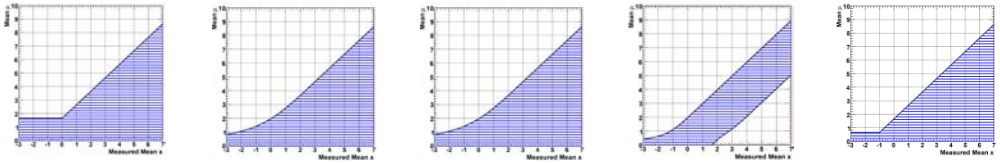


Bayes, Fisher, Neyman, Neutrino Masses, and the LHC



Bob Cousins
Univ. of California, Los Angeles

Virtual Talk
12 September 2011

This talk tells a story of remarkable facts and controversies about a problem very simple to state:

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 $\mu \geq 0$?

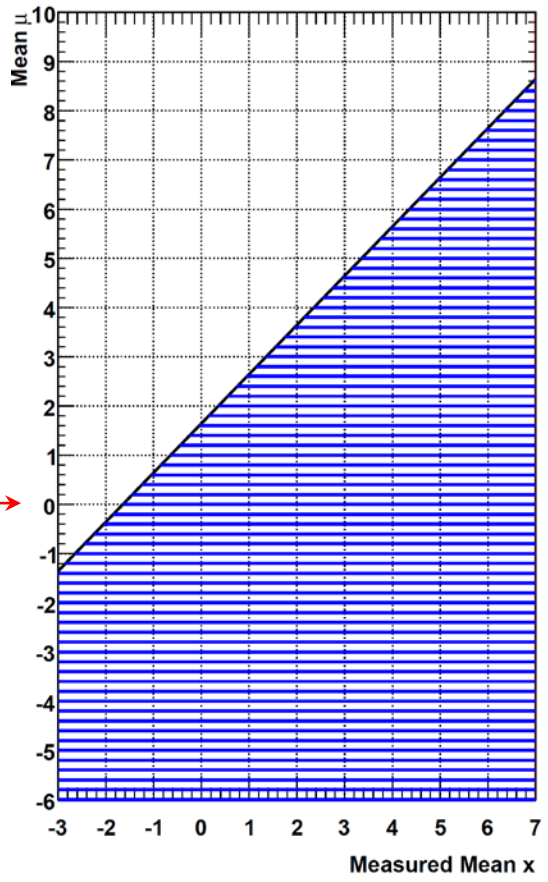
***Without the constraint on μ*, traditional frequentist and Bayesian methods both yield:**

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

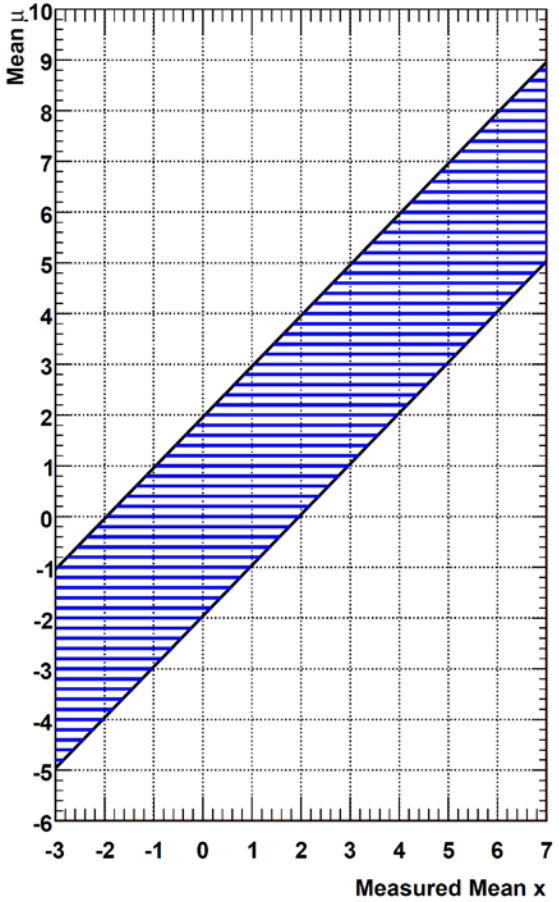
and 95% C.L. central confidence interval is $x \pm 1.96\sigma$. See next slide:

Graphical display of intervals is a *confidence belt:* Confidence interval include all values of μ for which horizontal blue line is intersected.

μ=0 →

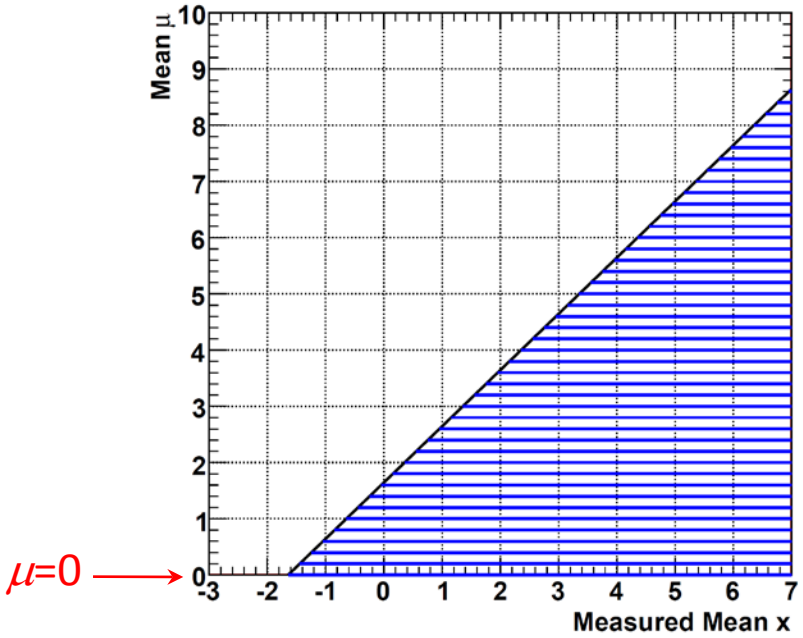


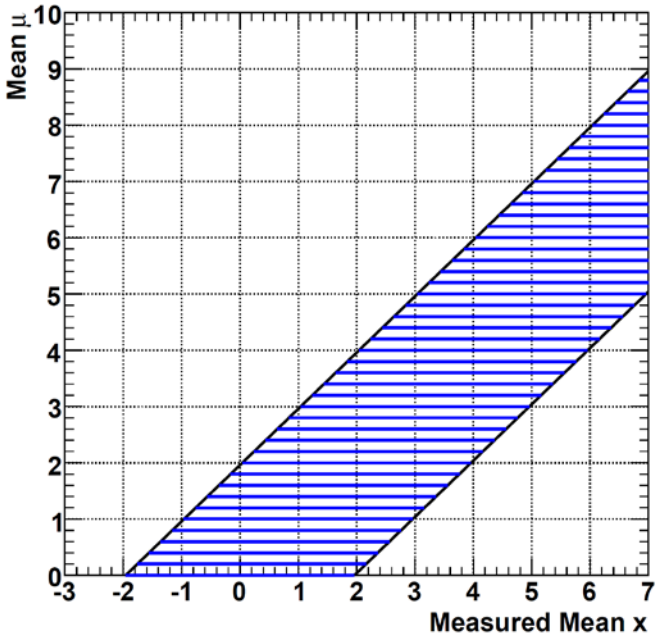
Upper limit = x + 1.64 σ



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!**



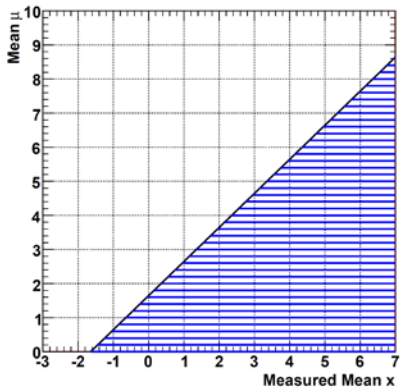


For $x < -1.64\sigma$ with UL, and for $x < -1.96\sigma$ with central intervals, the confidence interval is the *null* set!

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

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.

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 ± 2.5

Set x=0 and proceeded.

NEUTRAL DECAY BRANCHING RATIOS OF THE η^0 MESON

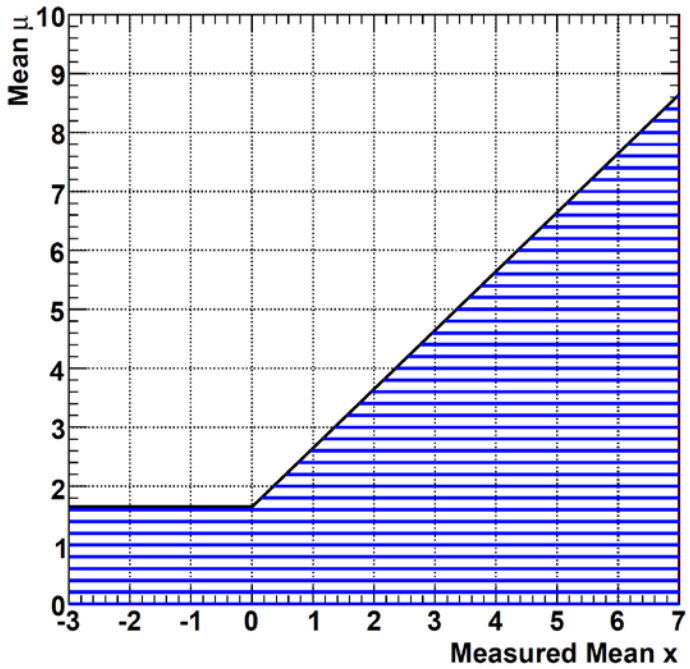
C. Baltay,† P. Franzini, J. Kim, R. Newman, and N. Yeh
Columbia University, New York, New York, and Brookhaven National Laboratory, Upton, New York
L. Kirsch
Brandeis University, Waltham, Massachusetts

PRL 19 1495 (1967)

x = – 0.06 ± 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.**

Much thought was stimulated by experiments directly measuring neutrino masses in the 1970's and 1980's:

- **ν_e mass (tritium β decay),**
- **ν_μ mass (π decay), and later**
- **ν_τ mass (τ decay).**

Here $x = m_\nu^2 = E_\nu^2 - p_\nu^2$ was typically Gaussian.

