Part 2 of Introduction to “Statistics in Theory”: Prelude to Statistics in Practice

Based on updated versions of some of the slides originally prepared for 2009 HCPSS,
http://indico.cern.ch/event/44587/contributions/1108838/attachments/943235/1337911/cousins_stats_hcpss_2009.pdf

Bob Cousins

Univ. of California, Los Angeles

February 26, 2017

What can be computed without using a prior, with only the frequentist definition of P?

Not P(constant of nature is in some *specific* interval | data)

Not P(SUSY is true | data) ; *Not* P(SM is false | data)

1) *Confidence Intervals* for constants of nature, parameter values, as defined in the 1930's by Jerzy Neyman.

   Statements are made about probability properties of ensembles of intervals (what fraction contains unknown true value)

2) Likelihood *ratios*, the basis for a large set of techniques for point estimation, interval estimation, and hypothesis testing.

Both can be constructed using the frequentist definition of P.

Notation: x is observable, μ is parameter;
p(x|μ) is pdf characterizing the experiment apparatus, called "the statistical model", or simply "the model", by statisticians.

Confidence Intervals

"Confidence intervals", and this phrase to describe them, were invented by Jerzy Neyman in 1934-37. Statisticians mean Neyman's intervals (or an approximation) when they say "confidence interval". In HEP the language is a little loose.

I highly recommend using "confidence interval" (and "confidence regions" when multi-D) only to describe intervals and regions corresponding to Neyman's construction (or good approximations thereof), described below.

Confidence Intervals

Next many slides:

1. **Introduce basic notions, illustrated by upper/lower limits and closely related *central* confidence intervals**
2. **Discuss Neyman's more general construction (used e.g. by Feldman and Cousins).**
3. **Make connection to hypothesis testing of particular value of parameter vs other values.**

Basic notions of confidence intervals

Given the model $p(x|\mu)$ and the observed value x_0, for what values of μ is x_0 an "extreme" value of x? Include in the confidence interval $[\mu_1,\mu_2]$ those values of μ for which x_0 is *not* "extreme".

In order to define "extreme", one needs to choose an *ordering principle* for x applicable to each μ: *high rank means not extreme*.

Basic notions of confidence intervals (cont.)

Some common ordering choices in 1D (when $p(x|\mu)$ is such that higher μ implies higher average x):

1. **Order x from largest to smallest.**
   So smallest values of x are most extreme.
   Given x_0, the confidence interval containing μ for which x_0 is not extreme will typically not contain largest values of μ.
   Leads to confidence intervals known as *upper limits* on μ.
2. **Order x from smallest to largest. Leads to *lower limits* on μ.**
3. **Order x using smallest central quantile of $p(x|\mu)$ containing x_0.**
   Leads to *central* confidence intervals for μ.

N.B. These three apply only when x is 1D.

(4th ordering, LR ratio used by F-C, still to come.)

Basic notions of confidence intervals (cont.)

Given model $p(x|\mu)$ and ordering of x, one chooses a fraction of highest-ranked values of x that are *not* considered as "extreme".

This fraction is called the *confidence level* (C.L.), say 68% or 95%.

We also define α = 1 – C.L., the lower-ranked fraction, "extreme".

The *confidence interval* $[\mu_1,\mu_2]$ contains those values of μ for which x_0 is *not* "extreme" at the chosen C.L. (given the ordering).

E.g., at 68% C.L., $[\mu_1,\mu_2]$ contains those μ for which x_0 is in the highest-ranked (least extreme) 68% values of x.*

In this talk, 68% is more precisely 68.27; 84% is 84.13%; etc.

Basic notions of confidence intervals (cont.)

The endpoints of *central* confidence intervals at C.L. are the same as upper/lower limits with 1 – (1 – C.L.)/2. E.g.:

84% C.L. *upper* limit μ_2 excludes μ for which x_0 is in the lowest 16% values of x.

84% C.L. *lower* limit μ_1 excludes μ for which x_0 is in the highest 16% values of x.

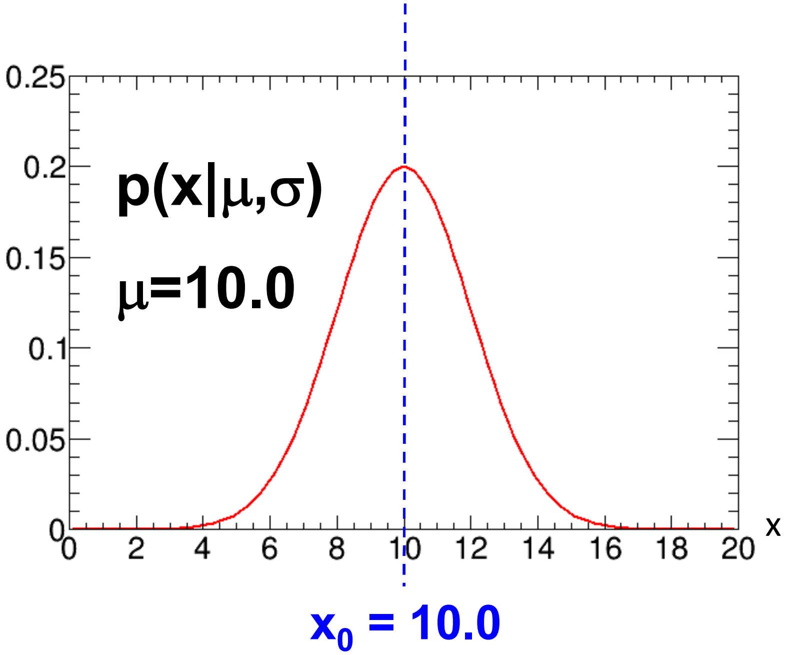
Then $[\mu_1,\mu_2]$ includes the central 68% quantile of x values ordered from high to low; it is a 68% C.L. *central* confidence interval (!)

Gaussian pdf p(x|μ,σ) with σ a function of μ: σ = 0.2 μ

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}$$

$\sigma(\mu) = (0.2)\ \mu$

p(x|μ,σ) with μ=10.0, σ = 0.2 :



Suppose x_0 = 10.0 is observed.

What can one say about μ ?

μ

Gaussian pdf p(x|μ,σ) with σ a function of μ: σ = 0.2 μ

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}$$

$\sigma(\mu) = (0.2)\,\mu$

p(x|μ,σ) with μ=10.0, σ = 0.2 :

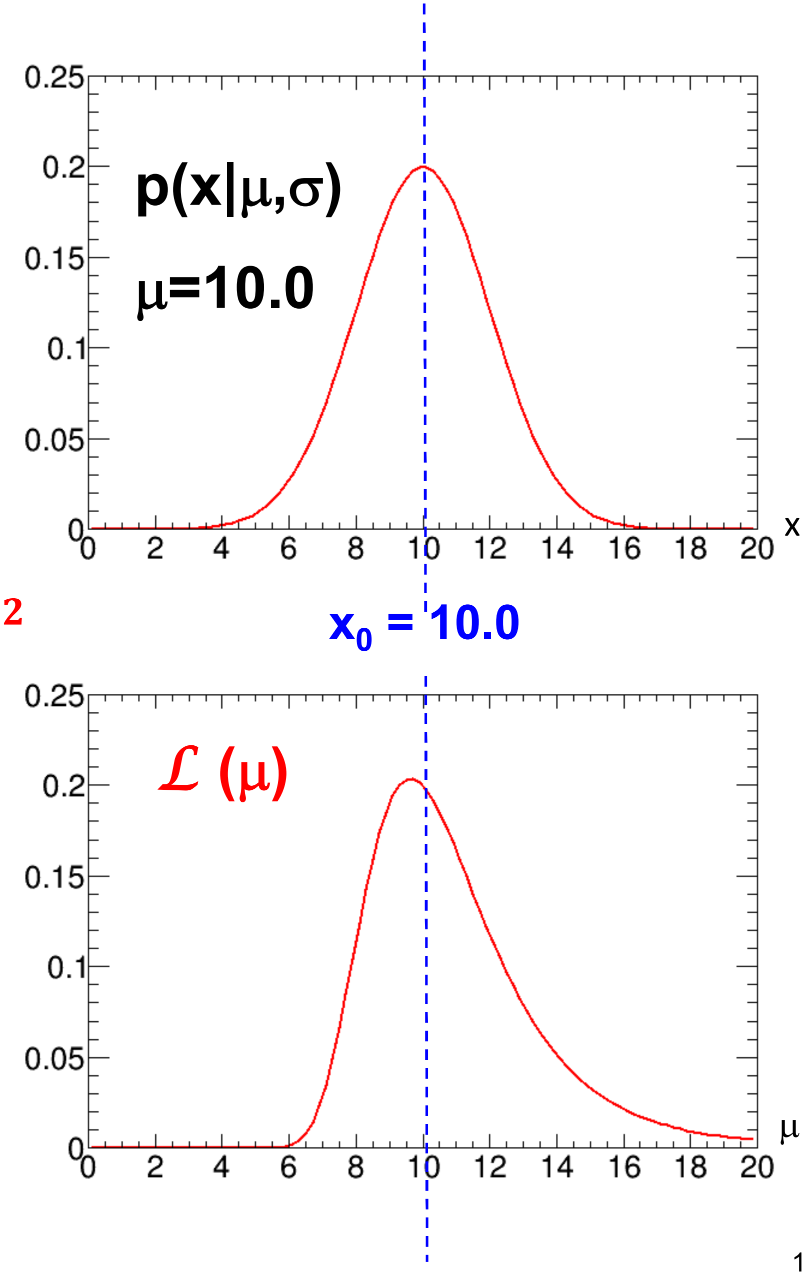
Suppose x_0 = 10.0 is observed.

$$\mathcal{L}\,(\mu) = \frac{1}{\sqrt{2\pi(0.2\mu)^2}}\, e^{-(x-\mu)^2/2(0.2\mu)^2}$$

$\mathcal{L}$ (μ) for observed x_0 = 10. :

μ_{ML}= 9.63

What is confidence interval for μ?



