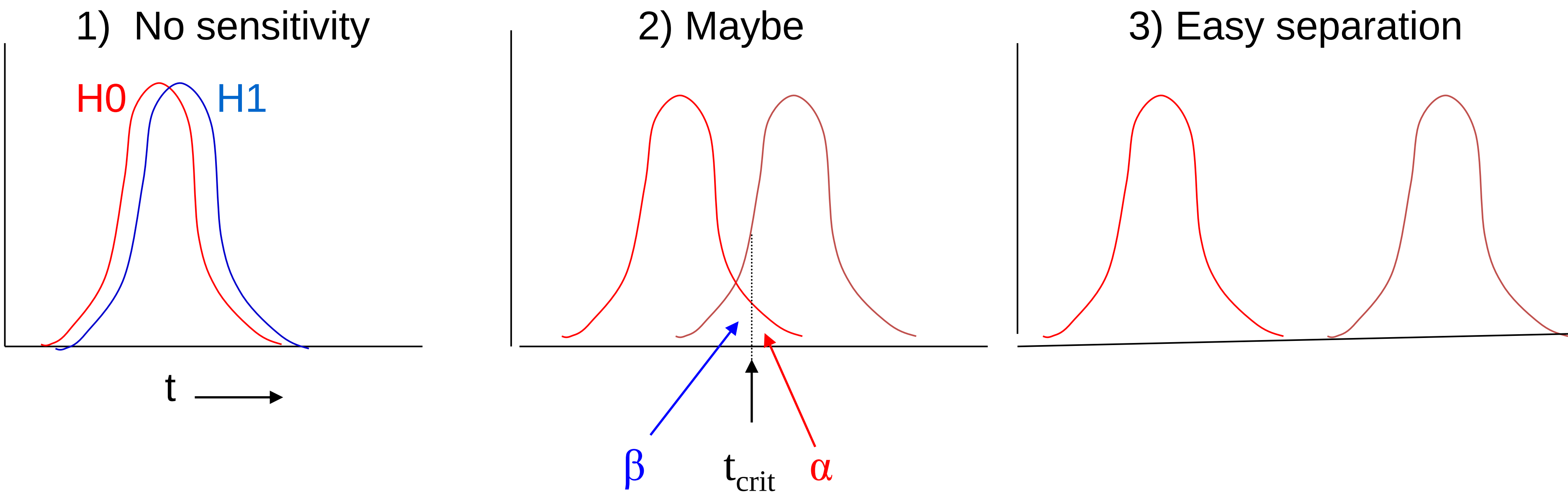


First define your data statistic t (n, L-ratio, etc.)



With 2 hypotheses, each with own pdf, p-values are defined as tail areas, pointing in towards each other

Procedure for choosing between 2 hypotheses



Procedure: Obtain expected distributions for data statistic (e.g. $\mathcal{L}$-ratio) for H0 and H1

Choose α (e.g. 95%, 3σ, 5σ ?) and CL for p_1 (e.g. 95%)

Given b, α determines t_{crit}

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

1-β = Power of test

Now data: If $t_{obs} \geq t_{crit}$ (i.e. $p_0 \leq \alpha$), discovery at level α

If $t_{obs} < t_{crit}$, no discovery. If $p_1 <$ 1– CL, exclude H1

Significance

(Number of σ = p-value converted to Gaussian one-sided tail)

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

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, bins (Blind analyses)

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

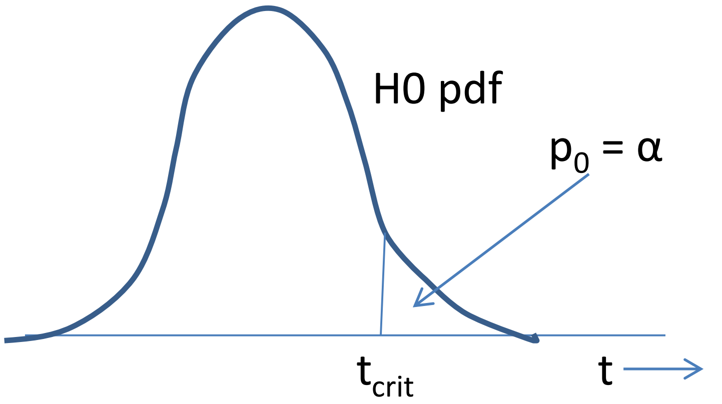
or Prob{speed of $\nu > c$ | data} ~ 1

Everyday situation:

p(eat bread|murderer) ~ 99%

p(murderer|eat bread) ~ 10^{-6}

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, $S = z * (1 - \ln z) \geq z$)

Problems:

1) Recipe is not unique (Uniform dist in n-D hypercube → uniform in 1-D)

2) 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".

3) Small p's due to different discrepancies

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

BLIND ANALYSES

Why blind analysis? Data statistic, 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

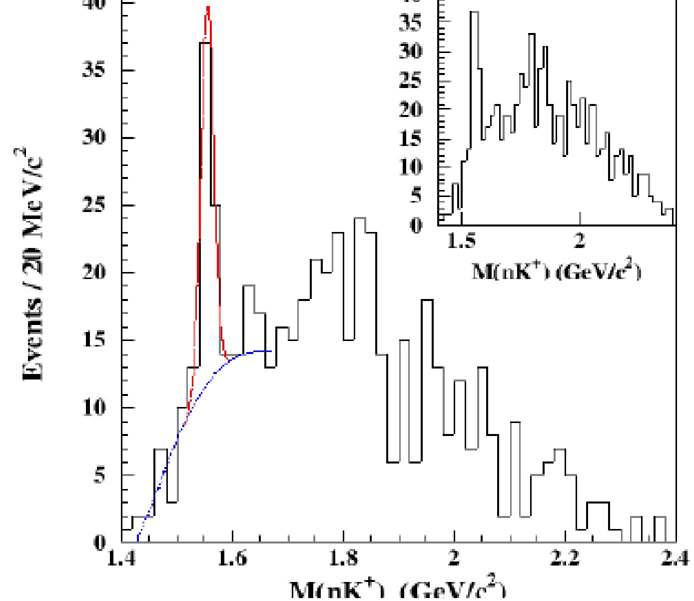
Disadvantages

- Takes longer time
- Usually not available for searches for unknown

After analysis is unblinded, don't change anything unless ....

