Statistics for HEP

Roger Barlow
Manchester University

Lecture 1: Probability

Definition 1: Mathematical

P(A) is a number obeying the Kolmogorov axioms

$$P(A) \geq 0$$

$$P(A_1 \vee A_2) = P(A_1) + P(A_2)$$

$$\sum P(A_i) = 1$$

Problem with Mathematical definition

No information is conveyed by P(A)

Definition 2: Classical

- The probability P(A) is a property of an object that determines how often event A happens.
- It is given by symmetry for equally-likely outcomes
- Outcomes not equally-likely are reduced to equally-likely ones
- Examples:
- Tossing a coin:

  P(H)=1/2
- Throwing two dice

  P(8)=5/36

Problems with the classical definition…

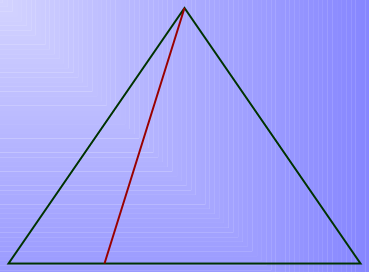
1. When are cases ‘equally likely’?

- If you toss two coins, are there 3 possible outcomes or 4?

Can be handled

2. How do you handle continuous variables?

- Split the triangle at random:



Cannot be handled

Bertrand's Paradox

A jug contains 1 glassful of water and between 1 and 2 glasses of wine

Q: What is the most probable wine:water ratio?

A: Between 1 and 2 $\rightarrow$3/2

Q: What is the most probable water:wine ratio?

A: Between 1/1 and 1/2 $\rightarrow$3/4

$$(3/2) \neq (3/4)^{-1}$$

Definition 3: Frequentist

- The probability P(A) is the limit (taken over some ensemble)

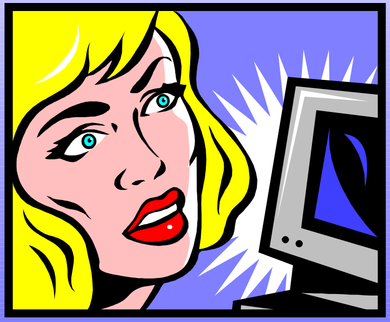
$$P(A) \underset{N\to\infty}{=} {N(A)}\big/{N}$$

Problem (limitation) for the Frequentist definition

P(A) depends on A and the ensemble

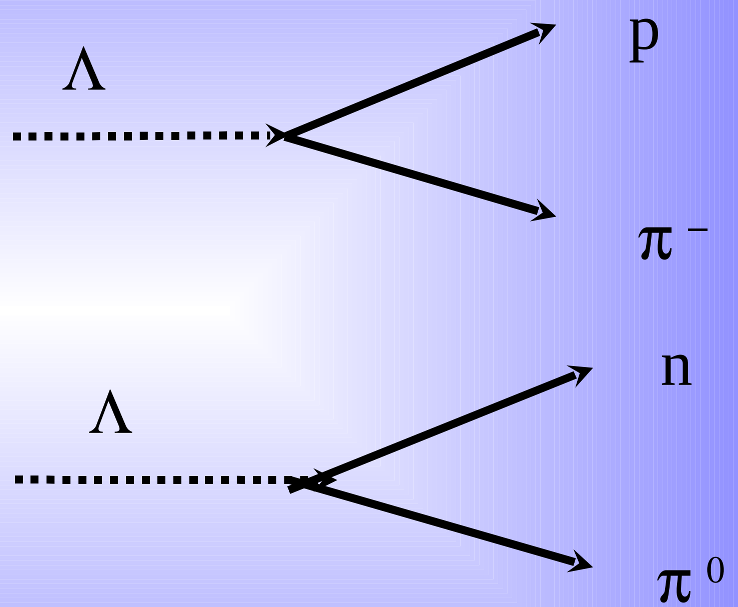
Eg: count 10 of a group of 30 with beards.

P(beard)=1/3



Aside: Consequences for Quantum Mechanics

- QM calculates probabilities
- Probabilities are not ‘real’ – they depend on the process and the ensemble



PDG:

$P(p\pi^-)=0.639$,

$P(n\pi^0)=0.358$

Big problem for the Frequentist definition

Cannot be applied to unique events

~~‘It will probably rain tomorrow’~~

Is unscientific

`The statement

“It will rain tomorrow”

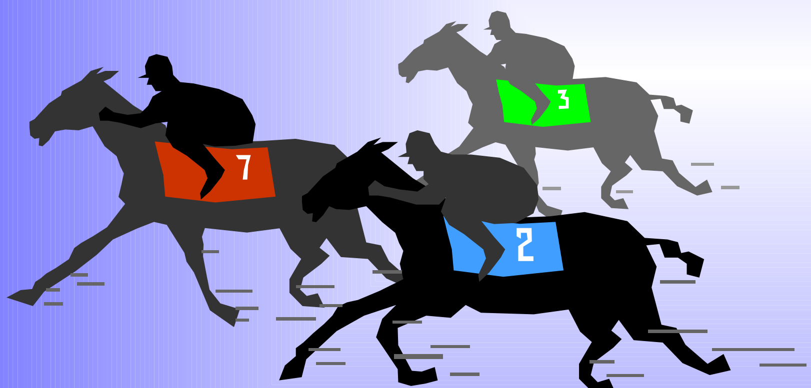
is probably true.’

Is quite OK

But that doesn’t always work

- Rain prediction in unfamiliar territory
- Euler’s theorem
- Higgs discovery
- Dark matter
- LHC completion

Definition 4: Subjective (Bayesian)



P(A) is your degree of belief in A;

You will accept a bet on A if the odds are better than

1-P to P

A can be Anything :

Beards, Rain, particle decays, conjectures, theories

Bayes Theorem

Often used for subjective probability

Conditional Probability P(A|B)

P(A & B)= P(B) P(A|B)

P(A & B)= P(A) P(B|A)

$$P(A|B) = \frac{P(B|A)}{P(B)} P(A)$$

Example:

W=white jacket B=bald

P(W&B)=(2/4)x(1/2) or (1/4)x(1/1)

P(W|B)

= $\frac{1}{(2/4)}$ x (1/4) =1/2

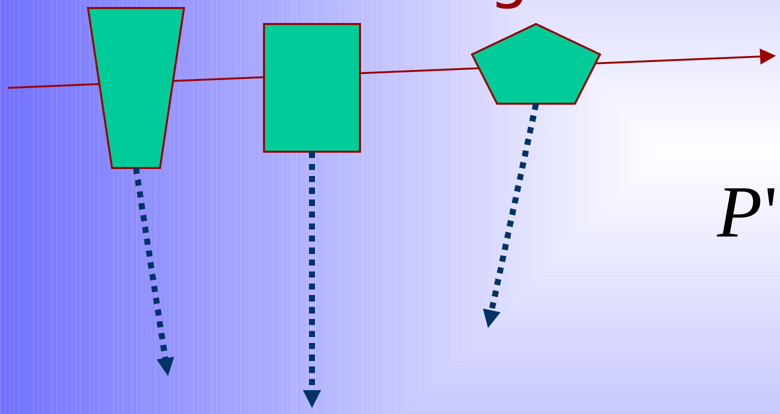
Start digression

Frequentist Use of Bayes Theorem

Example: Particle Identification

Particle types e,π ,μ,K,p

Detector Signals: DCH,RICH,TOF,TRD



$$P'(e) = P(e \mid DCH) = \frac{P(DCH \mid e)}{P(DCH)} P(e)$$

$$P(DCH) = P(DCH \mid e)P(e) + P(DCH \mid \mu)P(\mu) + P(DCH \mid \pi)P(\pi)\ldots$$

Then repeat for P(e|RICH) using P’(e) etc

Warning Notice

To determine P'(e) need P(e), P(μ) etc

('*a priori* probabilities')

If/when you cut on Probability, the Purity depends on these *a priori* probabilities

Example: muon detectors.

P(track| μ)$\approx$ 0.9 P(track|π) $\approx$ 0.015

But P(μ)$\approx$ 0.01 P(π) $\approx$ 1

Quantities like $$\frac{P(data\,|\,\mu)}{P(data\,|\,e)+P(data\,|\,\mu)+P(data\,|\,\pi)+P(data\,|\,K)}$$

Have no direct meaning –

use with care!

End digression

Bayes' Theorem and subjective probability

$$P(Theory|Result) = \frac{P(Result|Theory)}{P(Result)} P(Theory)$$

Your (posterior) belief in a Theory is modified by experimental result

If P(Result|Theory)=0 belief is killed

Large P(Result|Theory) increases belief, modified by general P(Result)

Applies to successive results

Dependence on prior P(Theory) eventually goes

Problem with subjective probability

It is subjective

My P(A) and your P(A) may be different

Scientists are supposed to be objective

Reasons to use subjective probability:

- Desperation ✓
- Ignorance ✗
- Idleness ?

Can Honest Ignorance justify P(A)=flat?

Argument:

- you know nothing
- every value is as believable as any other
- all possibilities equal

How do you count discrete possibilities?

SM true or false?

SM or SUSY or light Higgs or Technicolor?

For continuous parameter

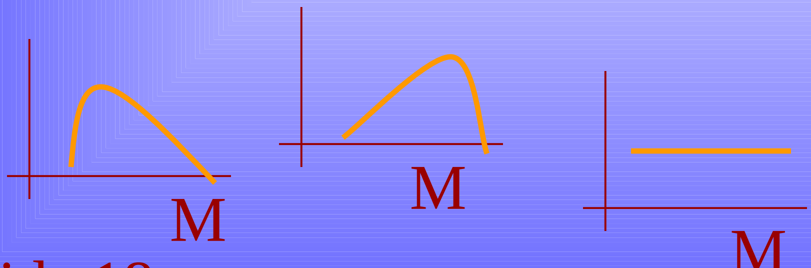
(e g. M_{higgs})

Take. $P(M_{higgs})$ as flat

Actually has to be zero as $\int P(M)dM=1$ but never mind… 'improper prior'

Working with √M or lnM will give different results

Real Statisticians accept this and test for



‘Objective Prior’ (Jeffreys)

Transform to a variable q(M) for which the Fisher information is constant

$$I(q) = -\left\langle \frac{\partial^2 \ln P(x;q)}{\partial q^2} \right\rangle = const$$

For a location parameter with P(x;M)=f(x+M) use M

For scale parameter with P(x;M)=Mf(x) use ln M

For a Poisson λ use prior 1/√λ

For a Binomial with probability p use prior 1/√p(1-p)

This has never really caught on

Conclusion
What is Probability?