For $m_{\nu_e}^2$, the issue became acute (with >1 lesson):

VALUE (eV^2)			DOCUMENT ID		TECN	COMMENT
– 54	**± 30**	**OUR AVERAGE**				
– 39	± 34	± 15	[14] WEINHEIMER	93	SPEC	^{3}H β decay
– 24	± 48	± 61	[15] HOLZSCHUH	92B	SPEC	^{3}H β decay
– 65	± 85	± 65	[16] KAWAKAMI	91	SPEC	$\overline{\nu}_e$, tritium
–147	± 68	± 41	[17] ROBERTSON	91	SPEC	$\overline{\nu}_e$, tritium

1995 PDG RPP: "Caution is urged in interpreting this result" for UL.

But even when obtaining x > 0, the presence of the boundary influenced some physicists.

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

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

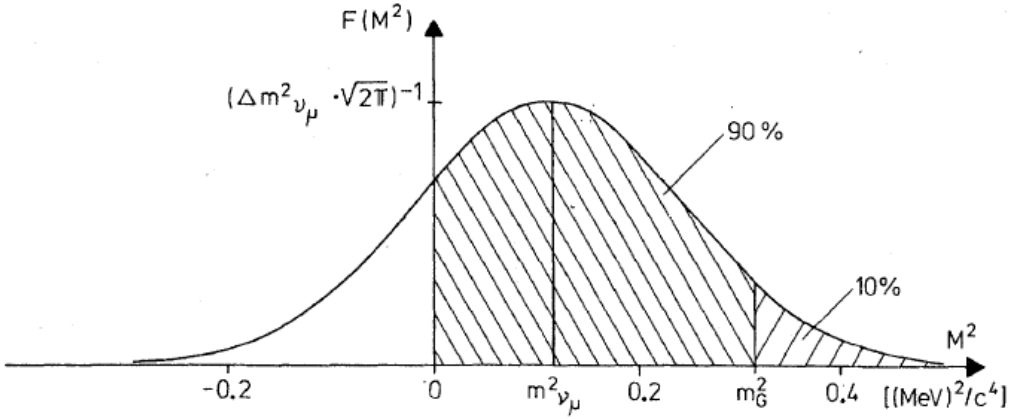


FIG. 22. According to the prescription of the Particle Data Group (Ref. 33) the upper limit m_G of the muon-neutrino mass is calculated from the squares mass $m_{\nu_\mu}{}^2$ and its uncertainty $\Delta(m_{\nu_\mu}{}^2)$ by setting the probability function $F(M^2)$ to zero for $M^2 < 0$, as indicated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

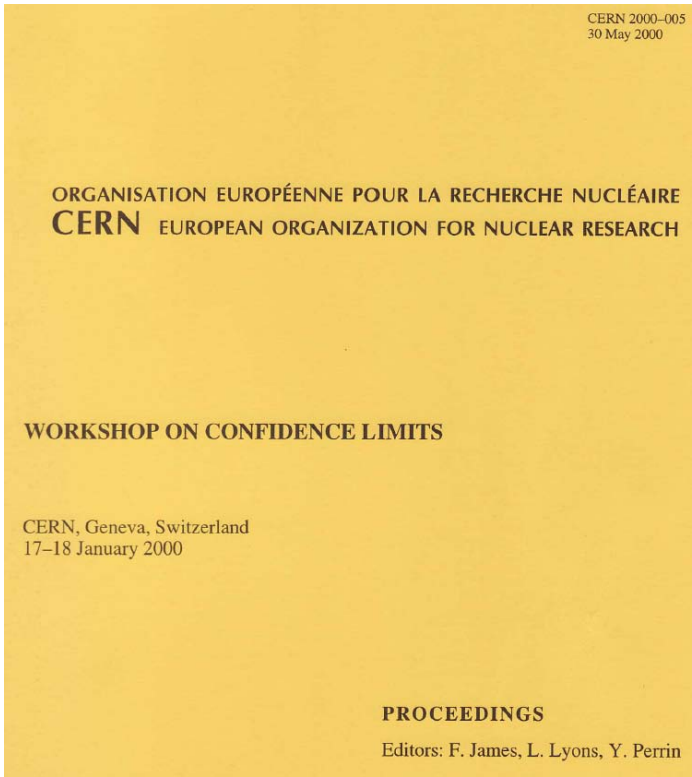
[33] T. G. Trippe, private communication, 1976.

These physicists, while perhaps unschooled in foundations of statistics, had important insights in the 1960's through 1980's. More progress with better foundations followed in the 1990's.

Gary Feldman and I concluded in 1998 that part of the problem was in rigidity of the question asked, advocating a Unified Approach with 2-sided intervals.

Confidence Limits Workshops at CERN and Fermilab in 2000 brought together many of us (filled CERN Council Chamber), with adherents of three main methods, all in PDG RPP since 2002.

http://cdsweb.cern.ch/record/411537/files/CERN-2000-005.pdf



Since 2000, statisticians have pointed us to yet more insights in their 50-year-old (!) literature that we in HEP had missed, and made fresh comments.

My conclusion, strengthened by these more recent insights: it is *not* wise for HEP to depart from the 2000-era methods for upper limits and the Unified Approach.

The argument is deep and brings in more than one dispute among giants of 20th-century statistics.

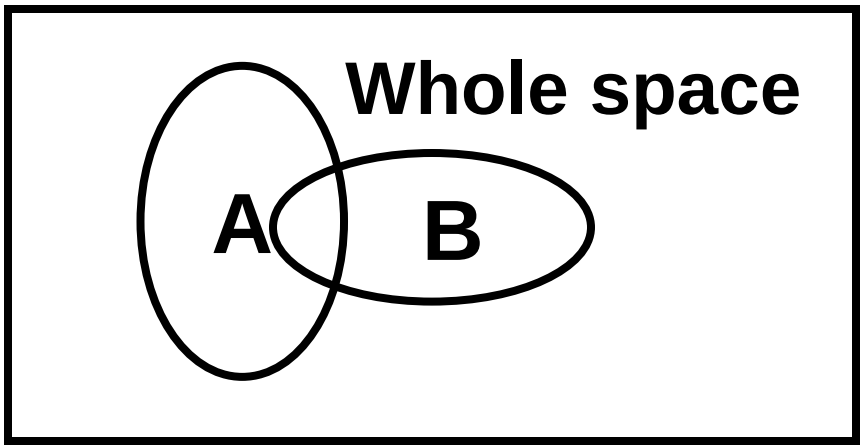
So to explain why some of us have reached this conclusion, I'll ask you to understand five ingredients, and then we'll put them together.

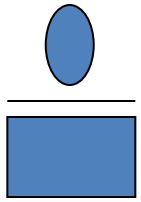
Outline of Five Ingredients

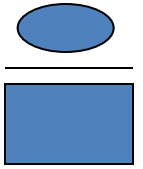
1) **Bayes's Theorem and Bayesian credible intervals**
2) **Neyman's construction for confidence intervals, and the concept of coverage**
3) **Neyman-Pearson hypothesis testing, and concepts of Type I and Type II errors, and power**
4) **The equivalence between Neyman's intervals and N-P hypothesis testing**
5) ***** Pre-data vs post-data inference, and probabilities *conditioned* on the observed data: *Frequentist* criticisms of most-powerful tests.**

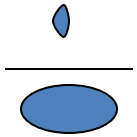
A bit of a "crash course", but this simplest example is rich in statistical issues!

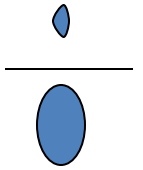
P, Conditional P, and Derivation of Bayes' Theorem in Pictures

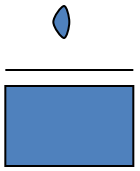


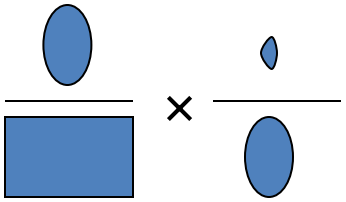
$P(A) =$ 

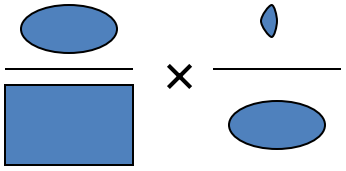
$P(B) =$ 

$P(A|B) =$ 

$P(B|A) =$ 

$P(A \cap B) =$ 

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

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

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

What is the “Whole Space”?

For probabilities to be well-defined, the “whole space” needs to be defined, which in practice introduces assumptions and restrictions.

Thus the “whole space” itself is conditional on the assumptions going into the model (possible outcomes.)

Furthermore, in frequentist statistics, restricting the “whole space” to a *relevant subspace* can sometimes improve the quality of statistical inference – this is a crucial point in the discussion below.

Bayes' Theorem Generalized to Probability Densities

Recall $P(B|A) \propto P(A|B)\, P(B)$.
For Bayesian P, parameters are random variables which can appear in conditional probabilities.

Let $p(x|\mu)$ be conditional pdf for data x, given parameter μ.
Then Bayes' Theorem becomes

$p(\mu|x) \propto p(x|\mu)\, p(\mu)$.

Substituting in a particular set of observed data, x_0 :

$p(\mu|x_0) \propto p(x_0|\mu)\, p(\mu)$. **Recognizing the likelihood,**

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

$p(\mu|x_0)$ = posterior pdf for μ, given the results of this expt

$\mathcal{L}(x_0|\mu) = \mathcal{L}(\mu)$ = Likelihood function of μ from this expt

$p(\mu)$ = prior pdf for μ, before updating with result of this expt

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.

