Statistical methods

Jim Libby IITM

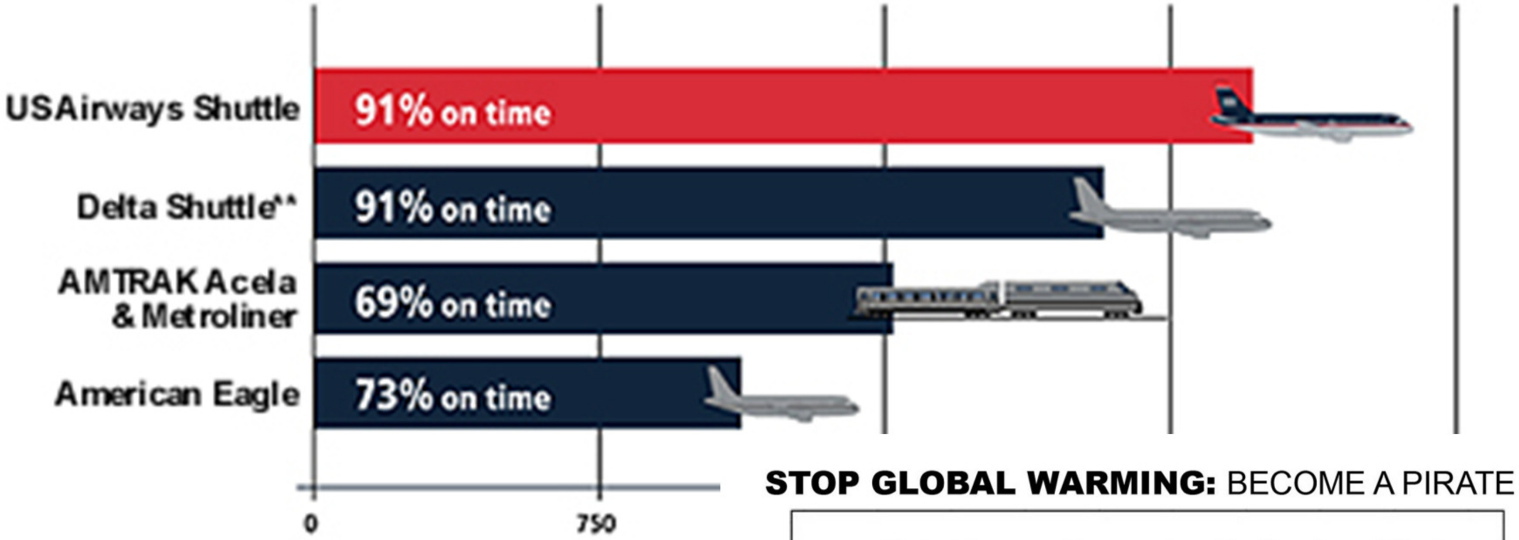
Contents and resources

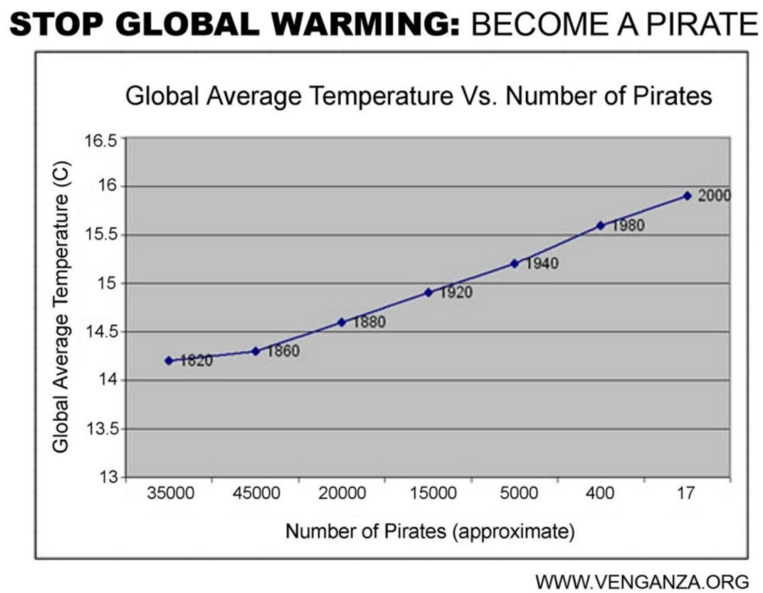
1. Probability distributions
2. Statistical and systematic uncertainties
3. Estimators - fitting
4. Probability and confidence intervals
5. Multivariate techniques – will not be covered – instead a tutorial

- **Statistics – R. J. Barlow (John Wiley & Sons)**
- A Practical Guide to Data Analysis for Physical Science Students – L. Lyons (Cambridge University Press)
- Leo - chapter 4
- Data analysis techniques for HEP, Fruhwirth et al (Cambridge University Press)
- Particle Data Group, Review of Particle Properties, Sections 35 and 36
- SLUO lectures on statistics (Frank Porter and Roger Barlow) http://www-group.slac.stanford.edu/sluo/lectures/Stat_Lectures.html
- RooFit: http://roofit.sourceforge.net/
- RooStats: https://twiki.cern.ch/twiki/bin/view/RooStats/

Lies, damn lies and statistics

Benjamin Disraeli (British politician, 1804-1881)





http://en.wikipedia.org/wiki/Flying_Spaghetti_Monster

Why you need to know some statistics?

- Most of you will be measuring some parameter during your graduate studies
  - branching fraction
  - mass
  - coupling
  - differential cross section
  - .....
- In reality you never measure a single value but an interval expressed as

$$R_{DK} = [\underbrace{1.98}_{\text{Central value}} \pm \underbrace{0.62}_{\text{Statistical uncertainty}} \pm \underbrace{0.24}_{\text{Systematic uncertainty}}] \times 10^{-2}$$

[M. Nayak et al., (Belle Collaboration), PRD **88**, 091104]

Which is the <u>most</u> important number?

You will need

1. A method of estimating the central value and its related statistical uncertainty in a **consistent**, **unbiased** and **efficient** way
2. To identify sources and estimate the magnitude of the systematic uncertainty
3. Combine measurements and uncertainties, even if they are **correlated**
4. Interpret your result **degree of belief/confidence** in your result
   - all intervals correspond to some probability
   - $\pm$ one standard deviation should indicate that if you repeat your measurement many times your result will lie in that range 68% of the time
   - Interpret your result

Will try to give a flavour of how to go about the above

- **But statistical methods are tools, which you must learn to use practically**
- **'A bad workman blames his tools'**

Part 1

PROBABILITY DISTRIBUTIONS

Assumed knowledge/revision I

- Classical definition of probability
  - If I toss an unbiased coin many times the no. of heads divided by number of tosses $\rightarrow$1/2 $\equiv$ Probability of a coin toss giving heads
- Definition of mean and standard deviation for a sample and distributions (discrete and continuous)

$$\text{Sample: } \bar{x} = \frac{1}{N}\sum x_i \quad \sigma = \sqrt{V(x)} = \sqrt{\overline{x^2} - \bar{x}^2}$$

$$\text{Discrete distribution: } \langle r \rangle = \sum_r rP(r) \quad \sigma = \sqrt{\langle r^2 \rangle - \langle r \rangle^2}$$

$$\text{Continuous distribution: } \langle x \rangle = \int xP(x)\,dx \quad \sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$$

- A probability density function

$$P(x) = \lim_{\delta x \to 0} \frac{\text{Probability result lies between } x \text{ and } x + \delta x}{\delta x}$$

Assumed knowledge/revision II

- Covariance and correlations

$$V_{ij} = \mathrm{cov}(x_{(i)}, x_{(j)}) = \overline{x_{(i)}x_{(j)}} - \overline{x_{(i)}}\;\overline{x_{(j)}}$$

If $x_{(i)}$ and $x_{(j)}$ independent this is zero

$$V_{ij} = \rho_{ij}\sigma_i\sigma_j$$

where $\rho_{ij} \in (-1,1)$ is the correlation coefficient

- Uncorrelated: pp vertex position and jet energy
- Partially correlated: electron energy and momentum
  - why only partially?
- Fully correlated: number of pp collisions and luminosity