* Luis Alvarez suggestion re "discovery" of free quarks

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

Example of LEE: Stonehenge

12 is the number of constellations

6 is the number of ages (2160) we spend on each side of the galactic equator

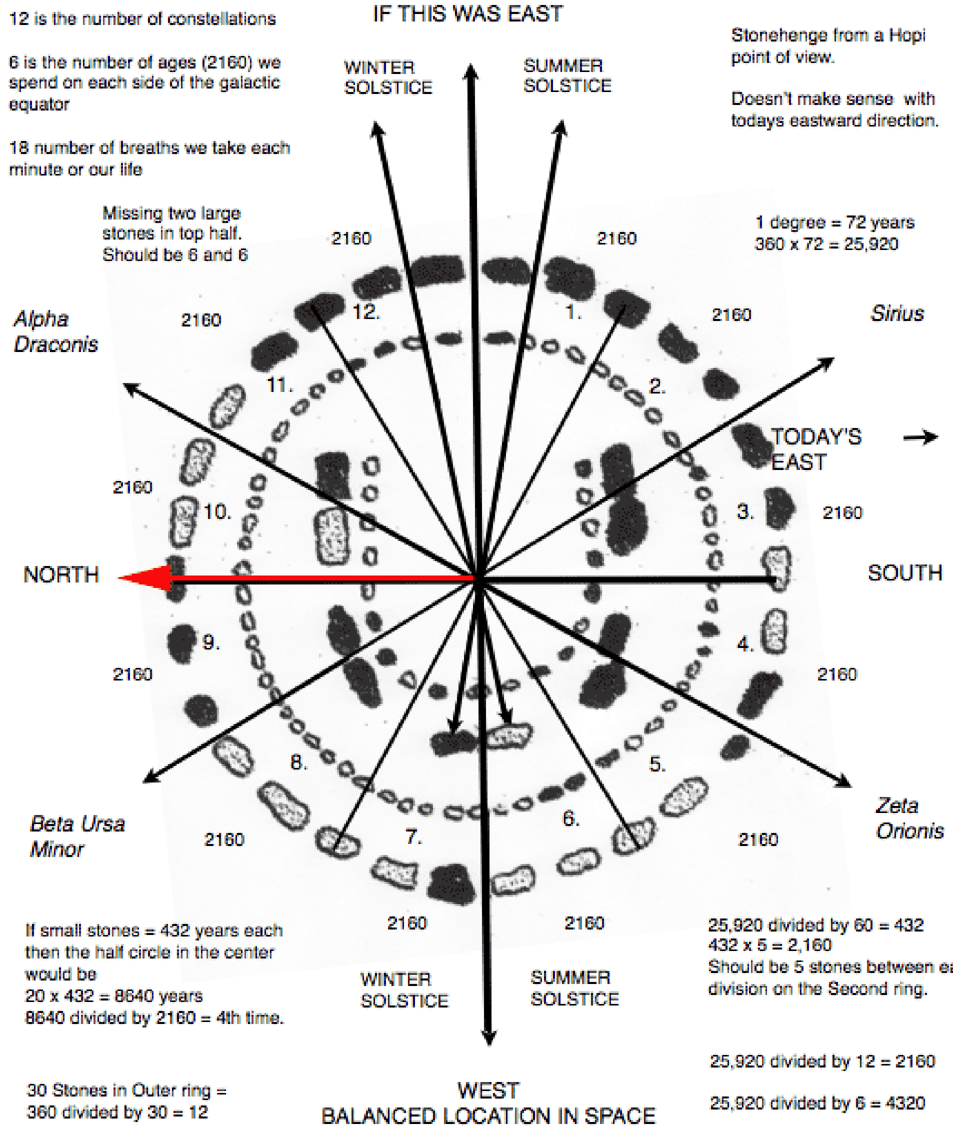
18 number of breaths we take each minute or our life

Missing two large stones in top half. Should be 6 and 6

Stonehenge from a Hopi point of view.

Doesn't make sense with todays eastward direction.

1 degree = 72 years
360 x 72 = 25,920



If small stones = 432 years each then the half circle in the center would be
20 x 432 = 8640 years
8640 divided by 2160 = 4th time.

25,920 divided by 60 = 432
432 x 5 = 2,160
Should be 5 stones between each division on the Second ring.

25,920 divided by 12 = 2160

25,920 divided by 6 = 4320

25,920 divided by 18 = 1440

30 Stones in Outer ring =
360 divided by 30 = 12

60 Stones in Second ring =
360 divided by 60 = 6

20 Stones in Center ring =
360 divided by 20 = 18

Center Stone in Center Ring would be divided in half by sun rays when Earth in perfect balance. Nine on each side + 2 = 20.

STONEHENGE

The Book of Truth
A New Perspective on the Hopi Creation Story
by Thomas O. Mills

Are alignments significant?

- Atkinson replied with his article "Moonshine on Stonehenge" in *Antiquity* in 1966, pointing out that some of the pits which ….. had used for his sight lines were more likely to have been natural depressions, and that he had allowed a margin of error of up to 2 degrees in his alignments. Atkinson found that the probability of so many alignments being visible from 165 points to be close to 0.5 rather that the "one in a million" possibility which ….. had claimed.

- ….. had been examining stone circles since the 1950s in search of astronomical alignments and the megalithic yard. It was not until 1973 that he turned his attention to Stonehenge. He chose to ignore alignments between features within the monument, considering them to be too close together to be reliable. He looked for landscape features that could have marked lunar and solar events. However, one of …..'s key sites, Peter's Mound, turned out to be a twentieth-century rubbish dump.

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

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 $\neq 0$	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"

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

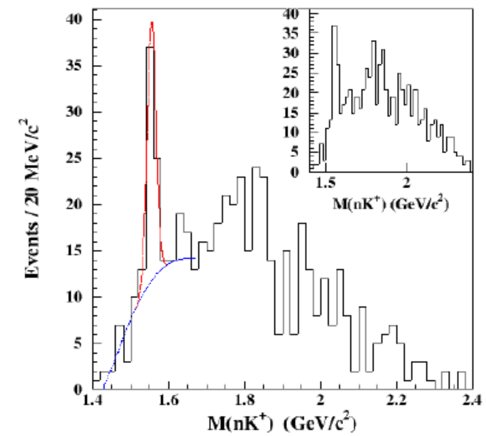
H1 = polynomial of degree 5

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 $\Delta S = S_0 - S_1$? Expect not large. Is it χ^2?

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

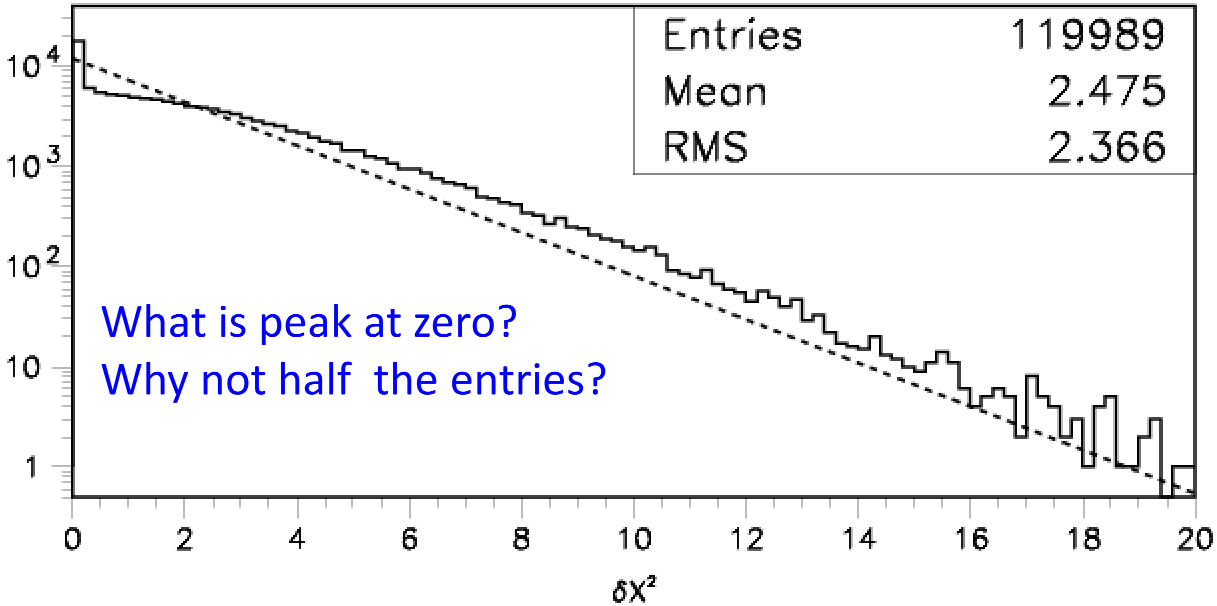
3) H0 = normal neutrino hierarchy *********
   H1 = inverted hierarchy *********

