Upper Limits: A Personal View of Some History and Foundations

Bob Cousins

Univ. of California, Los Angeles

Terascale School on Data Combination and Limit Setting, Oct. 6, 2011

For more complete version, see my lectures at

Hadron Collider Physics Summer School 2009

http://hcpss.web.cern.ch/hcpss/2009/

"Virtual Talk" 12 Sep 2011

http://www.physics.ucla.edu/~cousins/stats/cousins_bounded_gaussian_virtual_talk_12sep2011.pdf

“Statistics” is HARD

- **It can be very *complicated* as well, but the deep reasons that *statistical inference* is HARD can be shown with alarmingly simple problems on Upper Limits.**
- ***I will assume that we really care about the answer:*** **This is often *not* the case for Upper Limits, but suppose we start to “exclude” all masses of the S.M. Higgs (!).**
- **I hope it becomes clear that one should perform three classes of caclulations (Bayesian credible intervals, likelihood ratio intervals, and Neyman’s confidence intervals) and compare.**

Start with simple problem, add complications:

Adapted from R. Cousins, Am. J. Phys. 63 398 (1995)

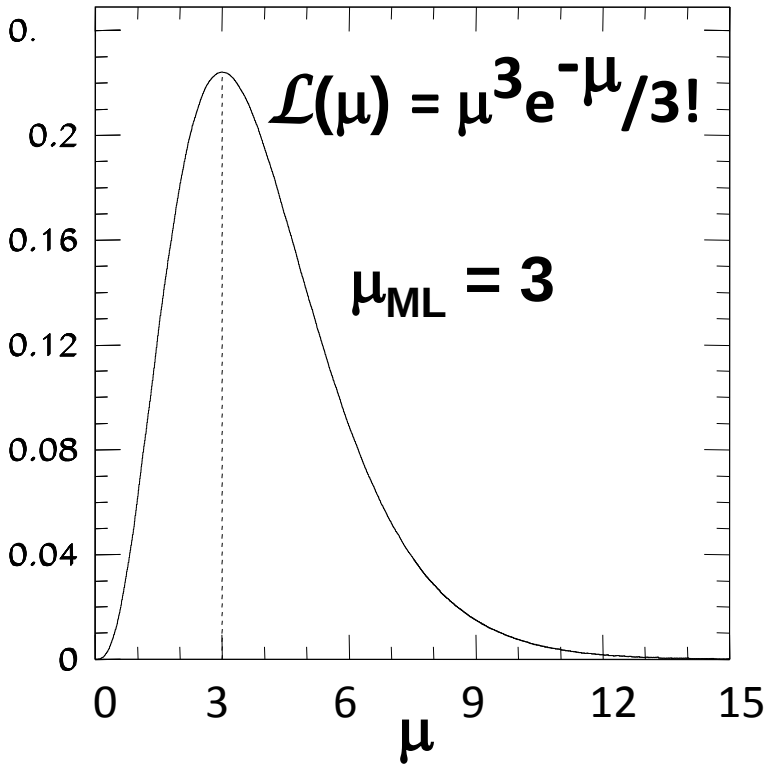
Poisson process $P(n|\mu) = \mu^n e^{-\mu}/n!$

Measurement of n yields n=3.

Substituting n=3 into $P(n|\mu)$ yields the *Likelihood function* $\mathcal{L}(\mu)$.

It is tempting to consider area under $\mathcal{L}$, but $\mathcal{L}(\mu)$ is *not* a probability density in μ:

Area under $\mathcal{L}$ is meaningless.

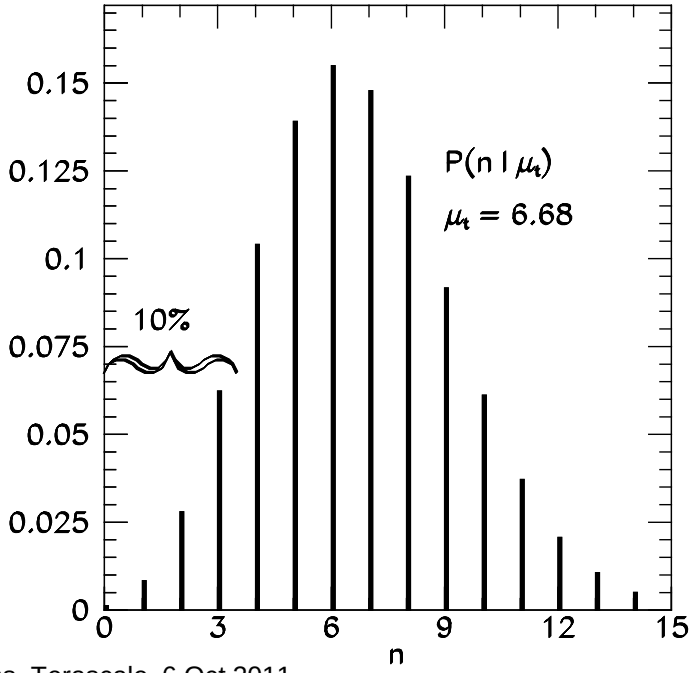


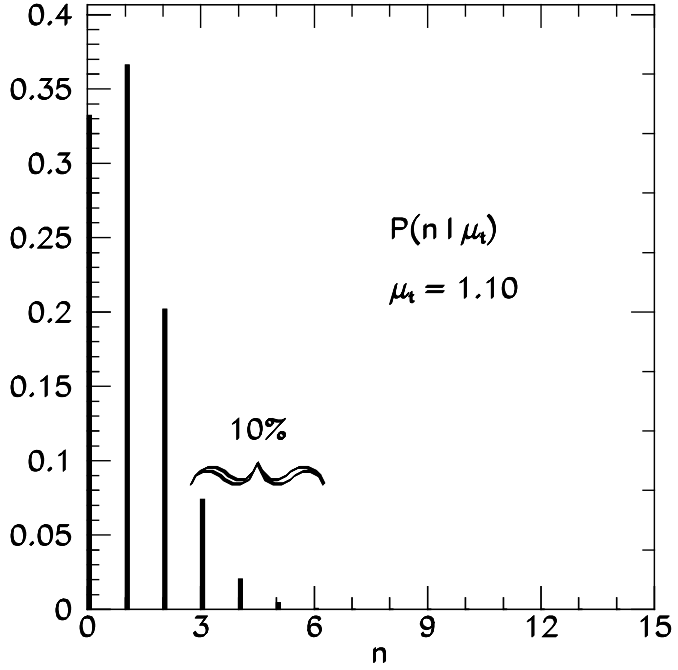
How to get upper (or lower) limit on μ ?
Consider 90% *upper* and 90% *lower* limits on μ.
Together they form an 80% *central interval* for μ.

1) *Frequentist confidence limit* method:

Find μ_u s.t. Poisson $P(n \leq 3 \mid \mu_u) = 0.1$. $\mu_u = 6.68$

Find μ_ℓ s.t. Poisson $P(n \geq 3 \mid \mu_\ell) = 0.1$. $\mu_\ell = 1.10$

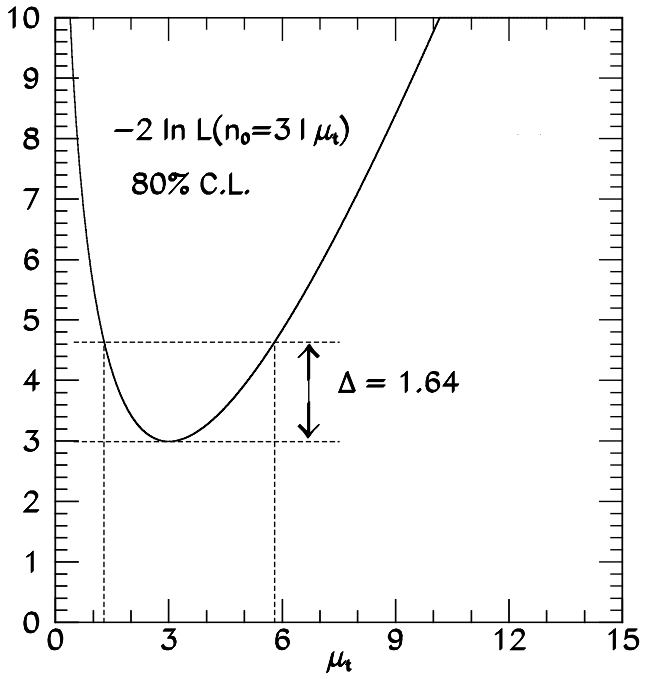




2) *Likelihood ratio* method.
Based on $\mathcal{L}(\mu)\,/\mathcal{L}(\mu_{ML})$, equivalently:

$$-2\ln\mathcal{L}(\mu) - (-2\ln\mathcal{L}(\mu_{ML})) \leq Z^2, \text{ for Z real.}$$

Asymptotically (note regularity conditions) this interval approaches a frequentist central confidence interval with C.L. corresponding to ± Z Gaussian standard deviations.



For 80% central interval, Z=1.28.
90% upper and lower limits are:
$\mu_u = 5.80$
$\mu_\ell = 1.29$

3) *Bayesian* method.

Different definition of probability: *degree of belief*.

With that definition, one can have pdf's in μ (!)

