Frequentist Confidence Limits and Intervals

Roger Barlow

SLUO Lectures on Statistics

August 2006

Confidence Intervals

A common part of the physicist's toolkit

Especially relevant for results which are basically null

- Upper limits for rare processes BR < 10^{-7} @ 90% CL
- Lower limits for exotic masses M_X>1.2 TeV @90% CL

(though a conventional measurement is actually a 68% central confidence interval)

Trade-off confidence interval and confidence limit

Example…

BR($B_s \rightarrow \mu\mu$) < 1.0×10^{-7} @ 95% CL
< 8.0×10^{-8} @ 90% CL

BR($B_d \rightarrow \mu\mu$) < 3.0×10^{-8} @ 95% CL
< 2.3×10^{-8} @ 90% CL

CDF:
at
ICHEP06

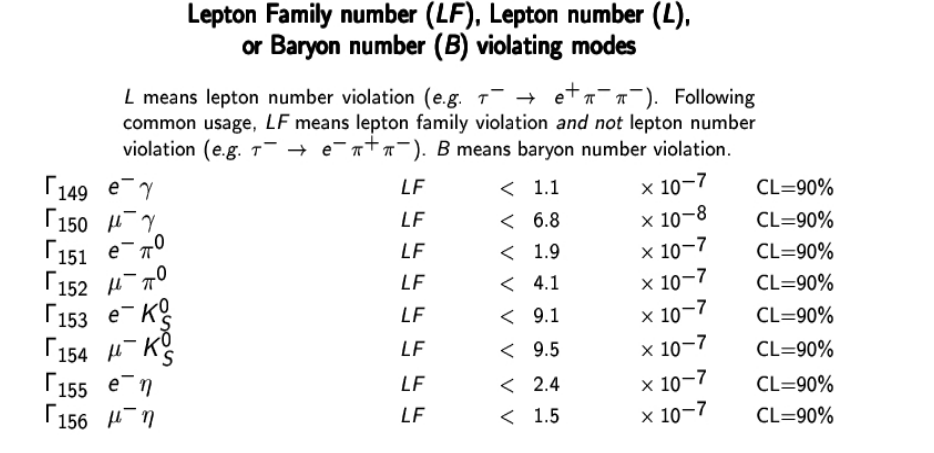
What does it mean?

Not just “the probability that the result is true”

52 LFV violating tau decays

52 confidence limits at 90%

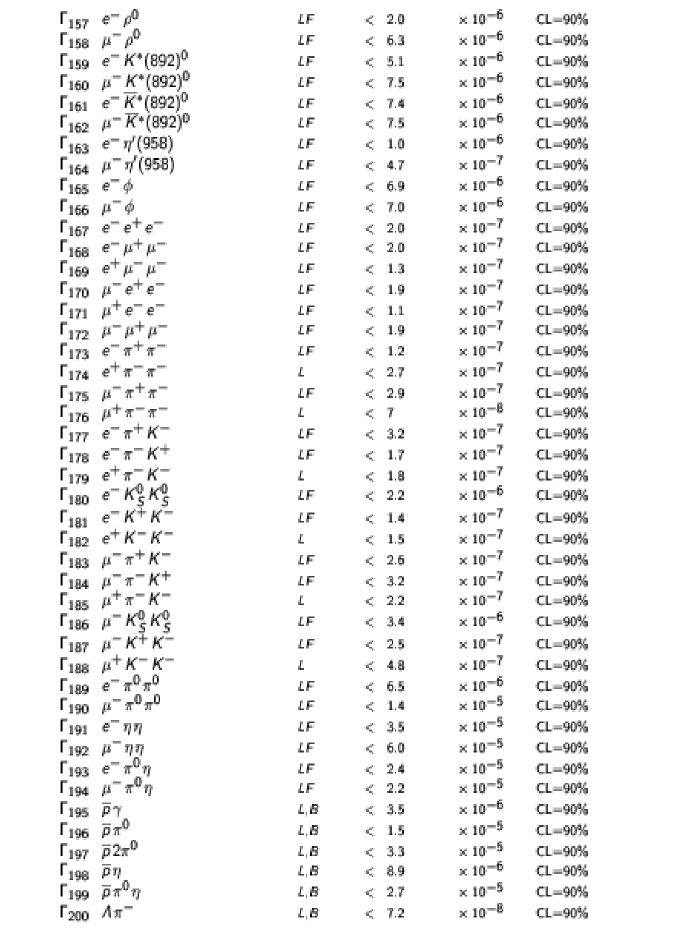
Anyone believe ~5 of these limits are exceeded?