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

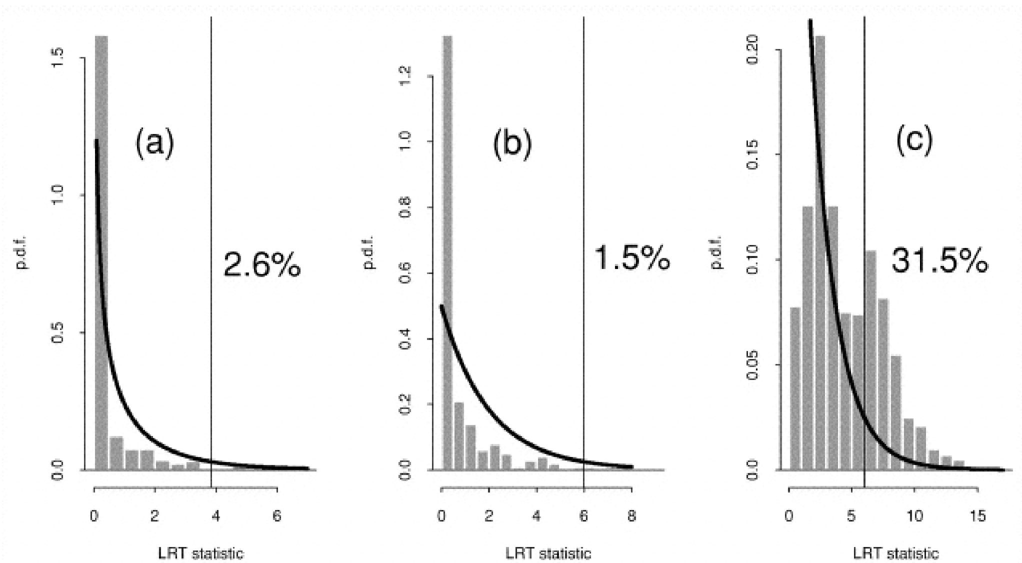
N.B. 1: Even when W. Th. does not apply, it does not mean that ΔS is irrelevant, but you cannot use W. Th. for its expected distribution.

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

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

So need to determine the ΔS distribution by Monte Carlo

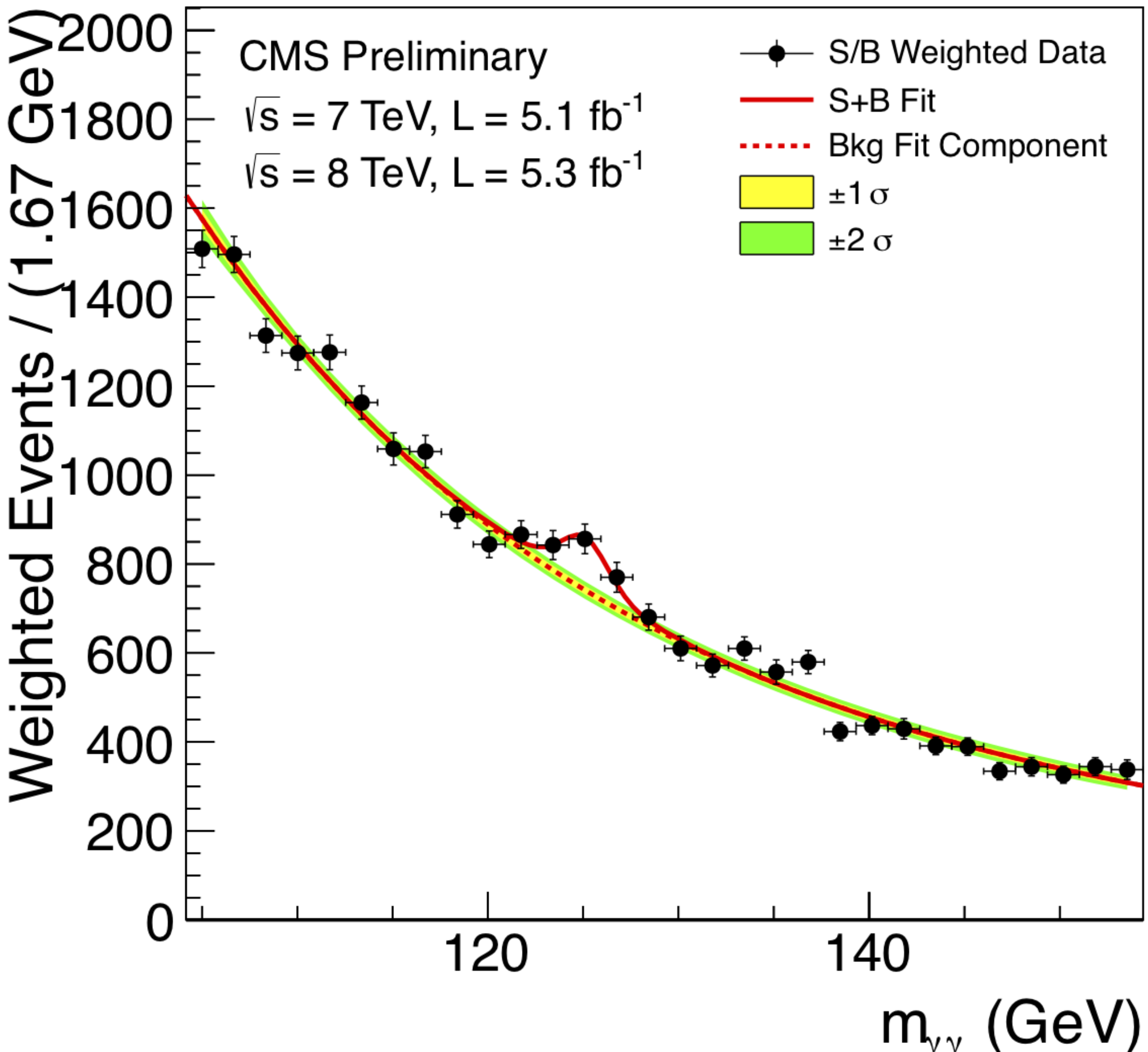
N.B.

1) For mass spectrum, determining ΔS 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)

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]