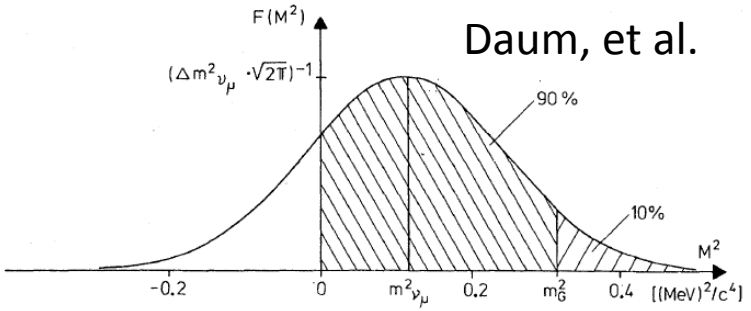
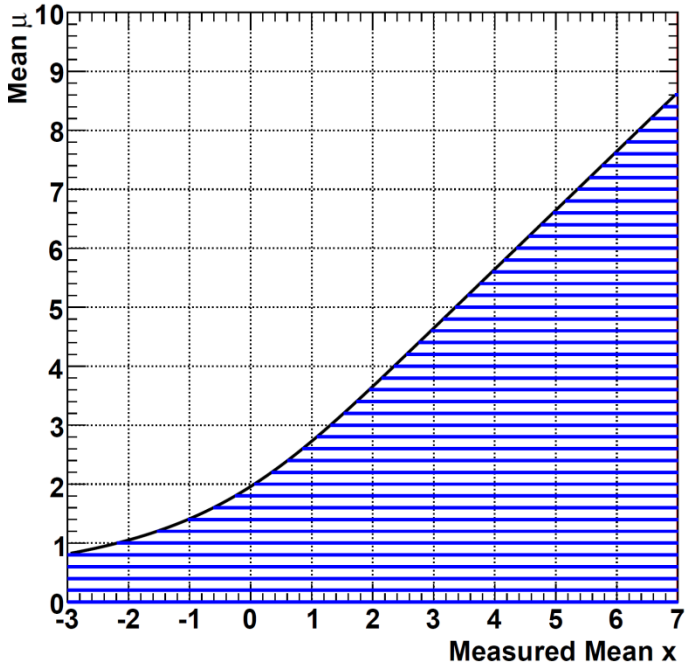


FIG. 22. According to the prescription of the Particle Data Group (Ref. 33) the upper limit m_G of the muon-neutrino mass is calculated from the squares mass $m_{\nu_\mu}^2$ and its uncertainty $\Delta(m_{\nu_\mu}^2)$ by setting the probability function $F(M^2)$ to zero for $M^2 < 0$, as indicated in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



Confidence Intervals

"Confidence intervals", and this phrase were invented by Jerzy Neyman in 1934-37.
They use the frequentist definition of P.

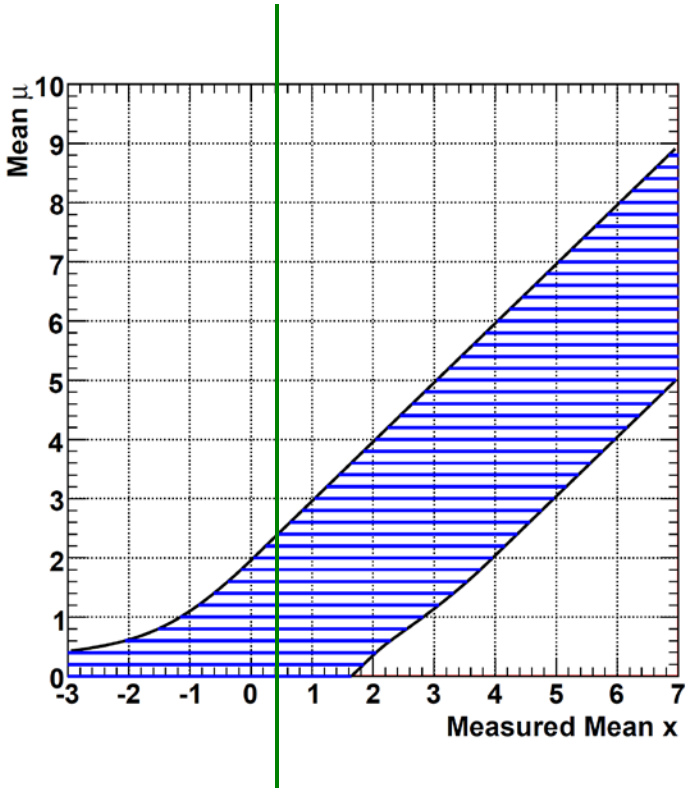
The next two slides give some basic points.
It takes a bit of time to sink in – given how often confidence intervals are misinterpreted, the argument is perhaps a bit too ingenious.

Neyman's Confidence Interval construction

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.



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.

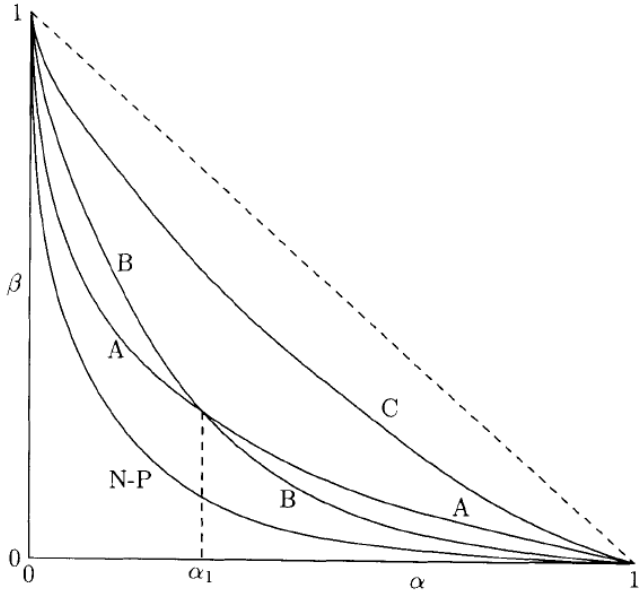
Neyman-Pearson Hypothesis Testing (1933)

Frame discussion in terms of null hypothesis, e.g.,
H_0 = S.M., and an alternative H_1 = your favorite SUSY model.

- **α: probability (under H_0) of rejecting H_0 when it is true, i.e., false discovery claim (Type I error)**
- **β: probability (under H_1) of accepting H_0 when it is false, i.e., not claiming a discovery when there is one (Type II error)**
- **θ: parameters in the hypotheses**

Competing hypothesis tests A, B, and C can be compared by looking at graphs of β vs α at various θ, and at graphs of β vs θ at various α (power function).

N-P Hypothesis Testing (cont.)



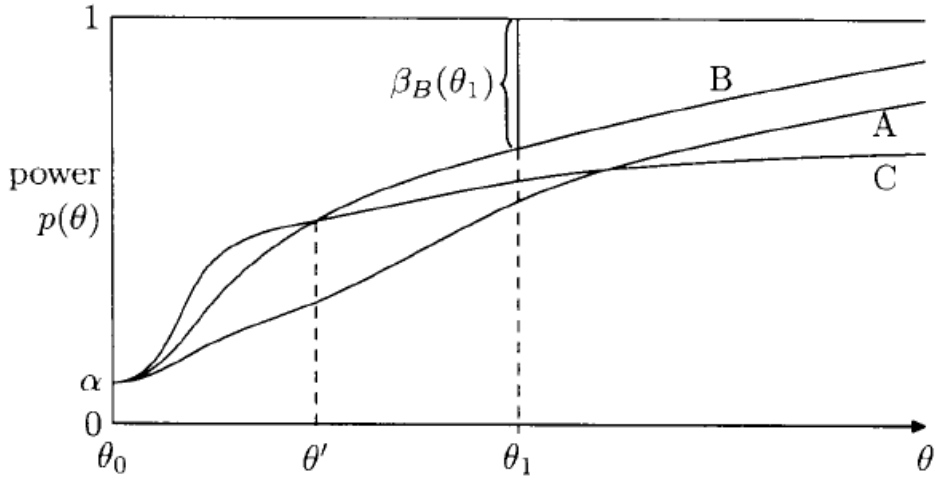


Fig. 10.3. Power functions of tests A, B, and C at significance level α. Of these three tests, B is the best for $\theta > \theta'$. For smaller values of θ, C is better.

Where to live on the β vs α curve is a *long* discussion.

***Decision* to declare discovery requires two more inputs:**

1) **Prior belief in H_0 vs H_1**
2) **Cost of Type I error (false discovery claim) vs cost of Type II error (missed discovery)**

Which test is most powerful can depend on value of unknown θ.

(With *no* parameters, N-P Lemma proved L.R. test is m.p.)

F. James, *Statistical Methods in Experimental Physics*, pp. 258, 262

N-P tests and Neyman's construction are equivalent

The N-P test for $\theta=\theta_0$ with Type I error probability α is equivalent to "Accept H_0 if θ_0 is in the confidence interval for θ with C.L. = 1 – α"

"There is thus no need to derive optimum properties separately for tests and for intervals; there is a one-to-one correspondence between the problems…"
– Kendall & Stuart

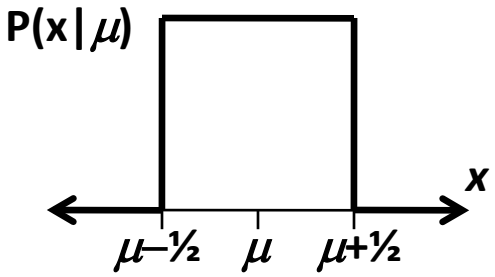
Insights by Sir Ronald Fisher in 1956 and Sir David Cox in 1958 pointed to situations in which Most Powerful Neyman-Pearson tests gave answers clearly not relevant to the data at hand!

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.

