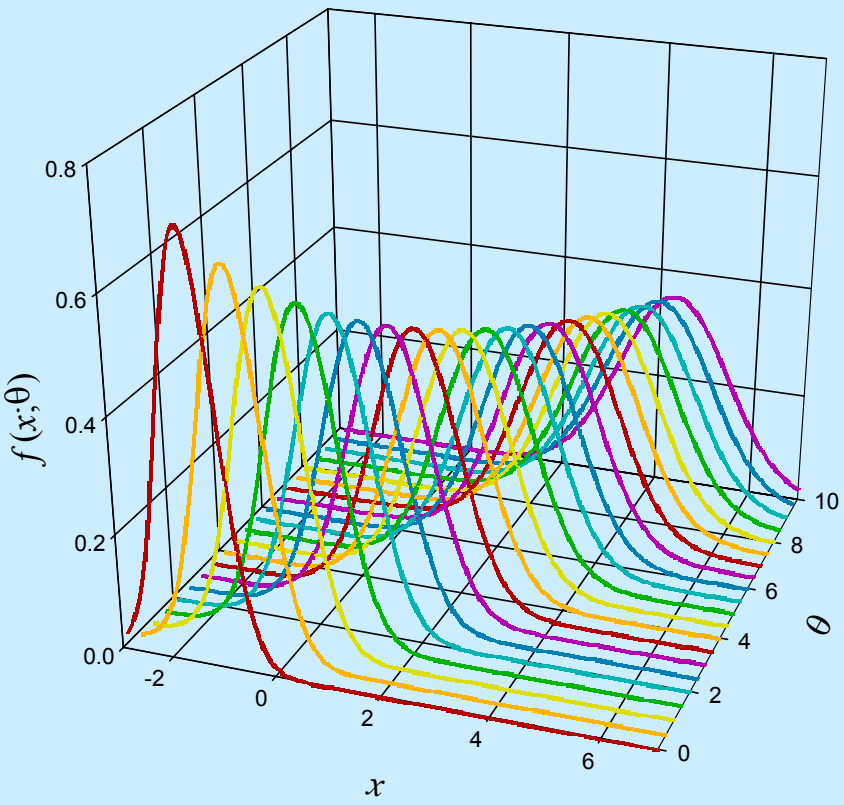
Credibility of confidence intervals

Dean Karlen / Carleton University

Advanced Statistical Techniques
in Particle Physics
Durham, March 2002

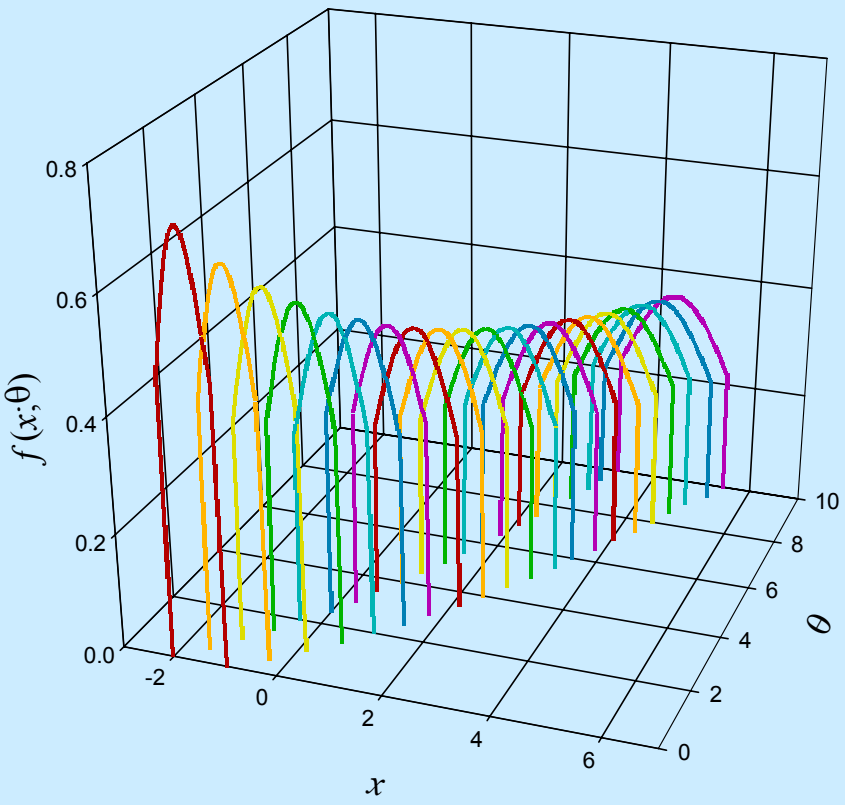
Classical confidence intervals

- Classical confidence intervals are well defined, following Neyman's construction:



