“Statistics in Theory”: Prelude to Statistics in Practice (“long” 4-hour version)

Bob Cousins
Univ. of California, Los Angeles

Hadron Collider Physics Summer School
August 20-21, 2018

For more complete writeup with references, see arXiv post
https://arxiv.org/abs/1807.05996 **.**

Preface

Many of us teach advanced "data analysis" courses that last an entire academic term. How to condense?

Here I concentrate on the "theoretical" underpinnings:
What you must know in order to choose appropriate methods.

My hope is that by studying these slides you will learn to avoid common pitfalls (and even silly statements) that frequently trip up professionals in the field.

This is a dense talk – you will not pick it all up in real time. It should however be extremely useful to you to study this talk, referring to the references.

I initially focus on definitions and the Bayesian approach. That helpe to understand what the frequentist approach is *not* ! Comparison of approaches comes later.

Thanks to many (F. James, G. Feldman, L. Lyons, L. Demortier, et al., + many statisticians) from whom I learned.

Why Foundations Matter

In the “final analysis”, we often make approximations, take a pragmatic approach, or follow a convention.

To inform such actions, it is important to understand some foundational aspects of statistical inference.

In Quantum Mechanics, we are used to the fact that for all of our practical work, one’s philosophical interpretation (e.g., of collapse of the wave function) does not matter.

Why Foundations Matter (cont.)

In statistical inference, however, *foundational differences result in different answers*: one cannot ignore them!

The professional statistics community went through the topics of many of our discussions starting in the 1920's, and revisited them in the resurgence of Bayesian methods in recent decades.

I will attempt to summarize some of the things we should understand from that debate.

***Most importantly*: understand both approaches!**

Definitions are Important

As in physics, much confusion can be avoided by being precise about definitions, and much confusion can be generated by being imprecise, or by assuming every-day definitions in a technical context.

You should see just as much confusion in these two statements:

1) **"The *confidence* level tells you how much confidence one has that the true value is in the confidence interval."**
2) **"A *noninformative* prior probability density does not insert any information."**

...as you have learned to see in the statement,

"I did a lot or *work* today by carrying this big stone around the building and then putting it back in its original place."

Example adapted from Eadie et al. (James06, p. 2)

Physicists say…	when Statisticians say:
Determine	Estimate
Estimate	(Informed) Guess
Gaussian	Normal
Breit-Wigner, Lorentzian	Cauchy

Key tasks: Important to distinguish!

- ***Point estimation:*** **what single "measured" value of a parameter do you report?**

Key tasks: Important to distinguish!

- ***Point estimation:*** **what single "measured" value of a parameter do you report?**
- ***Interval estimation:*** **what interval (giving a measure of uncertainty of the parameter inference) do you report?**

Key tasks: Important to distinguish!

- ***Point estimation:*** **what single "measured" value of a parameter do you report?**
- ***Interval estimation:*** **what interval (giving a measure of uncertainty of the parameter inference) do you report?**
- ***Hypothesis testing:*** **Many special cases:**
  - ***a) A given functional form ("model") vs another functional form. Also known as "model selection".***
  - ***b) A single value of a parameter (say 0 or 1) vs all other values***
  - ***c) Goodness of Fit: A given functional form against all other (unspecified) functional forms (aka "model checking")***

Key tasks: Important to distinguish!

- ***Point estimation:*** **what single "measured" value of a parameter do you report?**
- ***Interval estimation:*** **what interval (giving a measure of uncertainty of the parameter inference) do you report?**
- ***Hypothesis testing:*** **Many special cases:**
  - ***a) A given functional form ("model") vs another functional form. Also known as "model selection".***
  - ***b) A single value of a parameter (say 0 or 1) vs all other values***
  - ***c) Goodness of Fit: A given functional form against all other (unspecified) functional forms (aka "model checking")***
- ***Decision making:*** **What action should I take (tell no one, issue press release, propose new experiment, ...) based on the observed data? Rarely done formally in HEP, but important to understand outline of formal theory, to avoid confusion with inference and to inform informal application.**

Key tasks: Important to distinguish! (cont.)

