

Statistics and Data analysis

Bill Murray
RAL, CCLRC
w.murray@rl.ac.uk

RAL Graduate Lectures
29th November 2006

- Why use statistics?
- Have I found something?
- Parameter extraction
- Goodness of fit



What do we want to achieve?

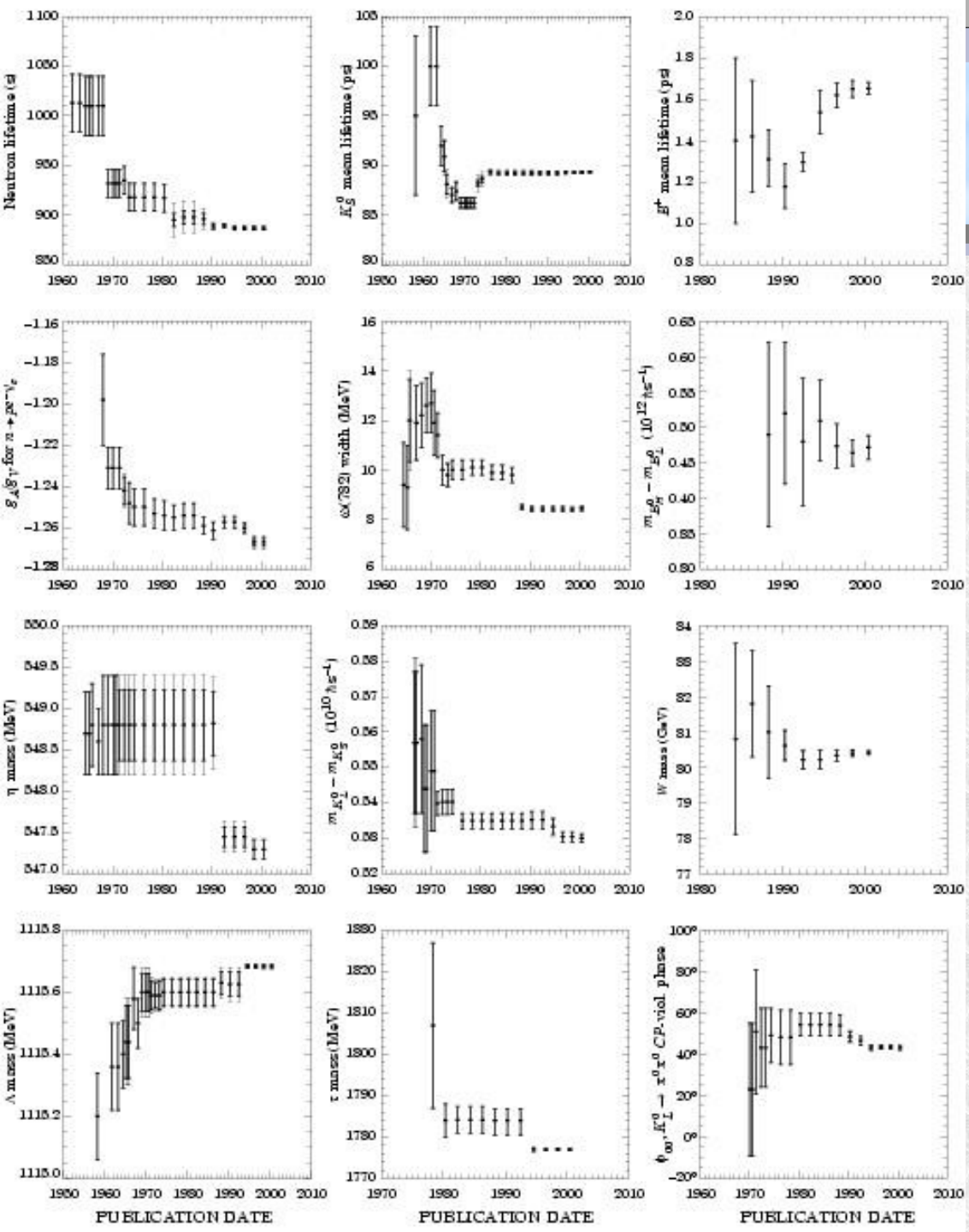
- We aim to make optimal use of the data collected
 - Data are expensive:
 - Use powerful techniques
 - Data processing is also expensive:
 - Mathematical perfection is not the only criterion
 - Systematic errors may well dominate
 - We need to be able to justify our results.



Three Classes of problem

- Hypothesis testing
 - Does this signal exist?
 - Bayes and Frequentist limits
- Parameter extraction
 - What is the mass of the W
 - Systematic and statistical errors
- Goodness of fit
 - Chi2 test
 - Other tests

History of measurements



Each measurement agrees with preceding one!

Publications which disagree with the standard model/previous estimates are checked more carefully.

If you look you can usually find something

BABAR use blind analysis

Figure 2: An historical perspective of values of a few particle properties tabulated in this Review as a function of date of publication of the Review. A full error bar indicates the quoted error; a thick-lined portion indicates the same but without the "scale factor."



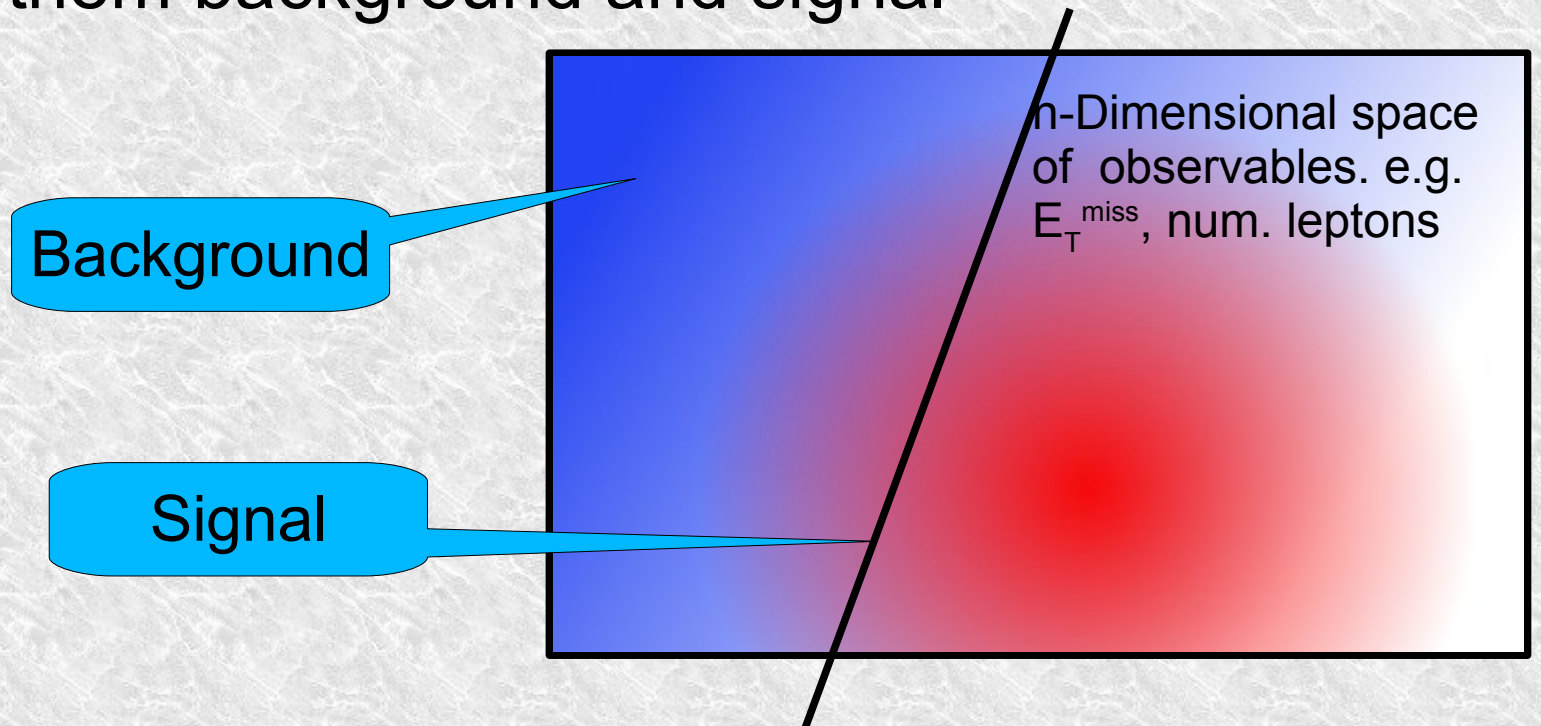
Hypothesis Testing

- Have we found a new signal?