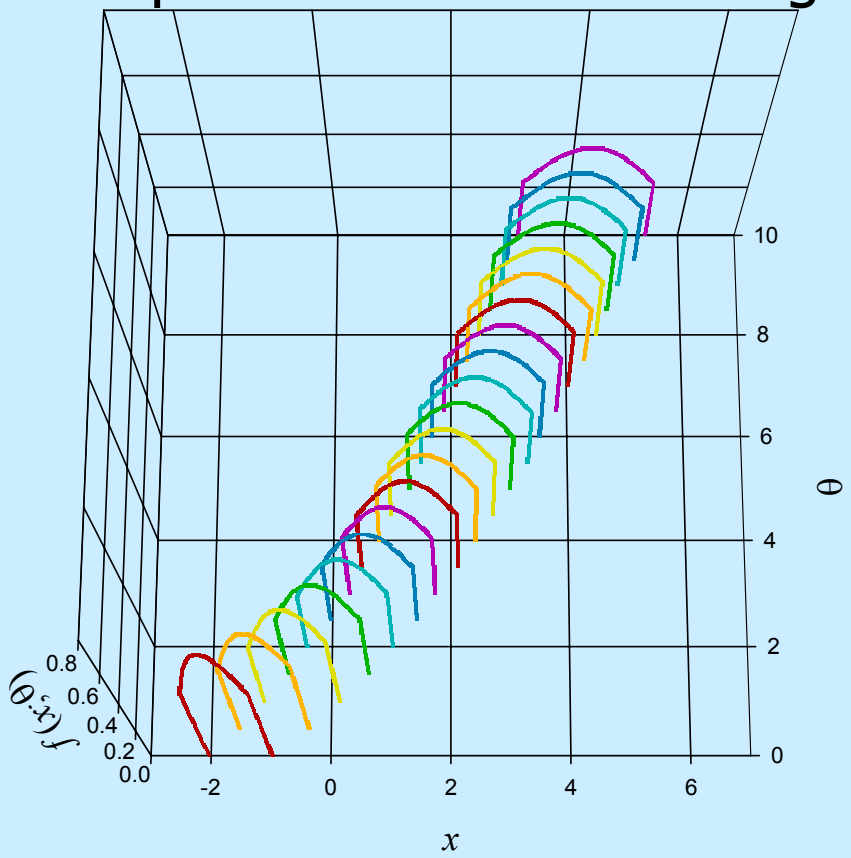
Classical confidence intervals

- select a portion of the pdfs (with content α)
  - for example the 68% central region:



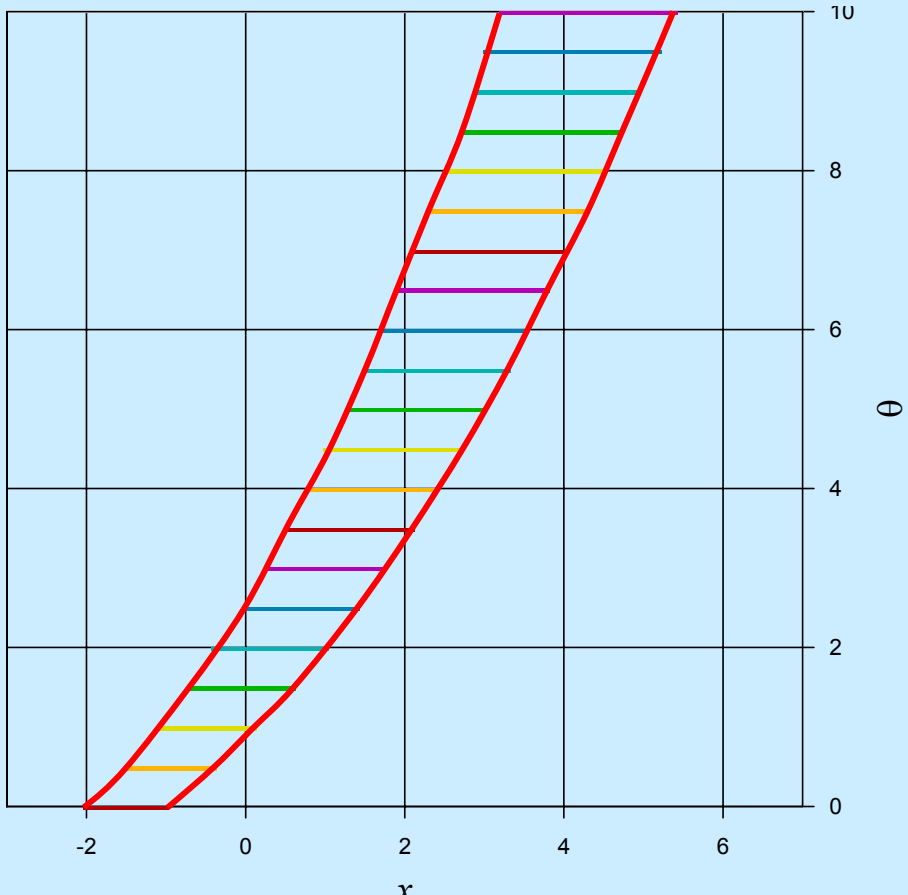
Classical confidence intervals

- select a portion of the pdfs (with content α)
  - for example the 68% central region:



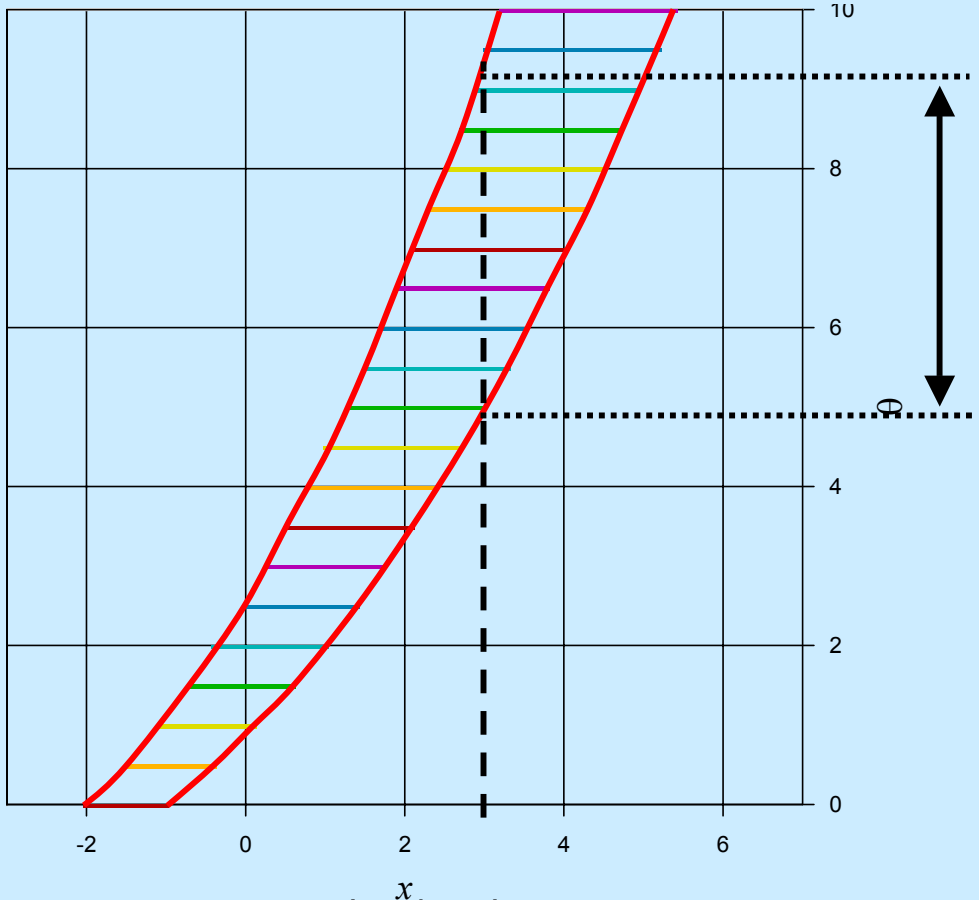
Classical confidence intervals

- gives the following confidence belt:



Classical confidence intervals

- The (frequentist) probability for the random interval to contain the true parameter is α



Problems with confidence intervals

- Misinterpretation is common, by general public and scientists alike…
  - Incorrect: α states a "degree of belief" that the true value of the parameter is within the stated interval
  - Correct: α states the "relative frequency" that the random interval contains the true parameter value
  - Popular press gets it wrong more often than not
    - "The probability that the Standard Model can explain the data is less than 1%."