IXth SERC School on EHEP

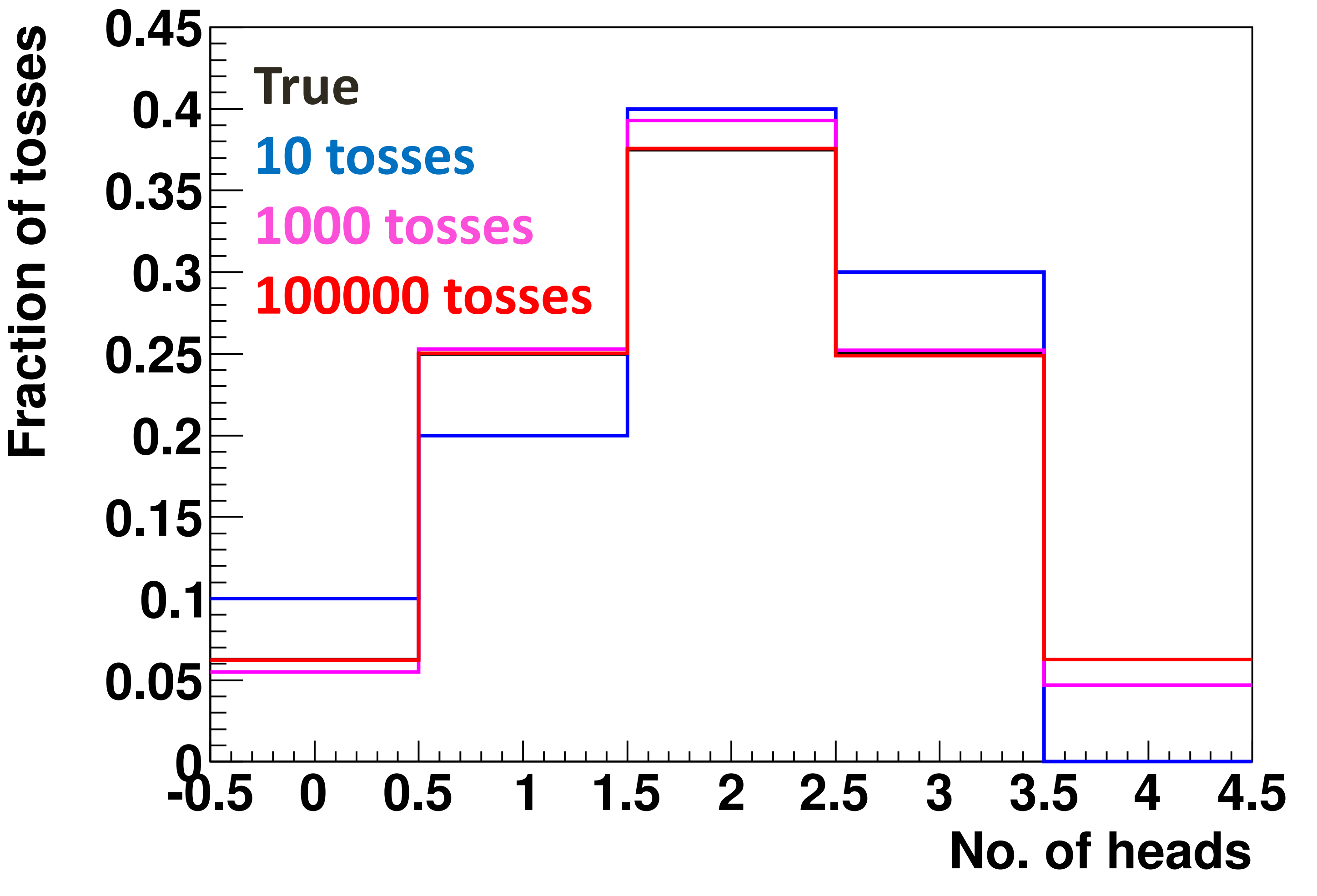
Law of large numbers

- Something we all know but it is worth emphasizing
  - We are always trying to measure some true parameter or distribution
  - However, a few pieces of data are unlikely to give you a good estimate of that parameter/distribution due to the fluctuations
  - Example: tossing a coin four times

No. of Heads	0	1	2	3	4
Probability	1/16	4/16	6/16	4/16	1/16

  - Now do the experiment and estimate the probability after
    - 10 tosses
    - 1000 tosses
    - 100000 tosses

Law of large numbers: example



Probability distributions

- We are now going to review four important ones which often describe physical processes of interest
  - Binomial
  - Poisson
  - Gaussian
  - Uniform
- Not exhaustive
  - Multinomial
  - Exponential – lifetimes
  - Breit-Wigner/Cauchy – resonances
  - Landau – dE/dx in a thin piece of material
  - Polynomials – particularly orthogonal sets - Legendre, Hermite, Chebyshev
  - χ^2

Binomial distribution

- Applies to pass-fail situations
  - Coin toss
  - **Event selection**
  - Forward-backward asymmetries (or similar)
- From *n* attempts there 2^n ways to put together the successes and failures
- The number of ways that contain *r* successes is

$$\text{Binomial coefficient} = {}_nC_r = \frac{n!}{r!(n-r)!}$$

- Therefore, probability of *r* successes from *n* trials is

$$P(r;p,n) = p^r(1-p)^{n-r}\,{}_nC_r$$

where *p* is the individual probability of success at each trial

Binomial distribution II

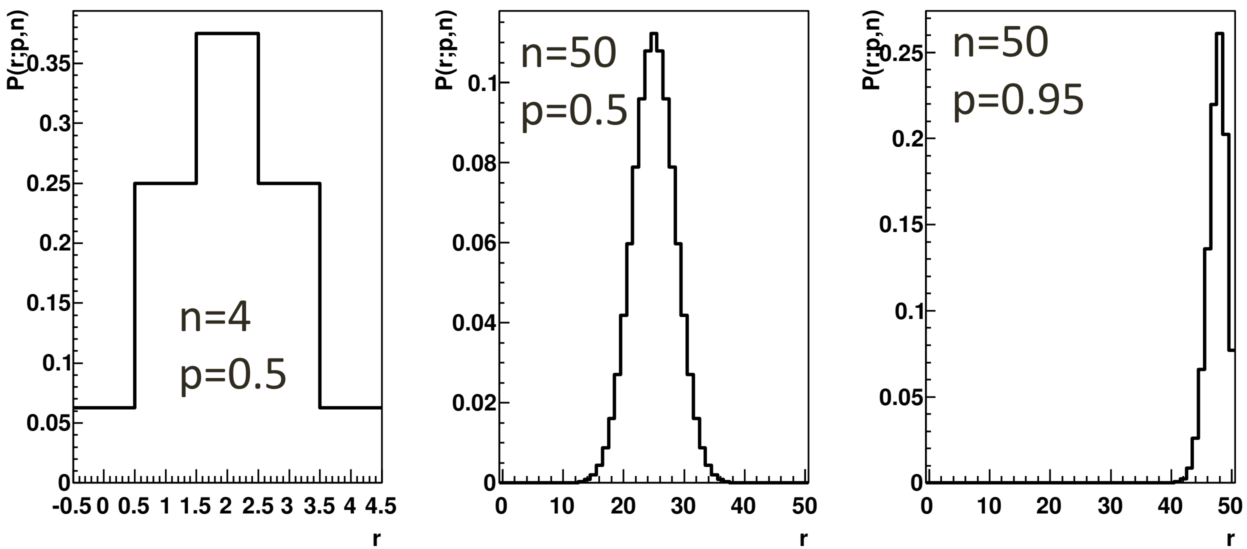
- Check of the total probability

$$\sum_{r=0}^{n} p^r (1-p)^{n-r} \, {}_nC_r = [p + (1-p)]^n = 1$$

- Mean and standard deviation

$$\langle r \rangle = pn \quad \text{and} \quad \sigma = \sqrt{np(1-p)}$$

- Example distributions



Binomial example: efficiencies

- Often you need to estimate a selection efficiency from a sample of simulated events:
  - Efficiency = ε = no. selected (m)/no. in sample (n)
- No. selected follows a binomial distribution
  - Therefore, uncertainty is

$$\sigma_\varepsilon = \sqrt{\frac{\varepsilon(1-\varepsilon)}{n}}$$

- Common mistake is to say error is sqrt(m)/n
  - 98% from a sample of 1000 events $\Rightarrow$
    - (98.0±3.1)% (efficiency greater than 100% !)
    - **(98.0±0.4)% correct binomial error**
    -

Poisson

- Describes the a case where there are still particular outcomes like the binomial but you don't know the number of trials
- Sharp events in a continuum of nothing happening
  - Radioactive decay
  - Flashes of lightening
  - **Signal produced in a collision**
- One knows the average number of events over some period
  - Want to know the probability of observing a given number in a certain period
- Analysis of binomial distribution, in which the number of trials n becomes large while the probability p becomes small but their product is constant
  - **On board**