Signal Recognition

- Consider separating a dataset into 2 classes
 - Call them background and signal

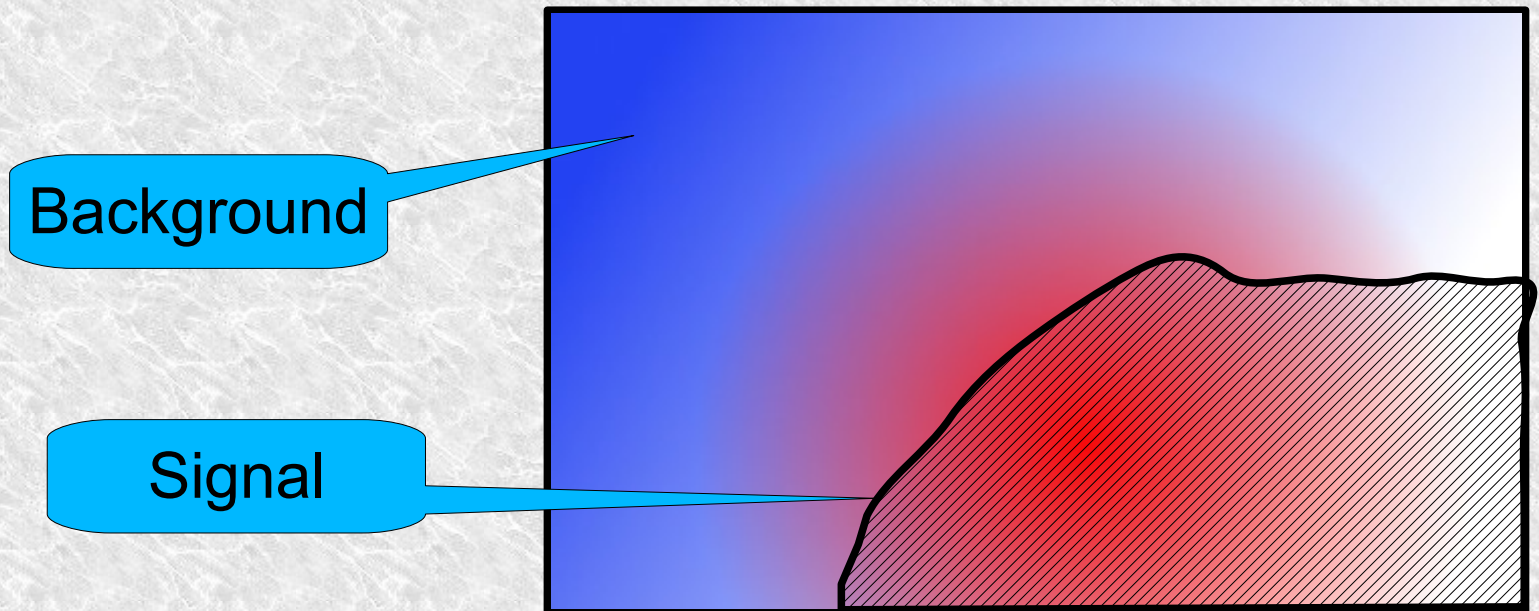


- A simple cut is not optimal



The right answer II

- What is optimal?

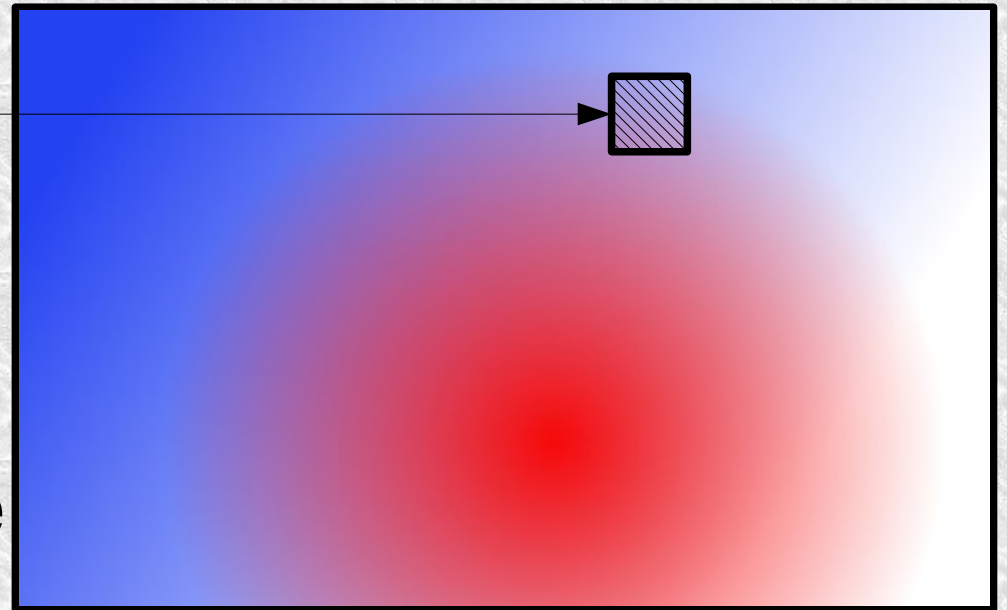


- Maybe something like this might be...



The right answer III

- For a given efficiency, we want to minimize background
- Sort by signal to background ratio around event
- Accept all areas with s/b above some threshold
- Leading to the ***Likelihood Ratio***





But is that the best we can do?

- Ordering events by L_s/L_b and selecting above some threshold gives best possible sample
- But when we ask the question “is there a signal there?” we can extract more information.
- Do a likelihood-ratio fit:
 - Use: $1+L_s/L_b = (L_s+L_b)/L_b$
 - How much more probable if signal also present?
- Take product over all events
 - One event with $L_b=0$ disproves b
 - 2 events with $(s+b)/b=10$ same as 1 with 100
- This product is the most sensitive estimator



Determination of s, b densities

- We may know matrix elements
 - Not for e.g. a b-tag
 - But anyway there are detector effects
- Usually taken from simulation



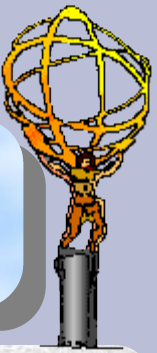
Using MC to calculate density

- Brute force:
 - Divide our n-D space into hypercubes with m divisions of each axis
 - m^n elements, need $100 m^n$ events for 10% estimate.
 - e.g. 1,000,000,000 for 7 dimensions and 10 bins in each
- This assumed a uniform density – actually need far more
 - The purpose was to separate different distributions



Better likelihood estimation

- Clever binning
 - Starts to lead to tree techniques
- Kernel density estimators
 - Size of kernel grows with dimensions
 - Edges are an issue
- Ignore correlations in variables
 - Very commonly done **'I used likelihood'**
- Pretend measured=true, correct later
 - Linked to OO techniques, bias correction

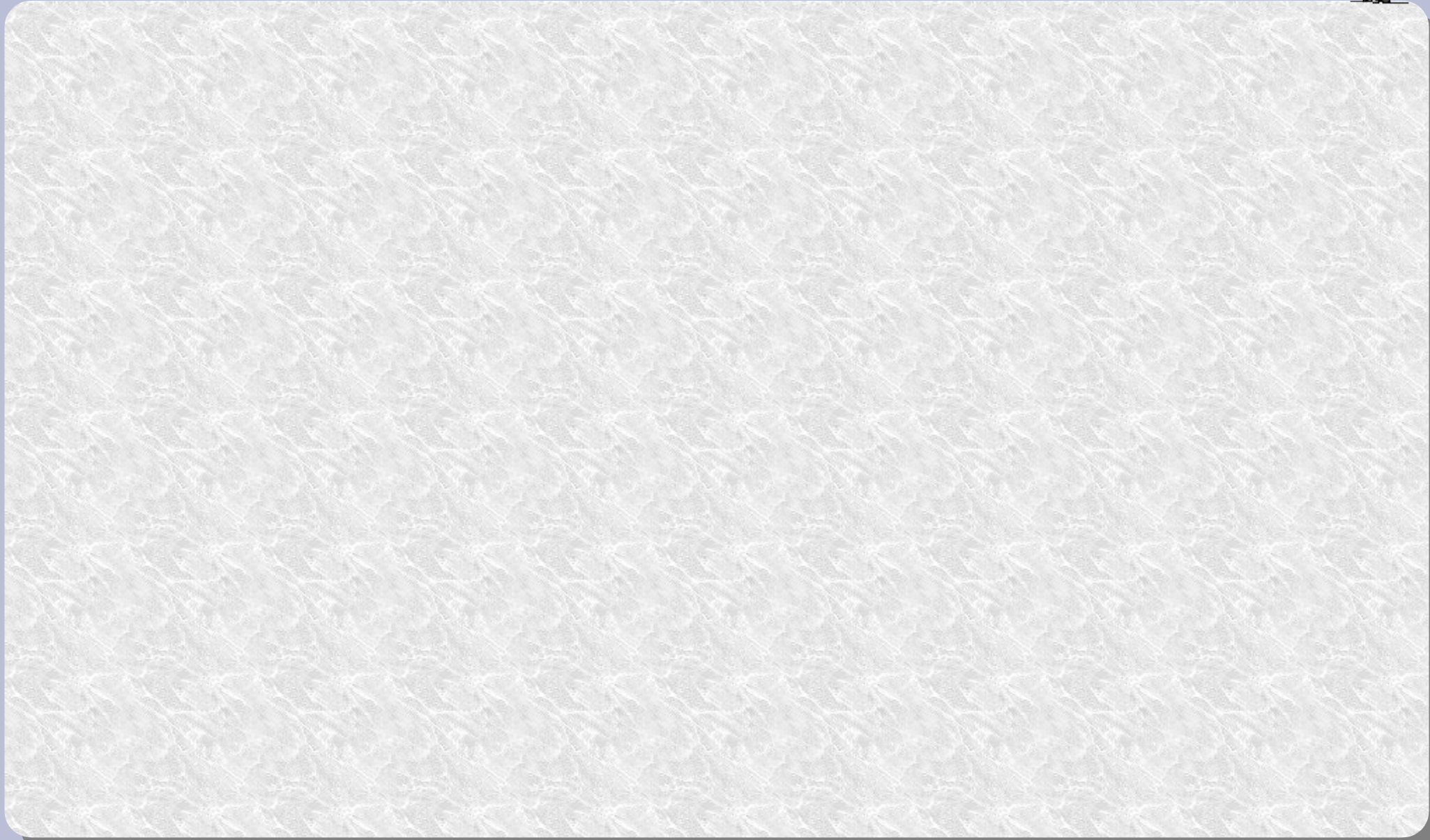


Alternative approaches

- Neural nets
 - Well known, good for high-dimensions
- Support vector machines
 - Computationally easier than kernel
- Decision trees
 - Boosted or not?