4 ways to define it

- Mathematical
- Classical
- Frequentist
- Subjective

Each has strong points and weak points

None is universally applicable

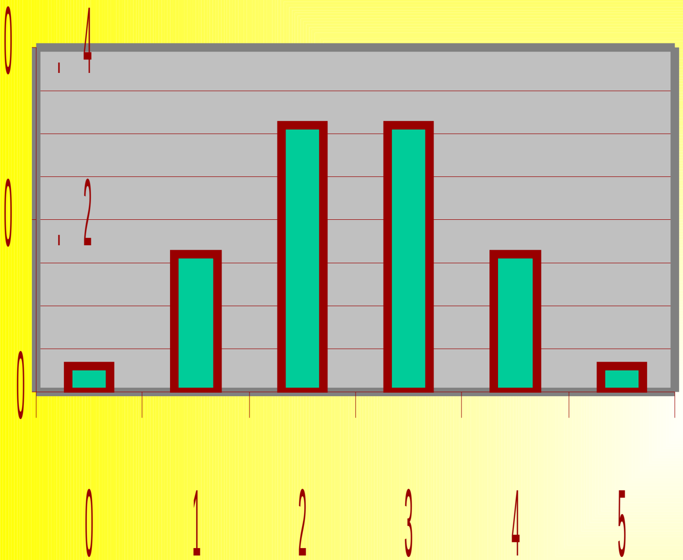
Be prepared to understand and use them all -

Statistics for HEP

Roger Barlow
Manchester University

Lecture 2: Distributions

The Binomial



n trials r successes

Individual success probability p

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

Mean

$\mu = \langle r \rangle = \Sigma r P(r)$

$= np$

Variance

$V \equiv \sigma^2 = \langle (r-\mu)^2 \rangle = \langle r^2 \rangle - \langle r \rangle^2$

$= np(1-p)$

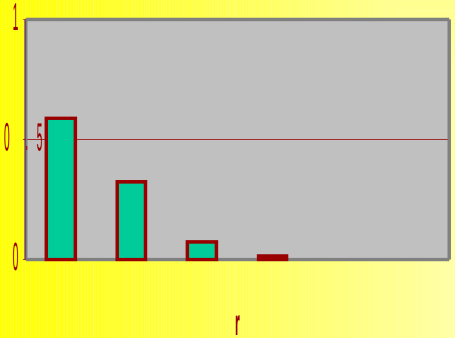
Met with in Efficiency/Acceptance calculations

$1-p \equiv \bar{p} \equiv q$

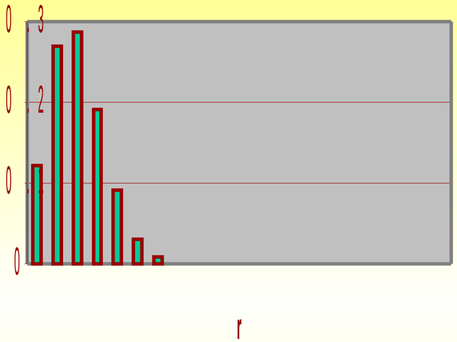
Binomial Examples

p=0.1

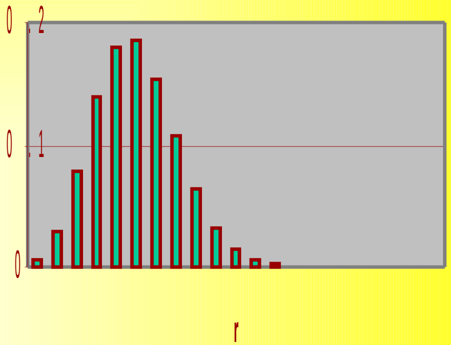
n=5



n=20

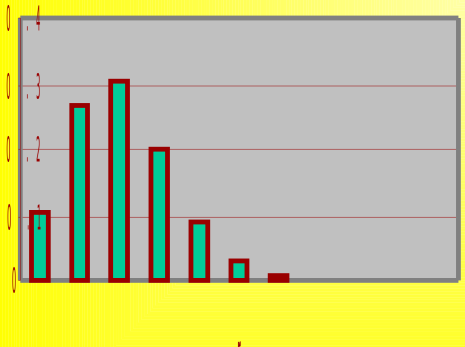


n=50

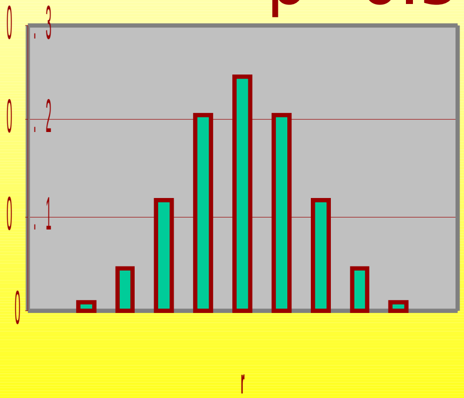


n=10

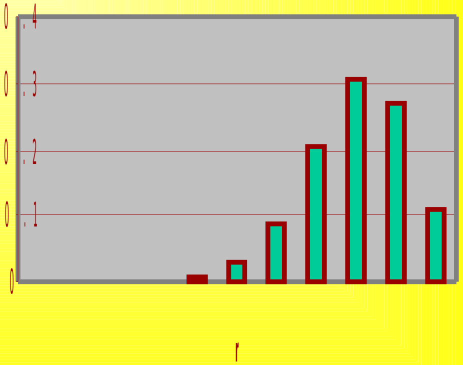
p=0.2



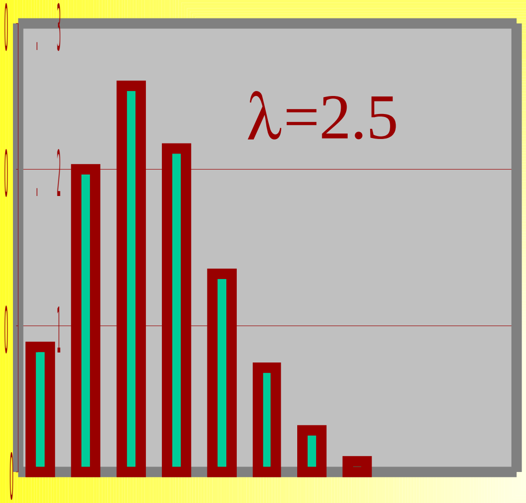
p=0.5



p=0.8



Poisson



'Events in a continuum'

e.g. Geiger Counter clicks

Mean rate λ in time interval

Gives number of events in data

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

Mean

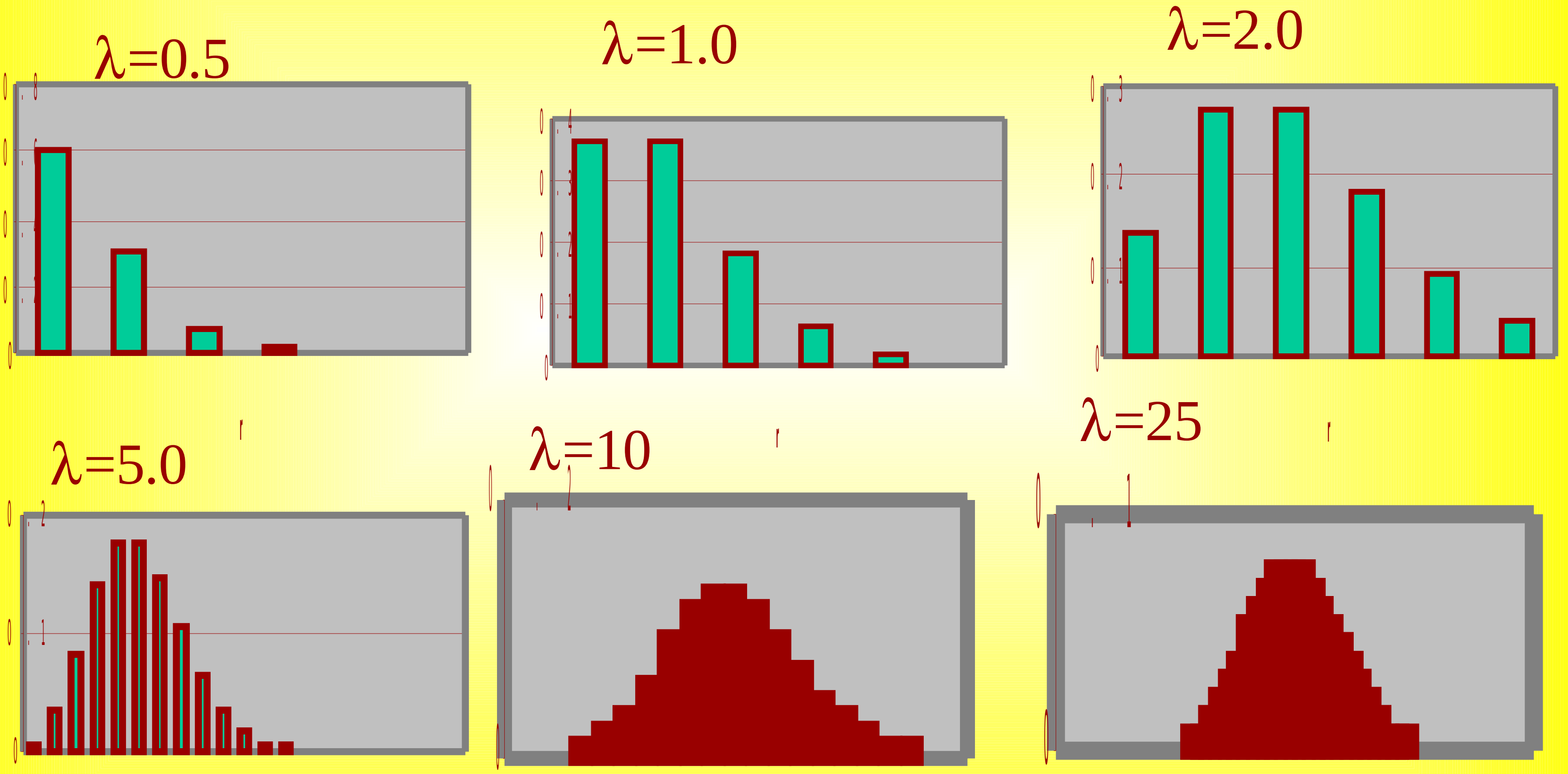
$\mu = \langle r \rangle = \Sigma r P(\ r\ )$

$= \lambda$

Variance

$V \equiv \sigma^2 = \langle (r-\mu\ )^2 \rangle = \langle r^2 \rangle - \langle r \rangle^2$

Poisson Examples



Binomial and Poisson

From an exam paper

A student is standing by the road, hoping to hitch a lift. Cars pass according to a Poisson distribution with a mean frequency of 1 per minute. The probability of an individual car giving a lift is 1%. Calculate the probability that the student is still waiting for a lift

(a) After 60 cars have passed

(b) After 1 hour

a) $0.99^{60}=0.5472$

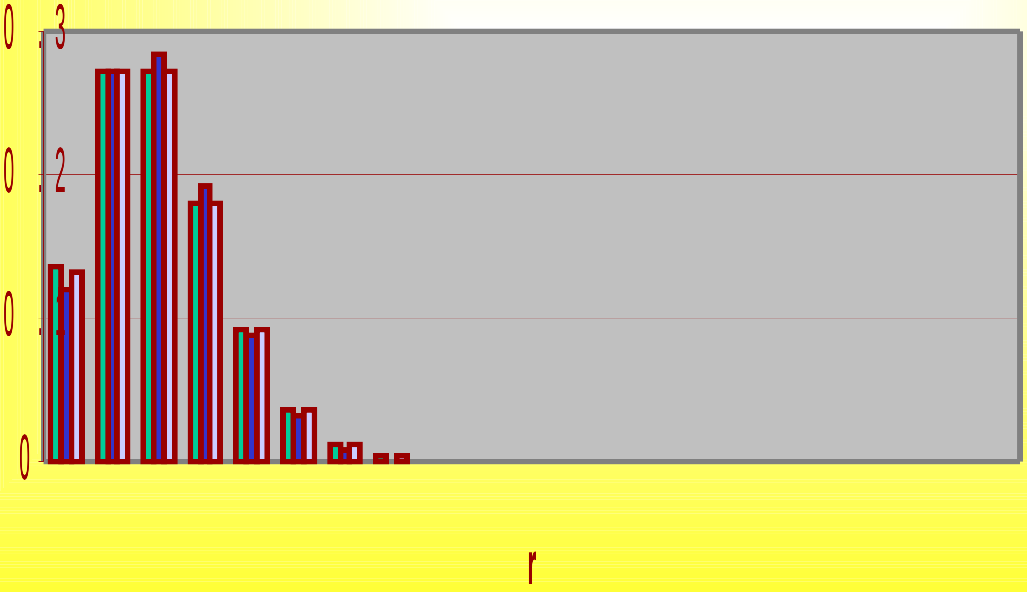
b) $e^{-0.6}=0.5488$

Poisson as approximate binomial

Poisson mean 2

Binomial: p=0.1, 20 tries
Binomial: p=0.01, 200 tries

Use: MC simulation (Binomial) of Real data (Poisson)