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%

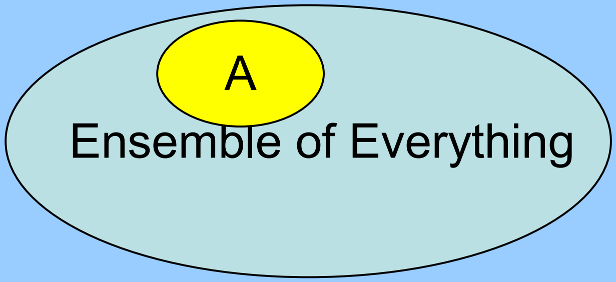
What do we mean by “@90% CL”?

- Confidence Levels are not probabilities for results
- But they are linked to probability. Need to revisit what we mean by that.

Probability

The probability of A is the large N limit fraction of cases in which A is true

$$P(A) = Limit_{N\to\infty} \frac{N_A}{N}$$



Lots of examples:

- Toss a coin: P_{head}=0.50
- J/ψ decays: $P_{\mu\mu}$=5.9%

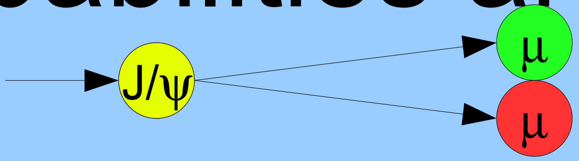
The standard (frequentist) definition. Taught at school. It has (i) an interesting property and (ii) an interesting limitation – not taught at school.

A property: There can be many 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
  - German insured male racing drivers
  - …

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

But surely probabilities are hard numbers…?



- The probability of a J/ψ decaying to μμ is 5.9% It's in the PDG so it must be true…
- Actually if we take the J/ψ particles in our BaBar data, the probability of a μμ decay is more than 5.9% (due to our trigger and selection)
- If a J/ψ is produced in p-Be, the probability of a μμ decay is less than 5.9% (it may interact with another nucleon)

Even in the PDG pages: the probability is 5.9% for a particular ensemble. For other ensembles it is something else

A 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 (Wednesday). There is NO ensemble. P(rain) is either 0/1 =0 or 1/1 = 1

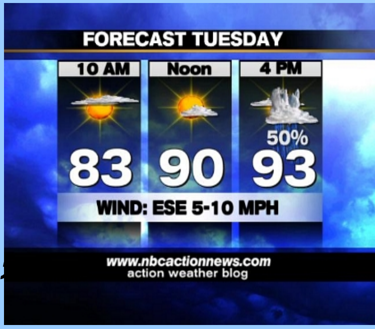
Strict frequentists cannot say 'It will probably rain tomorrow'.

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 70% are true.

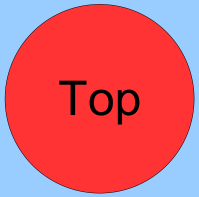
(E.g. Weather forecasts with a verified track record)

The Ensemble matters

- P(Rain)=50% and P(Rain)=90% can both be true
- P(Rain)=90% and P(fine)=90% can both be true

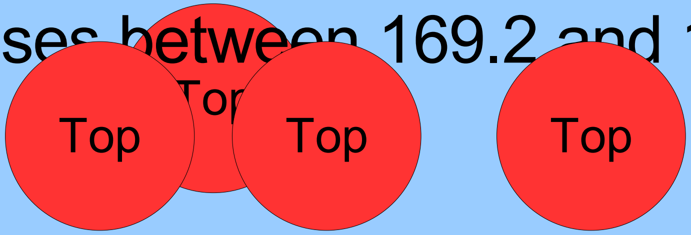
Hopfully all suitably softened up and confused. Back to some physics.

$$M_{top} = 174.3 \pm 5.1 \quad GeV/c^2$$



What does this mean?