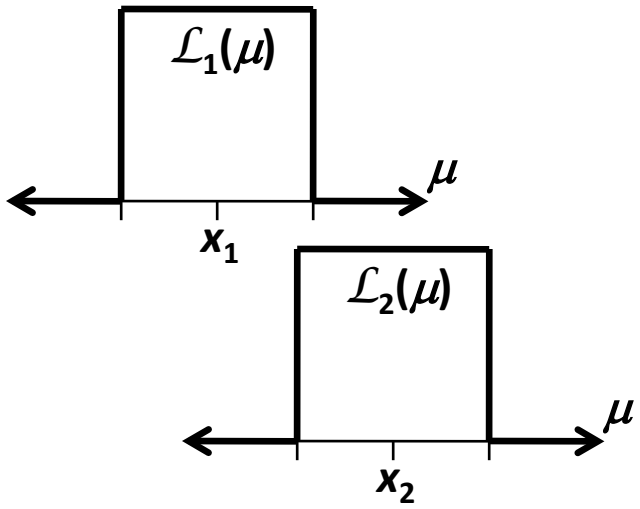
What is a central confidence interval for μ ?

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.

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

Setting Confidence Intervals for Bounded Parameters

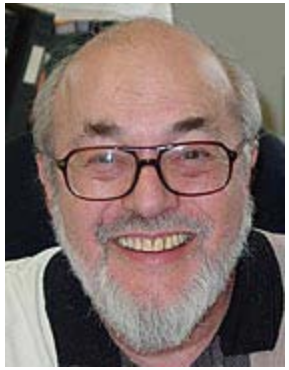
Mark Mandelkern

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."

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."

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."

Betting Game Inspired by Buehler (1959)

Suppose Peter uses a set of confidence intervals with perfect coverage, $P(\mu_t \in [\mu_1, \mu_2]) = \text{C.L.} = 1-\alpha$.

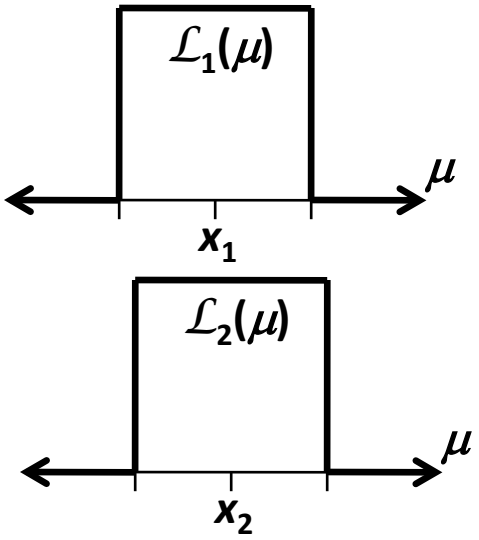
Paula proposes that Peter be willing to bet at odds $(1-\alpha)/\alpha$ that $\mu_t \in [\mu_1, \mu_2]$ in repetitions of the experiment, and that Paula gets to decide whether or not to accept the bet *using only the information available to Peter*, namely $P(x|\mu)$ and *the value of x observed*.

Peter accepts… his intervals have exact coverage!

Betting Game Inspired by Buehler (cont.)

Suppose Paula identifies a set C in the *sample* space such that if $x \in C$, then $P(\mu_t \in [\mu_1, \mu_2]) < 1 - \alpha$ for all μ_t .

In the rectangular example above, such a set C can be simply defined by $|x_1 - x_2|$ being small enough.

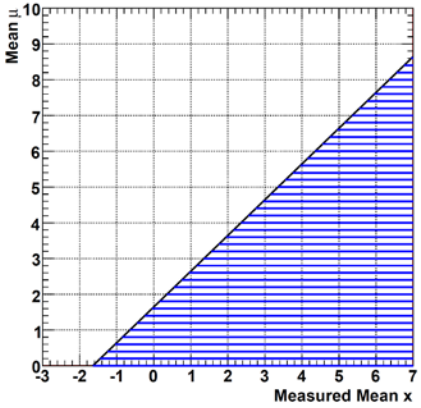


$\mu = (x_1 + x_2)/2 \pm 0.22$ under-estimates uncertainty when $|x_1 - x_2| << 1$

Paula can win in long term by betting (only) when $x \in C$!

The existence of *such set C, called a negatively biased relevant subset,* strongly calls into doubt the use of pre-data uncertainty (coverage) as post-data uncertainty!

So let's return to the upper limits of the "original diagonal line", with new insights.



We always knew that Paula could win by accepting bet when x < –1.64.
Fred James and I called this the problem of

"What do I do when I *know* I am in the wrong 5%?"

So Paula can select bets with 0% chance of losing.
But to win in the long run, she need only select bets with probability of losing strictly less than 95%!

This can be easily done: Choose any constant K, and consistently accept the bet if and only if x<K.

For example, Paula can define a relevant subset C by x : x<0.7. So she bets against Peter's assertion that $\mu_t \leq$ UL at 19:1 odds whenever x<0.7.

To see how she fares, we need to calculate, for each μ, the conditional coverage probability

$P(\mu_t \leq UL \mid x<0.7) = P(x \geq \mu_t - 1.64 \mid x<0.7)$.

This probability is *maximum* for μ_t =0, in which case it is

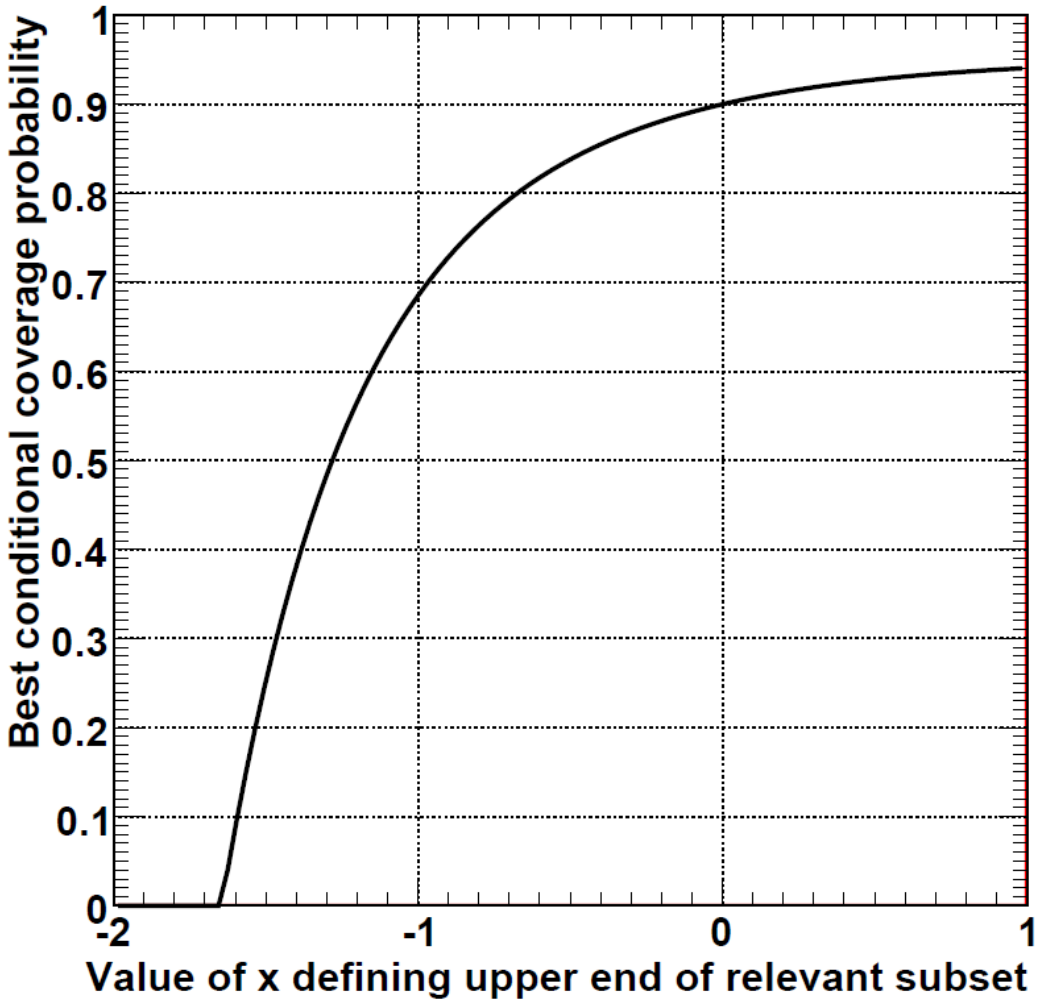
$P(x \geq -1.64 \mid x<0.7)$, for $P(x)$ = Gaussian with mean 0.

Answer: 1 – (0.05/0.758) = 93.4%, *negatively biased*.

The true conditional odds in Peter's favor are *at most* 0.934/(1 – 0.934), about 14:1, so Paula will win in the long run if Peter pays out at 19:1 odds on the bets she makes.

Fraction of Bets won by Peter depends on true μ_t.

Maximum Fraction of Bets Won by Peter when Paula bets against His 95% C.L. UL for *x*<K, as Function of K:



For K=0, failure to cover is twice (1 – C.L.).

This result quantifies the difficulty understood intuitively by past physicists, and connects it to a body of statistics literature going back 50+ years!

Our simply stated problem is in one of the thorniest corners of the statistics literature: what to do when one *knows* post-data that the pre-data coverage probability is inapplicable to the "recognizable subset" containing the observed *x*.

The BIG LESSON: if all your discussions/arguments consider only N-P coverage and power, you can be missing important considerations about post-data inference.

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.

Jim Berger:

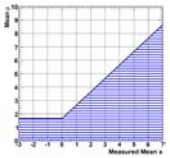
M. Kendall, giving the 'old' frequentist viewpoint of Bayesian analysis;

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

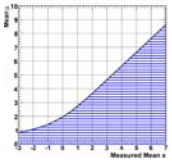
What should be the view today;

Objective Bayesian analysis is the best frequentist tool around.

