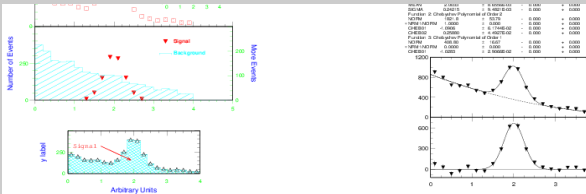


Statistical Methods in Particle Physics

Heidelberg+LHCb Workshop
Neckarzimmern

Ian C. Brock
22nd February 2012

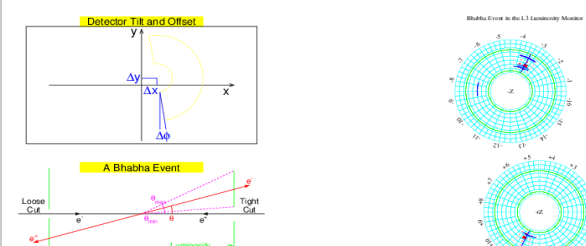
Part 2



Mn_Fit

Fitting and Plotting Package Using MINUIT

Version 5_15
21st February 2012
BONN-MS-99-02
Ian C. Brock
Physikalisches Institut der Universität Bonn
Nußallee 12, D-53115 Bonn
E-Mail: Ian.Brock@cern.ch or brock@physik.uni-bonn.de
Homepage: http://pi.physik.uni-bonn.de/~brock/mn_fit.html



Detector Tilt and Offset

A Bhabha Event

Loose Cut

Tight Cut

Luminosity Monitor

Overview

- ♦ Part 1:
 - Tools and literature
 - Measurements and presentation of data
 - Distributions
 - Central limit theorem
 - Error propagation
- ♦ Part 2:
 - Systematic errors
 - Estimation
 - Likelihood
 - Maximum likelihood examples
 - Least squares
 - Straight line fit
 - Bayesian statistics
 - Confidence levels
 - (Hypothesis testing)

Systematic Errors

- ▶ How should we handle systematic errors both within a single measurement and when we combine measurements?
- ▶ Error usually due to a flaw or inaccuracy in measuring apparatus, e.g. a voltage offset
- ▶ With systematic errors repeating measurements does not help, i.e. measurements are not independent of each other and CLT does not apply when combining them
- ▶ More data can help you to determine them better
- ▶ Systematic error is not a systematic mistake – if we know a measurement should be corrected by 1.05 ± 0.03
 - Not making the correction is a systematic mistake
 - Should correct data by 1.05 and take uncertainty (0.03) as systematic error
 - Often correction made and then for systematic error uncorrected data used – this is probably an overestimate

Avoiding or Minimising Systematic Errors

- ▶ Order of measurements
 - Do not measure current at voltages of 0,1,2,...,10 V
 - Much better 7,3,4,6,9,1,... V, i.e. in a random order to minimise effects of apparatus drift – does not get rid of effects, but randomises them
 - Worry about hysteresis
 - For absolute value measure sometimes coming from above and sometimes from below
 - For slope always come from same side
- ▶ Cross-checks
 - Redundant triggers
 - Different ways of determining energy scale – test beam, particle masses, p_T balance etc.

Sanity (Consistency) Checks

- ▶ Decide beforehand if a procedure is a consistency check or evaluation of an uncertainty
- ▶ Make as many sanity checks as you can
 - Do not blindly rely on Monte Carlo – check it with data
 - Use a sample of events where no effect is expected
 - e.g. wrong charge combinations for a mass peak
 - Measure CP asymmetry for sample of events known to have no asymmetry
- ▶ Vary your fitting procedure and look for changes in result
 - Is the variation consistent with a statistical fluctuation?
 - Yes: consistency check
 - No: source of systematic uncertainty
 - What is consistent?

$$\sigma_{A-B}^2 = \sigma_A^2 - \sigma_B^2$$

If better technique, B, saturates the MVB
– see later

Sanity Checks

- ▶ What to do if a sanity check fails?
 - Check your analysis
 - Check it again
 - Explain exactly what you do to someone else and then realise what your mistake is (**amazing how often this works!**)
 - **Incorporate as a systematic error as last resort**
- ▶ What if sanity (consistency) check OK?
 - **Do not incorporate it into your systematic error**

Evaluating Systematic Errors

- ▶ Consider possible sources at planning stage of experiment
 - Make intelligent guesses on their size;
 - Consider how to calibrate your devices and check the calibrations
 - Do not forget environmental factors (temperature, pressure, humidity, ...)
 - **Make sure you record them!**
 - Check for effects by repeating under same nominal conditions at random times, then e.g. plot vs. temp. and evaluate correlation coefficient

Evaluating Systematic Errors

- ♦ Theory errors are tough to evaluate
- ♦ You have 2 possible models – what is the systematic error?
 - Take average and spread as error – rarely done
 - Use one model and deviation of 2nd model as systematic error
 - Quote as one-sided or two-sided error?
 - **Once again probably an overestimate**

Evaluating Systematic Errors

- ▶ What about tolerances?
 - Your technician constructs chamber with a tolerance of 0.2 mm
 - What do you give for error on a length of 10 mm? ± 0.2 mm?
 - NO! Prob. distribution is flat, so use error of $0.4/\sqrt{12} = 0.12$ mm
- ▶ Suppose you have 2 (theory/MC) models A,B
 - If they are extreme scenarios (truth always between A and B), take variance to be that of a uniform distribution with width A-B

$$\sigma = \frac{|A - B|}{\sqrt{12}}$$

- If they are typical scenarios (JETSET vs. HERWIG) error given by:

$$\sigma = \frac{|A - B|}{2} \cdot \sqrt{\textcircled{a}} = \frac{|A - B|}{\sqrt{2}}$$

Factor $\sqrt{[N/(N-1)]}$ to get unbiased estimate of σ_{parent}

Cut variation

- Most used (and abused?) way of assigning systematic uncertainty due to incomplete detector simulation etc

p	105	0.835	125.8
$p + \Delta p$	110	0.865	127.2
$p - \Delta p$	100	0.803	124.5

$$\sigma_{\text{sys}} = (127.2 - 124.5) / 2 = 1.4$$

$$x = 125.8 \pm 1.4$$

- Vary your cuts by a bit (**how much?**)
 - Measure new data yield
 - Measure new MC efficiency
 - Take difference in corrected yield as systematic uncertainty
- Common recommendation is to vary cut by resolution of variable

- Is the variation consistent with a statistical fluctuation?

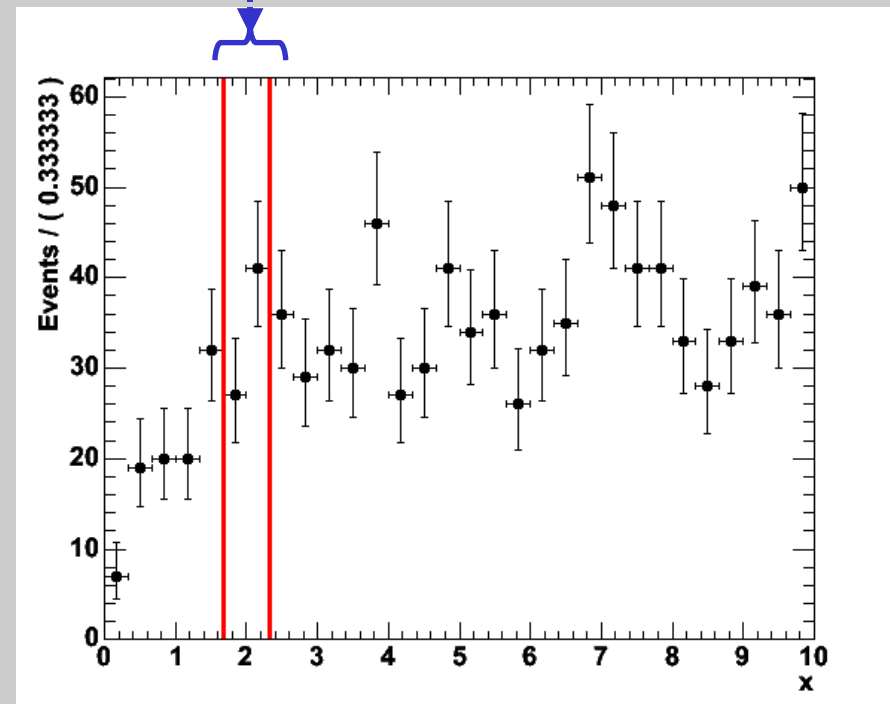
$$\sigma_{(p+\Delta p)-p}^2 = \frac{110}{0.865^2} - \frac{105}{0.835^2}$$

$$\sigma_{(p+\Delta p)-p} = 1.9$$

- I would say yes, and it should not be counted as a systematic effect**

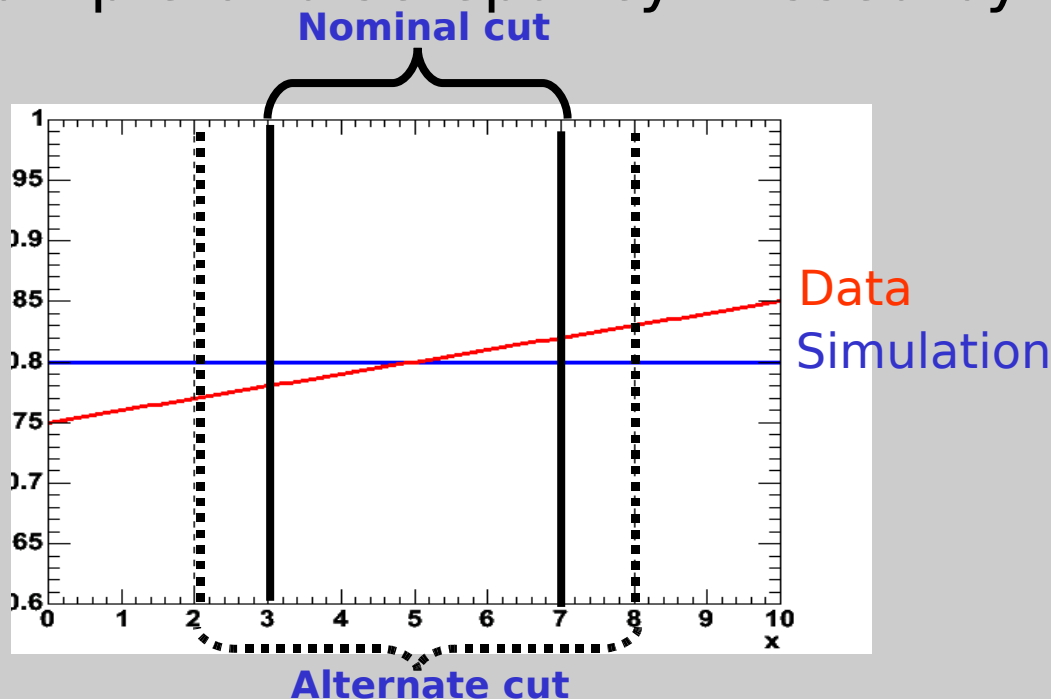
Low statistics

- Warning I: Cut variation does not give an precise measure of the systematic uncertainty due data/MC disagreement!
- Your **systematic error** is **dominated** by a potentially **large statistical error** from the **small** number of events in data between your two cut alternatives
 - This holds independent of your MC statistics
- You could see a large statistical fluctuation → **error overestimated**
- You could see no change due to statistical fluctuation → **error underestimated**



Different dependence

- Warning II: Cut variation doesn't catch all types of data/MC discrepancies that may affect your analysis
 - Error may be fundamentally underestimated
 - Example of discrepancy missed by cut variation:



Data and Simulation give same efficiency for nominal and alternate cut, so

zero systematic is evaluated (in limit $N \rightarrow \infty$)

even though data and MC are clearly different

Cut variation is a good sanity check, but not necessarily a good estimator for systematic uncertainty

Combining Systematic Errors

- ▶ Evaluated and estimated the size of errors, now combine them
- ▶ As the errors are independent of each other combine them in quadrature
- ▶ In many case 1 or 2 dominate – try to reduce them
- ▶ Don't forget that most of the values you give are not much better than educated guesses!
- ▶ As statistical and systematic errors are again independent of each other can use CLT and combine in quadrature; better to quote them separately:
$$A = -10.2 \pm 0.4 (\text{stat.}) \pm 0.3 (\text{sys.})$$
- ▶ See at a glance their relative size

Combining Systematic Errors

Common (but not statistically correct) is to estimate systematic errors *conservatively*, i.e. probably overestimated. Probably OK if you have forgotten a source, but who says that the size of the source you have forgotten is compensated by your *conservative* estimate?

Propagating Systematic Errors

- ▶ Use techniques learned above
- ▶ Consider 2 measurements x_1, x_2 common systematic error, S , and random errors σ_1, σ_2
- ▶ Set up covariance matrix?
 - Treat x_1 as if is made of 2 parts: x_1^R with error σ_1 and x_1^S with common error S ; same for x_2
 - x_1^R, x_2^R uncorrelated
 - x_1^S, x_2^S 100% correlated
 - x_1^R, x_2^S are independent

$$\begin{aligned}V(x_1) &= \langle x_1^2 \rangle - \langle x_1 \rangle^2 \\ &= \langle (x_1^R + x_2^S)^2 \rangle - \langle (x_1^R + x_2^S) \rangle^2 \\ &= \sigma_1^2 + S^2\end{aligned}$$

Propagating Systematic Errors

- Have variance, what about covariance?

$$\begin{aligned} \text{cov}(x_1, x_2) &= \langle x_1 x_2 \rangle - \langle x_1 \rangle \langle x_2 \rangle \\ &= \langle (x_1^R + x_1^S)(x_2^R + x_2^S) \rangle - \langle (x_1^R + x_1^S) \rangle \langle (x_2^R + x_2^S) \rangle \\ &= \langle x_1^S x_2^S \rangle - \langle x_1^S \rangle \langle x_2^S \rangle = S^2 \end{aligned}$$

- 3 of 4 cross products involve x^R
- x_1^S and x_2^S are 100% correlated

- Error matrix is:

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

- We are now set – can propagate errors, calculate correlation coefficients etc.

Propagating Systematic Errors

- ◆ If systematic error is fraction of the value we have

$$S_1 = \epsilon x_1, S_2 = \epsilon x_2$$

- Follow same procedure to calculate covariance between x_1 and x_2 – still completely correlated:

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

- ◆ More generally, if we have several measurements and some are correlated between all, while others are correlated between a subset, follow same procedure and split measurement into more parts.
- This is the standard procedure. You always try to split the measurements into parts that are either independent or completely correlated – much easier than trying to determine correlation coefficients.

