Statistics
Probability and Likelihood

Roger Barlow

Manchester University

Hadron Collider Physics Summer School, Fermilab

20th August 2010

Problem #1

Particle physics is Random -

You are measuring some number of events.

'Theory' prediction is 6.7

What can you say about the actual number you will observe?

Problem #1

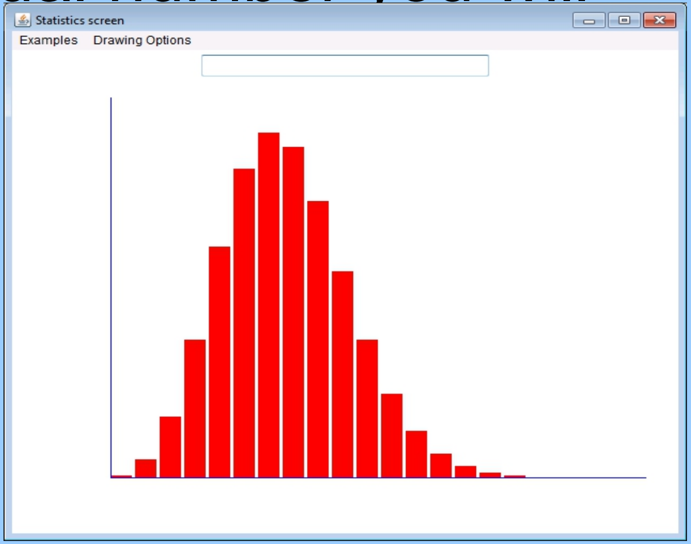
Particle physics is Random -

You are measuring some number of events.

'Theory' prediction is 6.7

What can you say about the actual number you will observe?

$$P(n;\mu)=e^{-\mu}\frac{\mu^n}{n!}$$



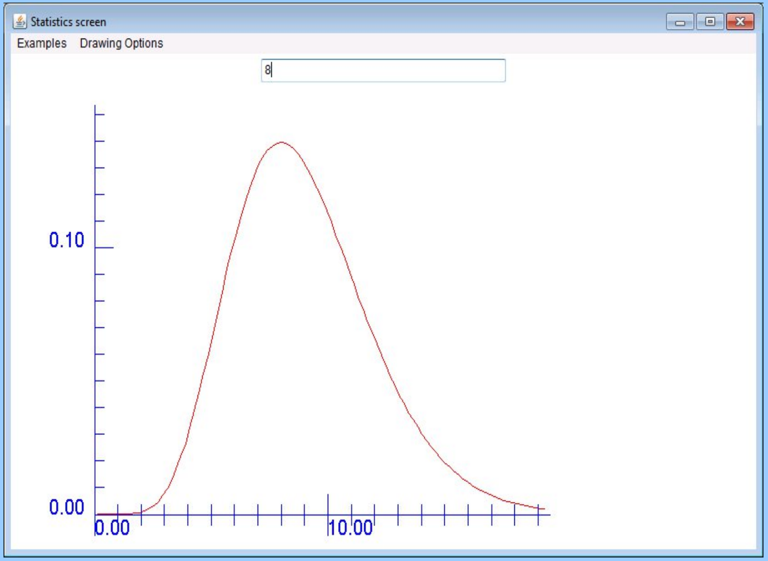
Problem #2

You are measuring some number of events.

You observe 8

What can you say about the actual number?

This is inference, not prediction



$$P(n;\mu) = e^{-\mu}\frac{\mu^n}{n!}$$

What is Probability?

A is some possible event. What is P(A)?

What is Probability?

A is some possible event. What is P(A)?

Frequentist: Limit $_{N\to\infty}$ N(A) / N

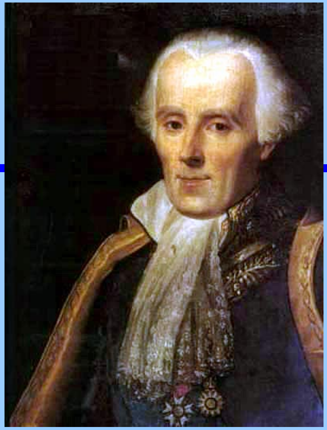
Mathematical: Some number between 0 and 1 obeying certain rules.

Classical: An intrinsic property or strength of A

Bayesian: My degree of belief in A

All 4 answers are true

Classical (Laplace and others)



Symmetry factor

- Coin – ½
- Cards – $^{1}/_{52}$
- Dice – $^{1}/_{6}$
- Roulette – $^{1}/_{32}$

Equally likely outcomes

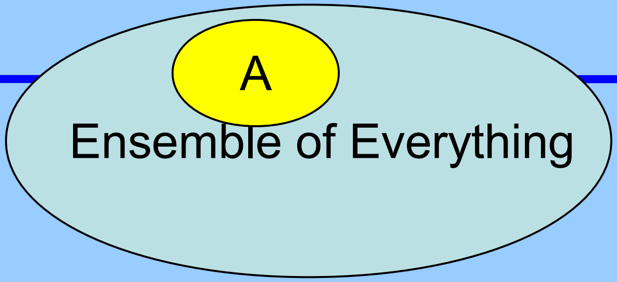
The probability of an event is the ratio of the number of cases favourable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible.

Théorie analytique des probabilités

Extend to more complicated systems of several coins, many cards, etc.

Does not (easily) extend to continuous choices, and other situations.

Frequentist Probability (von Mises, Fisher)



Limit of frequency

$P(A) = \text{Limit}_{N \to \infty} N(A)/N$

This was a property of the classical definition, now promoted to become a definition itself

P(A) depends not just on A but on the ensemble – which must be specified.

A property: There can be several Ensembles

- Probabilities belong to the event and the ensemble
- Insurance company data shows P(death) for 40 year old male clients = 1.4% (Classic example due to von Mises)
- Does this mean a particular 40 year old German has a 98.6% chance of reaching his 41st Birthday?
- No. He belongs to many ensembles
  - German insured males
  - German males
  - Insured nonsmoking vegetarians
  - Overweight alcohol-consuming physicists
  - …

Each of these gives a different number. All equally valid.

Limitation: There may be no ensemble

Some events are unique. Consider

“It will probably rain tomorrow.”

or even

“There is a 70% probability of rain tomorrow”

There is only one tomorrow (Saturday). There is NO ensemble. P(rain) is either 0/1 =0 or 1/1 = 1

Strict frequentists cannot say 'It will probably rain tomorrow'.

This presents severe social problems.

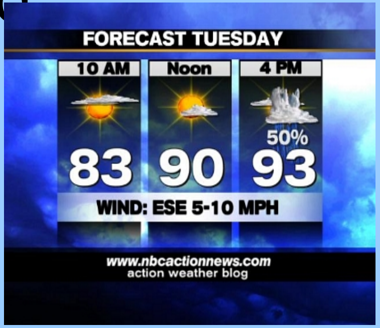
Circumventing the limitation

A frequentist can say:

“The statement ‘It will rain tomorrow’ has a 70% probability of being true.”

by assembling an ensemble of statements and ascertaining that at least 70% are true.

(E.g. Weather forecasts with a verified track record)



Say “It will rain tomorrow” with 70% confidence

For unique events, confidence level statements replace probability statements.

What is a Confidence Level?

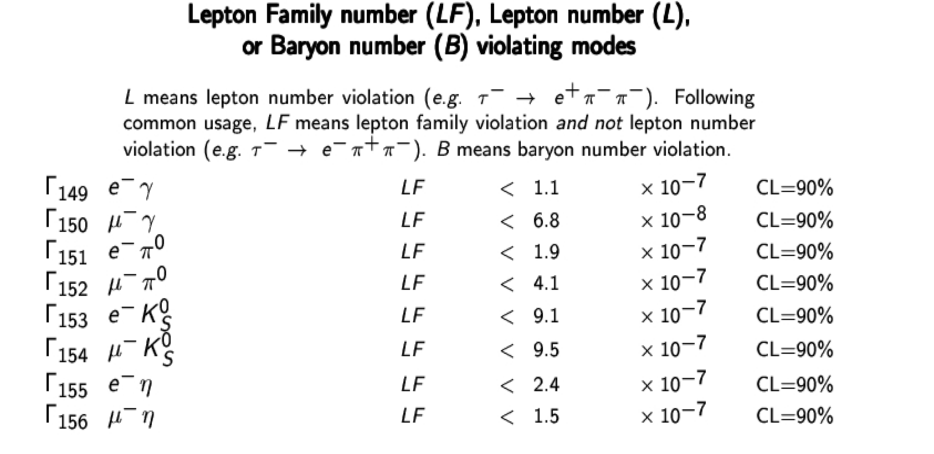
Not just “the probability that the result is true”

52 LFV violating tau decays

52 confidence limits at 90%

Anyone believe ~5 of these limits are exceeded?

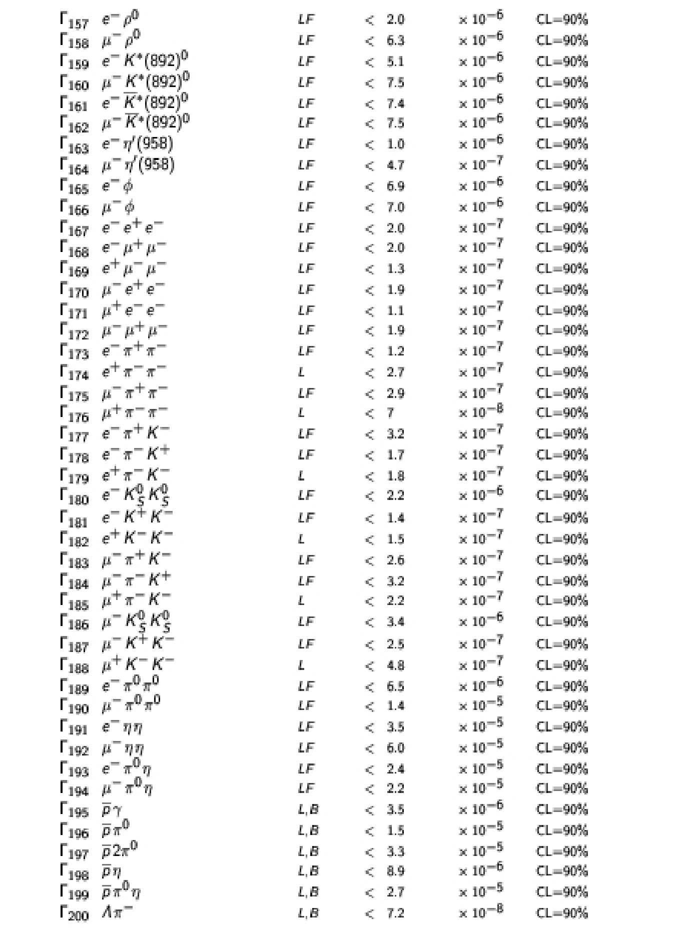
That 'at least' can be important...



Lepton Family number (LF), Lepton number (L), or Baryon number (B) violating modes

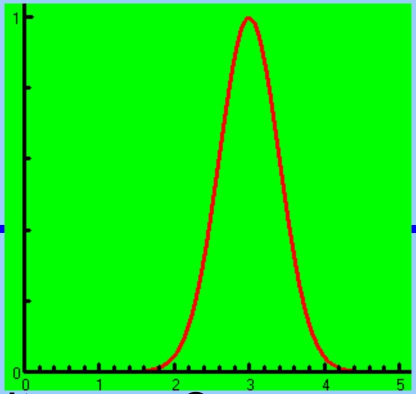
L means lepton number violation (*e.g.* $\tau^- \rightarrow e^+\pi^-\pi^-$). Following common usage, LF means lepton family violation *and not* lepton number violation (*e.g.* $\tau^- \rightarrow e^-\pi^+\pi^-$). B means baryon number violation.