Problems with confidence intervals

- People are justifiably concerned and confused when confidence intervals
  - are empty; or
  - reduce in size when background estimate increases (especially when n=0); or
  - turn out to be smaller for the poorer of two experiments; or
  - exclude parameters for which an experiment is insensitive

"confidence interval pathologies"

Source of confusion

- The two definitions of probability in common use go by the same name
  - relative frequency: probability
  - degree of belief: probability
- Both definitions have merit

- Situation would be clearer if there were different names for the two concepts
  - proposal to introduce new names is way too radical
- Instead, treat this as an education problem
  - make it better known that two definitions exist

A recent published example…

[redacted] Collaboration / Physics Letters B 504 (2001) 218–224

4 events selected, background estimate is 0.34 ± 0.05

Therefore, the total sample background of tau-like events generated by charm or interactions is 0.34 ± 0.05. The Poisson probability of the background fluctuating to the signal level is 4.0×10^{-4}.

frequency

degree of belief

The probability that the four events are from background sources is 4×10^{-4}, and we conclude that these events are evidence that τ neutrino charged current interactions have been observed.

And an unpublished one…

ii) What is the Higgs discovery potential if LEP operates in 2001?

Although the statistical significance would suggest a probability of only about 0.2% that the present excess is due to a background fluctuation, the committee considers the conservative likelihood of a Higgs near 115 GeV to be about "50/50" based on the present data.

Problems with confidence intervals

- Even those who understand the distinction find the "confidence interval pathologies" unsettling
  - Much effort devoted to define approaches that reduce the frequency of their occurrence
- These cases are unsettling for the same reason:

  The degree of belief that these particular intervals contain the true value of the parameter is significantly less than the confidence level

  - furthermore, there is no standard method for quantifying the pathology

Problems with confidence intervals

- The confidence interval alone is not enough to
  - define an interval with stated coverage; and
  - express a degree of belief that the parameter is contained in the interval

- F. – C. recommend that experiments provide a second quantity: *sensitivity*
  - defined as the average limit for the experiment
  - consumer's degree of belief would be reduced if observed limit is far superior to average limit

Problems with *sensitivity*

- *Sensitivity* is not enough – need more information to compare with observed limit
  - variance of limit from ensemble of experiments?

- Use (*Sensitivity* – observed limit)/σ ?
  - not a good indicator that interval is "pathological"