$p(\mu|n=3) \propto \mathcal{L}(\mu)\, p(\mu)$,

$p(\mu|n=3)$ = *posterior* pdf for μ, given n=3

$\mathcal{L}(\mu)$ = Likelihood function from above for n=3

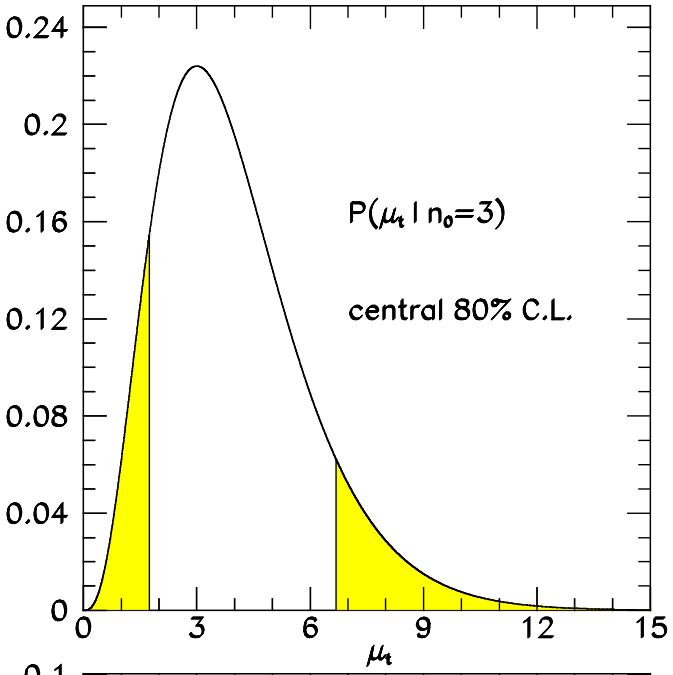
$p(\mu)$ = *prior* pdf for μ, before incorporating n=3.

Vast literature on Bayesian methods and priors.
This literature has largely been ignored in HEP, where most papers use uniform prior for μ.
In HEP, practice is generally what Bayesian statisticians call "pseudo-Bayesian".

Graph of posterior pdf *is* a density, so area under it has meaning.

With 10% of area under posterior in each tail, obtain 90% upper and lower *credible* limits.

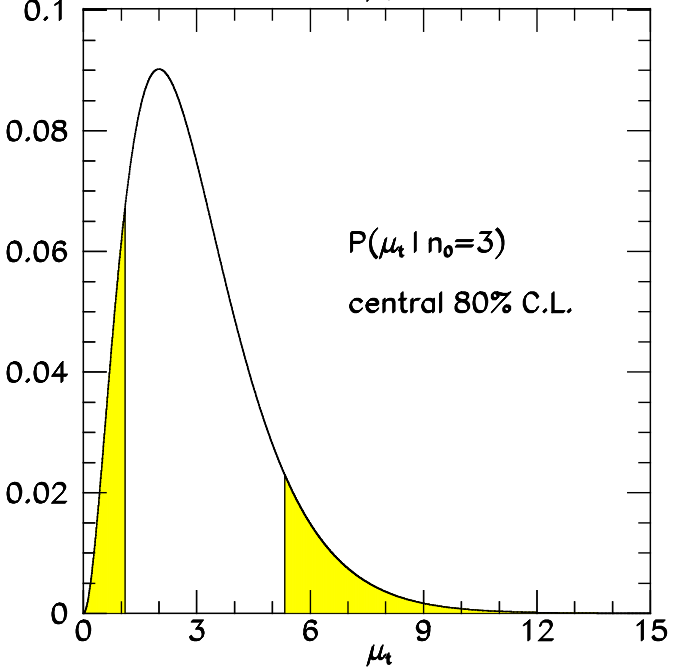
Interval of course depends on prior.



$p(\mu) \propto 1$

$\mu_u = 6.68$

$\mu_\ell = 1.74$



$p(\mu) \propto 1/\mu$

$\mu_u = 5.32$

$\mu_\ell = 1.10$

Deep Foundational Issue: Confidence Principle (Frequentist Coverage) vs Likelihood Principle

The Likelihood ratio interval and the Bayesian interval use $\mathcal{L}(\mu)$ given the *observed* n=3, but make *no* use of $P(n|\mu)$ for any $n \neq 3$. This is the essence of the *Likelihood Principle*.

The confidence interval relying on $P(n \leq 3 \mid \mu)$ and $P(n \geq 3 \mid \mu)$ used *probabilities of data not observed*.

This violates the L.P.

This turns out to be a *huge deal*:
In general, cannot have both coverage and L.P.
Whole approach of tail probabilities violates L.P. !

Summary of 80% Central Intervals, n=3 (Endpoints are 90% lower and upper limits)

	Frequentist Confidence	Likelihood Ratio	Bayesian Credible
$[\mu_\ell, \mu_u]$	[1.10, 6.68]	[1.29, 5.80]	$p(\mu) \propto 1$ [1.74, 6.68] $p(\mu) \propto 1/\mu$ [1.10, 5.32]
Requires prior pdf?	No	No	Yes
Provides P(parameter|data)?	No	No	Yes
Random variable in "$P(\mu_t \in [\mu_\ell, \mu_u])$":	μ_ℓ, μ_u	μ_ℓ, μ_u	μ_t
Coverage guaranteed? "Confidence Principle"	**Yes (but over-coverage…)**	**No**	**No**
Obeys "Likelihood Principle"?	**No**	**Yes**	**Yes (exception re Jeffreys prior)**

Now add complications:
First, a *known* mean background b, say b=2.8.

Central frequentist confidence interval shifts downward by 2.8. As n decreases or b increases, interval can "reject" regions where no sensitivity, and even reject *all* values of μ (null interval!).

Likelihood-ratio interval hits vertical axis before going up by Delta: running into violation of regularity conditions.

Bayesian interval is at least superficially well-behaved: historically was adopted by PDG (following Helene paper). *But how to interpret P?*

Likelihood Principle Example

The "Karmen Problem"

- You expect background events sampled from a Poisson mean b=2.8, assumed known precisely.
- For signal mean μ, the total number of events n is then sampled from Poisson mean μ+b.
- So $P(n) = (\mu+b)^n \exp(-\mu-b)/n!$
- Then you see no events at all! I.e., n=0.
- $\mathcal{L}(\mu) = (\mu+b)^0 \exp(-\mu-b)/0! = \exp(-\mu) \exp(-b)$

Likelihood Principle Example

The "Karmen Problem"

- You expect background events sampled from a Poisson mean b=2.8, assumed known precisely.
- For signal mean μ, the total number of events n is then sampled from Poisson mean μ+b.
- So $P(n) = (\mu+b)^n \exp(-\mu-b)/n!$
- Then you see no events at all! I.e., n=0.
- $\mathcal{L}(\mu) = (\mu+b)^0 \exp(-\mu-b)/0! = \exp(-\mu) \exp(-b)$

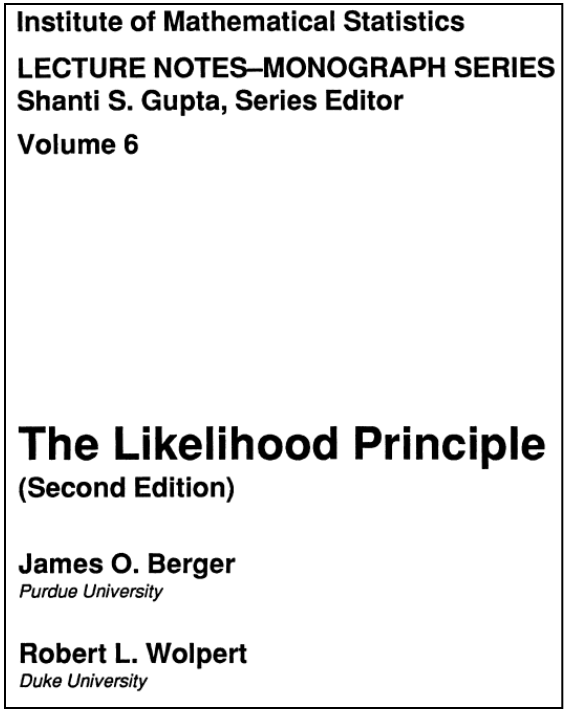
Note that changing b from 0 to 2.8 changes $\mathcal{L}(\mu)$ only by the constant factor exp(-b). This gets renormalized away in any Bayesian calculation, and is irrelevant for likelihood *ratios*. So for zero events observed, likelihood-based inference about signal mean μ is *independent of expected b*.

For essentially all frequentist confidence interval constructions, the fact that n=0 is less likely for b=2.8 than for b=0 results in *narrower* confidence intervals for μ as b increases. Clear violation of the L.P.

Likelihood Principle Discussion

We will not resolve this issue, but should be aware of it.

- **See book by Berger & Wolpert, but be prepared for the "Stopping Rule Principle" to set your head spinning.**
- **When frequentist intervals and limits badly violate the L.P., use great caution in interpreting them!**
- **And when Bayesian inferences badly violate the Confidence Principle (frequentist coverage), again use great caution!**



Poisson with Known Mean Background

1996 PDG RPP, a la Helene or Zech

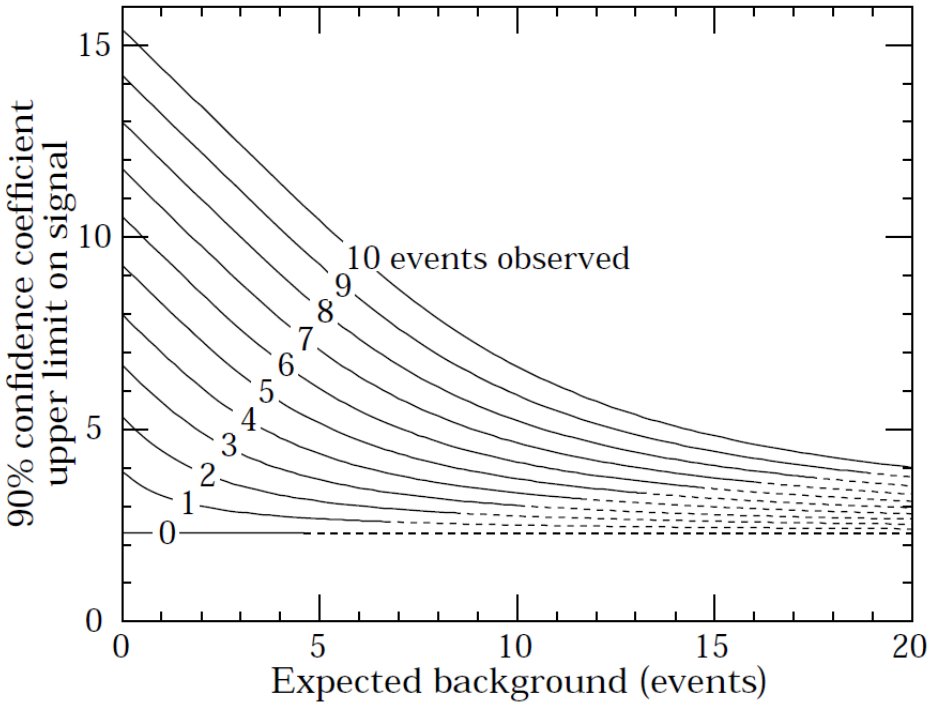


Figure 28.8: 90% confidence coefficient upper limit on the number of signal events as a function of the expected number of background events. For example, if the expected background is 8 events and 5 events are observed, then the signal is 4.0 (approximately) or less with 90% confidence. Dashed portions indicate regions where it is to be expected that the number observed would exceed the number actually observed $\geq 99\%$ of the time, even in the complete absence of signal.