Γ_{149}	$e^-\gamma$	LF	$< 1.1 \times 10^{-7}$	CL=90%
Γ_{150}	$\mu^-\gamma$	LF	$< 6.8 \times 10^{-8}$	CL=90%
Γ_{151}	$e^-\pi^0$	LF	$< 1.9 \times 10^{-7}$	CL=90%
Γ_{152}	$\mu^-\pi^0$	LF	$< 4.1 \times 10^{-7}$	CL=90%
Γ_{153}	$e^-K^0_S$	LF	$< 9.1 \times 10^{-7}$	CL=90%
Γ_{154}	$\mu^-K^0_S$	LF	$< 9.5 \times 10^{-7}$	CL=90%
Γ_{155}	$e^-\eta$	LF	$< 2.4 \times 10^{-7}$	CL=90%
Γ_{156}	$\mu^-\eta$	LF	$< 1.5 \times 10^{-7}$	CL=90%



Γ_{157}	$e^-\rho^0$	LF	$< 2.0 \times 10^{-6}$	CL=90%
Γ_{158}	$\mu^-\rho^0$	LF	$< 6.3 \times 10^{-6}$	CL=90%
Γ_{159}	$e^-K^*(892)^0$	LF	$< 5.1 \times 10^{-6}$	CL=90%
Γ_{160}	$\mu^-K^*(892)^0$	LF	$< 7.5 \times 10^{-6}$	CL=90%
Γ_{161}	$e^-\overline{K}^*(892)^0$	LF	$< 7.4 \times 10^{-6}$	CL=90%
Γ_{162}	$\mu^-\overline{K}^*(892)^0$	LF	$< 7.5 \times 10^{-6}$	CL=90%
Γ_{163}	$e^-\eta'(958)$	LF	$< 1.0 \times 10^{-6}$	CL=90%
Γ_{164}	$\mu^-\eta'(958)$	LF	$< 4.7 \times 10^{-7}$	CL=90%
Γ_{165}	$e^-\phi$	LF	$< 6.9 \times 10^{-6}$	CL=90%
Γ_{166}	$\mu^-\phi$	LF	$< 7.0 \times 10^{-6}$	CL=90%
Γ_{167}	$e^-e^+e^-$	LF	$< 2.0 \times 10^{-7}$	CL=90%
Γ_{168}	$e^-\mu^+\mu^-$	LF	$< 2.0 \times 10^{-7}$	CL=90%
Γ_{169}	$e^+\mu^-\mu^-$	LF	$< 1.3 \times 10^{-7}$	CL=90%
Γ_{170}	$\mu^-e^+e^-$	LF	$< 1.9 \times 10^{-7}$	CL=90%
Γ_{171}	$\mu^+e^-e^-$	LF	$< 1.1 \times 10^{-7}$	CL=90%
Γ_{172}	$\mu^-\mu^+\mu^-$	LF	$< 1.9 \times 10^{-7}$	CL=90%
Γ_{173}	$e^-\pi^+\pi^-$	LF	$< 1.2 \times 10^{-7}$	CL=90%
Γ_{174}	$e^+\pi^-\pi^-$	L	$< 2.7 \times 10^{-7}$	CL=90%
Γ_{175}	$\mu^-\pi^+\pi^-$	LF	$< 2.9 \times 10^{-7}$	CL=90%
Γ_{176}	$\mu^+\pi^-\pi^-$	L	$< 7 \times 10^{-8}$	CL=90%
Γ_{177}	$e^-\pi^+K^-$	LF	$< 3.2 \times 10^{-7}$	CL=90%
Γ_{178}	$e^-\pi^-K^+$	LF	$< 1.7 \times 10^{-7}$	CL=90%
Γ_{179}	$e^+\pi^-K^-$	L	$< 1.8 \times 10^{-7}$	CL=90%
Γ_{180}	$e^-K^0_SK^0_S$	LF	$< 2.2 \times 10^{-6}$	CL=90%
Γ_{181}	$e^-K^+K^-$	LF	$< 1.4 \times 10^{-7}$	CL=90%
Γ_{182}	$e^+K^-K^-$	L	$< 1.5 \times 10^{-7}$	CL=90%
Γ_{183}	$\mu^-\pi^+K^-$	LF	$< 2.6 \times 10^{-7}$	CL=90%
Γ_{184}	$\mu^-\pi^-K^+$	LF	$< 3.2 \times 10^{-7}$	CL=90%
Γ_{185}	$\mu^+\pi^-K^-$	L	$< 2.2 \times 10^{-7}$	CL=90%
Γ_{186}	$\mu^-K^0_SK^0_S$	LF	$< 3.4 \times 10^{-6}$	CL=90%
Γ_{187}	$\mu^-K^+K^-$	LF	$< 2.5 \times 10^{-7}$	CL=90%
Γ_{188}	$\mu^+K^-K^-$	L	$< 4.8 \times 10^{-7}$	CL=90%
Γ_{189}	$e^-\pi^0\pi^0$	LF	$< 6.5 \times 10^{-6}$	CL=90%
Γ_{190}	$\mu^-\pi^0\pi^0$	LF	$< 1.4 \times 10^{-5}$	CL=90%
Γ_{191}	$e^-\eta\eta$	LF	$< 3.5 \times 10^{-5}$	CL=90%
Γ_{192}	$\mu^-\eta\eta$	LF	$< 6.0 \times 10^{-5}$	CL=90%
Γ_{193}	$e^-\pi^0\eta$	LF	$< 2.4 \times 10^{-5}$	CL=90%
Γ_{194}	$\mu^-\pi^0\eta$	LF	$< 2.2 \times 10^{-5}$	CL=90%
Γ_{195}	$\overline{p}\gamma$	L,B	$< 3.5 \times 10^{-6}$	CL=90%
Γ_{196}	$\overline{p}\pi^0$	L,B	$< 1.5 \times 10^{-5}$	CL=90%
Γ_{197}	$\overline{p}2\pi^0$	L,B	$< 3.3 \times 10^{-5}$	CL=90%
Γ_{198}	$\overline{p}\eta$	L,B	$< 8.9 \times 10^{-6}$	CL=90%
Γ_{199}	$\overline{p}\pi^0\eta$	L,B	$< 2.7 \times 10^{-5}$	CL=90%
Γ_{200}	$\Lambda\pi^-$	L,B	$< 7.2 \times 10^{-8}$	CL=90%

Gaussian Measurement and Frequentist probability



M_T=174±3 GeV :What does it mean?

For true value μ the probability (density) for a result x is (for the usual Gaussian measurement)

$$P(x\,;\,\mu,\,\sigma)=(1/_{\sigma\sqrt{2\pi}})\ \exp-[(x-\mu)^2/2\sigma^2]$$

For a given μ, the probability that x lies within $\pm\sigma$ is 68%.

P(x; μ, σ) cannot be used as a probability for μ.

M_T=174±3 GeV

Is there a 68% probability that M_T lies between 171 and 177 GeV?

No. M_T is unique. It is either in the range or outside.

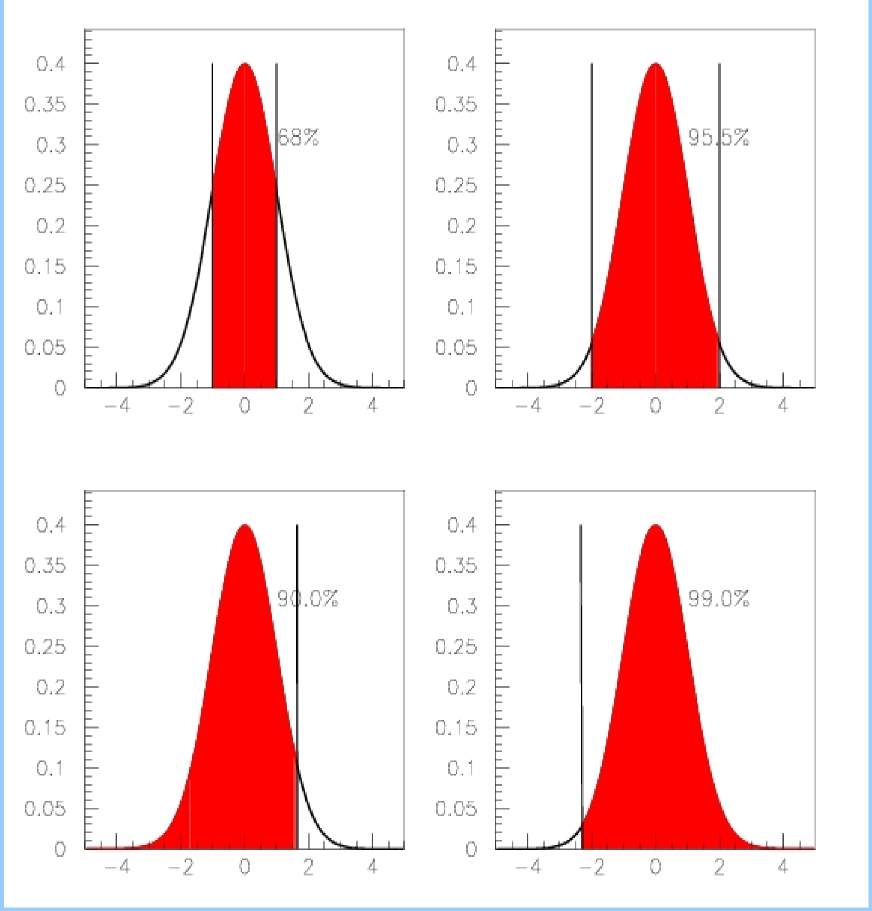
But $\mu \pm 3$ does bracket x 68% of the time: The statement ‘M_T lies between 171 and 177 GeV’ has a 68% probability of being true.

M_T lies between 171 and 177 GeV with 68% confidence

Choices, choices!

You can choose

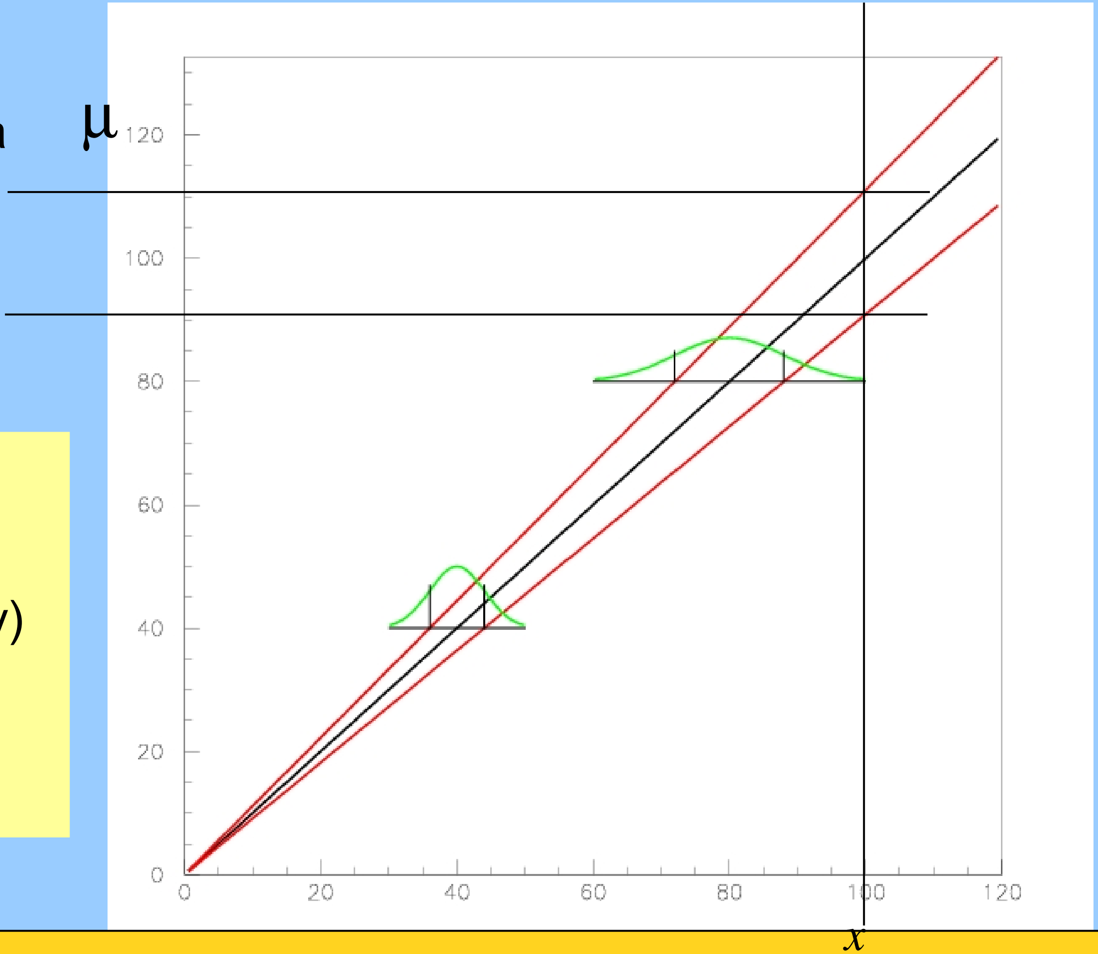
- The Confidence Level
- Whether to quote an upper limit or a lower limit or a 2-sided limit
- What sort of 2 sided limit (central, shortest,...)