In frequentist statistics, the above hypothesis testing case,

(b) A single value of a parameter (say 0 or 1) vs all other values,

maps identically onto interval estimation.

This is called the duality of "inversion of a hypothesis test to get confidence interval", and vice versa. I just mention it now but discuss it in more detail later.

In contrast, in Bayesian statistics, testing case (b) is an especially controversial form of case (a) model selection. The model with fixed value of parameter is lower-dimensional in parameter space than the model with parameter not fixed.

Again, I just mention this now to foreshadow a very deep issue, where frequentist and Bayesian methods do not converge in the limit of large data sets.

Comments on key tasks

- ***Point estimation:*** **Long history. In the end it is not clear what the criteria are for "best" estimator. Decision Theory can be used to specify criteria and choose among point estimators.**
  IMO, point estimation is not a key issue; typically the ML point estimator serves our needs in HEP.
- ***Interval estimation:*** **In HEP, it is fairly mandatory that there is a *confidence level* that gives *frequentist coverage* probability of a method, even if it's a Bayesian-inspired recipe. For many problems in HEP, there is reasonable hope of approximate reconciliation between Bayesian and frequentist methods,.**

Point estimation and interval estimation can be approached consistently by insisting that the interval estimate contain the point estimate; in that case one constructs the point estimate by taking the limit of interval estimates as intervals get smaller (limit of confidence level going to zero).

Comments on key tasks (cont.)

- ***Hypothesis testing:***
  - **Bayesian methods attempt to calculate probability that a hypothesis is true.**
  - **Frequentist methods use p-values (often bashed).**

Can be dramatic differences between frequentist and Bayesian hypothesis testing methods, even asymptotically. Beware!

See my paper on Jeffreys-Lindley paradox,
https://arxiv.org/abs/1310.3791 .

Much more in subsequent slides.

"Probability"

- **Abstract mathematical probability P can be defined in terms of sets and axioms that P obeys. Conditional probabilities are related by Bayes' Theorem (or "Bayes' Rule") -- see next slide,**

  $P(B|A) = P(A|B)\, P(B) / P(A)$.

- **Two established* incarnations of P are:**

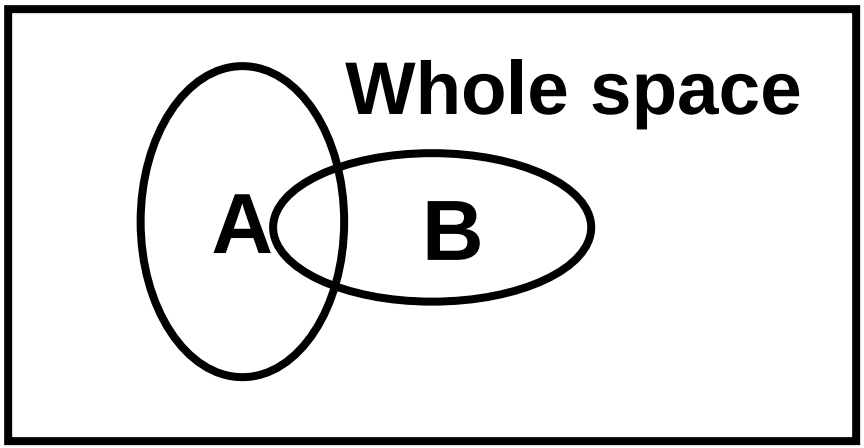
  ***1) Frequentist P*: limiting frequency in ensemble of imagined repeated samples (as usually taught in Q.M.).**
  P(constant of nature) and P(SUSY is true) do not exist (in a useful way) for this definition of P (at least in one universe).

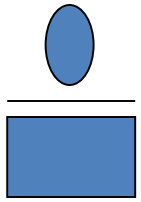
  ***2) (Subjective) Bayesian P*: subjective (personalistic) *degree of belief.* (de Finetti, Savage)**
  P(constant of nature) and P(SUSY is true) exist for You.
  Shown to be basis for coherent personal decision-making.