Estimation

- ▶ Have a set of measurements
- ▶ Know how to combine them
- ▶ What is the “best” way to combine them
- ▶ Process is called estimation in statistics

An estimator is a procedure applied to the data sample which gives a numerical value for a property of the parent population or, as appropriate, a property or parameter of the parent distribution

Properties of Estimators

- ◆ Make a measurement or series of measurements; results are samples drawn from a *parent population* that contains all possible results
- ◆ Want to measure a property of the parent distribution, but can only sample it
- ◆ From sample want to *estimate* the *true* value of the property and how far wrong our measurement may be
- ◆ Could be that sample is drawn from a distribution function that arises from some basic law. Function has both properties and parameters
 - Often not distinguished, for Poisson μ is both, for Gaussian μ and σ are both
 - For binomial p is a parameter, but its mean and width are np and $np(1-p)$

Properties of Estimators

- ▶ Isn't best method obvious? Often but by no means always the case.
- ▶ As soon as distributions get non-Gaussian things get more complicated. May not even be possible to define “best” estimator!
- ▶ Important properties for an estimator:
 - Consistency
 - Bias
 - Efficiency

Properties of Estimators

- ▶ Define a few (more) symbols
 - Trying to measure property a
 - Estimator of a denoted by \hat{a}
- ▶ Apply estimator \hat{a} to N measurements of sample
- ▶ \hat{a} can and will vary from true value.
Law of large numbers says effect should get smaller and smaller as $N \rightarrow \infty$, provided measurements are independent

Estimators for Average Student Height

- 1) Sum heights and divide by # of measurements, N
- 2) Sum heights of 1st 10, divide by 10, ignore the rest
- 3) Sum heights and divide by $(N-1)$
- 4) Ignore data, answer is 1.80 m
- 5) Add tallest and shortest and divide by 2
- 6) Height for which $\frac{1}{2}$ students above and $\frac{1}{2}$ below (median)
- 7) Multiply heights and take N th root (geometric mean)
- 8) Take most popular height (mode)
- 9) Add 2nd, 4th, 6th and divide by $N/2$ (or $(N-1)/2$)

All are valid estimators (even 2,3,4)
Which is “best”?

Consistency

- ▶ *Consistency* means that difference between estimator and true value should vanish for large N

$$\lim_{N \rightarrow \infty} \hat{a} = a$$

- ▶ Are estimators 1) to 9) consistent?

- 2) and 4) not consistent,

- 3) OK as for large N dividing by N or $(N-1)$ gives same result

- 1) $\hat{\mu} = \frac{x_1 + x_2 + \dots + x_N}{N} = \bar{x}$ Law of large numbers implies $\bar{x} \rightarrow \mu$ as $N \rightarrow \infty$

- 1), 3), 9) therefore consistent

- 5), 6), 7), 8) need more work. Whether they are consistent depends on how we want to define average and shape of distributions. For symmetric distributions all the same, but not for asymmetric

Bias

- ▶ \hat{a} may be smaller or larger than a , but hope that chances that it overestimates a are same as underestimates a
- ▶ This is called an *unbiased* estimator: $\langle \hat{a} \rangle = a$
 - Apply to 1) clearly unbiased
 - 3) biased
 - 2) and 9) unbiased
- ▶ Still not enough to find “best” estimator

$$\begin{aligned}\langle \hat{\mu} \rangle &= \left\langle \frac{x_1 + x_2 + \dots + x_N}{N} \right\rangle \\ &= \frac{\langle x \rangle + \langle x \rangle + \dots + \langle x \rangle}{N} \\ &= \frac{N \langle x \rangle}{N} = \mu\end{aligned}$$

Efficiency + Robustness

- ▶ Estimator is *efficient* if its variance is small, i.e. want estimator that is in general as close as possible to true value
 - cf 9) and 1):
 - 9) only uses half of data, so its variance will be twice as large
- ▶ Harder to quantify is *robustness*.
 - How stable is estimator against large measurement fluctuations, wrong data, wrong assumptions for the underlying p.d.f. etc?
 - Maximum likelihood discussed below is usually most efficient estimator, but can be very sensitive to form of p.d.f.
- ▶ Efficiency of estimator depends on p.d.f. Often the case that the most efficient estimator is biased – needs to be corrected

Likelihood

- ◆ Set of measurements x_1, x_2, \dots, x_N + estimator \hat{a} of the true quantity a
- ◆ Apply estimator and hope that $\hat{a}(x_1, x_2, \dots, x_N)$ is close to a ; depends on the measurements x_i
- ◆ To decide on estimator proceed in other direction:
 - Assume a form for the p.d.f.; data values x_i are drawn from this distribution
 - Distribution is a function of x and of a , i.e. we treat a as a parameter rather than a property of the distribution
 - Can then make statements such as “If I take a sample of N values from this distribution and use it to calculate $\hat{a}(x_1, x_2, \dots, x_N)$ on average the value \hat{a} of that I find will be $\langle \hat{a} \rangle$ ”

Likelihood

- Probability to get a series of measurements x_1, x_2, \dots, x_n

$$\begin{aligned}L(x_1, x_2, \dots, x_N; a) &= P(x_1; a)P(x_2; a)\cdots P(x_N; a) \\ &= \prod_{i=1}^N P(x_i; a)\end{aligned}$$

- Called likelihood and depends on both measurements and true value a
- Can define expectation value for any function of whole sample

$$\begin{aligned}\langle f(x_1, x_2, \dots, x_N) \rangle &= \int \cdots \int f(x_1, x_2, \dots, x_N) L(x_1, x_2, \dots, x_N; a) dx_1 dx_2 \cdots dx_N \\ &= \int f L dX\end{aligned}$$

Expectation
value of
estimator

$$\begin{aligned}\langle \hat{a} \rangle &= \int \hat{a} L dX \\ \langle \hat{a}^2 \rangle &= \int \hat{a}^2 L dX\end{aligned}$$

Likelihood



- ▶ Consistency for estimator: $\lim_{N \rightarrow \infty} \langle \hat{a} - a \rangle = 0$
 - Often depends on p.d.f
 - 6), 7), 8) all consistent if distribution is symmetric
 - 5) also then consistent
 - 1), 3) and 9) always consistent
- ▶ Bias can be evaluated in same way. Often easy to correct.
 - If $\langle \hat{a} \rangle = b$, then $(\hat{a} - b)$ is unbiased
 - If estimator is consistent, bias vanishes for $N \rightarrow \infty$
- ▶ Efficient estimator? $V(\hat{a}) = \langle \hat{a}^2 \rangle - \langle \hat{a} \rangle^2$
 - If we know $P(x; a)$ can calculate $V(\hat{a})$

Minimum Variance Bound (MVB)

- Can show that $V(\hat{a})$ is limited by:

$$V(\hat{a}) \geq \frac{1}{\langle (d \ln L / da)^2 \rangle}$$

MVB

- Useful relationship:

$$\left\langle \frac{d^2 \ln L}{da^2} \right\rangle = - \left\langle \left(\frac{d \ln L}{da} \right)^2 \right\rangle$$

- If for some estimator, \hat{a} , $V(\hat{a})$ is equal to MVB, estimator is “efficient” – it satisfies the MVB
- Efficiency given by $MVB/V(\hat{a})$

Example 1 - Estimators of Mean

- ▶ Sample mean is a consistent and unbiased estimate of true mean:

$$\hat{\mu} = \bar{x}$$
$$V(\hat{\mu}) = \sigma^2 / N$$

- ▶ σ is standard deviation of parent distribution,
 N is number of data values

Example 1 - Estimators of Mean

- ▶ Gaussian:

$$P(x_i; \mu) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x_i - \mu)^2 / 2\sigma^2}$$

$$\ln L = -\sum_i \frac{(x_i - \mu)^2}{2\sigma^2} - N \ln(\sigma \sqrt{2\pi})$$

- ▶ Likelihood depends on parameter μ , hence:

$$\frac{d^2 \ln L}{d\mu^2} = -\frac{N}{\sigma^2}$$

- ▶ Expression does not depend on x_i , so:

Variance from CLT,
so estimator is efficient

$$MVB = \frac{\sigma^2}{N}$$

Example 1 - Estimators of Mean

- ▶ Uniform distribution, where limits are not known
 - Sample mean is consistent and unbiased, but not most efficient.
 - More efficient is 2 most extreme measurements:

$$\hat{\mu} = \frac{1}{2} [\min(x_i) + \max(x_i)]$$

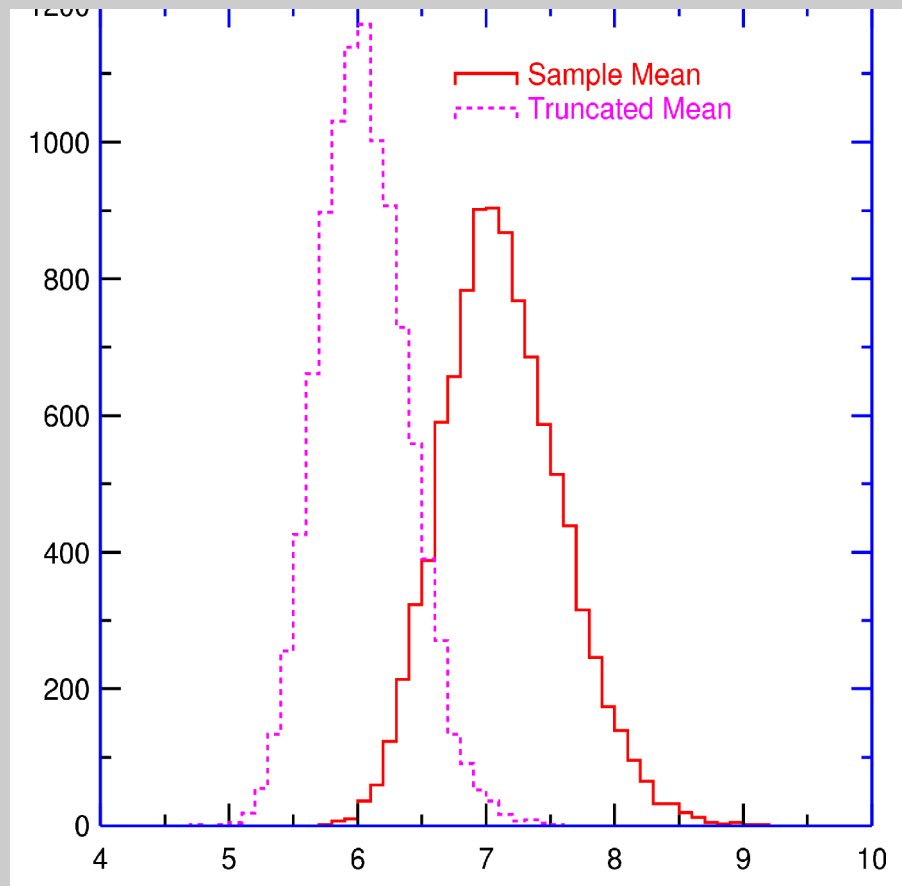
$$V(\hat{\mu}) = \frac{W^2}{2(N+1)(N+2)}$$

- Variance goes as $1/N^2$ instead of $1/N$ for sample mean!

Reason to use head/tail for position in microstrip detectors

Landau Distribution

- Landau - particle passing through material. dE/dx follows Landau - large tail to high energies



- Use sample mean to estimate true mean
 - Long tails lead to inefficient estimator
- Better is to *ignore* a certain fraction of the measurements in the tail
- e.g. mean energy loss 7.5, width 1.0, 20 measurements
 - Simulation shows efficiency of mean can be increased from 44% to 55% (width of estimator distribution is 0.35), if highest 9/20 measurements are ignored

Landau Distribution

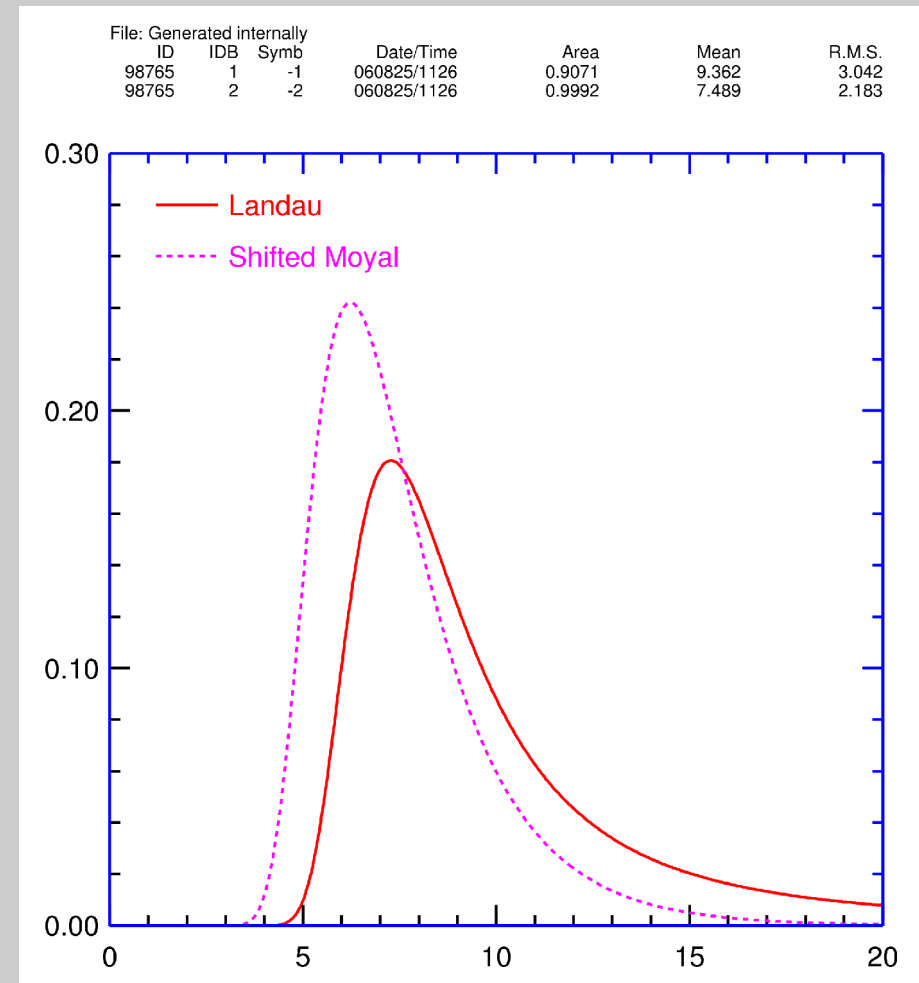
- Use Moyal to simulate Landau

$$f(x) dx = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x + \exp(-x))\right)$$

- Peaks at 0 and has mean 1.27
- Use transformation to generate any \bar{E} and width, W