Problems with *sensitivity*

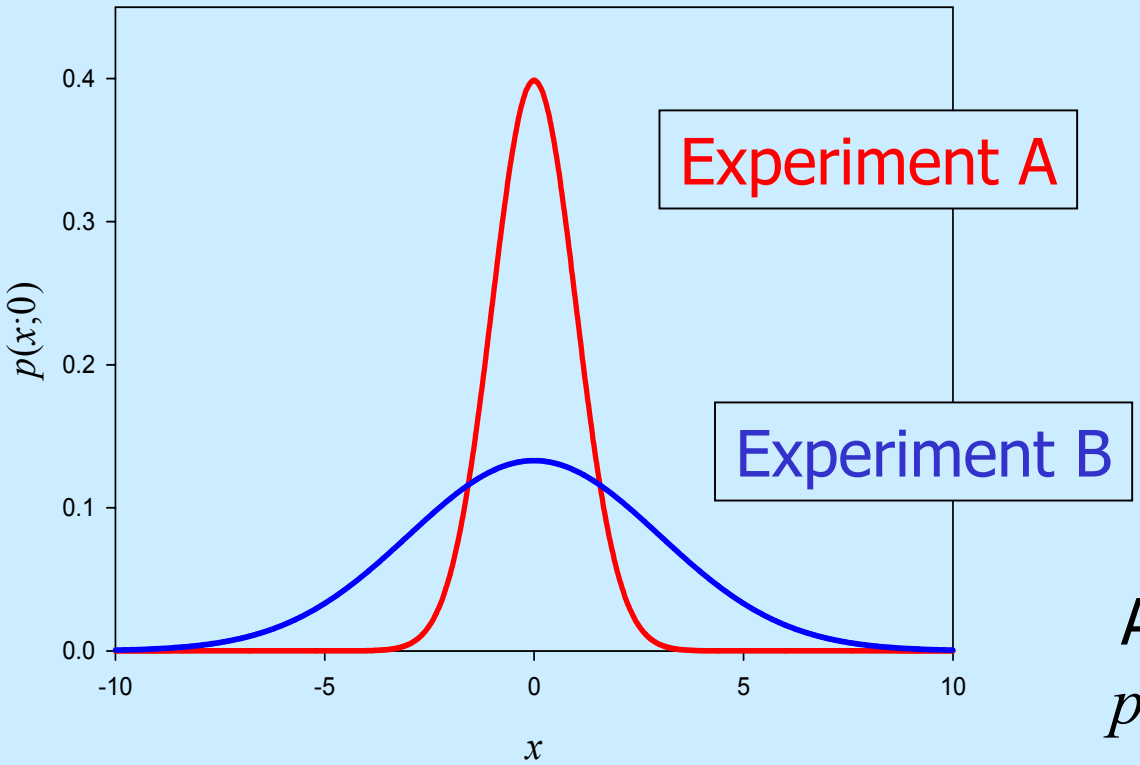
- Example: $m_{\nu\tau}$ analysis
  - $\tau \rightarrow$ 3 prong events contribute with different weight depending on:
    - mass resolution for event
    - nearness of event to $m_{\nu\tau} = 0$ boundary
  - ALEPH observes one clean event very near boundary $\rightarrow$ Limit is much better than average
  - Any reason to reduce degree of belief that the true mass is in the stated interval? NO!

Proposal

- When quoting a confidence interval for a frontier experiment, also quote its *credibility*
  - Evaluate the degree of belief that the true parameter is contained in the stated interval
    - Use Bayes thereom with a reasonable prior
      - recommend: flat in physically allowed region
    - call this the "credibility"
  - report credibility (and prior) in journal paper
    - if credibility is much less than confidence level, consumer would be warned that the interval may be "pathological"