How to calculate densities



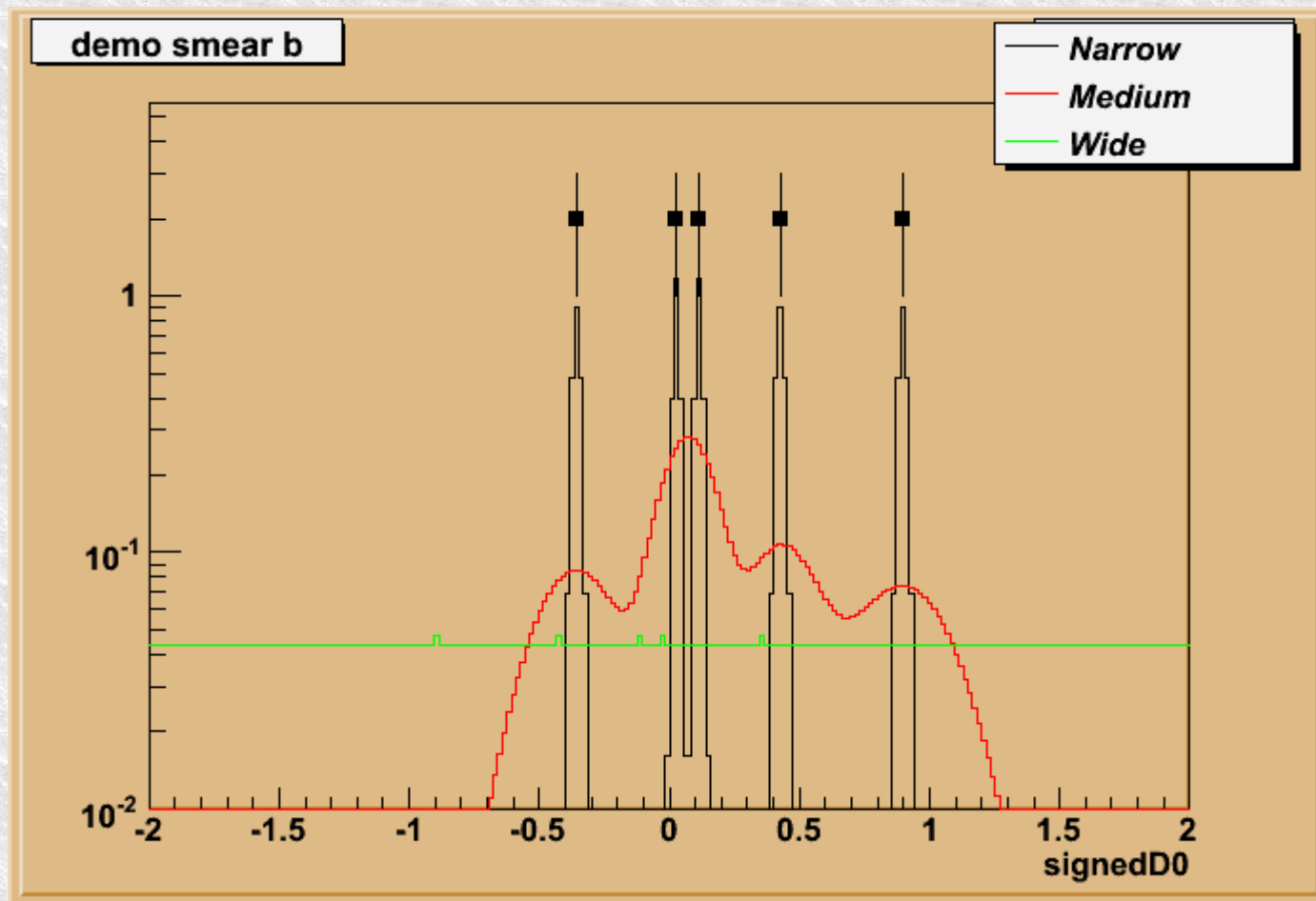


Kernel Likelihoods

- Directly estimate Probability Density Function of distributions based upon training sample events.
- Some kernel, usually Gaussian, smears the sample
 - increases widths
 - Width of kernel must be optimised
- Fully optimal *if* infinite MC statistics
 - Then we can use narrow kernels



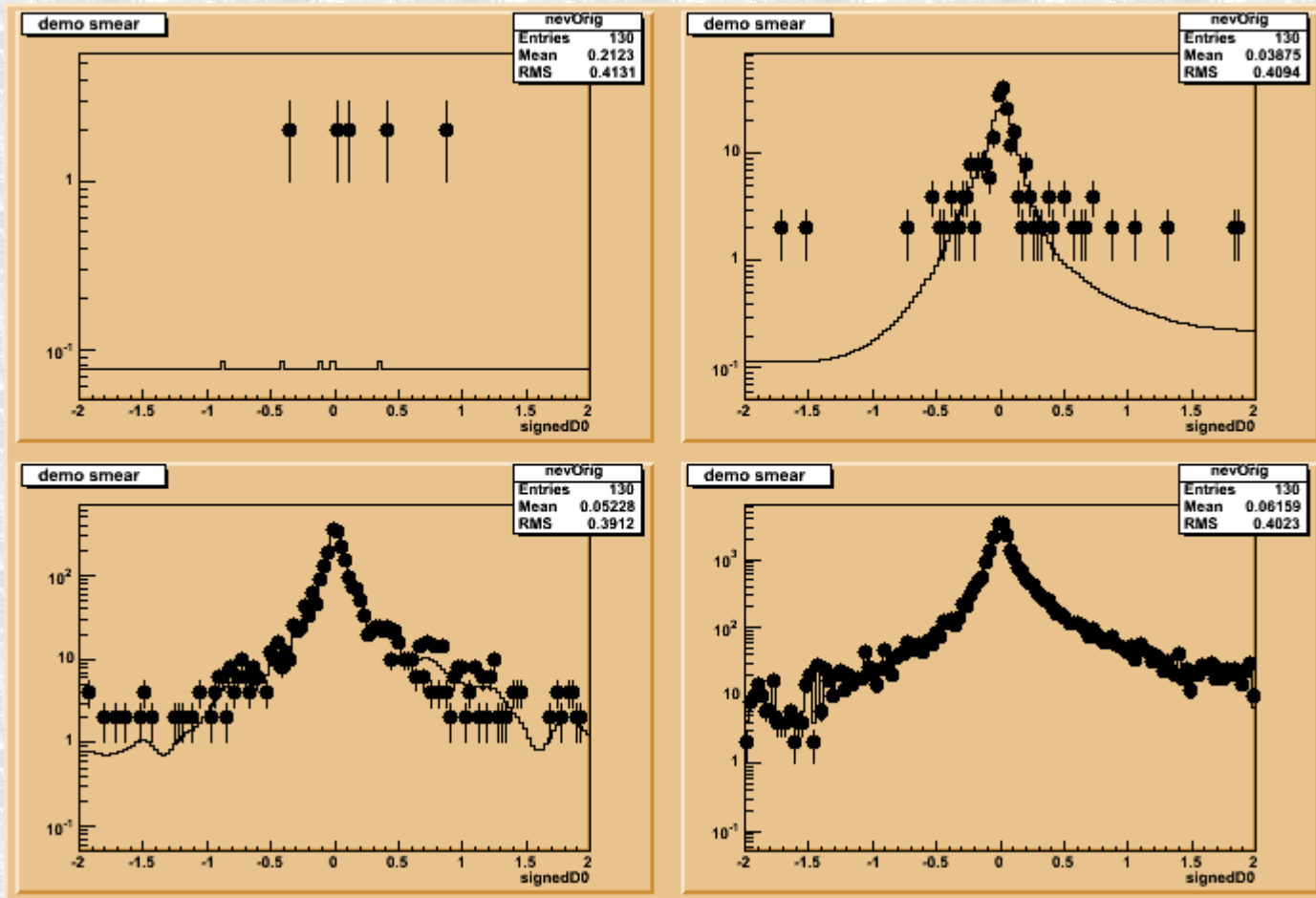
Smearing 5 events, 1D



- The kernel width is crucial
- Problem much worse in higher dimensions



Varying event nos.



- More events always helps



Kernel Likelihoods: nDim

- Metric of kernel, (**size, aspect ratio**) hard to optimize
 - Watch kernel size dependence on stats.
- Kernel size must grow with dimensions;
 - Lose precision if unnecessary dimensions added
 - Need to choose which variables to use
- Big storage/computational requirements



Taming m^n

Use of an approximately sufficient statistic or likelihood estimate

- No large resolution and acceptance effects:

Perform fit with uncorrected data and undistorted likelihood function.

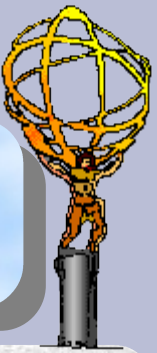
- Acceptance losses but small distortions:

Compute global acceptance by MC and include in the likelihood function.

- Strong resolution effects:

Perform crude unfolding.

All approximations are corrected by the Monte Carlo simulation. The loss in precision introduced by the approximations is usually completely negligible.