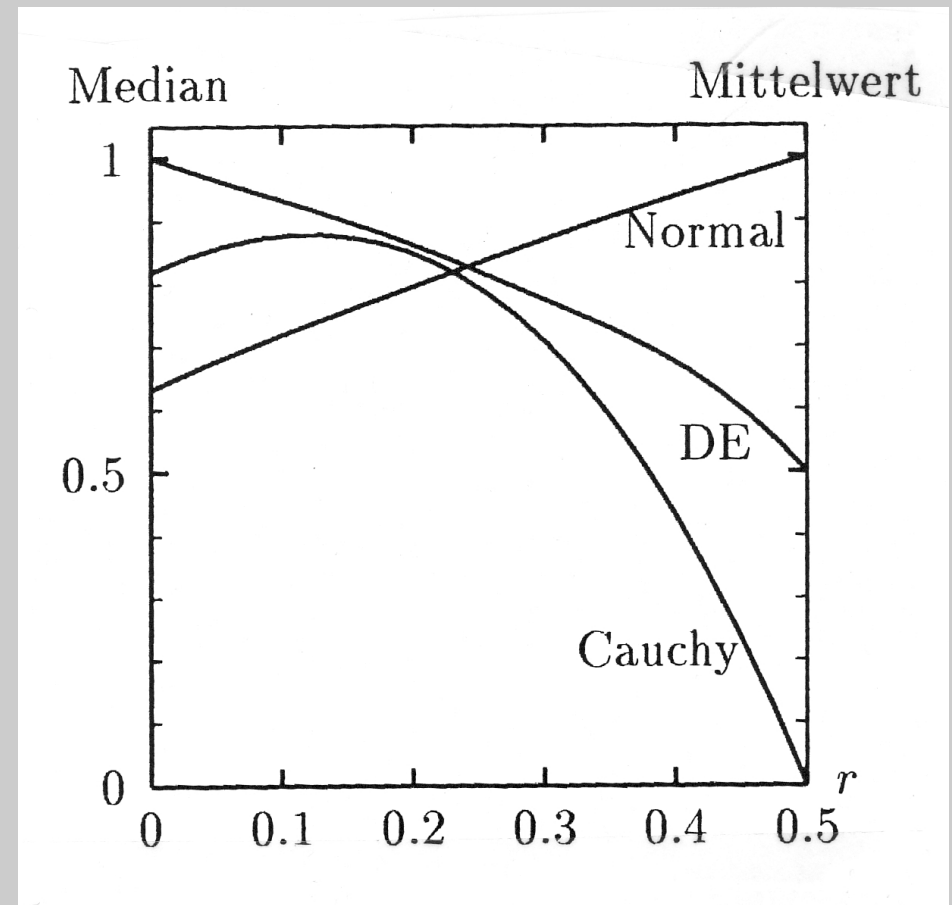
$$x = 1.27 + (E - \bar{E})/W$$

- Distributions similar, but not such a great match!
- CERN library has a better function G110:
Landau Distribution DENLAN



Truncated Means

- ▶ Truncated mean usually used as estimator of dE/dx in tracking chambers
- ▶ Similar arguments apply for other distributions with long tails, e.g. Double exponential and Breit-Wigner (Cauchy)
- ▶ Plot efficiency vs. r , $(1-2r)N/2$ largest and smallest measurements ignored
 - $r=0$ means median,
 - $r=0.5$ means sample mean



Estimator of Variance

- Assuming true mean is known use:

$$\widehat{V}(\mathbf{x}) = \frac{1}{N} \sum (x_i - \mu)^2$$

- Consistent and unbiased

- If μ not known use sample mean?

- Biased:

$$\langle \widehat{V}(\mathbf{x}) \rangle = \frac{N-1}{N} V(\mathbf{x}) \neq V(\mathbf{x})$$

- Bessel's correction:

$$\widehat{V}(\mathbf{x}) = s^2 = \frac{1}{N-1} \sum (x_i - \bar{x})^2$$

- Is a consistent and bias free estimator

- Variance of estimator?

- Have to evaluate:

$$\langle V(\mathbf{x})^2 \rangle - \langle V(\mathbf{x}) \rangle^2$$

- Complicated

- For large N simplifies to

$$\begin{aligned} V(\widehat{V}(\mathbf{x})) &= \frac{1}{N} [\langle (x-\mu)^4 \rangle - \langle (x-\mu)^2 \rangle^2] \\ &= \frac{2\sigma^4}{N} \text{ for Gaussian} \end{aligned}$$

- Can also apply Bessel's correction if necessary

Estimator of Standard Deviation

- Obvious estimator:

$$\begin{aligned}\hat{\sigma} &= \sqrt{V(\bar{x})} \\ &= s \text{ with Bessel}\end{aligned}$$

- Consistent, but may be biased
- Calculate variance by propagating errors – for N reasonably large

$$V(\hat{\sigma}) = \frac{\langle (x-\mu)^4 \rangle - \langle (x-\mu)^2 \rangle^2}{4N\sigma^2}$$

$$= \frac{\sigma^2}{2N} \text{ for Gaussian}$$

$$\sigma_s = \frac{\sigma}{\sqrt{2(N-1)}} \text{ for unbiased est.}$$

- Make 40 measurements find estimate for σ of 6.0, can quote result as:

$$\hat{\sigma} = 6.0 \pm 0.7$$

The Different σ

- Standard deviation of a data sample:

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

- Unbiased estimator:

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

- For parent distribution (same symbol, but in terms of expectation values):

$$\sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$$

- Know the true mean:

$$\sigma = \sqrt{\frac{1}{N} \sum_i (x_i - \mu)^2}$$

Effect of differences small for large N, but be careful for small N

Maximum Likelihood

- Value of \hat{a} which maximises

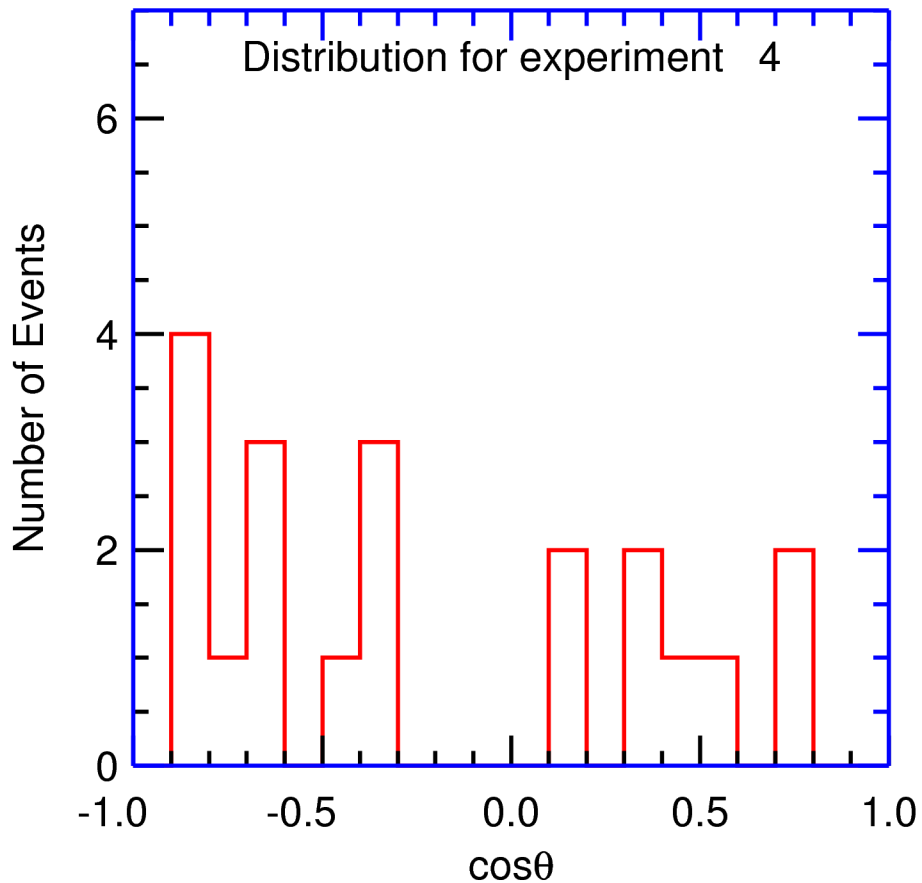
$$L(x_1, x_2, \dots, x_N) = \prod P(x_i; a)$$

- Determine value of a for which probability of measurements x_1, x_2, \dots, x_N is maximum. Sum is easier to use than product, so usually take logarithm
- In fact use $-\ln L$, as most programs minimise rather than maximise
- Consider a set of 20 measurements of angular distribution which should follow

$$P(x; a) = \frac{1}{2}(1 + ax) \text{ with } x = \cos \theta$$

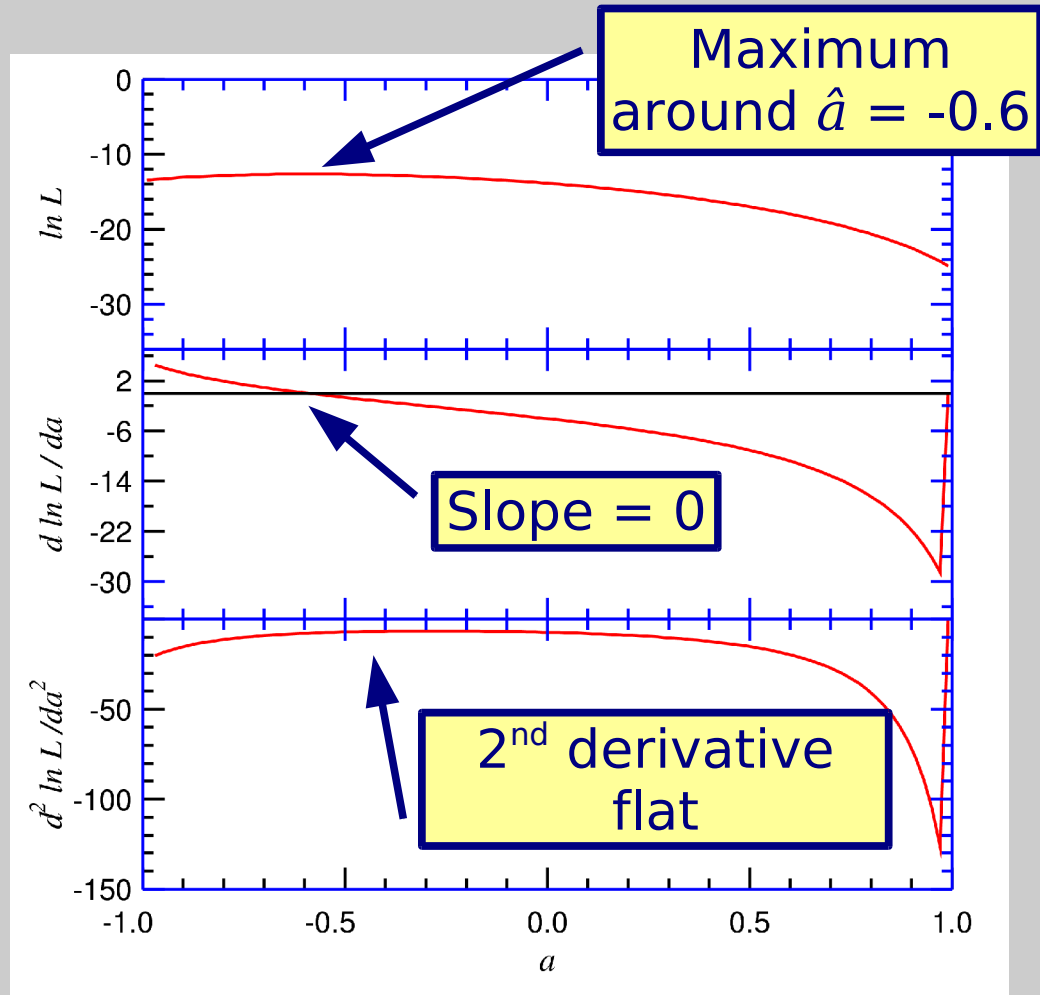
- Normalised for range $-1 \leq \theta \leq +1$, so gives probability directly

Angular Distribution



If 2nd derivative flat,
likelihood follows a parabola,
so likelihood function is Gaussian

- Look at likelihood as a function of a :



Particle Lifetime

- ▶ Unstable particle with lifetime τ . Measure decay time - should follow:

$$P(t; \tau) = \frac{1}{\tau} e^{-t/\tau}$$

- ▶ How to we estimate lifetime from measurements, t_i ?

$$\begin{aligned} \ln L &= \sum_i \ln \left(\frac{1}{\tau} e^{-t_i/\tau} \right) \\ &= \sum_i \left[-\frac{t_i}{\tau} - \ln \tau \right] \end{aligned}$$

- ▶ Differentiate w.r.t. the “true” value τ :

$$\begin{aligned} \frac{d \ln L}{d \tau} &= \sum_i \left[-\frac{t_i}{\tau^2} - \frac{1}{\tau} \right] \\ 0 &= \sum_i \left[-\frac{t_i}{\tau^2} - \frac{1}{\tau} \right] \quad \text{at max.} \\ \hat{\tau} &= \frac{1}{N} \sum_i t_i \end{aligned}$$

- ▶ Can therefore estimate ML analytically; it is just sample mean of measurements

Particle Lifetime

- ▶ Experiment unrealistic, as it is assumed that all possible times can be measured.
- ▶ Suppose we have an upper limit, T (probably have both):

$$P(t; \tau) = \frac{1}{\tau} \frac{e^{-t/\tau}}{(1 - e^{-T/\tau})}$$

- Change in p.d.f. is small, but ...

$$\ln L = \sum_i \left[-\ln \tau - \frac{t_i}{\tau} + \ln(1 - e^{-T/\tau}) \right]$$

- ▶ Differentiating and setting to 0, we obtain:

$$\hat{\tau} = \frac{1}{N} \sum_i t_i + \frac{1}{N} \sum_i \frac{T e^{-t_i/\hat{\tau}}}{1 - e^{-T/\hat{\tau}}}$$

- Can't be solved analytically anymore
- Extend in a similar way for lower and upper limits

ML for a Gaussian

- Suppose we have $\{x_i\}$ measurements of a quantity, each with a different precision, σ_i
- x_i taken from Gaussian with mean μ and σ_i - we want to find $\hat{\mu}$, and σ_i are known
- Differentiate w.r.t. $\hat{\mu}$ and set to 0:
$$\sum_i \left(\frac{x_i - \hat{\mu}}{\sigma_i^2} \right) = 0$$
$$\hat{\mu} = \frac{\sum_i x_i / \sigma_i^2}{1 / \sigma_i^2}$$
- Exactly same as form derived before for weighted mean

$$\ln L = \sum_i -\ln(\sigma_i \sqrt{\pi}) - \sum_i \frac{(x_i - \mu)^2}{2\sigma_i^2}$$

ML for a Gaussian

- ◆ Set of measurements from a distribution with a mean μ ; resolution of each measurement is same but unknown.
- Want to estimate μ and σ
- Differentiate $\ln L$ w.r.t μ and σ

$$\begin{aligned}\sum (x_i - \hat{\mu}) &= 0 \\ \sum \frac{(x_i - \hat{\mu})^2}{\hat{\sigma}^3} - \sum \frac{1}{\hat{\sigma}} &= 0 \\ \hat{\mu} &= \frac{1}{N} \sum x_i \\ \hat{\sigma}^2 &= \frac{1}{N} \sum (x_i - \bar{x})^2\end{aligned}$$

This was 1st guess at estimator for variance, but it is biased!