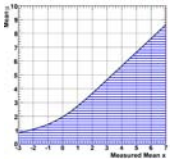
Five methods used for bounded Gaussian mean problem



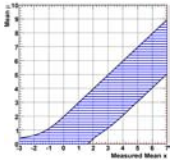
1) **1960's and beyond:**
   UL = max(x, 0) + 1.64σ



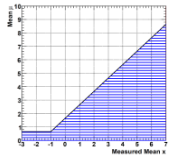
2) **1979 "PDG" (real 1986 PDG) and beyond:**
   Bayesian with uniform prior



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)

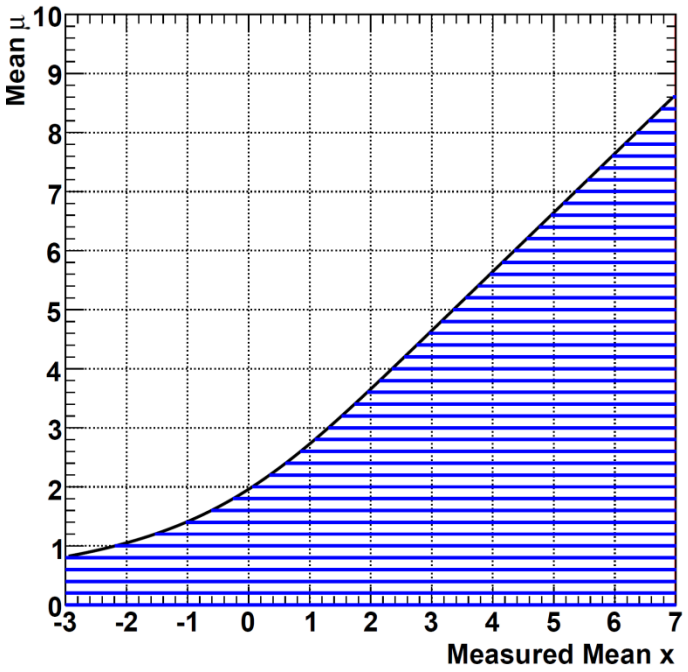
Bayesian with Uniform Prior

Tradition in HEP is to use uniform prior for both Gaussian mean and Poisson mean. Leads to over-coverage in Poisson with bkgnd.

Modern "objective" Bayesians use uniform prior for Gaussian mean but Jeffreys prior $1/\sqrt{\mu}$ for Poisson mean. This leads to undercoverage for some μ.

Bayesian derivation replaces sensitivity to irrelevant data with sensitivity to prior.

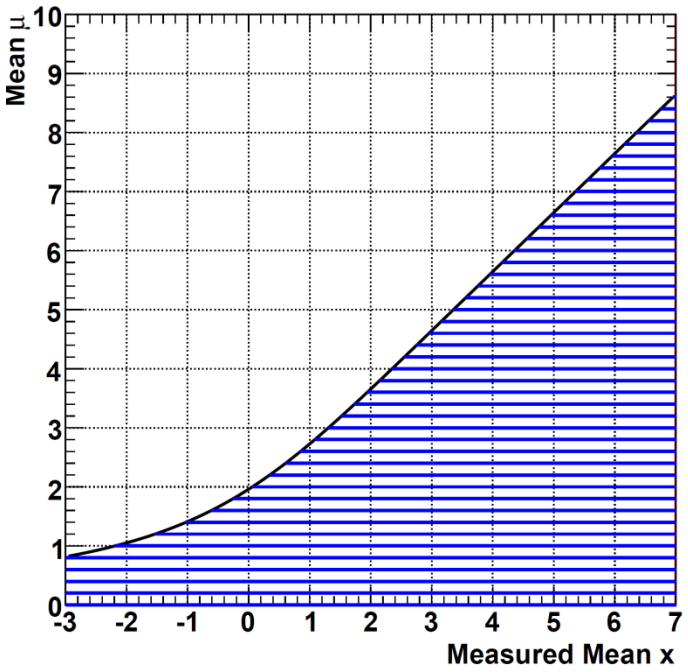
In HEP, there is good experience with priors in low dimensions, some naïveté in high D.



Recent (2010) work on "reference priors" (Demortier, Jain, Prosper) may lead to shift in priors used in HEP.

CL_S

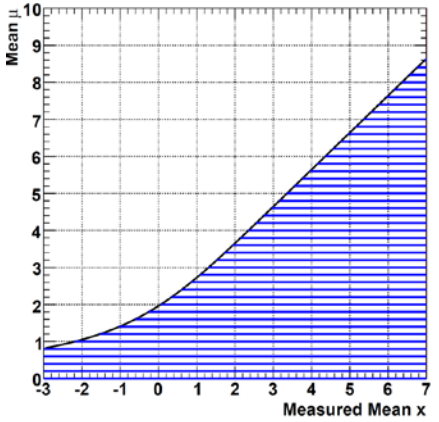
In 1988, for Poisson-with-background problem, Günter Zech performed a frequentist construction of upper limits using non-standard conditional probabilities. It gave the same numerical results as O. Helene's Bayesian calculation (P(μ) flat).



In 1997, Alex Read et al interpreted Zech's formula as ratio of two tail probabilities and applied it more generally: CL_S.

For the current problem, it again gives same result as Bayesian with P(μ) flat (!).

Foundations (?) of CL_s



As an intuitive development in HEP, it appears that a firm foundation for CL_s is lacking.

From Alex Read in 2000 CLW Yellow Report:

A confidence limit is an upper limit if the exclusion confidence is less than the specified confidence level for all values of the population parameter above the confidence limit.

"Note that confidence intervals obtained in this manner do not have the same interpretation as traditional frequentist confidence intervals nor as Bayesian credible intervals." (italics in original)

BC: both CL_s's "reasonable" performance and its over-coverage are probably related to its roots in Bayesian answers for simple problems, with flat prior in Poisson mean avoiding problems for UL.

Unified Approach of Feldman and Cousins

Starting points:

1) Remove null intervals

2) 95% coverage for *all* μ.

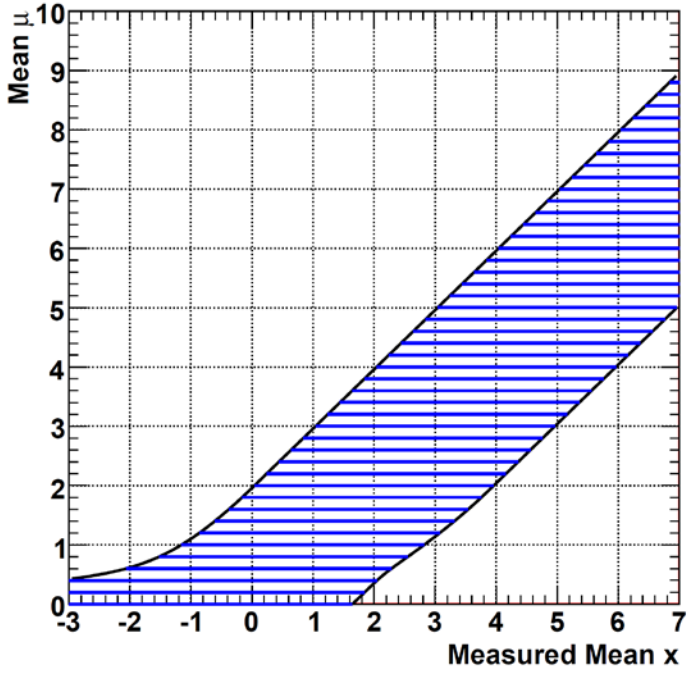
***Immediately:* 95% acceptance interval for μ=0 is [– ∞, 1.64].**

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)

Unified Approach of Feldman and Cousins (cont.)

With diagonal line, interval uses χ^2 instead of $\Delta\chi^2$. Recall above: "How to make correspondence between $\Delta\chi^2$ and C.L.?"

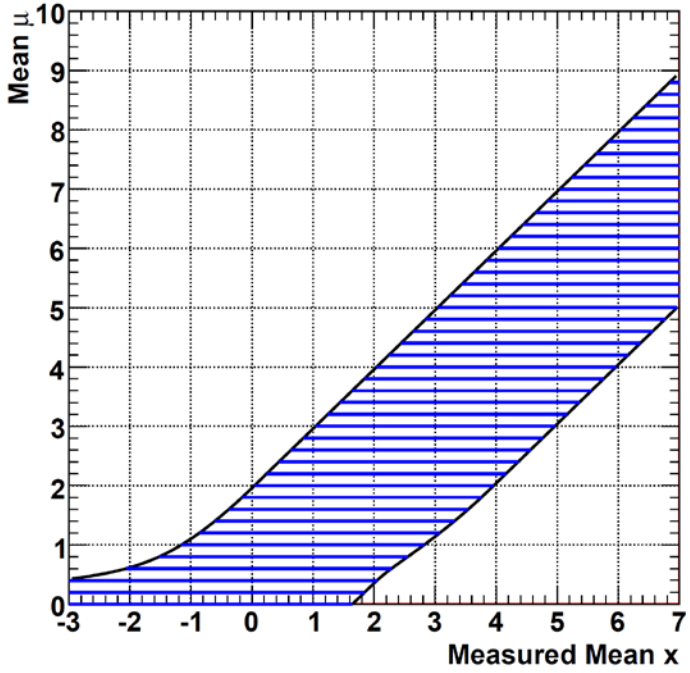
F-C: associate a value of $\Delta\chi^2$ = $\chi^2(\mu_t) - \chi^2(\mu_{best})$ with each *true* value μ_t. The endpoints of its acceptance interval have that $\Delta\chi^2$.

Acceptance interval has those values of x ranked in 95% by $\Delta\chi^2$.

Given x_0, confidence interval contains those values of μ for which x_0 is in top 95% rank by $\Delta\chi^2$.

Automatically includes or excludes μ=0 based on $\Delta\chi^2$.

Works for 3σ, 5σ as one wishes.