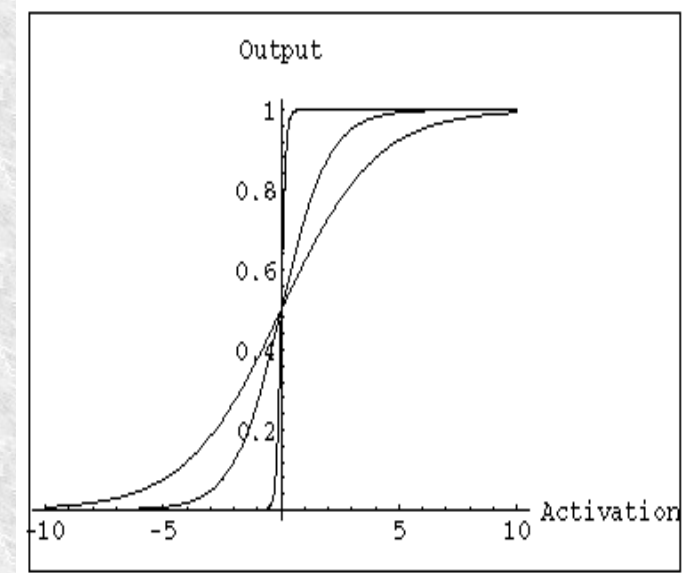
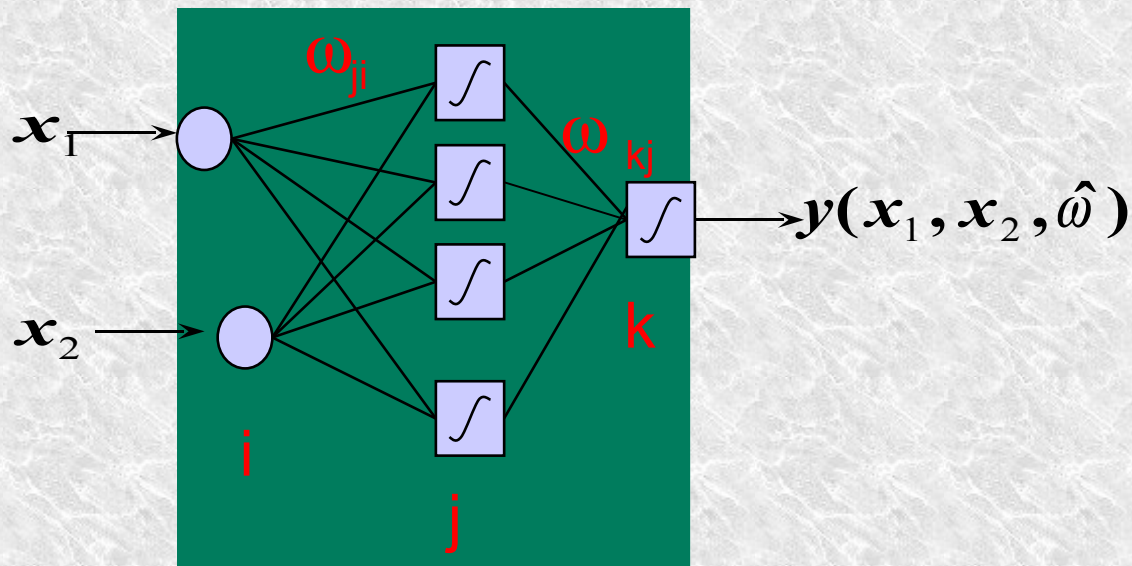
Neural Nets

- Well known in HEP



Multi-Layer Perceptron NN

- A popular and powerful neural network:
(in root)



$$F = \sum_j \omega_{kj} f\left(\sum_i \omega_{ji} x_i + \theta_j\right) + \theta_k;$$

$$y = \frac{1}{1 + e^{-F}}$$

Need to find ω 's and θ 's, the free parameters of the model



Neural Network Features

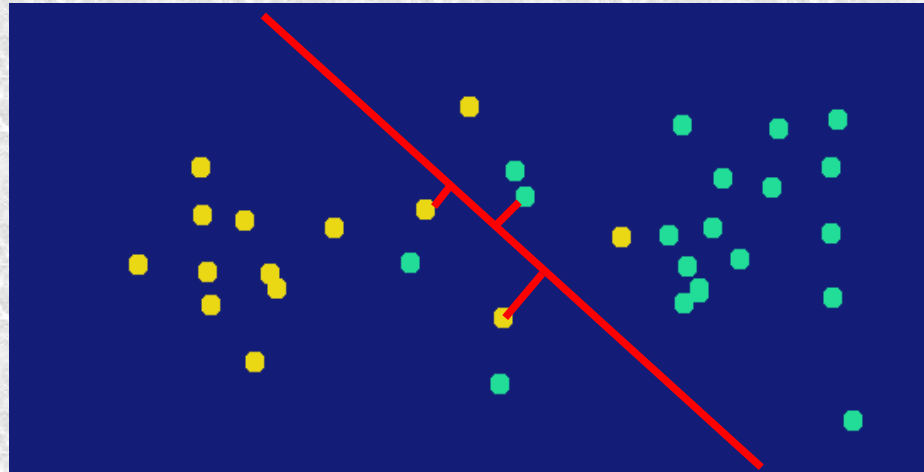
- Easy to use – packages exist everywhere
- Performance is good
 - Especially at handling higher dimensionality
 - No need to define a metric
 - But not optimal (we aim to approx. likelihood)
 - And only trained at one point
- Training is an issue
 - Optimization of nodes/layers can be difficult
 - Can over-focus on fluctuations
 - This is a problem for all machine learning
- Often worth a try
 - But it is solving the wrong problem



Support Vector Machines

Vapnik 1996

- Simplify storage/computation of separation by storing the 'support vector'

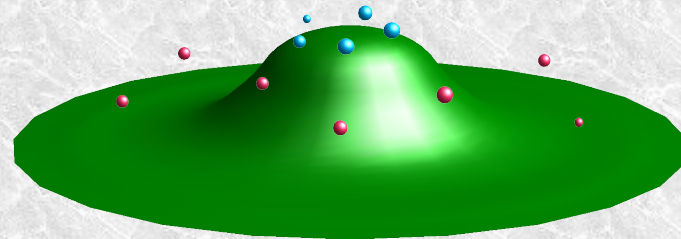


- The straight line is defined by closest points – the support vectors



Straight lines???

- Straight lines are not adequate, in general. Trick is to project from observed space into higher (infinite?) dimensionality space, such that a simple hyperplane defines the surfaces



- Projection done implicitly by kernel choice
- Only inner-products are ever evaluated, and these are metric independent, so can be calculated in normal space.
- Never need to explicitly define the higher dimensional space



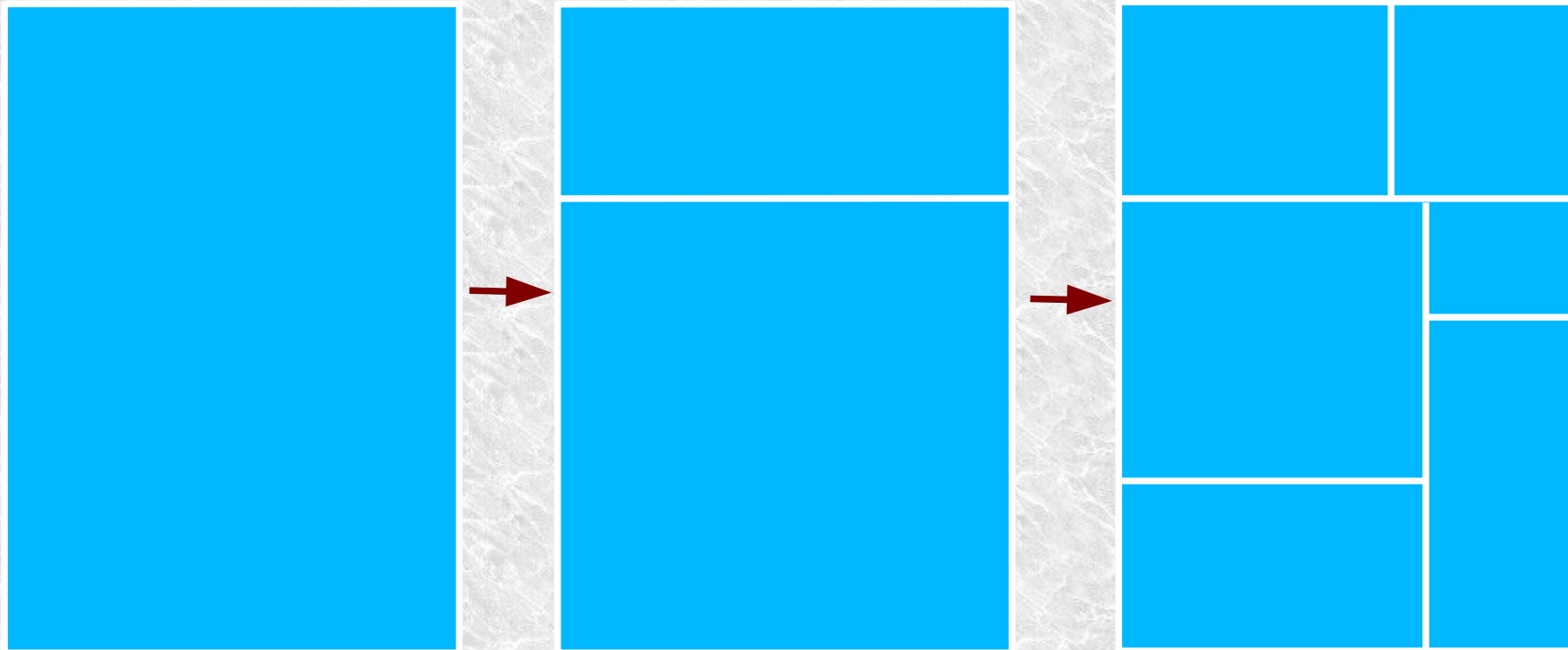
Support Vector Machines

- Storing just those points lying closest to the line is much easier than storing the entire space
- Defect is that the cut is only well defined near the line
- Computationally much easier than kernel likelihood



Decision Trees

- A standard decision tree divides a problem in a series of steps.