Example: Gaussian with boundary

- x is an unbiased estimator for θ
- parameter, θ, physically cannot be negative

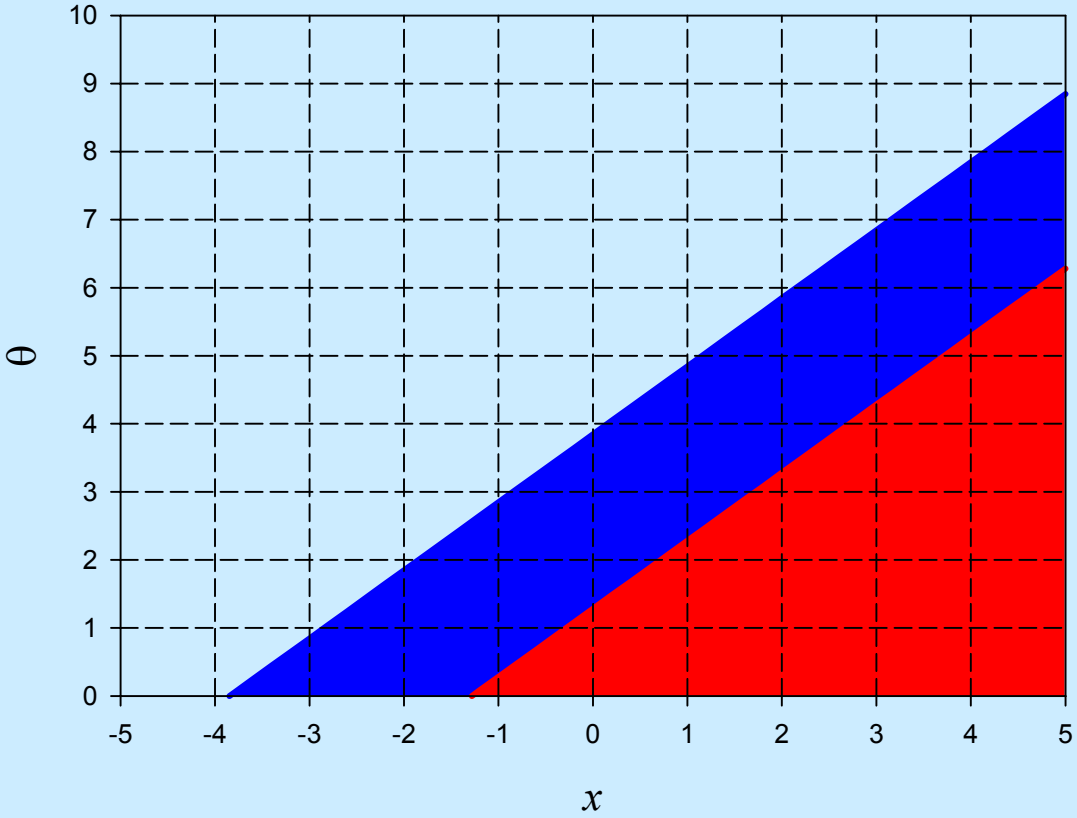


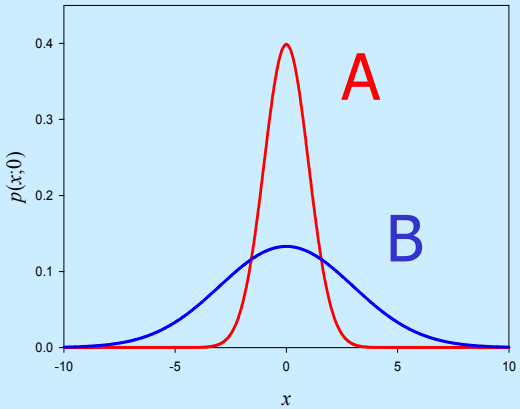
Assume

$$p(x;\theta) = p(x-\theta;0)$$

Example: 90% C.L. upper limit

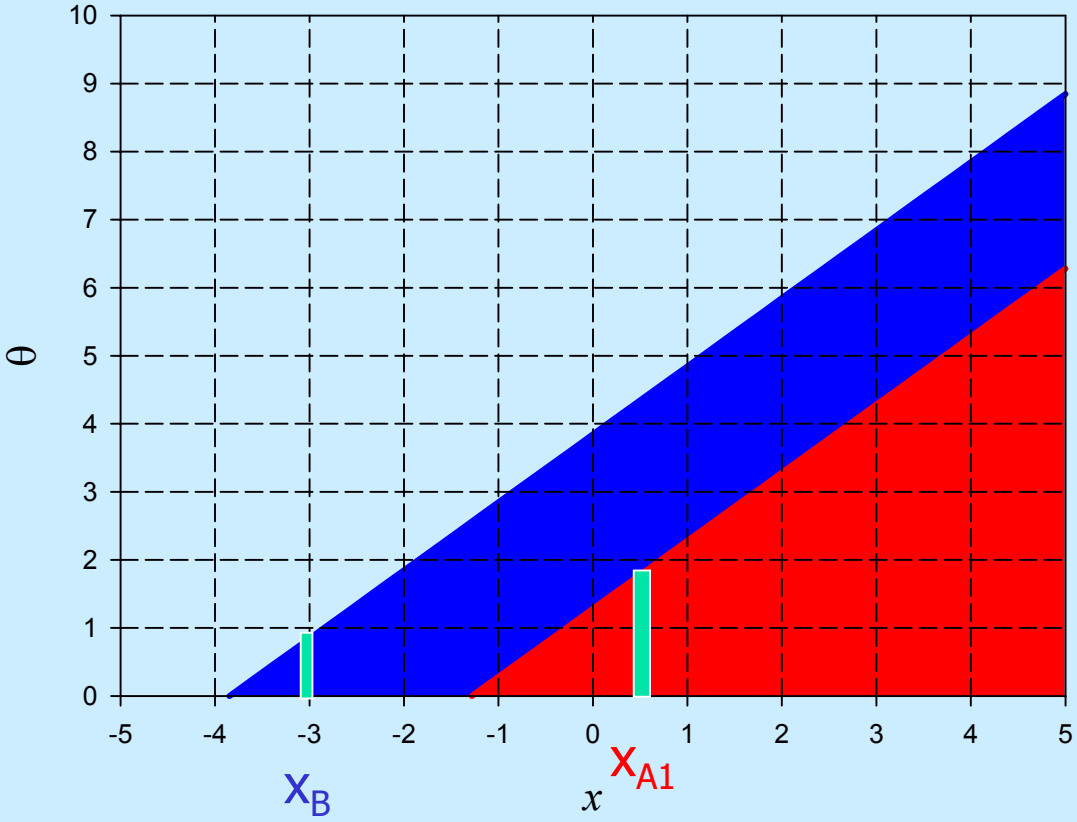
- Standard confidence belts:

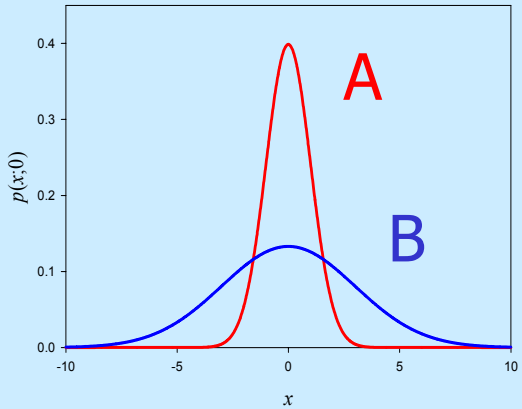




Example: 90% C.L. upper limit

- Consider 3 measurements



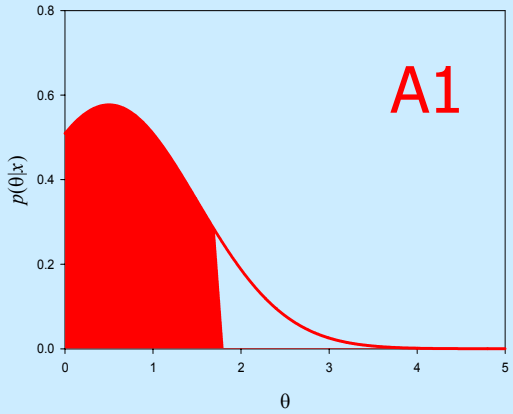


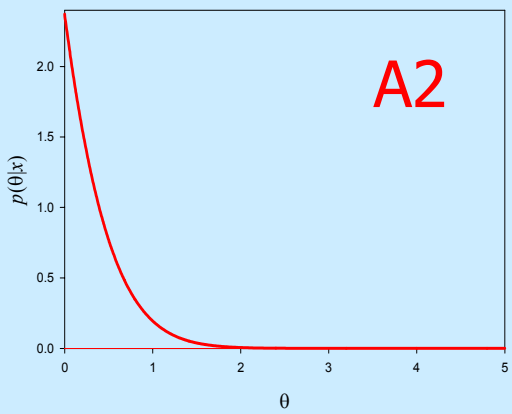
exp.	x	interval
A1	0.5	[0,1.78]
A2	-2.0	empty
B	-3.0	[0,0.85]

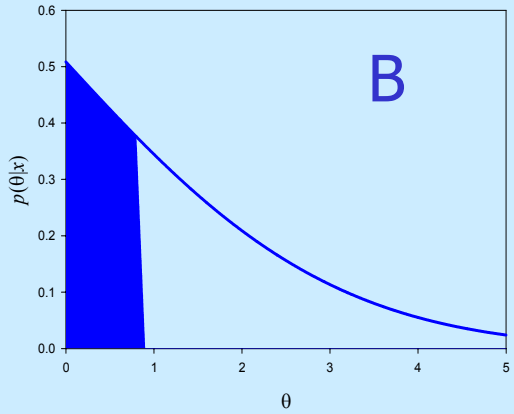
Example: 90% C.L. upper limit

- Calculate credibility of the intervals:
  - prior: $\pi(\theta) = \begin{matrix} \text{constant if } \theta \geq 0 \\ 0 \text{ if } \theta < 0 \end{matrix}$
  - Bayes theorem: $p(\theta \mid x) \propto L(x \mid \theta)\, \pi(\theta)$
  - Credibility: $\gamma = \int_{\theta_{low}}^{\theta_{up}} p(\theta \mid x) dx$