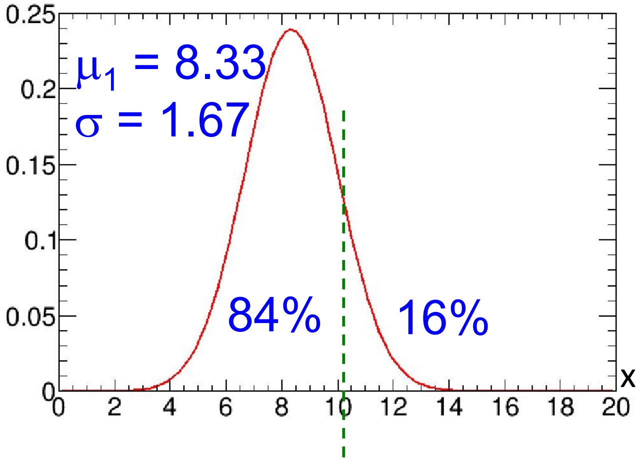
Gaussian pdf $p(x|\mu,\sigma)$ with σ a function of μ: $\sigma = 0.2\,\mu$ Observed $x_0 = 10.0$.

Find μ_1 such that 84% of $p(x|\mu_1,\sigma=0.2\mu_1)$ is below $x_0 = 10.0$; 16% of prob is above.

Solve: $\mu_1 = 8.33$.

$[\mu_1,\infty]$ is 84% C.L. confidence interval

μ_1 is 84% C.L. *lower* limit for μ.



Gaussian pdf $p(x|\mu,\sigma)$ with σ a function of μ: $\sigma = 0.2\,\mu$
Observed $x_0 = 10.0$.

Find μ_1 such that 84% of $p(x|\mu_1,\sigma{=}0.2\mu_1)$ is below $x_0 = 10.0$; 16% of prob is above.

Solve: $\mu_1 = 8.33$.

$[\mu_1,\infty]$ is 84% C.L. confidence interval

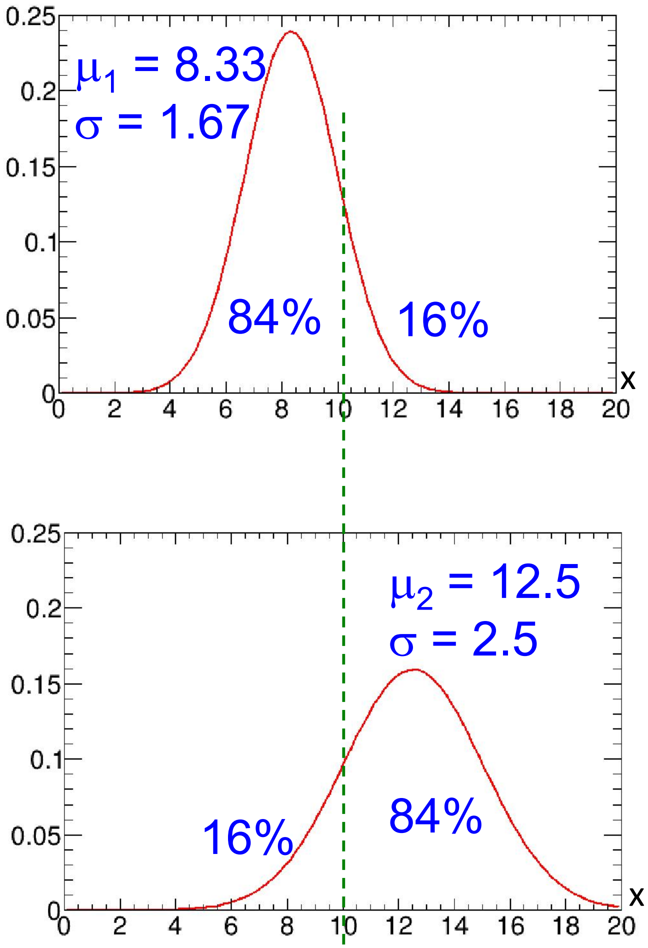
μ_1 is 84% C.L. *lower* limit for μ.

Find μ_2 such that 84% of $p(x|\mu_2,\sigma{=}0.2\mu_2)$ is above $x_0 = 10.0$; 16% of prob is below.

Solve: $\mu_2 = 12.5$.

$[-\infty,\mu_2]$ is 84% C.L. confidence interval

μ_2 is 84% C.L. *upper* limit for μ.



Then 68% C.L. *central* confidence interval is $[\mu_1,\mu_2] = [8.33,12.5]$.

Gaussian pdf $p(x|\mu,\sigma)$ with σ a function of μ: $\sigma = 0.2\ \mu$ Observed x_0 = 10.0.

So the 68% C.L. *central* confidence interval is [8.33,12.52].

This is "exact". Follows reasoning of E.B. Wilson, JASA 1927!

Note difference from reasoning that proceeds as:

1) For x_0 = 10.0, minimum-χ^2 point estimate of μ is $\hat{\mu}$ = 10.0.
2) Then estimate $\hat{\sigma} = 0.2 \times \hat{\mu}$ = 2.0.
3) Then $\hat{\mu} \pm \hat{\sigma}$ yields interval [8.0,12.0].

For ("exact") confidence intervals, the reasoning must always involve probabilities for x, calculated *considering particular possible true values of parameters*, as on previous slide!

Clearly the validity of above approximate reasoning depends on how much $\sigma(\mu)$ changes for μ relevant to problem at hand. Beware!

Confidence intervals for binomial parameter ρ
Directly relevant to efficiency calculation in HEP

Let $\mathrm{Bi}(n_{on} \mid n_{tot}, \rho)$ denote binomial probability of n_{on} successes in n_{tot} trials, each with binomial parameter ρ:

$$\mathrm{Bi}(n_{on} \mid n_{tot}, \rho) = \frac{n_{tot}!}{n_{on}!\,(n_{tot}-n_{on})!}\ \rho^{n_{on}}\ (1-\rho)^{(n_{tot}-n_{on})}$$

In repeated trials, n_{on} has mean $n_{tot}\,\rho$ and

rms deviation $\sqrt{n_{tot}\,\rho\,(1-\rho)}$

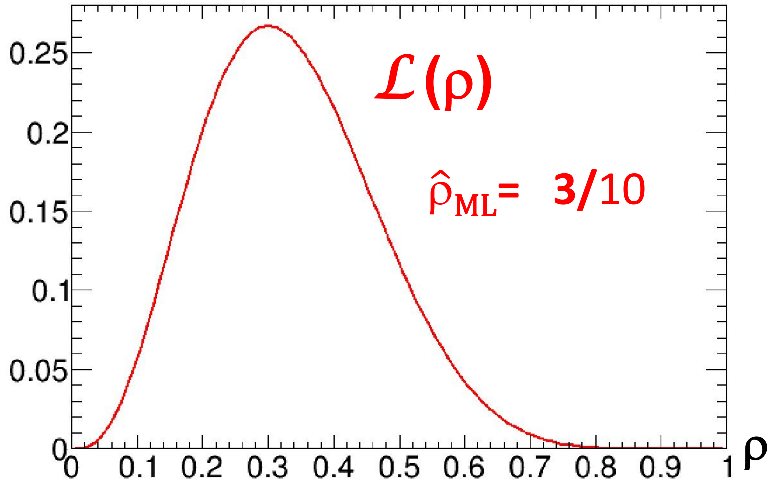
With observed successes n_{on}, the M.L. point estimate $\hat{\rho}$ of ρ is

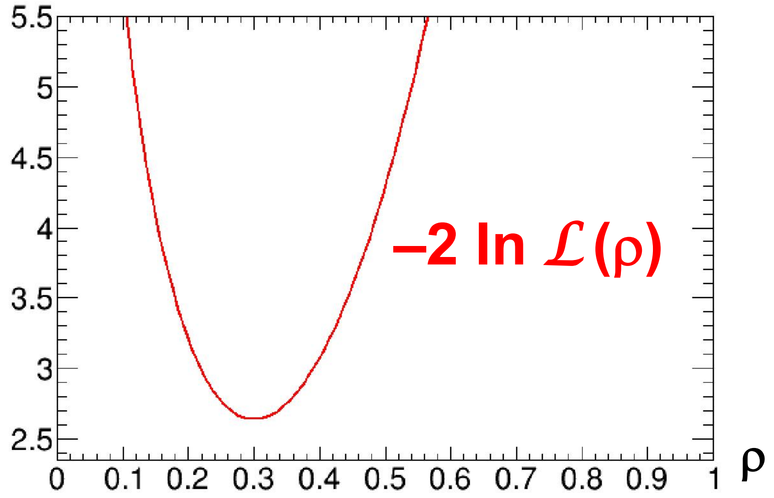
$$\hat{\rho} = n_{on} / n_{tot}\ .$$

What confidence interval $[\rho_1,\rho_2]$ should we report for ρ?

Confidence intervals for binomial ρ (cont.)

Suppose n_{on}=3 successes in n_{tot}=10 trials.





Let's find exact 68% C.L.* *central* confidence interval [ρ_1,ρ_2].

Recall shortcut above for central intervals:

Find lower limit ρ_1 with C.L. = 1 – (1 – 68%)/2. = 84%

I.e., Find ρ_1 such that Bi(n_{on} < 3 | n_{tot}=10, ρ_1) = 84%

Find upper limit ρ_2 with C.L. = 84%

I.e., Find ρ_2 such that Bi(n_{on} > 3 | n_{tot}=10, ρ_2) = 84%

****Recall in this talk, 68% is more precisely 68.27; 84% is 84.13%; etc.***

n_{on} = 3 , n_{tot}=10.

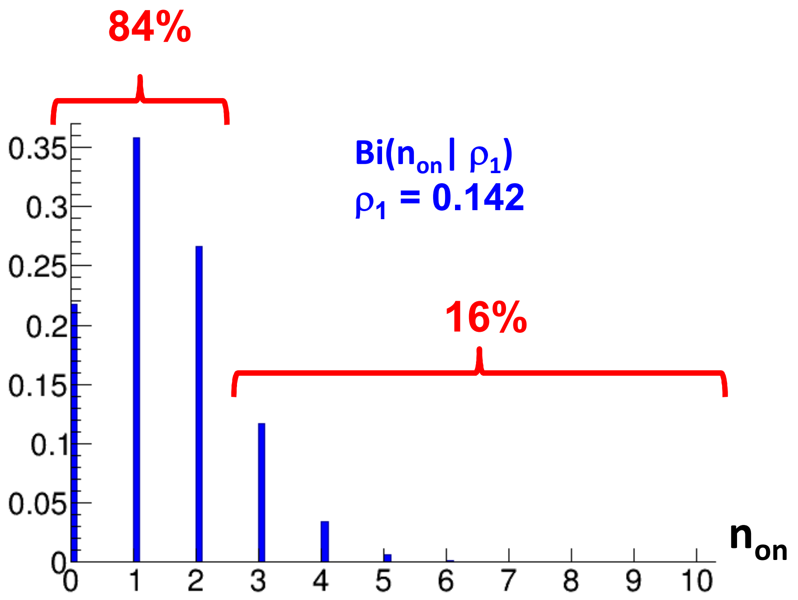
Find ρ_1 such that

Bi(n_{on} < 3 | ρ_1) = 84%

Bi($n_{on} \geq$ 3 | ρ_1) = 16%

(lower limit at 84% C.L.)