- 68% of top quarks have masses between 169.2 and 179.4 GeV



WRONG. All top quarks have the same mass

- The probability of M_{top} being in the range 169.2-179.4 GeV is 68%.

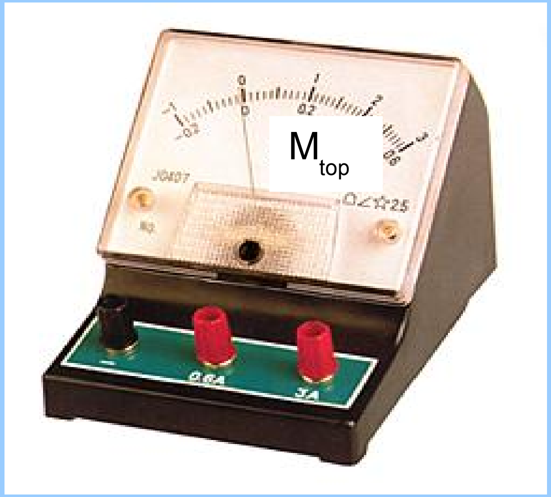
WRONG It either is or it isn't. P is 0 or 1

- M_{top} has been measured to be 174.3 GeV using a technique which has a 68% probability of being within 5.1 GeV of the true result

RIGHT.

What we mean by: $M_{top} = 174.3 \pm 5.1 \quad GeV / c^2$

The statement "M$_{top}$ is in the range 169.2-179.4 GeV" has a 68% probability of being true



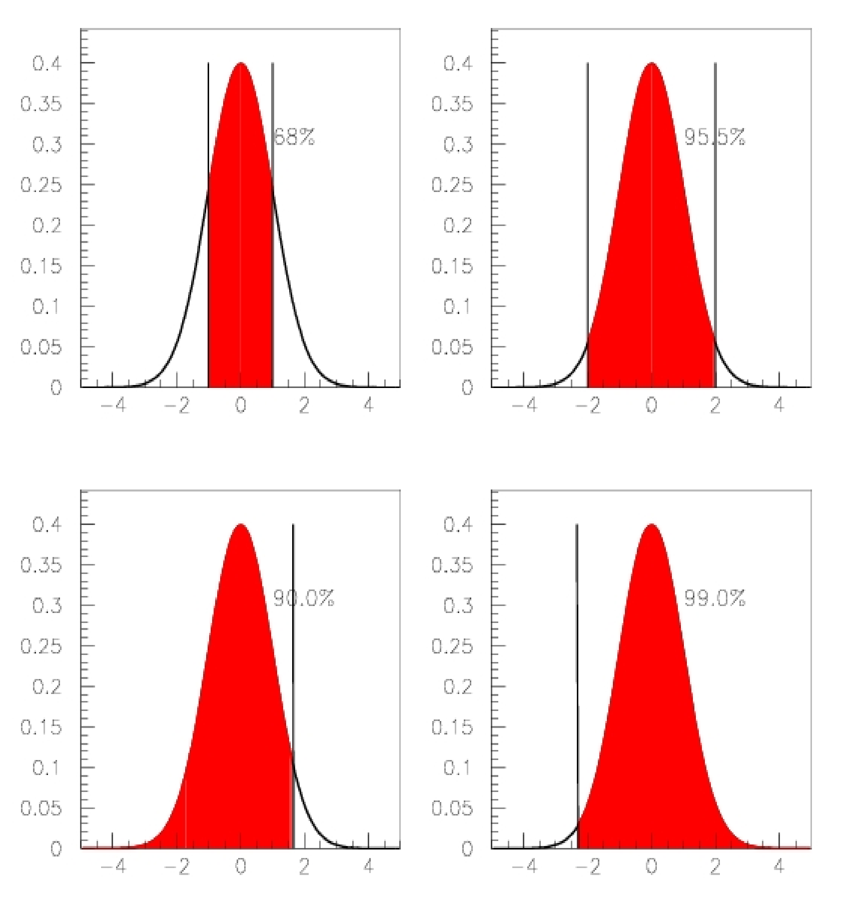
We make the statement "M$_{top}$ is in the range 169.2-179.4 GeV" with 68% confidence