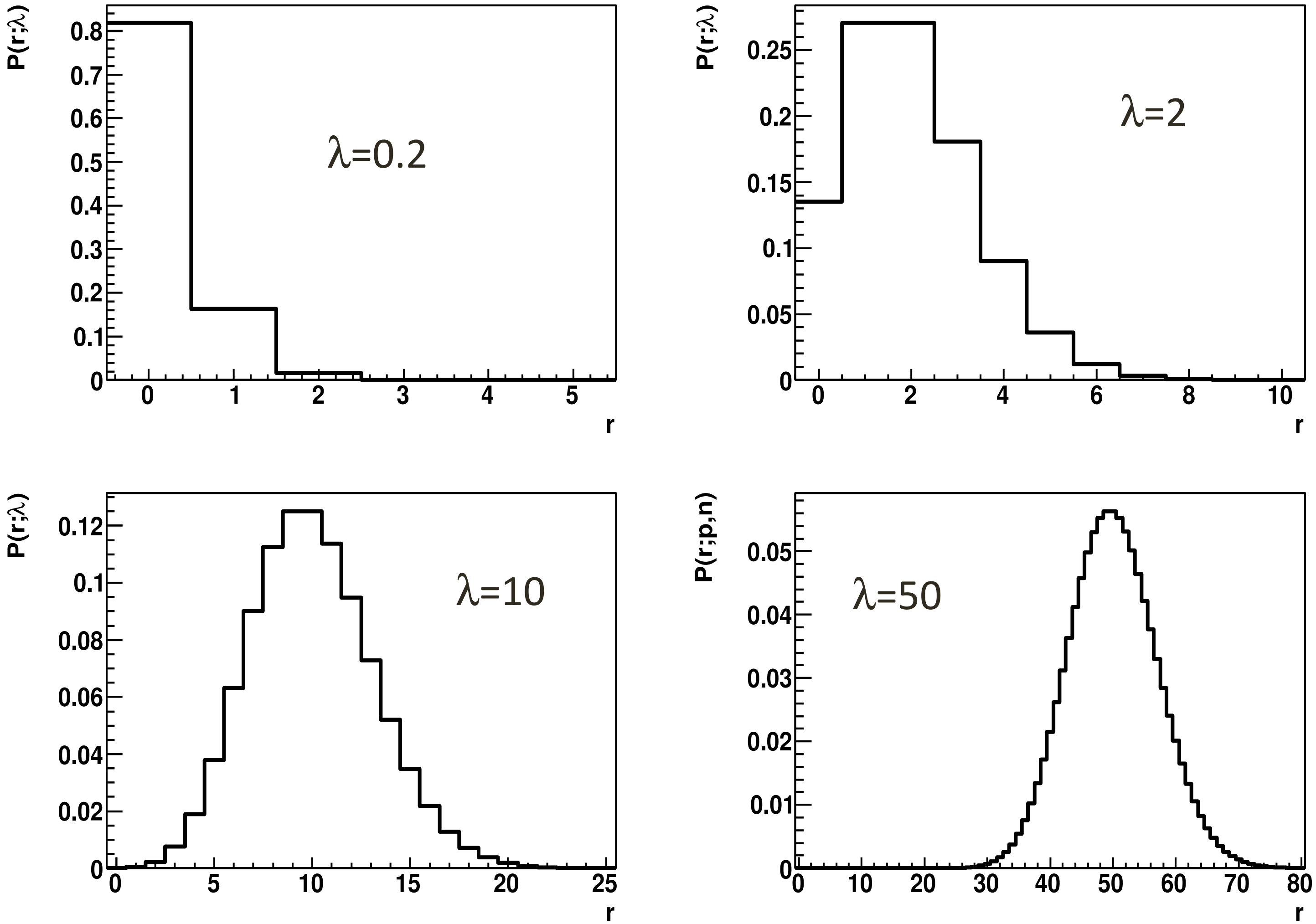
Poisson distribution

$$P(r;\lambda)=\frac{e^{-\lambda}\lambda^{r}}{r!}$$

mean

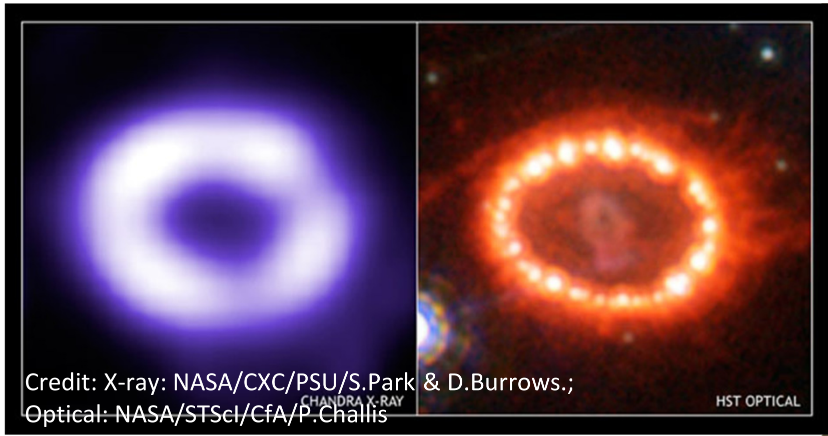
observed

$$\langle r\rangle=\lambda \text{ and } \sigma=\sqrt{\lambda}$$



Poisson example

Phys. Rev. Lett. 58, 1494–1496 (1987) and Barlow



No. ν	0	1	2	3	4	5	6	7	8	9
Intervals	1042	860	307	78	15	3	0	0	0	**1**
Predict.	1064	823	318	82	16	2	0.3	0.03	0.003	0.0003

Data collected in 10 second intervals on 23rd February 1987 around the time of the first observation of SN1987A

Ignoring the interval with nine neutrinos the average is 0.77

The Poisson prediction for λ=0.77 is given which is in excellent agreement with the observed counts

Therefore, probability that the interval with nine events is a fluctuation of the background rate is tiny

Gaussian

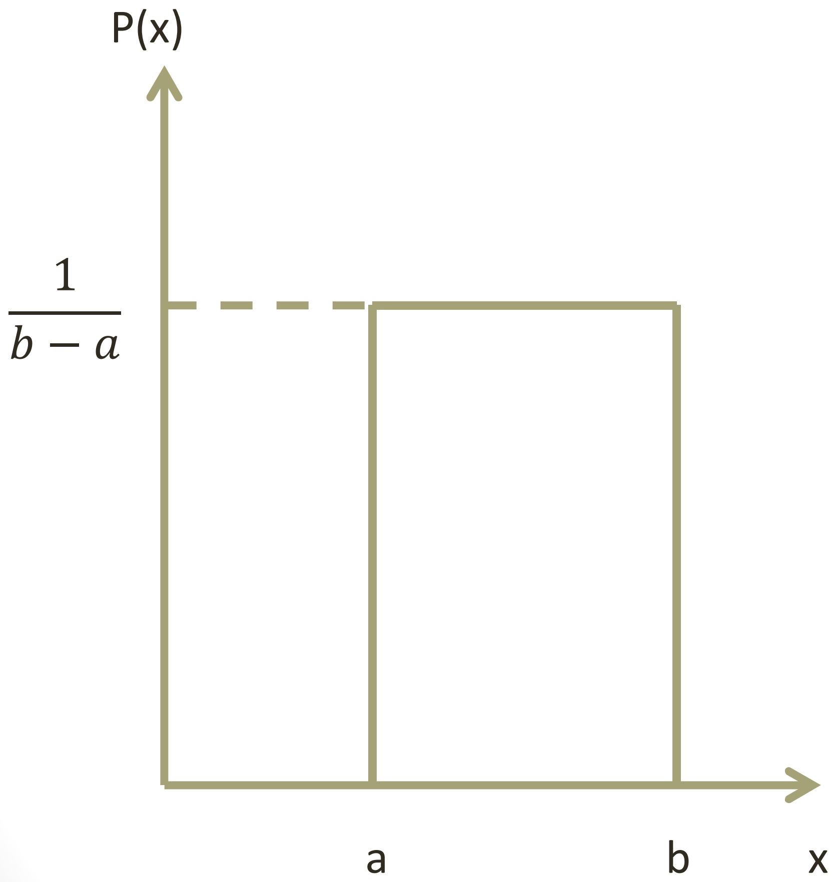
- The most common distribution – there is a reason which we will discuss in the next class
  - Unlike binomial or Poisson it describes a continuous distribution
- Appropriately normalised

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

$$\langle x\rangle=\mu \text{ and } \sigma=\sigma$$

- Will show in problem set that as n$\rightarrow\infty$ the Poisson distribution tends toward Gaussian distribution
- Will assume you have used this distribution

Uniform distribution:
parameterizing ignorance?



- Use this if you don't know a value but only the range in which may have fallen with equal probability
  - Binned data
  - A hit on Si strip
  - Some parameter which is bounded i.e. a phase 0-2π
- Self evident that <x>=(a+b)/2
- Standard deviation is (proof on board)