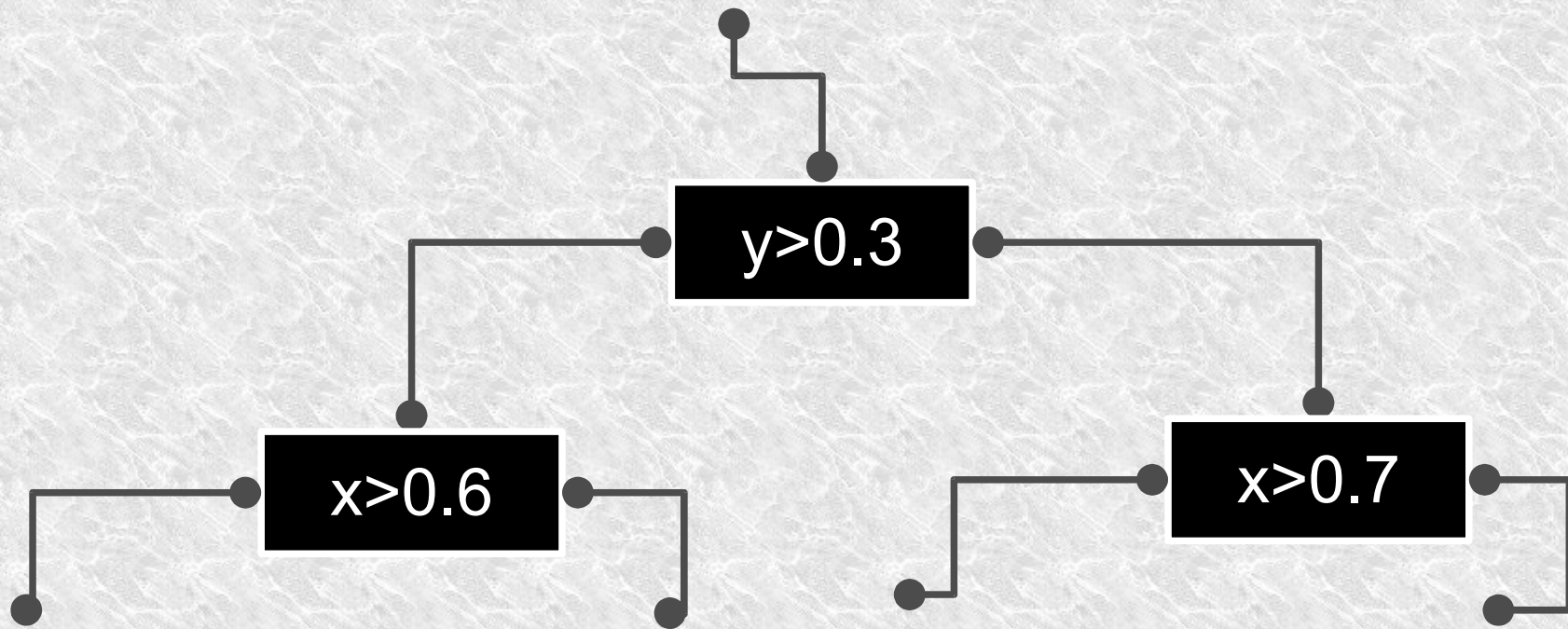


Signal/background evaluated in each box



Decision Trees II

- Very clear: Each decision is binary, and whole tree can be represented as a tree:



This display works for any dimensionality of problem
But how much do we value clarity?



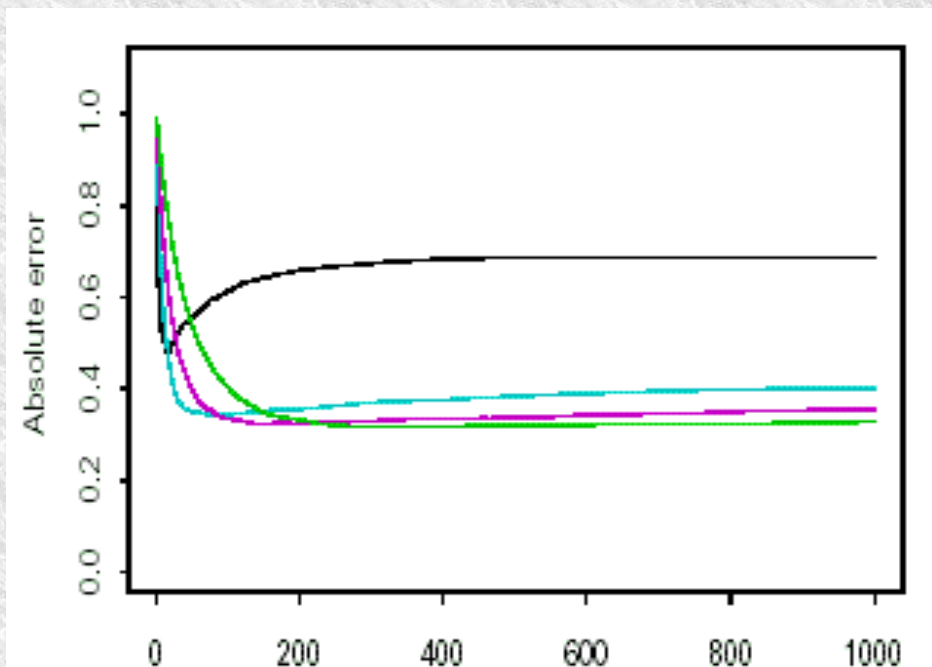
Decision Trees III

- Downside is lack of stability
 - First 'cut' affects all later ones, classification can vary widely with a different training set
- Power somewhat below NN/Kernel likelihood for typical HEP problems
- Stopping rule important, affected by sample size.



Boosted Decision Trees

- A first tree is made
- Add more, constrained not to mimic existing trees
- Final s/b is average (in some sense) over all trees.



Number of trees

Black, cyan, purple and green reflect increasing down-weighting of new trees



Boosted Decision Trees

- Lack of stability removed by averaging
- Computer intensive – but not N^3
- Power very good
- Trees individually small, whole data set is in each – good use of statistics
- Fairly fast.

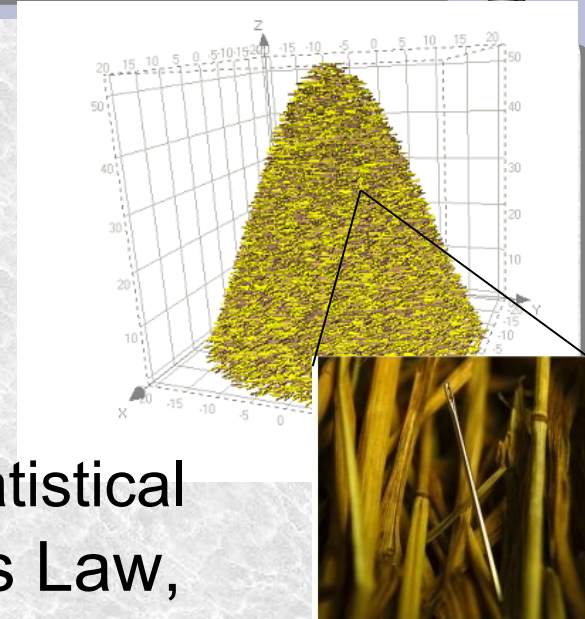
Breiman: Boosted trees best off the shelf classifier in the world

I do not have direct experience here



Summary of classification:

- Looking for
 - Needles in haystacks – the Higgs particle
- Needles are easier than haystacks
- ‘Optimal’ statistics have poor scaling
 - likelihood techniques N^3
 - For large data sets main errors are not statistical
- As data and computers grow with Moore’s Law, we can only keep up with $N \log N$
- A way out?
 - Discard notion of optimal (data is fuzzy, answers are approximate)
 - Don’t assume infinite computational resources or memory
- Requires combination of statistics & computer science



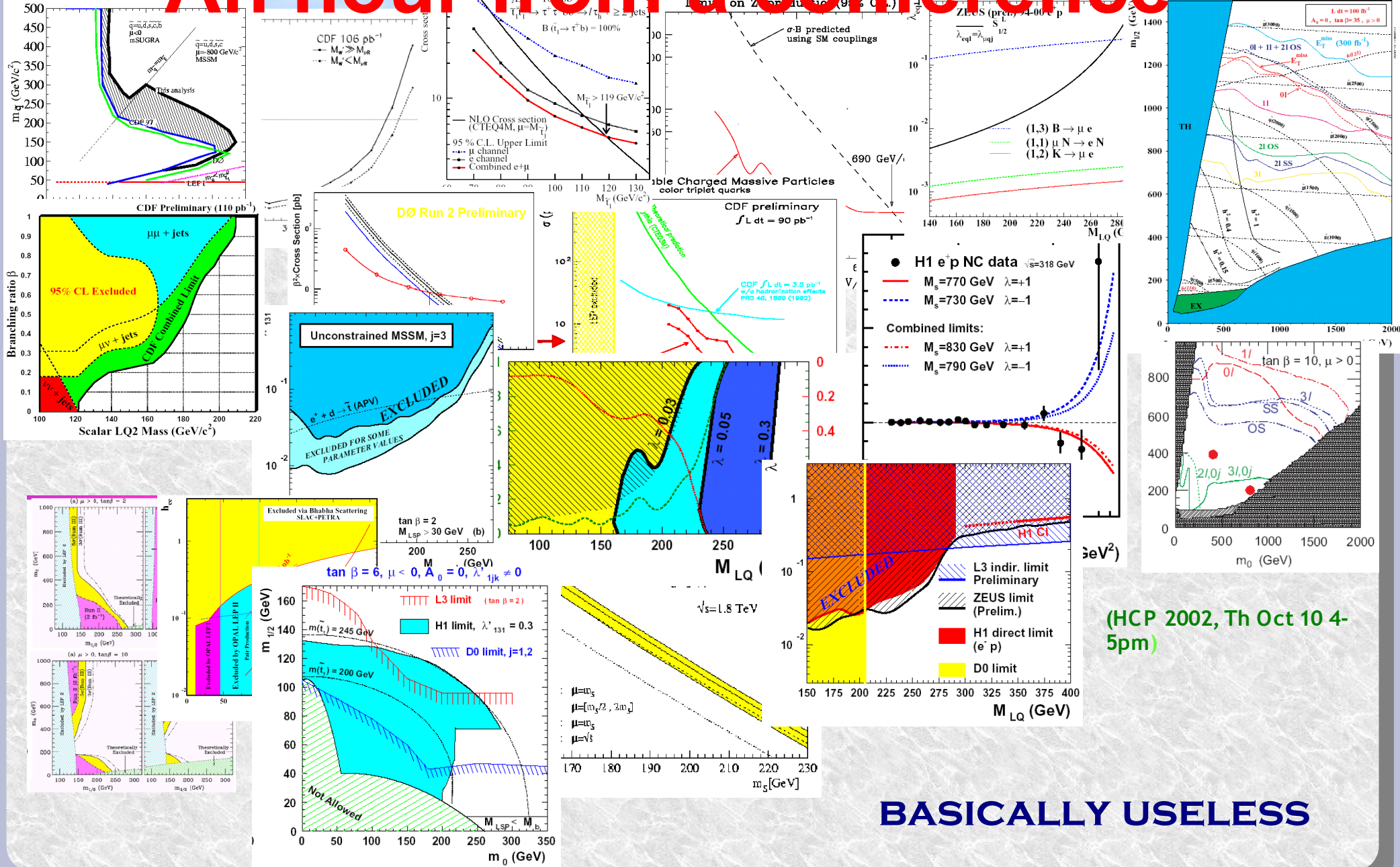


Limits

- Definition of a limit
 - Bayes vs Frequentist
- Errors on limits
- Combining limits
- You need to know: What is being limited?