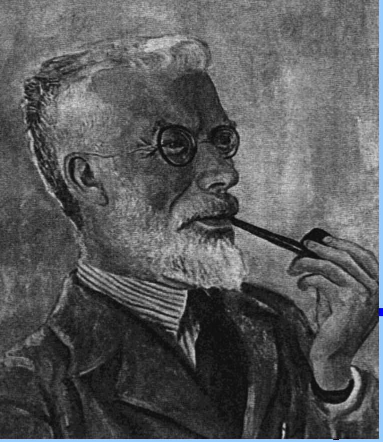
If we repeated the experiment many times, we would get different ranges. These are the ensemble of statements. They would bracket the true value in 68% of all cases.

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,...)





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

A bit more complicated

$$M_x = 100 \text{ GeV} \pm 10 \text{ GeV}$$

means M_x lies between 90 and 110 (@68% CL)

Now take M measured by a proportional Gaussian:

$$M_x = 100 \text{ GeV} \pm 10\%$$

A bit more than 1σ up from 90 (± 9)

A bit less than 1σ down from 110 (± 11)

$M_{upper} - 0.1 \text{ x } M_{upper} = 100$

$M_{lower} + 0.1 \text{x } M_{lower} = 100$

Range is 90.909 - 111.111

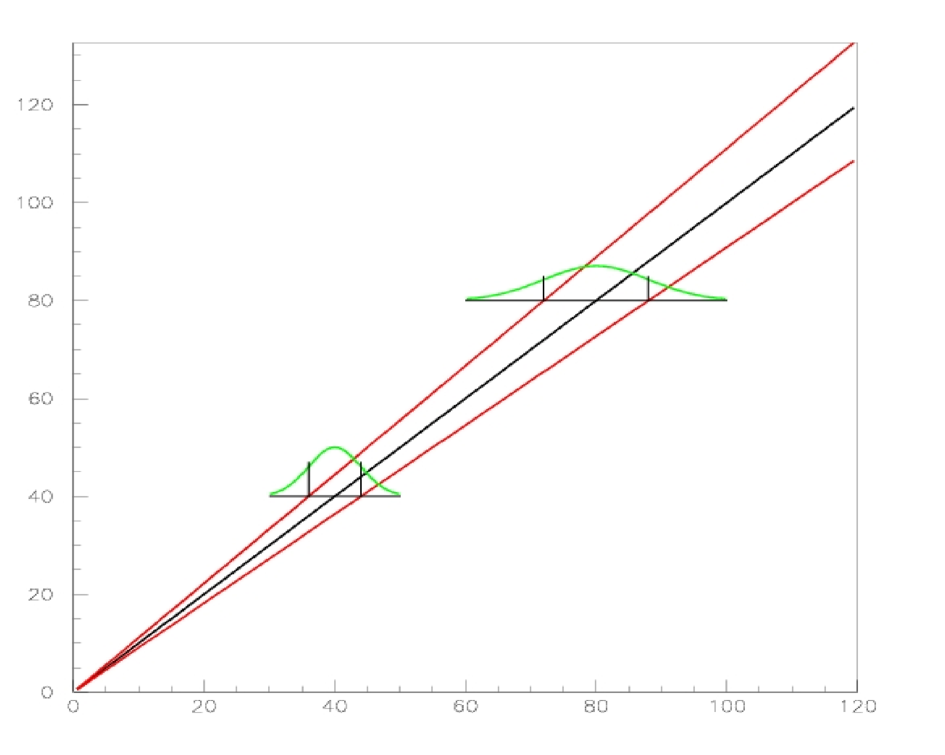
Confidence Belts

Neyman construction

1: Construct horizontally

2: Read Vertically

Works for any (reasonable) pdf P(x;μ)



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

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

λ=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%
.....	...