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

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,

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

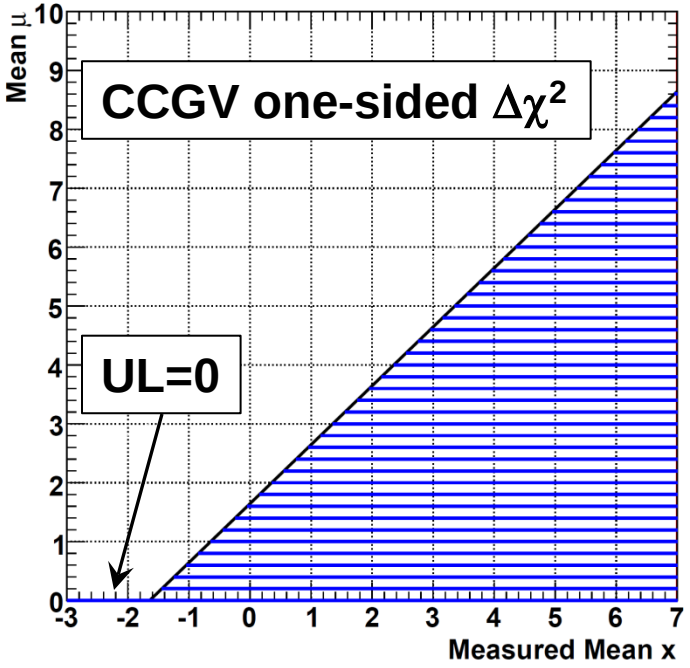
Neither maximum value of the LF is affected by a change of parameter from $\boldsymbol{\theta}$ to $\tau(\boldsymbol{\theta})$, the ML estimator of $\tau(\boldsymbol{\theta})$ being $\tau(\hat{\boldsymbol{\theta}})$ – cf. **18.3**. Thus the LR statistic is invariant under reparametrization.

Limits of Cowan, Cranmer, Gross, and Vitells (CCGV)

***Before* Power Constraint: uses $\Delta\chi^2$ but forces acceptance interval to be one-sided, even on boundary.**

Result: same old diagonal line as with *absolute* χ^2, except with null intervals replaced by UL=0 for x < –1.64.

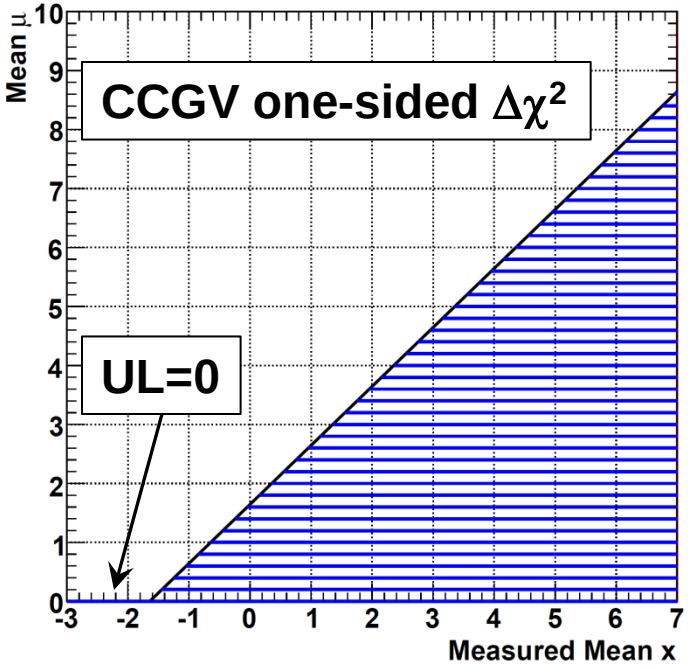
Coverage of μ=0 is 100% (!)



http://arxiv.org/abs/1105.3166

CCGV and the Betting Game

Recall: With original diagonal line, Paula can guarantee that odds in Peter's favor are no better than e.g., 14:1 odds by betting when x<0.7.



In whole "relevant subset" literature, acceptance regions have $p(x \in [x_1,x_2] \mid \mu)$ = C.L.

But CCGV acceptance region for $\mu = 0$ has p=100%, not 95% !

With 100% unconditional coverage for a single value of μ, all the math based on "suprema" of conditional coverage of course breaks down. How (or whether) to adapt whole literature is not agreed on. I think it is hard to claim that the problem simply disappeared. N.B. Paula wins if μ_t has acceptance region $p(x \in [x_1,x_2] \mid \mu_t)$ = C.L. (!)

Power Constrained Limits of CCGV

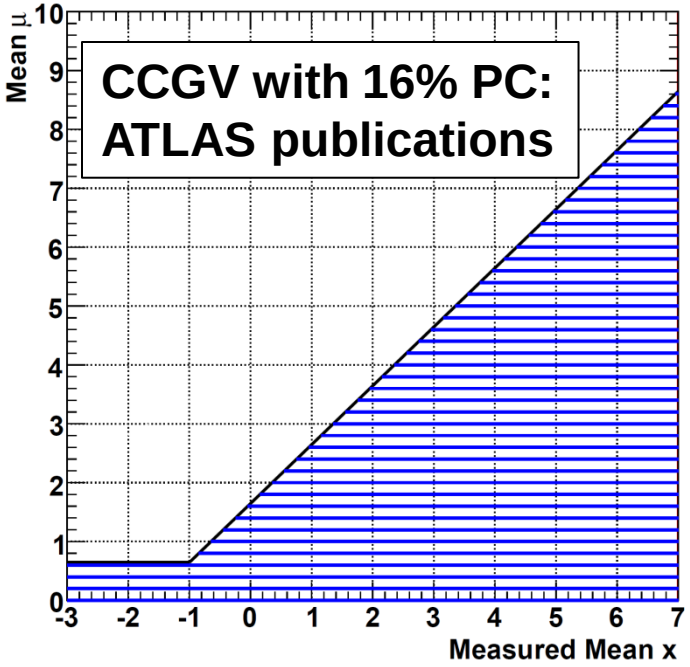
Power Constraint (PC) added by hand to avoid excluding values of μ for which N-P power below cutoff.

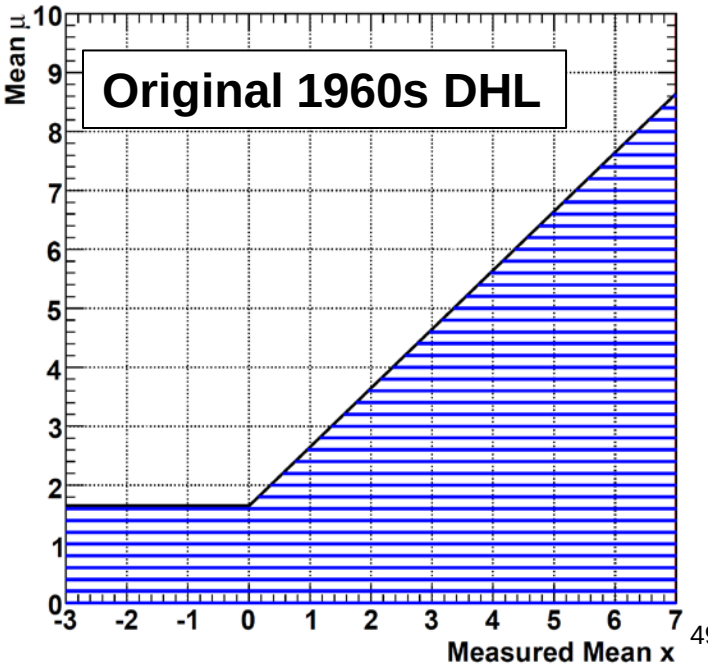
ATLAS PCL at first used PC=16%, i.e., UL = 0.64 for $x \in [-\infty, -1]$ (shown)

Compare with 1960s *original DHL*, UL = 1.64 for $x \in [-\infty, 0]$.

ATLAS revisited the value of PC, more recently used 50% PC: corresponds to original DHL (!).

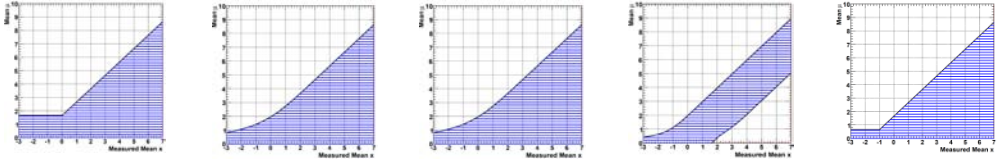
DHL is what set everyone looking for better alternatives in 1970's, 1980's, and 1990s (!)





http://arxiv.org/abs/1105.3166

Discussion

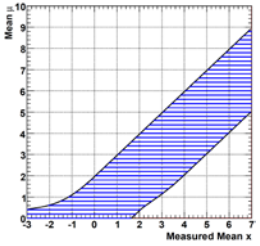


From three rather different perspectives, authors of Bayesian, CL_S, and F-C rejected original diagonal line. Recent insight into conditional coverage supports this conclusion.

Adding μ values with 100% coverage muddies situation, but I see no advantage to returning to Diagonal plus Horizontal Line methods.
DHL with 16% power constraint was a material change in HEP traditions.

Of all above methods, only F-C Unified Approach has coverage = C.L. for *all* μ; generalizes well; and ameliorates several issues.

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

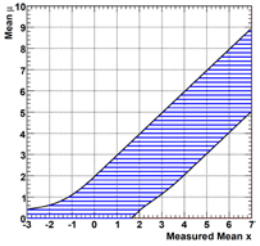


Bob: Let's consider two deep points.

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!”

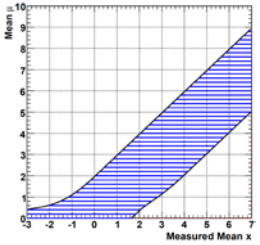


Bob:

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! "

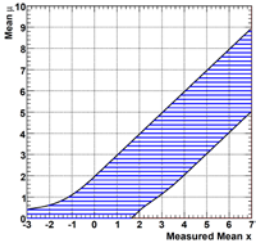


Bob: Let's consider two more points.

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! "

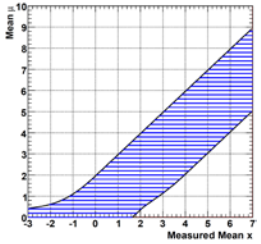


2) A very useful number to report is that value of C.L. for which μ=0 is just included in the F-C interval.

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!)