$$\sigma = \frac{b-a}{\sqrt{12}}$$

Part 2

LIVING WITH ERRORS

Uncertainties/errors

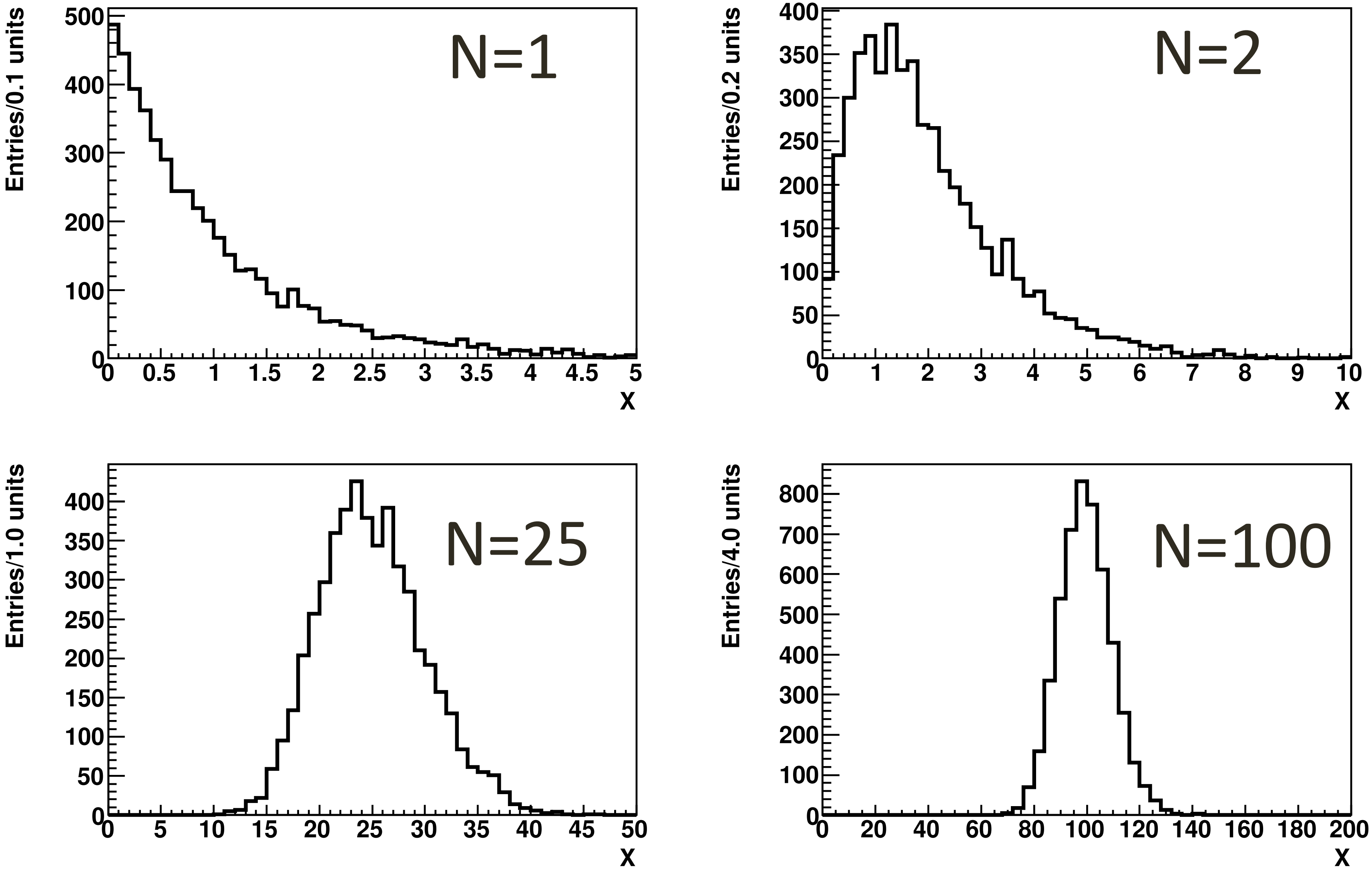
- Are everywhere
- Personally I prefer uncertainty/resolution to error, some of your collaborators maybe be religious about this
  - Error suggest you are doing something **wrong**
  - But by carefully considering your uncertainties you are being **righteous**
- What we need to **know** are the different types, how to evaluate them and to combine them
- Otherwise you result can lead to a lot of **mischief**

Why are uncertainties Gaussian?

- The **Central Limit Theorem** is the answer
  - If you take the sum X of N independent variables x_i , each taken from a distribution of mean μ_i and variance V_i the distribution for X
    1. has an expectation value $\langle X \rangle = \Sigma\, \mu_i$
    2. has a variance $V(X) = \Sigma\, V_i$ and
    3. **becomes a Gaussian as $N \rightarrow \infty$**
- I will prove 1 and 2 on the board now
- But the most startling of these 3 requires a more formal definition which I frankly do not have time to do:
  - Appendix 2 Barlow contains a proof
  - However, I recommend Chapter 30 of the 3rd Edition of Riley, Hobson and Bence 'Mathematical Methods for Physics and Engineering' to understand moments and hence the proof of the CLT

Central limit theorem: example

$$X = \sum_{i=1}^{N} x_i \text{ where } x_i \text{ is drawn from } P(x) = e^{-x}$$



Uncertainty on the mean: do more!

- If we make many independent measurements of the same quantity with a true mean μ
  - $\langle X \rangle = \Sigma\, \mu = N\mu$ (from CLT)
  - Therefore, your estimate of the mean is X/N
  - $V(X) = \Sigma\, \sigma^2 = N\sigma^2$ (from CLT)
  - Therefore,

$$V(\text{mean}) = \frac{1}{N^2} V(X) = \frac{\sigma^2}{N}$$

  - Taking more measurements is good for you
    - But to halve an uncertainty four times more measurements!
    - Only systematic uncertainties can mess this up

Combining measurements

- There are two measurements of the top mass with different resolutions one with (175$\pm$2) GeV/c^2 and the other (176$\pm$1) GeV/c^2
  - How do we combine them?
  - I would need four more of the first measurement to get the same precision as the second
- **Switch it:** second measurement is equivalent to four of the first so should be weighted by a **factor 4**
  - Therefore $$\overline{m_t} = \frac{1}{5}\times 175 + \frac{4}{5}\times 176 = 175.8\ \mathrm{GeV}$$

Proof in Problem set

  - In general $$\overline{x} = \frac{\sum x_i/\sigma_i^2}{\sum 1/\sigma_i^2} \text{ and } V(\overline{x}) = \frac{1}{\sum 1/\sigma_i^2}$$

Caveats

- CLT
  - Works well in the central part of a distribution
    - **Core** – 2 or 3 sigma from the mean
  - But the events outside this – **outliers, tails or wings** – do not tend to Gaussian as fast
  - As N never really tends to ∞ **beware outliers**
- When averaging we assume measurements are uncorrelated
  - Must modify combination to include correlations (more in a moment)
- Also, averaging measurements that are incompatible with one another makes **no sense**
  - Two other top mass measurements (175 ± 2) GeV/c^2 and the other with (186 ± 1) GeV/c^2
  - One (or both) are very likely to be wrong

Error propagation

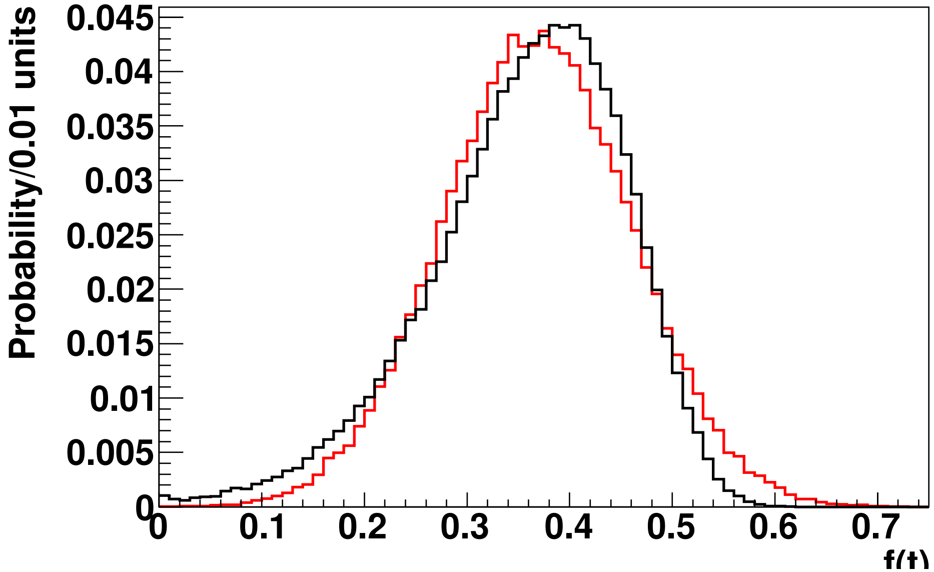
- On board variance of f=ax+b where a and b are exact constants
  - $\sigma_f = |a|\ \sigma_x$
- Now in general

$$f(x) = f(x_0) + (x - x_0)\left.\frac{df}{dx}\right|_{x_0}$$

$$\therefore \sigma_f = \left|\frac{df}{dx}\right|\sigma_x$$

- For 'small errors' – when df/dx approximately constant for a few standard deviations about the point.
- If this is not true you have to use a Monte Carlo or a higher order expansion
  - Example next slide: $f=\exp(-t/\tau)$ with $\tau=(2 \pm 0.5)$ sec and $\tau=(2 \pm 0.1)$ with t = 2 sec

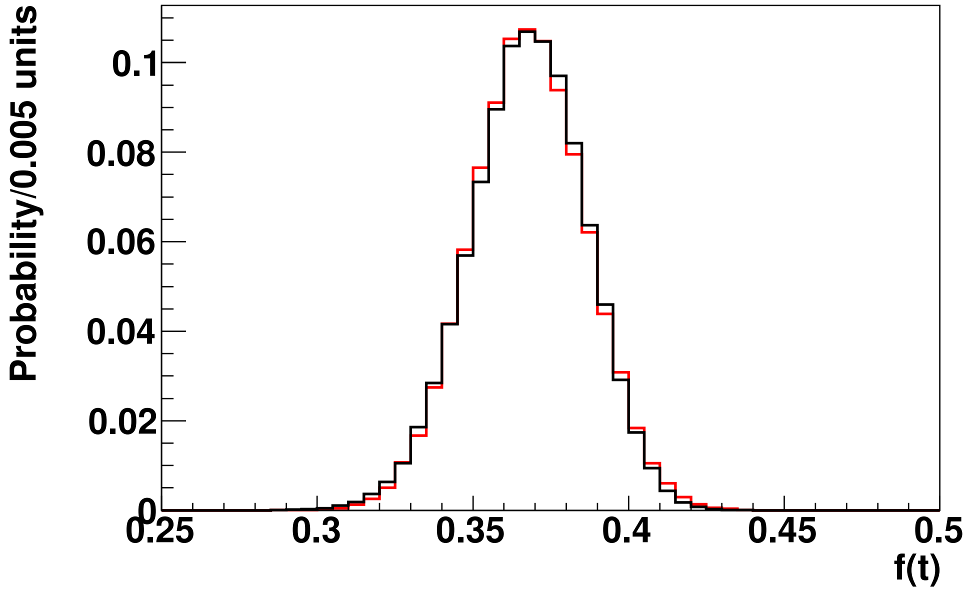
Error propagation: example



$\tau=(2 \pm 0.5)$

Propagation

Simulation



$\tau=(2 \pm 0.1)$

More than one variable and function

- Now we have m different functions f_k of n different variables x_i for which there are means and variances μ_i and σ^2_i respectively
  - Note the functions will be correlated even if the variables are not
- The variance of f is given by $V(f_k) = \langle {f_k}^2 \rangle - \langle f_k \rangle^2$
- Expanding as a Taylor series

$$f_k(x_i) \simeq f_k(\mu_i) + \sum_{i=1}^{n} \frac{\partial f_k}{\partial x_i}(x_i - \mu_i)$$

$$\Rightarrow V(f_k) = \sum_{i=1}^{n} \left(\frac{\partial f_k}{\partial x_i} \right)^2 V(x_i) + \sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} \left(\frac{\partial f_k}{\partial x_i} \right) \left(\frac{\partial f_k}{\partial x_j} \right) \mathrm{cov}(x_i, x_j)$$

$$\text{and } \mathrm{cov}(f_k, f_l) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{\partial f_k}{\partial x_i} \right) \left(\frac{\partial f_l}{\partial x_j} \right) \mathrm{cov}(x_i, x_j)$$