Two Poissons

2 Poisson sources, means λ_1 and λ_2

Combine samples

e.g. leptonic and hadronic decays of W

Forward and backward muon pairs

Tracks that trigger and tracks that don't

What you get is a *Convolution*

$$P(r)=\sum P(r'; \lambda_1)\, P(r-r'; \lambda_2)$$

Turns out this is also a Poisson with mean $\lambda_1+\lambda_2$

Avoids lots of worry

The Gaussian

Probability Density

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

Mean

$\mu = \langle x \rangle = \int xP(x)\,dx$

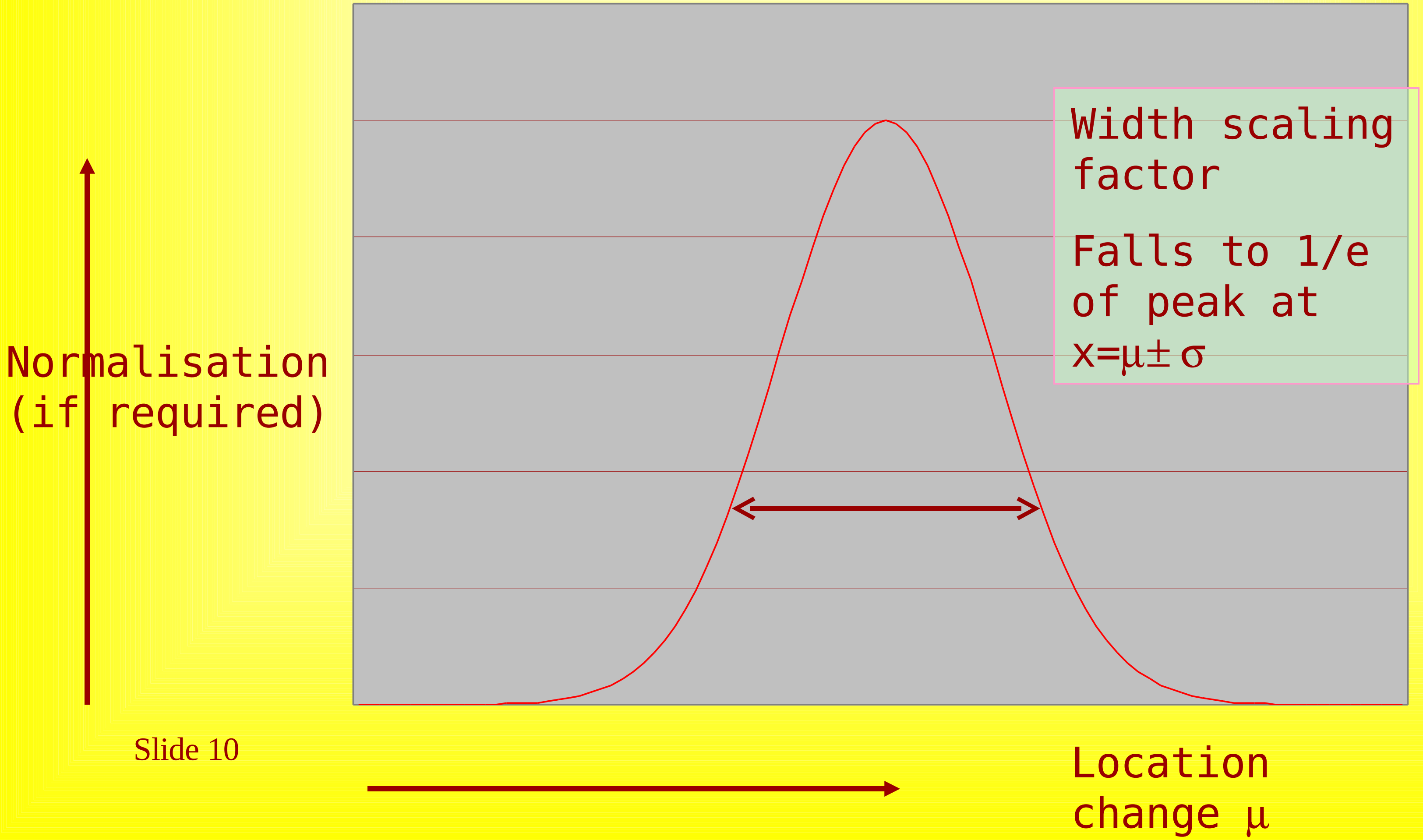
$= \mu$

Variance

$V \equiv \sigma^2 = \langle (x-\mu)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2$

Different Gaussians

There's only one!



Probability Contents

68.27% within 1σ

95.45% within 2σ

99.73% within 3σ

90% within 1.645 σ

95% within 1.960 σ

99% within 2.576 σ

99.9% within 3.290σ

These numbers apply to Gaussians and only Gaussians

Other distributions have equivalent values which you could use of you wanted

Central Limit Theorem

Or: why is the Gaussian Normal?

If a Variable x is produced by the convolution of variables $x_1, x_2 \ldots x_N$

I) $\langle x \rangle = \mu_1 + \mu_2 + \ldots \mu_N$

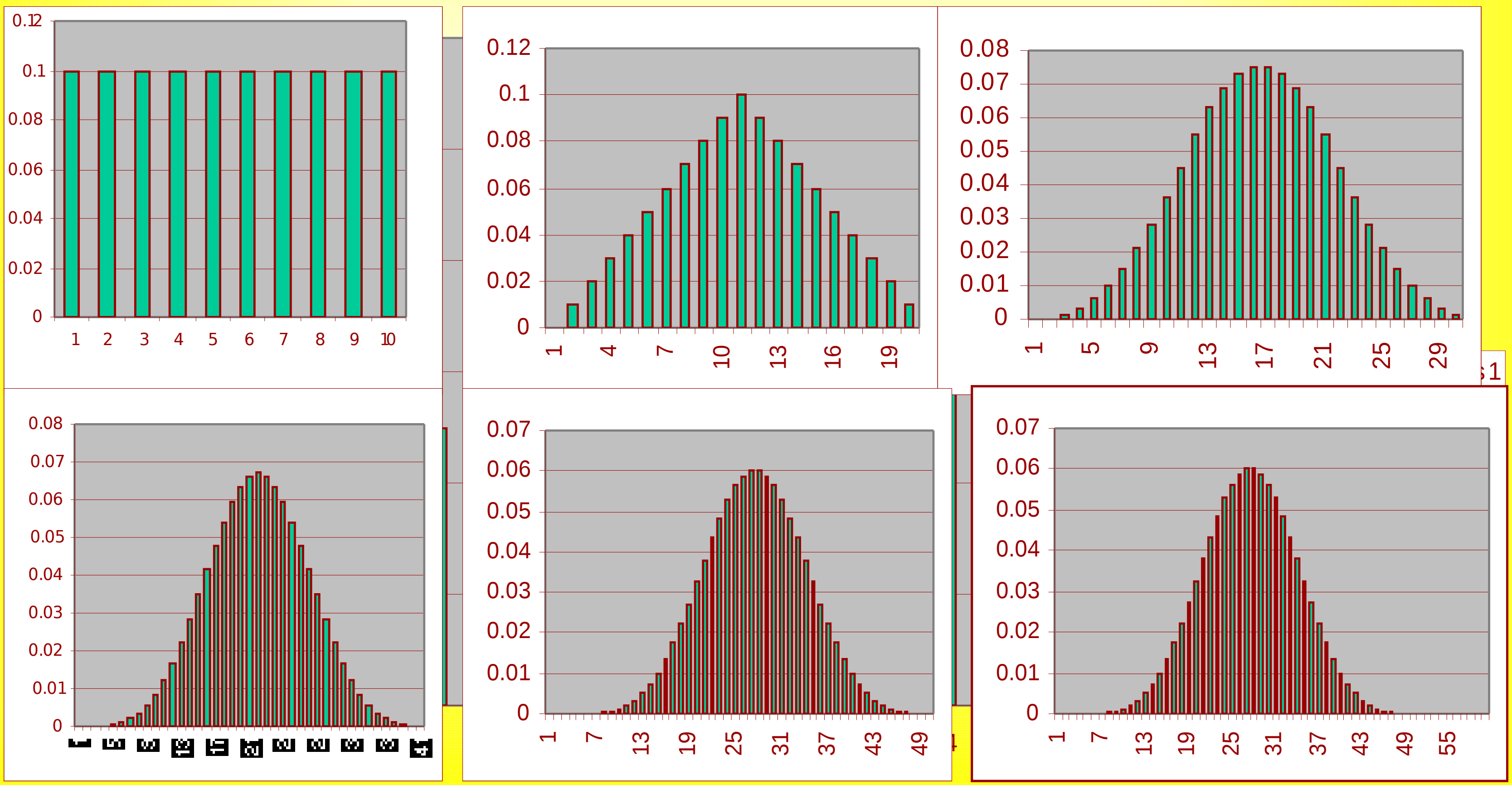
II) $V(x) = V_1 + V_2 + \ldots V_N$

III) P(x) becomes Gaussian for large N

There were hints in the Binomial and Poisson examples

CLT demonstration

Convolute Uniform distribution with itself



CLT Proof (I) Characteristic functions

Given P(x) consider

$$\left\langle e^{ikx} \right\rangle = \int e^{ikx} P(x) dx = \tilde{P}(k) = \Phi(k)$$

The Characteristic Function

For convolutions, CFs multiply

If $f(x) = g(x) \otimes h(x)$ then $\tilde{f}(k) = \tilde{g}(k)\tilde{h}(k)$

Logs of CFs Add

CLT proof (2) Cumulants

CF is a power series in k

$$\langle 1\rangle + \langle ikx\rangle + \langle (ikx)^2/2!\rangle + \langle (ikx)^3/3!\rangle + \ldots$$

$$1 + ik\langle x\rangle - k^2\langle x^2\rangle/2! - ik^3\langle x^3\rangle/3! + \ldots$$

Ln CF can then be expanded as a series

$$ikK_1 \; + \; (ik)^2K_2/2! \; + \; (ik)^3K_3/3! \ldots$$

K_r : the "semi-invariant cumulants of Thiele"

Total power of x^r

If $x \rightarrow x+a$ then only $K_1 \rightarrow K_1+a$

If $x \rightarrow bx$ then each $K_r \rightarrow b^r K_r$

CLT proof (3)

- The FT of a Gaussian is a Gaussian

$$e^{-x^2/2\sigma^2} \to e^{-k^2\sigma^2/2}$$

Taking logs gives power series up to k^2

$K_r=0$ for $r>2$ defines a Gaussian

- Selfconvolute anything n times: $K_r'=n\,K_r$

Need to normalise – divide by n

$K_r''=n^{-r}\,K_r'= n^{1-r}\,K_r$

- Higher Cumulants die away faster

If the distributions are not identical but similar the same argument applies

CLT in real life

Examples

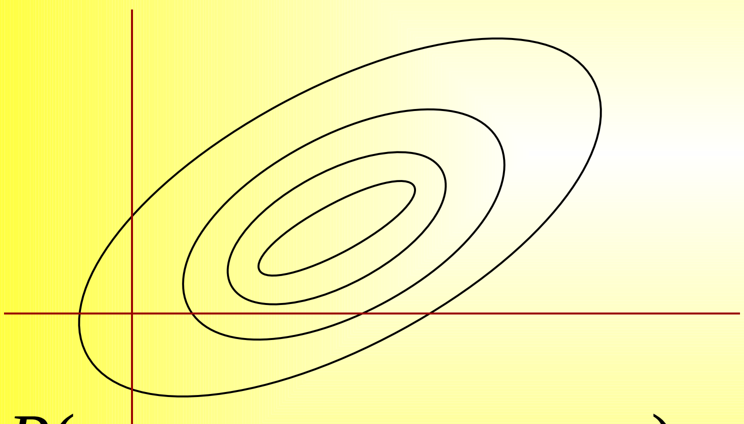
- Height
- Simple Measurements
- Student final marks