Beyond the Simple Gaussian: Confidence Belt

Constructed horizontally such that the probability of a result lying inside the belt is 68%(or whatever)

Read vertically using the measurement



Example: proportional Gaussian $\sigma = 0.1\ \mu$

(Measures with 10% accuracy)

Result (say) 100.0

μ_{LO}=90.91 μ_{HI}= 111.1

Whatever the value of the ordinate (true value),the probability of the result falling in the belt is 68%

Given a result (abscissa) we say with 68% confidence that it falls in the belt

Next complication: Discrete observations

Poisson Formula

μ=1.1

n	p(N)
0	33.3%
1	36.6%
2	20.1%
3	7.4%
4	2.0%
5	0.5%
6	0.1%
.....	...

$$P(n;\mu)=e^{-\mu}\frac{\mu^n}{n!}$$

To make a 95% upper limit:
n=0,1,2 with probability 90.0%
n=0,1,2,3 with probability 97.4%

Play safe: include 3

Upper Limit μ_{HI} : n is small. μ can't be very large. If the true value is μ_{HI} (or higher) then the chance of a result this small (or smaller) is only (1-CL) (or less)

Poisson table

Found by solving

$$\Sigma_0^n P(n,\lambda)=\alpha$$

For high limit

$$\Sigma_0^{n-1} P(n,\lambda)=1-\alpha$$

For low limit

90% limits

n	lo	hi
0	-	2.30
1	.105	3.89
2	.532	5.32
3	1.10	6.68
4	1.74	7.99
5	2.43	9.27

95% limits

n	lo	hi
0	-	3.00
1	.051	4.74
2	.355	6.30
3	0.818	7.75
4	1.37	9.15
5	1.97	10.51
….		

Technical point: Coverage

How often will your limit statements be true?

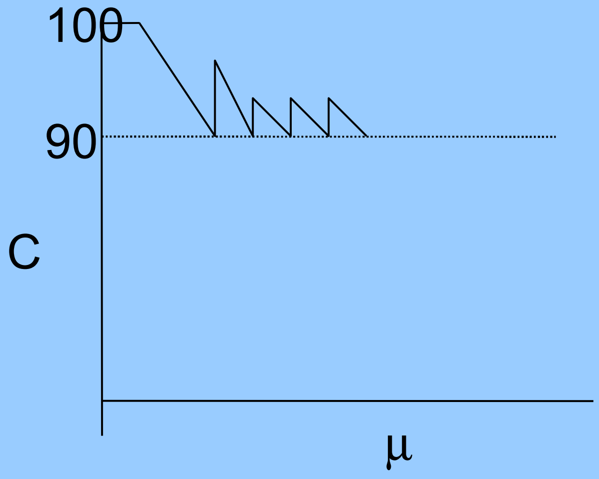
Should be same as CL, surely?

Yes. Unless you fall foul of the 'more than' stuff

Coverage is a function of μ (etc)

A (frequentist) test may “overcover” - coverage greater than CL

It should never undercover (by construction)



Confidence and significance

For historical reasons CL = 1-α

α is the Significance. Language of Hypothesis Testing:

Suppose the pdf really has this form. Then the probability that it would give a measurement this far (or further!) from the true one is α.

'Improvement among patients taking the treatment was significant at the 5% level' means that if the treatment does nothing, the probability of getting an effect this large (or larger) is 5% (or less).

Given a measurement, the corresponding probability is called the p-value. The null hypothesis is rejected if the p-value is smaller than the significance

Significance and p value have the same formula – but one is constructed before the data are seen, the second afterwards

Goodness of Fit

An instance of 'Hypothesis testing'.

Hypothesis being tested is that the theory describes the data.

Some measure of agreement is constructed, often χ^2

p-value of (say) 2.3% => If the theory truly does describe the data the probability of an agreement this bad (or worse) is only 2.3%

The χ^2 distribution and what it can do

Form sum of N results

$$\chi^2 = \Sigma \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2 = (\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{f}}) \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{f})$$

Distribution: integrated multidimensional Gaussian.
Depends on N, and has mean N (but χ^2/N not too useful)

Tables/functions exist for 'χ^2 probability'

i.e. p-value for this χ^2

i.e. Integral from χ^2 to infinity

i.e. probability of getting this bad an agreement by chance.

N sigma results

P-values (from χ^2 and elsewhere) are often converted into Gaussian discrepancies:

$2.7\ 10^{-3}$	$3\ \sigma$	'Evidence for'
$5.7\ 10^{-7}$	$5\ \sigma$	'Discovery of'

Question: Why don't particle physicists accept 99.73% probability as good enough?

N sigma results

Question: Why don't particle physicists accept 99.73% probability as good enough?'

Answer: Past experience!

Pentaquarks, Y(5.97), Top discovery at UA1...

Techniques for getting False Discoveries

1. Creativity.("Michaelangelo Method") Now controlled by the Blind Analysis technique

2. Reflections. Particle mis-ID or the effect of some kinematic or detector constraint.

3. Sheer hard work. Plot everything you can think of.

4. "Look Elsewhere effect." Applying statistical tools appropriate to a simple hypothesis to a range of hypotheses.

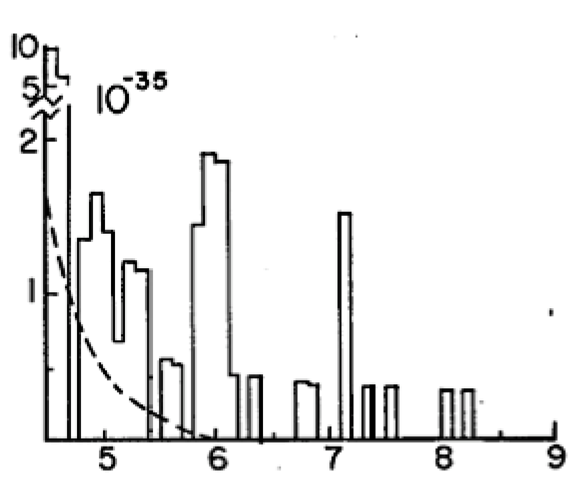
P.R.L. 36: 1236–1239 revisited

27 high mass events between 5.5 and 10 GeV.

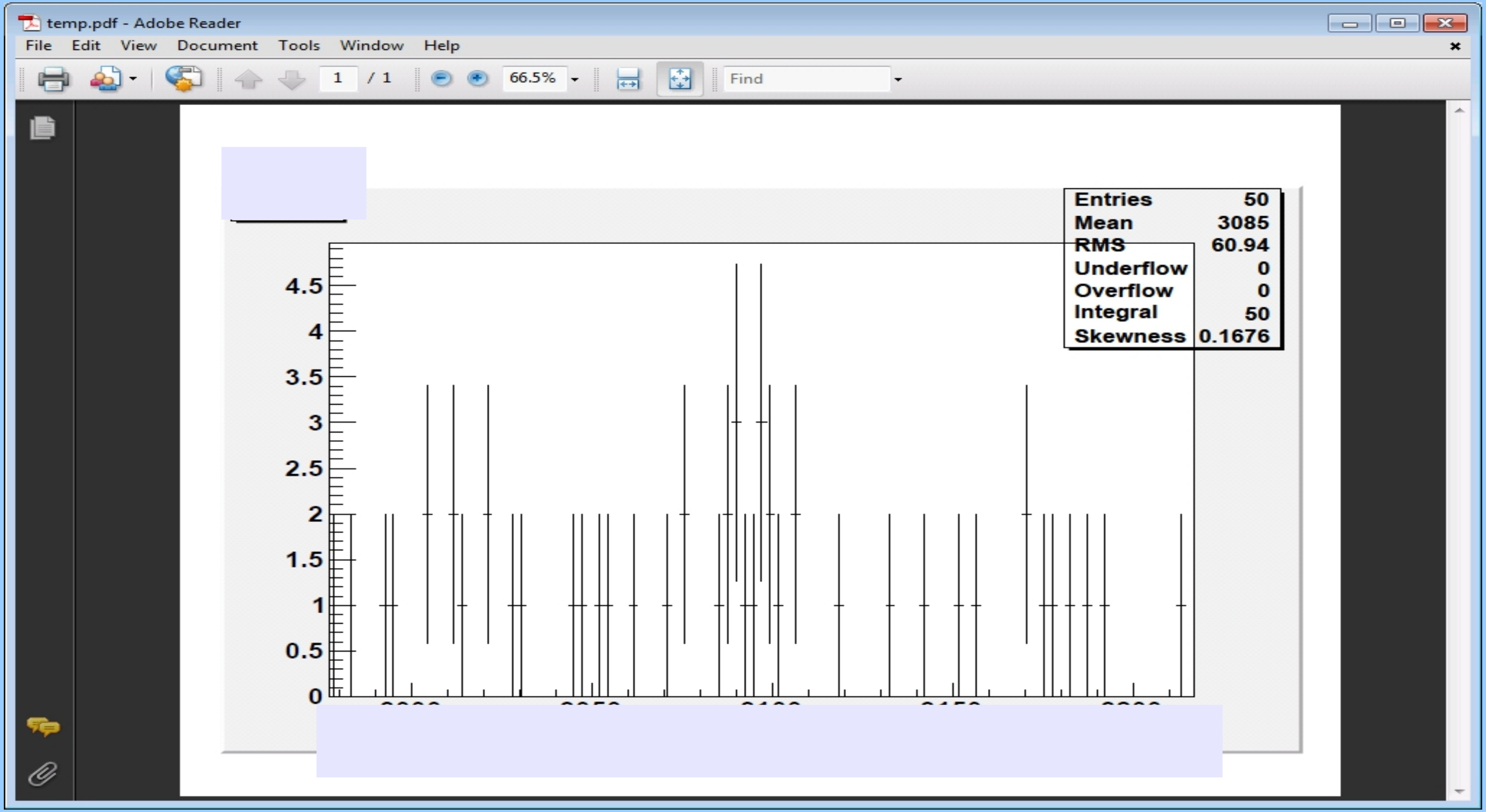
11 events between 5.8 and 6.1

'less than one chance in fifty

that this is a coincidence'