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

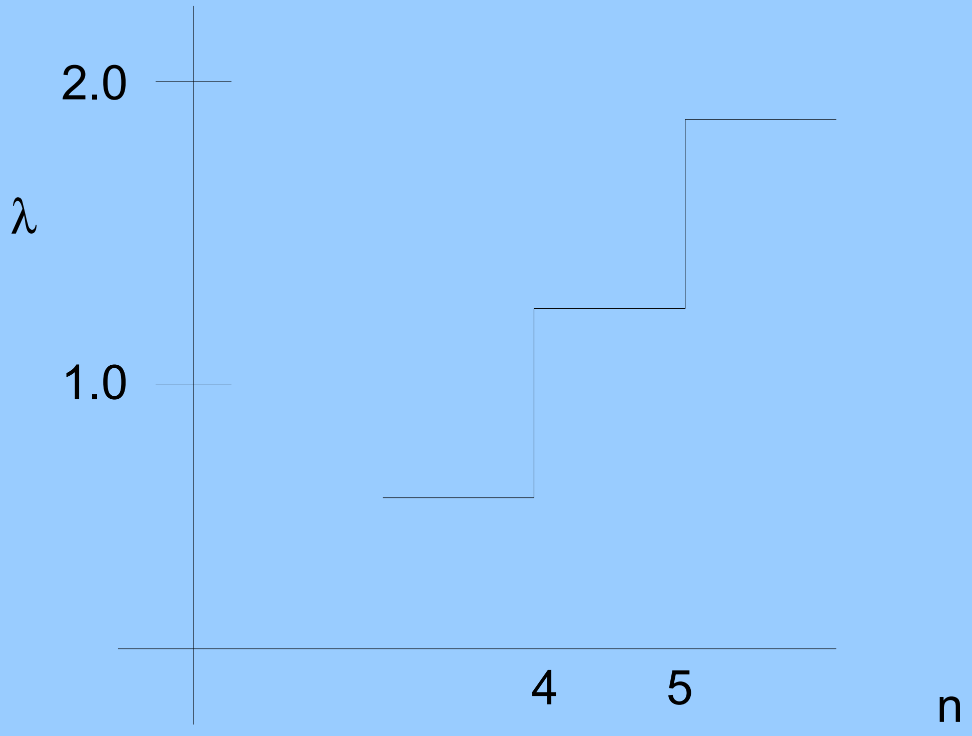
Play safe: include 3

If the true value is 1.1, or less, the probability of getting a result of 4 counts, or more, is only 5%, or less

Staircase

For λ from 0.8177 to 1.3663, P(0,1,2,3)>95%

from 1.3663 to 1.9702, P(0,1,2,3,4)>95%



Given n, quote highest λ for lower limit

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

Coverage

How often will your limit statements be true?

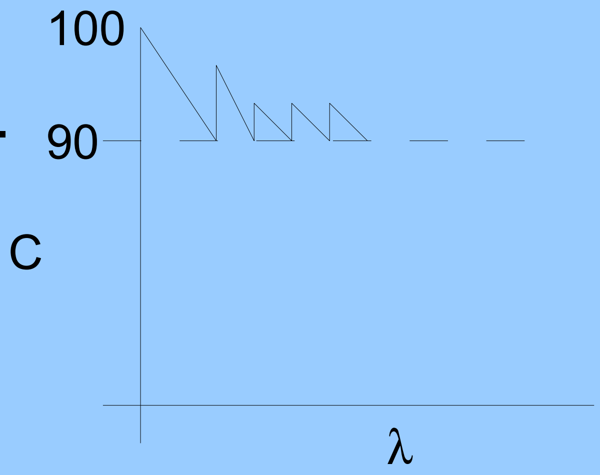
Should be same as CL, surely?