Solve: ρ_1 = 0.142

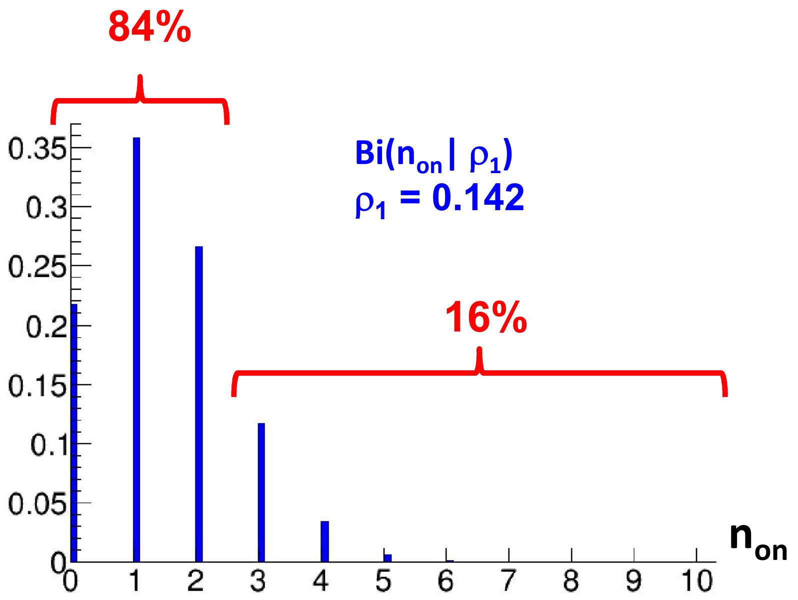


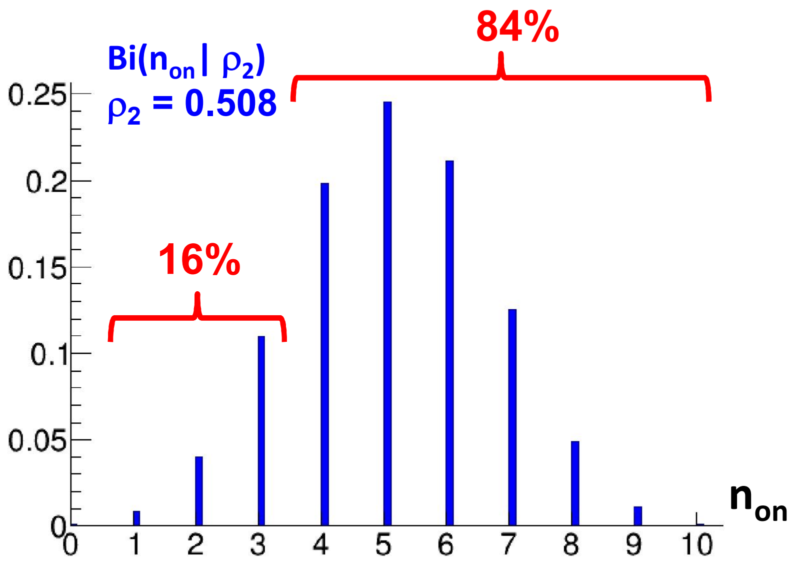
84%

$n_{on} = 3$, n_{tot}=10.
Find ρ_1 such that
$\text{Bi}(n_{on} < 3 \mid \rho_1) = 84\%$
$\text{Bi}(n_{on} \geq 3 \mid \rho_1) = 16\%$
(lower limit at 84% C.L.)
Solve: $\rho_1 = 0.142$

And find ρ_2 such that
$\text{Bi}(n_{on} > 3 \mid \rho_2) = 84\%$
$\text{Bi}(n_{on} \leq 3 \mid \rho_2) = 16\%$
(upper limit at 84% C.L.)
Solve: $\rho_2 = 0.508$

Then $[\rho_1,\rho_2] = (0.142, 0.508)$
is *central* confidence interval
with 68% C.L. Same as
Clopper and Pearson (1934)





Poisson example: Fig. 3a,b; R. Cousins, Am. J. Phys. 63 398 (1995) DOI: 10.1119/1.17901

Gaussian approximation for binomial conf. int.

As above, n_{on} has mean $n_{tot}\,\rho$ and rms deviation $\sqrt{n_{tot}\,\rho\,(1-\rho)}$.
So approximate binomial by Gaussian with mean and rms

$$\mu(\rho) = n_{tot}\,\rho$$
$$\sigma(\rho) = \sqrt{n_{tot}\,\rho\,(1-\rho)}$$

Idea is *not* to substitute $\hat{\rho}$ for ρ (big mistake), but rather follow E.B. Wilson (1927), use above recipe for upper and lower limits:

1) Find ρ_1 such that Gauss($x \geq 3$ | mean ρ_1, $\sigma(\rho_1)$) = 0.16
2) Find ρ_2 such that Gauss($x \leq 3$ | mean ρ_2, $\sigma(\rho_2)$) = 0.16

This consistently uses the σ associated with each ρ. Leads to a quadratic equation with solution $[\rho_1,\rho_2] = [0.18, 0.46]$ which is the approximate 68% C.L. confidence interval known as the *Wilson score interval*.

Avoid the Wald interval – no reason to use it

The "Wilson score interval" needs only the quadratic formula but is for some reason relatively unknown. It is tempting instead to substitute $\hat{\rho} = n_{on}/n_{tot}$ for ρ in the expression for σ:

$\hat{\sigma} = \sqrt{n_{tot}\,\hat{\rho}\,(1 - \hat{\rho})}$, obtaining the potentially disastrous "Wald interval": $[\rho_1,\rho_2] = \hat{\rho} \pm \hat{\sigma}$.

The Wald interval does not use the correct logic for frequentist confidence! In fact $\hat{\sigma} = 0$ when $n_{on} = 0$ (or $n_{on} = n_{tot}$). Incredibly, failure of the Wald interval when $n_{on} = 0$ (or $n_{on} = n_{tot}$) has been used as a *foundational argument* in favor of Bayesian intervals in at least four public HEP postings (one retracted) and one published astro paper! (Typically the authors did not understand Bayesian statistics either, and used flat prior...)

Clopper-Pearson is the standard in HEP

In HEP, Clopper-Pearson intervals are the traditional standard: in Particle Data Group's Review of Particle Physics since 2002.

Many tables and online calculators for C-P exist, e.g., http://statpages.org/confint.html .

But C-P is criticized by some as "wastefully conservative" – see our paper below.

For a comprehensive review of both central and non-central confidence intervals for a binomial parameter and for the ratio of Poisson means, see Cousins, Hyme, and Tucker, http://arxiv.org/abs/0905.3831 . Many are implemented in https://root.cern.ch/doc/master/classTEfficiency.html .

For related construction of upper/lower limits and central interval for Poisson mean, see R. Cousins, Am. J. Phys. 63 398 (1995)

HEP applications of conf. intervals for binomial param

1. **As mentioned, directly relevant to efficiency calculations.**
2. **Using a famous math identity, directly applicable to confidence intervals for *ratio of Poisson means.***
3. **Then, applicable to significance (Z_{Bi}) of excess in a signal bin when sideband is used to estimate background. Cousins, Linnemann, and Tucker, http://arxiv.org/abs/physics/0702156 .**
4. **Can even stretch #3 (using "rough correspondence") to problem of signal bin when Gaussian estimate of mean bkgnd exists.**

Issues for upper-lower limits and central conf. ints.

For decades, problems with upper limits and central confidence intervals. Prototype problems:

1. Gaussian measurement resolution near a physical boundary (e.g. neutrino mass-squared is positive)
2. Poisson signal mean measurement when observed number of events is less than mean expected background (so naïve “background-subtracted” cross section is negative)

Many ideas put forward, PDG settled on three. Some history:
http://www.physics.ucla.edu/~cousins/stats/cousins_bounded_gaussian_virtual_talk_12sep2011.pdf

Today in Part 2, I stick to frequentist confidence intervals.

Beyond upper/lower limits and *central* confidence intervals

More general ordering choices for ordering x in $p(x|\mu)$:

- **Order x_0 using the likelihood ratio $\mathcal{L}(x_0|\mu) \,/\, \mathcal{L}(x_0|\mu_{\text{best fit}})$. Advocated in HEP by Feldman and Cousins in 1998 (and in Kendall and Stuart long before and since). Applicable in both 1D and multi-D for x.**

N.B. Ordering x by the probability *density* $p(x|\mu)$ is dependent on metric of x, and hence *not* recommended! Jacobian of transformation to y(x) alters ordering.

(Recall from Part 1 that likelihood *ratios* as in F-C are independent of metric in x since Jacobian cancels.)

Neyman's Construction of Confidence Intervals

The general method for constructing "Confidence intervals", and the name, were invented by Jerzy Neyman in 1934-37.

The next few slides give basic outline.

It takes a bit of time to sink in – given how often confidence intervals are misinterpreted, the argument is perhaps a bit too ingenious.