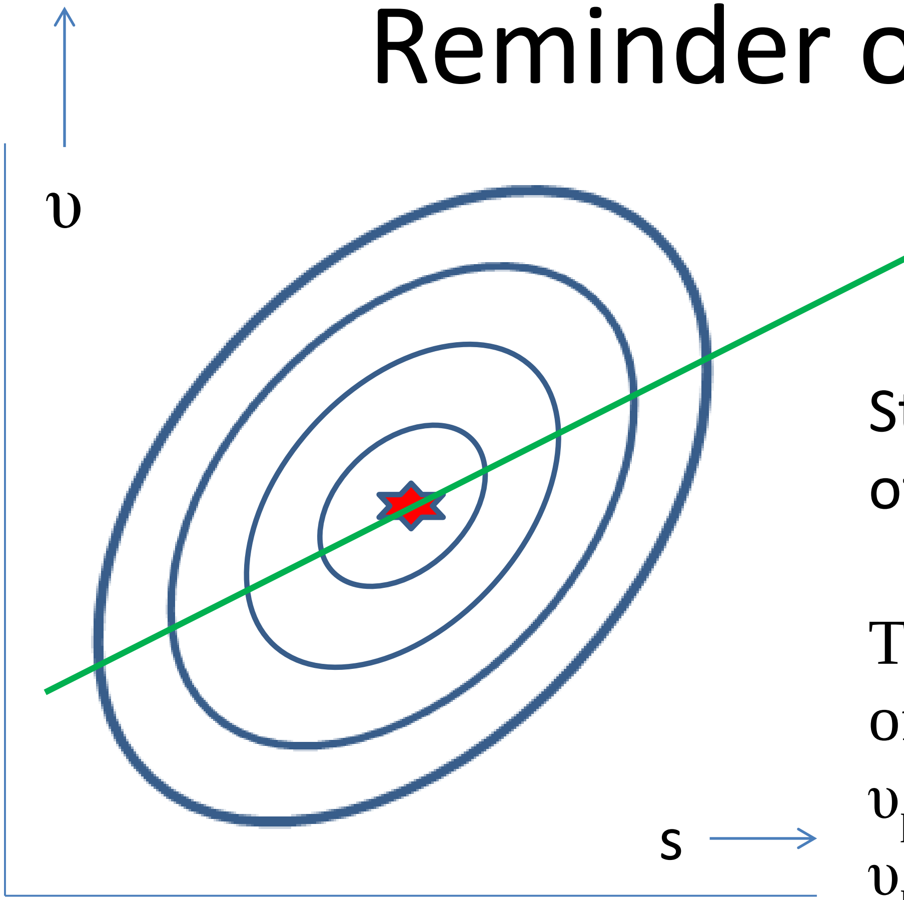
Has been used in CMS analysis of H→γγ

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 $\mathcal{L}$

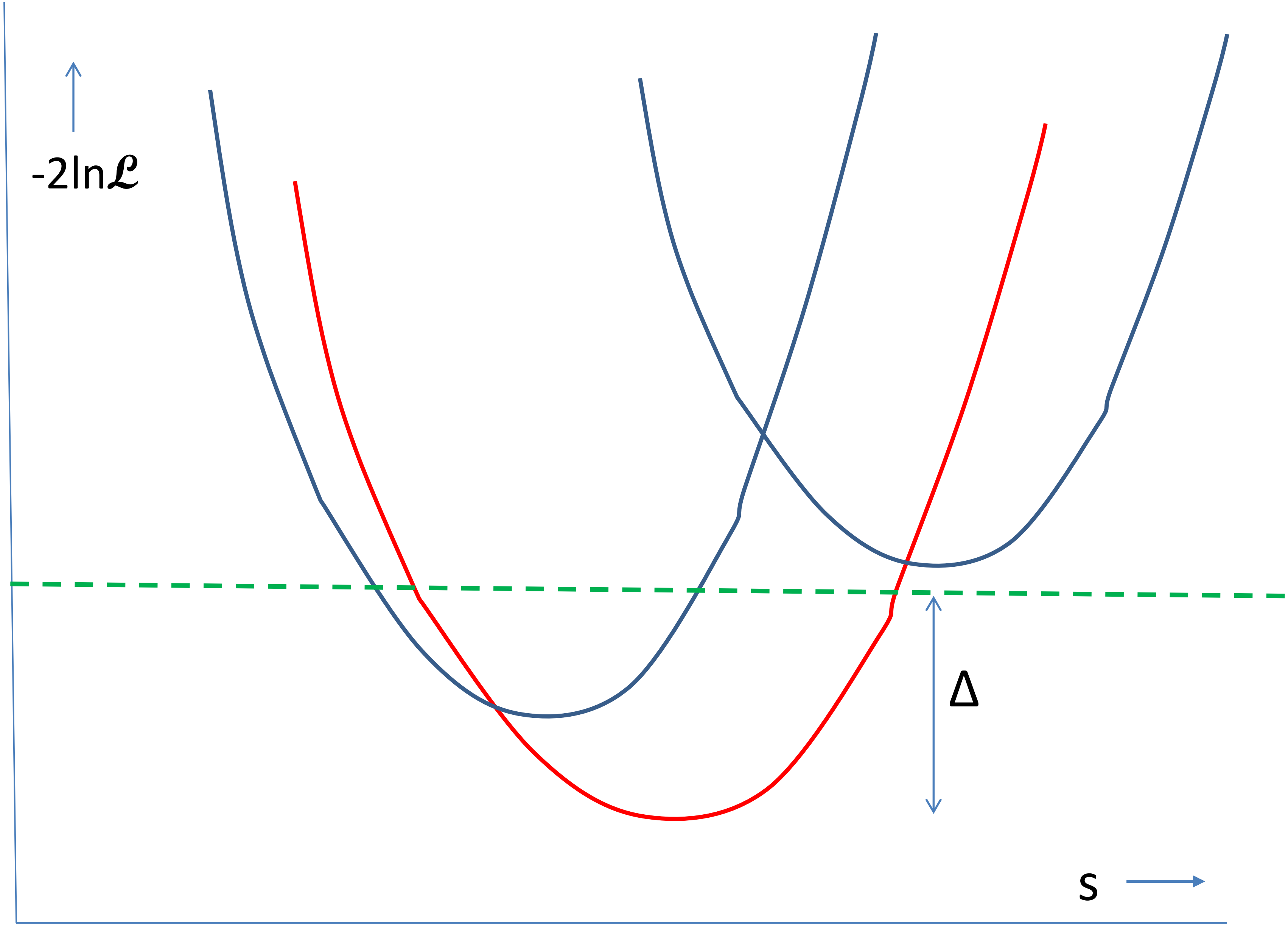


Contours of ln$\mathcal{L}$(s,υ)
s = physics param
υ = nuisance param

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


-2lnℒ
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 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, as in 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

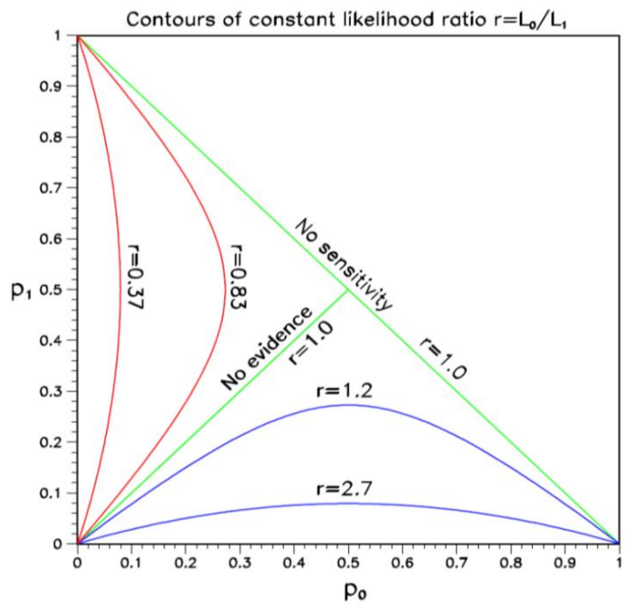
p_0 v p_1 plots

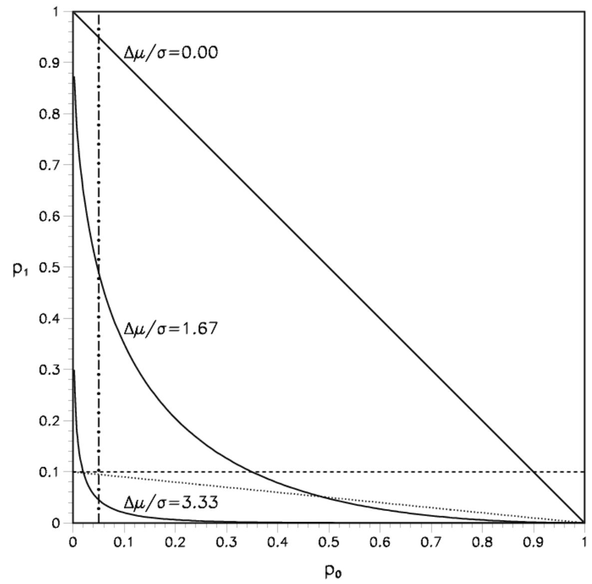
Preprint by Luc Demortier and LL, "Testing Hypotheses in Particle Physics: Plots of p_0 versus p_1"
http://arxiv.org/abs/1408.6123