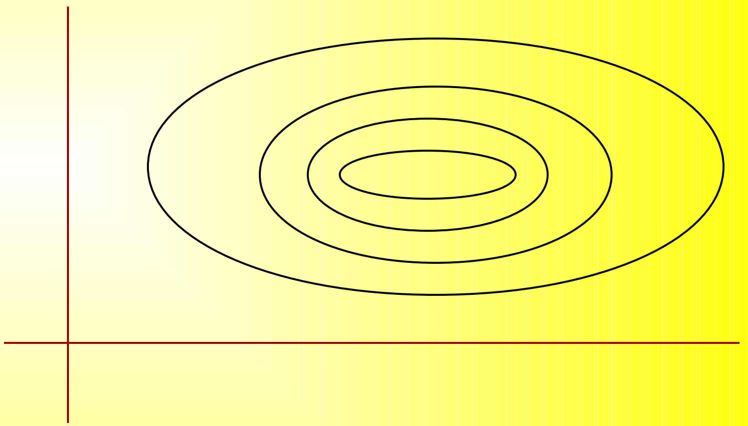
Counterexamples

- Weight
- Wealth
- Student entry grades

Multidimensional Gaussian

$$P(x, y; \mu_x, \mu_y, \sigma_x, \sigma_y) = \frac{1}{\sigma_x \sigma_y 2\pi} e^{-(x-\mu_x)^2/2\sigma_x{}^2} e^{-(y-\mu_y)^2/2\sigma_y{}^2}$$





$$P(x, y; \mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$$

$$= \frac{1}{\sigma_x \sigma_y 2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left((x-\mu_x)^2/\sigma_x^2 + (y-\mu_y)^2/\sigma_y^2 - 2\rho(x-\mu_x)(y-\mu_y)/\sigma_x\sigma_y\right)}$$

Chi squared

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2$$

Sum of squared discrepancies, scaled by expected error

Integrate all but 1-D of multi-D Gaussian

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

Mean n

Variance 2n

CLT slow to operate

Generating Distributions

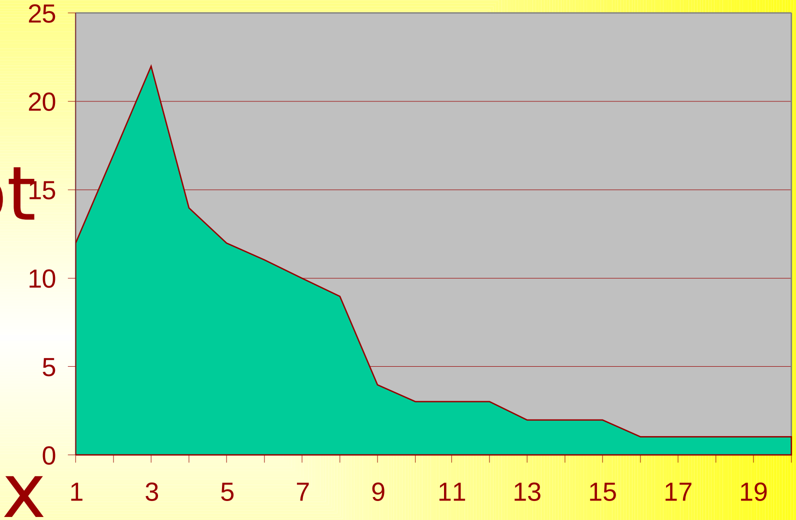
Given `int rand()` in `stdlib.h`

```
float Random()
                {return ((float)rand())/RAND_MAX;}
float uniform(float lo, float hi)
                     {return lo+Random()*(hi-lo);}
float life(float tau)
                      {return -tau*log(Random());}
float ugauss()        // really crude. Do not use
   {float x=0; for(int i=0;i<12;i++) x+=Random();
                                    return x-6;}
float gauss(float mu, float sigma)
                       {return mu+sigma*ugauss();}
```

A better Gaussian Generator

```
float ugauss(){
    static bool igot=false;
    static float got;
    if(igot){igot=false; return got;}
    float phi=uniform(0.0F,2*M_PI);
    float r=life(1.0f);
    igot=true;
    got=r*cos(phi);
    return r*sin(phi);}
```

More complicated functions



Find P_0=max[P(x)].

Overestimate if in doubt

Repeat :

Repeat:

Generate random x

Find P(x)

Generate random P in range 0-P_0

till P(x)>P

till you have enough data

If necessary make x non-random and compensate

Other distributions

Uniform(top hat)

σ=width/√12

Breit Wigner (Cauchy)

Has no variance – useful for wide tails

Landau

Has no variance or mean

Not given by $e^{e^{-\lambda}}$ .Use CERNLIB

Functions you need to know

- Gaussian/Chi squared
- Poisson
- Binomial
- Everything else

Statistics for HEP

Roger Barlow
Manchester University

Lecture 3: Estimation

About Estimation

Theory

Probability

Calculus

Data

Given these distribution parameters, what can we say about the data?

Given this data, what can we say about the properties or parameters or correctness of the distribution functions?

Statistical

Inference

Data

Theory

What is an estimator?

$$\hat{\mu}(\{x\}) = \frac{1}{N}\sum_i x_i$$

$$\hat{\mu}(\{x\}) = \frac{x_{\max} + x_{\min}}{2}$$

$$\hat{V}(\{x\}) = \frac{1}{N}\sum_i (x_i - \hat{\mu})^2$$

$$\hat{V}(\{x\}) = \frac{1}{N-1}\sum_i (x_i - \hat{\mu})^2$$

An estimator is a procedure giving a value for a parameter or property of the distribution as a function of the actual data values

What is a good estimator?

One often has to work with less-than-perfect estimators

A perfect estimator is:

- Consistent $\underset{N\to\infty}{Limit}\left(\hat{a}\right)=a$
- Unbiassed

$$\langle\hat{a}\rangle=\int\int...\int\hat{a}(x_1,x_2,...)P(x_1;a)P(x_2;a)P(x_3;a)...dx_1dx_2...=a$$

- Efficient

$V(\hat{a})=\left\langle\left(\hat{a}-\langle\hat{a}\rangle\right)^2\right\rangle$ minimum

Minimum Variance Bound

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

The Likelihood Function

Set of data $\{x_1, x_2, x_3, \ldots x_N\}$

Each x may be multidimensional – never mind

Probability depends on some parameter a

a may be multidimensional – never mind

Total probability (density)

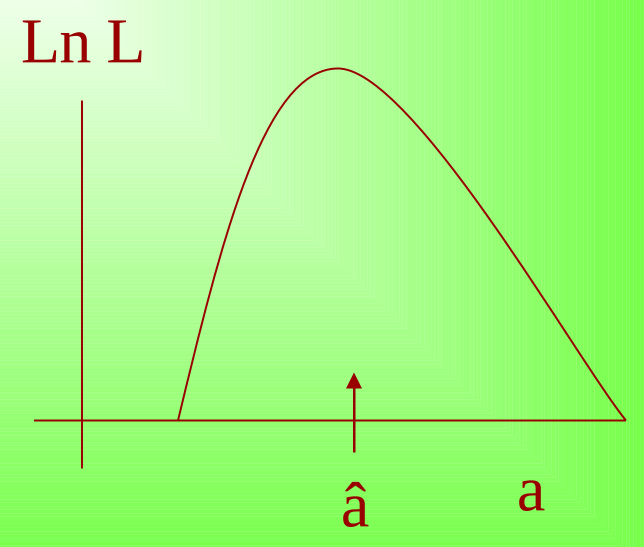
$P(x_1;a)\ P(x_2;a)\ P(x_3;a)\ \ldots P(x_N;a)=L(x_1, x_2, x_3, \ldots x_N\ ;a)$

The Likelihood

Maximum Likelihood Estimation

Given data $\{x_1, x_2, x_3, \ldots x_N\}$ estimate a by maximising the likelihood $L(x_1, x_2, x_3, \ldots x_N ;a)$

$$\left.\frac{dL}{dA}\right|_{a=\hat{a}} = 0$$



In practice usually maximise ln L as it's easier to calculate and handle; just add the ln $P(x_i)$

ML has lots of nice properties

Properties of ML estimation

- It's consistent

(no big deal)

- It's biassed for small N

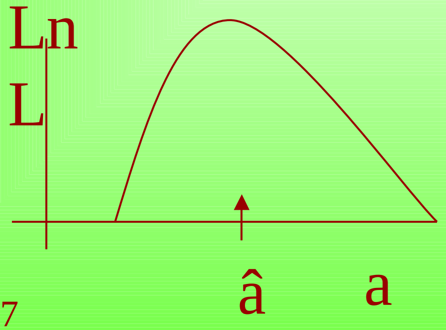
May need to worry

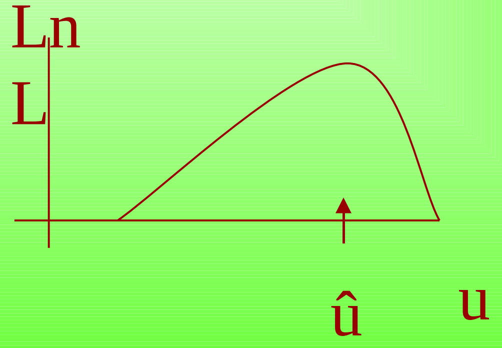
- It is efficient for large N

Saturates the Minimum Variance Bound

- It is invariant

If you switch to using u(a), then û=u(â)



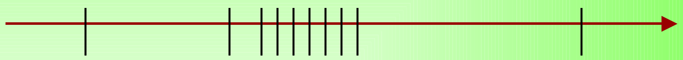


More about ML

- It is not ‘right’. Just sensible.
- It does not give the ‘most likely value of a’. It’s the value of a for which this data is most likely.
- Numerical Methods are often needed
- Maximisation / Minimisation in >1 variable is not easy
- Use MINUIT but remember the minus sign

ML does not give goodness-of-fit

- ML will not complain if your assumed P(x;a) is rubbish
- The value of L tells you nothing

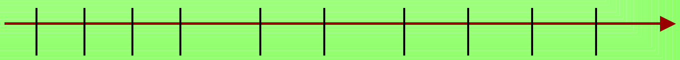


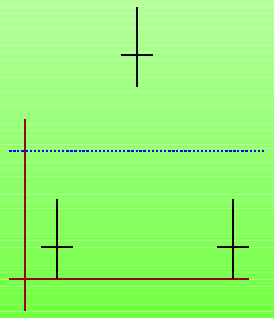
Fit $P(x)=a_1x+a_0$

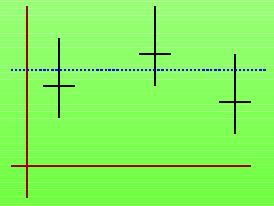
will give $a_1=0$; constant P

$L= a_0^N$

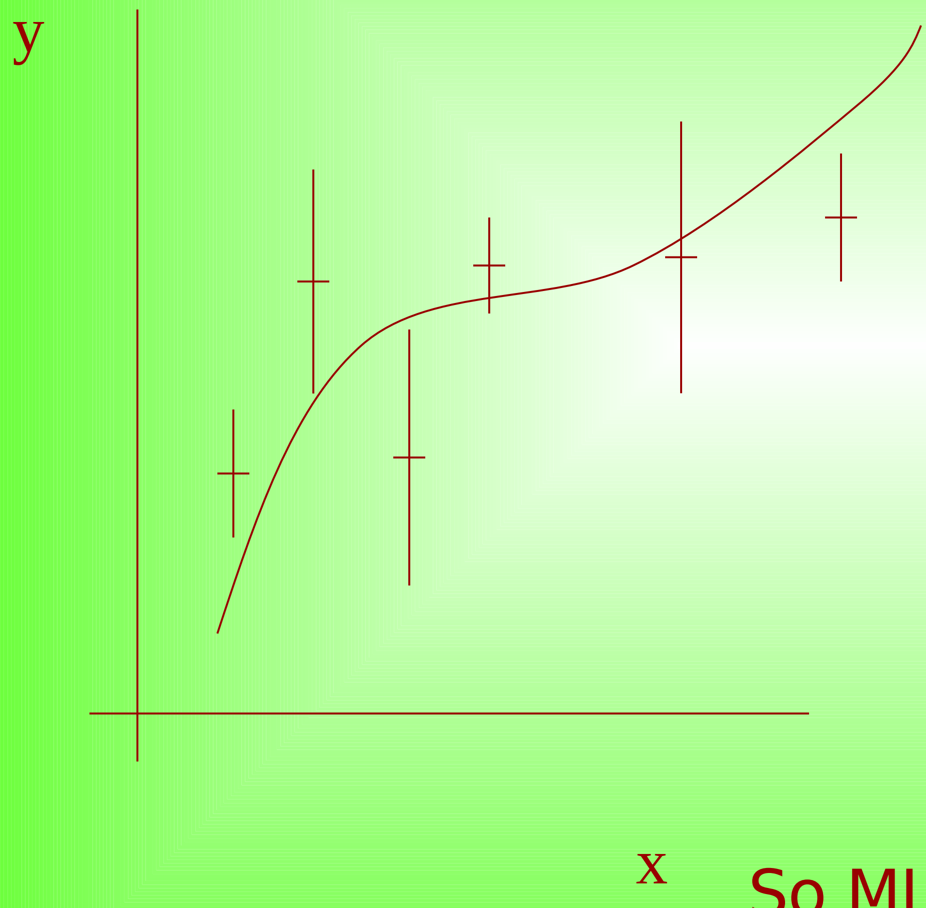
Just like you get from fitting







Least Squares



- Measurements of y at various x with errors σ and prediction f(x;a)
- Probability $\propto e^{-(y-f(x;a))^2/2\sigma^2}$
- Ln L $-\frac{1}{2}\sum_i \left(\frac{y_i - f(x_i;a)}{\sigma_i} \right)^2$
- To maximise ln L, minimise χ^2

So ML ‘proves’ Least Squares. But what ‘proves’ ML? Nothing

Least Squares: The Really nice thing

- Should get $\chi^2 \approx 1$ per data point
- Minimise χ^2 makes it smaller – effect is 1 unit of χ^2 for each variable adjusted. (Dimensionality of MultiD Gaussian decreased by 1.)