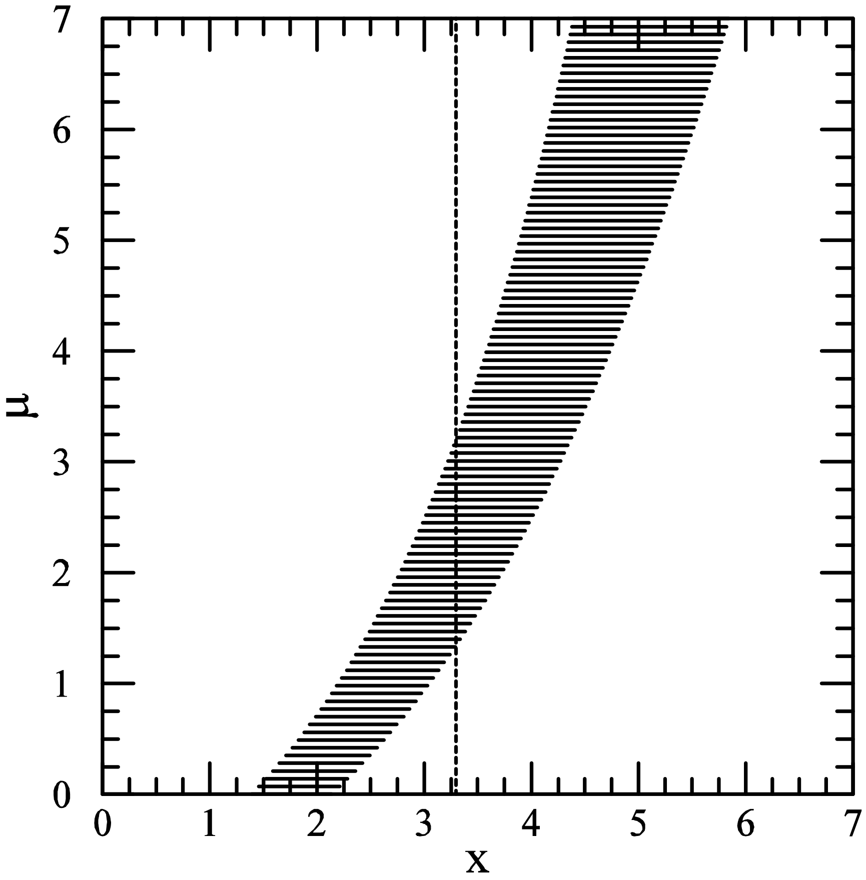
In particular, you should understand that the confidence level does *not* tell you "how confident you are that the unknown true value is in the interval" – only a *subjective* Bayesian credible interval has that property!

Neyman's Construction of Confidence Intervals

Given p(x|μ) from a model: For each value of μ , one draws a horizontal *acceptance interval* $[x_1,x_2]$ such that $p(x \in [x_1,x_2] \mid \mu) =$ C.L. = 1 - α. ("Ordering principle" is used to well-define.)

Upon observing x, obtaining the value x_0, one draws the vertical line through x_0.

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

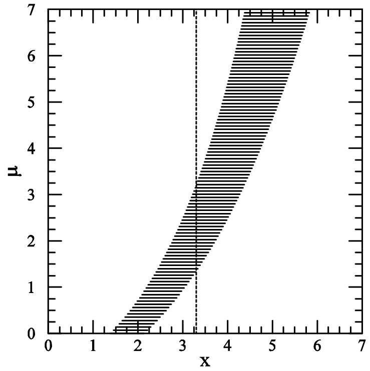


Note: x and μ need not have the same range, units, or (in generalization to higher dimensions) dimensionaliity!

Figure from G. Feldman, R Cousins, Phys Rev D57 3873 (1998)

Important note regarding *x* and μ

Note : x and μ need not have the same range, units, or (in generalization to higher dimensions) dimensionaliity!



I think it is *much* easier to avoid confusion when x and μ are qualitatively different.

Louis Lyons gives the example where x is the flux of solar neutrinos and μ is the temperature at the center of the sun.

I like examples where x and μ have different dimensions: Neyman's original paper has 2D observation space and 1D parameter space – see backup.

Famous confusion re Gaussian p(x|μ) where μ is mass ≥ 0

It is *crucial* to distinguish between the data *x*, which *can* be negative (no problem), and the mass parameter μ, for which negative values *do not exist in the model*.

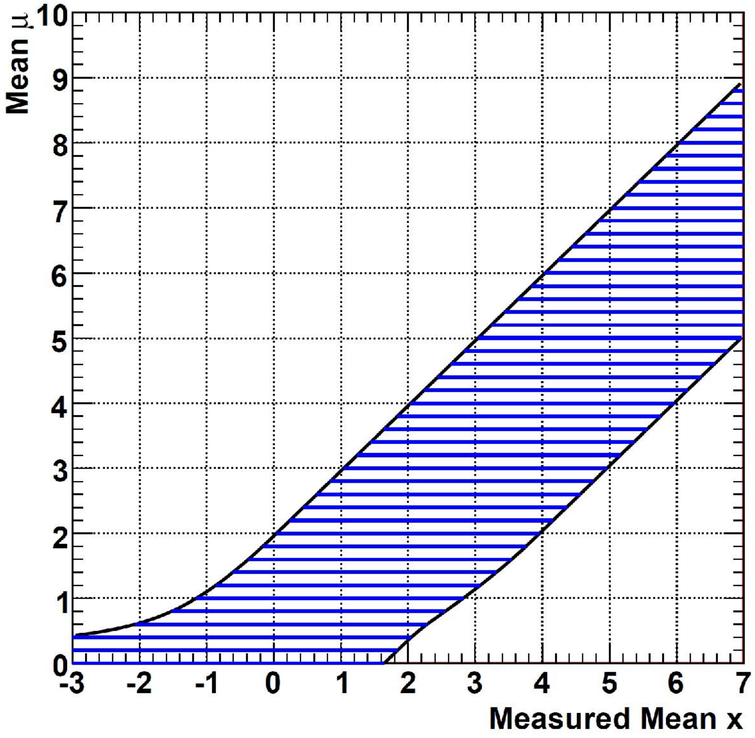
I.e., for mass μ <0, p(*x*|μ) does not exist! You would not know how to simulate the physics of detector response for *mass* < 0.

Constraint μ ≥ 0 has *nothing* to do with a Bayesian prior for μ !!!

It's in the *model* (and hence in $\mathcal{L}(\mu)$).

The confusion is encouraged since we often refer to x as the "measured value of μ", and say that x<0 is "unphysical" – bad habits!

A proper Neyman construction graph has x of both signs but only non-negative μ ≥ 0. Example: Construction on right is LR ordering advocated by Feldman-Cousins



Famous 1934 Construction of Clopper and Pearson: Central Confidence Intervals for a Binomial Parameter

Biometrika, Vol. 26, No. 4. (Dec., 1934), pp. 404-413

x = number of successes (here, integer 0-10 out of 10 trials)

Inner corners of the steps give the intervals; traditional to draw the curved "belts" connecting them, but only evaluated at the integers. Tricky to draw, read!

Discreteness of x typically requires horizontal acceptance intervals to contain more than 95% probability, so there is *over-coverage* in the vertical confidence intervals.

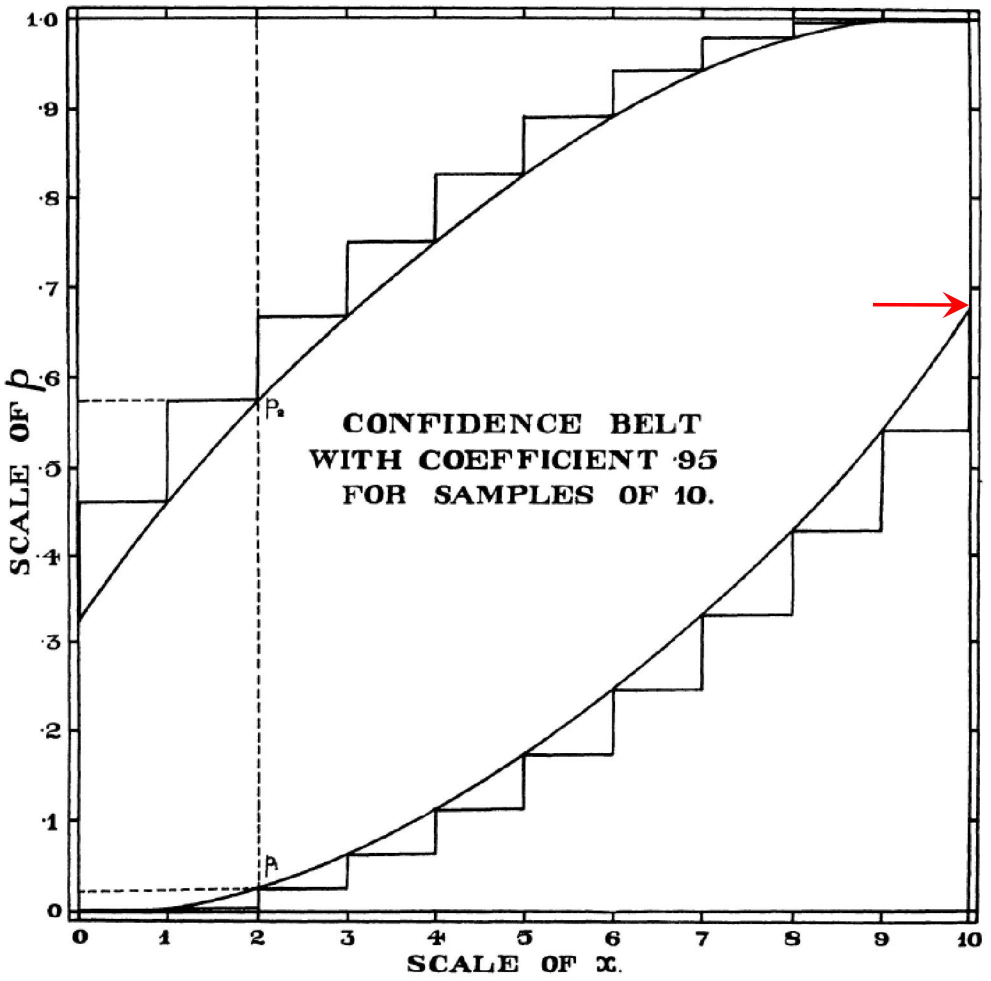


FIG. 1

E.g. 95% C.L. central interval for p if 10/10 successes/trials: (0.69,1.0)

Partial details of construction:

Blue lines are two of the acceptance intervals having central 95% or more prob, at continuous ρ.

Note data x is discrete, so graph is only read at discrete x.

If you stare at it long enough, you will see connection between upper/lower limits and central intervals, for discrete data.