Properties of ML Estimator

- ▶ Usually consistent, but often biased!
- ▶ Invariant under parameter transformations
 - In above example estimator for variance is:
$$\hat{\sigma}^2 = \frac{1}{N} \sum (x_i - \bar{x})^2$$
 - Same as that of standard deviation squared
 - Reason is that we are always looking for turning point of likelihood. At this point $a = a_1$, but also $a^2 = a_1^2$
 - Such an invariance is incompatible with an unbiased estimator as df/da is involved in transformation
- ▶ Can show that for large N the ML estimator fulfils the MVB and is therefore an efficient estimator

Errors on ML Estimator

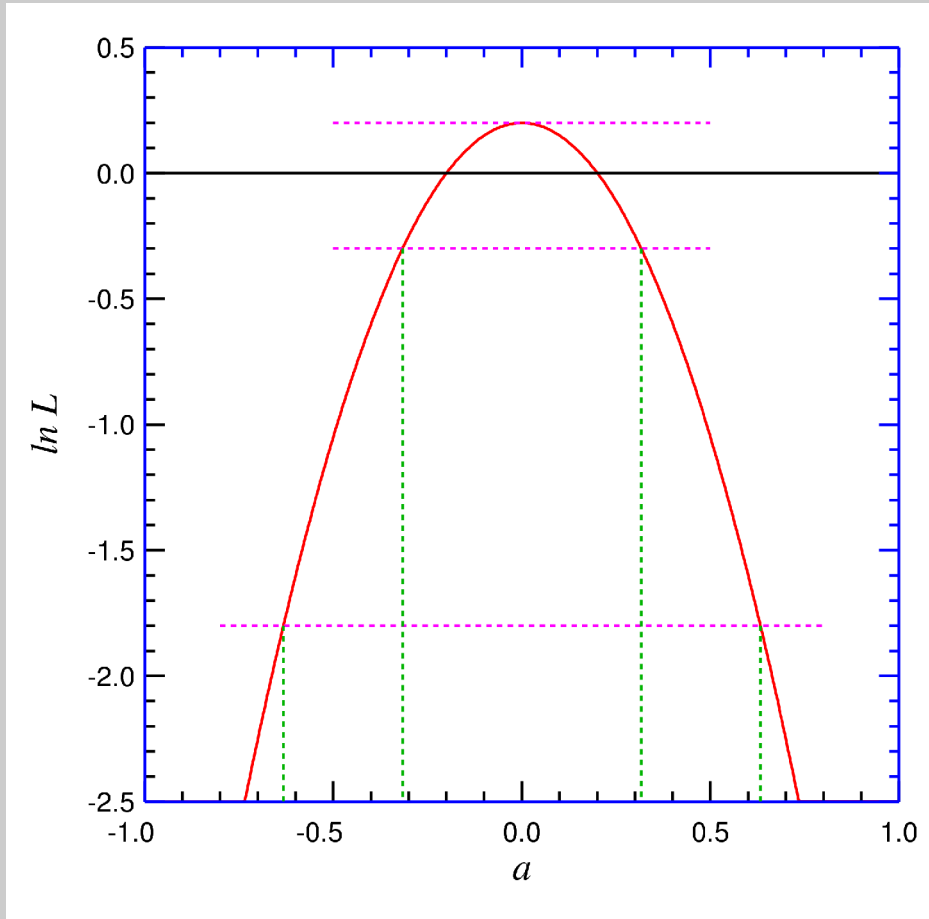
- ▶ Can show that

$$L(x_1, x_2, \dots, x_N; a) \propto \exp(-A[a - \hat{a}(x_1, x_2, \dots, x_N)]^2/2)$$

$$A = - \left\langle \frac{d^2 \ln L}{da^2} \right\rangle = - \left. \frac{d^2 \ln L}{da^2} \right|_{a=\hat{a}}$$

- provided $d^2 \ln L / da^2$ is almost constant for a close to a_0
- ▶ In large N limit log likelihood is a parabola and likelihood follows a Gaussian. Standard deviation of Gaussian is $1/\sqrt{A}$, which is also the standard deviation of the estimator \hat{a}
- ▶ Can read off errors on from plot of $\ln L$ vs. a
- ▶ Note that this does not only apply to large N limit, as even if $\ln L(a)$ is not parabolic we can presumably find a variable a' which is

Errors on ML Estimator



- ▶ 1σ error reduces $\ln L$ by 0.5
- ▶ 2σ error reduces $\ln L$ by 2.0
- ▶ 3σ error reduces $\ln L$ by 4.5
- ▶ In case of transformation to a' corresponds to $\ln L(a')$ changing by 0.5
- Calculate corresponding a ; errors are no longer necessarily symmetric
- Same procedure for 2σ limits, but 2σ errors not necessarily 2x bigger for finite N

ML and Histograms

- So far considered each measurement individually
- Measurements often binned into histogram with bin width Δx
- Want to fit prob. dist. to contents of each bin
- Distribution gives # events in each bin
- Actual number distributed according to a Poisson with expectation value μ_j for each bin

$$P(n_j; \mu_j) = \frac{\mu_j^{n_j} e^{-\mu_j}}{n_j!}$$

- How do we obtain μ_j ?
- Depends on parameter a . Write down $\ln L$:

$$\begin{aligned}\ln L &= \sum_{j=1}^{Nbin} \left(\ln \frac{\mu_j^{n_j} e^{-\mu_j}}{n_j!} \right) \\ &= \sum_{j=1}^{Nbin} \left[n_j \ln \mu_j - \mu_j - \ln(n_j!) \right] \\ &= \sum_{j=1}^{Nbin} \left[n_j \ln \frac{\mu_j}{n_j} - (\mu_j - n_j) \right] + \text{const}\end{aligned}$$

Nice form for numerical maximisation - also works for $n_j = 0$

ML for Several Variables

- Generalise previous method. Have to solve p simultaneous equations, where p is number of variables:

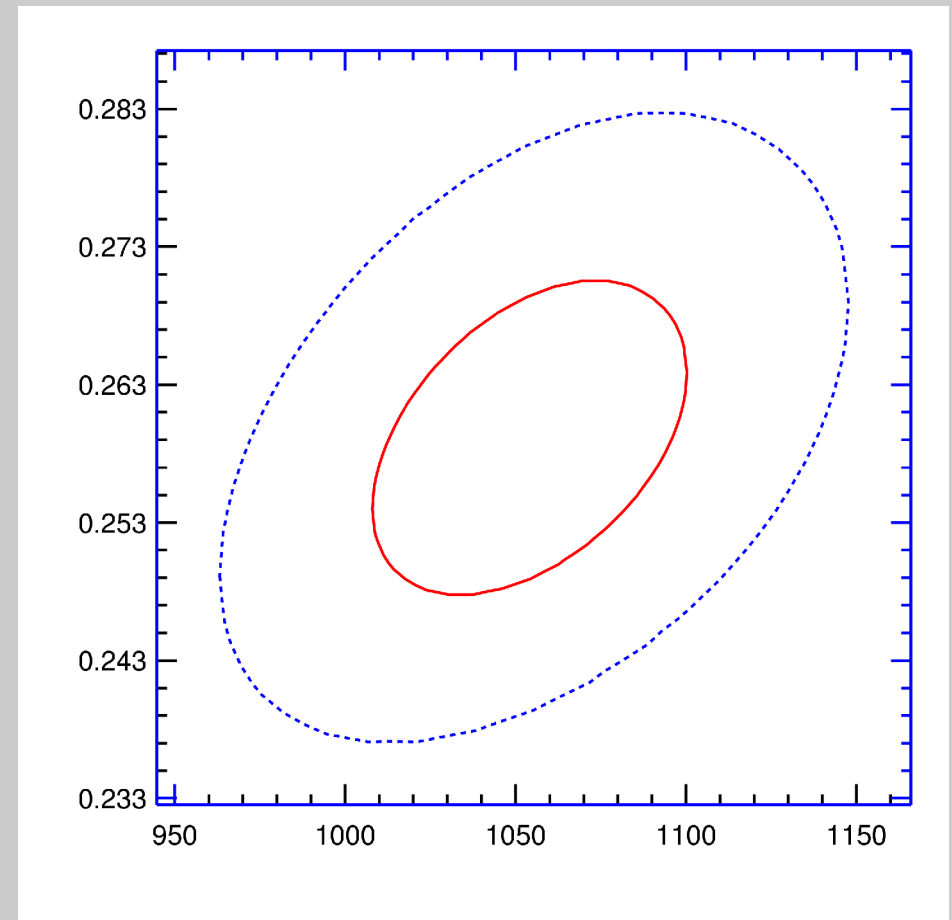
$$\frac{\partial \ln L(x_1, x_2, \dots, x_N; a_1, a_2, \dots, a_p)}{\partial a_j} = 0$$

- Covariance is given by matrix inversion of the matrix of 2nd derivatives:

$$\text{cov}^{-1}(a_i, a_j) = - \left\langle \frac{\partial^2 \ln L}{\partial a_i \partial a_j} \right\rangle = \left\langle \frac{\partial \ln L}{\partial a_i} \frac{\partial \ln L}{\partial a_j} \right\rangle = - \frac{\partial^2 \ln L}{\partial a_i \partial a_j} \Big|_{a=\hat{a}}$$

ML for Several Variables

- ▶ Look at fit results etc. using forms such as contour plots
- ▶ Such plots also illustrate if the parameters are correlated or not



Extended Maximum Likelihood

- ▶ Up to now ML worked with a total prob. of 1
- ▶ Implicitly assumes we know number of events
- ▶ If we measure something for fixed length of time, number of events no longer fixed
- ▶ Introduce extended maximum likelihood (EML)
- ▶ Total number of events given by:

$$\int Q(x; a) dx = \nu$$

- ▶ Expect ν events, probability to observe N is given by:

$$e^{-\nu} \frac{\nu^N}{N!}$$

- ▶ Form likelihood:

$$\begin{aligned} \ln L &= \sum [\ln P(x_i; a)] - \nu + N \ln \nu \\ &= \sum [\ln \nu P(x_i; a)] - \nu \\ &= \sum \ln Q(x_i; a) - \nu \end{aligned}$$

- ▶ Maximise to determine a . Increasing normalisation of Q increases L ; $-\nu$ term compensates, maximisation finds balance

Extended Maximum Likelihood

- ▶ Properties of EML and ML very similar: bias for finite N and asymptotically efficient
- ▶ Can also generalise to 2 or more parameters
- ▶ As for lifetime, acceptance affects normalisation of prob. dist. If algebra too hard, use EML instead of ML and let overall normalisation vary
- ▶ Arrange things so that normalisation changes, but shape does not
- ▶ ML and EML solutions will then have same solution for max and fitted # of events ν is same as “true” number N
 - See that this is empirically the case for numerical maximisation
- ▶ Drawback is errors are larger than for ML, as problem believes that number of events can fluctuate

Comments on ML

- ◆ ML is sensible way to estimate values of unknown parameters. Returns the values for which measurements are most probable (not the most probable values for the parameters)
- ☺ In limit $N \rightarrow \infty$ ML estimator is unbiased and Gauss distributed about a with a variance that fulfils MVB
- ☺ No loss of information. Binning not necessary
- ☺ Straightforward to extend to many parameters
- ☹ For small N , ML estimator usually biased
- ☹ You have to know form for $P(x; a)$ and it had better be correct. ML gives you no information about how well your data fit the form you are using
- ☹ Rarely possible to evaluate $d \ln L / da$ analytically. Have to use numerical minimisation techniques, e.g. MINUIT. Don't forget to include normalisation factors when differentiating

Least Squares

- ▶ G.U. Yule & M.G. Kendall, An Introduction to Statistics (1958):

It was formerly the custom, and it is still so in works on the theory of observations to derive the method of least squares from certain theoretical considerations, the assumed normality of the errors of observations being one such. It is, however, more than doubtful whether the conditions for the theoretical validity of the method are realised in statistical practice, and the student would do well to regard the method as recommended chiefly by its comparative simplicity and by the fact that it has withstood the test of experience.

Least Squares

- ◆ Consider a set of r data (x, y) pairs. x_i known exactly, y_i have errors σ_i
- ◆ y are given by $f(x)$, which depends on one or more parameters to be estimated
- ◆ Invoking CLT, distribution of y_i about true y is given by Gaussian. Prob. for a certain y_i for a given x_i is:

$$P(y_i; a) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-(y_i - f(x_i; a))^2 / 2\sigma_i^2}$$

- ◆ Form $\ln L$:

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

- ◆ To maximise likelihood minimise 1st term;

$$\sum \left[\frac{(y_i - f(x_i; a))}{\sigma_i} \right]^2$$

Least Squares

- Can either derive principle or regard it as a sensible estimator.
- Need to minimise:

$$\chi^2 = \sum \left[\frac{(y_i - f(x_i; a))}{\sigma_i} \right]^2$$

- Find value of a for which χ^2 is minimum

$$\left. \frac{d\chi^2}{da} \right|_{a=\hat{a}} = 0$$
$$\sum \frac{1}{\sigma_i^2} \frac{df(x_i; a)}{da} [y_i - f(x_i; a)] \Big|_{a=\hat{a}} = 0$$

- Solve this for a to find estimator \hat{a}
- Also use formula to find error on estimator as solution gives \hat{a} as a function of y_i . Errors on y_i known so can use error propagation
- Easy to extend to several variables a_1, a_2, \dots, a_p with p simultaneous eqn. with p unknowns

A Word of Caution

- ▶ What are σ_i ?
- ▶ Take a binned histogram - number of entries in each bin follows a Poisson
- ▶ Variance is then c_i ; as this is not known used n_i instead?
- ▶ Suppose we expect 5 entries in a bin - Poisson error is $\sqrt{5} = 2.23$
- ▶ Fluctuations mean we sometimes have 3 entries and sometimes 7
- ▶ Assign errors of $\sqrt{3} = 1.73$ and $\sqrt{7} = 2.64$.
- ▶ Weight given to entries differs by a factor $7/3 = 2.33$!