Poisson with Background Problem

Not all difficulties are exposed with the Gaussian problem, and the best-founded objection to F-C is in the Poisson problem with zero events observed.

That's another talk! I just note that the PCL version of the battle-ground plot has not been publicized by PCL advocates, as far as I know.

1996 PDG RPP, a la Helene or Zech

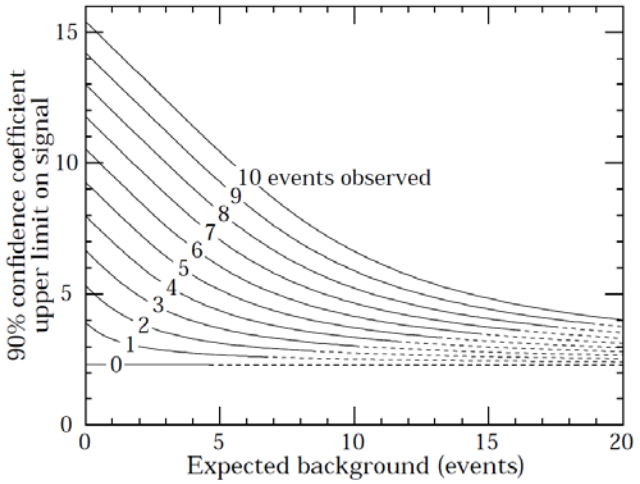


Figure 28.8: 90% confidence coefficient upper limit on the number of events as a function of the expected number of background events. For exa the expected background is 8 events and 5 events are observed, then the si 4.0 (approximately) or less with 90% confidence. Dashed portions indicate where it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

1998 PDG RPP, a la Feldman & Cousins

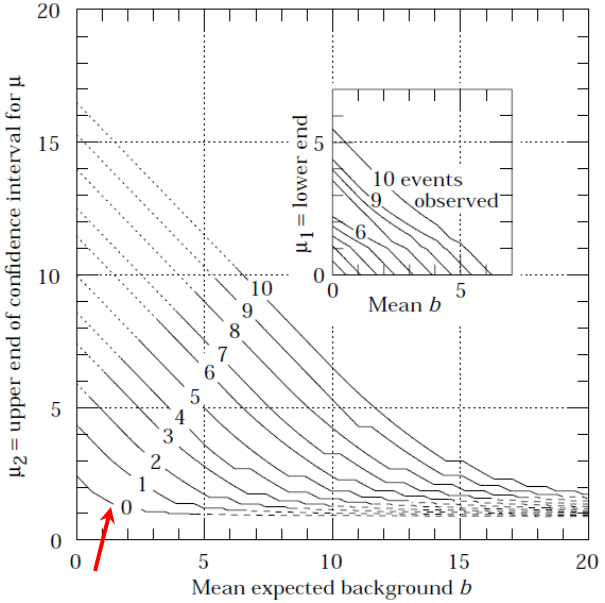
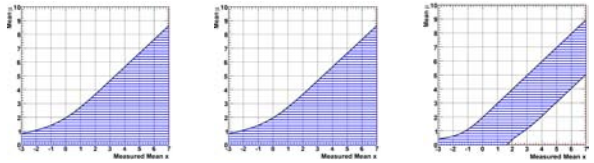


Figure 29.5: 90% confidence intervals $[\mu_1, \mu_2]$on the number of signal events as a function of the expected number of background events b. For example, if the expected background is 8 events and 5 events are observed, then the signal is 2.60 or less with 90% confidence. Dotted portions of the 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.

Conclusion

Since 2002, the PDG RPP has had a menu of choices that is quite sufficient from my point of view.



Acknowledgments

For over 25 years, I have discussed these issues with too many people to recall. Among those helping me to understand these issues some years ago were Gary Feldman, Don Groom, Virgil Highland, Fred James, and Louis Lyons. More recently, discussions with CMS and ATLAS colleagues have shed more light on the issues and stimulated this review.

References

A starting point in the literature is in the works cited by my recent arxiv post, http://arxiv.org/abs/1109.2023.