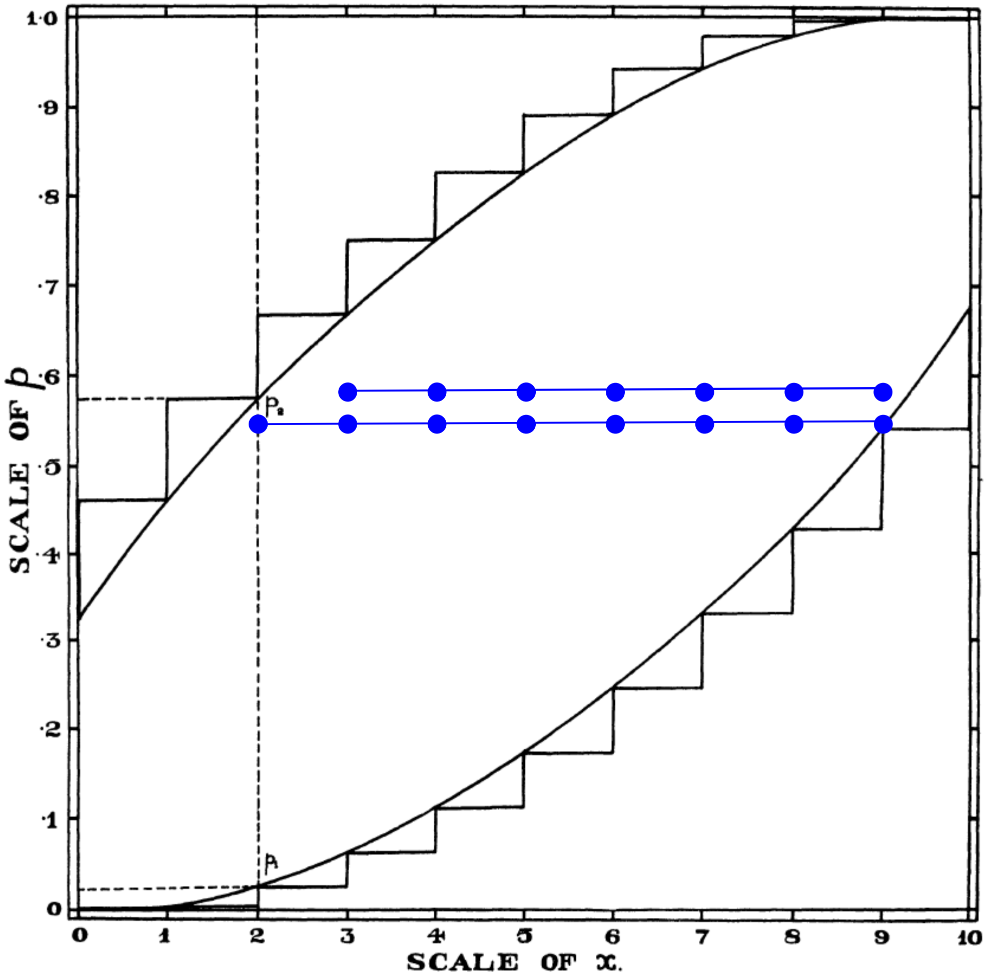


FIG. 1

Confidence Intervals and Coverage

Recall: In math, one defines a *vector space* as a set with certain properties, and then the definition of a *vector* is "an element of a vector space".
(A vector is not defined in isolation.)

Similarly, whether constructed in practice by Neyman's construction or some other technique, a *confidence interval* is defined to be "a element of a confidence set", where the *confidence set* is a set of intervals defined to have the property of frequentist *coverage* under repeated sampling:

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. = $1 - \alpha$ 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. = $1 - \alpha$. (Definition of coverage)

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.

Confidence Intervals and Coverage (cont.)

$P(\mu_t \in [\mu_1, \mu_2]) = C.L. = 1 - \alpha$. (Definition of coverage)

One of the complaints about confidence intervals is that the consumer often forgets (if he or she ever knew) that
the random variables in this equation are μ_1 and μ_2, and not μ_t ,
and that *coverage is a property of the set*, not of an individual interval!
Please don’t forget!

It *is* true (in precisely the sense defined by the ordering principle used in the Neyman construction) that the confidence interval consists of those values of μ for which the observed x is among the least extreme values to be observed.

A lot of confusion might have been avoided if Neyman had chosen the name “*coverage intervals*”!

Classical Hypothesis Testing

In Neyman-Pearson hypothesis testing (James06), frame discussion in terms of null hypothesis H_0 (e.g. S.M.), and an alternative H_1 (e.g., some BSM model).

For the null hypothesis, order possible observations x from least extreme to most extreme, using an ordering principle (which can depend on H_1 as well). Choose a cutoff α (smallish number).

Then "reject" H_0 if observed x_0 is in the most extreme fraction α of observations x (generated under H_0). Then

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)

μ : parameters in the hypotheses (statisticians like name θ)

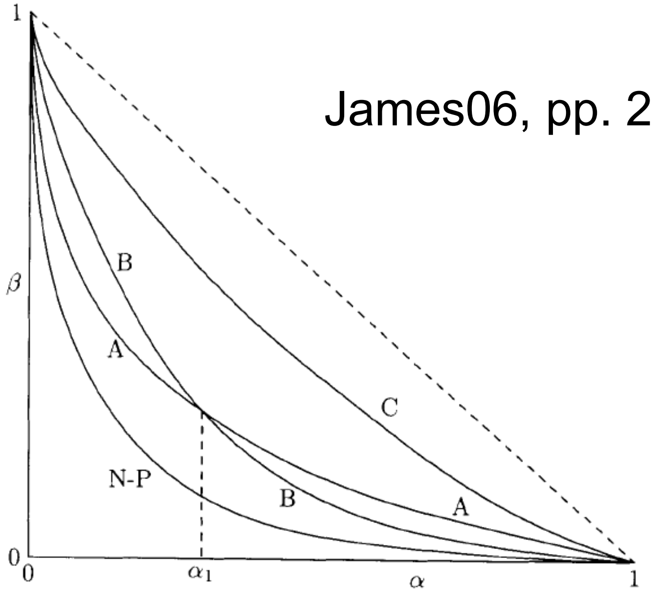
Classical Hypothesis Testing (cont.)

Common for H_0 to be *nested* in H_1 to, i.e. H_0 corresponds to particular parameter μ value μ_0 (e.g., zero, 1, or ∞) in H_1.

Competing analysis methods can be compared by looking at graphs of β vs α at various μ, and at graphs of β vs μ at various α (power function).

Similar to comparing b-tagging efficiency for signal and background, at different p_T. Equivalent to ROC curve.

Classical Hypothesis Testing (cont.)



James06, pp. 258, 262

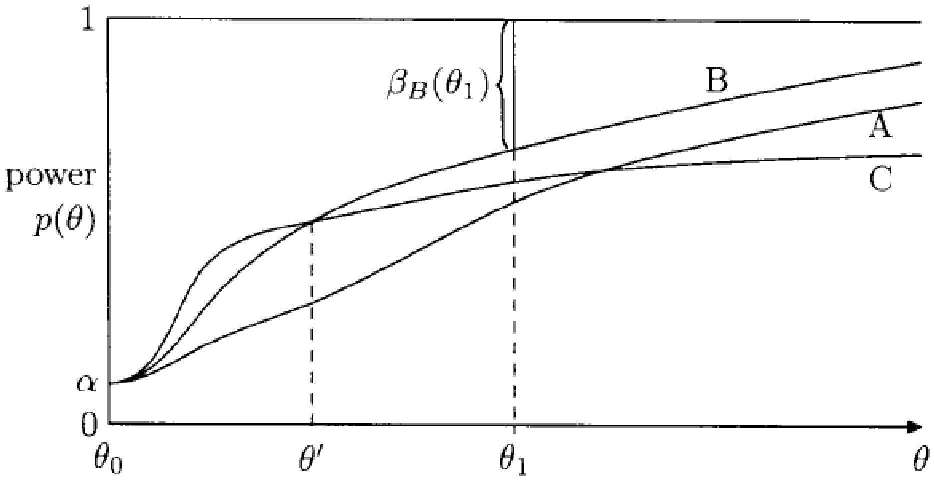


Fig. 10.3. Power functions of tests A, B, and C at significance level α. Of these three tests, B is the best for $\theta > \theta'$. For smaller values of θ, C is better.

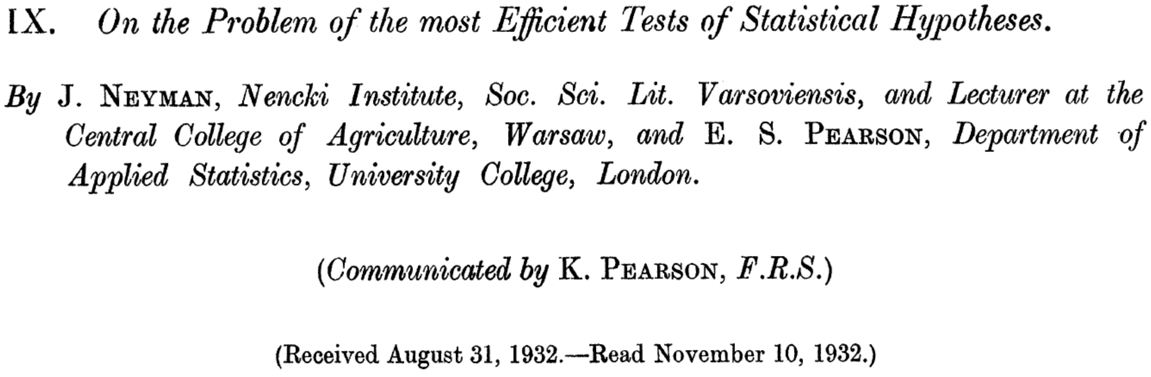
Where to live on the β vs α curve is a *long* discussion. (Even longer when considered as N events increases, so curve moves toward origin.)

***Decision* on whether 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)**

A one-size-fits-all criterion of α corresponding to 5σ is without foundation.

Classical Hypothesis Testing: Neyman-Pearson Lemma



IX. *On the Problem of the most Efficient Tests of Statistical Hypotheses.*

By J. Neyman, *Nencki Institute, Soc. Sci. Lit. Varsoviensis, and Lecturer at the Central College of Agriculture, Warsaw, and* E. S. Pearson, *Department of Applied Statistics, University College, London.*

(*Communicated by* K. Pearson, *F.R.S.*)

(Received August 31, 1932.—Read November 10, 1932.)

Phil. Transactions of the Royal Society of London. Vol. 231, (1933), pp. 289-337

If Type I error probability α is specified in a test of *simple* hypothesis H_0 against *simple* hypothesis H_1 ,

then the Type II error probability β is minimized by ordering according to the *likelihood ratio* $\lambda = \mathcal{L}(x|\ H_0)\ /\mathcal{L}(x|\ H_1)$.

One finds cutoff k_α for that α and rejects H_0 if $\lambda \leq k_\alpha$

The “lemma” applies only to a very special case: no nuisance parameters, not even undetermined parameters of interest!

But it has inspired many generalizations, and likelihood ratios are a oft-used component of both frequentist and Bayesian methods.