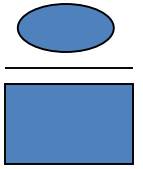
- ***It is important to be able to work with either definition of P, and to know which one you are using!***

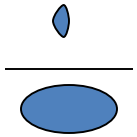
***Of course they are still argued about, but to less practical effect, I think.**

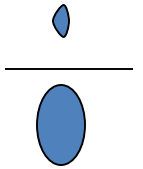
P, Conditional P, and "Derivation" of Bayes' Theorem

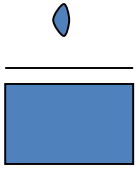


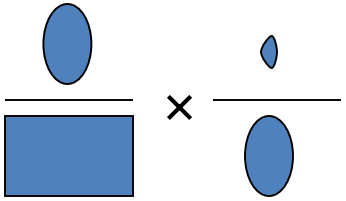
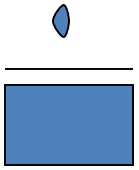
$P(A) =$ 

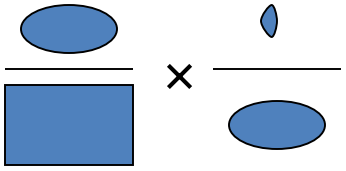
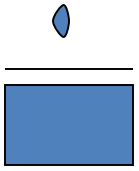
$P(B) =$ 

$P(A|B) =$ 

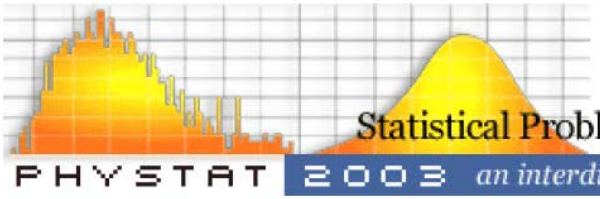
$P(B|A) =$ 

$P(A \cap B) =$ 

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

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

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



Statistical Problems in Particle Physics, Astrophysics, and Cosmology

an interdisciplinary gathering of statisticians and physicists in these fields

Bayesians, Frequentists, and Physicists

Bradley Efron

Department of Statistics and Department of Health Research and Policy, Stanford University, Stanford, CA 94305, USA

"Bayes' rule is satisfying, convincing, and fun to use. But using Bayes' rule does not make one a Bayesian; *always* using it does, and that's where difficulties begin."

I'll give a simple example for each definition of P. One of the sillier things one sometimes sees in HEP is the use of a frequentist example of Bayes' Theorem as a foundational argument for "Bayesian" statistics.

Aside: What is the “Whole Space”?

For probabilities to be well-defined, the “whole space” needs to be defined. Can be hard for both frequentists and Bayesians!

Thus the “whole space” itself is more properly thought of as a conditional space, conditional on the assumptions going into the model (Poisson process, whether or not total number of events was fixed, etc.).

Furthermore, it is widely accepted that restricting the “whole space” to a relevant (“conditional”) subspace can sometimes improve the quality of statistical inference. The important topic of such “conditioning” in frequentist inference will be discussed in detail later.

I will not clutter the notation with explicit mention of the assumptions defining the “whole space”, but some prefer to do so – in any case, it is important to keep them in mind.

Example of Bayes' Theorem Using Frequentist P

A b-tagging method is developed and one measures:

P(btag | b-jet), i.e., efficiency for tagging b's

P(btag | not a b-jet), i.e., efficiency for background

P(no btag | b-jet) = 1 - P(btag | b-jet),

P(no btag | not a b-jet) = 1 - P(btag | not a b-jet)

Question: Given a selection of jets tagged as b-jets, what fraction of them is b-jets? I.e., what is P(b-jet | btag) ?

Example of Bayes' Theorem Using Frequentist P

A b-tagging method is developed and one measures:

P(btag | b-jet), i.e., efficiency for tagging b's

P(btag | not a b-jet), i.e., efficiency for background

P(no btag | b-jet) = 1 - P(btag | b-jet),

P(no btag | not a b-jet) = 1 - P(btag | not a b-jet)

Question: Given a selection of jets tagged as b-jets, what fraction of them is b-jets? I.e., what is P(b-jet | btag) ?