1998 PDG RPP, a la Feldman & Cousins

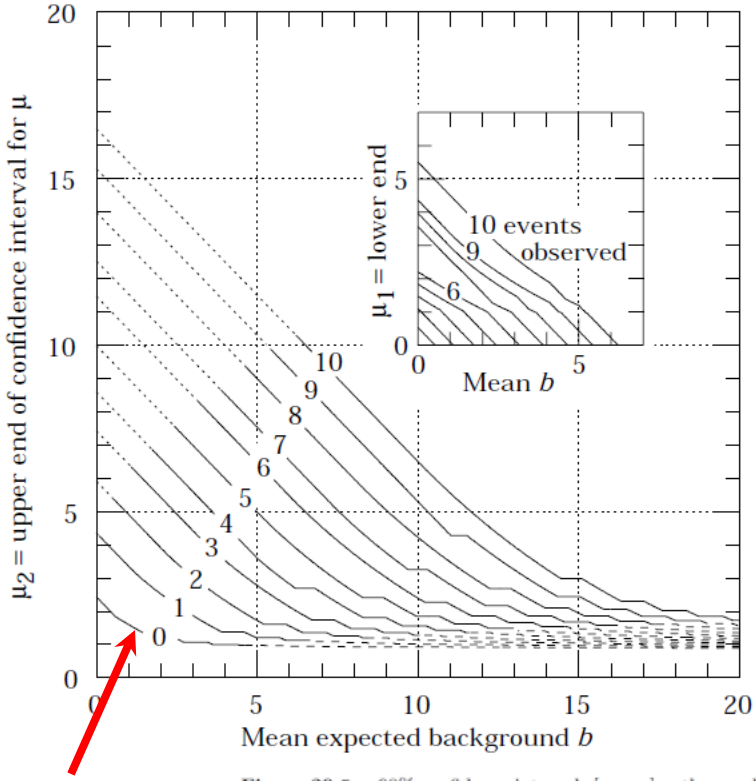


Figure 29.5: 90% confidence intervals $[\mu_1, \mu_2]$on the number of signal events as a function of the expected number of background events b. For example, if the expected background is 8 events and 5 events are observed, then the signal is 2.60 or less with 90% confidence. Dotted portions of the μ_2 curves on the upper left indicate regions where μ_1 is non-zero (as shown by the inset). Dashed portions in the lower right indicate regions where the probability of obtaining the number of events observed or fewer is less than 1%, even if $\mu = 0$. Horizontal curve sections occur because of discrete number statistics. Tables showing these data as well as the CL = 68.27%, 95%, and 99% results are given in Ref. 11.

Now add systematic uncertainties:

1) Uncertainty on mean background b.

2) Uncertainty on product of luminosity and efficiencies.

This complicates things enormously!

Takes us into territory of "nuisance parameters" and research problems in professional statistics literature for which there is still no clearly preferred solution.

Cousins review at PhyStat05: http://www.physics.ox.ac.uk/phystat05/proceedings/
Demortier review at PhyStat07: http://phystat-lhc.web.cern.ch/phystat-lhc/2008-001.pdf

Remember: we would like a numerical answer for which "90%" corresponds, at least approximately, to some definition of probability!

Treatment of Nuisance Parameters within Each Paradigm

1) Frequentist Confidence Intervals.
***(Full) Neyman construction*. (See my HCPSS lectures.)**
For each point in the subspace of nuisance parameters, treat them as fixed true values and perform a Neyman construction for multi-D confidence regions in the full space of all parameters. Project these regions onto the subspace of the parameter of interest.

Problem(s): Typically intractable and causes overcoverage, and therefore rarely attempted. Tractability recovered by doing the construction in the lower dimensional space of the profile likelihood function. Not well-studied.

Nuisance Parameters within Each Paradigm (Cont.)

2) Likelihood intervals: "Profile likelihood Method".

For each value of the parameter of interest, search the full subspace of nuisance parameters for the point at which the likelihood is maximized. Associate that value of the likelihood with that value of the parameter of interest. The set of such likelihoods is called the *profile likelihood*, and is a function only of the parameter of interest. The math is now reduced to the case of no nuisance parameters.

(Familiar to many as MINUIT MINOS.)

Problem(s): This has a reputation of underestimating the true uncertainties. In Poisson problems, this is partially compensated by effect due to discreteness of n.

In HEP, profile likelihood (MINUIT MINOS) gives good performance in many problems.

Nuisance Parameters within Each Paradigm (Cont.)

3) Bayesian credible intervals:

Construct a multi-D prior pdf P(parameters) for the space spanned by all parameters; multiply by P(data|parameters) for the data obtained; integrate over the full subspace of all nuisance parameters; you are left with the posterior pdf for the parameter of interest. The math is now reduced to the case of no nuisance parameters.

Problem(s): The multi-D prior pdf is a problem for both subjective and non-subjective priors. Until very recently, in HEP there has been almost no use of the favored non-subjective priors (reference priors of Bernardo and Berger). The high-D integral can be a technical problem, more and more overcome by Markov Chain Monte Carlo.

Hybrid Techniques: Introduction to Pragmatism

Given these difficulties, it is common in HEP to relax foundational rigor and:

- **Treat nuisance parameters in a Bayesian way while treating the parameter of interest in a frequentist way, or**
- **Treat nuisance parameters by profile likelihood while treating parameter of interest another way, or**
- **Use the Bayesian framework (even without the priors recommended by statisticians), but evaluate the frequentist performance.**
  In effect (as in profile likelihood) one gets approximate coverage while respecting the L.P.
- **The properties of the result are more important than the "derivation" !**

Example of treating nuisance parameter in a Bayesian way while treating parameter of interest in a frequentist way

Nuclear Instruments and Methods in Physics Research A320 (1992) 331–335

Incorporating systematic uncertainties into an upper limit

Robert D. Cousins
Physics Department, University of California, Los Angeles, CA 90024, USA

Virgil L. Highland
Physics Department, Temple University, Philadelphia, PA 19122, USA

Our statistical approach includes both classical and Bayesian elements [1]. Our treatment of the Poisson parameter is classical, the type of statistics we generally prefer. Because we average over a probability distribution for the experimental sensitivity, our treatment of that quantity is necessarily Bayesian.

A main point of this "C-H" paper was that for small n, effect of syst error on upper limit went as square of relative syst error: 10% syst error has negligible effect on limit.

Luc Demortier pointed out that result is same as G. Box's *prior predictive p-value* (1980).

Problems with treating nuisance parameters in a Bayesian way while treating parameter of interest in a frequentist way

1) Inherits all the unresolved issues of priors from Bayesian methods.

2) Since method mixes definitions of P, results have no guaranteed properties and must be studied on case-by-case basis.

a) **Numerous studies have shown that results for upper limits at 90-95% C.L. (the C-H case) are reasonable, though typically over-covering.**

b) **Kyle Cranmer showed at Oxford PhyStat (2005) that claimed 5-sigma discovery could really be 4.2.**

About Those Priors…

- **There are many flavors of Bayesians among statisticians, in two broad categories:**
  - **"Subjective": P is personalistic degree of belief. Prior encodes that. Strong foundational arguments of "coherence". (B. DeFinetti, J. Savage, et al.)**
  - **"Objective" (self-description): uses "formal rules" for priors, attempting to "let the data speak as loud as possible". (H. Jeffreys, J. Bernardo, J. Berger, et al.)**
- **"Non-informative" priors *do not exist*: a prior *always* inputs information!**
- **Improper priors (e.g., uniform on [0,∞]) can cause all kinds of trouble: Equalities become proportionalities! Stats literature has important insights on how to avoid some traps.**

Can "subjective" be taken out of "degree of belief"?

- **A bright idea, vigorously pursued by physicist Harold Jeffreys in in mid-20th century: *Can one define a prior $p(\mu)$ which contains as little information as possible, so that the posterior pdf is dominated by the likelihood?***
- **The really *really* thoughtless idea*, recognized by Jeffreys as such, but dismayingly common in HEP: just choose $p(\mu)$ uniform in whatever metric you happen to be using!**
- **The "objective" priors from Jeffreys's rule and from "reference priors of Bernardo define the prior based on properties of the *measuring apparatus*, not from thinking about the parameter!**

***In spite of having a fancy name, Laplace's Principle of Insufficient Reason**

"Uniform Prior" Requires a Choice of Metric

"Jeffreys Prior" uses a prior uniform in a metric related to the Fisher information (technical term).

- Poisson signal mean μ, no background: $p(\mu) = 1/\sqrt{\mu}$
- Poisson signal mean μ, mean background b: $p(\mu) = 1/\sqrt{\mu+b}$
- Unbounded mean μ of gaussian: $p(\mu) = 1$
- RMS deviation of a Gaussian when mean fixed: $p(\sigma) = 1/\sigma$
- Binomial parameter ρ, $0 \leq \rho \leq 1$: $p(\rho) = \rho^{-1/2}(1 - \rho)^{-1/2} = \text{Beta}(1/2,1/2)$

If measuring apparatus has Gaussian resolution in m, the prior is uniform in m.

If it has Gaussian resolution in m^2, the prior is uniform in m^2.

Jeffreys prior yields pdfs which are consistent under transformation into different metrics.