All you need to know about error propagation

- The new covariance can be neatly defined in terms of a matrix multiplication defining m × n matrix **G** as

$$G_{ki} = \frac{\partial f_k}{\partial x_i}$$

- Then $$\mathbf{V}_f = \mathbf{G}\,\mathbf{V}_x\,\mathbf{G}^T$$

- Where the covariance matrices are n × n for the variables and m × m for the functions.
- Example: f(x,y,z) with x, y, z uncorrelated on board

Systematic uncertainties

Those of you doing analysis will one day have to produce a table like this

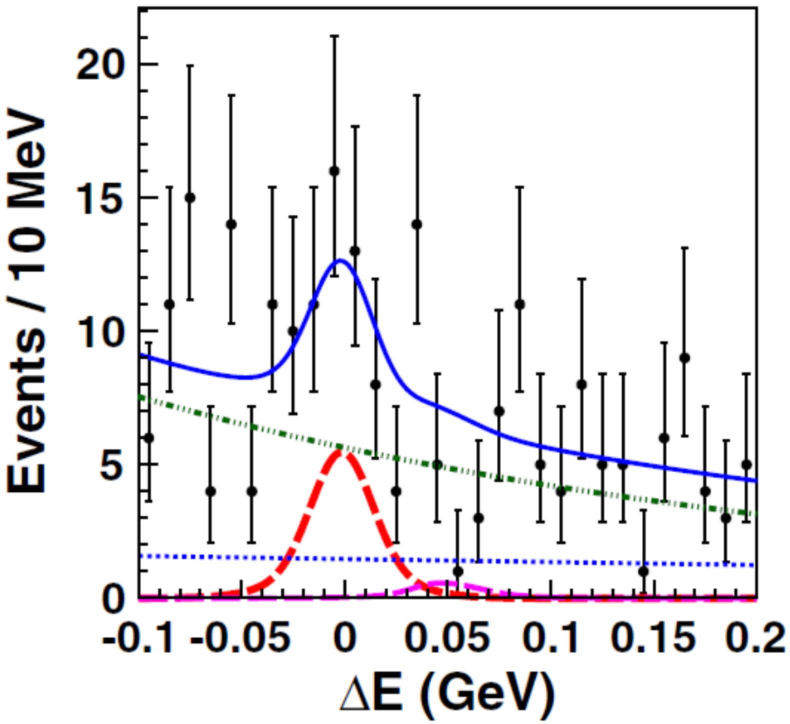
TABLE II. Summary of the systematic uncertainties for R_{Dh} and A_{Dh}. Negligible contributions are denoted by "$\cdots$."

Source	R_{DK} (%)	$R_{D\pi}$ (%)	A_{DK}	$A_{D\pi}$
ΔE and C'_{NB} PDFs	$^{+6.5}_{-7.1}$	$^{+8.3}_{-10.3}$	$^{+0.03}_{-0.02}$	$^{+0.02}_{-0.03}$
Fit bias	$+0.1$	$+0.4$	$\cdots$	$\cdots$
Due to $B\bar{B}$ and $q\bar{q}$ bias	± 3.0	$\cdots$	$\cdots$	$\cdots$
Peaking background	± 9.5	± 8.2	± 0.04	± 0.01
Efficiency	± 0.1	± 0.1	$\cdots$	$\cdots$
Detector asymmetry	$\cdots$	$\cdots$	± 0.02	± 0.02
Total	$^{+11.9}_{-12.2}$	$^{+11.7}_{-13.2}$	± 0.05	$^{+0.03}_{-0.04}$

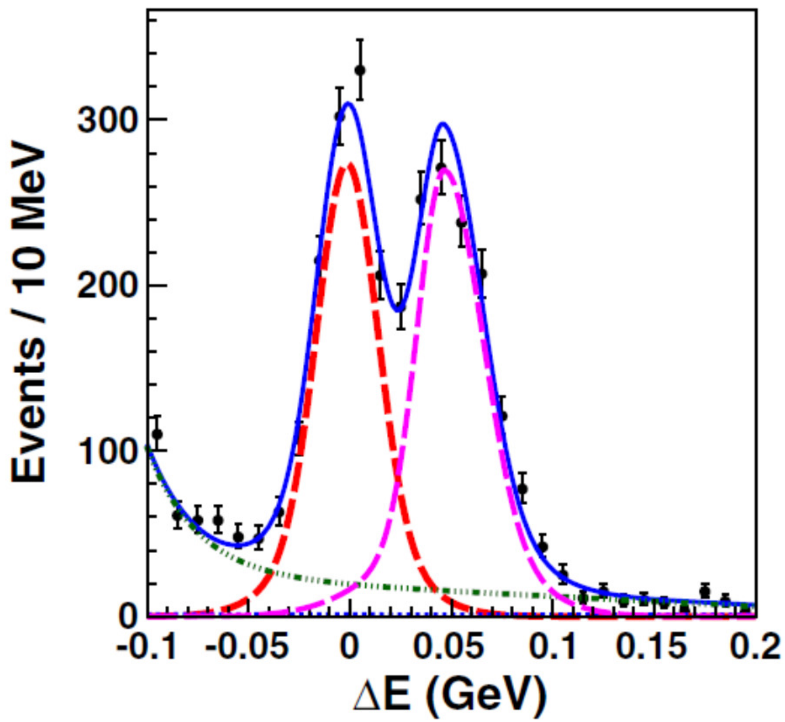
[M. Nayak et al., (Belle Collaboration), PRD **88**, 091104]

This analysis is essentially

Area under red curve in this plot…



…divided by the area under the red curve in this plot



Ratios are often cool for reducing systematics (why?)

Systematic uncertainties: definition

- Take a calorimeter with an energy resolution of 5%
  - measurements are sometimes too high sometimes too low
  - however repeated measurements of the same thing (i.e. mass of π^0) still leads to a reduced uncertainty
- If it always gives 5% too high it doesn't matter how often you repeat the measurement it will always be off by 5%
  - In reality you have to calibrate this away
- These are systematic uncertainties which are essentially **ones that do not scale with 1/sqrt(N)**
- Another problem is non-independence at different points
  - For example measuring a differential cross section the luminosity uncertainty moves all points up or down by the same relative amount
- If you ignore these everything looks consistent **but** your answer is wrong

Things are not so bad

- **Many systematics are easy to deal with:**
  - Calibrations, efficiency, luminosities etc which will come with some associated uncertainty which you can propagate
- **More problematic are unknowns where some intelligent guesswork is required:**
  - Often saved by Gaussian quadrature sum of uncertainties
    - If error is poorly known but small, larger better controlled systematics or your statistical uncertainty will dominate
    - **so don't sweat the small stuff just be conservative**
- Example of a systematically limited measurement on board
- **Barlow's advice is be paranoid about everything and perform checks**
  - Fitters on MC samples
  - Divide the data into different periods of data taking
  - Vary analysis procedure
  - Note: these checks do not necessarily lead to systematic uncertainties – only if they throw up a discrepancy

Systematic uncertainties: look again at our local example

TABLE II. Summary of the systematic uncertainties for R_{Dh} and A_{Dh}. Negligible contributions are denoted by "$\cdots$."