Warning: Should really have ≥ 10 entries per bin in order for least squares method to be valid, if measured frequency distribution used to estimate the errors!

Straight Line Fit

- Function:

$$f(x; m, c) = mx + c$$

- Assume errors on data points are all same; minimise:

$$\sum_i (y_i - mx_i - c)^2$$

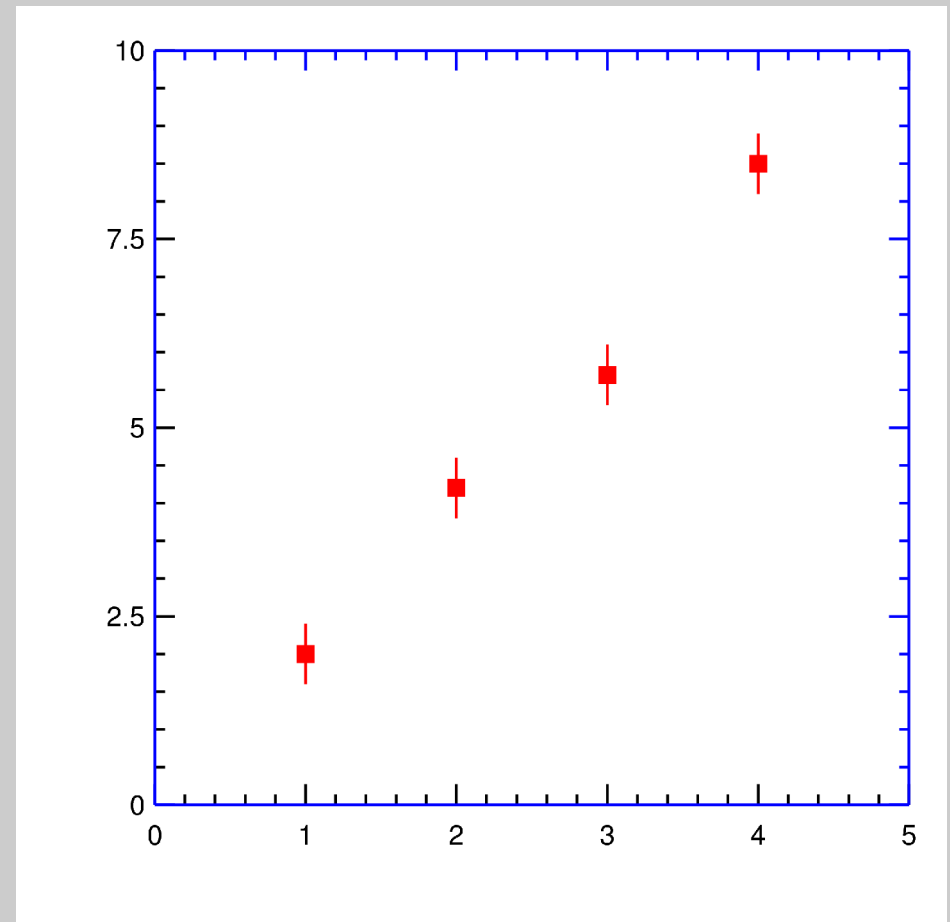
- Differentiate w.r.t. c , then m and set to 0:

$$-2 \sum_i (y_i - \hat{m} x_i - \hat{c}) = 0$$

$$-\overline{y} - \hat{m} \overline{x} - \hat{c} = 0$$

$$-2 \sum_i x_i (y_i - \hat{m} x_i - \hat{c}) = 0$$

$$\overline{x} \overline{y} - \hat{m} \overline{x}^2 - \hat{c} \overline{x} = 0$$



Straight Line Fit

- Solve for \hat{m} and \hat{c}

$$\begin{aligned}\hat{m} &= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} \\ &= \frac{\text{cov}(x, y)}{V(x)} \\ \hat{c} &= \frac{\overline{x^2} \bar{y} - \bar{x} \overline{xy}}{\overline{x^2} - (\bar{x})^2} \\ &= \bar{y} - \hat{m} \bar{x}\end{aligned}$$

- Last form is much easier to remember and shows that line goes through centre of gravity of points (\bar{x}, \bar{y})

- Determine errors via error propagation:

$$\begin{aligned}\hat{m} &= \sum_i \frac{(x_i - \bar{x})}{N(\overline{x^2} - (\bar{x})^2)} y_i \\ V(\hat{m}) &= \sum_i \left[\frac{(x_i - \bar{x})}{N(\overline{x^2} - (\bar{x})^2)} \right]^2 \sigma^2 \\ &= \frac{\sigma^2}{N(\overline{x^2} - (\bar{x})^2)} \\ V(\hat{c}) &= \sum_i \left[\frac{(\overline{x^2} - \bar{x} x_i)}{N(\overline{x^2} - (\bar{x})^2)} \right]^2 \sigma^2 \\ &= \frac{\sigma^2 \overline{x^2}}{N(\overline{x^2} - (\bar{x})^2)}\end{aligned}$$

Straight Line Fit

- Covariance given by:

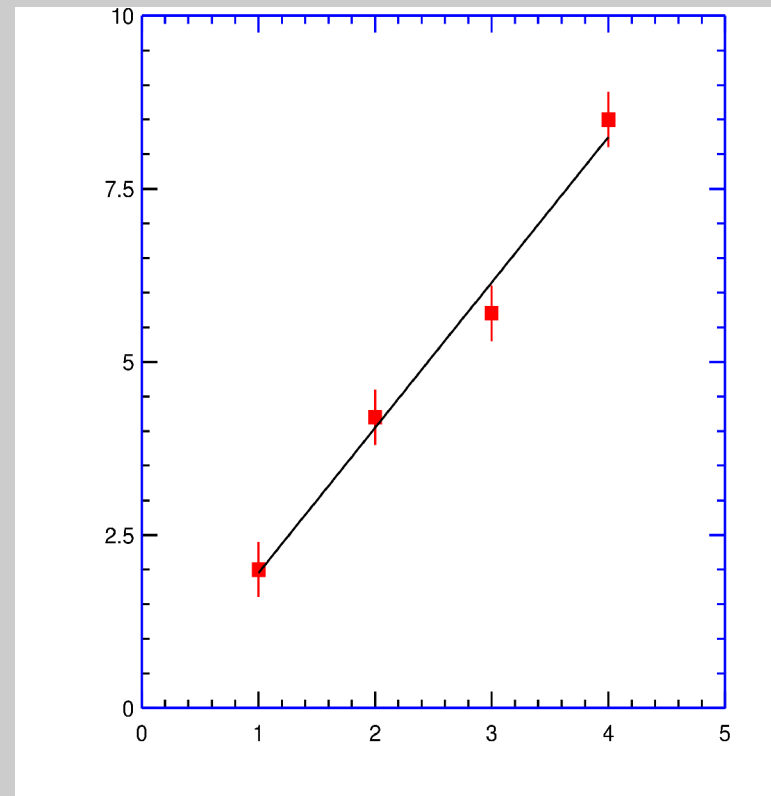
$$\text{cov}(\hat{m}, \hat{c}) = \frac{\sigma^2 \bar{x}}{N(\overline{x^2} - (\bar{x})^2)}$$

$$\rho_{\hat{m}, \hat{c}} = -\frac{\bar{x}}{\sqrt{\overline{x^2}}}$$

- See that if x_i are scaled such that $\bar{x} = 0$ the \hat{m}, \hat{c} estimators are uncorrelated
- Note that quantities such as $V(x), V(y)$ are for whole sample and not a single measurement

- Can calculate χ^2 without determining \hat{m}, \hat{c}

$$\chi^2 = \frac{V(y)}{\sigma^2} (1 - \rho_{x,y}^2)$$



Weighted Straight Line Fit

- Now have to minimise:

$$\sum_i \frac{(y_i - mx_i - c)^2}{\sigma_i^2}$$

- Equations are same as before except simple averages become weighted averages and N becomes sum of weights:

$$\frac{\sum y_i}{N} \rightarrow \frac{\sum y_i / \sigma_i^2}{\sum 1 / \sigma_i^2}$$
$$\sigma^2 \rightarrow \bar{\sigma}^2 = \frac{\sum \sigma_i^2 / \sigma_i^2}{\sum 1 / \sigma_i^2} = \frac{N}{\sum 1 / \sigma_i^2}$$

Extrapolation

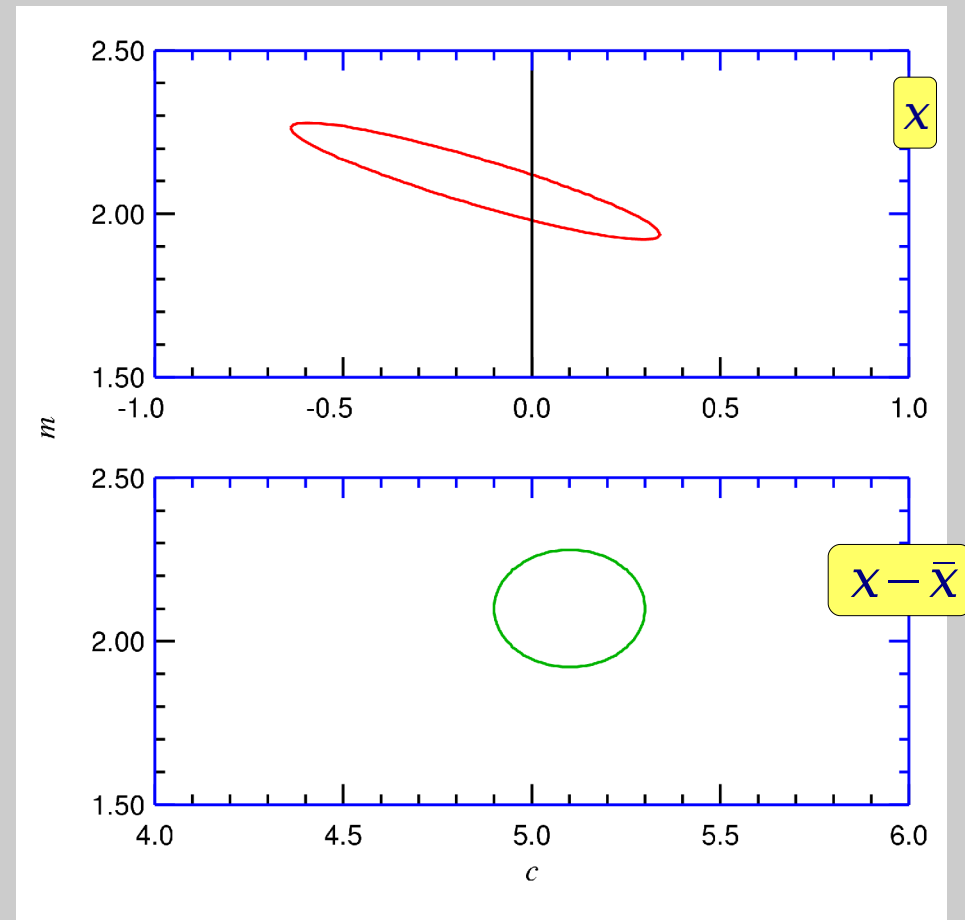
- Want to know value of y for a given value of x including error:

$$y = \hat{m}x + \hat{c}$$

$$V(Y) = V(\hat{c}) + X^2 V(\hat{m}) + 2X \text{cov}(\hat{m}, \hat{c})$$

- Covariance term is important \hat{m}, \hat{c} are anticorrelated
- If $\bar{x} = 0$ \hat{m}, \hat{c} are not correlated, so better to fit: $y = \hat{m}(x - \bar{x}) + \hat{c}'$

$$V(Y) = \frac{\sigma^2 (X - \bar{x})^2}{N(\overline{x^2} - (\bar{x})^2)} + \frac{\sigma^2}{N}$$



Systematic Errors

- ◆ In addition to random error we have a common systematic error, S , on all points?
 - Points are now correlated, so have to use full error propagation formula $V_f = G V_x G$

- Apply to estimator for m : $\hat{m} = \sum_i \frac{(x_i - \bar{x})}{N(\bar{x}^2 - (\bar{x})^2)} y_i$

$$V(\hat{m}) = \frac{1}{N(\bar{x}^2 - (\bar{x})^2)} \sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x}) \text{cov}(y_i, y_j)$$

$$= \frac{1}{N(\bar{x}^2 - (\bar{x})^2)} \left[\sum_i (x_i - \bar{x})^2 \sigma^2 + \sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x}) S^2 \right]$$

- as $\text{cov}(y_i, y_j) = \delta_{ij} \sigma^2 + S^2$
- If S is constant 2nd term is 0, so we get previous error from slope – as expected as systematic error can only move all points up or down

Systematic Errors

- ▶ Variance for estimator of c is more work...

$$V(\hat{c}) = \frac{\overline{x^2} \sigma^2}{N(\overline{x^2} - (\overline{x})^2)} + S^2$$

- Result is as one would probably naively expect
- ▶ Analysis can be extended to case where different points have different sizes of error (or errors are relative and not absolute) etc.