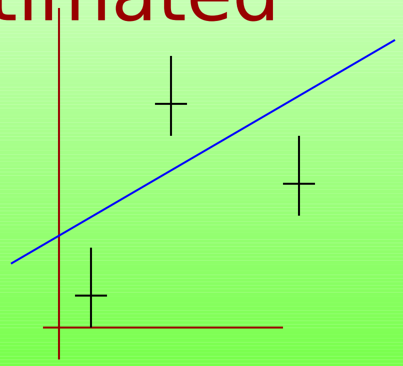
$$N_{\text{degrees Of Freedom}} = N_{\text{data pts}} - N_{\text{parameters}}$$

- Provides 'Goodness of agreement' figure which allows for credibility check

Chi Squared Results

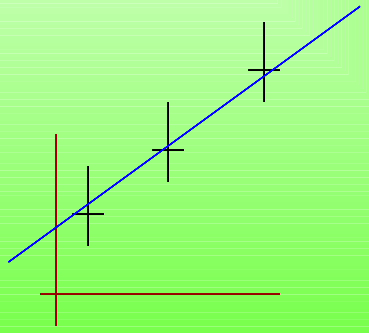
Large χ^2 comes from

1. Bad Measurements
2. Bad Theory
3. Underestimated errors
4. Bad luck



Small χ^2 comes from

1. Overestimated errors
2. Good luck



Fitting Histograms

Often put $\{x_i\}$ into bins

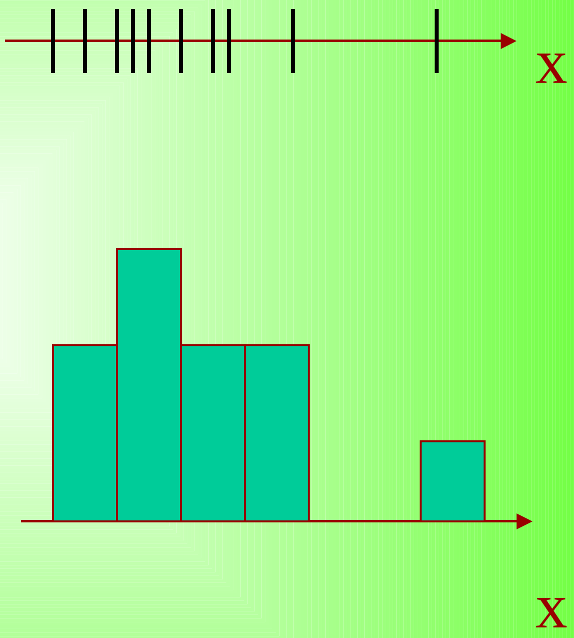
Data is then $\{n_j\}$

n_j given by Poisson,

mean $f(x_j) = P(x_j)\Delta x$

4 Techniques

- Full ML
- Binned ML
- Proper χ^2
- Simple χ^2



What you maximise/minimise

- Full ML $$\ln L = \sum_i \ln P(x_i; a\ )$$
- Binned ML $$\ln L = \sum_j \ln Poisson(n_j; f_j) \approx \sum_j n_j \ln f_j - f_j$$
- Proper χ^2 $$\sum_j \frac{(n_j - f_j)^2}{f_j}$$
- Simple χ^2 $$\sum_j \frac{(n_j - f_j)^2}{n_j}$$

Which to use?

- Full ML: Uses all information but may be cumbersome, and does not give any goodness-of-fit. Use if only a handful of events.
- Binned ML: less cumbersome. Lose information if bin size large. Can use χ^2 as goodness-of-fit afterwards
- Proper χ^2 : even less cumbersome and gives goodness-of-fit directly. Should have n_j large so Poisson$\rightarrow$Gaussian
- Simple χ^2 : minimising becomes linear. Must have n_j large

Consumer tests show

- Binned ML and Unbinned ML give similar results unless binsize > feature size
- Both χ^2 methods get biassed and less efficient if bin contents are small due to asymmetry of Poisson
- Simple χ^2 suffers more as sensitive to fluctuations, and dies when bin contents are zero

Orthogonal Polynomials

Fit a cubic: Standard polynomial

$$f(x)=c_0+ c_1x+ c_2x^2+ c_3x^3$$

Least Squares $[\Sigma(y_i-f(x_i))^2]$ gives

$$\begin{pmatrix} 1 & \overline{x} & \overline{x^2} & \overline{x^3} \\ \overline{x} & \overline{x^2} & \overline{x^3} & \overline{x^4} \\ \overline{x^2} & \overline{x^3} & \overline{x^4} & \overline{x^5} \\ \overline{x^3} & \overline{x^4} & \overline{x^5} & \overline{x^6} \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} \overline{y} \\ \overline{xy} \\ \overline{x^2 y} \\ \overline{x^3 y} \end{pmatrix}$$

Invert and solve? Think first!

Define Orthogonal Polynomial

$P_0(x) = 1$

$P_1(x) = x \;\; + a_{01}P_0(x)$

$P_2(x) = x^2 + a_{12}P_1(x) + a_{02}P_0(x)$

$P_3(x) = x^3 + a_{23}P_2(x) + a_{13}P_1(x) + a_{03}P_0(x)$

Orthogonality: $\Sigma_r P_i(x_r)\, P_j(x_r) = 0$ unless $i=j$

$$a_{ij} = -(\;\; \Sigma_r\, x_r^{\,j}\, P_i\,(x_r))/\; \Sigma_r\, P_i\,(x_r)^2$$

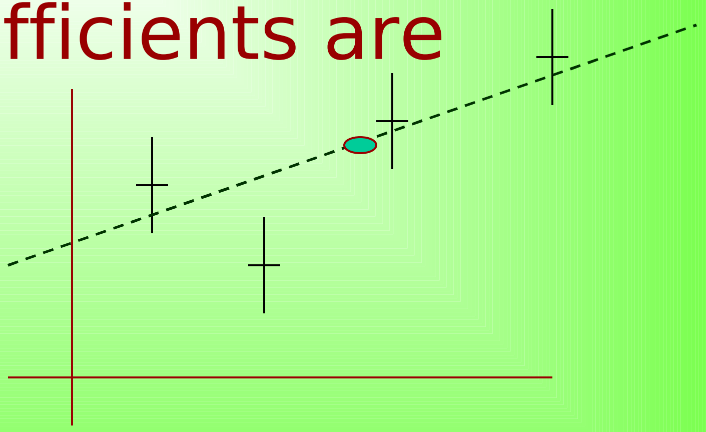
Use Orthogonal Polynomial

$$f(x)=c'_0P_0(x)+ c'_1P_1(x)+ c'_2P_2(x)+ c'_3P_3(x)$$

Least Squares minimisation gives

$$c'_i=\Sigma yP_i / \Sigma P_i^2$$

Special Bonus: These coefficients are UNCORRELATED

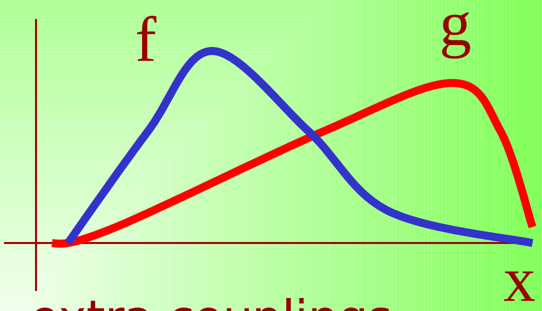


Simple example:

Fit y=mx+c or

$y=m(x - \bar{x})+c'$

Optimal Observables



Function of the form

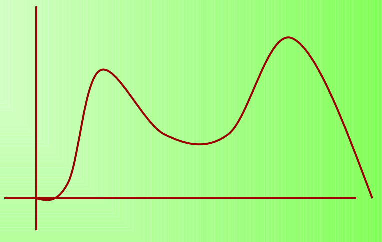
P(x)=f(x)+a g(x)

e.g. signal+background, tau polarisation, extra couplings

A measurement x contains info about a

Depends on f(x)/g(x) ONLY.

Work with O(x)=f(x)/g(x)



Write

$$\overline{O} = \int \frac{f^2}{g} dx + a \int f dx$$

Use

$$\hat{a} = \left(\overline{O} - \int \frac{f^2}{g} dx \right) \Big/ \int f dx$$

Why this is magic

$$\hat{a} = \left(\overline{O} - \int \frac{f^2}{g} dx \right) \Big/ \int f dx$$

It’s efficient. Saturates the MVB. As good as ML

x can be multidimensional. O is one variable.

In practice calibrate $\overline{O}$ and â using Monte Carlo

If a is multidimensional there is an O for each

If the form is quadratic then use of the mean OO is not as good as ML. But close.

Extended Maximum Likelihood

- Allow the normalisation of P(x;a) to float
- Predicts numbers of events as well as their distributions

$$N_{pred} = \int P(x;a)dx$$

- Need to modify L

$$\ln L = \sum_i \ln P(x_i;a) - \int P(x;a)dx$$

- Extra term stops normalistion shooting up to infinity

Using EML

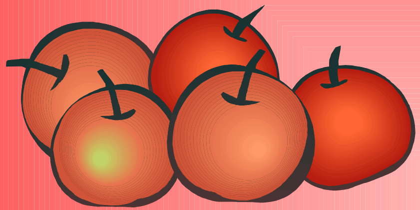
- If the shape and size of P can vary independently, get same answer as ML and predicted N equal to actual N
- If not then the estimates are better using EML
- Be careful of the errors in computing ratios and such

Statistics for HEP

Roger Barlow
Manchester University

Lecture 4: Confidence Intervals

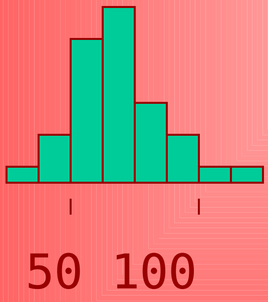
The Straightforward Example



Apples of different weights

Need to describe the distribution

μ = 68g σ = 17 g



All weights between 24 and 167 g (Tolerance)

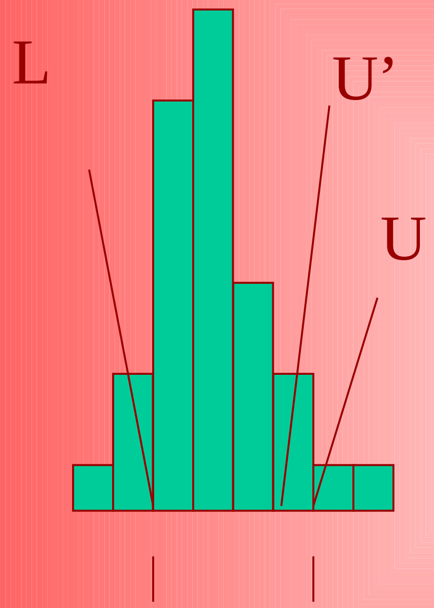
90% lie between 50 and 100 g

94% are less than 100 g

96% are more than 50 g

Confidence level statements

Confidence Levels



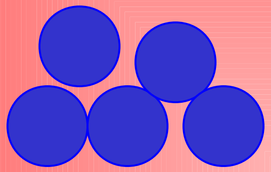
- Can quote at any level

  (68%, 95%, 99%...)
- Upper or lower or twosided

  (x<U x<L L<x<U)
- Two-sided has further choice

  (central, shortest...)