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



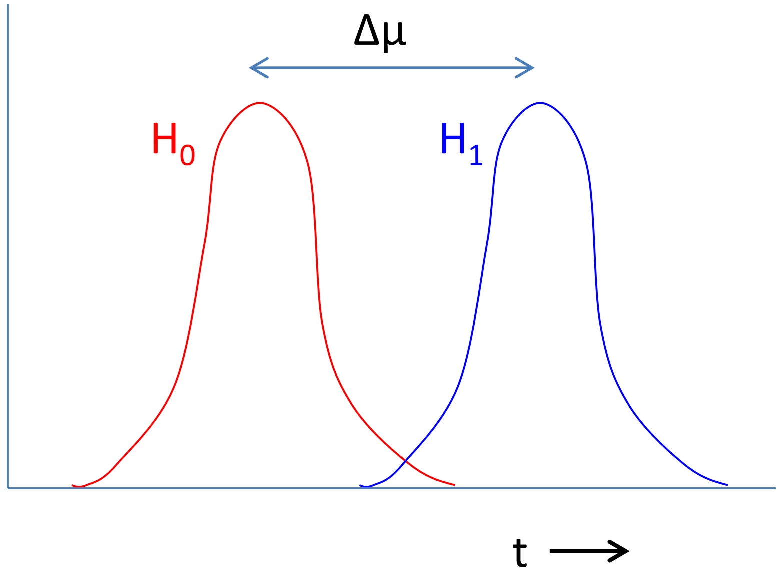


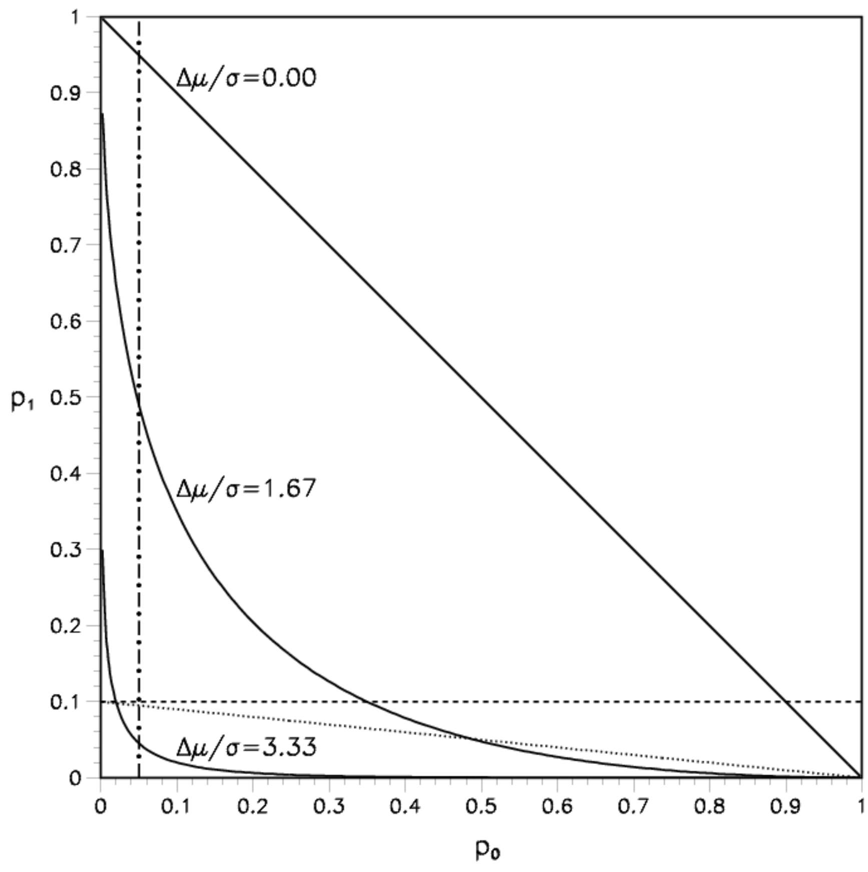
$CLs = p_1/(1-p_0)$ → diagonal line

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





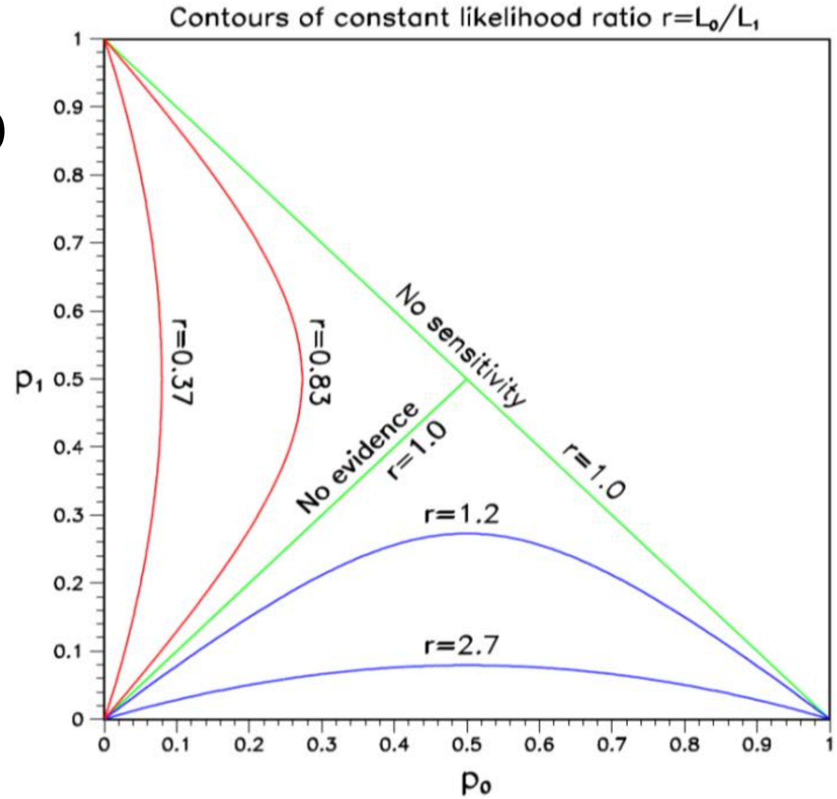
Can read off power of test

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; $\mathcal{L}_{01}$ compares H0 and H1

Depends on amount of data:

e.g. Poisson counting expt little data:

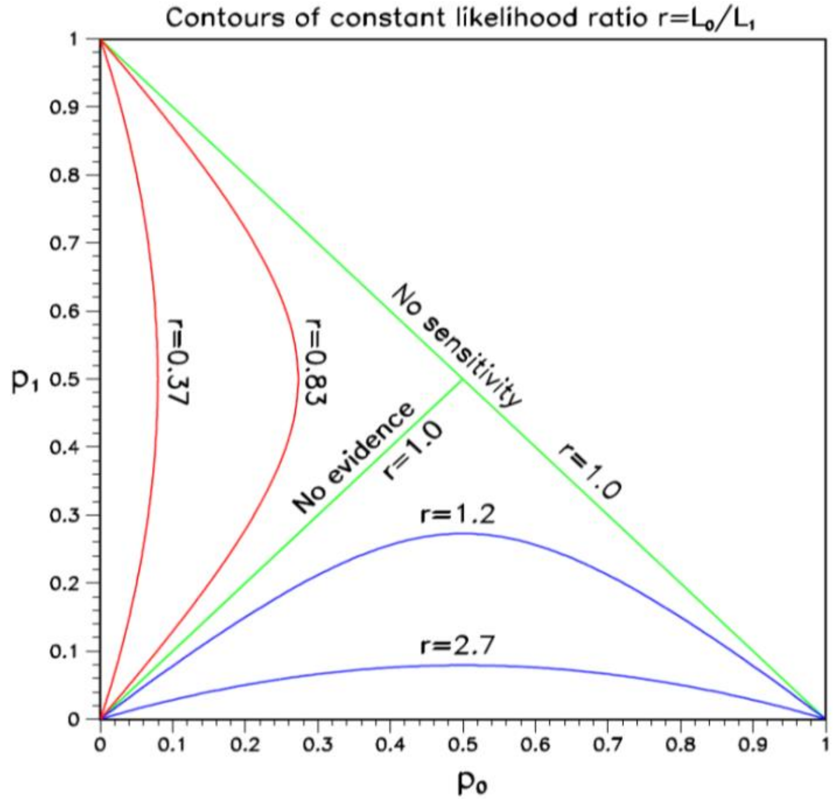
For H0, $\mu_0 = 1.0$. For H1, $\mu_1 = 10.0$

Observe n = 10 $p_0 \sim 10^{-7}$ $\mathcal{L}_{01} \sim 10^{-5}$

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

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

Jeffreys-Lindley Paradox



H0 = simple, H1 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

WHY LIMITS?