Least Squares and Low Statistics

- True χ^2 to be minimised

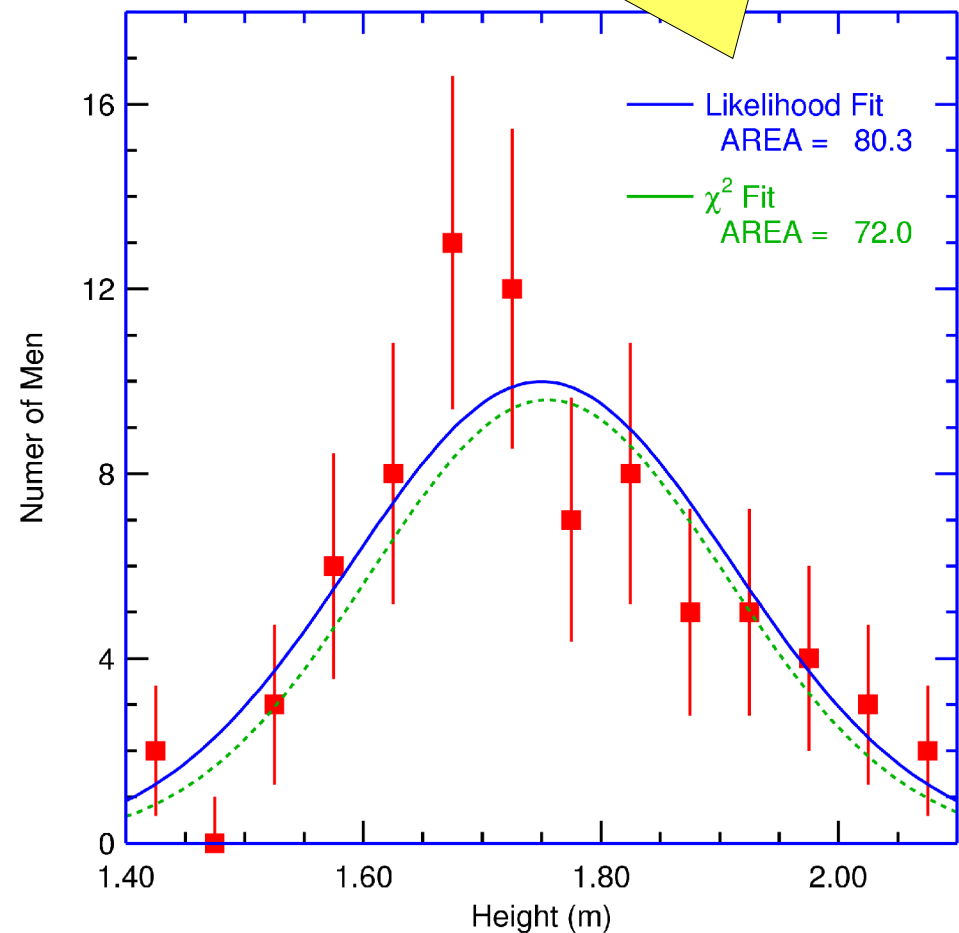
$$\chi^2 = \sum_{j=1}^{N_{bin}} \frac{(n_j - f_j)^2}{f_j}$$

- Much simpler numerically is:

$$\chi^2 = \sum_{j=1}^{N_{bin}} \frac{(n_j - f_j)^2}{n_j}$$

- Be careful with # entries per bin!
- What to do about bins with zero entries?
 - Usual solution (PAW/Root?) is to ignore them!
 - Mn_Fit gives 0 entries and error of 1 - better but not totally satisfactory

Area from χ^2 fit 10% too low!



χ^2 Distribution

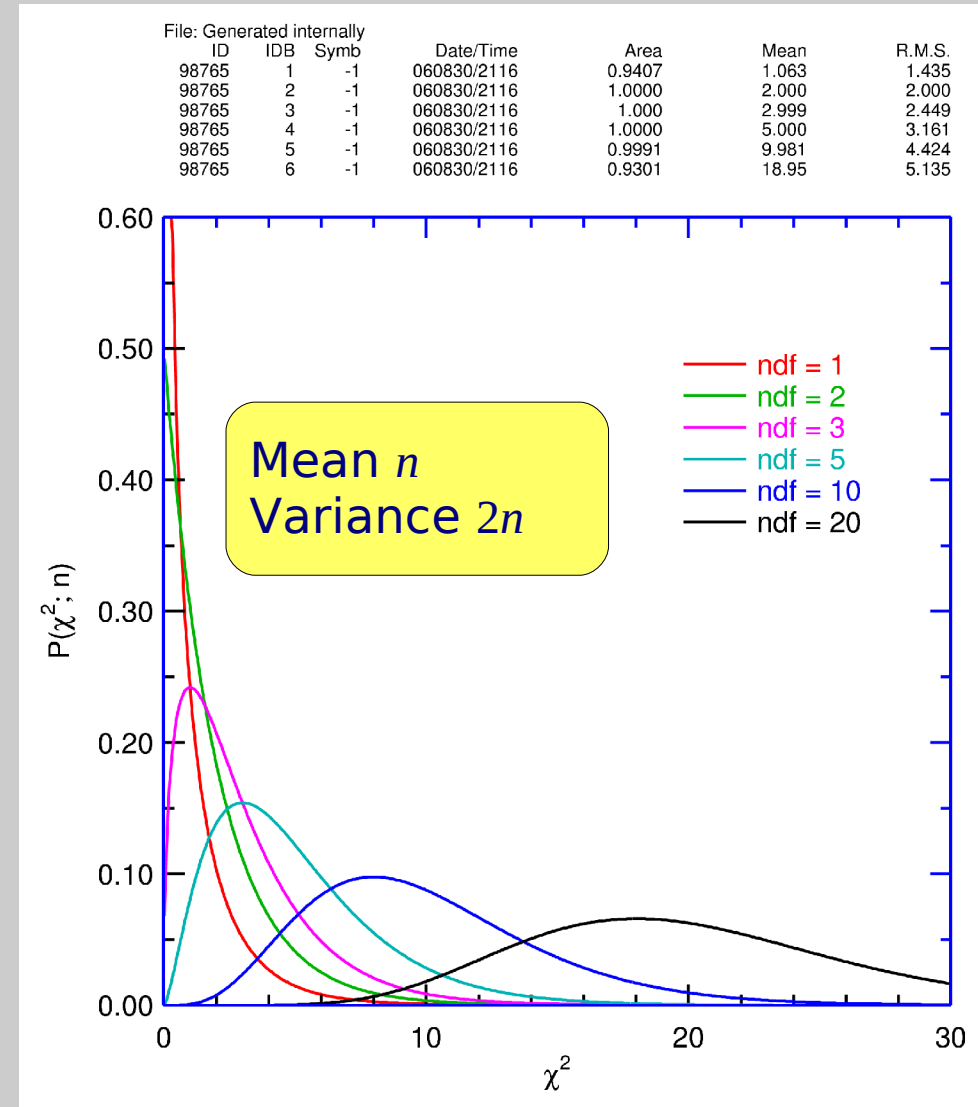
- m^2 small means measurements are close to expectations, large indicates something may have gone wrong. To quantify it need form:

$$P(\chi^2; n) = \frac{2^{-n/2}}{\Gamma(n/2)} \chi^{2(n/2-1)} e^{-\chi^2/2}$$

$$\Gamma(x+1) = x!$$

$$x! = \int_0^{\infty} u^x e^{-u} du$$

- n is number of points in sum minus number of variables adjusted to minimise m^2 , called *number of degrees of freedom (n.d.f)*



Probability Definitions

- ▶ **Frequency definition:**
 - Perform experiment N times
 - Outcome A occurs M times
 - As $N \rightarrow \infty$, M/N tends to limit defined as $P(A)$
- ▶ Set of all possible cases called *collective* or *ensemble*
- ▶ Probability is a property of both experiment and collective
- ▶ Supposes it is possible to repeat experiment many times under identical conditions
- ▶ Conditional probability $p(a|b)$ is probability for a given that b is true
- ▶ Bayes theorem (Rev. Thomas Bayes 1763) uses construction
$$p(a|b)p(b) = P(a \wedge b) = p(b|a)p(a)$$
 - ▶ gives
$$p(a|b) = \frac{p(b|a)p(a)}{p(b)}$$
- ▶ Probability for b regardless of a is:
$$p(b) = p(b|a)p(a) + p(b|\bar{a})[1 - p(a)]$$

Bayesian Statistics

- ▶ Example (from Barlow):
 - Cherenkov counter to separate π from K .
 - Arrange refractive index such that π produces Cherenkov light while K does not – threshold Cherenkov
 - Probability for detector to give a signal for π 95%
 - Probability to get accidental signal from K 6%
 - Particle beam contains 90% π and 10% K
 - Use conditional probability to evaluate probability that particle is π or K if we see a signal and if we do not:

Bayesian Statistics

- See a signal – probability that it is from a π ?

$$\begin{aligned} p(\pi|\text{signal}) &= \frac{p(\text{signal}|\pi)}{p(\text{signal}|\pi)p(\pi)+p(\text{signal}|K)p(K)} p(\pi) \\ &= \frac{0.95}{0.95 \times 0.90 + 0.06 \times 0.10} \times 0.90 = 99.3\% \end{aligned}$$

- With signal pretty sure it is a pion. $p(K|\text{signal})=0.7\%$
Probability also not very sensitive to beam purity and efficiency.
- What is probability that particle is a K ?

$$\begin{aligned} p(K|\text{no signal}) &= \frac{p(\text{no signal}|K)}{p(\text{no signal}|\pi)p(\pi)+p(\text{no signal}|K)p(K)} p(K) \\ &= \frac{0.94}{0.05 \times 0.90 + 0.94 \times 0.10} \times 0.10 = 67.6\% \end{aligned}$$

- Fine to use method to identify π , but not for K

Bayesian Statistics

- ◆ Above calculation used frequency definition
- ◆ What about question such as:
 - Given a particular experimental result, what is probability that theory is true – try to quantify our “degree of belief”:

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

- If prob. for a result given a certain theory is 0, then result disproves theory as then also $p(\text{theory}|\text{result})=0$
- Result unlikely given theory, reduces prob. theory is correct
- Qualified by initial prob. that theory is correct
- Probability that one gets a particular result – to get this necessary to consider all alternatives and sum probabilities

Bayesian Statistics

- ▶ Often think we are using frequency definition, but be aware that Bayesian interpretation often implied
- ▶ Measure mass of electron as $520 \pm 10 \text{ keV}/c^2$
 - Claim true mass close to $520 \text{ keV}/c^2$ and lies with 68% probability between 510 and 530 keV/c^2
 - Does not agree with freq. definition - electron has only one mass and measurement either agrees with it or not
- ▶ Have to use Bayesian statistics to say something about true mass. If we know nothing about electron mass, take probability distribution for mass to be flat:

$$p(m|m_e) \propto e^{-(m-m_e)^2/2\sigma^2}$$

- ▶ Use Bayes theorem to turn it around:

$$p(m_e|m) \propto e^{-(m-m_e)^2/2\sigma^2}$$

- Assumes that $p(m_e)$ is constant - if not information should be included before inverting

Bayesian Statistics

- ◆ If we are trying to measure the value of a parameter a and have measured a value x , Bayes theorem can be written as:

$$p(a|x) = p(x|a) \cdot \frac{p(a)}{p(x)}$$

- ◆ Gives prob. that a parameter has a value a , given a measurement x , in terms of the prob. that we measure x given a particular value of the parameter
- ◆ Can determine the latter using a likelihood fit

- ◆ To say something about “true” value need to say something about $p(x)$ and $p(a)$.
- ◆ $p(x)$ is prod. density for data – fixed and can be absorbed into normalisation
- ◆ $p(a)$ interpreted as “degree of belief”, usually called the *prior*
- ◆ $p(a|x)$ called *posterior*
- ◆ Know nothing about $p(a)$ – take a uniform distribution
- ◆ Otherwise have to include information
 - Betting is a good example, odds are way of quantifying degree of belief

Confidence Levels

- Want to be able to say with a certain level of certainty that a result lies in a particular range or is less than or greater than a certain value
- e.g. bags of rice filled with a mass of 500g with a standard deviation of 5g, can say with 95% confidence level (c.l.) that weight of any particular bag is between 490 and 510g

- In general want to say a value x lies with a certain confidence, C , between 2 values x_- and x_+
$$P(x_- \leq x \leq x_+) = \int_{x_-}^{x_+} P(x) dx = C$$

- 3 possible conventions
 - Symmetric interval
$$x_+ - \mu = \mu - x_-$$
 - Shortest interval
 - Central interval

$$\int_{-\infty}^{x_-} P(x) dx = \int_{x_+}^{+\infty} P(x) dx = (1 - C)/2$$

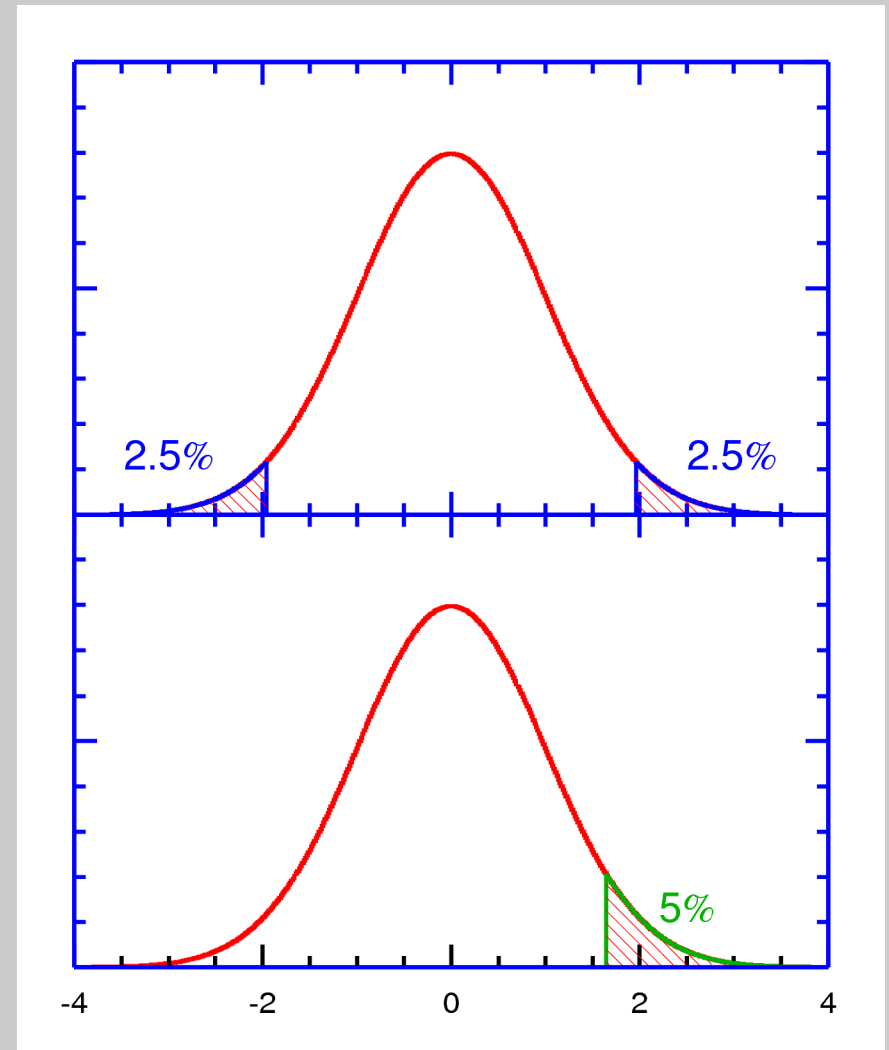
Confidence Levels

- ▶ Central interval mostly used
- ▶ For symmetric distributions definitions are equivalent
- ▶ Often interested in one-sided limits:

$$P(x \leq x_+) = \int_{-\infty}^{x_+} P(x) dx = C \text{ or}$$

$$P(x \geq x_-) = \int_{x_-}^{+\infty} P(x) dx = C$$

- ▶ For Gaussian take $+1.64\sigma$ to define 95% c.i. upper limit and $\pm 1.96\sigma$ to define a 95% central confidence interval