Conceptual proof in Second lecture of Kyle Cranmer, February 2009 http://indico.cern.ch/event/48426/ . See also Stuart99, p. 176

Classical Hypothesis Testing (cont.)

For rest of talk concentrate on:
H_0: $\mu = \mu_0$ (the "point null", or "sharp hypothesis") vs
H_1: $\mu \neq \mu_0$ (the "continuous alternative").

Common examples:
Signal strength μ of new physics: $\mu_0 = 0$, alternative $\mu>0$

$B_s^0 \rightarrow \mu^+\mu^-$ before discovery:
Null hypothesis is zero rate, alternative is positive rate;

$B_s^0 \rightarrow \mu^+\mu^-$ after discovery (essentially at same time): null is SM rate, alternative is any other rate

In classical/frequentist formalism (in contrast to Bayesian formalism), theory of these tests maps to that of confidence intervals!

Classical Hypothesis Testing: Duality

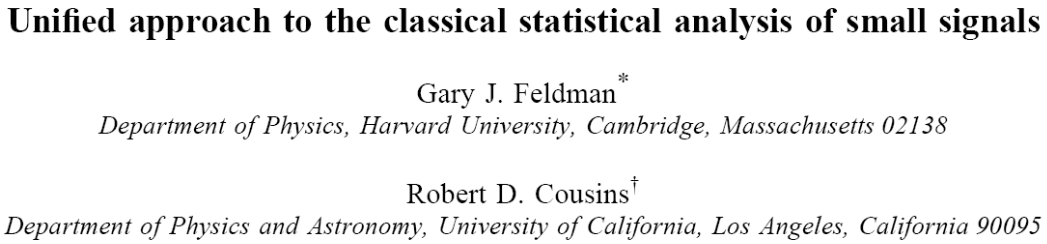
Given an ordering:
Test if $\mu=\mu_0$ vs $\mu\neq\mu_0$ at significance level α
↔ Is μ_0 in confidence interval for μ with C.L. = 1- α ?

"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 as in the dictionary in Table 20.1" Stuart99, p. 175. [Table in backup slides] E.g.,
α ↔ 1 – C.L.
Equal-tailed test ↔ central confidence intervals
One-tailed tests ↔ Upper/lower limits

Referred to as "inverting a test" to obtain intervals, and vice versa.

Classical Hypothesis Testing (cont.)

Test $\mu=\mu_0$ at $\alpha \leftrightarrow$ Is μ_0 in conf. int. for μ with C.L. = 1- α



Unified approach to the classical statistical analysis of small signals

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

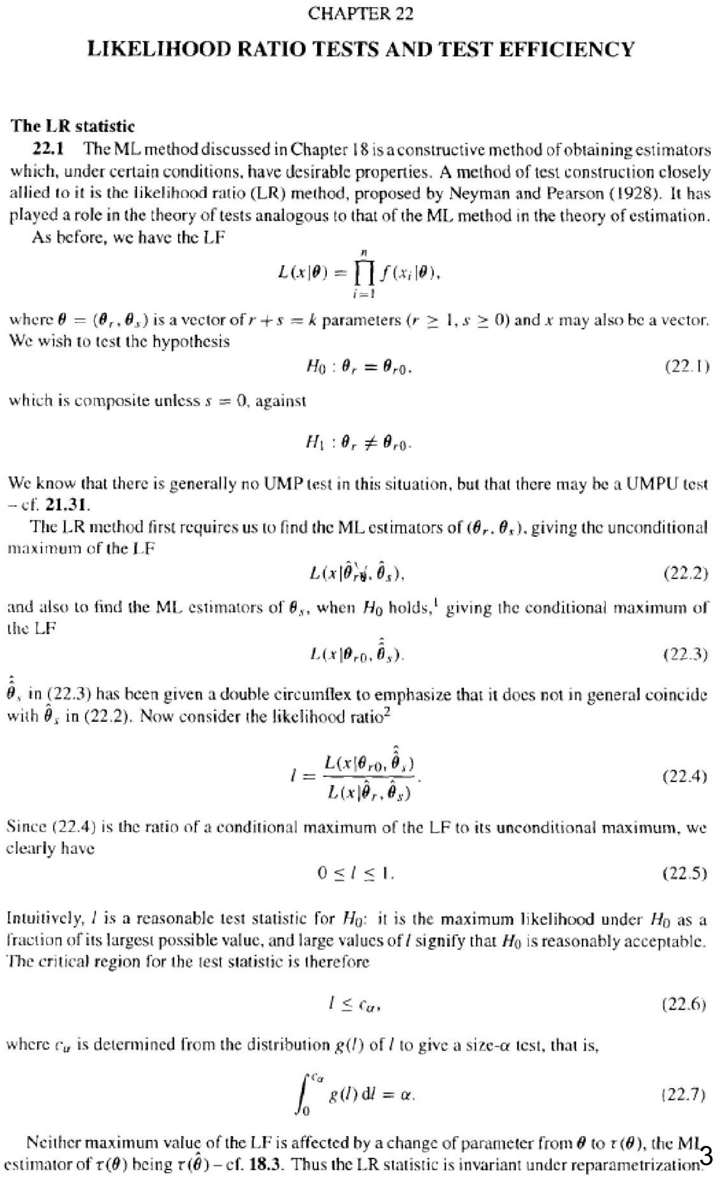
Robert D. Cousins†
Department of Physics and Astronomy, University of California, Los Angeles, California 90095

Phys. Rev. D57 3873 (1998):

We emphasized "new" ordering principle based on LR. While paper was "in proof", Gary realized that "our" intervals were simply those obtained by "inverting" the LR hypothesis test. In fact it was all on 1¼ pages of "Kendall and Stuart", plus nuisance paramers ! →

This was of course *good* !

It led to rapid inclusion in PDG RPP.



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.

Above is all “pre-data” characterization of the test

How to characterize *post-data*? P-values and Z-values

In N-P theory, α is *specified in advance*.

Suppose after obtaining data, you notice that with α=0.05 previously specified, you reject H_0, but with α=0.01 previously specified, you accept H_0. In fact, you determine that with the data set in hand, H_0 would be rejected for $\alpha \geq 0.023$. This interesting value has a name:

After data are obtained, the *p-value* is the smallest value of α for which H_0 would be rejected, *had it been specified in advance*.

Numerically (if not philosophically) the same as usual “value obtained or more extreme” due to Fisher.

Large literature bashing p-values.

I defend HEP: http://arxiv.org/abs/1310.3791

Interpreting p-values and Z-values

It is crucial to realize that that value of α was typically *not* specified in advance, so p-values do *not* correspond to Type I error rates of the experiments which report them.

Interpretation of p-values is a long, contentious story – beware!

In HEP, typically converted to Z-value (unfortunately commonly called "the significance S"), equivalent number of Gaussian sigma. (E.g.., for one-tailed test, p=2.87E-7 is Z=5.)

Whatever they are, p-values are not the probability that H_0 is true!

- **They are calculated *assuming that* H_0 *is true*, so they can hardly tell you the probability that H_0 is true!**
- **Calculation "probability that H_0 is true" requires prior(s)!**

Please help educate press officers and journalists!

Tentative stopping point

Likelihood (Ratio) Intervals

Recall from above: Likelihood $\mathcal{L}(\theta)$ is invariant under reparametrization from θ to $u(\theta)$: $\mathcal{L}(\theta) = \mathcal{L}(u(\theta))$.

So *likelihood ratios* $\mathcal{L}(\theta_1) / \mathcal{L}(\theta_2)$ and *log-likelihood differences* $\ln\mathcal{L}(\theta_1) - \ln\mathcal{L}(\theta_2)$ are also invariant.

Thus, after using maximum-likelihood method to obtain estimate $\hat{u}$ which maximizes $\mathcal{L}(u)$, one can obtain a likelihood interval $[u_1,u_2]$ as the union of all u for which

$$2\ln\mathcal{L}(\hat{u}) - 2\ln\mathcal{L}(u) \leq Z^2, \text{ for Z real.}$$

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

Convergence to Gaussian is faster than you might expect. See James06 for interesting explanation why.

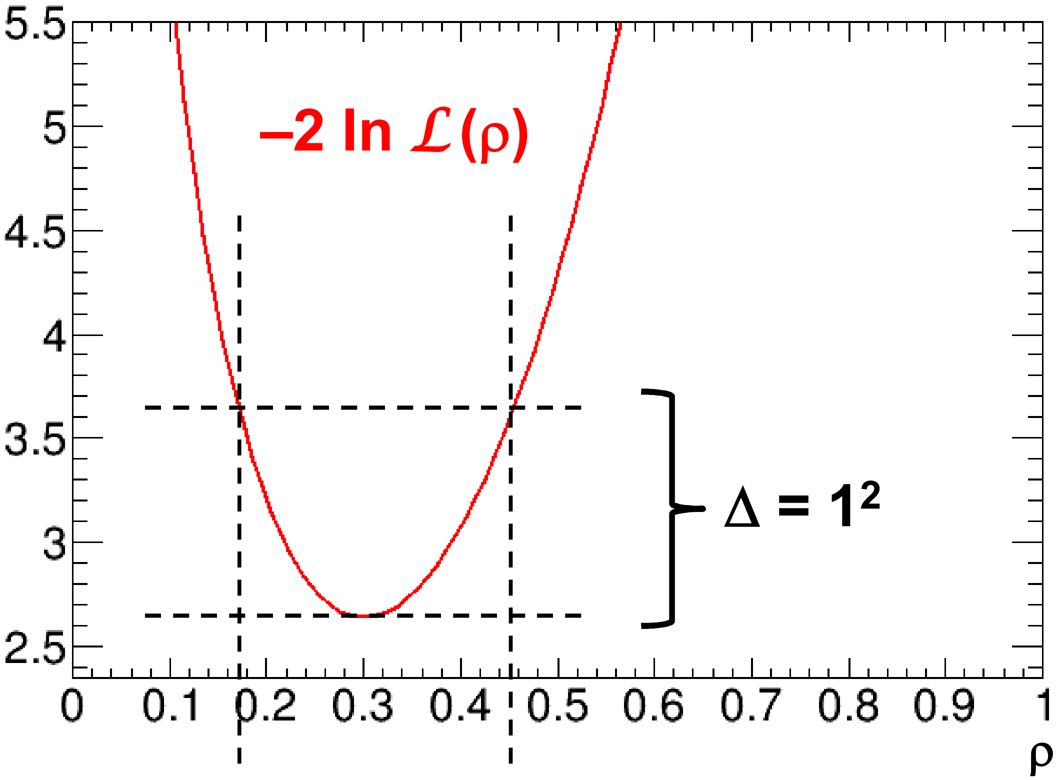
But! Regularity conditions, in particular requirement that $\hat{u}$ not be on the boundary, need to be carefully checked.
(E.g., if $u \geq 0$ on physical grounds, then $\hat{u}=0$ requires care.)