Michelson-Morley experiment → death of aether

HEP experiments: If UL on rate for new
particle $<$ expected, exclude particle

CERN CLW (Jan 2000)
FNAL CLW (March 2000)
Heinrich, PHYSTAT-LHC, "Review of Banff Challenge"

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 ($\sigma^2 = \mu$)

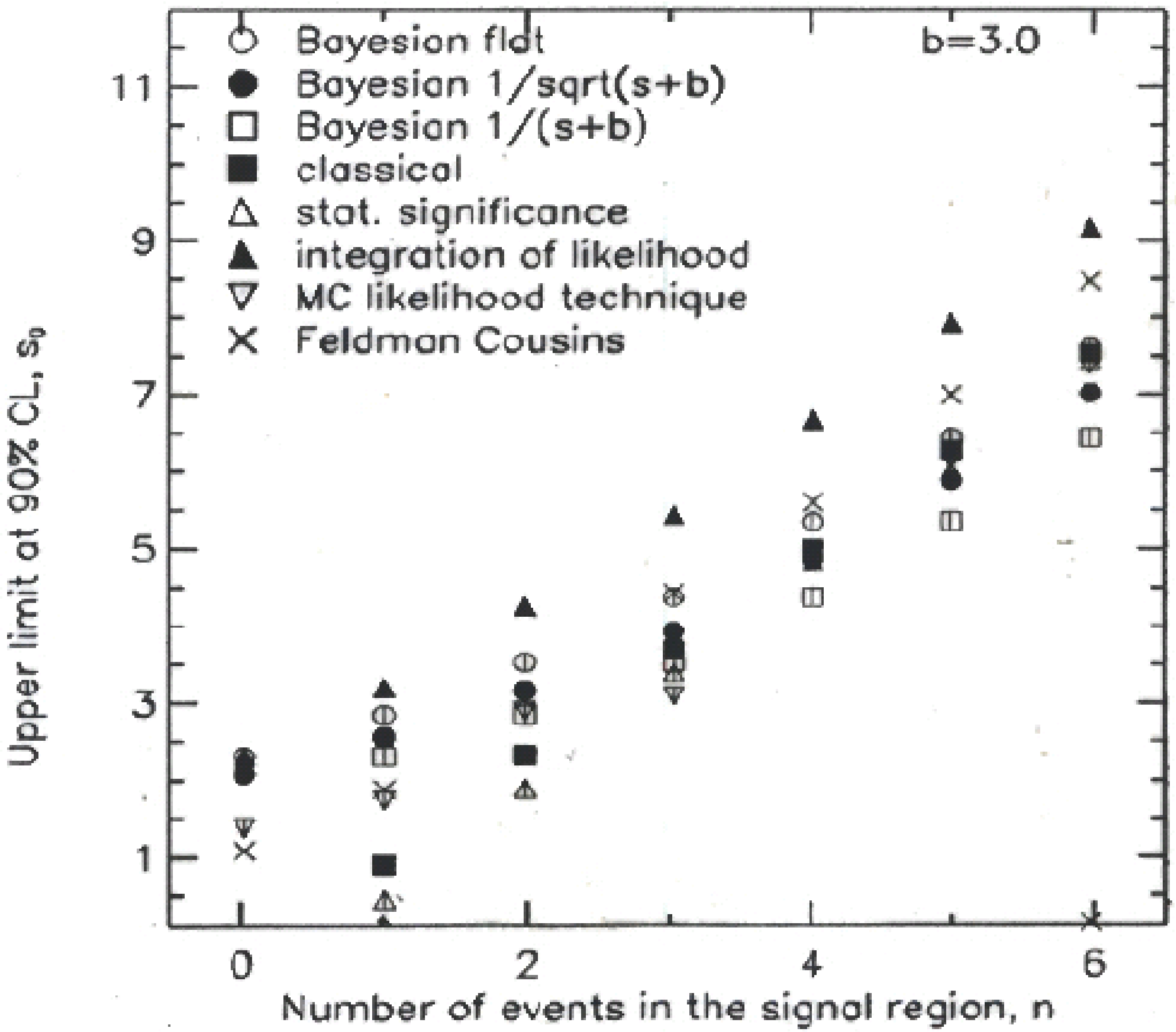
χ^2(σ^2 = n)

Recommendation 7 from CERN CLW (2000): "Show your $\mathcal{L}$"