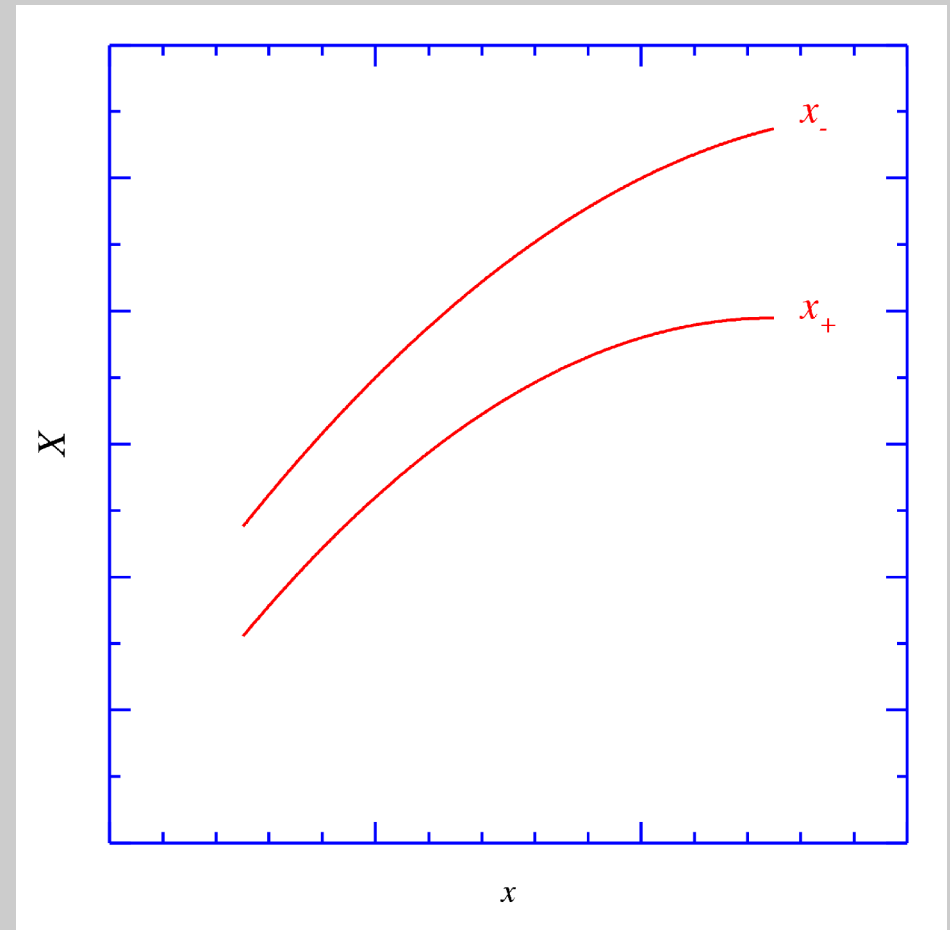


Confidence Levels and Estimation

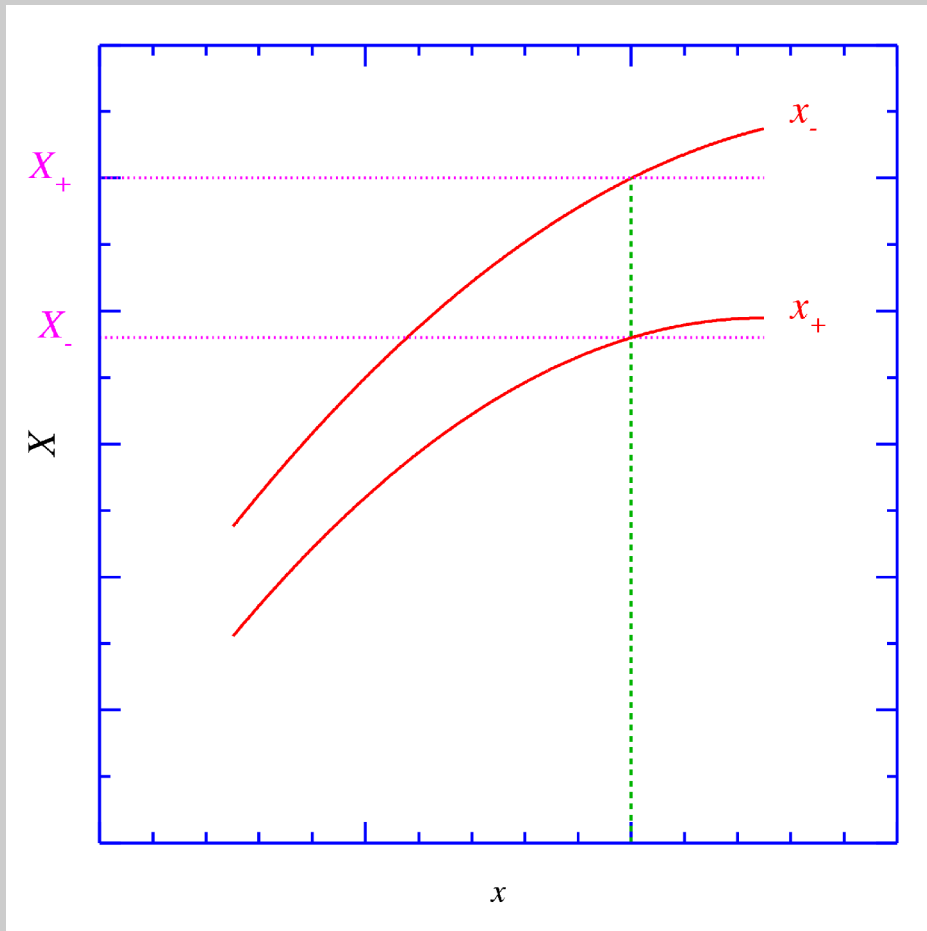
- ▶ Want to know value of a parameter X , have measured value x , know resolution of apparatus $V(x)$
- ▶ Want to turn our knowledge about x and $V(x)$ into a c.l. statement about X
- ▶ Simply saying that with 68% X lies between $x-\sigma$ and $x+\sigma$ contains Bayesian assumption about prior
- ▶ Example:
 - Weight of an empty dish is 25.30 ± 0.14 g
 - Add sample of powder and weigh again
 - Weight now 25.50 ± 0.14 g
 - Powder 0.20 ± 0.20 g?
16% chance weight of powder is negative?

Confidence Levels and Estimation

- ◆ Proceed more carefully
- ◆ Parameter has a value X , which we are trying to measure
- ◆ Measurements, x , follow a prob. dist. $P(x; X)$
- ◆ For each value of X , can construct a confidence interval such that 90% measurements in range
- ◆ Limits vary for different X - draw as a function of x
- ◆ Region between curves called confidence belt
- ◆ Construct diagram before you see the data!



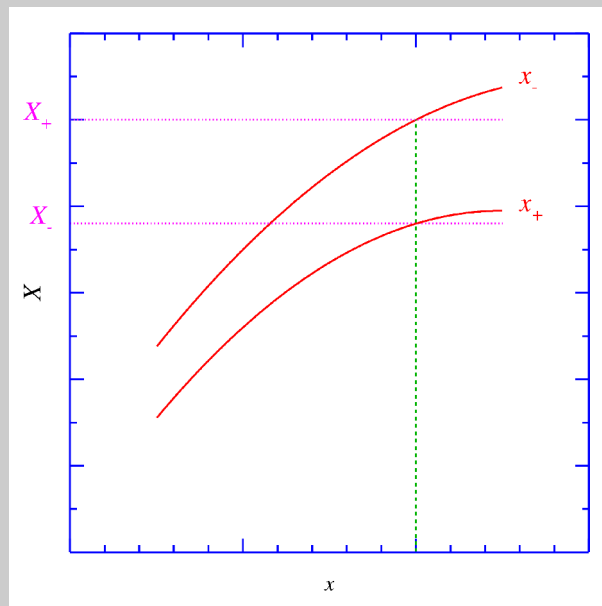
Confidence Levels and Estimation



- ◆ Make a measurement
- ◆ Draw vertical line from this value up to x_- curve
- ◆ Read off upper limit X_+
- ◆ Meaning is that if real X is X_+ or greater, prob. to get a measurement x_- or smaller is 5%
- ◆ Similarly for x_+ curve
- ◆ Quote 90% confidence interval as X_- to X_+

Confidence Levels and Estimation

- ▶ Quote 90% confidence interval for X as range X_- to X_+
- ▶ Means in the frequency interpretation, if you make a large # measurements, 90% of them will be between x_- and x_+ by construction
- ▶ If we use each of measurements to set limits on true value of X , statement on range will be true 90% of time



Confidence Levels for Gaussians

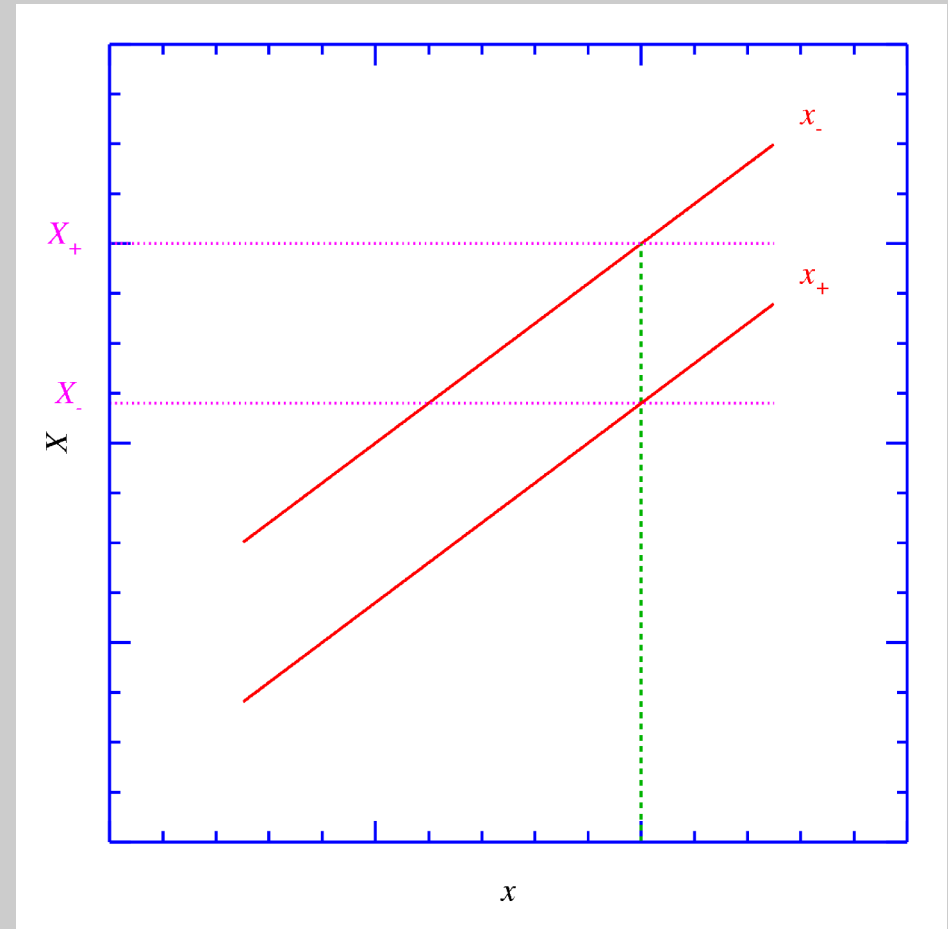
- Conversion from horizontal to vertical scale when $P(x;X)$ follows a Gaussian
- Measurement x with known resolution σ , have to find values X_- and X_+ such that

$$\int_x^{\infty} \frac{1}{\sigma \sqrt{2\pi}} e^{-(x'-X_-)^2/2\sigma^2} dx' = 0.05 \text{ and}$$

$$\int_{-\infty}^x \frac{1}{\sigma \sqrt{2\pi}} e^{-(x'-X_+)^2/2\sigma^2} dx' = 0.05$$

- Turn around and say x lies 1.64σ (90% c.l.) above X_- :

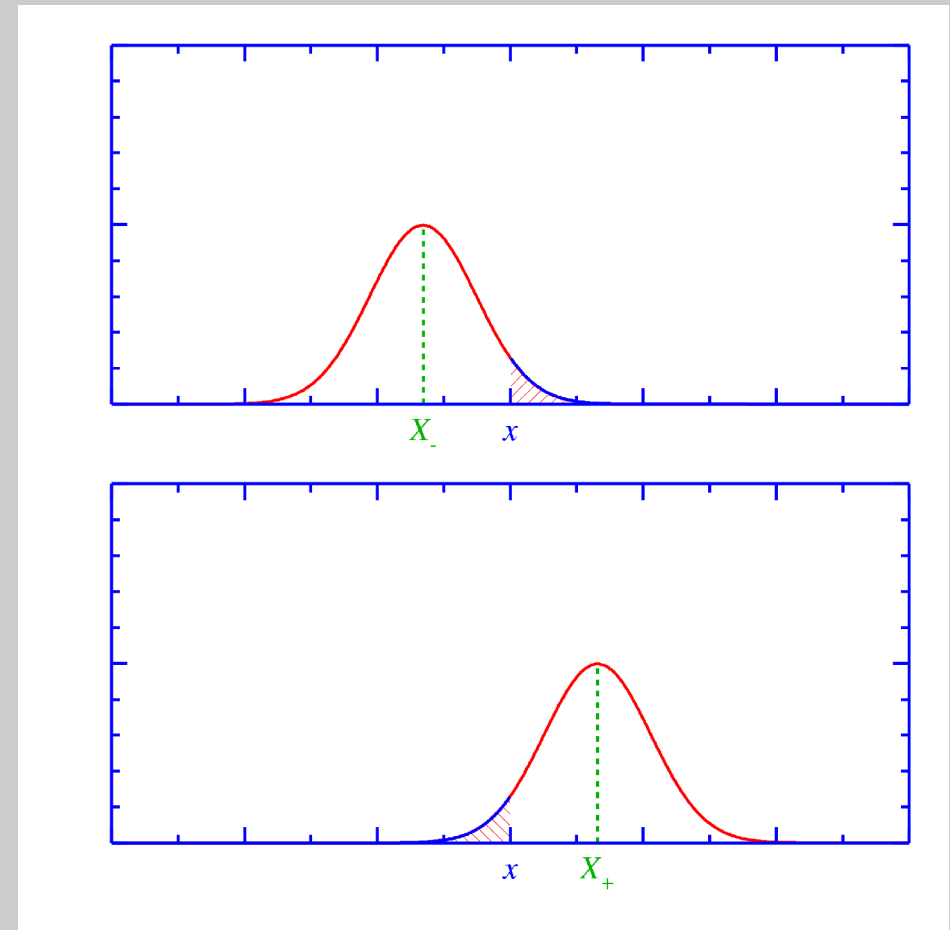
$$\int_{-\infty}^{X_-} \frac{1}{\sigma \sqrt{2\pi}} e^{-(x'-x)^2/2\sigma^2} dx' = 0.05$$