Welch and Peers famously showed that Bayesian intervals with Jeffrey's prior have good coverage (to order 1/n).

Workshop on Confidence Limits

27-28 March, 2000
Fermilab 1-West Conference Room

Jim Berger:

M. Kendall, giving the 'old' frequentist viewpoint of Bayesian analysis;

"If they [Bayesians] would only do as he [Bayes] did and publish posthumously, we should all be saved a lot of trouble."

What should be the view today;

Objective Bayesian analysis is the best frequentist tool around.

Sensitivity Analysis

- **Since a Bayesian result depends on the prior probabilities, which are either personalistic or with elements of arbitrariness, it is widely recommended by Bayesian statisticians to study the *sensitivity* of the result to varying the prior.**

Bounded Gausian problem:

Measurement x is unbiased Gaussian estimate of μ:

$$p(x|\mu) \sim e^{-(x-\mu)^2 / 2\sigma^2}.$$

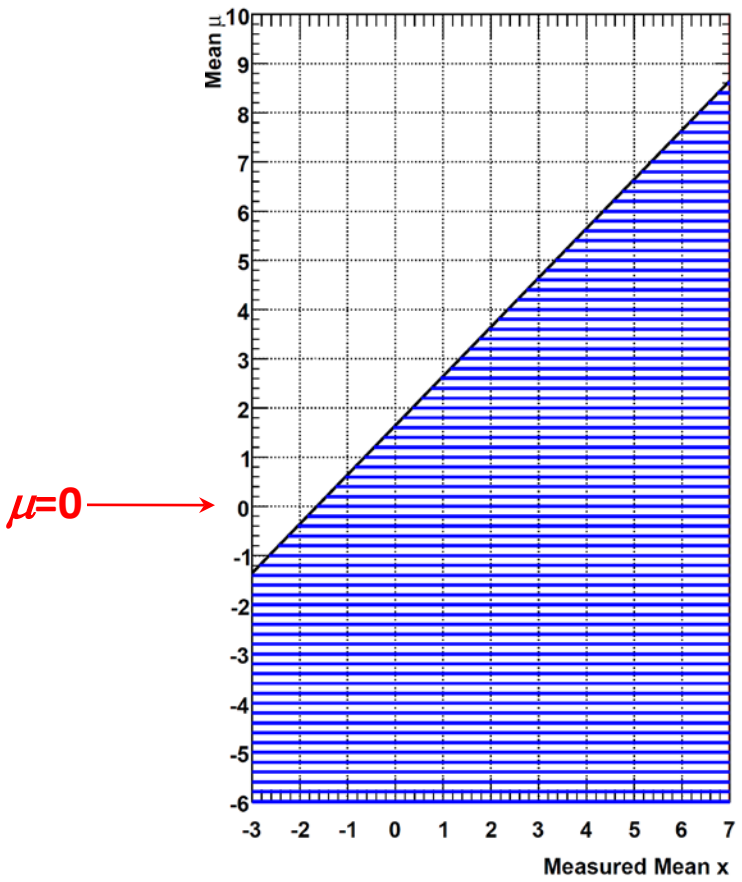
What is the 95% C.L. Upper Limit (UL) for μ if the physical model for $p(x|\mu)$ exists only for $\mu \geq 0$?

***Without the constraint on μ*, traditional frequentist and Bayesian methods both yield:**

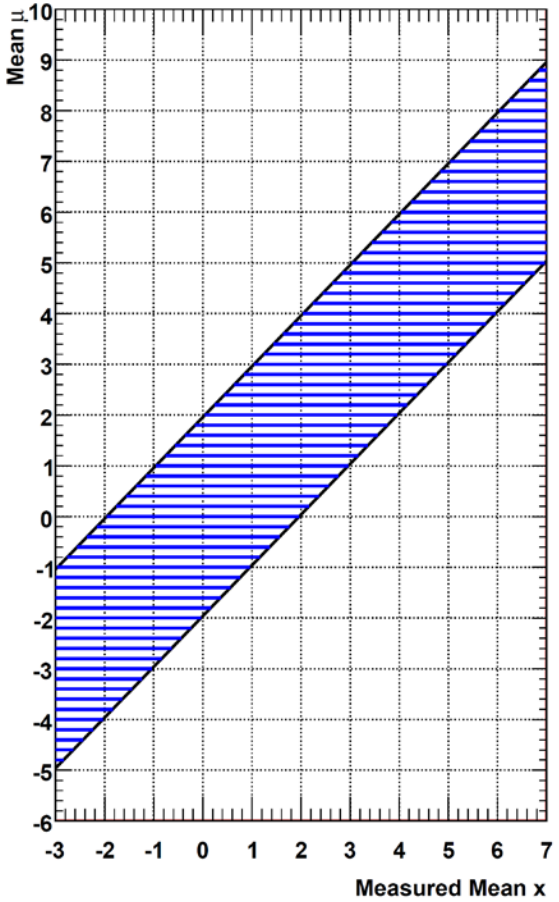
$$UL = x + 1.64\sigma,$$

and 95% C.L. central confidence interval is $x \pm 1.96\sigma$.
See next slide:

Graphical display of intervals is a *confidence belt:* Confidence interval include all values of μ for which horizontal blue line is intersected by vertical line drawn at measured value of *x*.

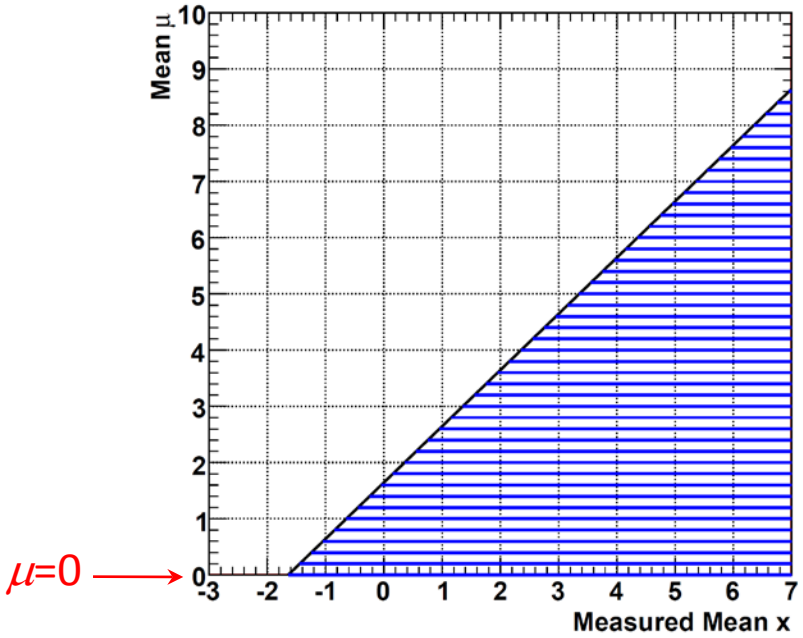


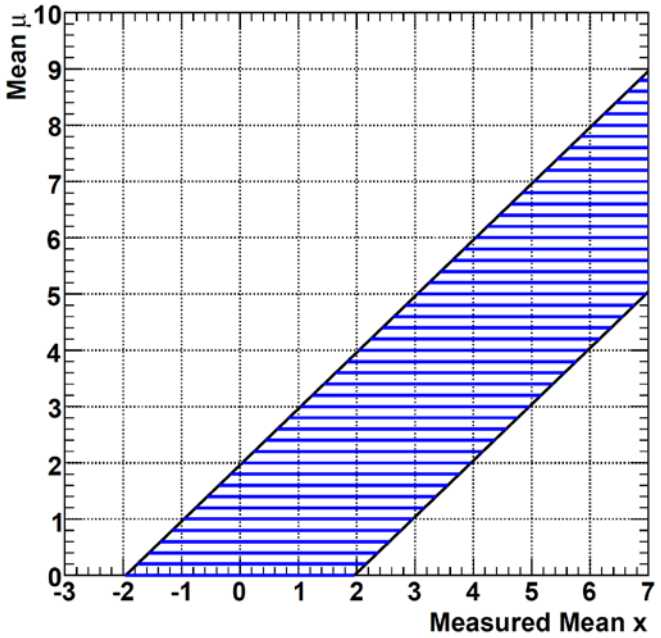
Upper limit = $x + 1.64\,\sigma$



Central interval = $x \pm 1.96\,\sigma$

***With the constraint $\mu \geq 0$*, the story takes us not only to the heart of Bayesians-frequentist disputes, but also to *frequentist* criticisms of Neyman & Pearson by Sir Ronald Fisher and Sir David Cox!**





For $x < -1.64\sigma$ with UL, and for $x < -1.96\sigma$ with central intervals, the confidence interval is the *null* set!

I refer to the plot on left as the "*diagonal line*".

So, what did people in HEP do?
The problem arose in experiments with true μ << σ, so that measured x<0 was common.
Some chose to move $x<0$ to physical boundary of μ.

A SEARCH FOR THE DECAY $\pi^0 \rightarrow 3\gamma$ *

J. DUCLOS **, D. FREYTAG, K. SCHLÜPMANN and V. SOERGEL
CERN, Geneva, Switserland

J. HEINTZE and H. RIESEBERG
I. Physikalisches Institut des Universität Heidelberg, Germany

Phys Lett 19 253 (1965)

$x = -0.5 \pm 2.5$

Set x=0 and proceeded.