The Frequentist Twist



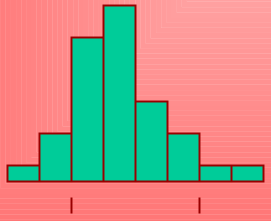
Particles of the same weight

Distribution spread by measurement errors

What can we say about M?

μ = 68 σ = 17

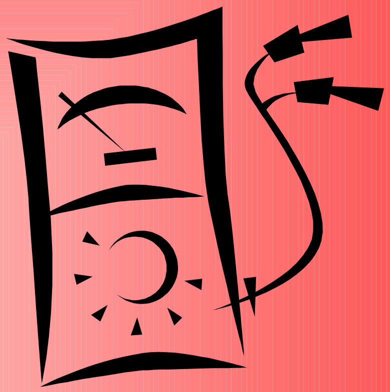
“M<90” or “M>55” or “60<M<85” @90% CL



These are each always true or always false

Solution: Refer to ensemble of statements

Frequentist CL in detail



You have a meter: no bias,

Gaussian error 0.1.

For a value X_T it gives a value X_M according to a Gaussian Distribution

X_M is within 0.1 of X_T 68% of the time

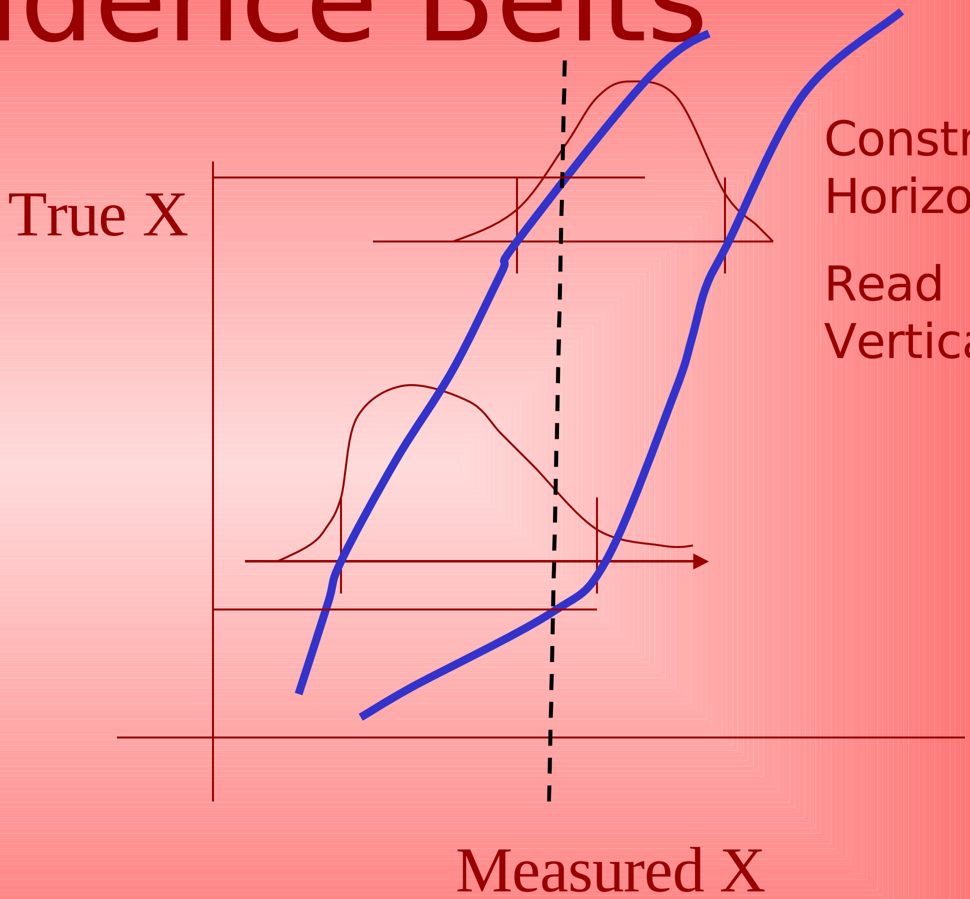
X_T is within 0.1 of X_M 68% of the time

Can state $X_M-0.1<X_T<X_M+0.1$ @68% CL

Confidence Belts

For more complicated distributions it isn't quite so easy

But the principle is the same



Construct Horizontally

Read Vertically

Coverage

“L<x<U” @ 95% confidence

(or “x>L” or “x<U”)

This statement belongs to an ensemble of similar statements of which at least* 95% are true

95% is the *coverage*

This is a statement about U and L, not about x.

*Maybe more. Overcoverage

Think about: the difference between a 90% upper limit and the upper limit of a 90% central interval.

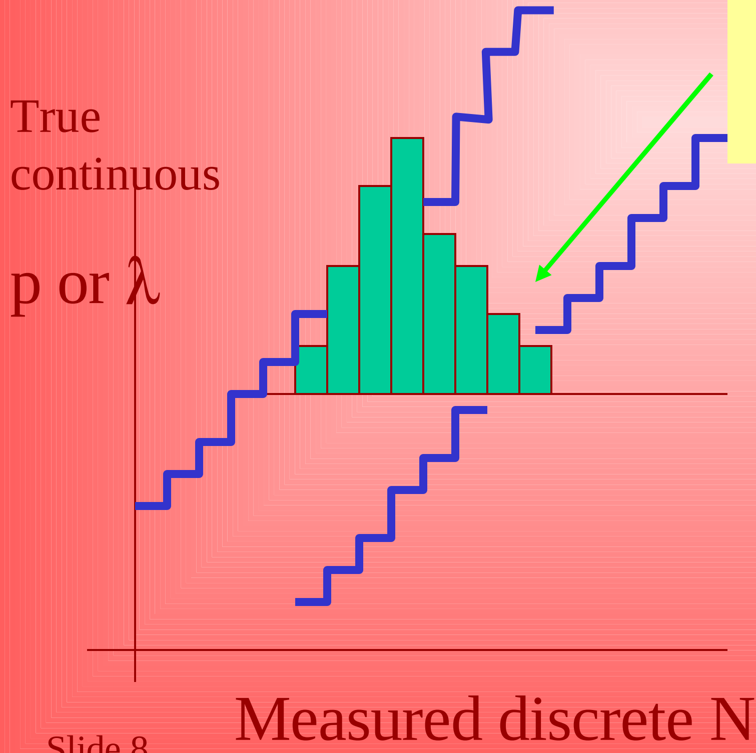
Discrete Distributions

CL belt edges become steps

May be unable to select (say) 5% region

Play safe.

Gives overcoverage



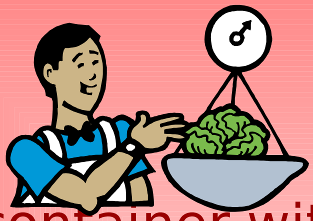
Binomial: see tables

Poisson

		Upper			Lower	
	90	95	99	90	95	99
0	2.3	3	4.61			
1	3.89	4.74	6.64	0.11	0.05	0.0
2	5.32	6.3	8.41	0.53	0.36	0.1
3	6.68	7.75	10.05	1.1	0.82	0.4

Given 2 events, if the true mean is 6.3 (or more) then the chance of getting a fluctuation this low

Problems for Frequentists



Weigh object+container with some Gaussian precision

Get reading R

$R-\sigma < M+C < R+\sigma$ @68%

$R-C-\sigma < M < R-C+\sigma$ @68%

E.g. C=50, R=141, $\sigma=10$

81<M<101 @68%

E.g. C=50, R=55, $\sigma=10$

-5 < M < 5 @68%

E.g. C=50, R=31, $\sigma=10$

-29 < M < -9 @68%

Poisson: Signal + Background

Background mean 2.50

Detect 3 events:

Total < 6.68 @ 95%

Signal<4.18@95%

Detect 0 events

Total < 2.30 @ 95%

Signal < -0.20 @ 95%

These statements are OK. We are allowed to get 32% / 5% wrong. But they are stupid

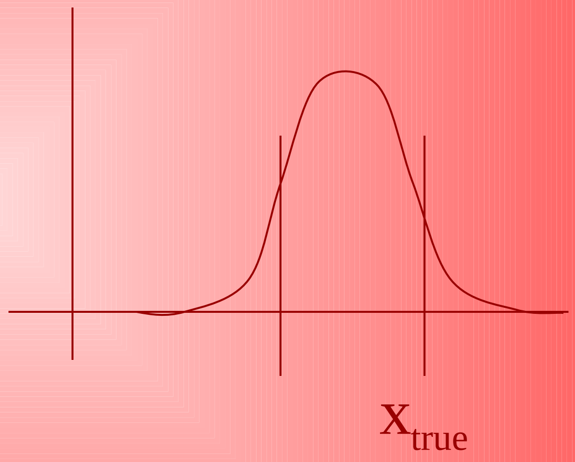
Bayes to the rescue

$$P(Theory\,|\,Data) = \frac{P(Data\,|\,Theory)}{P(Data)}\,P(Theory)$$

Standard (Gaussian) measurement

- No prior knowledge of true value
- No prior knowledge of measurement result
- P(Data|Theory) is Gaussian
- P(Theory|Data) is Gaussian

Interpret this with Probability statements in any way you please



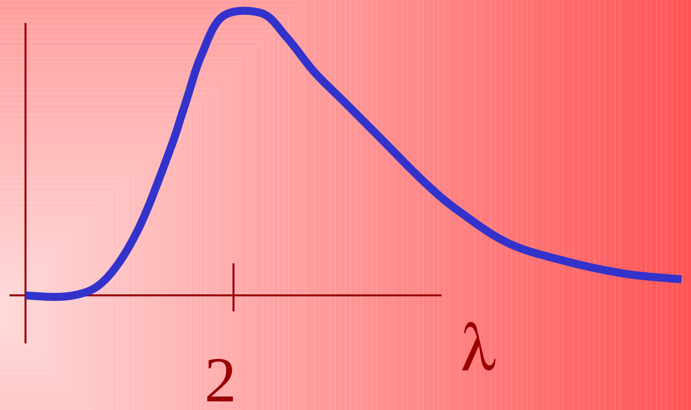
Gives same limits as Frequentist method for simple Gaussian

Bayesian Confidence Intervals (contd)

Observe (say) 2 events

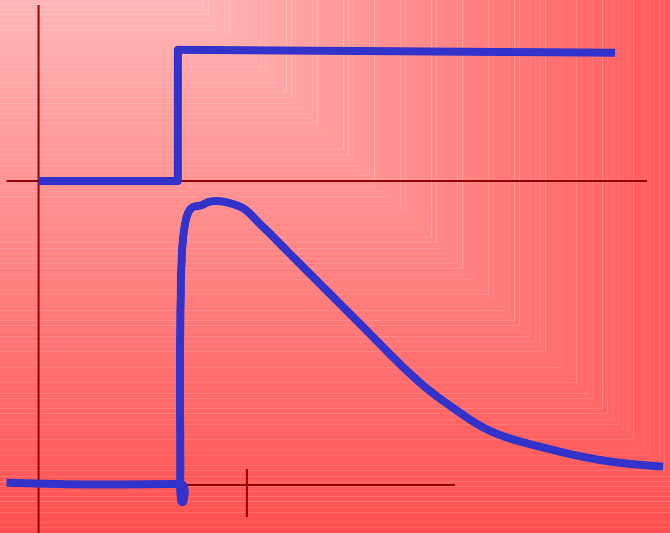
$P(\lambda;2) \propto P(2;\lambda) = e^{-\lambda}\lambda^2$