Curves in general figure are straight lines with unit gradient for Gaussian

Confidence Levels for Poisson

- Observe n events, which follow a Poisson with (unknown) mean N - want confidence interval
- 90% c.i. upper limit defined by
by
$$\sum_{r=n+1}^{\infty} P(r; N_+) = 0.90 \text{ or}$$
$$\sum_{r=0}^{n} P(r; N_+) = 0.10$$
- Meaning: If true value of N really N_+ , prob. of getting n or smaller is 10%. If larger prob. of getting n is smaller
- Invert things for lower limit



Confidence Levels with Constraints

- ▶ Measure a quantity that you know cannot be less than or greater than a certain value for physics reasons
- ▶ Measure mass of neutrino from tritium endpoint – you can set a limit on m_ν^2 , so it must be positive
- ▶ Suppose we measure $m_\nu^2 = -0.32 \pm 0.20 \text{ eV}^2 / c^4$
- ▶ Want to set a 95% c.l. limit on mass squared – go up from measured value by 1.64σ
 - Limit is 0.01?!
- ▶ If I has measured -0.22, limit would be 0.11 and with -0.39 limit would be negative
- ▶ Such a low limit is not strictly wrong, but it is certainly dishonest!
- ▶ Way out of dilemma is to use Bayesian approach.
- ▶ Have measurement, x , with a Gaussian error and a true value X . Invoke:
$$p(X|x) = \frac{p(x|X)p(X)}{p(x)}$$

Confidence Levels with Constraints

- ▶ $p(x|X)$ is prob. dist. for getting a result x given a true value X and can be calculated
- ▶ $p(x)$ is probability of result and can be absorbed into normalisation
- ▶ $p(X)$ usually assumed to be flat, hence:

$$p(X|x) = \frac{e^{-(x-X)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$

- ▶ Know X cannot be < 0

- ▶ $p(X)$ is therefore a step function and we have:

$$p(X|x) = \frac{e^{-(x-X)^2/2\sigma^2}}{\int_0^{\infty} e^{-(x-X')^2/2\sigma^2} dX'}$$

- ▶ With this equation can calculate c.l.
- ▶ For case above find that Prob. of exceeding 1.6σ is 0.055
Prob. of exceeding 2.78σ is 0.00274 (5% of 0.055).
95% c.l. upper limit is $-0.32 + 2.78 \times 0.2 = 0.24$

Confidence Levels with Constraints

- ▶ This is an honest upper limit.
- ▶ If another variables used, X^2 or $1/X$ upper limit would be different, as assumption of ignorance means different things for different variables
- ▶ A flat distribution in X is no longer flat as a function of X^2
- ▶ Confidence levels may appear to be simple, but in practice you have to be very careful and sometimes only sensible way to proceed is to use Bayesian statistics

Student's t Distribution

- ▶ Up to now assumed that we know the resolution of our measurements
- ▶ What if make several measurements (e.g. of mass) and use them to estimate resolution?
- ▶ Have to use measured mean as true mean not known. Unbiased estimator is:

$$\hat{\sigma} = s = \sqrt{\frac{N}{N-1} \overline{(x - \bar{x})^2}}$$

- ▶ $(x - \mu)/\sigma$ is distributed according to unit Gaussian

$$t = \frac{x - \mu}{\hat{\sigma}}$$

- ▶ Different distribution, due to uncertainty on σ
- ▶ Distribution of t follows Student's t distribution:

$$f(t; n) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi} \Gamma(n/2)} \frac{1}{(1 + (t^2/n))^{(n+1)/2}}$$

- Gaussian-like for large N , larger tails for small N
- n is n.d.f. N if μ is known, $(N-1)$ if unbiased estimator used

Hypothesis Testing

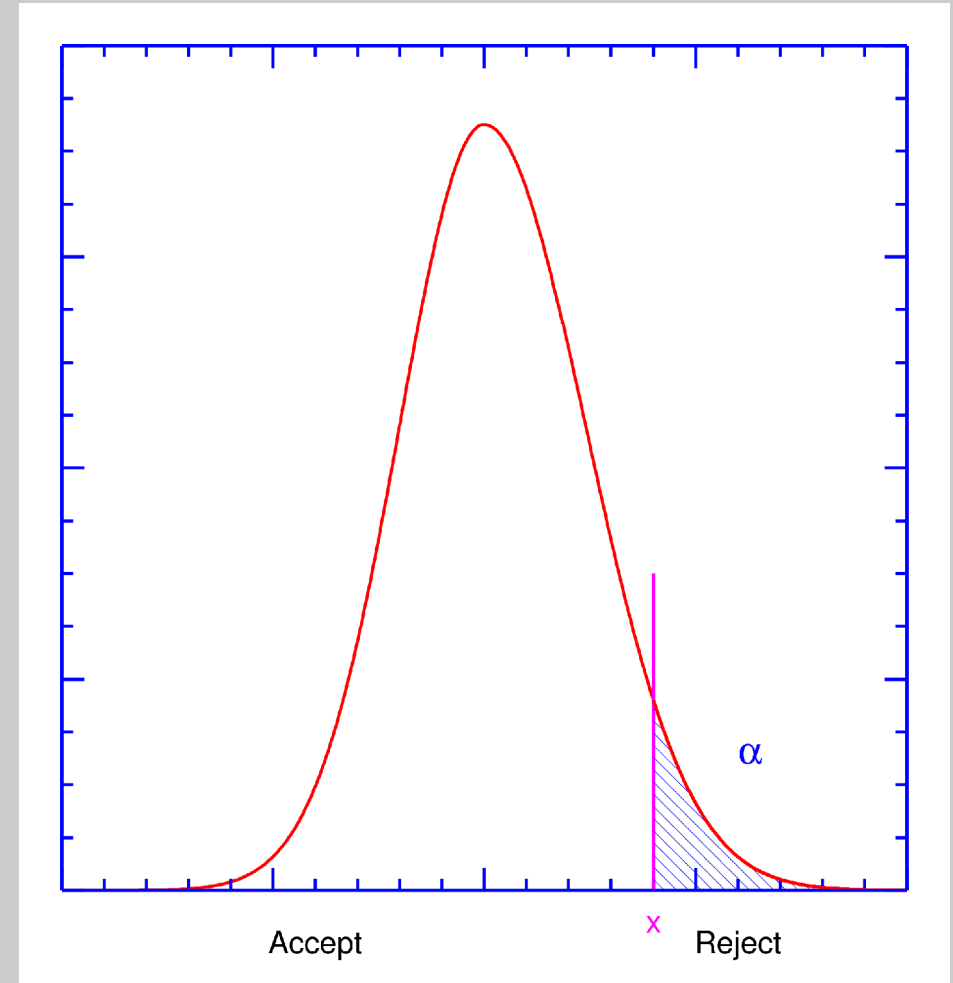
- ▶ Want to know if a certain hypothesis or theory agrees or disagrees with measurement.
- ▶ Need to pose question (*hypothesis*) and then *test* whether hypothesis is *accepted* or *rejected*
- ▶ Note that most of the time this can only be done with a certain level of confidence
- ▶ Often necessary to consider *alternative* hypothesis

Hypothesis Testing

- ▶ Hypotheses can be simple:
 - Data drawn from a Poisson with mean 4.5
- ▶ Composite:
 - Data drawn from Poisson with mean >4.5
 - Data drawn from Poisson with mean that has to be estimated from data
- ▶ Sometimes make a wrong decision:
 - **Type I Error:** You reject a true hypothesis
 - **Type II Error:** You accept a false hypothesis

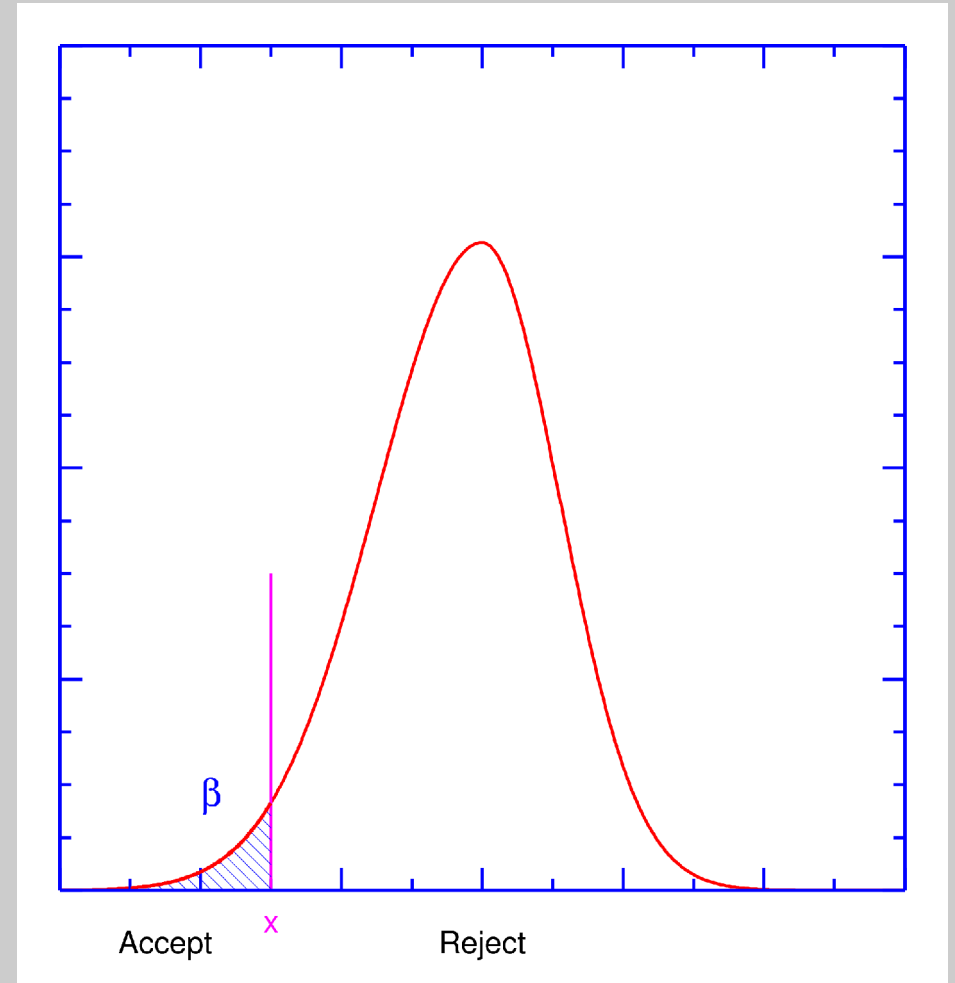
Significance

- ▶ How often a Type I error happens is called significance of the test
- ▶ Evaluate some quantity x to test hypothesis that prob. dist. has some form $P_H(x)$.
- ▶ Divide range of x into a region we accept and one we reject. Acceptance region $P_H(x)$ is large
- ▶ Probability to reject true hypothesis determined by integrating over rejection region



Power

- ▶ If alternative hypothesis is a simple hypothesis (no free parameters), its prob. dist. $P_A(x)$ is known
- ▶ Want to know prob. we will accept this false hypothesis
- ▶ Evaluate $\beta = \int_A P_A(x) dx$
integrate over everywhere outside rejection region
- ▶ Power of test defined as $(1 - \beta)$



Tests

- ▶ Ideally want test for which both α and β small
- ▶ Possible if prob. dist. very different
 - Often hard to achieve
 - Have to decide which error you are most happy with
- ▶ Beware of test with small α
 - 20 people make a measurement wanting to test if a hypothesis is true
 - One of them likely to reject hypothesis at 5% level
 - OK, except he publishes and other 19 don't bother!

Tests

- ▶ Test for which α and β small as possible called Neyman-Pearson test. Only possible if hypothesis and alternative are simple
- ▶ Want to make β small for given α
- ▶ Large $\int_A P_A(x) dx = 1 - \beta$
for given $\int_R P_H(x) dx = \alpha$
- ▶ Test sounds nice, but rarely used as alternative hypothesis is usually not simple

Null Hypothesis

- ◆ Sounds negative, but probably most important and most used hypothesis
- ◆ What does it mean when you say your data are consistent with a new effect being there: global warming, new elementary particle, ...
- ◆ Meaningless statement, unless you can rule out *all* other alternative hypotheses with a certain c.l.
- ◆ As so often in statistics you have to argue backwards!
- ◆ Make *Null Hypothesis* that there is no effect and see if your data can rule out this hypothesis
- ◆ If null hypothesis succeeds you have not shown it is right, you can just say that if some new effect is there, it is at a level too small to be observed by your experiment and you can set some c.l. on the size of it

Looking for New Physics

- ▶ How to look for new physics?
- ▶ Often histogram many quantities and look for a peak
- ▶ Statistical fluctuations mean if you look at enough histograms you will see a peak somewhere
 - Is it significant?
 - Is it consistent with null hypothesis?
- ▶ Simple example - radioisotope decays.
- ▶ See 87 events, expect 54 background
- ▶ Fluctuations $\sqrt{54} = 7.35$
- ▶ Difference $(87-54)/7.35 = 4.5\sigma$ probably have a signal
- ▶ Method OK, if you know expected energy.
- ▶ If not? Resolution of 3 keV and look for peak in 1 MeV range.
- ▶ Discovery c.l. lower than highest c.l for null hypothesis by:
$$C.L._{disc} \approx 1 - 1000/3(1 - C.L._{max})$$

(Some of the) Topics Left Out

- ◆ Separating signal and background
- ◆ Goodness of fit – Run test, Kolmogorov test
- ◆ Fraction fitting
- ◆ Linear least squares
- ◆ BLUE (Best Linear Unbiased Estimator)
Useful when combining different measurements with correlations
- ◆ Profile likelihood
- ◆ **MINUIT**
- ◆ Comparing samples
- ◆ Smoothing and spline fitting
- ◆ ...

Conclusions

- ▶ At all stages of data analysis think about what you are doing!
- ▶ Convince yourself that the statistical methods you are using are appropriate
- ▶ Likelihood is very powerful, but don't forget robustness
- ▶ Fitting:
 - Don't blindly use least squares or likelihood option
 - Try to avoid correlated parameters
 - If possible fit for the parameter you want m^2 and not m as c.l. limits are not the same
- ▶ What level should such lectures start at? Can one assume binomial, Poisson, etc. are known?