1) Not always practical

2) Not sufficient for frequentist methods

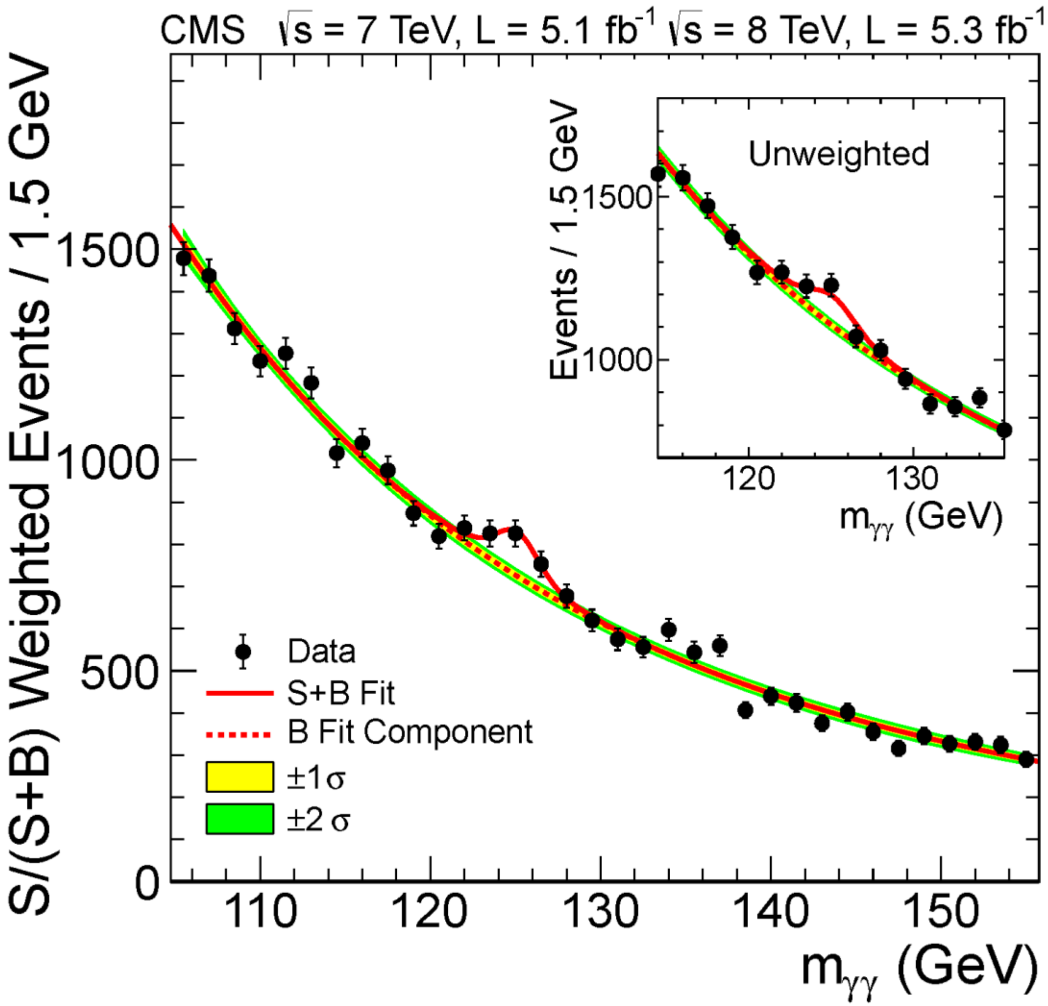
Ilya Narsky, FNAL CLW 2000



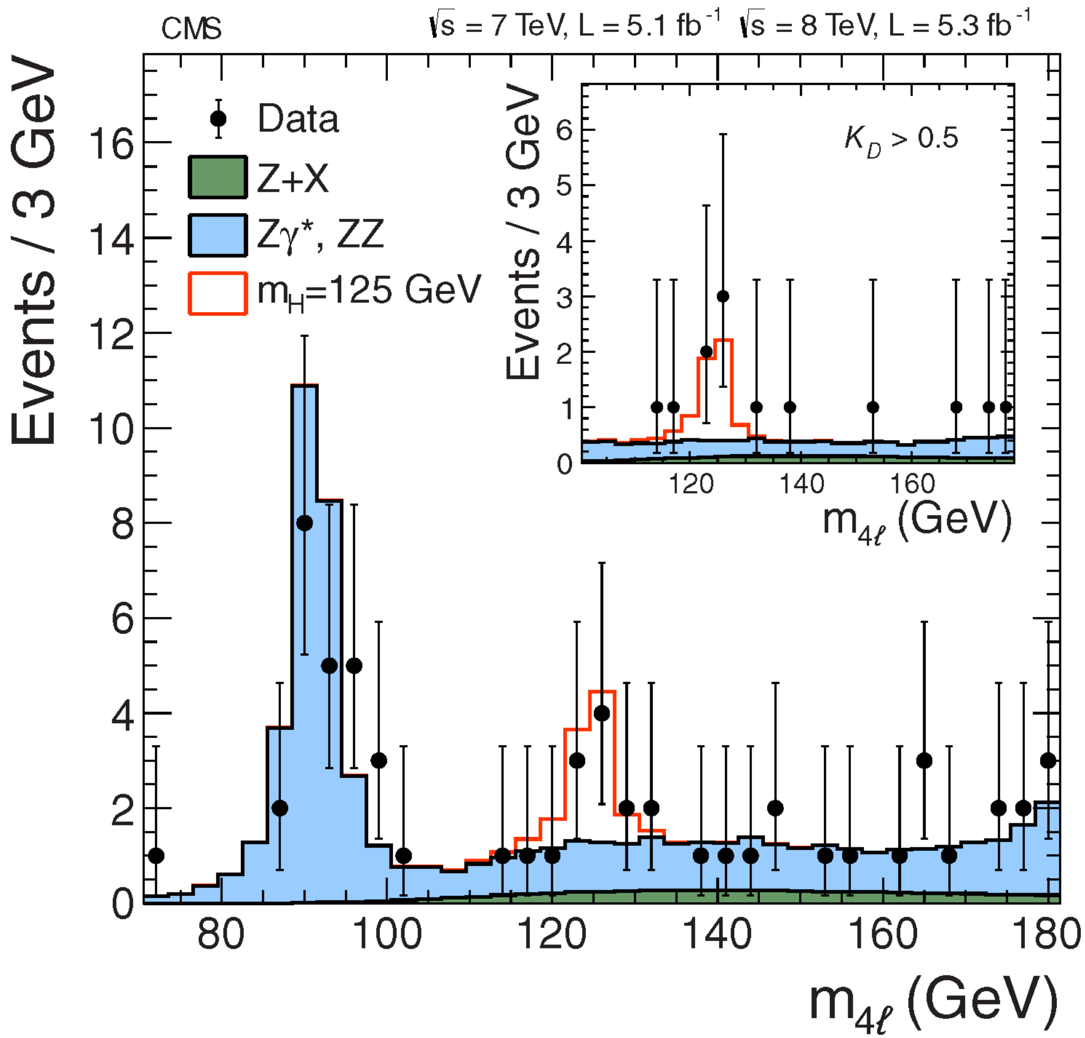
DESIRABLE PROPERTIES

- Coverage
- Interval length
- Behaviour when $n < b$
- Limit increases as σ_b increases
- Unified with discovery and interval estimation

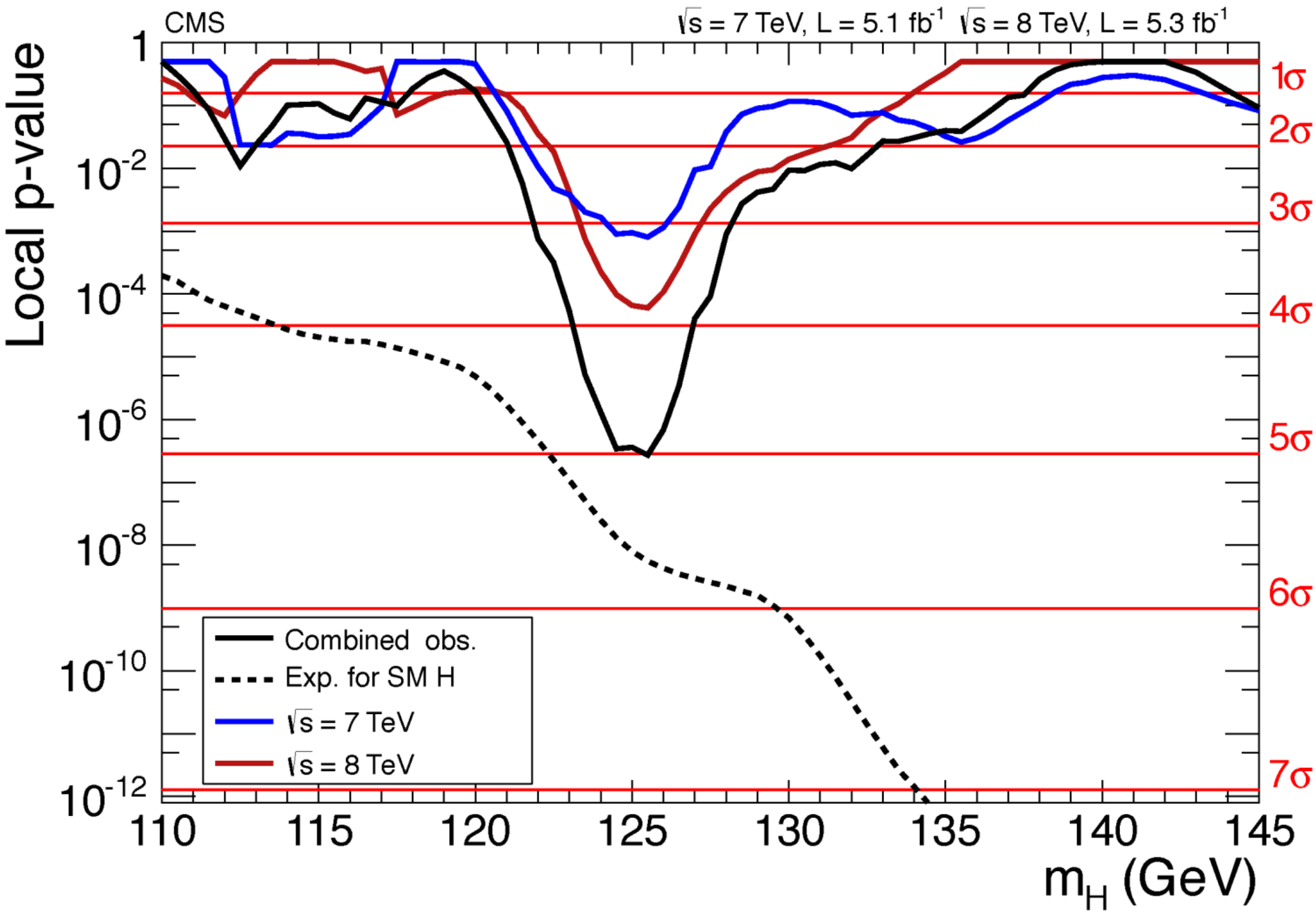
Search for Higgs:
H→ γ γ: low S/B, high statistics