NEUTRAL DECAY BRANCHING RATIOS OF THE η^0 MESON

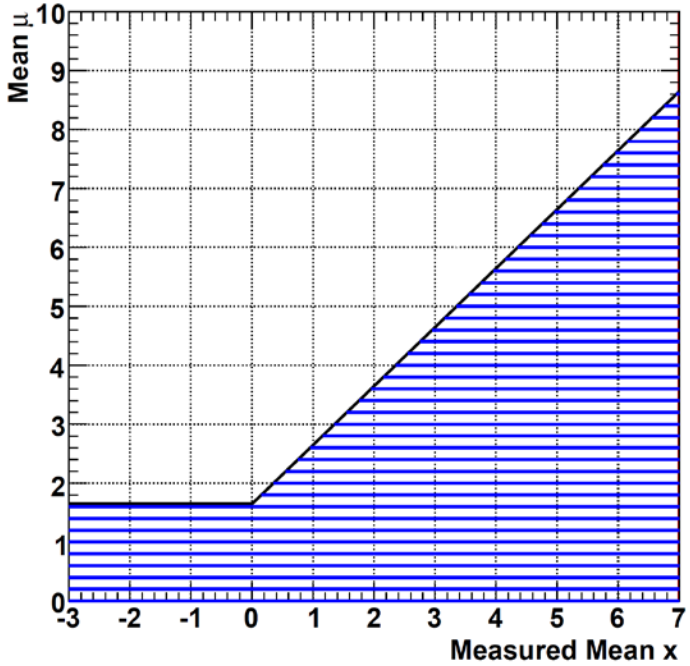
C. Baltay,† P. Franzini, J. Kim, R. Newman, and N. Yeh
Columbia University, New York, New York, and Brookhaven National Laboratory, Upton, New York
L. Kirsch
Brandeis University, Waltham, Massachusetts

PRL 19 1495 (1967)

$x = -0.06 \pm 0.14$

Set x=0 and proceeded.

With this ad hoc patch, UL = max(x,0) + 1.64σ.
"95% C.L." intervals had 100% coverage (!) if $\mu < 1.64$



I'll refer to this as the
"*original Diagonal plus Horizontal Line*",
***"DHL"* for short.**

Precision measurement of the muon momentum in pion decay at rest

M. Daum, G. H. Eaton, R. Frosch, H. Hirschmann, J. McCulloch,* R. C. Minehart,† and E. Steiner
Swiss Institute for Nuclear Research, SIN, 5234 Villigen, Switzerland

Phys Rev D20 2692 (1979)

$$m_{\nu_\mu}{}^2 = 0.13 \pm 0.14\ (\mathrm{MeV}/c^2)^2$$

Following the method recommended by the Particle Data Group,[33] illustrated in Fig. 22, we calculated the upper limit of the muon-neutrino mass. The result is

$$m_{\nu_\mu} \leq 0.57\ \mathrm{MeV}/c^2\ (90\%\ \text{confidence level}). \quad (9)$$

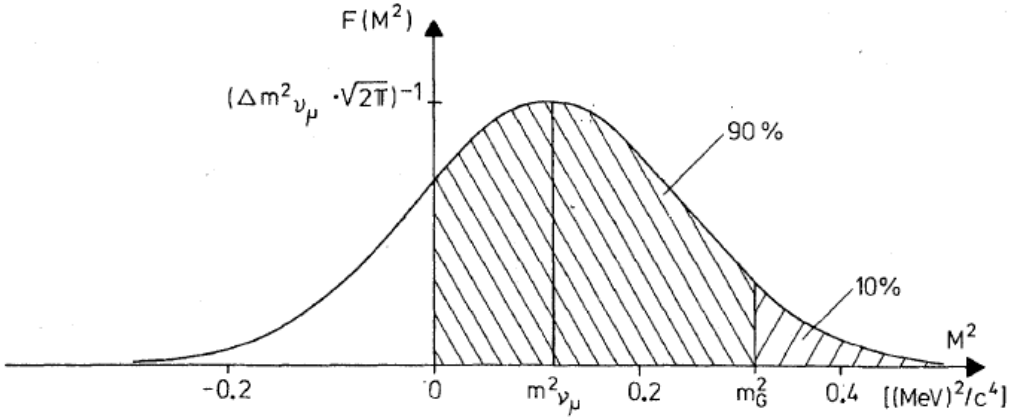


FIG. 22. According to the prescription of the Particle Data Group (Ref. 33) the upper limit m_G of the muon-neutrino mass is calculated from the squares mass $m_{\nu_\mu}{}^2$ and its uncertainty $\Delta(m_{\nu_\mu}{}^2)$ by setting the probability function $F(M^2)$ to zero for $M^2 < 0$, as indicated in the figure.

[33] T. G. Trippe, private communication, 1976.

The 1979 prescription alleged to be that of the PDG was numerically equivalent to:

$p(x|\mu) \sim e^{-(x-\mu)^2/2\sigma^2}$.

$\Rightarrow \mathcal{L}(x_0|\mu) \sim e^{-(x_0-\mu)^2/2\sigma^2}$.

Prior $p(\mu) \sim 1$ if $\mu \geq 0$, else 0.

Posterior $p(\mu|x_0) \propto \mathcal{L}(\mu)\, p(\mu)$.

This is a prob. density in μ. Renormalize and integrate to find μ_{UL} with 5% tail probability.

This prescription *did* appear in PDG Review of Particle Physics since 1986.

Belt of Bayesian UL at right.

Daum, et al.

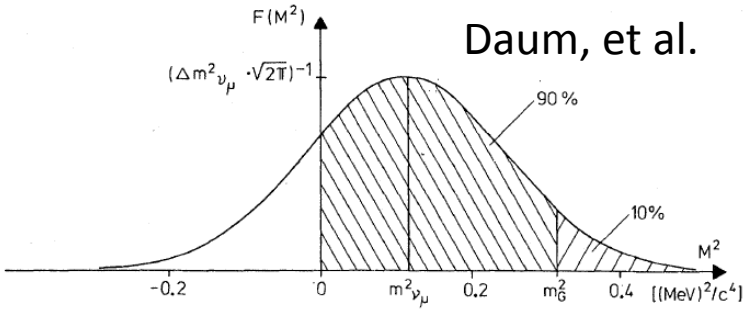
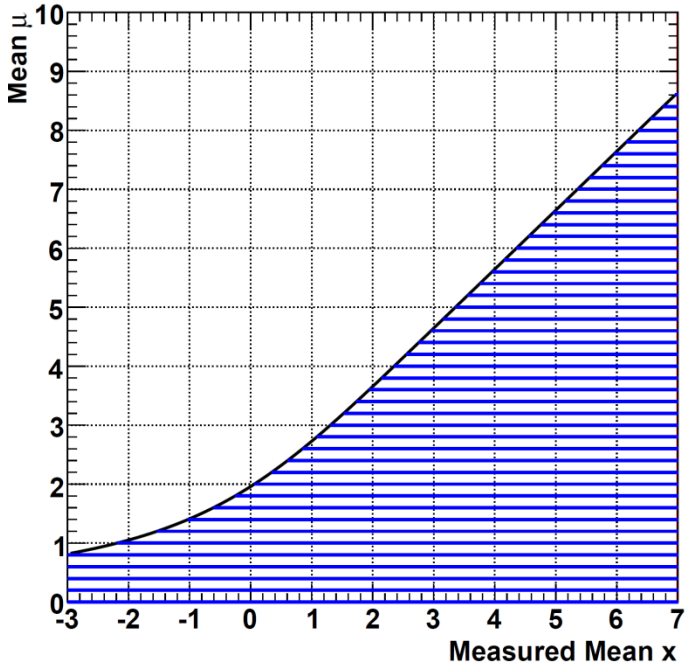


FIG. 22. According to the prescription of the Particle Data Group (Ref. 33) the upper limit m_G of the muon-neutrino mass is calculated from the squares mass $m_{\nu_\mu}^2$ and its uncertainty $\Delta(m_{\nu_\mu}^2)$ by setting the probability function $F(M^2)$ to zero for $M^2 < 0$, as indicated in the figure.



2002: Physicist Mark Mandelkern writes Statistics review article asking statisticians for advice (!)

Setting Confidence Intervals for Bounded Parameters

Mark Mandelkern

Abstract. Setting confidence bounds is an essential part of the reporting of experimental results. Current physics experiments are often done to measure nonnegative parameters that are small and may be zero and to search for small signals in the presence of backgrounds. ...

Editor asks five statisticians to Comment.
Leon Jay Gleser is truly incisive, emphasizing:

"…the predata measure of risk is not necessarily the correct postdata measure of uncertainty."

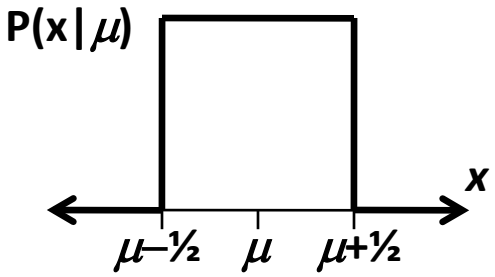
Insights by Sir Ronald Fisher in 1956 and Sir David Cox in 1958 pointed to situations in which Most Powerful Neyman-Pearson tests gave answers clearly not relevant to the data at hand!

The basic idea is that sometimes there are "recognizable subsets" of the *sample space* (*x*) for which the N-P C.L. (computed from the *whole* space) is in conflict with properties of the subset.

In our problem, we are clearly in this situation when the "upper limit" is null or unphysical: *conditional probability of coverage* within that *recognizable* part of the sample space is zero!

A whole literature. First, a simple clean example.

Let p(x|μ) = 1 if $\mu - ½ \le x \le \mu + ½$; 0 otherwise.



Two measurements x_1,x_2 are made.

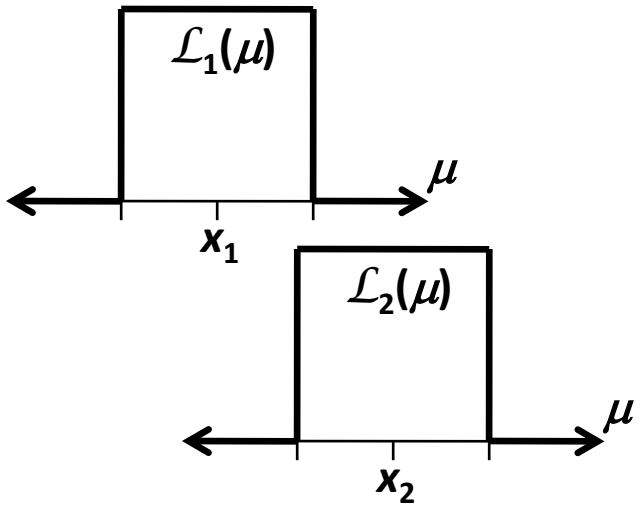
What is a central confidence interval for μ ?