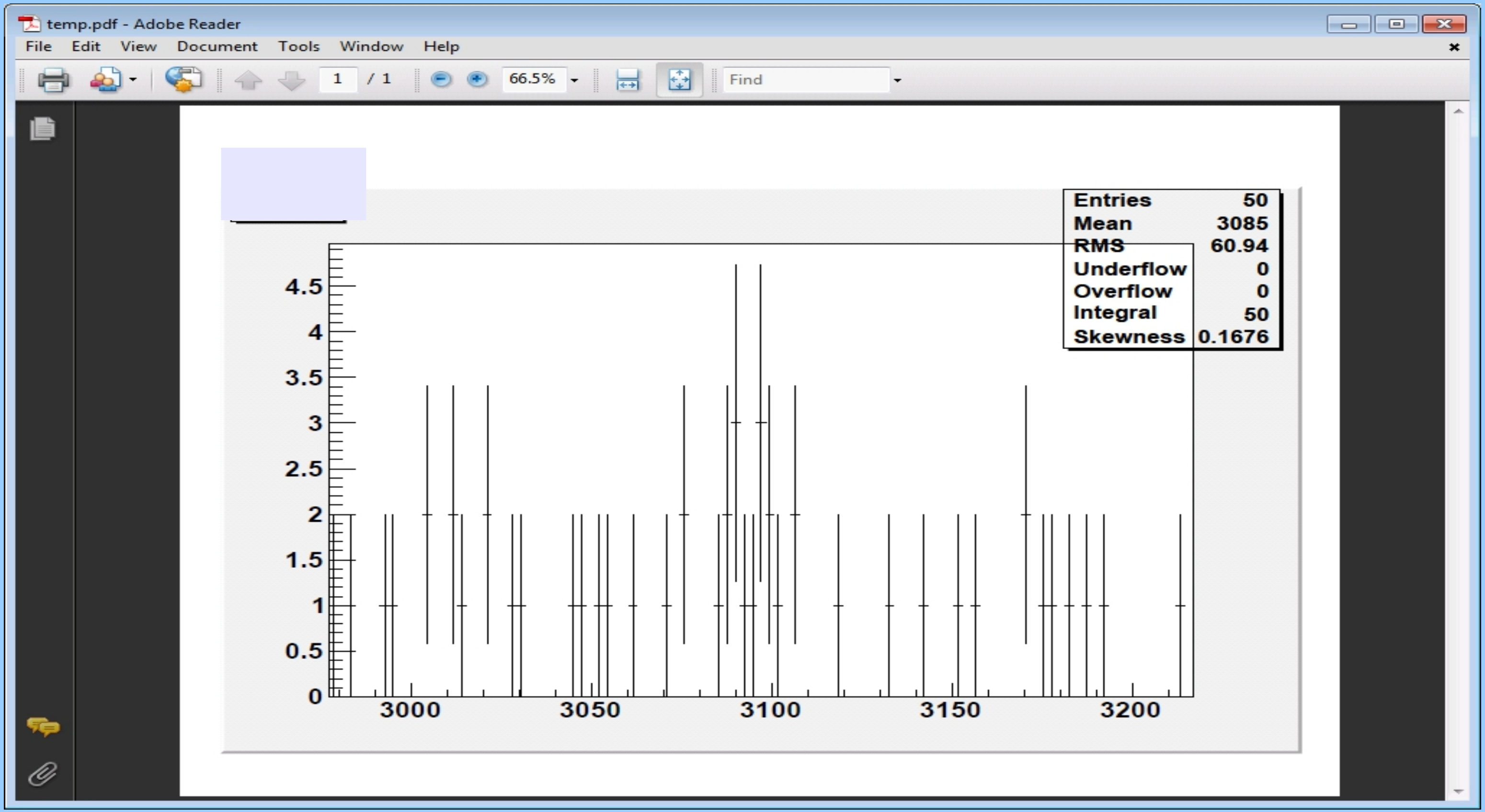


Is there a peak?



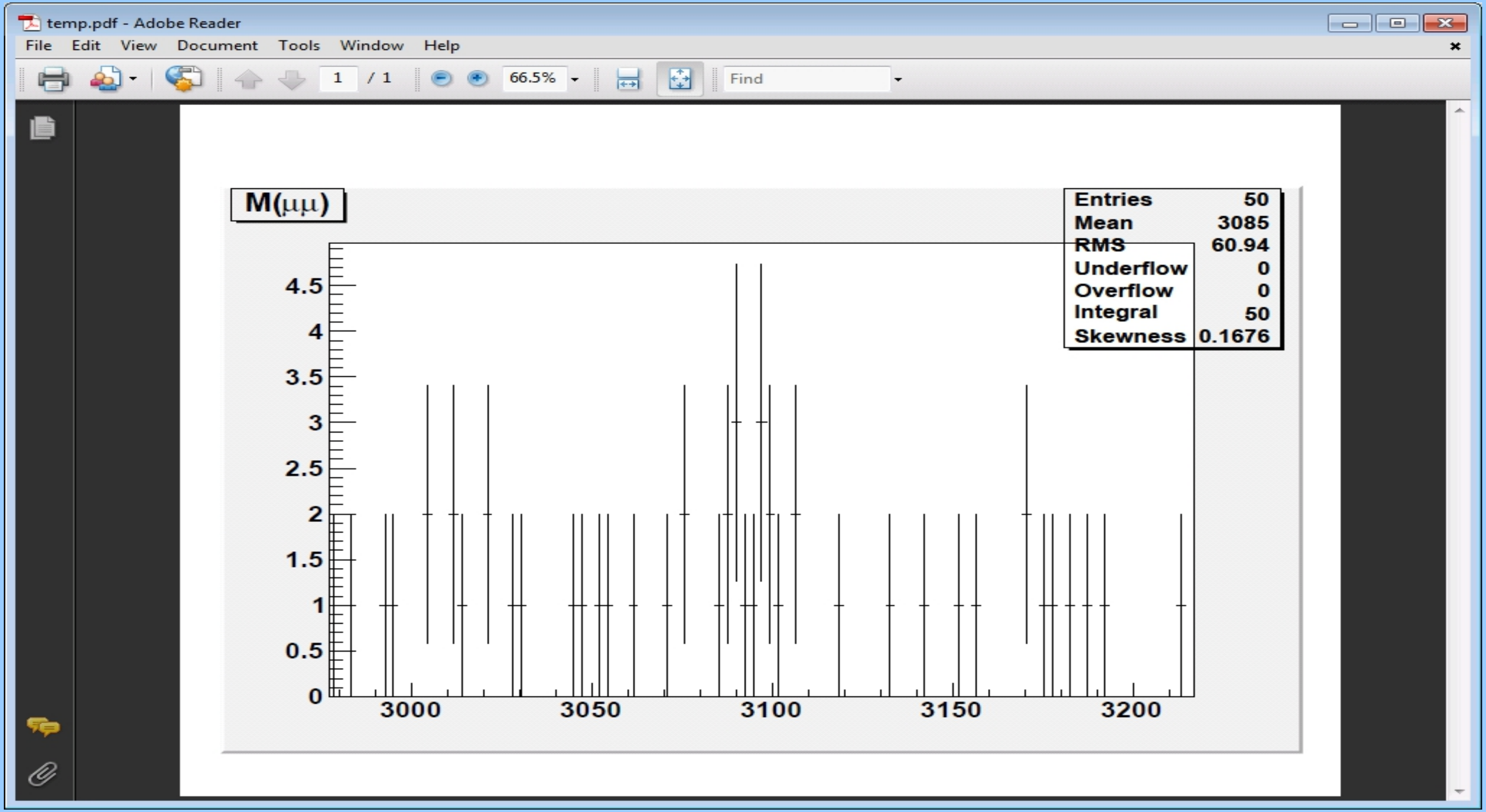
Entries	50
Mean	3085
RMS	60.94
Underflow	0
Overflow	0
Integral	50
Skewness	0.1676

Is there a peak?



Entries	50
Mean	3085
RMS	60.94
Underflow	0
Overflow	0
Integral	50
Skewness	0.1676

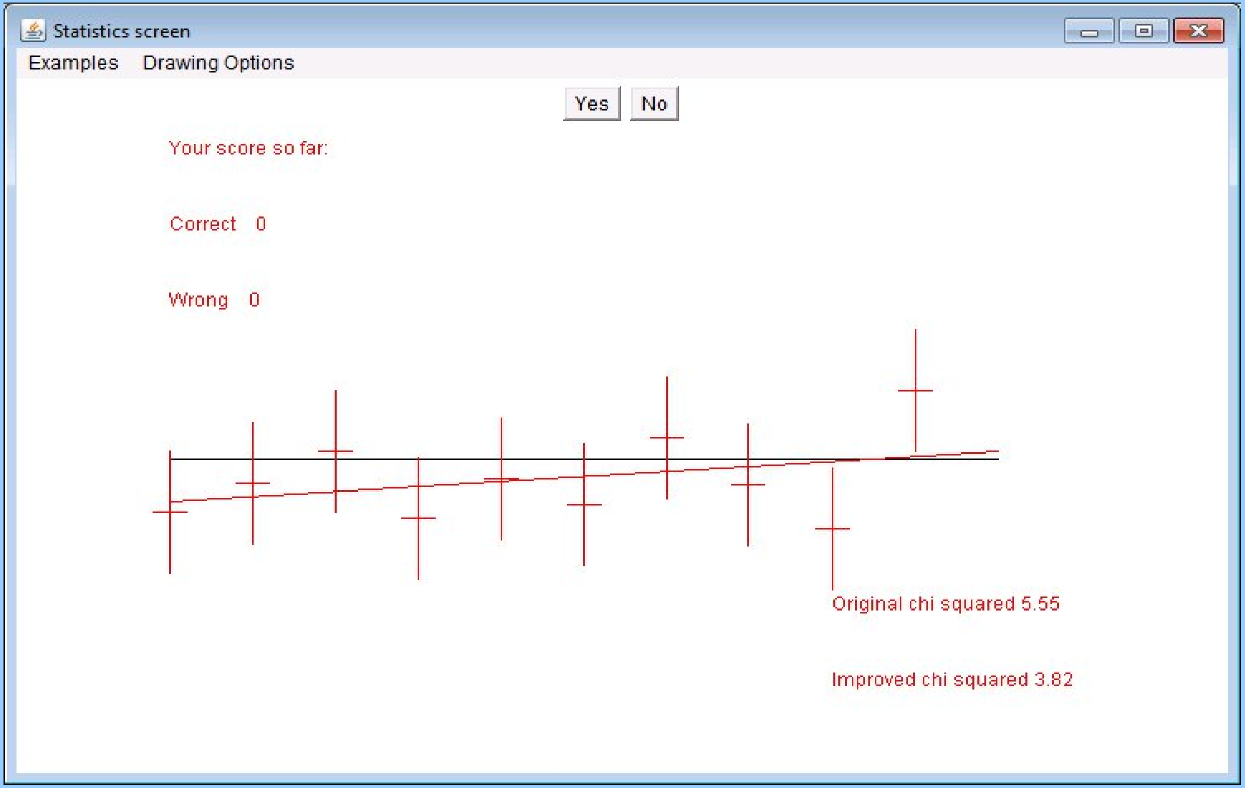
Is there a peak?



Entries	50
Mean	3085
RMS	60.94
Underflow	0
Overflow	0
Integral	50
Skewness	0.1676

Goodness of fit

Test the Standard Model...



Your score so far:

Correct 0

Wrong 0

Original chi squared 5.55

Improved chi squared 3.82

Chi squared: another nice thing

If you adjust parameters to fit the theory to the data, that improves χ^2 by (on average) 1.0 per parameter.

The improved distribution and the difference both have χ^2 distributions with appropriate N

This does not always apply – specifically if the improved model contains parameters which are meaningless under the old model.

Watch out!

Basic model – flat $f(x)=0.5$

Straight line

- $f(x)=m\,x + c$ OK

flat + bump

- $f(x)=c + n\,exp(-(x-.3)^2/.02)$ - OK

flat+bump

- $f(x)=c + n\,exp(-(x-m)^2/.02)$ - No. If $n=0$ then m is meaningless

Davies Biometrika papers

Last example is obvious when you think about a narrow peak.

Fixed-position peak enables you to eliminate the discrepancy at that peak. Reduce χ^2 by ~ 1

Variable-position peak enables you to eliminate your worst contribution to χ^2 . Reduction large and complicated to calculate.

Recommended remedy: Toy Monte Carlo.

Problem for Frequentists: Add a background $\mu=S+b$

b known*, μ measured through observing n events, S wanted

1. Find range for μ
2. Subtract b to get range for S

Examples:

See 5 events, background 1.2

95% Upper limit: 10.5 $\rightarrow$ 9.3

See 5 events, background 5.1

95% Upper limit: 10.5 $\rightarrow$ 5.4 ?

See 5 events, background 10.6

95% Upper limit: 10.5 $\rightarrow$ -0.1

This is technically correct. We are allowed to be wrong 5% of the time. But stupid. We know that the background happens to have a downward fluctuation but have no way of incorporating that knowledge

*We assume that the background is calculated correctly

Constrained parameters: 2 sad but true(ish) stories

Measure a mass

$M_X{}^2$=-2 ± 5 GeV

Or even

$M_X{}^2$=-5± 2 GeV

"$M_x{}^2$ lies between -7 and -3" with 68% confidence

?!