An hour from a conference



(HCP 2002, Th Oct 10 4-5pm)

BASICALLY USELESS



Bayes and Frequentist Statistics

- Bayes confidence level
 - Probability of theory given data
 - `There is a 5% probability $X < 0$ `
 - Right question, but **subjective**
- Frequentist/Classical P-value
 - Probability of data given theory
 - `If $X < 0$, probability of these (or more extreme) data is 5% or less`
 - Wrong question



Bayes Theorem

- Bayes theorem modifies probabilities in light of data
- But needs an a-priori probability

$$p(a \wedge b) = p(a) p(b|_a) = p(b) p(a|_b)$$

$$p(b|_a) = \frac{p(b)}{p(a)} p(a|_b)$$

Uncontroversial

Requires p(theory)

$$p(\text{theory}|_{\text{data}}) = \frac{p(\text{theory})}{p(\text{data})} p(\text{data}|_{\text{theory}})$$

- If data plentiful frequentist and Bayesian converge
- Bayesian statistics much easier to use



Frequentist limits

'For this theory, probability of these data is $<5\%$ '

- About as useful as:

'It was raining when I went to the theatre'

- e.g. a Higgs Search, background=3, observed=0
 - S=3: P = 0.2%
 - S=1: P=1.8%
 - S=0: P=5%
- So even a production of 0 Higgses is excluded at 95% CL. ($M_H=1000$ excluded!)
- Mathematically fine, but not useful



Bayes vs Frequentist II

- For measurement, almost all results are implicitly Bayesian - but could be justified frequentistically.
- Limits are much less clear - **read the small print**
 - Many modifications (e.g. Feldmann & Cousins, RPP 2000) try to give Bayesian properties to classical limits
 - Without such modification, frequentist limits can be meaningless. (e.g. Example on previous page)

Fundamental problem:

$$P(\text{data}|\text{hypothesis}) \neq P(\text{hypothesis}|\text{data})$$

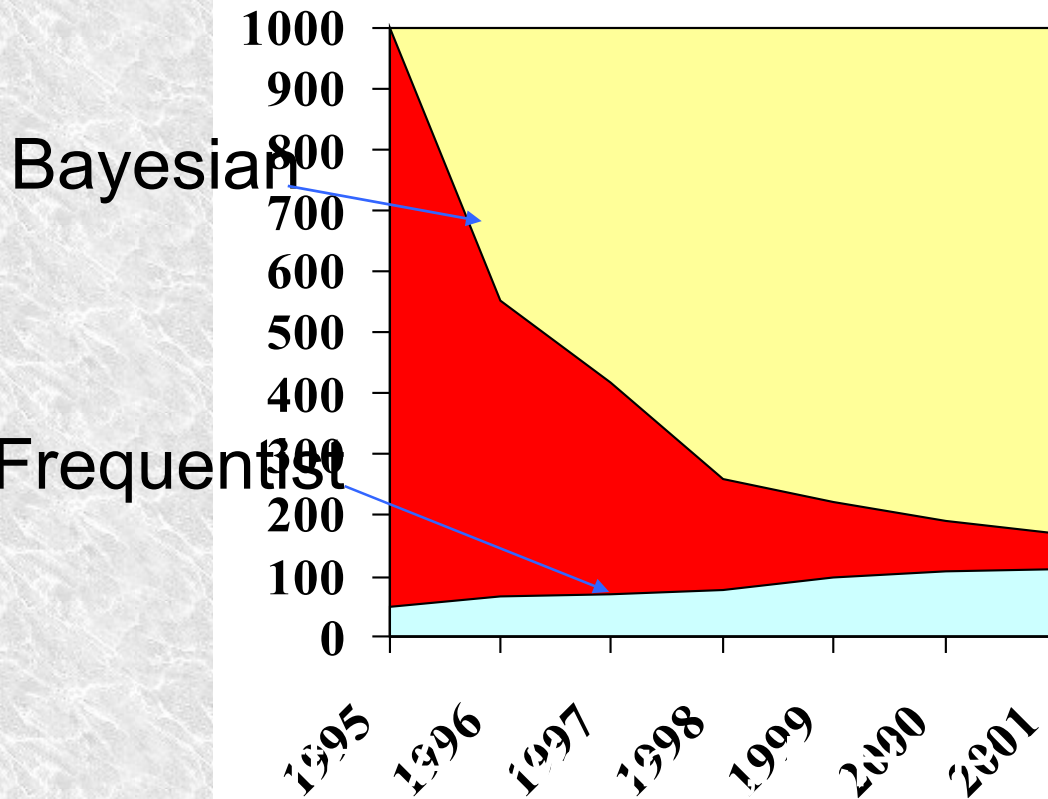


Error on a limit

- Sometimes people ask: What is the error on this limit?
- *This is the wrong question, it hides two others:*
 - How different would the limit be if
 - the data was a bit different?
 - you quoted 90% or 99% instead of 95%
- It is wrong, because $M_H > 113 \pm 1$ helps no-one. $M_H = 115 \pm 1$ is useful



Mixed up results!



Higgs limits

EW fits assume a Higgs

Search looks for one

- Excluded by EW fits
- Allowed
- Excluded by Direct Search

To be fair, he had no choice - this is what is provided



So what does a limit mean?

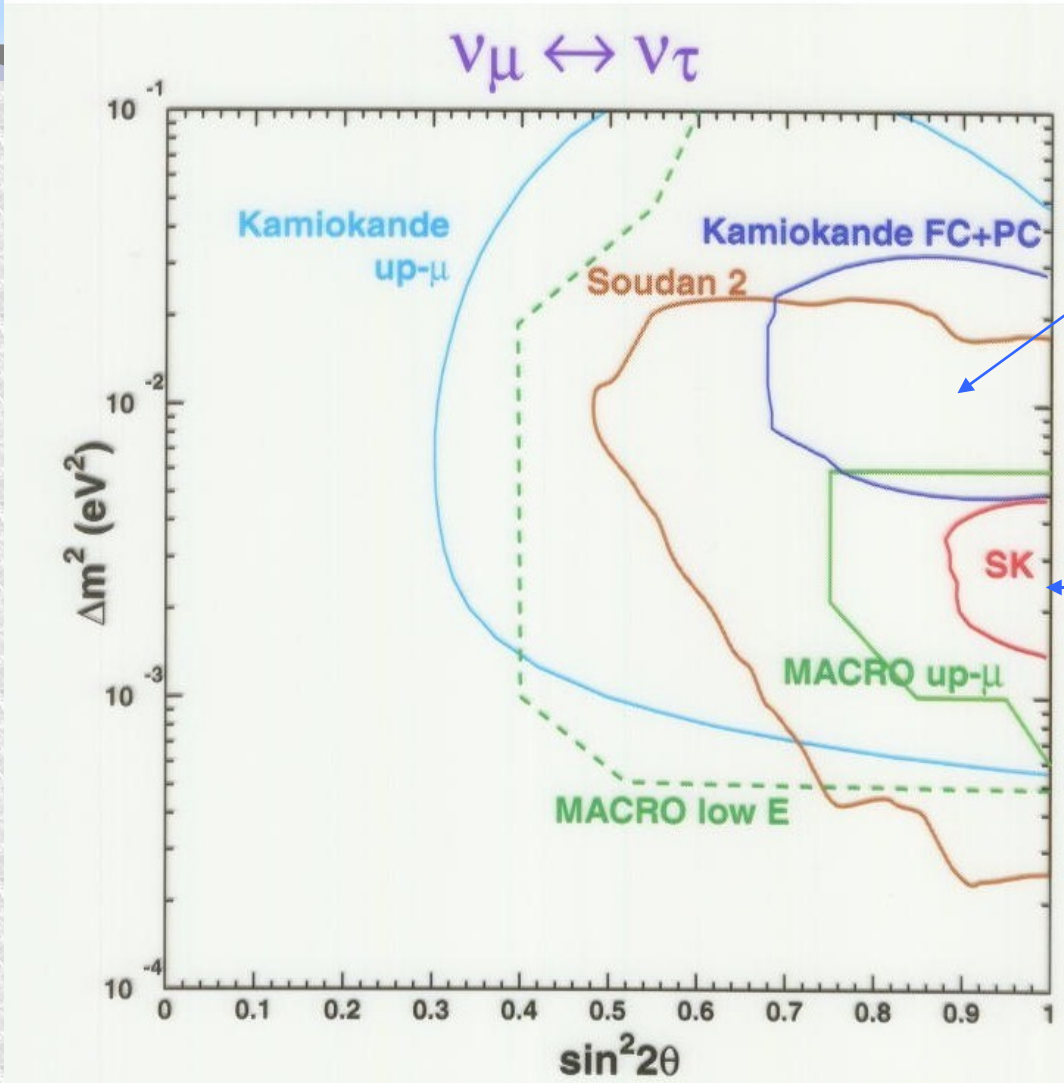
- Usually, a statement like:

$$M_H > 113.2 \text{ GeV}/c^2 @ 95\% \text{ CL}$$

- Does NOT Mean that there is a 95% probability that $M_H > 113.2$
- DOES mean that **IF** $M_H < 113.2$ **then** there was at most a 5% probability we missed it.
- But you should not CARE anyway - it is a *probability*



Combination of limits



Kamiokande 90%

Super-K 90%

Results do not overlap! How can you combine?

Please, don't do it!