Source	R_{DK} (%)	$R_{D\pi}$ (%)	A_{DK}	$A_{D\pi}$
ΔE and $C'_{\rm NB}$ PDFs	$^{+6.5}_{-7.1}$	$^{+8.3}_{-10.3}$	$^{+0.03}_{-0.02}$	$^{+0.02}_{-0.03}$
Fit bias	$+0.1$	$+0.4$	$\cdots$	$\cdots$
Due to $B\bar{B}$ and $q\bar{q}$ bias	±3.0	$\cdots$	$\cdots$	$\cdots$
Peaking background	±9.5	±8.2	±0.04	±0.01
Efficiency	±0.1	±0.1	$\cdots$	$\cdots$
Detector asymmetry	$\cdots$	$\cdots$	±0.02	±0.02
Total	$^{+11.9}_{-12.2}$	$^{+11.7}_{-13.2}$	±0.05	$^{+0.03}_{-0.04}$

[M. Nayak et al., (Belle Collaboration), PRD **88**, 091104]

Dealing with systematics

- The hard part is the estimation of the systematic uncertainties
- Once you have the errors you can use the covariance matrix in the usual way
- For two measurements $x_{1,2}$ with common correlated systematic uncertainties S and independent statistical uncertainties $\sigma_{1,2}$ – proof on board

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 + S^2 & S^2 \\ S^2 & \sigma_2^2 + S^2 \end{bmatrix}$$

- For a common fractional error f it is

$$\mathbf{V} = \begin{bmatrix} \sigma_1^2 + f^2 x_1^2 & f^2 x_1 x_2 \\ f^2 x_1 x_2 & \sigma_2^2 + f^2 x_2^2 \end{bmatrix}$$

Part 3

ESTIMATION

Properties of estimators

- An estimator is:
  - a procedure applied to the data sample which gives a numerical value for a property of a parent population or, as appropriate a parameter of the parent distribution function
    - Yields, masses, lifetimes, mixing angles or whatever
- Three things we want from an estimator
  - **Consistency:** the difference between the estimator and the true values vanishes for large samples
  - **It is unbiased:** the expectation value of the estimator gives the true value
    - Example of mean to show the difference between these statements on the board
  - **Efficiency**: gives a small variance

The likelihood function

- Consider some PDF P(x,a) which depends on parameter a which you wish to estimate
- The probability you get a particular set of data x_i drawn from P(x,a) is

$$L(x_1,\ldots,x_N,a)=\prod_{i=1}^{N}P(x_i,a)$$

- This is the **likelihood function**
- If we have an estimator of $\hat{a}(x_i)$ of our parameter *a* its expectation value is

$$\langle\hat{a}\rangle=\int\hat{a}L\,d\mathbf{x}\text{ where }d\mathbf{x}\equiv dx_1dx_2\ldots dx_N$$

- We will now use this to prove that there is a limit to the efficiency of an estimator
  - **The Minimum Variance Bound**

Estimating the variance: Bessel's correction

- Some of you may have seen differing versions of the definition of the standard deviation

$$s = \sqrt{\frac{1}{N-1}\sum_i \left(x_i - \bar{x}\right)^2}$$

- Rather than this

$$\sigma = \sqrt{\frac{1}{N}\sum_i \left(x_i - \bar{x}\right)^2}$$

- I will now explain the difference – on board

Maximum log likelihood

- This is one of the two commonest methods of estimation
  - Simply put you vary the parameter(s) in the likelihood until you find the global maximum
  - In practice one normally solves

$$\left.\frac{d\ln L}{da}\right|_{a=\hat{a}} = 0$$

  - Really you use MINUIT
    - http://seal.web.cern.ch/seal/snapshot/work-packages/mathlibs/minuit/

  and minimize $-2 \ln L$
    - I will explain the factor 2 in a bit
- But sometimes it can be solved analytically – example of the Gaussian weighted mean.

Max. likelihood is biased…

- On board will show that ML estimate of V is

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i \left(x_i - \overline{x}\right)^2$$

  which we know is **biased**

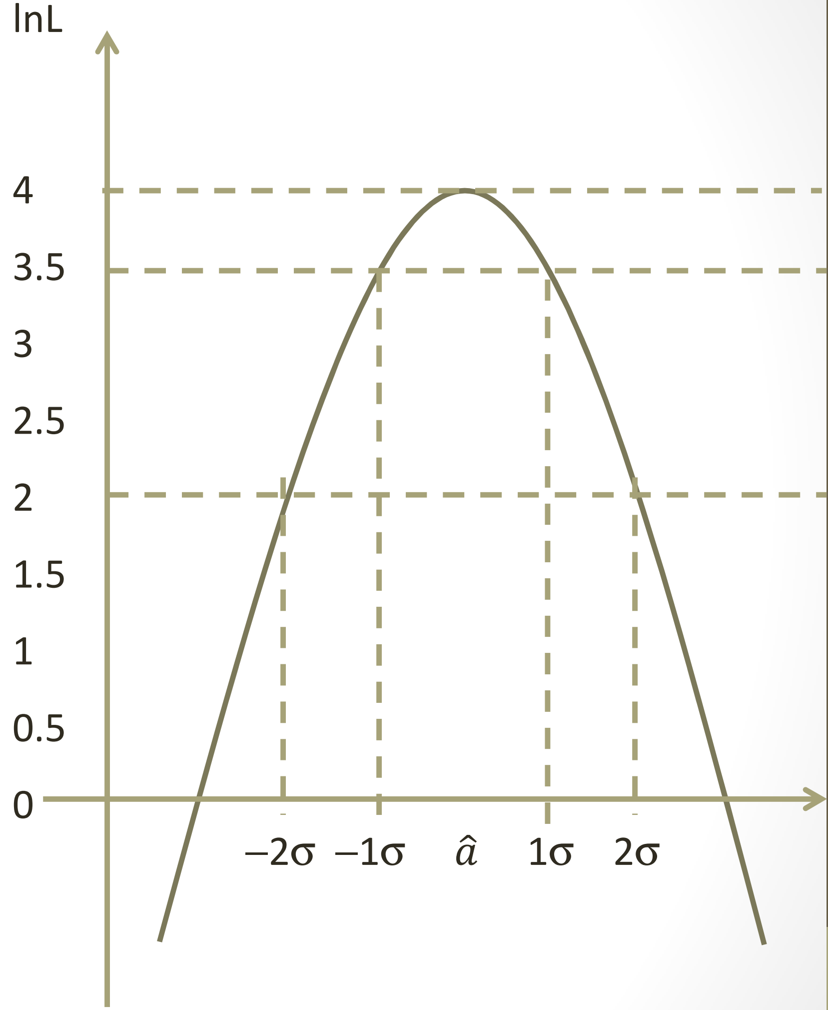
- The reason we use the likelihood is because it is **invariant** under transformations of a – i.e. if we differentiate the Gaussian likelihood we just used with respect to σ^2 we get $\widehat{\sigma^2} = \hat{\sigma}^2$ - try it yourselves
- In general $\hat{f}(a) = f(\hat{a})$

…but efficient

- We don't worry too much because in the limit of large N a **consistent** estimator becomes **unbiased**
- Further the variance of a maximum likelihood estimate lies on the minimum variance bound as N becomes large –
  - proof in Barlow 5.3.3
- Hence, the dominance of L
  - **it squeezes the maximum information out of your data**
- The **pull plot** is a useful tool if you are worried about bias in a likelihood fit
  - **Recipe:** make many simulation experiments (**toys**)
    1. varying sample size generated with some value of your parameter of interest
       - Sample size of toy should be taken from a Poisson distribution with a mean equal to your observed sample size
    2. Run your maximum likelihood fit on each of these samples
    3. Plot difference between generated and fitted value of parameter divided by the uncertainty on the parameter
    4. If distribution is normal (μ=0,σ=1) even your grumpiest collaborators will be convinced that the fit gives an unbiased estimate with reliable uncertainties

Uncertainties from the ML

- For large samples one can show (See Barlow Sec. 5.3.3) that L is Gaussian with $\sigma = \sqrt{V(\hat{a})}$
- Therefore, ln L is a parabola which has fallen
  - **0.5 at $\pm$1 sigma from $\hat{a}$**
  - **2 at $\pm$2 sigma from $\hat{a}$**
  - **4.5 at $\pm$3 sigma from $\hat{a}$**
- For small N ln L not parabolic but invariance of L means
  - For some alternate parameter a′(a) it is parabolic so still use ln L(max) – 0.5 to get $\pm$ 1σ
  - Uncertainties in a are **asymmetric**



Least squares

- Suppose you have a set of points $\{(x_i,y_i)\}$
- x_i are exact, but y_i have a resolution σ_i
- Suppose there is a hypothesis y=f(x;a) and we want to estimate parameter a
- CLT tells us measured y are Gaussian distributed about their true values so

$$P(y_i;a) = \frac{1}{\sigma_i\sqrt{2\pi}}\exp\left[-\left(y_i - f(x_i;a)\right)^2 / 2\sigma_i^2\right]$$