Most Powerful one-sided N-P tests lead to the 68% C.L. central interval $\mu = (x_1 + x_2)/2 \pm 0.22$.

This uncertainty is determined by the ensemble of *all possible* measurements x_1,x_2.

It is a *pre-data assessment of uncertainty*.

But once data is in hand, if $|x_1 - x_2|$ is close to 1, we *know* that we have a much more accurate measurement of μ for *our particular "lucky" data*.



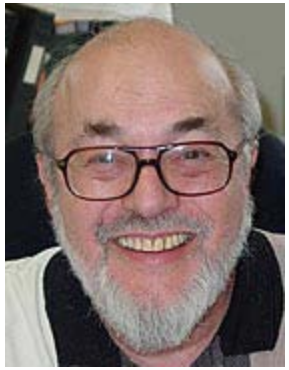
The "relevant" *post-data assessment of uncertainty about* μ depends on $|x_1 - x_2|$, which can be used to partition the sample space into *recognizable subsets.*

In clean cases with such as this, *the coverage of the conditional statements in the unconditional ensemble is exact*, though power is *less*.

In the 1980's, Günter Zech attempted (in the related Poisson problem) to build in exact conditional coverage from the beginning of the construction of upper limits on a bounded parameter. His calculation, which inspired CL_s, leads to *over-coverage in the unconditional ensemble*.

In 2002, statistician Gleser pointed us to 1959+ literature on *conditional coverage* as a tool for *evaluating* confidence sets built to have perfect unconditional coverage.

More from Leon Jay Gleser



“The subset of samples having the property that the sample mean is two standard deviations to the left of zero would have been called a ‘recognizable subset’ by Fisher (1956).”

More from Leon Jay Gleser

"Buehler (1959), and later Robinson (1979), introduced the notion of *conditionally admissible* tests and confidence intervals—those procedures whose frequentist control of error (coverage probability, level of significance) was not adversely affected by the realization that a given data set belonged to a recognizable subset of samples."

Very enlightening literature – see my recent post

Negatively Biased Relevant Subsets Induced by the Most-Powerful One-Sided Upper Confidence Limits for a Bounded Physical Parameter

http://arxiv.org/abs/1109.2023

More from Leon Jay Gleser

"…any confidence intervals that keep a constant width as *X* becomes more negative, as some of the physicists seem to desire, are indicating not necessarily what the data shows through the model and likelihood, but rather desiderata imposed external to the statistical model."

Deep Connections to Bayesian Statistics

Furthermore, a number of theorems have been proved in the last 50 years making connections between:

- **Good frequentist *conditional coverage* properties**
- **The existence of *any* prior for which the Bayesian credible set resembles the confidence set.**

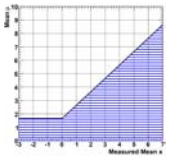
Taking "resembles" to the extreme leads to the likelihood principle and breakdown in unconditional coverage.

But as a useful guide for when post-data inference can be misleading, this is a remarkable deep connection between frequentist confidence intervals (statements about P(data|parameter)) and credible intervals (statements about P(parameter|data)) !

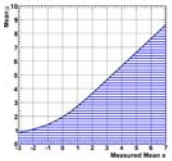
Deep Connections to Bayesian Statistics (cont.)

Beginning in 2000, statistician Jim Berger has argued at four of our meetings that bad conditional properties can be so hard to detect in frequentist methods that one is better off using Bayesian methods with priors known to have approximate unconditional coverage.

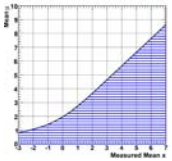
Five methods used for bounded Gaussian mean problem



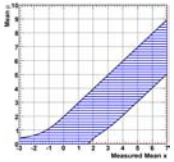
1) **1960's and beyond:**
   UL = max(x, 0) + 1.64σ



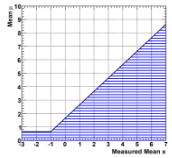
2) **1979 "PDG" (real 1986 PDG) and beyond:**
   Bayesian with uniform prior



3) **1997: Alex Read et al. (LEP)**
   CL_S



4) **1997: Feldman and Cousins (NOMAD)**
   Unified Approach



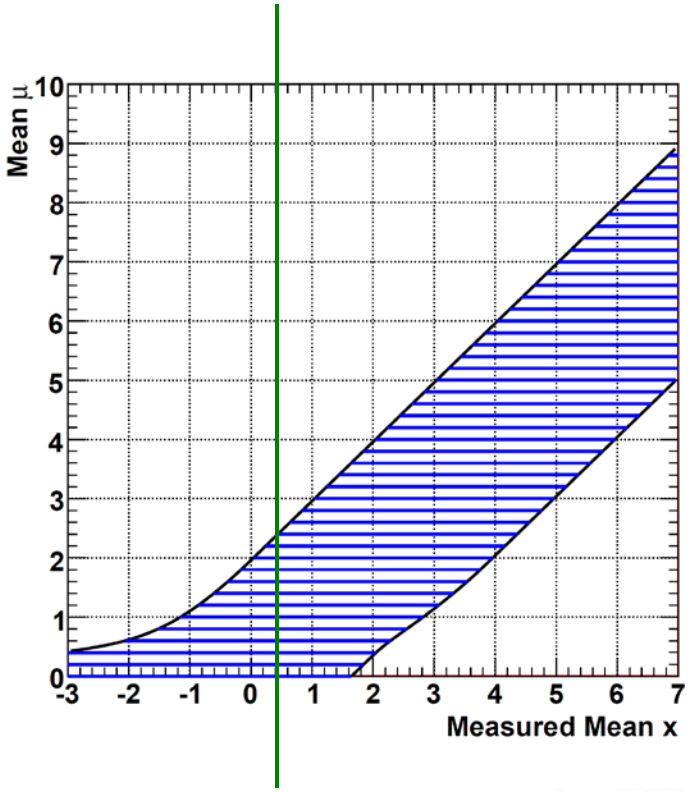
5) **2010: Power Constrained Limits;**
   Cowan, Cranmer, Gross, Vitells (ATLAS):
   UL = max(0, max(x, x_{PCL}) + 1.64σ)

Neyman's Confidence Interval construction, 1934-37

Given $p(x|\mu)$ from a model: For each value of μ, draw a horizontal *acceptance interval* $[x_1,x_2]$ such that $p(x \in [x_1,x_2] \mid \mu) = 1 - \alpha$.

Upon performing expt and obtaining the value x_0, draw the vertical line through x_0.

The vertical *confidence interval* $[\mu_1, \mu_2]$ with C.L. = $1 - \alpha$ is the union of all values of μ for which the corresponding *acceptance interval* is intercepted by the vertical line.



Unified Approach of Feldman and Cousins

Starting points:

1) Remove null intervals

2) 95% coverage for *all* μ.

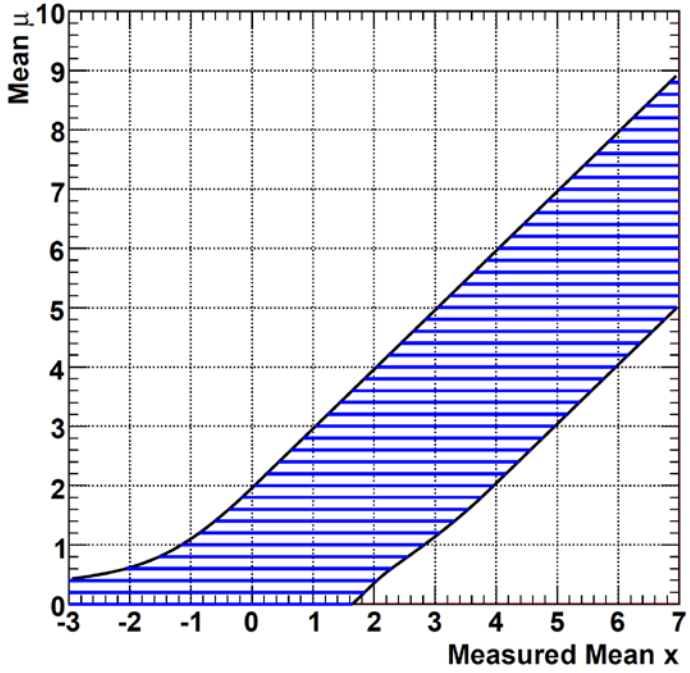
***Immediately:* 95% acceptance interval for μ=0 is [– ∞, 1.64].**

Leads to *Unified Approach*: [μ_1,μ_2]

1) **For low and negative *x*, μ_1=0.**
2) **μ=0 excluded when rejected by *one*-tailed test at 1–C.L. (!)**
3) **At large *x*, [μ_1,μ_2] converges to *central* interval.**

[Above seen by S. Ciampolillo, who also moved *x*<0 to 0.] F-C:

4) **Interval based on $\Delta\chi^2$ (L.R.)**
5) **Cures "flip-flop" problem.**



Phys Rev D57 3873 (1998)

Phys. Rev. D57 3873 (1998)

Unified approach to the classical statistical analysis of small signals

Gary J. Feldman*
Department of Physics, Harvard University, Cambridge, Massachusetts 02138

Robert D. Cousins†
Department of Physics and Astronomy, University of California, Los Angeles, California 90095

"Test for $\theta=\theta_0$" ↔
"Is θ_0 in confidence interval for θ"

Using the Likelihood Ratio Test, this correspondence is the basis of the "Unified Approach" intervals/regions of F-C.

In Gaussian problem, $-2\ln(LR) = \Delta\chi^2$.

"Unified Approach" solves "flip-flopping problem – see paper.

Generalizes well.

Kendall and Stuart

CHAPTER 22

LIKELIHOOD RATIO TESTS AND TEST EFFICIENCY

The LR statistic

22.1 The ML method discussed in Chapter 18 is a constructive method of obtaining estimators which, under certain conditions, have desirable properties. A method of test construction closely allied to it is the likelihood ratio (LR) method, proposed by Neyman and Pearson (1928). It has played a role in the theory of tests analogous to that of the ML method in the theory of estimation.