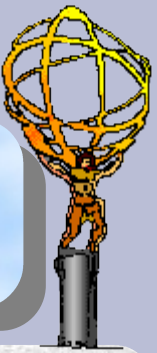


Parameter Extraction



Statistics and Systematics

- Data are collected very often in terms of **numbers of events**
- These are converted into cross-sections; rates at which things happen.
- This is true of almost all HEP measurements
- The statistical precision is fairly straightforward to estimate, and will have a Poisson or Binomial form which can often be approximated with a Gaussian (CLT)
- Various systematic effects will also affect the result.



Evaluation of Systematics

- **Usually the hardest part of a measurement**
- **Frequently done badly**
- **Can be under- or over- estimated**
 - **Sometimes an error source is forgotten**
 - **Frequently statistics get taken twice**



Systematics

- Typical experimental statements:
 - $M_W = 80.336 \pm 0.055(\text{stat}) \pm 0.028(\text{sys}) \pm 0.025(\text{FSI}) \pm 0.009(\text{LEP})$
(DELPHI)
 - $M_W = 80.329 \pm 0.029$ (EW group)
- First has statistical and systematic errors
- The second does not. **Why?**



Systematics in combination

- Take an example: Two measurements of 'x'

- $x_a = 50 \pm 10(\text{stat}) \pm 1(\text{sys})$ (expt. a)

- $x_b = 60 \pm 1(\text{stat}) \pm 10(\text{sys})$ (expt. b)

- To combine, use:

Correlations?

$$\frac{x_{tot}}{\sigma_{tot}^2} = \sum_i \frac{x_i}{\sigma_i^2}$$

$$\frac{1}{\sigma_{tot}^2} = \sum_i \frac{1}{\sigma_i^2}$$

- $x_{tot} = 55 \pm 7.11$ (total)

- But either stat. or syst. combined separately is <1!

Errors are completely mixed in combination
be evaluated on equal footing

- must



Why is fitting important?

- Experimental data usually used to obtain theoretical parameter values
- There are two distinct questions:
 - What is the best value of X (Parameter optimization)
 - Does the theory explain the data (Goodness of fit)
- The use of in very different

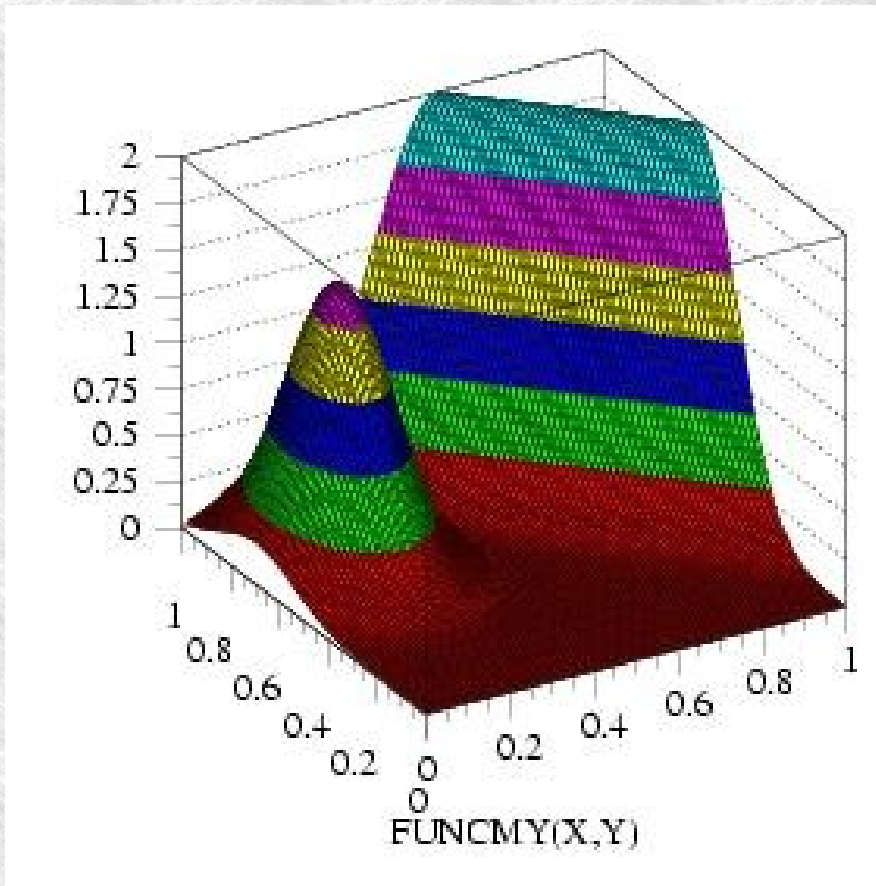


Parameter optimisation

- What value of the free parameters best describes the data?
- **'MINUIT'** from CERNLIB a very common minimization package
- User must specify WHAT is minimized (χ^2 , likelihood, etc.)
- **MINUIT** will (usually) find the optimal parameter set
- Complex functions may have **secondary maxima / convergence failure**



Example of a hard maximisation



- Subsidiary minimum may confuse algorithm
- Primary maximum has very high parameter correlation
- Likely convergence failure



The right way to fit?

	χ^2	<i>likelihood</i>
<i>Binned</i>	Yes	Yes
<i>Unbinned</i>	No	Yes
<i>Works low stats</i>	No	Yes
<i>Unbiased</i>	Sometimes	Yes
<i>Goodness-of-fit</i>	$P(\chi^2)$	Hard

The techniques converge for high statistics

Both can be handled by 'Minuit'

No *right* way



Definition of χ^2

→ Compatibility of results with expectation:

$$\chi^2 = \sum_i \frac{(x_i^{obs} - x_i^{pred})^2}{\sigma_x^2}$$

→ If counting events in bins then:

→ Beware: Is $\sigma \propto \sqrt{N^{obs}}$ or $\sqrt{N^{pred}}$? - Both are wrong!

→ $\sqrt{N^{obs}}$ Biased down if data down... (Mean 5 & 3 = 3.75)

→ $\sqrt{N^{pred}}$ Error depends upon theory - biased up (~4.12)



Definition of likelihood

$$L_r = e^{-r} \times \prod_i R_i$$

$$LR = \frac{L_{s+b}}{L_b} = e^{-s} \times \prod_i \frac{S_i + B_i}{B_i}$$

→r: The total rate

→R_i: The density at point i

→s: signal

→b: background

→Likelihood ratio compares two hypotheses.

→Or vary R to maximize L. Maximum likelihood powerful estimator.

$$\log(L_R) = -r + \sum_i \ln R_i \quad \leftarrow \text{weighted sum of events.}$$



Comparison of definitions

$$\chi^2 = \sum_i \frac{(x_i^{obs} - x_i^{pred})^2}{\sigma_x^2}$$

$$\log(L_R) = -r + \sum_i \ln R_i$$

Take a Gaussian centred on 0, width σ

$$\chi^2 = \sum_i \frac{x_i^2}{\sigma^2}$$

$$\log(L_R) = -r + \sum_i \ln \left(e^{-0.5 \frac{x_i^2}{\sigma^2}} \right)$$

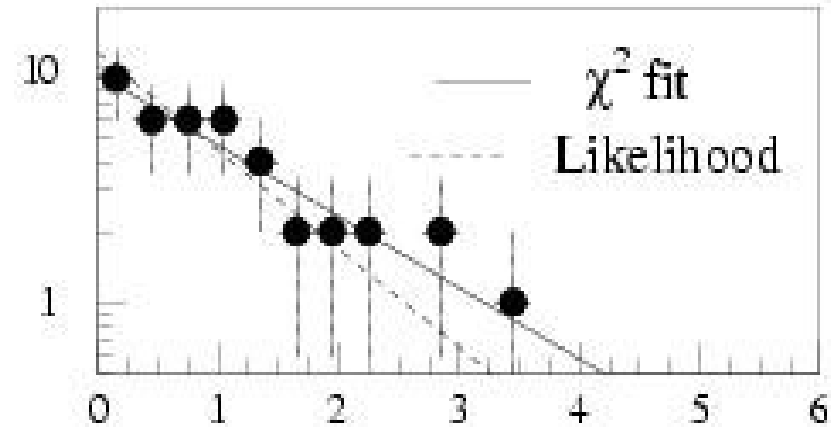
$$\log(L_R) = -r + \sum_i -0.5 \frac{x_i^2}{\sigma^2}$$

So: $\delta\chi^2 = -2*\delta\log(L_R)$ **IF GAUSSIAN**



Example: Fitting a lifetime

- Exponential function
- Events sparse at the end by definition
- A χ^2 fit will need careful binning

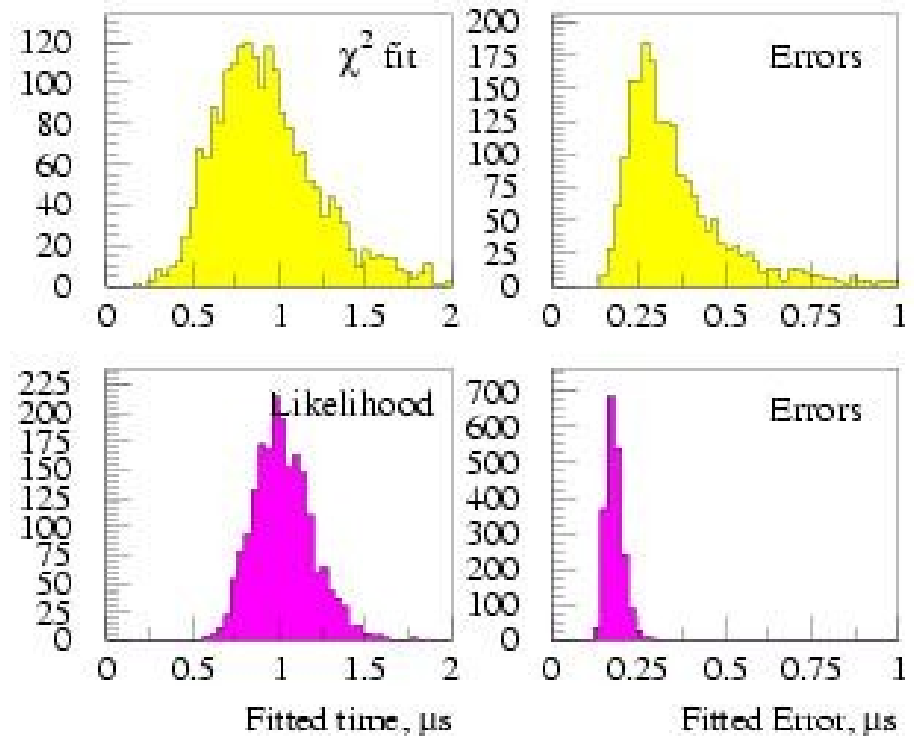


- Likelihood can give events individual errors
 - Useful IF errors vary a lot



Fitting a lifetime: II

- Use $1\mu\text{s}$ exponential
- Likelihood unbiased, reliable
- The χ^2 biased and variable. (Using default 'PAW' settings)