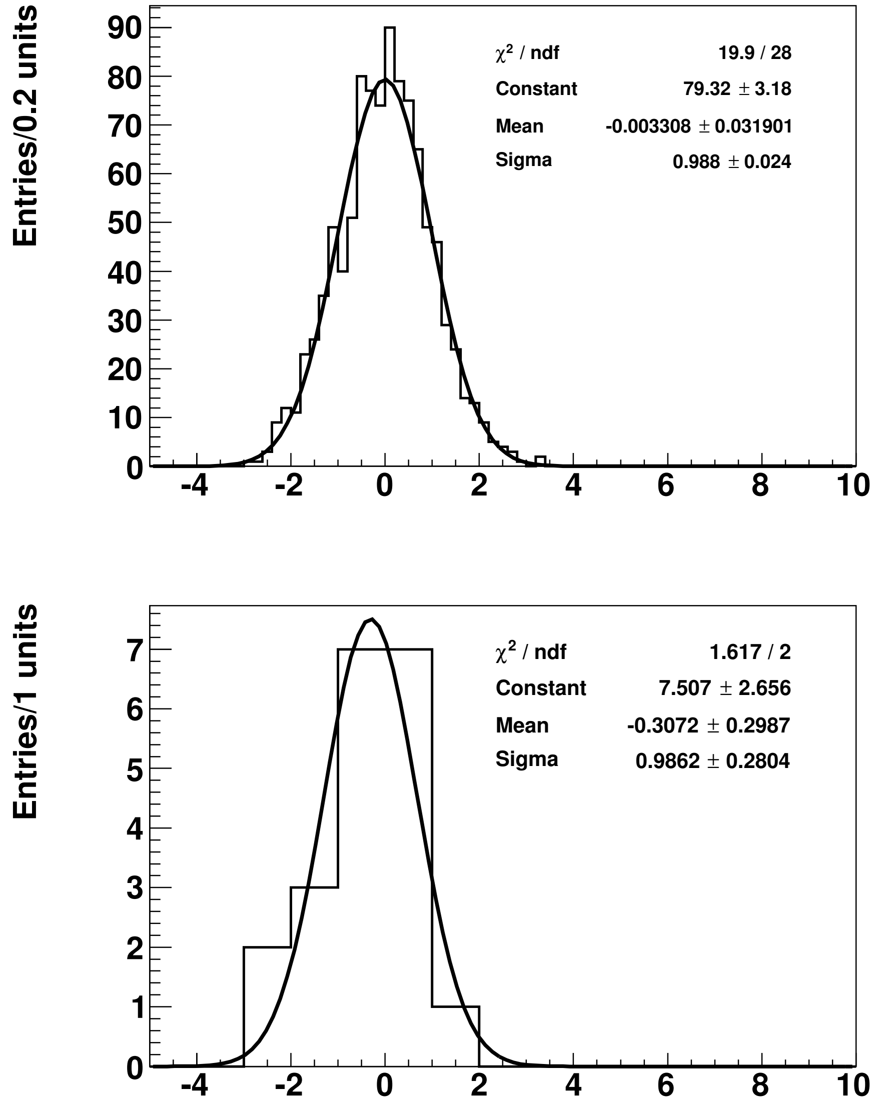
$$\Rightarrow \ln L = -\frac{1}{2}\sum\left[\frac{y_i - f(x_i;a)}{\sigma_i}\right]^2 + \text{constant}$$

Minimise this sum and you maximise ln L
method of least squares or χ^2

Can you explain your factor 2 in Minuit now?

χ^2 fitting binned data

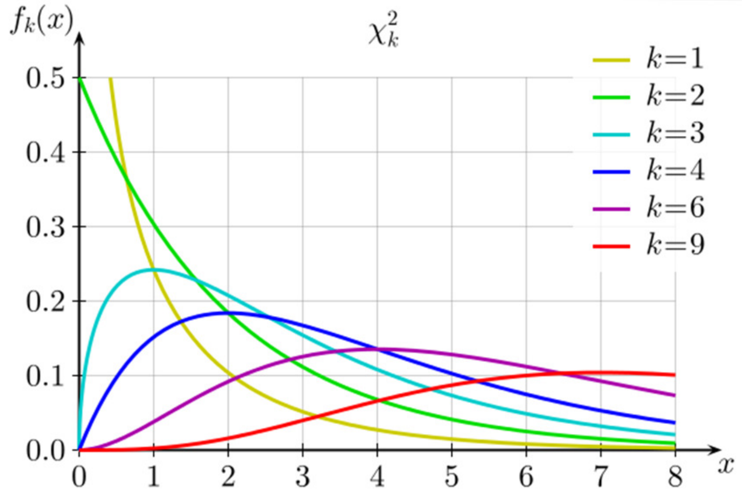
- When you do this
  - myHisto->Fit("gaus");
- Out pops your estimate of the width and mean with an uncertainty
- However, there is an important subtlety here
  - $\sigma = \sqrt{N_i}$
  - Should be the square root of the integral of the PDF over the bin $\times$ total number of events
    - This is the mean of the Poisson distribution from which you assume your event sample is drawn
  - Large N no difference but if statistics are small care needs to be taken



IXth SERC School on EHEP

χ^2 per degree of freedom

Fig. from Wikipedia



- The χ^2 follows a distribution:
  - proof on board if time allows

$$P(\chi^2, k) = \frac{2^{-k/2}}{\Gamma(k/2)} \chi^{k-2} e^{-\chi^2/2}$$

- Where n = number of degrees

  = number of data points – number of parameters
- Mean is k and variance 2k
- Therefore, $\chi^2/k \sim 1$ indicates a good fit
  - $\chi^2/k \ll 1$ overestimated the uncertainties
  - $\chi^2/k \gg 1$ wrong function or some outlier

Multi-dimensional Gaussian

- Let us consider a set of of variables **x**=$\{x_1,x_2,\ldots\ldots,x_n\}$ with means $\mu=\{\mu_1,\mu_2,\ldots\ldots,\mu_n\}$ which follow normal distributions with widths $\sigma=\{\sigma_1,\sigma_2,\ldots\ldots,\sigma_n\}$

$$P(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\sigma}) \propto \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{A}(\mathbf{x}-\boldsymbol{\mu})\right)$$

- Where **A** is a n $\times$ n matrix which depends upon $\boldsymbol{\sigma}$
- If the variables are **uncorrelated** then **A** is diagonal $1/\sigma_i^2$
- If the variables are correlated then matrix is symmetric

  with

$$\mathbf{A} = \mathbf{V}^{-1}$$

  - Section 3.4.6 of Barlow for proof

χ^2 – generalised

- The multidimensional Gaussian motivates the generalised χ^2 function

$$\chi^2 = (\mathbf{x} - \mathbf{f})^T \mathbf{V}(\mathbf{x})^{-1} (\mathbf{x} - \mathbf{f})$$

where $\mathbf{x}=(x_1,\ldots..,x_N)$ and $\mathbf{f}=(f(x_1,\mathbf{a}),\ldots\ldots, f(x_1,\mathbf{a})\}$ and the covariance matrix is among the elements of **x**

- If **f** is linear in **a** then **f=Ca** then you can show that (Barlow 6.6)

$$\mathbf{V}(\hat{\mathbf{a}}) = \left[\mathbf{C}^T \mathbf{V}(\mathbf{x})^{-1} \mathbf{C}\right]^{-1}$$

- If **f** non-linear in **a** then iterate linearly using the Taylor expansion

Part 4

PROBABILITY AND CONFIDENCE

Definition of probability

- **Mathematical (Kolmogorov)**
  1. P>0
  2. P(1 or 2) = P(1) + P(2) if 1 and 2 mutually exclusive
  3. $\Sigma P = 1$

  Uncontroversial but what is P
- **Empirical/Classical (Von Mises) – limit of frequency as number of trials/experiments tends to infinity**
  - But depends on the ensemble from which all events are chosen
  - Probability of a D meson being produced at sqrt(s)=M(Υ(4S)) is different to the probability it is produced in Υ(4S) decays
  - What about a single experiment – you cannot say anything
- **Objective probability (Popper) – it is a property of an object**
  - Does not depend upon ensemble:
    - quantum mechanical probability directly from wavefunction
  - But what about a continuous distribution: $P(\Delta\theta)$ vs $P(\Delta\cos\theta)$?

Bayesian statistics

- First conditional probability:
  - p(a|b) = probability of a given b
- Bayes theorem
  - p(a|b)p(b) = p(a and b)=p(b|a)p(a)

$$p(a\,|\,b) = \frac{p(b\,|\,a)\,p(a)}{p(b)} = \frac{p(b\,|\,a)\,p(a)}{p(b\,|\,a)\,p(a) + p(b\,|\,\overline{a})[1 - p(a)]}$$

- Example of threshold Cerenkov detector on board
- Now for **subjective probability (the controversial piece)**

$$p(\text{theory}\,|\,\text{result}) = \frac{p(\text{result}\,|\,\text{theory})\,p(\text{theory})}{p(\text{result})}$$

Example: spot fixing

What are the chances Sachin or Sreesanth bowl three no-balls in an over?