Counting Experiment

Expect 2.8 background events. See 0

Signal+background<2.3, so signal< -0.5 (at 90% CL)

?!

Similar problems

- Expected number of events must be non-negative
- Mass of an object must be non-negative
- Mass-squared of an object must be non-negative
- Higgs mass from EW fits must be bigger than LEP2 limit of 114 GeV

3 Solutions

- Publish a ‘clearly crazy’ result
- Use Feldman-Cousins or CL_S
- Switch to Bayesian analysis

Feldman Cousins Method

Works by attacking what looks like a different problem...

Also called* 'the Unified Approach'

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%)

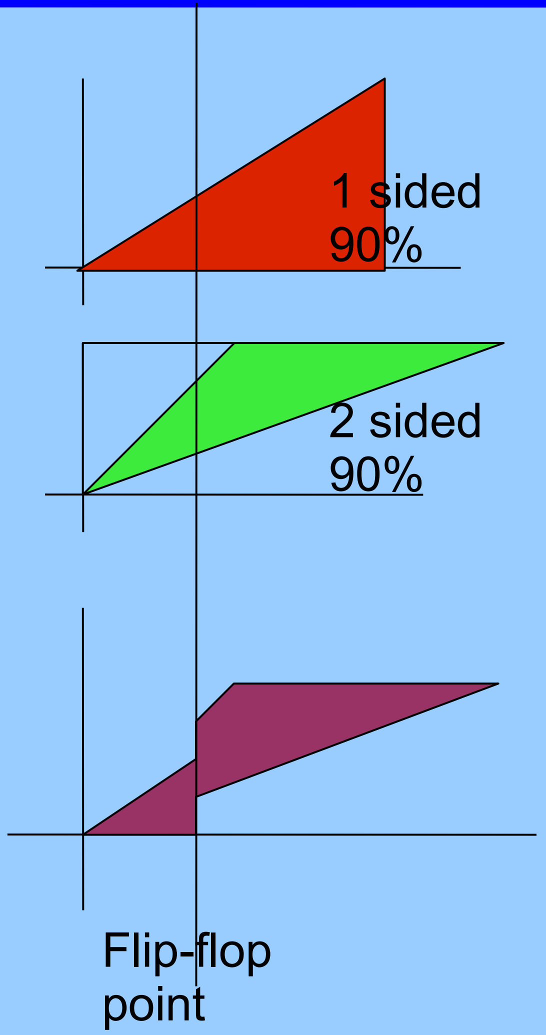
Observe 25 events? Quote two-sided limits

* by Feldman and Cousins, mostly

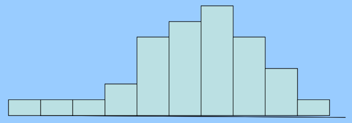
Feldman Cousins: μ=s+b

b is known. n is measured. s is what we're after

This is called 'flip-flopping and is BAD because it wrecks the whole idea of the Neyman confidence belt construction



Feldman Cousins: Ranking



First idea (almost right)

Sum/integrate over range of (s+b) values with highest probabilities for this observed n.

(advantage that this is the shortest interval)

Glitch: Suppose n small. (low fluctuation)

P(n;s+b) will be small for any s and never get counted

Instead: compare to 'best' probability for this n, at s=n-b or s=0 and rank on that number

Such a plot does an automatic ‘flip-flop’

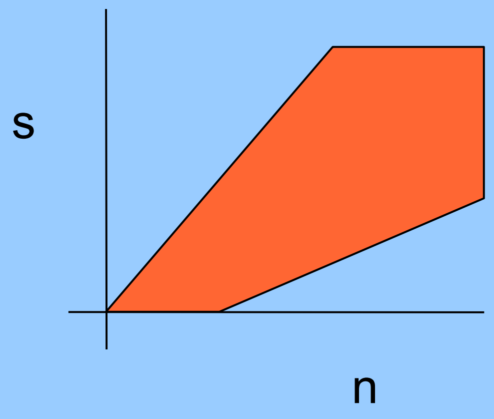
n~b single sided limit (upper bound) for s

n>>b 2 sided limits for s

How it works

Belt has to be computed for the appropriate value of background b. (Sounds complicated, but there is lots of software around)

As n increases, flips from 1-sided to 2-sided limits – but in such a way that the probability of being in the belt is preserved



Means that sensible 1-sided limits are quoted instead of nonsensical 2-sided limits!

Arguments against using Feldman Cousins

- Argument 1

It takes control out of hands of physicist. You might want to quote a 2 sided limit for an expected process, an upper limit for something weird

- Counter argument:

This is the virtue of the method. This control invalidates the conventional technique. In rare cases it is permissible to say "We set a 2 sided limit, but we're not claiming a signal"

Feldman Cousins: Argument 2

- Argument 2

If zero events are observed by two experiments, the one with the higher background b will quote the lower limit. This is unfair to hardworking physicists

- Counterargument

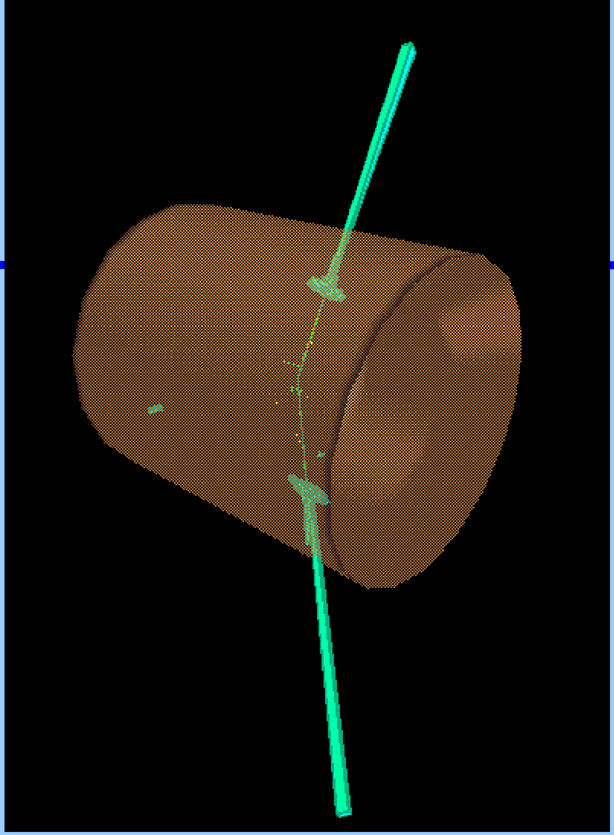
An experiment with higher background has to be ‘lucky’ to get zero events. Luckier experiments will always quote better limits. Averaging over luck, lower values of b get lower limits to report.

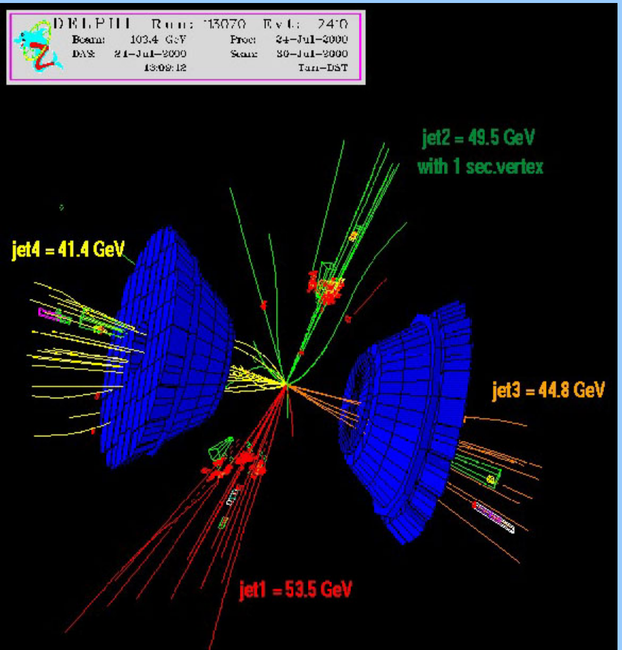
Example: you reward a good student with a lottery ticket which has a 10% chance of winning $100. A moderate student gets a ticket with a 1% chance of winning $200. They both win. Were you unfair?

The CL_s Technique

Used for Higgs searches by the combined LEP experiments.

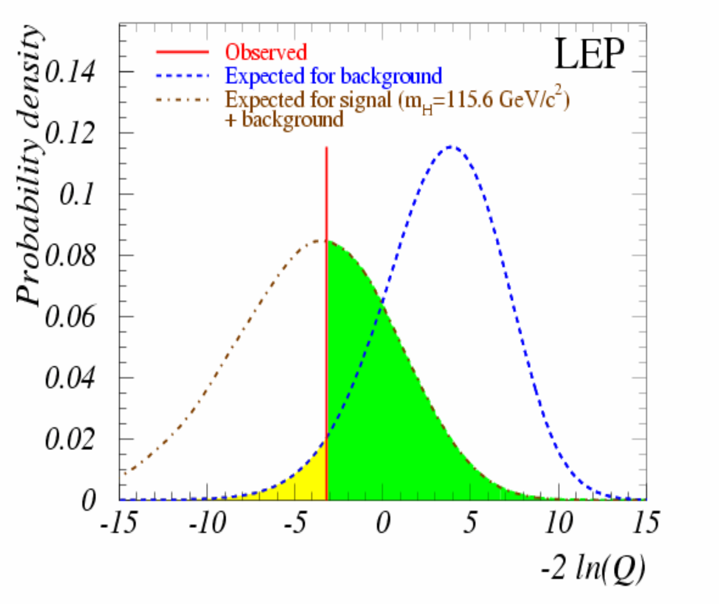
'Frequentist-motivated'





Different experiments selected events with Higgs hints

CL_s



Generalisation of Helene formula

Define some quantity Q. Could be number of events, or some more clever Higgsishness number. Larger values of Q imply a signal.

Standard frequentist CL numbers

CL_b=P(Q or less|b)

CL_{s+b}=P(Q or less|s+b)

Then take ratio – or difference of logs

$$CL_s=CL_{s+b}/CL_b$$

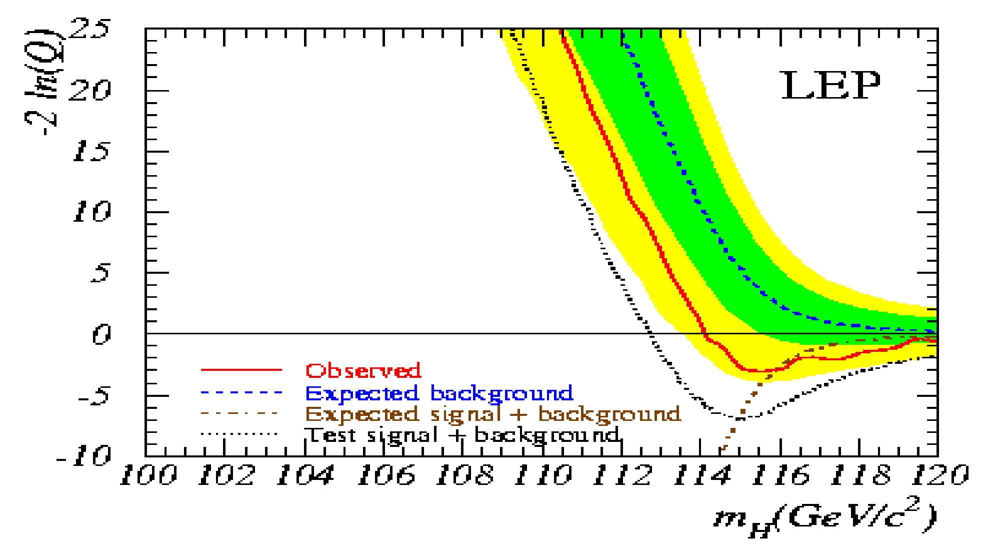
Used as confidence level (overcover). Optimise strategy using it and quote results

Yellow is 1-CL_b
Green is CL_{s+b} for given m_H

"One will be hard pressed to find a more robust frequentist-motivated presentation of results at the search frontier"
Alex Read:

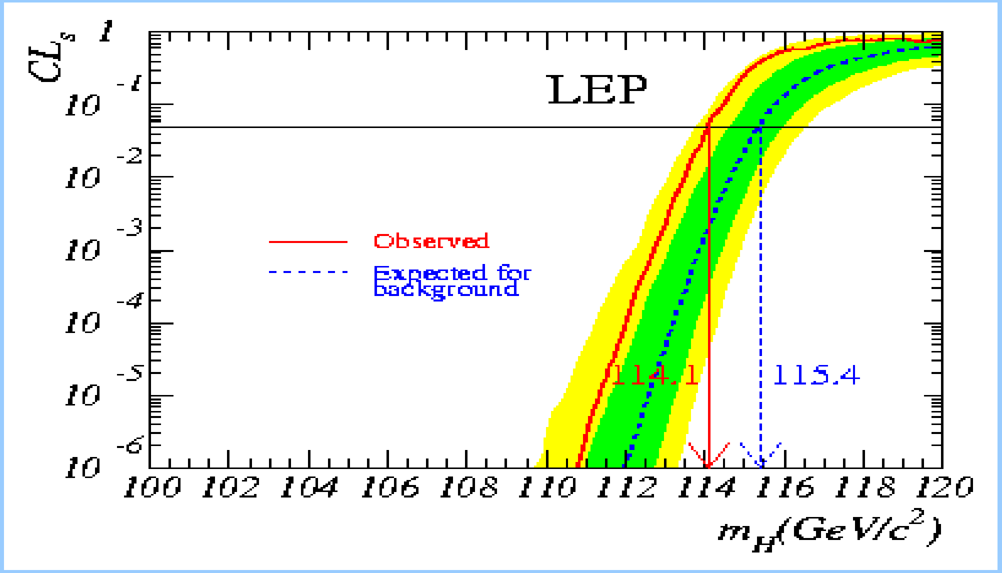
Results

(as of 2002)



Rule out M_H up to 114.1 GeV

(>114.1 GeV @ 95%)



Summary on CL_s

Used for several searches at LEP and elsewhere

Adaptive and sensible.

Frequentist but 'behaves like P(theory|Data)'

Well adapted to exclusion.

See Alex Read's talks at CERN and Durham workshops

Bayesian (Subjective) Probability

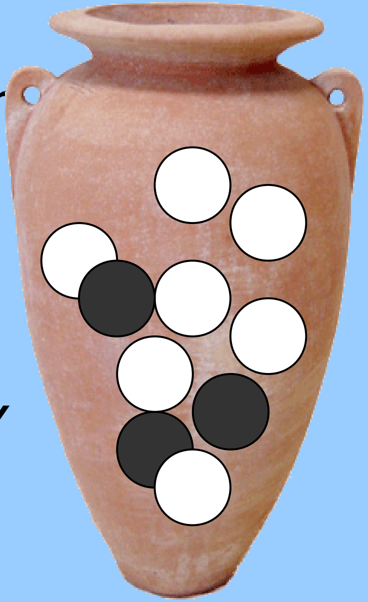
I can say:“The probability of rain tomorrow is 70%”

And I mean:

I regard 'rain tomorrow' and 'drawing a white ball from an urn containing 7 white balls and 3 black balls' as equally likely.

By which I mean:

If I were offered a choice of betting on one or the other I would be indifferent.



P(A) is a number describing my degree of belief in A

1=certain belief. 0=total disbelief

- A can be anything: rain, horses, existence of SUSY
- Is my P(A) is the same as your P(A). Subjective = unscientific?

Subjectivity check

What probability do you assign to the following:

- The Higgs will be seen at the LHC
- Obama will be re-elected
- SUSY will be seen at the LHC
- It will rain tomorrow
- The Standard Model is correct

Bayes' Theorem

General (uncontroversial) form

$$P(A|B)P(B) = P(A \,\&\, B) = P(B|A)\,P(A\ )$$

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

P(B) can be written $P(B|A)\,P(A) + P(B|\text{not } A)\,(1-P(A))$

Examples:

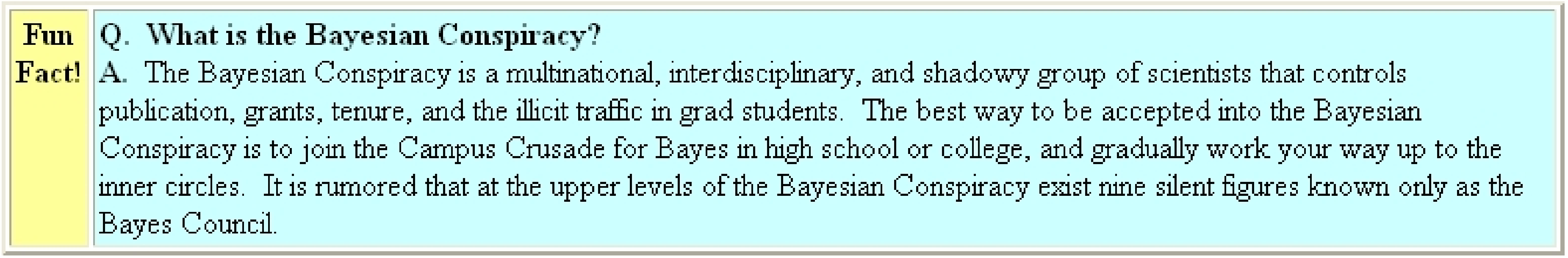
People $P(\text{Artist}|\text{Beard})=\frac{P(\text{Beard}|\text{Artist})}{P(\text{Beard})}\,P(\text{Artist})$

π /K Cherenkov counter $P(\pi|\text{signal})=\frac{P(\text{signal}|\ \pi)}{P(\text{signal})}\,P(\pi)$

0.9*0.5/(.9*.5+.01*.5)= 0.989

Medical diagnosis $P(\text{disease}|\text{symptom})=\frac{P(\text{symptom}|\text{disease})}{P(\text{symptom})}\,P(\text{disease})$

Misinformation abounds...



Fun Fact!	**Q. What is the Bayesian Conspiracy?** A. The Bayesian Conspiracy is a multinational, interdisciplinary, and shadowy group of scientists that controls publication, grants, tenure, and the illicit traffic in grad students. The best way to be accepted into the Bayesian Conspiracy is to join the Campus Crusade for Bayes in high school or college, and gradually work your way up to the inner circles. It is rumored that at the upper levels of the Bayesian Conspiracy exist nine silent figures known only as the Bayes Council.

http://yudkowsky.net/bayes/bayes.html

Bayes at work: modifying beliefs

Dr. A Sceptic thinks that Global Warming is probably a myth. P=10%

Data arrives showing loss of Antarctic ice coverage. Global warming said this would definitely happen (P=1). But it could happen as part of natural cyclical fluctuations (P=20%)

All numbers totally fictitious

Use Bayes Theorem

$$P_G' = \frac{P(melt|G)P_G}{P(melt|G)P_G + P(melt|\bar{G})\bar{P}_G} = \frac{0.1}{0.1 + 0.2\text{x}0.9} = 0.36$$

Priors and Posteriors

Can regard the function P(M) as a probability distribution a model parameter M confronting some result R

$$P(M)' = \frac{P(R|M)\,P(M)}{P(R)}$$

Prior distribution for M

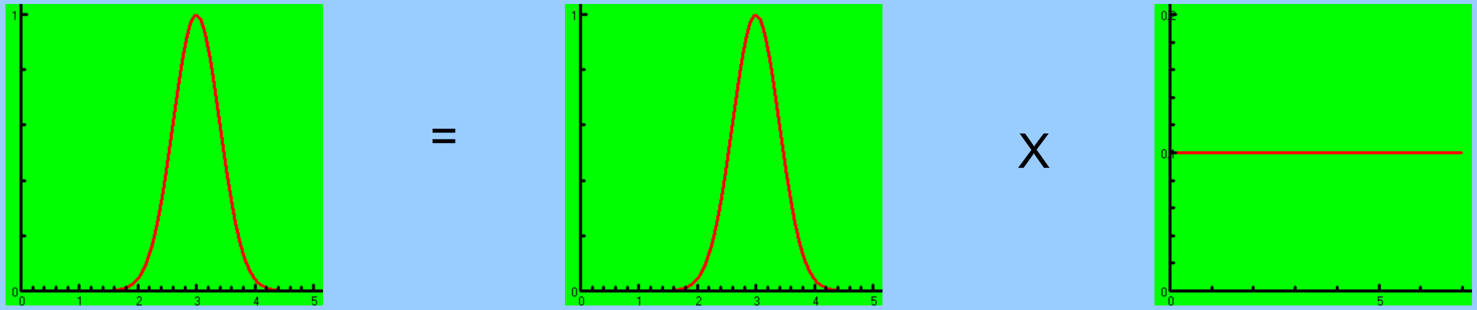
Posterior distribution for M

Probability distribution for R given M

distribution for R anyway

Measurements:Bayes at work

Result value x Theoretical ‘true’ value μ $P(\mu|x) \propto P(x|\mu)\, P(\mu)$



Prior is generally taken as uniform

Ignore normalisation problems

Construct theory of measurements – prior of second measurement is posterior of the first

$P(x|\mu)$ is often Gaussian, but can be anything (Poisson, etc)

For Gaussian measurement and uniform prior, get Gaussian posterior

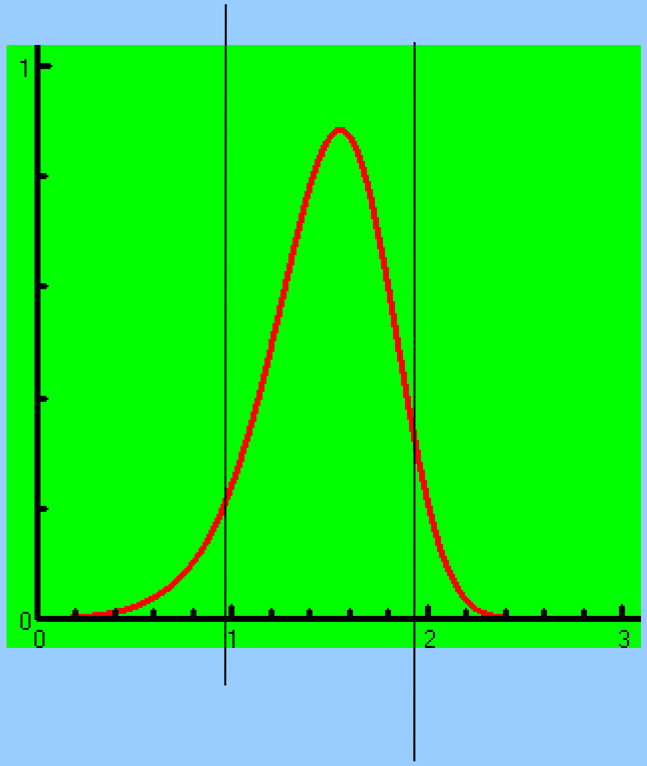
Bayesian Confidence Intervals

Trivial!

- Given the posterior P'(M|R) you choose a range [M_{lo},M_{hi}] for which

$$\int_{M_{lo}}^{M_{hi}} P'(M|R)\,dM = CL$$

Choice of strategies: central, upper limit lower limit, etc.



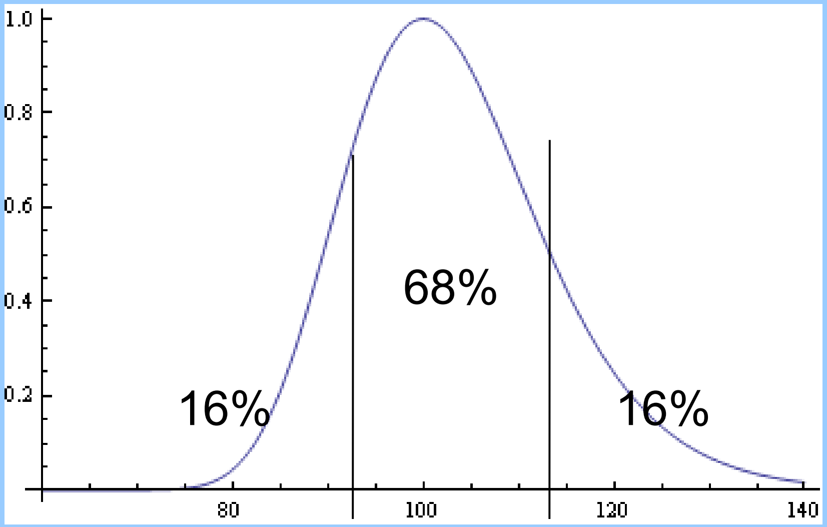
Pause for breath

For Gaussian measurements of quantities with no constraints/objective prior knowledge the same results are given by:

- Frequentist confidence intervals
- Bayesian posteriors from uniform priors

A frequentist and a Bayesian will report the same outcome from the same raw data, except one will say ‘confidence’ and the other ‘probability’. They mean something different but will never realise this.