As before, we have the LF

$$L(x|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_r, \boldsymbol{\theta}_s)$ is a vector of $r+s=k$ parameters ($r \geq 1$, $s \geq 0$) and x may also be a vector. We wish to test the hypothesis

$$H_0 : \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r0}, \tag{22.1}$$

which is composite unless $s = 0$, against

$$H_1 : \boldsymbol{\theta}_r \neq \boldsymbol{\theta}_{r0}.$$

We know that there is generally no UMP test in this situation, but that there may be a UMPU test – cf. **21.31**.

The LR method first requires us to find the ML estimators of $(\boldsymbol{\theta}_r, \boldsymbol{\theta}_s)$, giving the unconditional maximum of the LF

$$L(x|\hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s), \tag{22.2}$$

and also to find the ML estimators of $\boldsymbol{\theta}_s$, when H_0 holds,[1] giving the conditional maximum of the LF

$$L(x|\boldsymbol{\theta}_{r0}, \hat{\hat{\boldsymbol{\theta}}}_s). \tag{22.3}$$

$\hat{\hat{\boldsymbol{\theta}}}_s$ in (22.3) has been given a double circumflex to emphasize that it does not in general coincide with $\hat{\boldsymbol{\theta}}_s$ in (22.2). Now consider the likelihood ratio[2]

$$l = \frac{L(x|\boldsymbol{\theta}_{r0}, \hat{\hat{\boldsymbol{\theta}}}_s)}{L(x|\hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s)}. \tag{22.4}$$

Since (22.4) is the ratio of a conditional maximum of the LF to its unconditional maximum, we clearly have

$$0 \leq l \leq 1. \tag{22.5}$$

Intuitively, l is a reasonable test statistic for H_0: it is the maximum likelihood under H_0 as a fraction of its largest possible value, and large values of l signify that H_0 is reasonably acceptable. The critical region for the test statistic is therefore

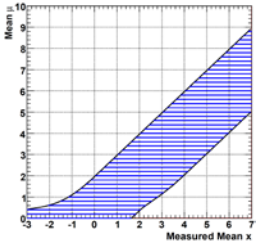
$$l \leq c_\alpha, \tag{22.6}$$

where c_α is determined from the distribution $g(l)$ of l to give a size-α test, that is,

$$\int_0^{c_\alpha} g(l)\,\mathrm{d}l = \alpha. \tag{22.7}$$

Neither maximum value of the LF is affected by a change of parameter from $\boldsymbol{\theta}$ to $\tau(\boldsymbol{\theta})$, the ML estimator of $\tau(\boldsymbol{\theta})$ being $\tau(\hat{\boldsymbol{\theta}})$ – cf. **18.3**. Thus the LR statistic is invariant under reparametrization.

"But Bob, I *insist* on an *upper* limit!"
"Do I need to define *upper* for you?"

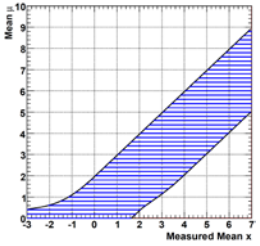


Bob: Let's consider two deep points.

1) Insisting on a CCGV *upper* limit means insisting on *not* rejecting $\mu = 0$ at 95% while simultaneously rejecting μ which has a better $\Delta\chi^2$ than $\mu = 0$ (say when $x = 2$). This is related to the "extra" power of CCGV upper limit when it rejects $\mu = 1$ when $x = -1$.

2) Insisting on an *upper* limit means insisting on over-coverage (unless null intervals are brought back). Intervals with correct coverage, based on $\Delta\chi^2$, allow for more relevant and interpretable post-data inference.

“But Bob, CCGV intervals have more power!”

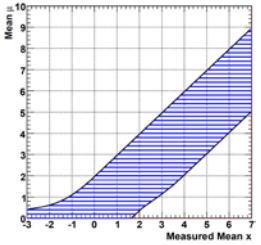


Bob:

The most powerful confidence belt is the original diagonal line with null intervals. It also has perfect coverage.

Yet it bothers most of us. **Power is a pre-data concept which must be supplemented by post-data considerations.**

"But Bob, I don't want to exclude μ=0 unless I have 5σ! "

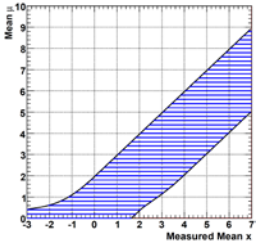


Bob: Let's consider two more points.

1) Reporting a 95% interval which does not include μ=0 is not declaring discovery (or evidence, or indication, or…).

The F-C interval is reporting those values of μ which have the best $\Delta\chi^2(\mu) = \chi^2(\mu) - \chi^2(\mu_{best})$ given the observed *x*. That would seem to be very useful!

"But Bob, I don't want to exclude μ=0 unless I have 5σ! "

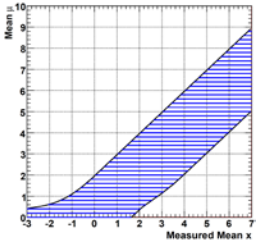


2) A very useful number to report is that value of C.L. for which μ=0 is just included in the F-C interval.

E.g., for *x*=2, μ=0 is in the 97.72% C.L. F-C interval. (1- C.L.$_{FC}$ is just the *one*-sided p-value for 2σ.)

Or one can quote the number of sigma.
This is in fact what we are used to doing!
It all falls out naturally from the "Unified" Approach.

"But Bob, isn't μ too tightly constrained when x<<0?"



Bob: Gleser (above) points out this behavior is consistent with the likelihood principle. It does however call into question the model: the assumption of Gaussian shape and value of σ.

Statistician Woodroofe commenting on Mandelkern: ``The unified method…clearly provides an improvement over the Neyman intervals…however, …it can produce unbelievably short intervals.''

Woodroofe & Sen (2009): add uncertainty to σ, leads to looser constraint for x<<0. This could be more fruitful approach than power constraint.

I think it's a better fit to physicist's thinking (and was in fact the answer for electron neutrino mass!)

Conclusion: Think Hard about This:

	Frequentist Confidence	Likelihood Ratio	Bayesian Credible
Requires prior pdf?	No	No	Yes
Provides P(parameter|data)?	No	No	Yes
Random variable in "$P(\mu_t \in [\mu_\ell, \mu_u])$":	μ_ℓ, μ_u	μ_ℓ, μ_u	μ_t
Coverage guaranteed? "Confidence Principle"	**Yes (but over-coverage…)**	**No**	**No**
Obeys "Likelihood Principle"?	**No**	**Yes**	**Yes (exception re Jeffreys prior)**

I hope you will reach the conclusion, as many of us have, that for "hard" problems one should compare the three methods. For the first column F-C (actually Kendall and Stuart) has many useful features.

Recommended reading

Books: Among the many books available, I usually recommend the following progression, reading the first three cover-to-cover, and consulting the last one as needed:

1) Philip R. Bevington and D.Keith Robinson, Data Reduction and Error Analysis for the Physical Sciences (Quick read for undergrad-level review)

2) Glen Cowan, Statistical Data Analysis (Solid foundation for HEP)

3) Frederick James, Statistical Methods in Experimental Physics, World Scientific, 2006. (This is the second edition of the influential 1971 book by Eadie et al., has more advanced theory, many examples)

4) A. Stuart, K. Ord, S. Arnold, Kendall's Advanced Theory of Statistics, Vol. 2A, 6th edition, 1999; and earlier editions of this "Kendall and Stuart" series. (Authoritative on classical frequentist statistics; anyone contemplating a NIM paper on statistics should look in here first!)

PhyStat conference series: Beginning with Confidence Limits Workshops in 2000, links at http://phystat-lhc.web.cern.ch/phystat-lhc/ and http://www.physics.ox.ac.uk/phystat05/

By now there are many many web pages with lists of statistics references – Google on your favorite topic.

My Bayesian reading list is the set of citations in my Comment, Phys. Rev. Lett. 101 029101 (2008), especially refs 2, 8, 9, 10, 11 (and 7 for model selection)

References Cited in Talk Slides

Berger00: Jim Berger, "Objective Bayesian Analysis and Frequentist Statistics", talk at Fermilab Confidence Limits Workshop, March 2000. See also his talk at PhyStat-LHC at CERN, 2007.

Cousins05: Robert Cousins, "Treatment of nuisance parameters in high energy physics, and possible justifications and improvements in the statistics literature", PhyStat05: Statistical Problems in Particle Physics, Astrophysics and Cosmology, Oxford, 12-15 Sept. 2005.

James06: Frederick James, Statistical Methods in Experimental Physics, World Scientific, 2006.

Kass96: Robert E. Kass, Larry Wasserman, "The Selection of Prior Distributions by Formal Rules" J. Amer. Stat. Assn. 91 1343 (1996)

Reid95: N. Reid, "The Roles of Conditioning in Inference", Statistical Science 10 138 (1995).

Stuart99: A. Stuart, K. Ord, S. Arnold, Kendall's Advanced Theory of Statistics, Vol. 2A, 6th edition, 1999; and earlier editions by Kendall and Stuart.

A Selected Reading List re Bayesian Priors

Robert E. Kass and Larry Wasserman, "The Selection of Prior Distributions by Formal Rules," J. Amer. Stat. Assoc. 91 1343 (1996).

Telba Z. Irony and Nozer D. Singpurwalla, "Non-informative priors do not exist: A dialogue with Jose M. Bernardo," J. Statistical Planning and Inference 65 159 (1997).