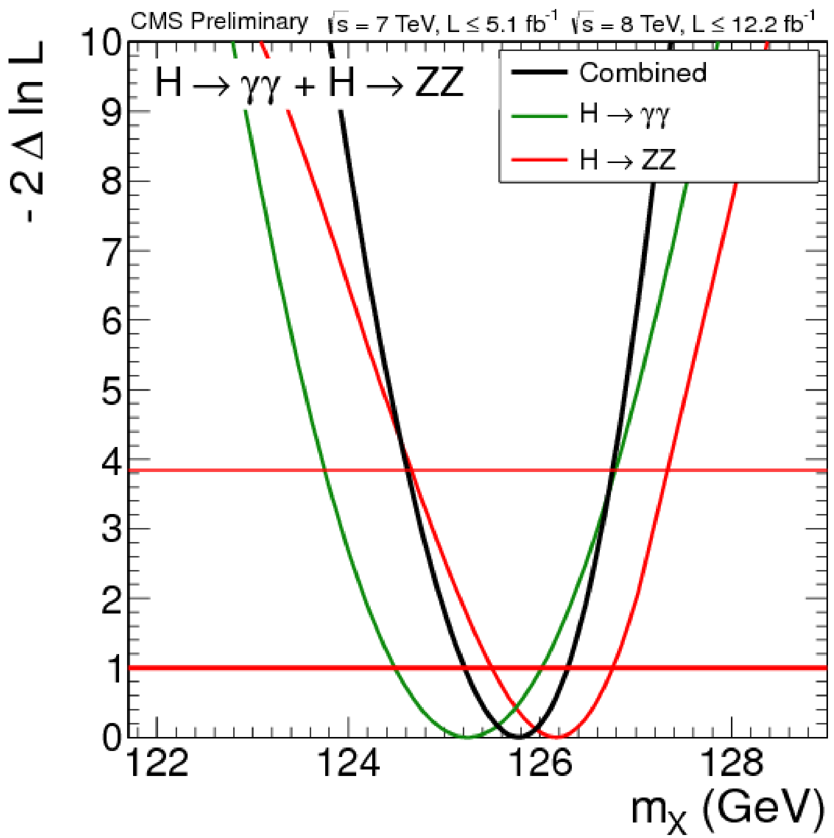
H→Z Z → 4 l: high S/B, low statistics

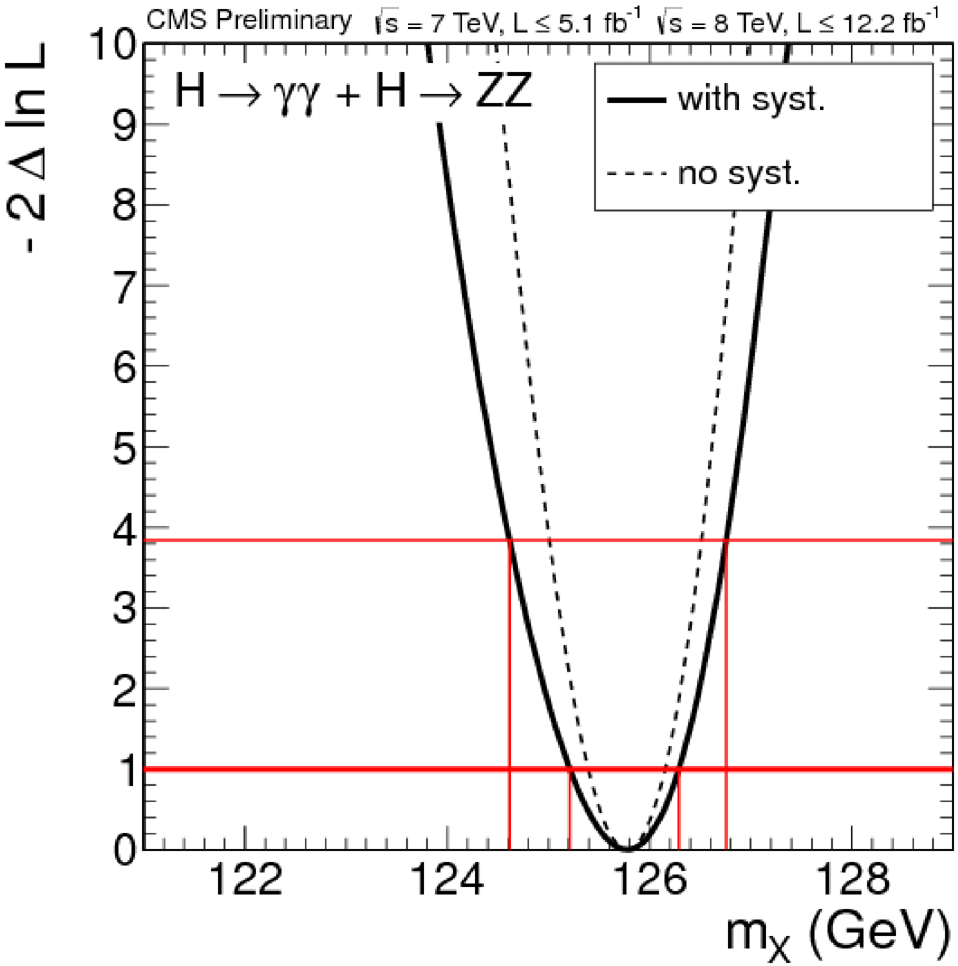


p-value for 'No Higgs' versus m_H



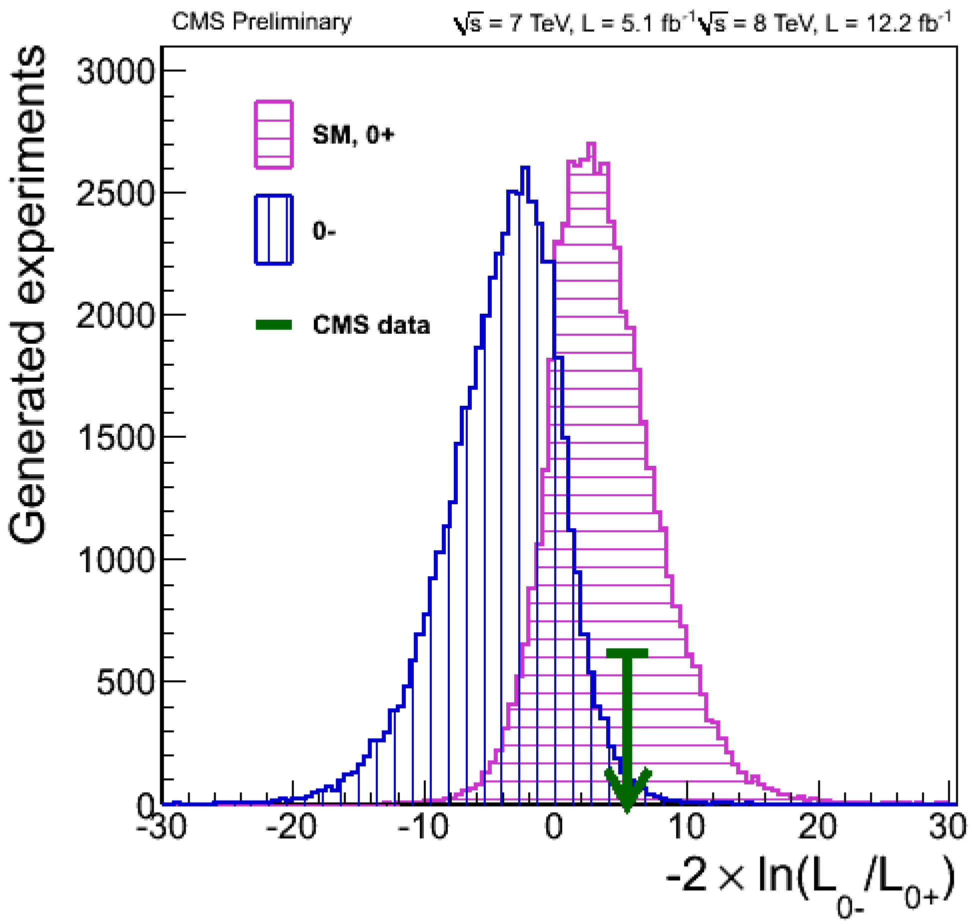
Mass of Higgs: Likelihood versus mass





Comparing 0^+ versus 0^- for Higgs

(like Neutrino Mass Hierarchy)



http://cms.web.cern.ch/news/highlights-cms-results-presented-hcp

Conclusions

Resources:

Software exists: e.g. RooStats

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

`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"