Bayesian: Proportional Gaussian



Likelihood function

C exp(- ½(μ-100)^2/(0.1 μ)^2)

Integration gives C=0.03888

68% (central) limits

92.6 and 113.8

Frequentist and Bayesian approaches give different answers

Bayesian limits from small number counts

$P(r,\mu)=\exp(-\mu)\,\mu^{r}/r!$

With uniform prior this gives posterior for μ

Shown for various small r results

Read off intervals...

Upper limit from n events

$$\int_0^{\mu_{HI}} \exp(-\mu)\,\mu^n/n!\;d\mu = CL$$

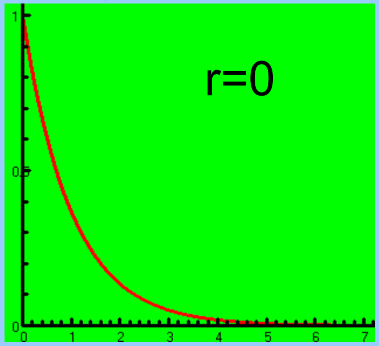
Repeated integration by parts:

$$\Sigma_0^n \exp(-\mu_{HI})\,\mu_{HI}^n/n! = 1-CL$$

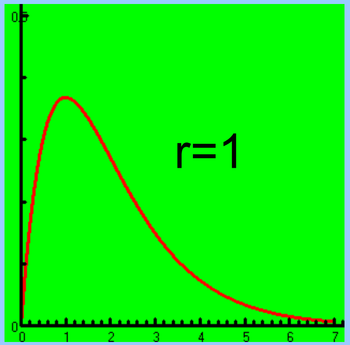
Same as frequentist limit

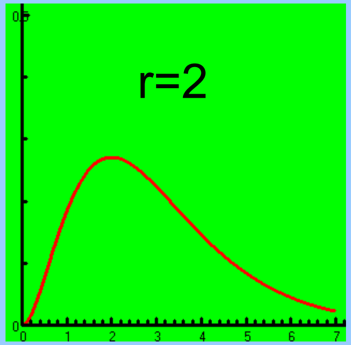
This is a coincidence! Lower Limit formula is not the same

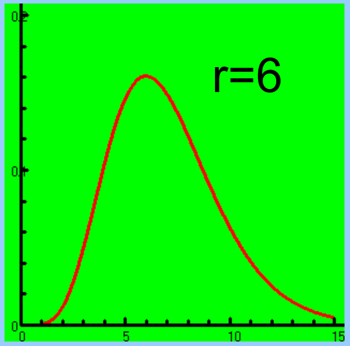
P(μ)



μ

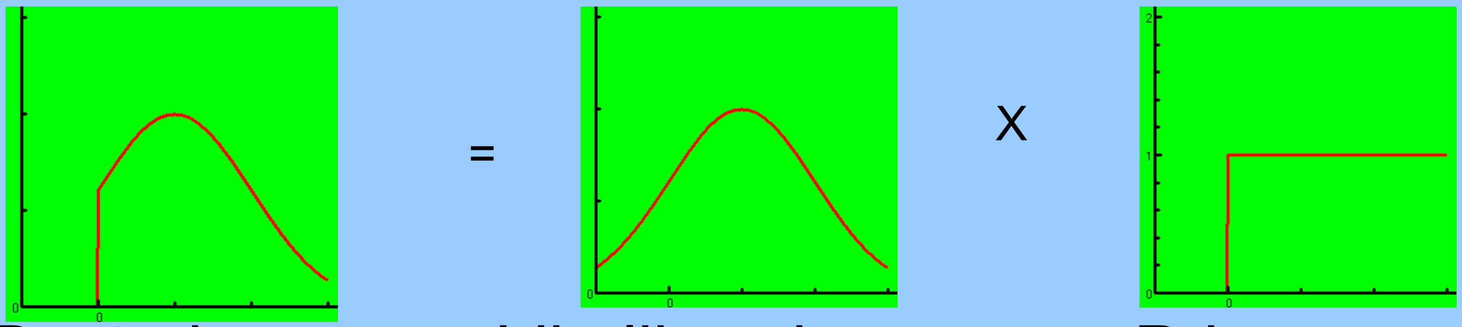






μ=S+b for Bayesians

- No problem!
- Prior for μ is uniform for S≥b
- Multiply and normalise as before



Posterior = Likelihood X Prior

Read off Confidence Levels by integrating posterior

Incorporating Constraints: Poisson

Work with total source strength (s+b) you know is not less than the background b

Need to solve

$$\alpha = \frac{\sum_0^n e^{-(s+b)}(s+b)^r / r!}{\sum_0^n e^{-b} b^r / r!}$$

Formula not as obvious as it looks.

Known as “the old PDG formula” or “Helene’s formula” or “that heap of crap”

Problem: the Uniform Prior

General usage: choose *P(a)* uniform in *a*

(principle of insufficient reason – actually usually laziness)

Often 'improper': $\int P(a)da = \infty$. Though posterior *P(a|x)* comes out sensible

BUT!

If *P(a)* uniform, $P(a^2)$, *P(ln a)* , $P(\sqrt{a})$.. are not

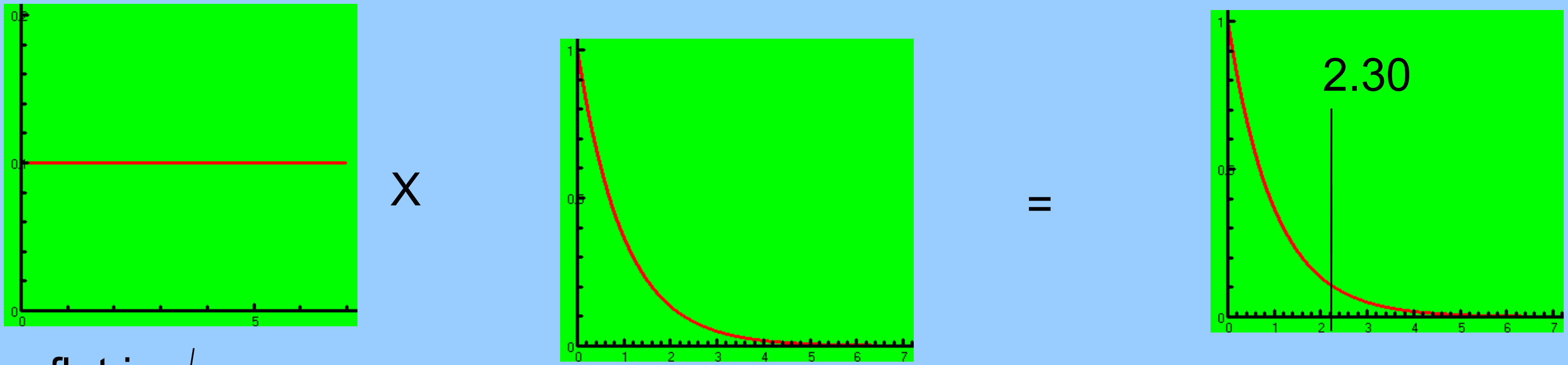
Insufficient reason not valid (unless *a* is 'most fundamental' – whatever that means)

Statisticians handle this: check results for 'robustness' under different priors

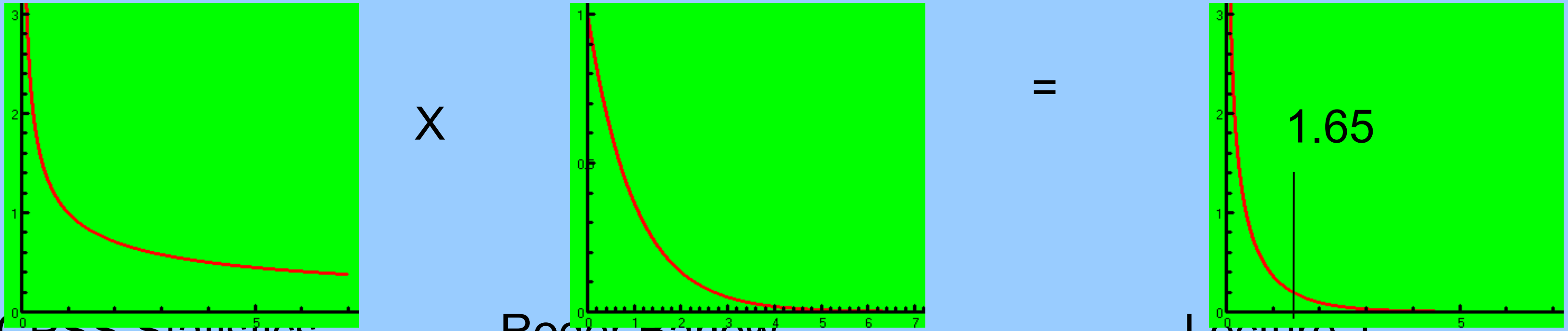
Result depends on Prior

Example: 90% CL Limit from 0 events

Prior flat in μ

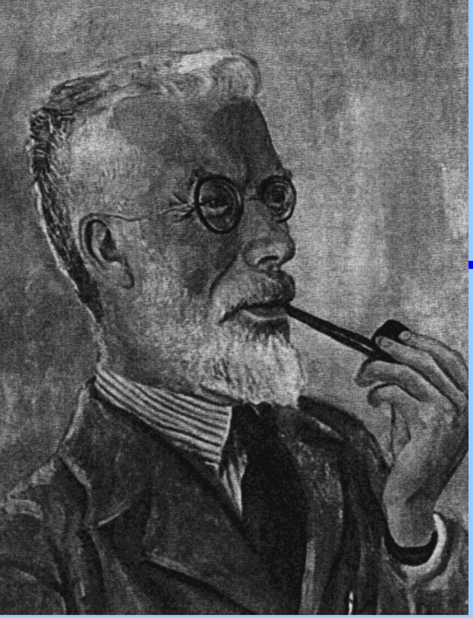


Prior flat in √μ



Robustness

- Result depends on chosen prior
- More data reduces this dependence
- Statistical good practice: try several priors and look at the variation in the result
- If this variation is small, result is robust under changes of prior and is believable
- If this variation is large, it's telling you the result is meaningless

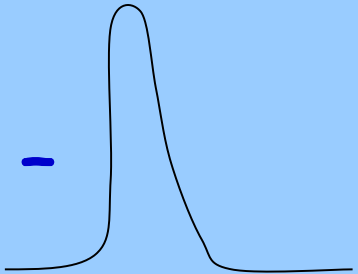


Fisher Information

An informative experiment is one for which a measurement of *x* will give precise information about the parameter *a*.