Normalise and interpret



If you know background mean is 1.7, then you know $\lambda > 1.7$

Multiply, normalise and interpret



Bayes: words of caution

Taking prior in λ as flat is not justified

Can argue for prior flat in ln λ or $1/\sqrt{\lambda}$ or whatever

Good practice to try a couple of priors to see if it matters

Feldman-Cousins Unified Method

Physicists are human

Ideal Physicist

1. Choose Strategy
2. Examine data
3. Quote result

Real Physicist

1. Examine data
2. Choose Strategy
3. Quote Result

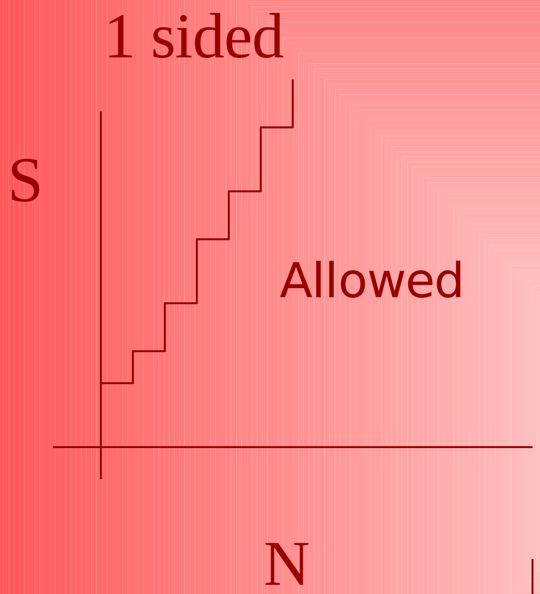
Example:

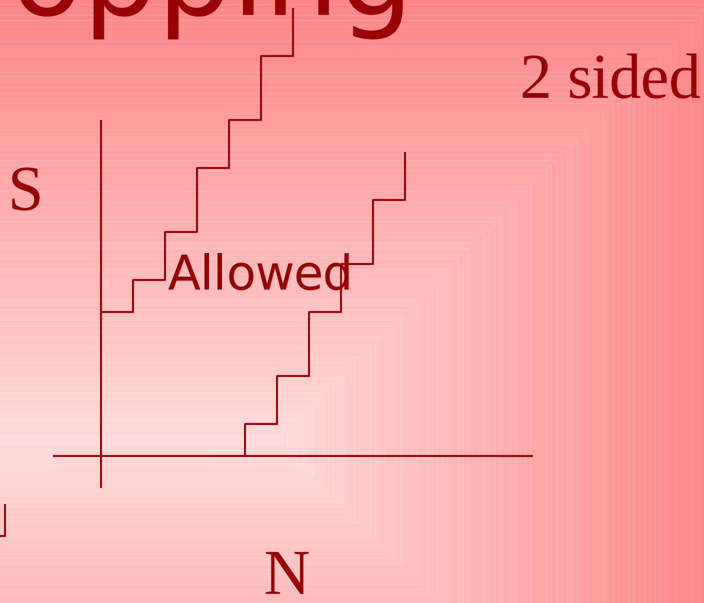
You have a background of 3.2

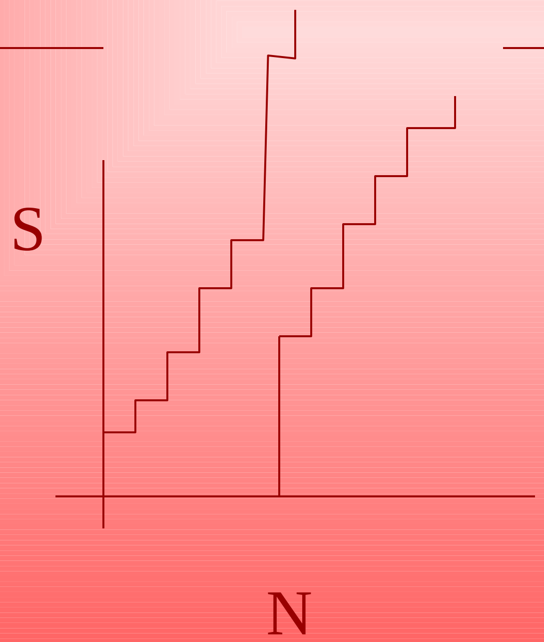
Observe 5 events? Quote one-sided upper limit (9.27-3.2 =6.07@90%)

Observe 25 events? Quote two-sided limits

“Flip-Flopping”

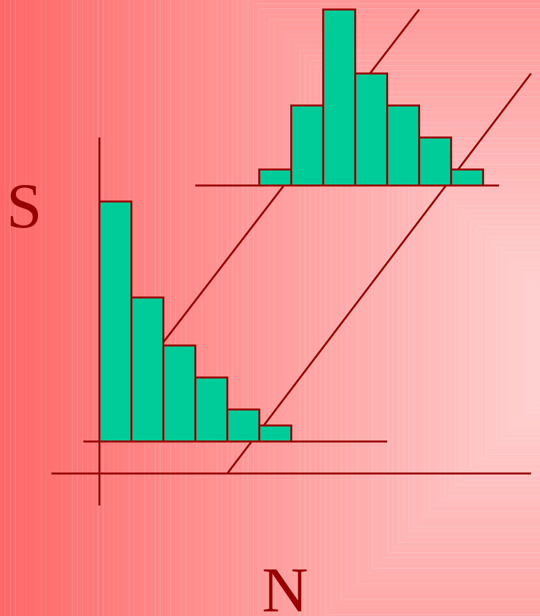






This is not a true confidence belt! Coverage varies.

Solution: Construct belt that does the flip-flopping



For 90% CL

For every S select set of N-values in belt

Total probability must sum to 90% (or more): there are many strategies for doing this

Crow & Gardner strategy (almost right):

Select N-values with highest probability $\rightarrow$ shortest interval

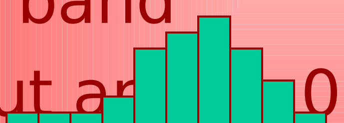
Better Strategy

N is Poisson from S+B

B known, may be large

E.g. B=9.2,S=0 and N=1

P=.1% - not in C-G band



But an 0 will be worse

Fair comparison of P is with best P for this N

Either at S=N-B or S=0

To construct band for a given S:

For all N:

Find P(N;S+B) and

P_{best}=P(N;N) if (N>B) else P(N;B)

Rank on P/P_{best}

Accept N into band until Σ P(N;S+B) $\geq$ 90%

Feldman and Cousins Summary

- Makes us more honest (a bit)
- Avoids forbidden regions in a Frequentist way

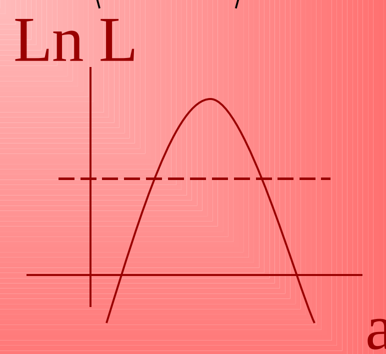
- Not easy to calculate
- Has to be done separately for each value of B
- Can lead to 2-tailed limits where you don't want to claim a discovery
- Weird effects for N=0; larger B gives lower (=better) upper limit

Maximum Likelihood and Confidence Levels

ML estimator (large N) has variance given by MVB $\sigma_{\hat{a}}{}^2 = V(\hat{a}) = \frac{-1}{\left\langle \frac{d^2 \ln L}{da^2} \right\rangle}$

At peak $\ln L \approx L_{\max} + \frac{(a-\hat{a})^2}{2} \left. \frac{d^2 \ln L}{da^2} \right|_{a=\hat{a}}$ For large N $\left\langle \frac{d^2 \ln L}{da^2} \right\rangle = \left. \frac{d^2 \ln L}{da^2} \right|_{a=\hat{a}}$

Ln L is a parabola (L is a Gaussian)



$$\ln L = L_{\max} - \frac{(a-\hat{a})^2}{2\sigma_{\hat{a}}^2}$$

Falls by ½ at $a = \hat{a} \pm \sigma_{\hat{a}}$

Falls by 2 at $a = \hat{a} \pm 2\sigma_{\hat{a}}$

Read off 68% , 95% confidence regions

MVB example

N Gaussian measurements: estimate μ

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

Ln L given by $-\sum_i \frac{(x_i-\mu)^2}{2\sigma^2} - N\ln\left(\sigma\sqrt{2\pi}\right)$

Differentiate twice wrt μ

$$-\frac{N}{\sigma^2}$$

Take expectation value – but it's a constant

Invert and negate:

$$V(\hat{\mu}) = \frac{\sigma^2}{N}$$

Another MVB example

N Gaussian measurements: estimate σ

Ln L still given by

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

Differentiate twice wrt σ

$$-\sum_i \frac{3(x_i-\mu)^2}{\sigma^4} + \frac{N}{\sigma^2}$$

Take expectation value $<(x_i-\mu)^2>=\sigma^2\ \forall\ i$

Gives

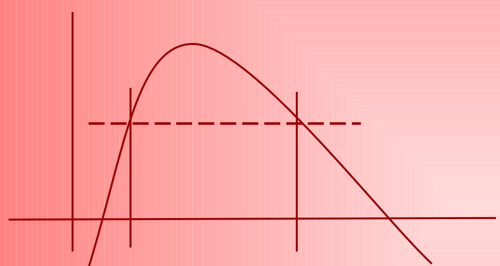
$$-\frac{2N}{\sigma^2}$$

Invert and negate:

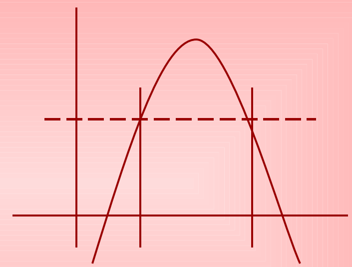
$$V(\hat{\sigma}) = \frac{\sigma^2}{2N}$$

ML for small N

ln L is not a parabola



a



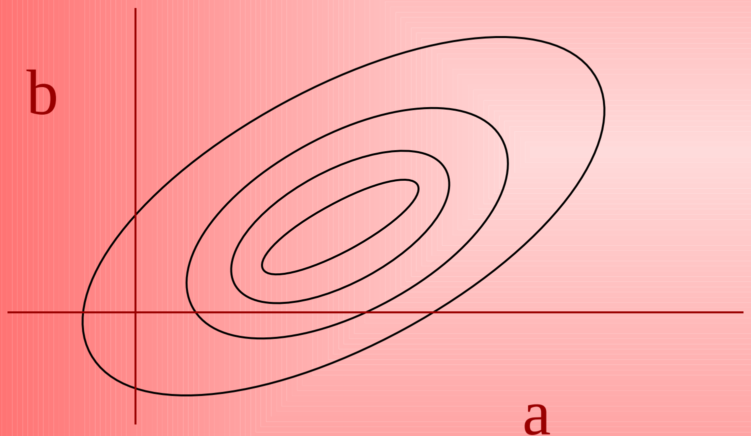
a'

- Argue: we could (invariance) transform to some a' for which it is a parabola
- We could/should then get limits on a' using standard L_{max}-½ technique
- These would translate to limits on a
- These limits would be at the values of a for which L= L_{max}-½

So just do it directly

Multidimensional ML

- L is multidimensional Gaussian



For 2-d 39.3% lies within 1σ i.e. within region bounded by $L=L_{max}-½$

For 68% need $L=L_{max}-1.15$

Construct region(s) to taste using numbers from integrated χ^2 distribution

Confidence Intervals

- Descriptive
- Frequentist
  - Feldman-Cousins technique
- Bayesian
- Maximum Likelihood
  - Standard
  - Asymmetric
  - Multidimensional

Statistics for HEP

Roger Barlow
Manchester University

Lecture 5: Errors

Simple Statistical Errors

$$f(x,y)$$

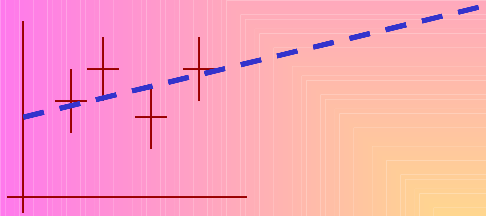
$$V(f)=\left(\frac{\partial f}{\partial x}\right)^2 V(x)+\left(\frac{\partial f}{\partial y}\right)^2 V(y)+2\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right)Cov(x,y)$$

$$V(x)=\sigma_x^2 \qquad V(y)=\sigma_y^2 \qquad Cov(x,y)=\rho\sigma_x\sigma_y$$

$$\mathbf{f}=\mathbf{G}\mathbf{x}$$

$$\mathbf{V_f}=\mathbf{G}\mathbf{V_x}\tilde{\mathbf{G}}$$

Correlation: examples



$$V(m) = \frac{\sigma^2}{N\left(\overline{x^2} - \bar{x}^2\right)}$$

$$V(c) = \frac{\sigma^2 \overline{x^2}}{N\left(\overline{x^2} - \bar{x}^2\right)}$$

$$Cov(m,c) = -\frac{\sigma^2 \bar{x}}{N\left(\overline{x^2} - \bar{x}^2\right)}$$