Binomial Likelihood-Ratio Interval example

Recall example of n_{on}=3 successes in n_{tot}=10 trials.

Min $-2 \ln \mathcal{L}(\rho)$ = 2.64.
Obtain interval from
$-2 \ln \mathcal{L}(\rho) = 2.64 + 1 = 3.64$

⇒ likelihood-ratio interval
$[\rho_1,\rho_2]$ = [0.17, 0.45]



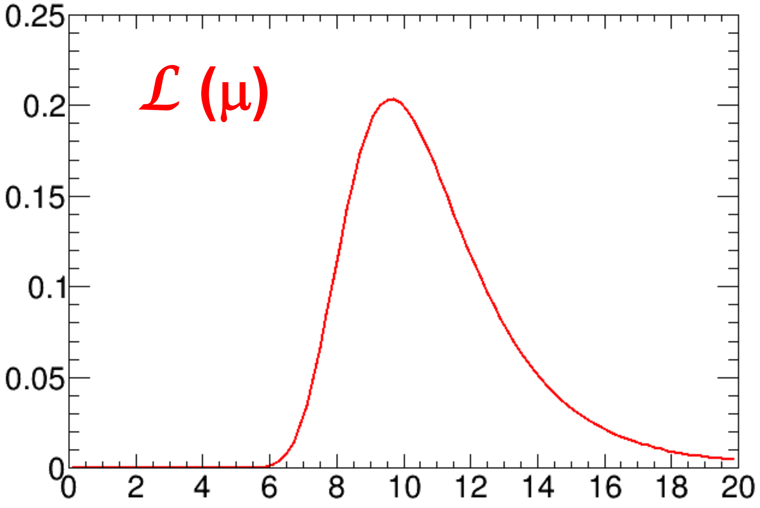
Recall:
Copper-Pearson $[\rho_1,\rho_2]$ = [0.14, 0.51]
Wilson $[\rho_1,\rho_2]$ = [0.18, 0.46]

Gaussian pdf $p(x|\mu,\sigma)$ with σ a function of μ: $\sigma = 0.2\ \mu$ Observed x_0 = 10.0.

Recall:

$\mathcal{L}(\mu)$ for observed x_0 = 10.0.

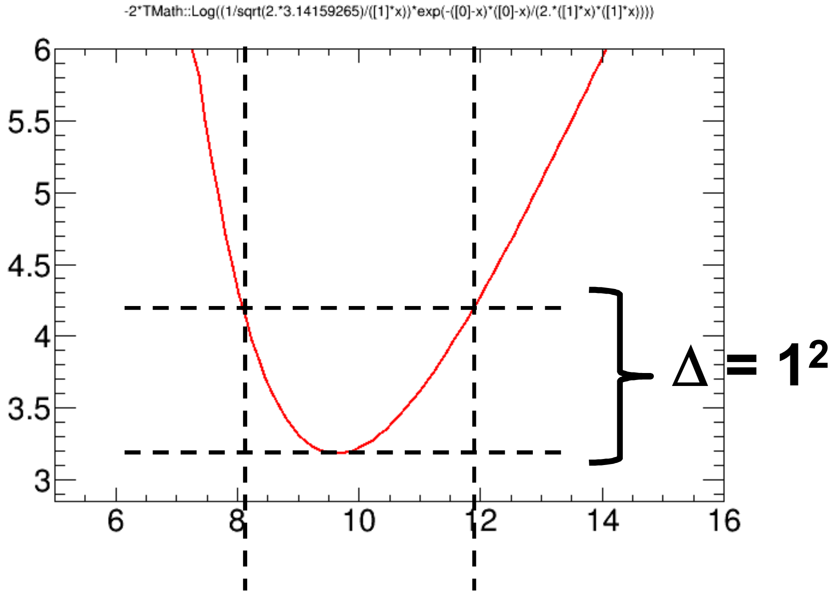
μ_{ML}= 9.63



Likelihood ratio interval for μ at approximate 68% C.L.:

$[\mu_1, \mu_2]$ = [8.10, 11.9].

Compare with exact confidence interval [8.33,12.5].



Poisson Likelihood-Ratio Interval example

Approx "68% C.L." likelihood-ratio interval for Poisson process with n=3 observed:

$\mathcal{L}(\mu) = \mu^3 \exp(-\mu)/3!$

Maximum at $\mu = 3$.

$\Delta 2\ln\mathcal{L} = 1^2$ yields LR interval $[\mu_1, \mu_2] = [1.58, 5.08]$

Neyman construction central: $[\mu_1, \mu_2] = [1.37, 5.92]$

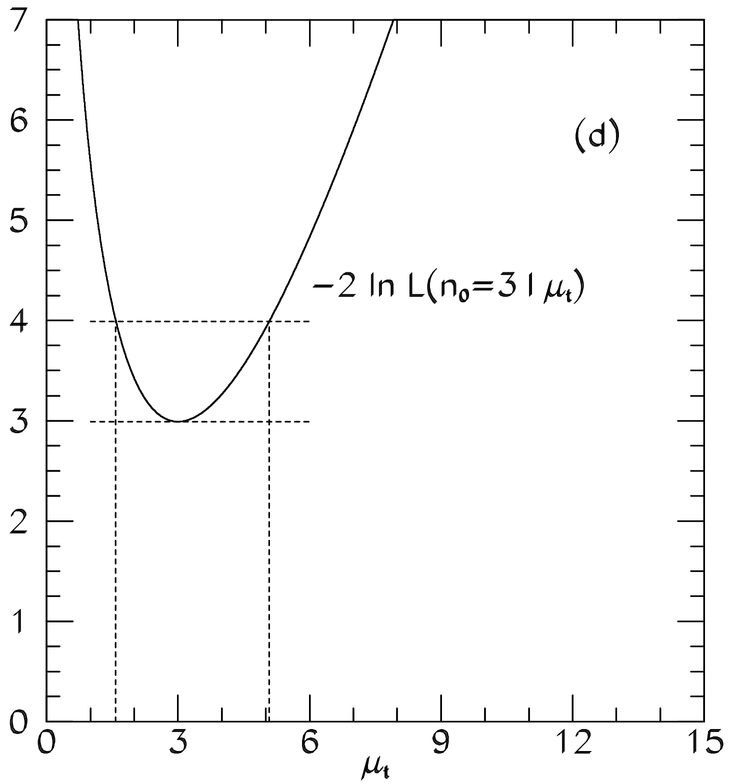


Figure from R. Cousins, Am. J. Phys. 63 398 (1995)

References Cited in Talk Slides

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

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.

Recommended reading

Books: Among the many books available, I usually recommend the following progression, reading the first three cover-to-cover, and consulting the next two 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. (Comprehensive old treatise on classical frequentist statistics; anyone contemplating a NIM paper on statistics should look in here first!)

5) George Casella and R.L. Berger, Statistical Inference, 2nd, Ed. 2002. A more modern, less dense text on similar topics as Kendall and Stuart.

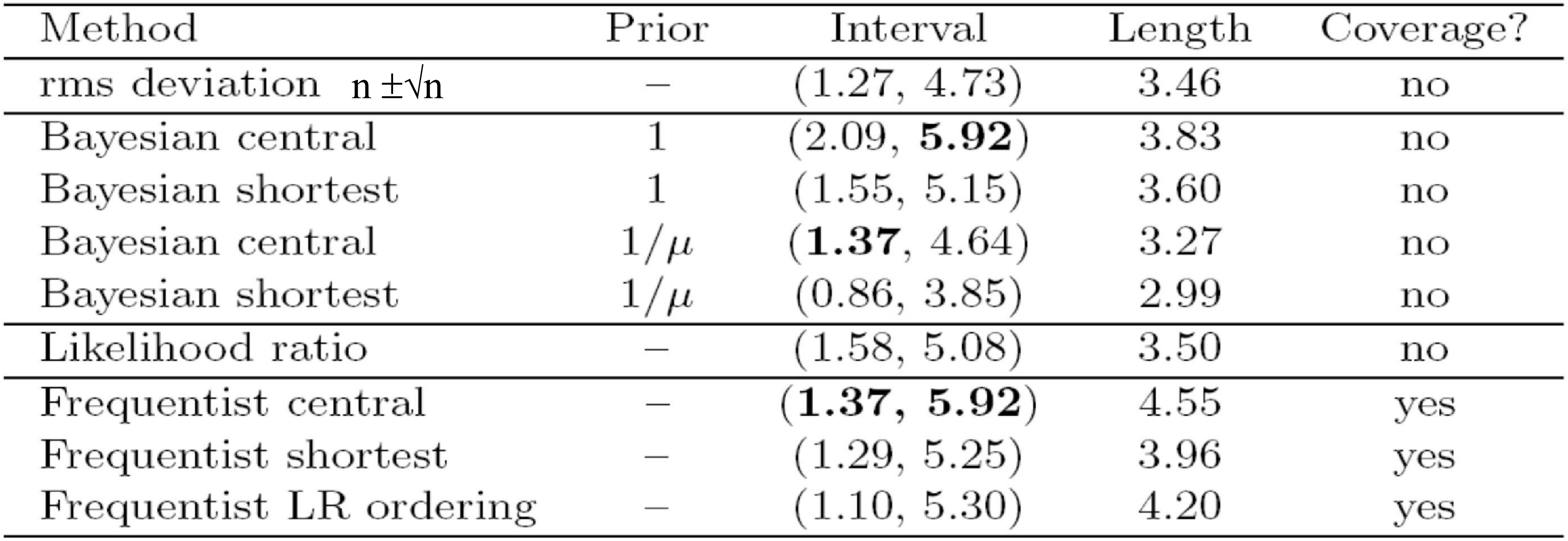
PhyStat conference series: Beginning with Confidence Limits Workshops in 2000, links at http://phystat-lhc.web.cern.ch/phystat-lhc/ and http://www.physics.ox.ac.uk/phystat05/

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)

End of Part 2

BACKUP

68% intervals by various methods for Poisson process with n=3 observed



Method	Prior	Interval	Length	Coverage?
rms deviation n ±√n	–	(1.27, 4.73)	3.46	no
Bayesian central	1	(2.09, **5.92**)	3.83	no
Bayesian shortest	1	(1.55, 5.15)	3.60	no
Bayesian central	$1/\mu$	(**1.37**, 4.64)	3.27	no
Bayesian shortest	$1/\mu$	(0.86, 3.85)	2.99	no
Likelihood ratio	–	(1.58, 5.08)	3.50	no
Frequentist central	–	(**1.37, 5.92**)	4.55	yes
Frequentist shortest	–	(1.29, 5.25)	3.96	yes
Frequentist LR ordering	–	(1.10, 5.30)	4.20	yes

For the Jeffreys prior (1/√μ), Bayesian central interval is (1.72, 5.27).