This is an example! Sometimes χ^2 is better



Goodness of fit



Averaging two numbers

- Suppose we have two measurements of x :
 10 ± 1 and 11 ± 1
- We know the average:
 10.5 ± 0.7
- But what about
 10 ± 1 and 20 ± 1
- Are we happy with:
 15.0 ± 0.7
- If the errors are Gaussian we should be happy
 - Combining two Gaussians gives a Gaussian
- Or we conclude we have a 'bad fit'



Goodness of fit

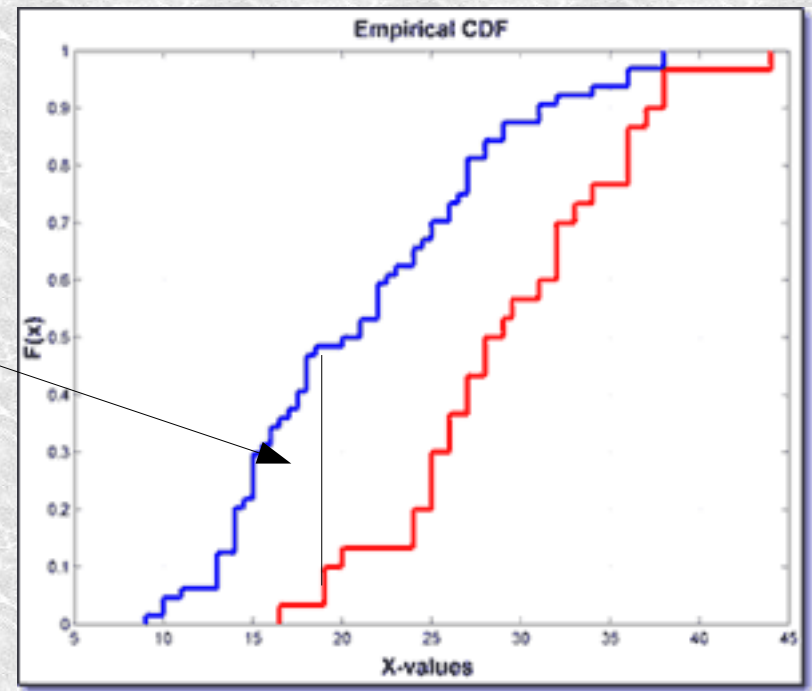
Do I believe my fitted results?

- For a χ^2 fit (much the most common) easy
 - Look up $\text{Prob}(\chi^2, \text{NDF})$ in a table.
 - Good general test
 - often abused, e.g. χ^2/DoF
- For a likelihood fit hard
 - Can sometimes use simulated trials to find Probability of getting observed result OR 'larger' one



Kolmogorov-Smirnov Test

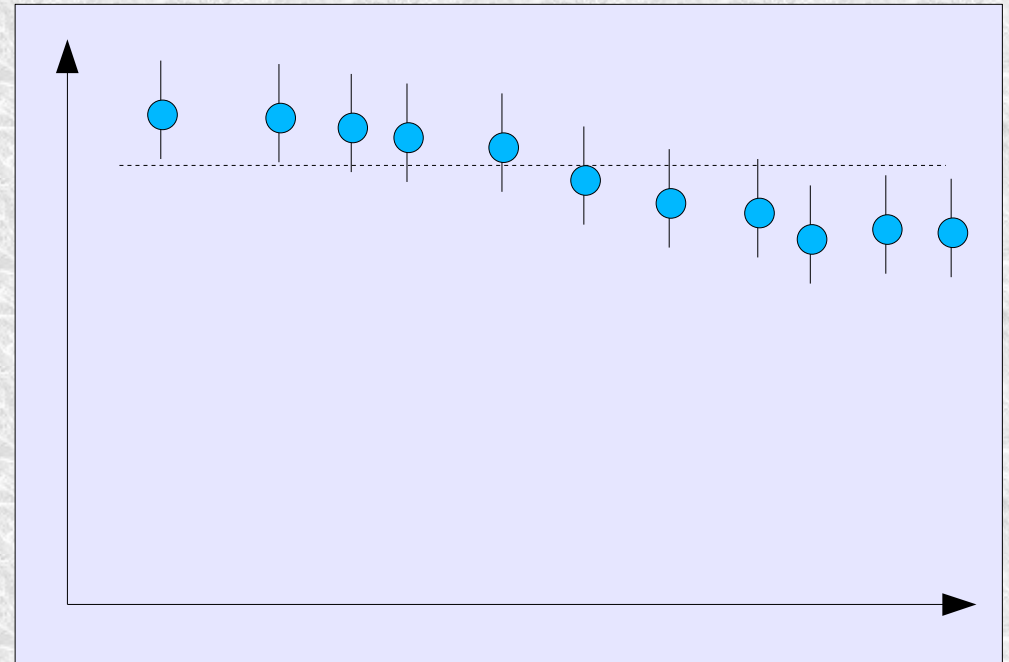
- Are two 1D distributions drawn from a common parent?
- Find integral distribution
- Observe maximum difference between two integrals
- Probability is calculable
- Very bad here!
- K-S very good
 - No allowance for fit
 - Or systematic error

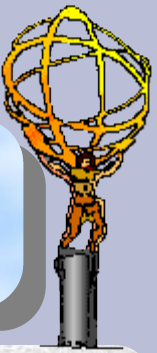




Run test

- Measures 'runs': How often one distribution is above (or below) the other
 - Find length of maximum run
 - Probability of this length is calculable
- This data has good χ^2
- Run test would be poor





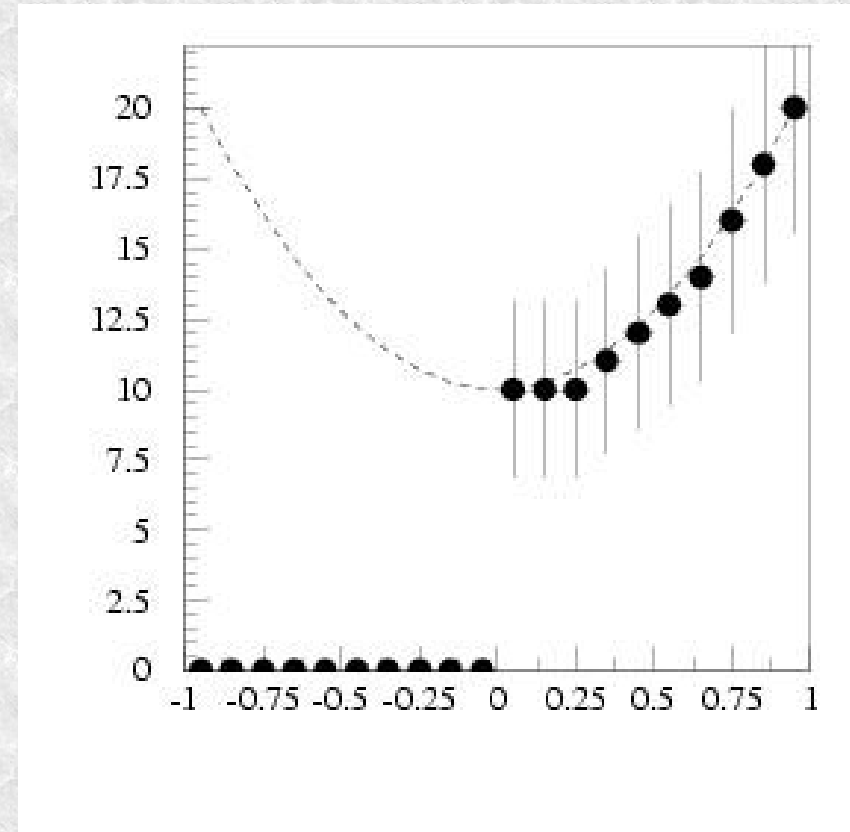
Is there a better test than χ^2 ?

- Yes...and no
- For any given problem, a more sensitive test can be defined
- But you can only define the sensitivity if you know what you are choosing between
- There is no 'most powerful' goodness of fit test.



How things can go wrong...

- Example: Fit function
$$y = a + b \cos \theta^2$$
- Likelihood fit is not sensitive to sign on θ
- Concludes that distⁿ shown is very good fit!
Nothing is perfect...





Neutrino Oscillations:

- Small Mixing Angle
- Large Mixing Angle
- 'Low' solution

Seen as hypotheses

- But V_{MNS} is a matrix of parameters to optimise !
- Use $p(\delta\chi^2, npar)$, not $p(\chi^2, DoF)$



χ^2 or $\delta\chi^2$?

- Suppose NDF=30 and
 - SMA $\chi^2=36$
 - LMA $\chi^2=28$
- Then $P(\chi^2, \text{NDF})$:
 - SMA: 21%
 - LMA: 57%
- But, probability of $\delta\chi^2$ is 1.8%
- So SMA is excluded in this example
 - Shows that χ^2 is a blunt instrument



Conclusions

- The likelihood ratio underpins everything
 - Use it
- Cost of computing becoming important
 - Optimal methods not *necessarily* optimal
 - But the data is very expensive too.
 - Systematic errors may dominate anyway
- Need to see statistics as a tool
- Please:
 - Ask me if you have statistical questions