Extrapolate

Y=mX+c

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

Avoid by using

y=m(x- $\bar{x}$) +c'

Efficiency (etc)

r=N/N_T

$$V(r) = \left(\frac{1}{N_T}\right)^2 N + \left(\frac{-N}{N_T{}^2}\right)^2 N_T + 2\left(\frac{1}{N_T}\right)\left(\frac{-N}{N_T{}^2}\right) N$$

$$= \frac{N\left(N_T - N\right)}{N_T{}^3}$$

Avoid by using

r=N/(N+N_R)

Using the Covariance Matrix

Simple χ^2 : $\sum \left(\frac{x_i - f_i}{\sigma_i} \right)^2$

For uncorrelated data

Generalises to

$$(\tilde{\mathbf{x}} - \tilde{\mathbf{f}}) \mathbf{V}^{-1} (\mathbf{x} - \mathbf{f})$$

Multidimensional Gaussian

$$P(\mathbf{x}; \boldsymbol{\mu}, \mathbf{V}) = \frac{1}{(2\pi)^{N/2} \sqrt{|\mathbf{V}|}} e^{-\frac{1}{2}(\tilde{\mathbf{x}} - \tilde{\boldsymbol{\mu}}) \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

Building the Covariance Matrix

Variables x,y,z…

x=A+B

y=C+A+D

z=E+B+D+F

$$\begin{pmatrix} \sigma_A^2+\sigma_B^2 & \sigma_A^2 & \sigma_B^2 \\ \sigma_A^2 & \sigma_A^2+\sigma_C^2+\sigma_D^2 & \sigma_D^2 \\ \sigma_B^2 & \sigma_D^2 & \sigma_E^2+\sigma_B^2+\sigma_D^2+\sigma_F^2 \end{pmatrix}$$

…..

A,B,C,D…

independent

If you can split into separate bits like this then just put the σ^2 into the elements

Otherwise use $V=GVG^T$

Systematic Errors

Systematic Error: reproducible inaccuracy introduced by faulty equipment, calibration, or technique

Bevington

Systematic effects is a general category which includes effects such as background, scanning efficiency, energy resolution, angle resolution, variation of couner efficiency with beam position and energy, dead time, etc. The uncertainty in the estimation of such as systematic effect is called a *systematic error*

Orear

Error=mistake?

Error=uncertainty?

Experimental Examples

- Energy in a calorimeter E=aD+b

  a & b determined by calibration expt
- Branching ratio $B=N/(\eta N_T)$

  η found from Monte Carlo studies
- Steel rule calibrated at 15C but used in warm lab

  If not spotted, this is a mistake

  If temp. measured, not a problem

If temp. not measured guess $\rightarrow$uncertainty

Repeating measurements doesn't help

Theoretical uncertainties

An uncertainty which does not change when repeated does not match a Frequency definition of probability.

Statement of the obvious

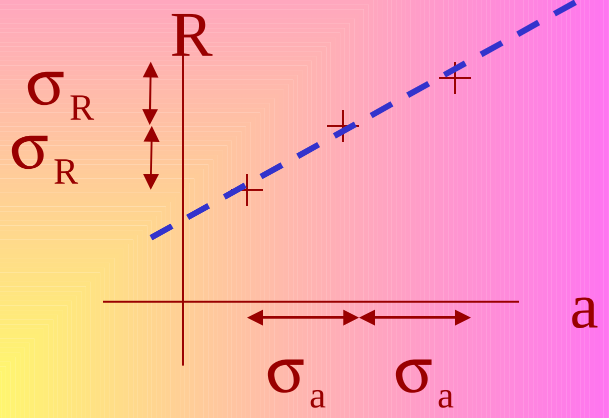
Theoretical parameters:

- B mass in CKM determinations
- Strong coupling constant in M_W
- All the Pythia/Jetset parameters in just about everything
- High order corrections in electroweak precision measurements
- etcetera etcetera etcetera.....

No alternative to subjective probabilities

But worry about robustness with changes of prior!

Numerical Estimation



Theory(?)

parameter a affects your result R

a is known only with some precision σ_a

Propagation of errors impractical as no algebraic form for R(a)

Use data to find dR/da and σ_a dR/da

Generally combined into one step

The ‘errors on errors’ puzzle

Suppose slope uncertain

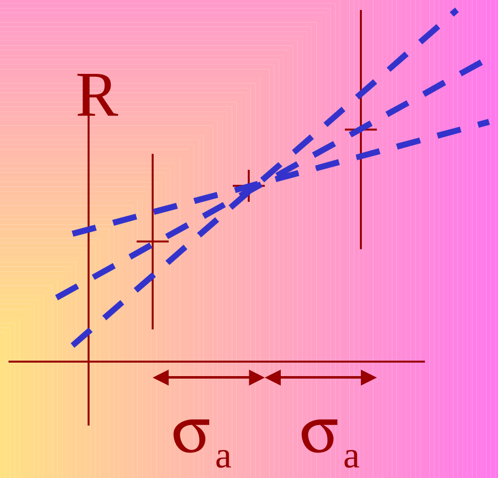
Uncertainty in σ_R.

Do you:

A. Add the uncertainty (in quadrature) to σ_R?

B. Subtract it from σ_R?

C. Ignore it?



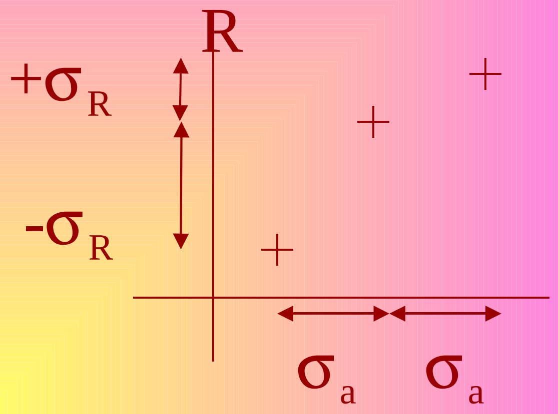
Timid and Wrong

Technically correct but hard to argue

Especially if $\sigma_{\sigma_R} > \sigma_R$

Strongly advised

Asymmetric Errors



- Can arise here, or from non-parabolic likelihoods
- Not easy to handle
- General technique for $x = y^{+\sigma_y^+}_{-\sigma_y^-} + z^{+\sigma_z^+}_{-\sigma_z^-}$ is to add separately

$$x^{+\sqrt{\left(\sigma_y^+\right)^2+\left(\sigma_z^+\right)^2}}_{-\sqrt{\left(\sigma_y^-\right)^2+\left(\sigma_z^-\right)^2}}$$

Not obviously correct

Introduce only if really justified

Errors from two values

Two models give results: R_1 and R_2

You can quote

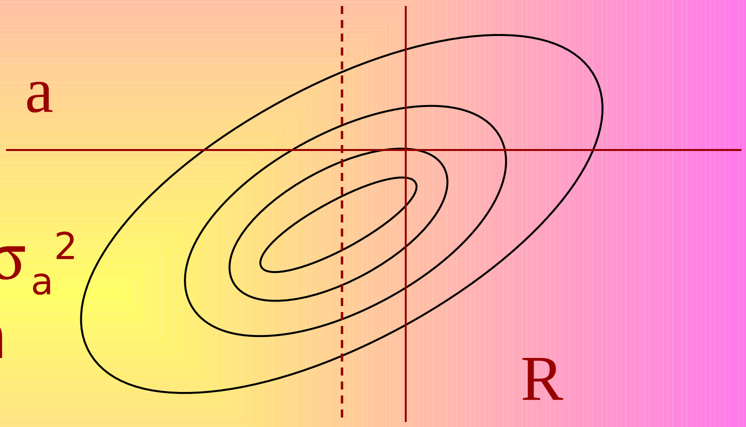
$R_1 \pm | R_1 - R_2 |$ if you prefer model 1

$\frac{1}{2}(R_1+R_2) \pm | R_1 - R_2 | /\sqrt{2}$ if they are equally rated

$\frac{1}{2}(R_1+R_2) \pm | R_1 - R_2 | /\sqrt{12}$ if they are extreme

Alternative: Incorporation in the Likelihood

- Analysis is some enormous likelihood maximisation
- Regard a as 'just another parameter': include $(a-a_0)^2/2\sigma_a^2$ as a chi squared contribution



- Can choose to allow a to vary. This will change the result and give a smaller error. Need strong nerves.
- If nerves not strong just use for errors
- Not clear which errors are 'systematic' and which are 'statistical' but not important

The Traditional Physics Analysis

1. Devise cuts, get result
2. Do analysis for statistical errors
3. Make big table
4. Alter cuts by arbitrary amounts, put in table
5. Repeat step 4 until time/money exhausted
6. Add table in quadrature
7. Call this the systematic error
8. If challenged, describe it as

Systematic Checks

- Why are you altering a cut?
- To evaluate an uncertainty? Then you know how much to adjust it.
- To check the analysis is robust? Wise move. But look at the result and ask ‘Is it OK?

Eg. Finding a Branching Ratio…

- Calculate Value (and error)
- Loosen cut
- Efficiency goes up but so does background. Re-evaluate them
- Re-calculate Branching Ratio (and error).
- Check compatibility

When are differences 'small'?

- It is OK if the difference is 'small' – compared to what?
- Cannot just use statistical error, as samples share data
- 'small' can be defined with reference to the difference in quadrature of the two errors

12 ± 5 and 8 ± 4 are OK.

18 ± 5 and 8 ± 4 are not

When things go right

DO NOTHING

Tick the box and move on

Do NOT add the difference to your systematic error estimate

- It’s illogical
- It’s pusillanimous
- It penalises diligence

When things go wrong

1. Check the test
2. Check the analysis
3. Worry and maybe decide there could be an effect
4. Worry and ask colleagues and see what other experiments did

99. Incorporate the discrepancy in the systematic

The VI commandments

Thou shalt never say ‘systematic error’ when thou meanest ‘systematic effect’ or ‘systematic mistake’

Thou shalt not add uncertainties on uncertainties in quadrature. If they are larger than chickenfeed, get more Monte Carlo data

Thou shalt know at all times whether thou art performing a check for a mistake or an evaluation of an uncertainty

Thou shalt not not incorporate successful check results into thy total systematic error and make thereby a shield behind which to hide thy dodgy result

Thou shalt not incorporate failed check results unless thou art truly at thy wits’ end

Thou shalt say what thou doest, and thou shalt be able to justify it out of thine own mouth, not the mouth of thy supervisor, nor thy colleague who did the analysis last time, nor thy mate down the pub.

Do these, and thou shalt prosper, and thine analysis likewise

Further Reading

R Barlow, Statistics. Wiley 1989

G Cowan, Statistical Data Analysis. Oxford 1998

L Lyons, Statistics for Nuclear and Particle Physicists, Cambridge 1986

B Roe, Probability and Statistics in Experimental Physics, Springer 1992

A G Frodesen et al, Probability and Statistics in Particle Physics, Bergen-Oslo-Tromso 1979

W T Eadie et al; Statistical Methods in Experimental Physics, North Holland 1971

M G Kendall and A Stuart; "The Advanced Theory of Statistics". 3+ volumes, Charles Griffin and Co 1979

Darrel Huff "How to Lie with Statistics" Penguin

CERN Workshop on Confidence Limits. Yellow report 2000-005

Proc. Conf. on Adv. Stat. Techniques in Particle Physics, Durham, IPPP/02/39

http://www.hep.man.ac.uk/~roger

The End