Frequentist intervals over-cover due to discreteness of n.

Adapted from Cousins05 and
R. Cousins, Am. J. Phys. 63 398 (1995)

Classical Goodness of Fit (g.o.f.)

If H_0 is specified but the alternative H_1 is not, then only the Type I error rate α can be calculated, since the Type II error rate β depends on a H_1.

A test with this feature is called a test for *goodness-of-fit* (to H_0).

The question "Which g.o.f. test is best?" is thus ill-posed. In spite of the popularity of tests with universal maps from test statistics to α (in particular χ^2 and Kolomogorov tests), they may be ill-suited for many problems (i.e., they may have poor power (1- β) against relevant alternative H_1's).

In 1D, unbinned g.o.f. test question is equivalent to:
"Given 3 numbers (e.g. neutrino mixing angles) in [0, 1], are they consistent with three calls to RAN() ?"

Have fun with that!

Goodness of Fit (cont.)

Issue in last 15 years: need for a multi-D unbinned test.

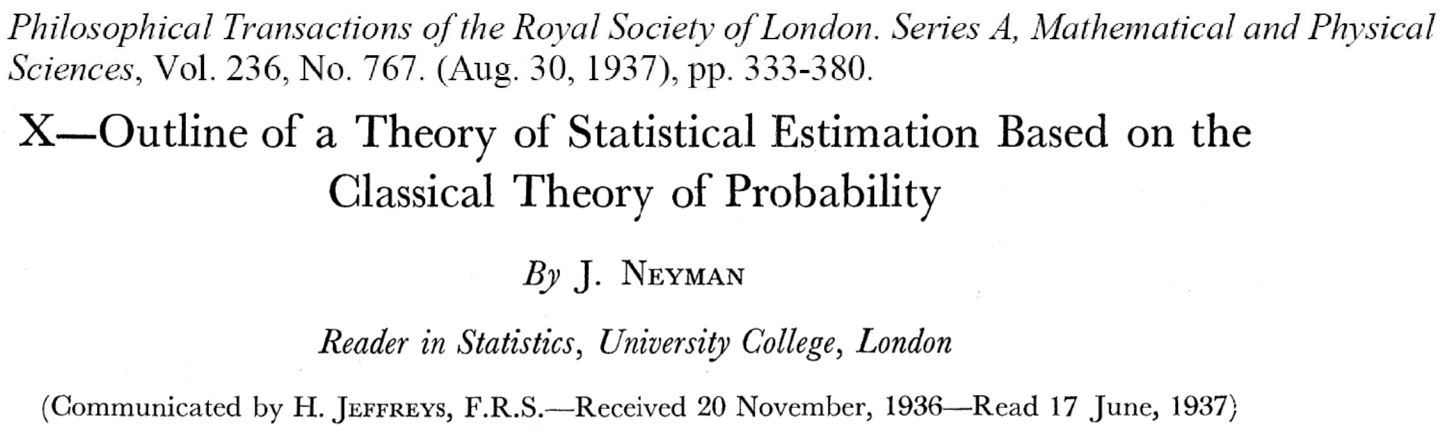
E.g., is it reasonable that 1000 events scattered in 5D have been drawn from a particular pdf (which may have parameters which were fit using an unbinned M.L. fit to those 1000 events.) ?

Of course this is an ill-posed question, but looking for good omnibus test. Getting the null distribution from M.C. is typically doable, it seems.

See Aslan02 and others at past PhyStats.

1D issues well-described in book by D'Agostino and Stephens (must-read for those wanting to invent a new test).

Recent review by Mike Williams, "How good are your fits? Unbinned multivariate goodness-of-fit tests in high energy physics", http://arxiv.org/abs/1006.3019



Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, Vol. 236, No. 767. (Aug. 30, 1937), pp. 333-380.

X—Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability

By J. NEYMAN

Reader in Statistics, University College, London

(Communicated by H. JEFFREYS, F.R.S.—Received 20 November, 1936—Read 17 June, 1937)

Original paper has one unknown parameter θ_1 and two observables x_1,x_2 per expt:

E is vector of observables x_1, x_2, …
$A(\theta)$ is acceptance region: $P(E \in A)$ = C.L.
θ_1 is unknown parameter

E′ is data actually observed in expt.

Prior to experiment , regions in E-space $A(\theta_1)$ are determined for each θ_1 (needs ordering principle). Upon obtaining data E′, confidence interval for θ_1 consists of all values of θ_1 for which E′ is in $A(\theta_1)$.

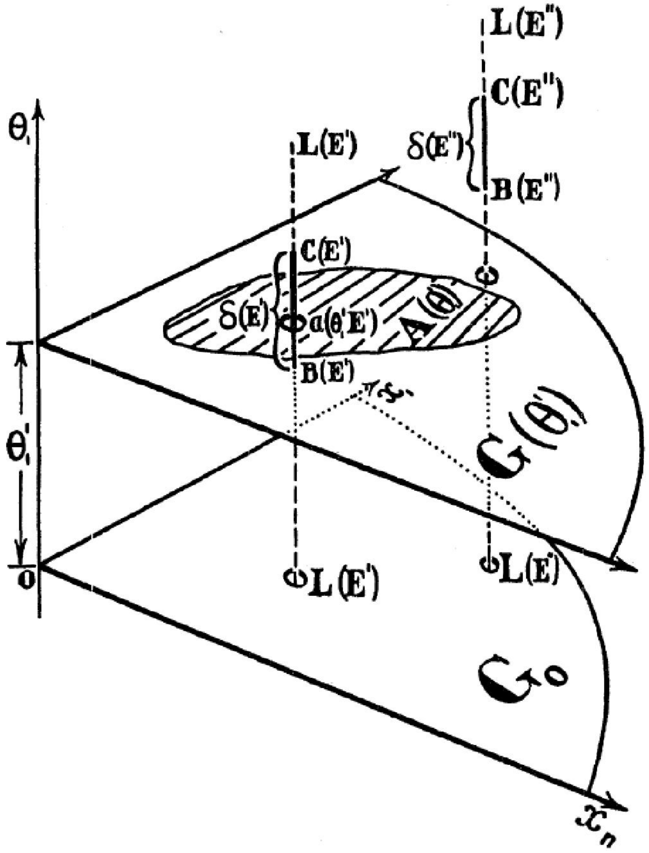


FIG. 1—The general space G.

Classical Hypothesis Testing: Duality

Test $\mu=\mu_0$ at $\alpha \leftrightarrow$ Is μ_0 in conf. int. for μ 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 as in the dictionary in Table 20.1” Stuart99, p. 175.

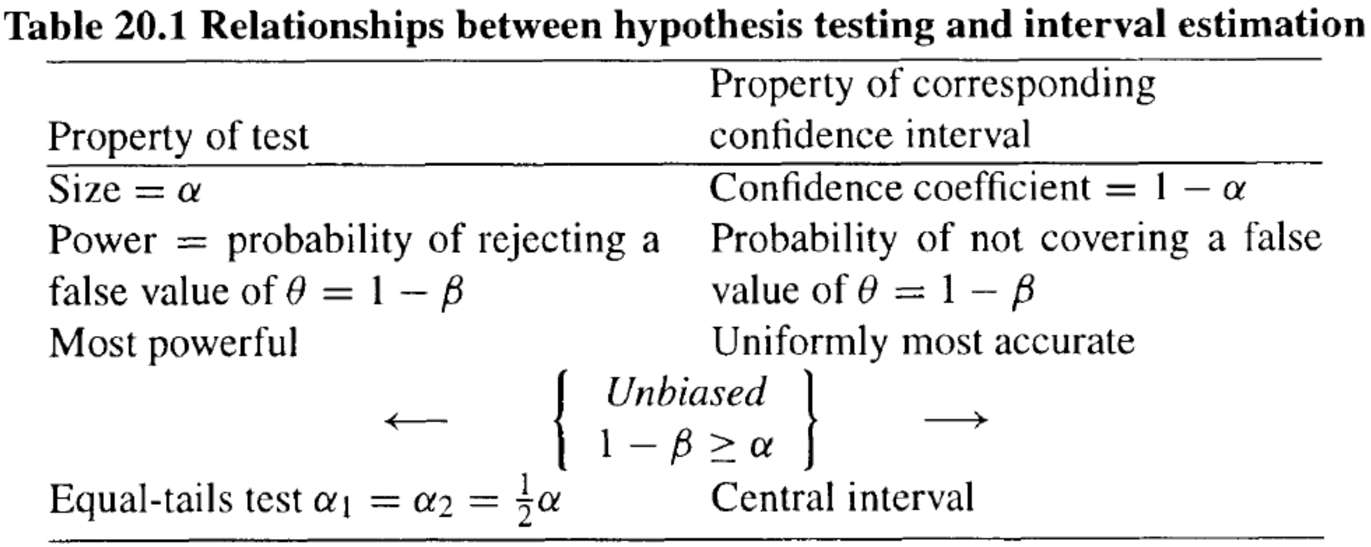


Table 20.1 Relationships between hypothesis testing and interval estimation

Property of test	Property of corresponding confidence interval
Size $= \alpha$	Confidence coefficient $= 1 - \alpha$
Power $=$ probability of rejecting a false value of $\theta = 1 - \beta$	Probability of not covering a false value of $\theta = 1 - \beta$
Most powerful	Uniformly most accurate
$\longleftarrow \left\{ \begin{array}{c} \textit{Unbiased} \\ 1 - \beta \geq \alpha \end{array} \right\} \longrightarrow$	
Equal-tails test $\alpha_1 = \alpha_2 = \frac{1}{2}\alpha$	Central interval

Referred to as “inverting a test” to obtain intervals; vice versa.