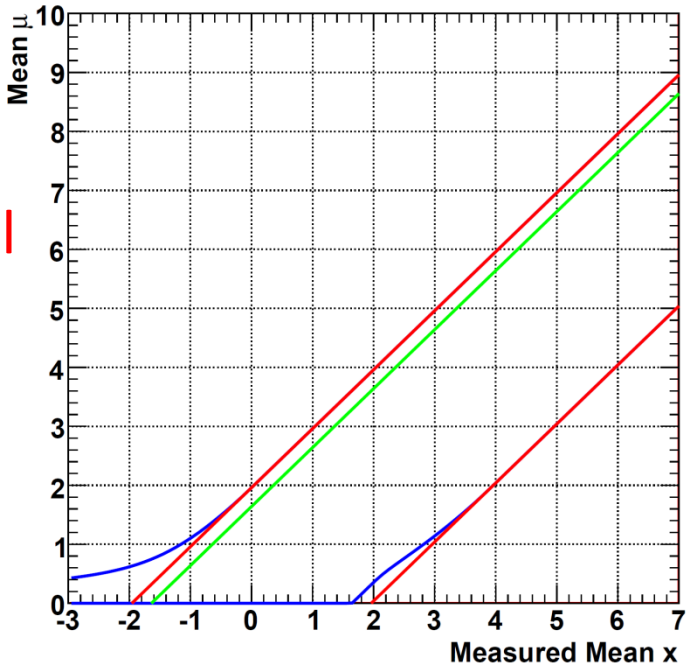
Backup

Unified and Un-Unified Intervals

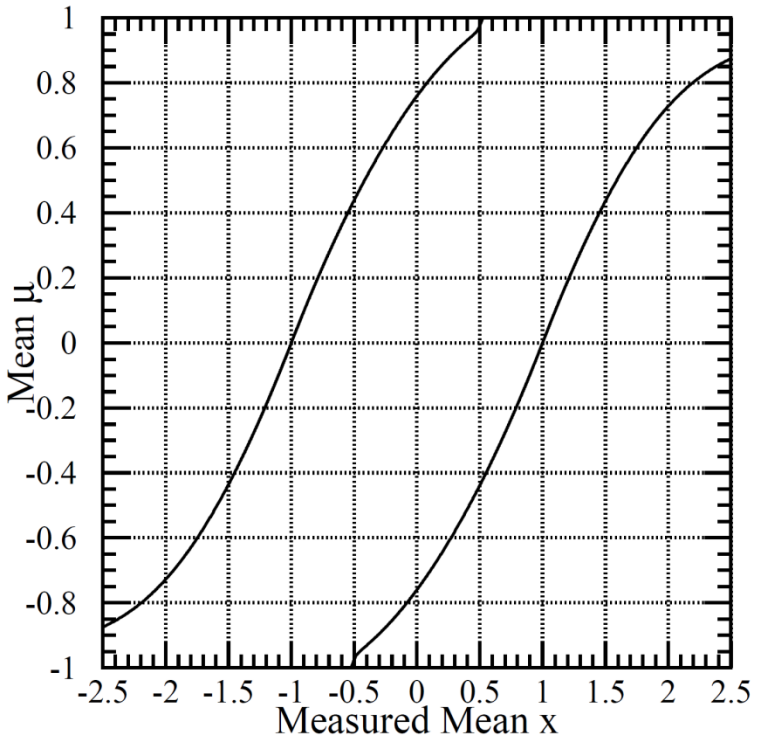
F-C

Traditional central

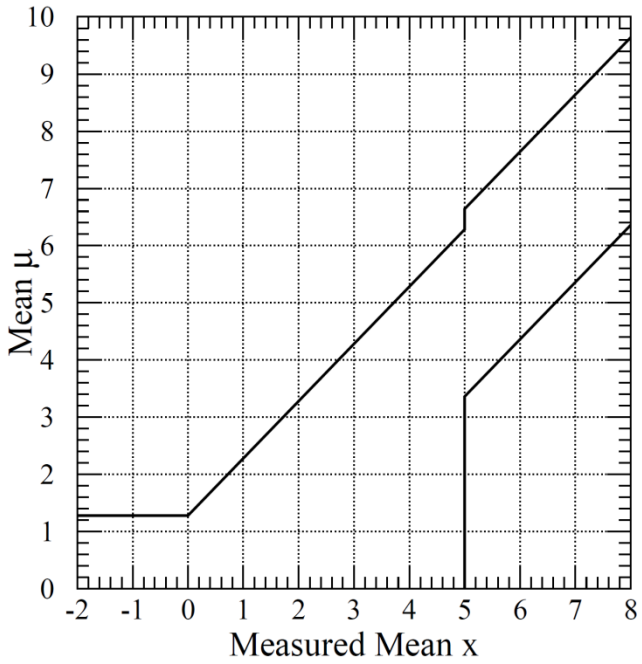
Traditional upper



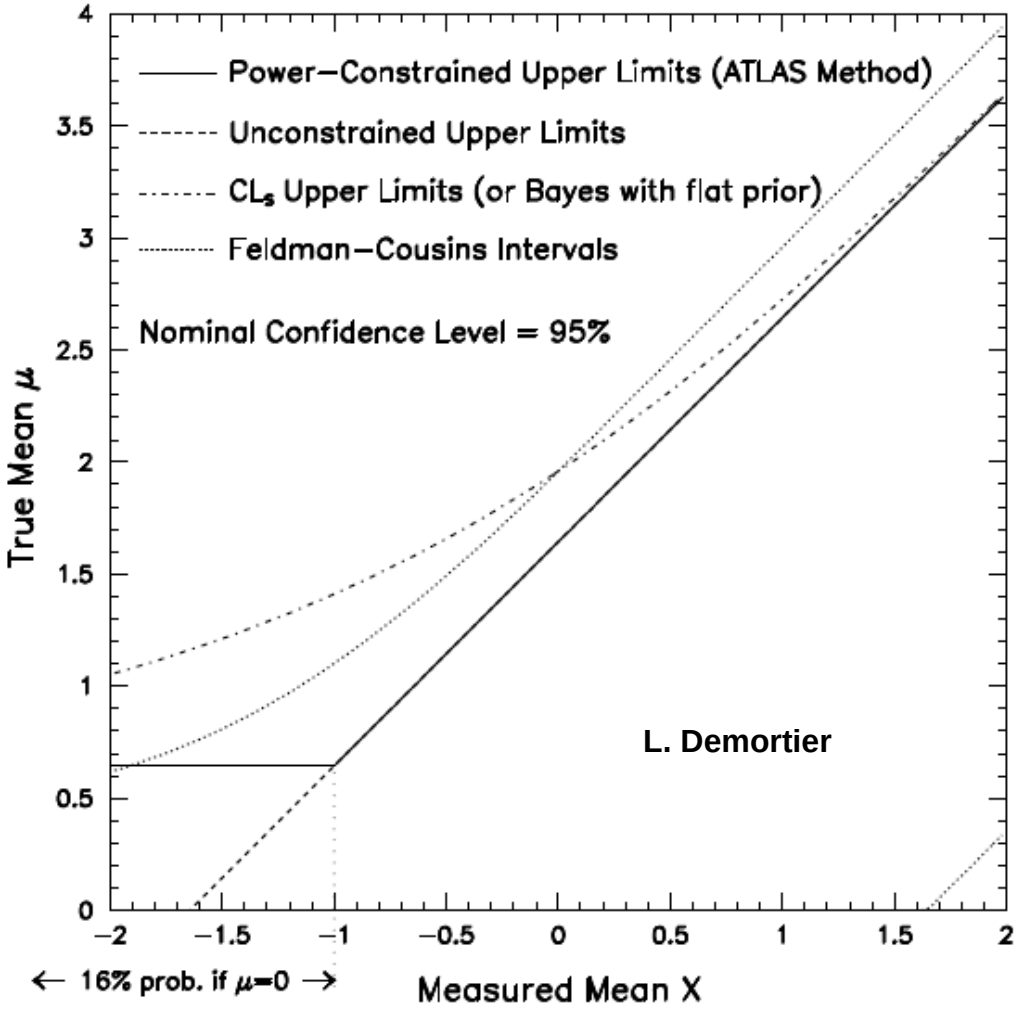
Feldman-Cousins for Two-sided Bound $-1 \leq \mu \leq 1$, $\sigma=1$



Flip-Flop Plot



Comparison of ATLAS 16% PCL with the 3 methods in PDG



(Atlas unconstrained U.L. is zero, not null, for x < -1.64)

Estimation of Upper Limits from Experimental Data

Report COO-3539-38

Virgil L. Highland

July 1986, Revised February 1987

Temple University

Philadelphia, PA 19122

Upper limit on mean of Gaussian based on one sample, x.

Physical values of mean are non-negative.

Numbers are in units of sigma (Gaussian rms).

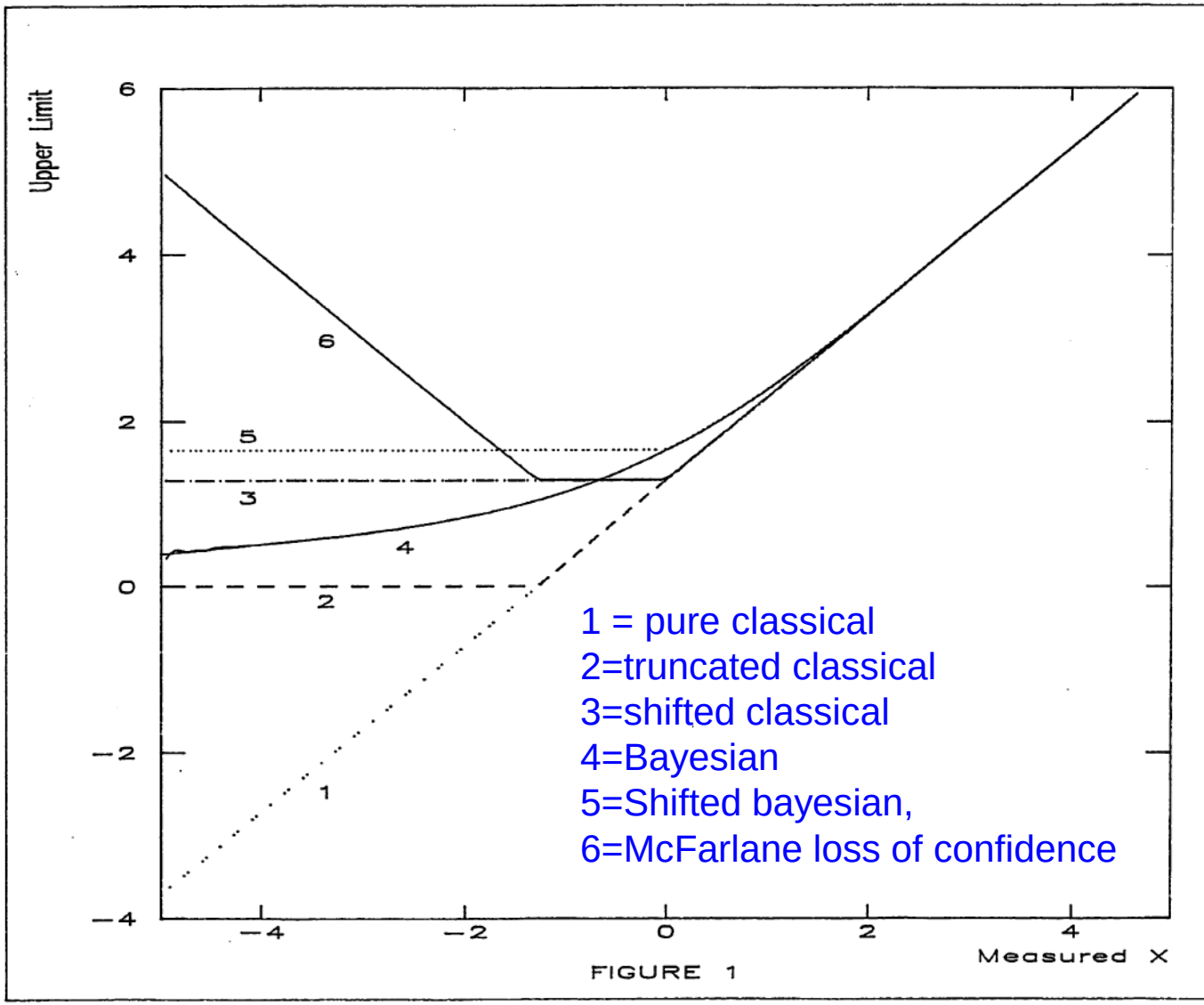


FIGURE 1

Efron Comparison of F-C and Bayesian Model Selection

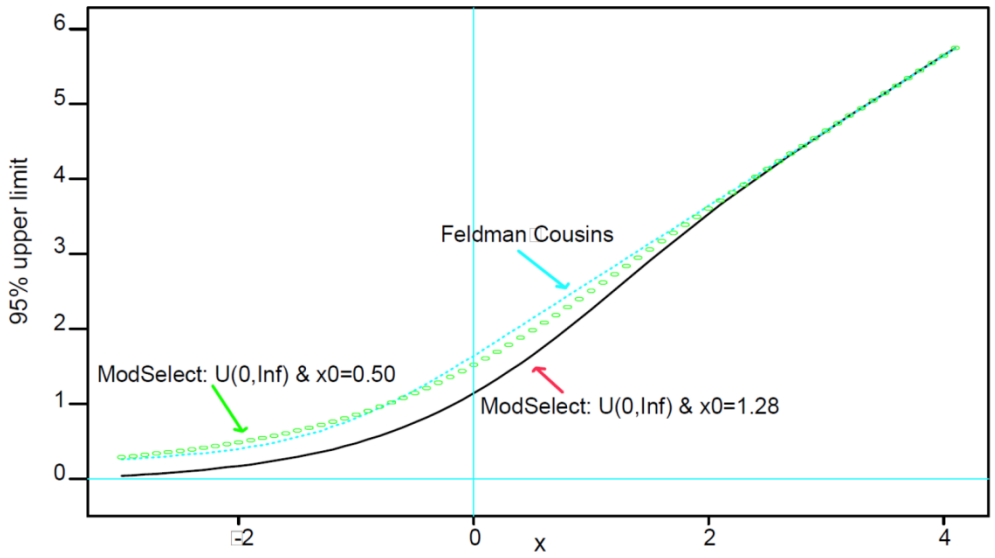


Figure 4: Moving the break-even point x_0, (17), closer to zero makes the Model Selection upper 95% point nearly match the Feldman-Cousins bound.

Features of 95% C.L. F-C intervals:

μ=0 is lower endpoint of interval until *x* is so large that *one*-tailed test excludes μ=0 at 95% (*x* > 1.64)

At x=0, interval is [0,1.96] (in contrast to μ < 1.64).

For *x*<0, interval becomes more restrictive as *x* becomes more negative.

For *x* >>1, the F-C interval converges to the 95% C.L. *central* interval, *x* $\pm$ *1.96.*

Coverage is exactly 95% for all values of μ.

Comparison of limit/interval for $x = -1$.

1) **UL = max(x,0) + 1.64 = 0 + 1.64 = 1.64**
2) **Bayesian with flat prior: UL= 1.41**
3) **CL_S: UL= 1.41**
4) **Unified Approach interval a la F-C: [0, 1.10].**
   $\mu_{best} = 0$; $\chi^2(\mu_{best}) = 1$. Interval includes μ_t for which $\Delta\chi^2 <$ critical $\Delta\chi^2$ for that μ_t .
5) **16% PCL with $x_{PCL} = -1$: UL = max(0, max(x, –1) + 1.64) = max(0, –1 + 1.64) = 0.64.**
   Note that $\chi^2(\mu = 0.64) = (-1 - 0.64)^2 = 2.70$. Interval includes μ_t for which χ^2 (*not* $\Delta\chi^2$!) is less than the "book value" $\Delta\chi^2 = 2.70$ for one-sided limit! In effect, this brings in goodness-of-fit to the model. (Sec. 4C of F-C paper).