J.O. Berger and L.R. Pericchi, "Objective Bayesian Methods for Model Selection: Introduction and Comparison," in Model Selection, Inst. of Mathematical Statistics Lecture Notes-Monograph Series, ed. P. Lahiri, vol 38 (2001) pp .135-207

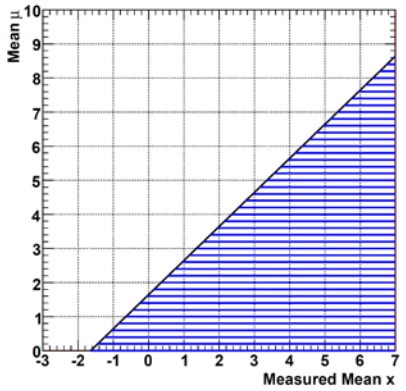
James Berger, "The Case for Objective Bayesian Analysis," Bayesian Analysis 1 385 (2006)

Michael Goldstein, "Subjective Bayesian Analysis: Principles and Practice," Bayesian Analysis 1 403 (2006)

BACKUP

The diagonal line rejects values of μ partially based on *absolute* χ^2 rather than $\Delta\chi^2$ with respect to best fit.

$\chi^2(\mu) = (x - \mu)^2$; $\mu \geq 0$.



For $x = -1$: min χ^2 is at μ=0: $\chi^2(\mu=0) = 1$.

UL from diagonal line is UL = 0.64.

Note that $\chi^2(\mu = 0.64) = (-1 - 0.64)^2 = 2.70$.

Interval only includes μ for which χ^2 itself (*not* $\Delta\chi^2$!) is less than "book value" $\Delta\chi^2 = 2.70$ for *1*-sided limit!

Such "goodness of fit" intervals are known to have problem in other contexts.

So: try to use $\Delta\chi^2(\mu) = \chi^2(\mu) - \chi^2(\mu_{best})$.

How to make correspondence between $\Delta\chi^2$ and C.L.? The answer to that would not come until 1998.

Confidence Intervals and Coverage

Let μ_t be the unknown true value of μ. In repeated experiments, confidence intervals will have different endpoints $[\mu_1, \mu_2]$, since the endpoints are functions of the randomly sampled *x*.

A little thought will convince you that a fraction C.L. = 1 – α of intervals obtained by Neyman's construction will contain ("cover") the fixed but unknown μ_t. I.e.,

$P(\mu_t \in [\mu_1, \mu_2])$ = C.L. = 1 – α.

The endpoints μ_1, μ_2 are the random variables (!).

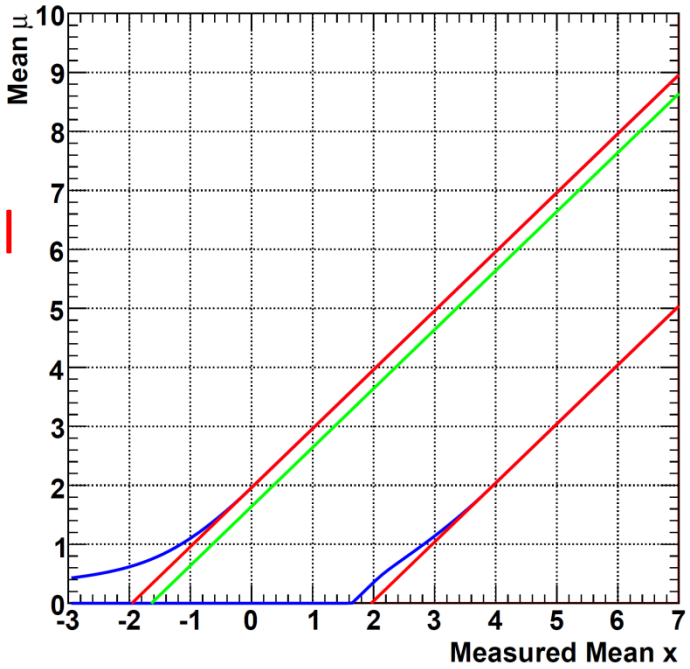
Coverage is a property of the *set* of confidence intervals, not of any one interval.

Unified and Un-Unified Intervals

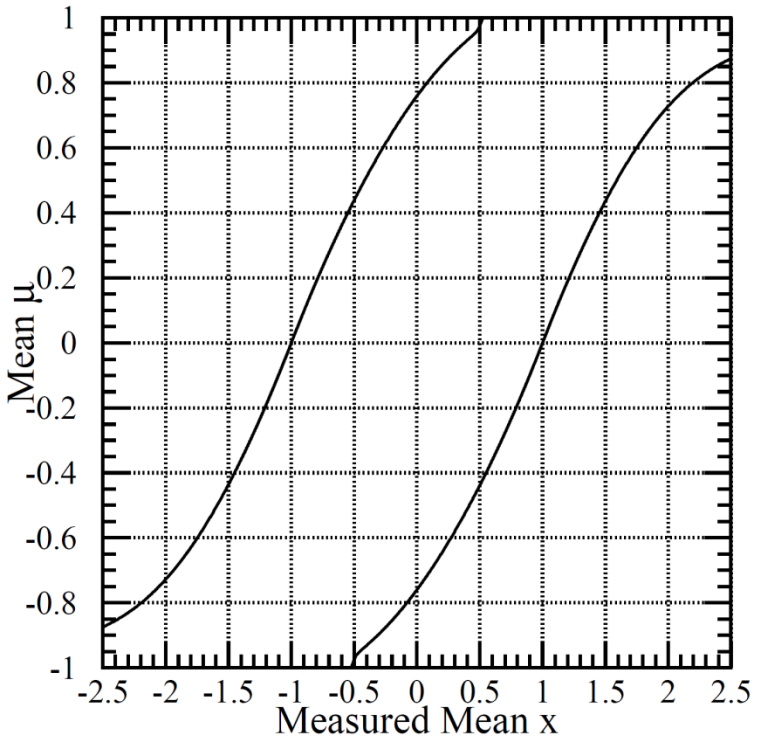
F-C

Traditional central

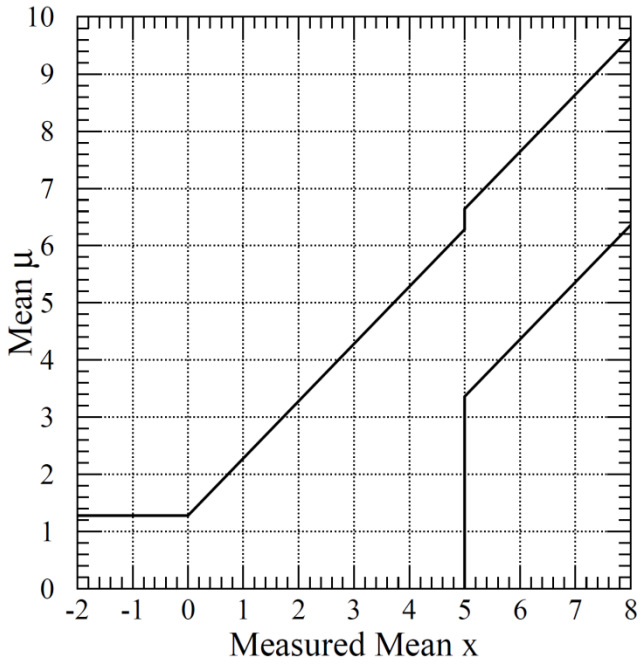
Traditional upper



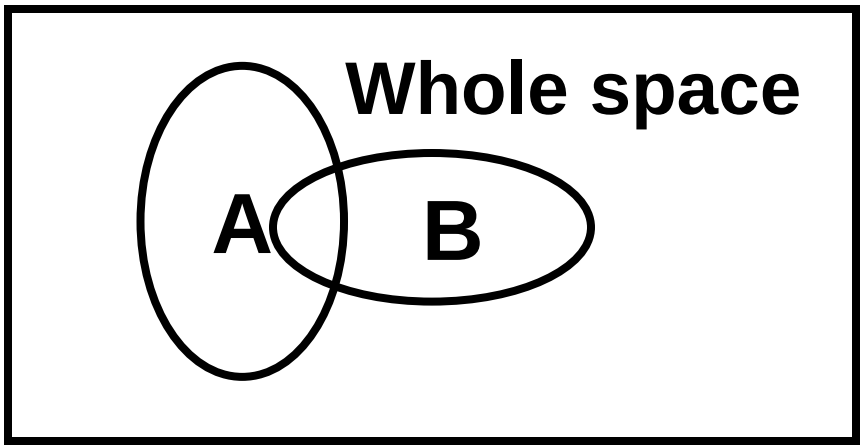
Feldman-Cousins for Two-sided Bound $-1 \leq \mu \leq 1$, $\sigma=1$

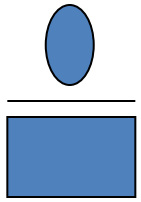


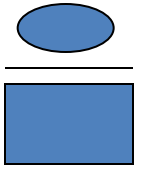
Flip-Flop Plot

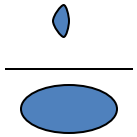


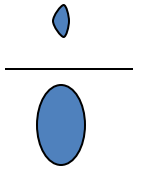
P, Conditional P, and Derivation of Bayes' Theorem in Pictures

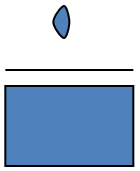


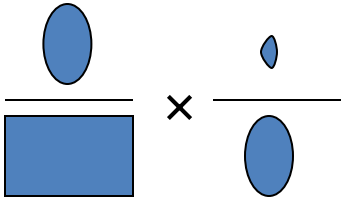
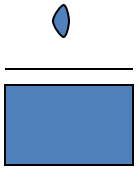
$P(A) =$ 

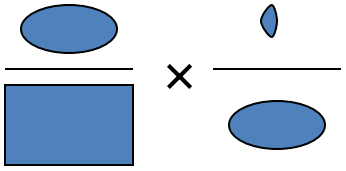
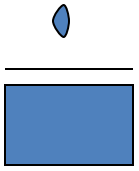
$P(B) =$ 

$P(A|B) =$ 

$P(B|A) =$ 

$P(A \cap B) =$ 

$P(A) \times P(B|A) =$  $=$  $= P(A \cap B)$

$P(B) \times P(A|B) =$  $=$  $= P(A \cap B)$

$\Rightarrow P(B|A) = P(A|B) \times P(B) / P(A)$