Quantify: $I(a) = -\langle \partial^2 \ln L / \partial a^2 \rangle$

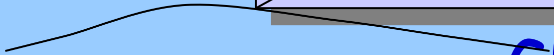
(Second derivative – curvature)



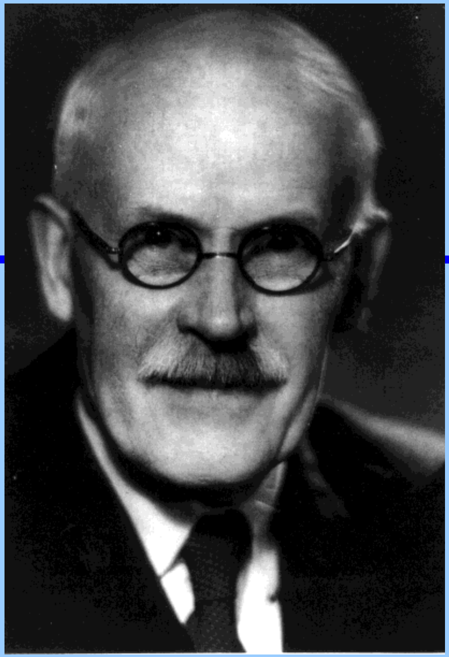
P(x,a): everything

$P(x)|_a$ is the pdf

$P(a)|_x$ is the likelihood L(a)



Jeffreys' Prior



A prior may be uniform in *a* – but if *I(a)* depends on *a* it's still not 'flat': special values of *a* give better measurements

Transform $a \rightarrow a'$ such that *I(a')* is constant. **Then** choose a uniform prior

- location parameter, uniform prior OK
- scale parameter – *a'* is *ln a.* prior *1/a*
- Poisson mean – prior *1/√a*

Why didn't it catch on?

It is 'objective' in the sense that everyone can agree on it. But they don't.

- It's more work than a uniform prior
- There are cases where it diverges and gives posterior functions that can't be normalised
- It does not work in more than one dimension (valiant attempts are being made to do this generalisation, under the name of Reference Priors)
- It depends on the form of L(R,M) which depends on the experiment. If you have an initial degree-of-belief prior function for (say) the Higgs mass, that should not depend on the measurement technique

Frequentist versus Bayesian?

Two sorts of probability – totally different.

Rivals? Religious differences?

Particle Physicists tend to be frequentists. Cosmologists tend to be Bayesians

No. Two different tools for practitioners

Important to be aware of the limits and pitfalls of both

Frequentist versus Bayesian?

Statisticians do a lot of work with Bayesian statistics and there are a lot of useful ideas. But they are careful about checking for robustness under choice of prior.

Beware snake-oil merchants in the physics community who will sell you Bayesian statistics (new – cool – easy – intuitive) and don't bother about robustness.

Use Frequentist methods when you can and Bayesian when you can't (and check for robustness.) But ALWAYS be aware which you are using.

Bayesian pitfall (1): Unitarity triangle

Measure CKM angle α by measuring B $\rightarrow\rho\rho$ decays (charged and neutral, branching ratios and CP asymmetries). 6 quantities.

Many different parametrisations suggested

Uniform priors in different parametrisations give different results from each other and from a Frequentist analysis (according to CKMfitter: disputed by UTfit)

For a complex number $z=x+iy=re^{i\theta}$ a flat prior in x and y is not the same as a flat prior in r and θ

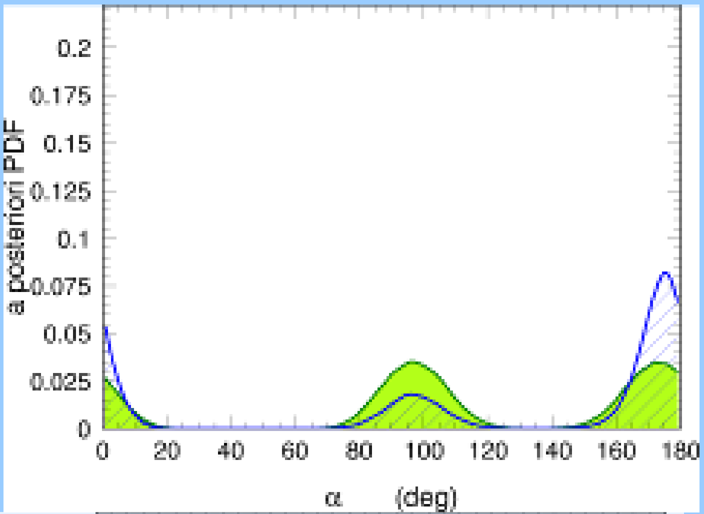
Results

Bayesian
Parametrise Tree and Penguin amplitudes

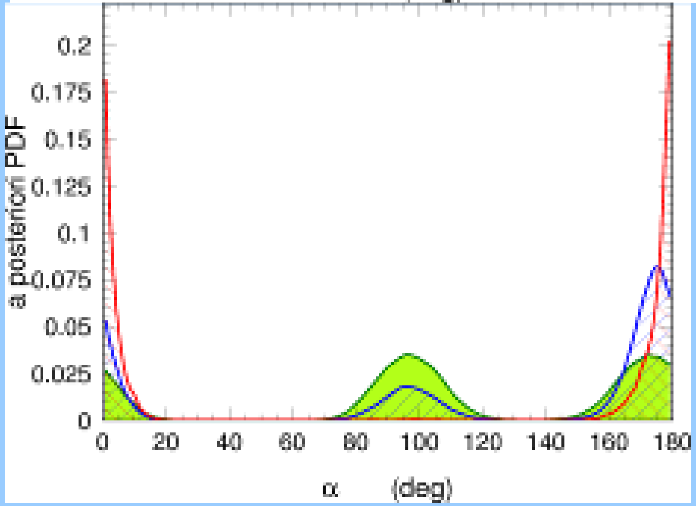
$$A^{+-} = -Te^{-i\alpha} + Pe^{i\delta_P}$$

$$A^{+0} = -\frac{1}{\sqrt{2}}e^{-i\alpha}\left(T + T_C e^{i\delta_{T_C}}\right)$$

$$A^{00} = -\frac{1}{\sqrt{2}}\left(e^{-i\alpha}T_C e^{i\delta_{T_C}} + Pe^{i\delta_P}\right)$$

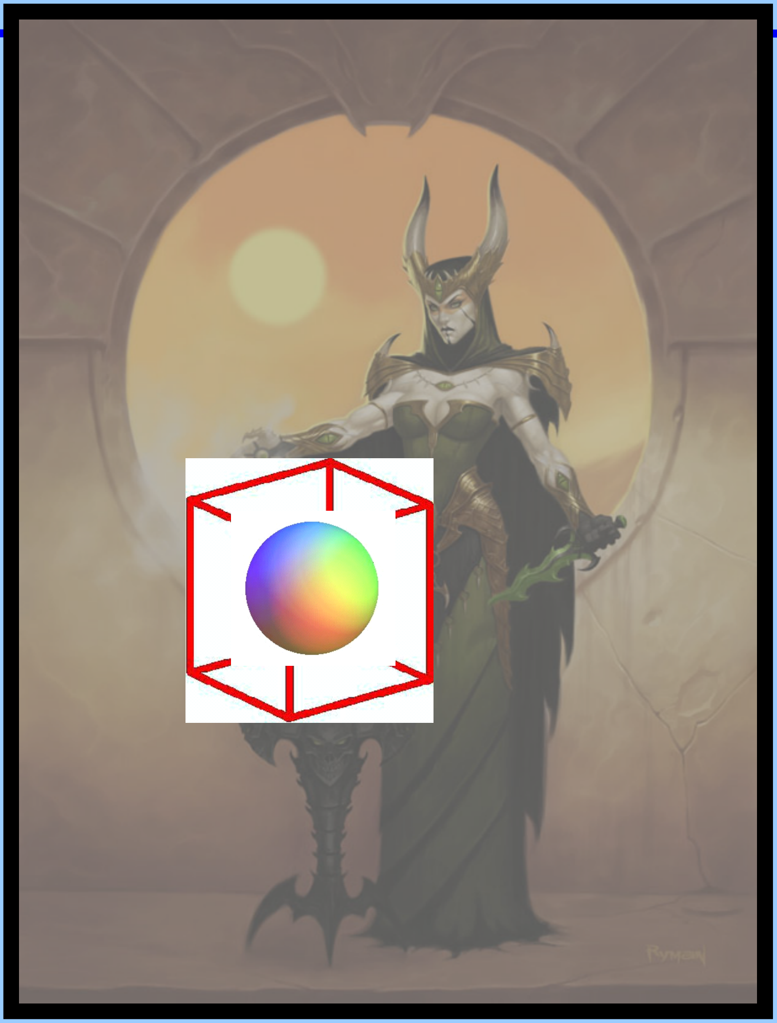


Bayesian
3 Amplitudes:
3 real parts, 3 Imaginary parts



Interpretation

- Removing all experimental info gives similar *P(α)*
- The curse of high dimensions is at work



Uniformity in *x,y,z* makes *P(r)* peak at large *r*

This result is not robust under changes of prior

Example – Efficiencies

CDF statistics group (Joel Heinrich) looking at problem of estimating signal cross section S in presence of background and efficiency.

$$N = \varepsilon S + b$$

Efficiency and Background from separate calibration experiments (sidebands or MC).

Everything done using Bayesian methods with uniform priors and Poisson statistics formula. Calibration experiments use uniform prior for *ε* and for *b*, yielding posteriors used for *S*

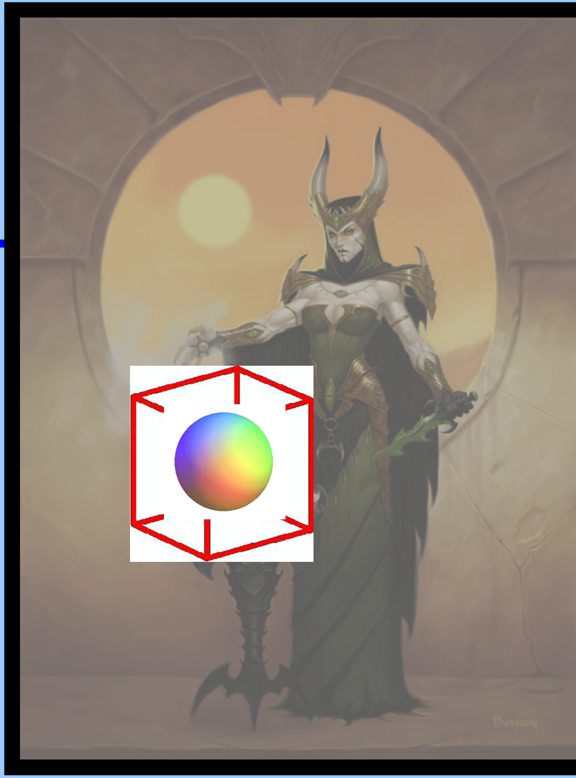
$$P(N|S)=({}^{1}/_{N!})\iint e^{-(\varepsilon S+b)}\ (\varepsilon S+b\ )^{N}\ P(\varepsilon)\ P(b)\ d\varepsilon\ db$$

Check coverage – all fine

Partition into classes (e.g. different run periods)

Coverage falls!

The curse strikes again



Uniform prior in ε: fine

Uniform prior in $\varepsilon_1, \varepsilon_2 \ldots \varepsilon_r$

$\rightarrow \varepsilon^{r-1}$ prior in total ε

Prejudice in favour of high efficiency

Signal size downgraded

Happy ending

Effect avoided by using Jeffreys' Priors instead of uniform priors for *ε* and *b*

Not uniform but like *1/ε, 1/b*

Uniform prior in S is not a problem – but maybe should consider 1/√S?

Coverage (a very frequentist concept) is a useful tool for Bayesians

Summary

Probability

- Frequentist
  - Confidence Levels
  - Small numbers (0 → <3 @ 95% CL)
  - Significance and p-values
  - Goodness of fit and χ^2
  - Problems with constraints
- Bayesian
  - Usage
  - Ambiguity of 'uniform prior'

Conclusions

Bayesian Statistics are

- Illuminating
- Occasionally the only tool to use
- Use with care: Results depend on choice of prior/choice of variable. Always check for robustness by trying a few different priors. Real statisticians do

If you're integrating the likelihood you are a Bayesian. I hope you know what you're doing.

Be suspicious of anything you don't understand

But always know what you are doing and say what you are doing.

Further reading

- The Particle Data Book
- Textbooks by Glen Cowan, Louis Lyons, Bohm and Zech, R.B.
- “Recommended Statistical Procedures for BaBar” BAD 318
- PHYSTAT proceedings (all Ed. Louis Lyons):
  - CERN 2000-05
  - Durham 2002 IPPP 02/39
  - SLAC 2003 SLAC-R-703
  - Oxford 2005 “Statistical problems in Particle Physics”, Imperial College Press (2006)