Example: 90% C.L. upper limit

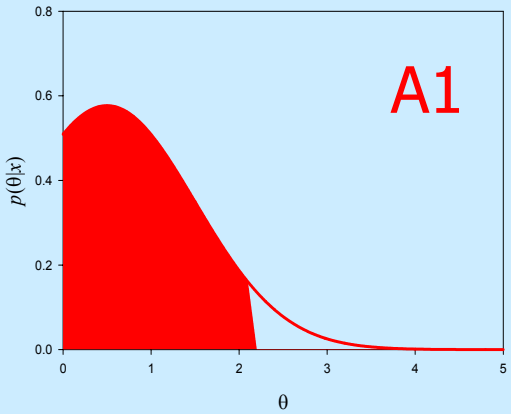


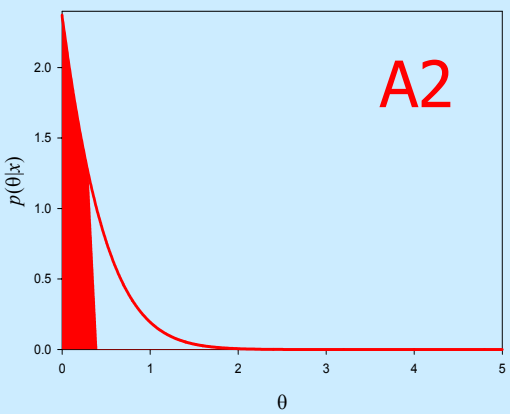


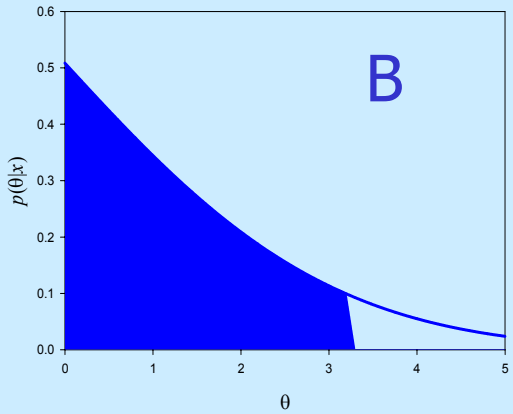


exp.	x	interval	credibility
A1	0.5	[0, 1.78]	0.86
A2	-2.0	empty	0.
B	-3.0	[0, 0.85]	0.37

Example: 90% C.L. unified interval







exp.	x	unified interval	credibility
A1	0.5	[0, 2.14]	0.93
A2	-2.0	[0, 0.40]	0.64
B	-3.0	[0, 3.30]	0.89

Example: Counting experiment

- Observe n events, mean background ν_b
  - Likelihood: $L(n \mid \nu_s) = \dfrac{(\nu_s + \nu_b)^n}{n!} e^{-(\nu_s + \nu_b)}$
  - prior: $\pi(\nu_s) = \begin{cases} \text{constant if } \nu_s \geq 0 \\ 0 \text{ if } \nu_s < 0 \end{cases}$

Example: $\nu_b = 3$

n	90% up lim	cred	90% unified	cred
0	empty	0.00	[0, 1.08]	0.66
1	[0, 0.89]	0.50	[0, 1.88]	0.78
3	[0, 3.68]	0.85	[0, 4.42]	0.90
6	[0, 7.53]	0.90	[0.15, 8.47]	0.93
10	[0, 12.41]	0.90	[2.63, 13.50]	0.91

Key benefit of the proposal

- Without proposal: experiments can report an overly small (pathological) interval without informing the consumer of the potential problem.

- With proposal: Consumer can distinguish credible from incredible intervals.

Other benefits of the proposal

- Education:
  - two different probabilities calculated – brings the distinction of coverage and credibility to the attention of physicists
- empty confidence intervals are assigned no credibility
- experiments with no observed events will be awarded for reducing their background (previously penalized)
- intervals "too small" (or exclusion of parameters beyond sensitivity) are assigned small credibility
- better than average limits not assigned small credibility if due to existence of rare, high precision events ($m_{\nu\tau}$)

Other benefits of the proposal

- Bayesian concept applied in a way that may be easy to accept even by devout frequentists:
  - choice of uniform prior appears to work well
  - does not "mix" Bayesian and frequentist methods
  - does not modify coverage
- Experimenters will naturally choose frequentist methods that are less likely to result in a poor degree of belief.
  - "Do you want to risk getting an incredible limit?"

Summary

- Confidence intervals are well defined, but
  - are frequently misinterpreted
  - can suffer from pathological problems when physical boundaries are present
- Propose that experiments quote credibility:
  - quantify possible pathology
  - reminder of two definitions of probabilities
  - encourages the use of methods for confidence interval construction that avoid pathologies