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

Coverage is a function of λ (etc)

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

It should never undercover (by construction)



Constrained parameters: 2 sad but true(ish) stories

Measure a mass

M_X=-2 ± 5 GeV

Or even

M_X=-5± 2 GeV

"M_x 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)

?!

What's happened? 2 Views

Nothing has gone wrong

You know that (up to)10% of your 90% CL statements can be wrong. This is one of them.

Indeed, you should publish this to avoid reporting bias

There are constraints on the parameters (Masses are non-negative. So are cross-sections.)

There is no way to input this information into the statistical apparatus.

We are not going to publish results that are manifestly wrong

This is broken and needs fixing

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: N=s+b

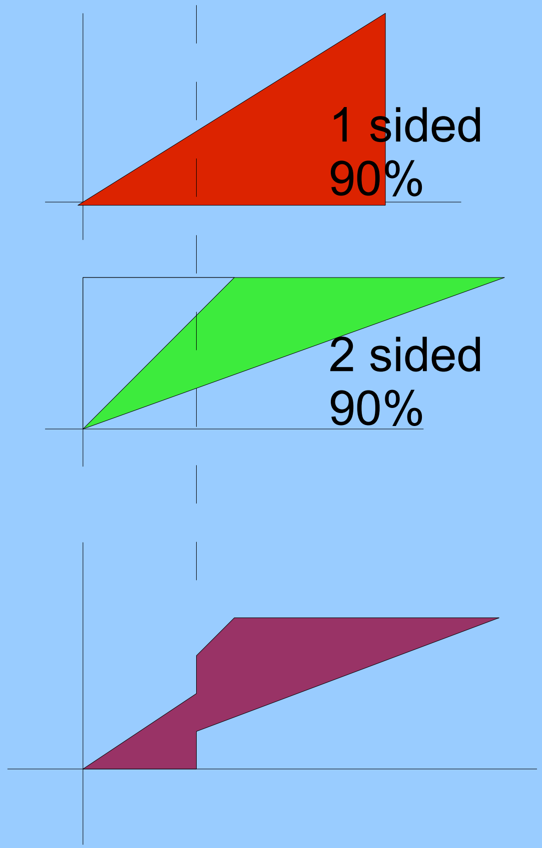
b is known. N is measured. s is what we're after

This is called 'flip-flopping' and BAD because is wrecks the whole design of the Confidence Belt

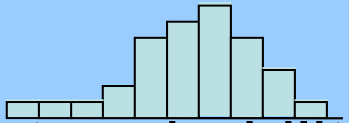
Suggested solution:

1) Construct belts at chosen CL as before

2) Find new ranking strategy to determine what's inside and what's outside



Feldman Cousins: Ranking



First idea (almost right)

Sum/integrate over outcomes with highest probabilities

(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

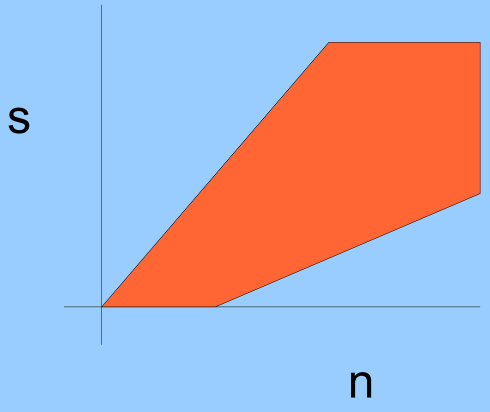
N~b single sided limit (upper bound) for s

N>>b 2 sided limits for s

How it works

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. The physicist can use their discretion over the CL. 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

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 $10. A moderate student gets a ticket with a 1% chance of winning $20. They both win. Were you unfair?

Summary

- A frequentist result "@ x% CL" is a member of an ensemble of results of which (at least) x% are true.
- There is a lot of choice in what to quote.
- The formalism enables us to interpret many results sensibly and present and discuss them
- Has problems with constrained quantities, but there are solutions