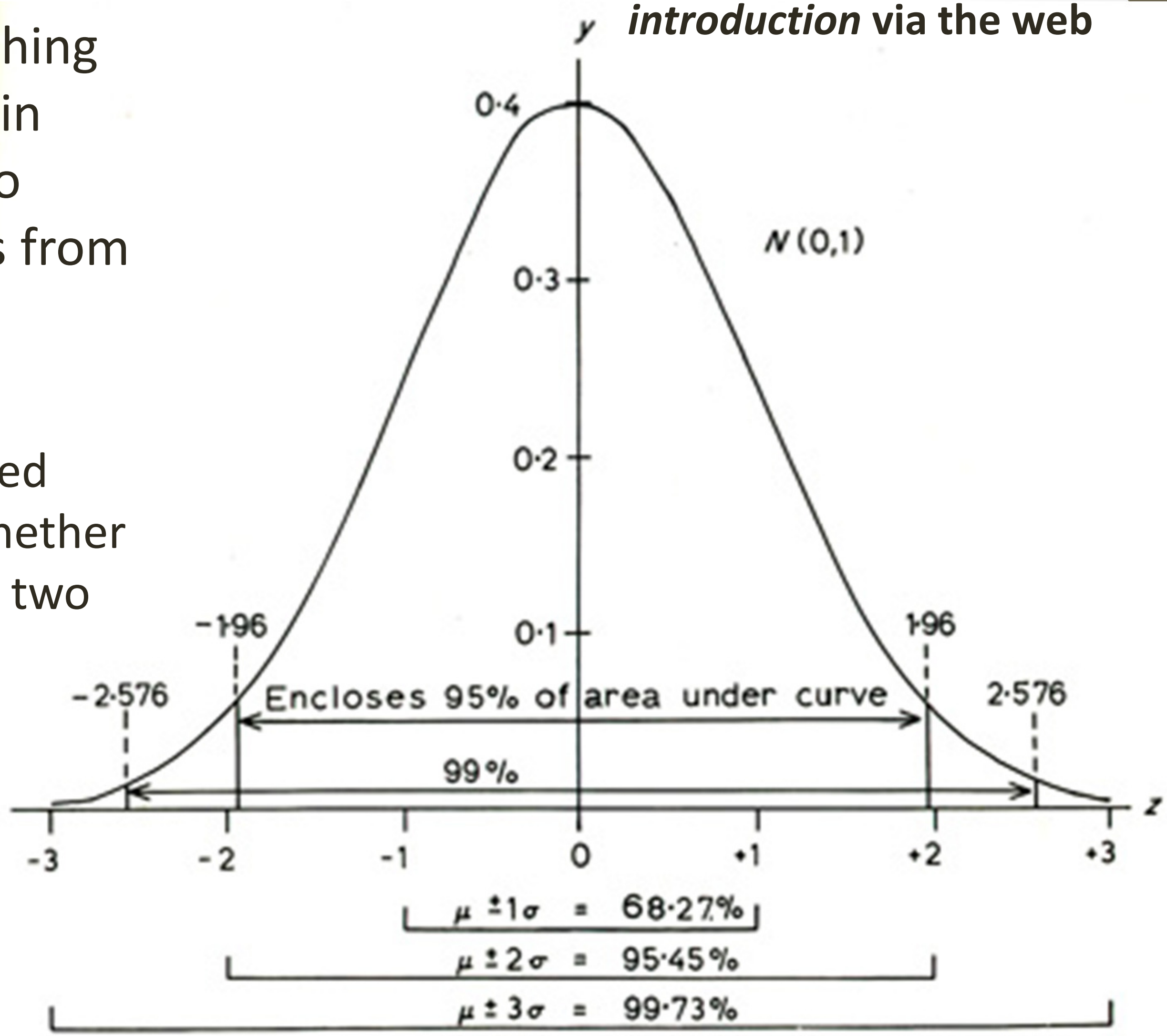
Probability – the bottom line

- Subjective probability looks **unscientific**
- Therefore, we are likely to identify ourselves as frequentists (classical probability) given the problems of objective probability
- But we should not move so fast
  - Certainly in QM most of us think of probabilities as **intrinsic objective numbers**
  - **Interpreting results** always leads us into a Bayesian approach: mass of the electron on board
- Philosophical wars rage on the frequentist vs Bayesian approach:
  - My opinion: don't worry about it just always explain clearly what you do when interpreting data and make sure it is consistent
  - If someone wants to interpret your data in a different way it is up to them to explain clearly what they are doing
  - If you do this you are just **different not wrong**

Confidence levels

Fig from Till, 1974, *Statistical Methods for the Earth Scientist: an introduction* via the web

- How often a something would be in a certain range if you were to sample many times from a given PDF
- Not controversial
  - Choose your desired probability and whether you want a one or two sided
  - Then integrate



Confidence levels - asymmetric

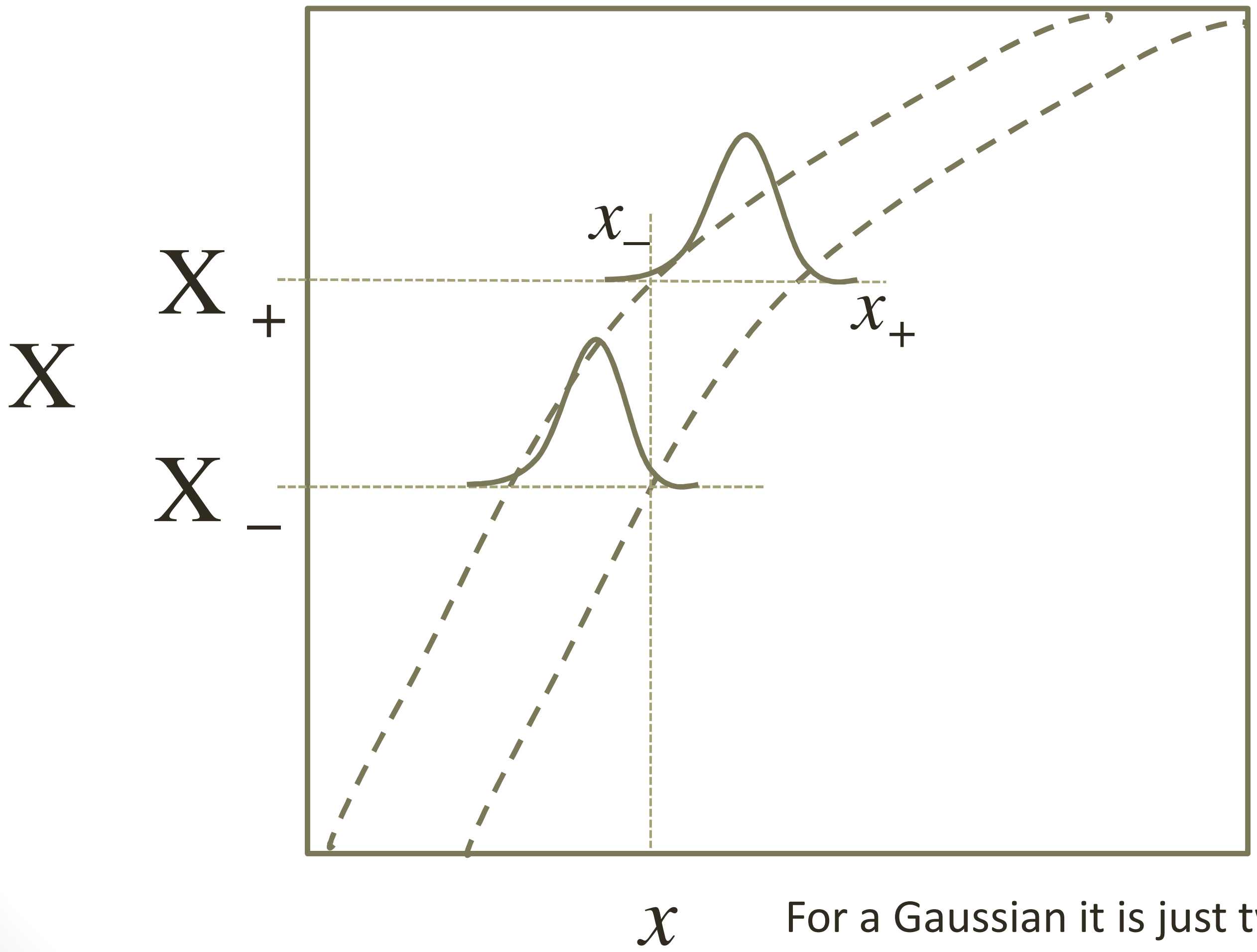
- If the probability distribution is not symmetric there are three possible ways to make your interval
  1. Demand a symmetric interval about the mean
  2. Make it as small as possible
  3. Central interval: have half the remaining probability in each tail
     - Illustrated on board
- Barlow, I and others prefer the latter despite the asymmetry about the mean
  - However, 1 and 2 are not wrong just make sure you have explained what you have done

Confidence in estimation

- Now we want to say something about an unknown parameter X given our measurement - estimation
- Naively for a measurement of Gaussian errors you measure x with a know σ so you say $x-2\sigma < X < x+2\sigma$ at 95% CL
  - As you will see this is often fine but
  - If you measure some branching fraction to be $(1\pm1)\%$ it means from the above approach 16% BF<0!
- Need to use a different approach and that is to build up a **confidence belt**

Confidence belt

For example for 90% central interval $0.05 = \int_{-\infty}^{x_-} p(X, x)dx = \int_{x_+}^{\infty} p(X, x)dx$



For a Gaussian it is just two straight lines and leads to the naïve result

Constrained quantities I

- Now why have I bothered with this
  - What about the case of measuring a negative branching fraction
  - True value 0.1% but you measure (– 0.80±0.41)% means the **physical range** (0,0.02%) has 95.4% probability if you treat this in a classical way - nonsense
  - Bayesian approach
    - P(X|x)=P(x|X)P(X)/P(x)
    - Assumed ignorance of X and the fact that P(x) disappears in the normalization (it is just a scalar number)
    - P(X|x) = P(x|X) which justifies what we did earlier with the mass of the electron
    - Also, frequentist and Gaussian limits are the same

Constrained quantities II

- What about the case of measuring a negative branching fraction
- Bayesian approach
  - P(X|x)=P(x|X)P(X)/P(x)
  - Now for the physical limit P(X|x)=P(x|X) θ(X) /P(x) where θ(X) is a step function which for a Gaussian distribution function gives

$$p(X\mid x)=\frac{\exp[-(x-X)^2/2\sigma^2]}{\int_0^\infty \exp[-(x-X')^2/2\sigma^2]dX'}\qquad (x>0)$$

  - Now for my example: BF<0.35% with 90% CL
- But I would of got a different answer if I worked with sqrt(BF)
  - **– so beware**

Summary

- We have reviewed
  1. Probability distributions
  2. Error analysis: including systematics
  3. Estimation
  4. Setting of confidence intervals
- These lectures contain things which are pretty much the minimum set you should know to be able to critique your own or anyone else's analysis
- **One must go on to study things further that are relevant to your particular analysis**
- Apologies to those interested in multivariate analysis:
  - I suggest the talks and tutorials from
  - http://tmva.sourceforge.net/