Answer: *Cannot be determined from the given information!*

Need in addition: P(b-jet), the true fraction of *all* jets that are b-jets. Then Bayes' Thm inverts the conditionality:

$$P(\text{b-jet} \mid \text{btag}) \propto P(\text{btag} \mid \text{b-jet})\, P(\text{b-jet})$$

Example of Bayes' Theorem Using Frequentist P (cont.)

In HEP, as noted,

P(btag | b-jet) is called the *efficiency* for tagging b's.

Meanwhile

P(b-jet | btag) is often called the *purity* of a sample of b-tagged jets.

As this is a pretty "easy" distinction, it is helpful to keep it in mind when one encounters cases where it is perhaps tempting to make the logical error of equating P(A|B) and P(B|A).

Example of Bayes' Theorem Using Bayesian P

In a *background-free* experiment, a theorist uses a "model" to predict a signal with Poisson mean of 3 events. From Poisson formula we know

P(0 events | model true) = $3^0 e^{-3}/0! = 0.05$

P(0 events | model false) = 1.0

P(>0 events | model true) = 0.95

P(>0 events | model false) = 0.0

The experiment is performed and *zero events are observed*.

Question: Given the result of the expt, what is the probability that the model is true? I.e., What is P(model true | 0 events) ?

Example of Bayes' Theorem Using Bayesian P

In a *background-free* experiment, a theorist uses a "model" to predict a signal with Poisson mean of 3 events. From Poisson formula we know

P(0 events | model true) = $3^0 e^{-3}/0! = 0.05$

P(0 events | model false) = 1.0

P(>0 events | model true) = 0.95

P(>0 events | model false) = 0.0

The experiment is performed and *zero events are observed*.

Question: Given the result of the expt, what is the probability that the model is true? I.e., What is P(model true | 0 events) ?

Answer: *Cannot be determined from the given information!*
Need in addition: P(model true), the *degree of belief* in the model *prior* to the experiment. Then Bayes' Thm inverts the conditionality:

P(model true | 0 events) $\propto$ P(0 events | model true) P(model true)

Example of Bayes' Theorem Using Bayesian P (cont.)

P(0 events | model true) = 0.05

P(0 events | model false) = 1.0

0 events observed

Apply Bayes' Thm in a little more detail, with normalization:

Let "A" $\leftrightarrow$ "0 events"; let "B" $\leftrightarrow$ "model true". Recall:

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

Similarly, with P(not B) = 1 – P(B)

P(not B|A) = P(A| not B) $\times$ P(not B) / P(A) .

P(B|A) + P(not B|A) = 1, so P(A) is normalization. Can easily show

P(B|A) = 0.05 P(B) / (1 – 0.95 P(B)), i.e.,

P(model true | 0 events)

= 0.05 P(model true) / (1 – 0.95 P(model true)).

Example of Bayes' Theorem Using Bayesian P (cont.)

P(model true | 0 events)

= 0.05 P(model true) / (1 – 0.95 P(model true)).

Limiting cases to lowest order in ε, with $\varepsilon << 1$:

Let "model" be S.M., prior P(model true) = $1 - \varepsilon$,

$\Rightarrow$ P(model true | 0 events) $\approx 1 - 20\varepsilon$

Still very high degree of belief! Even if someone says,

"P(0 events | model true) = 5%, and 0 events observed means there is 5% chance the S.M. is true." (UGH!)

Let "model" be large extra dimensions, prior P(model true) = ε ,

$\Rightarrow$ P(model true | 0 events) $\approx 0.05\ \varepsilon$

Low prior belief becomes even lower.

N.B. More realistic examples are of course more complex.

A Note re *Decisions*

Suppose that as a result of the previous experiment, your degree of belief in the model is P(model true | 0 events) = 99%, and you need to *decide* on an action (making a press release, or planning next experiment), based on the model being true.

Question: What should you *decide*?

A Note re *Decisions*

Suppose that as a result of the previous experiment, your degree of belief in the model is P(model true | 0 events) = 99%, and you need to *decide* on an action (making a press release, or planning next experiment), based on the model being true.

Question: What should you *decide*?

Answer: *Cannot be determined from the given information!*

Need in addition:

The *utility* function (or its negative, the *loss* function), which quantifies the relative costs (to You) of

- **Type I error (declaring model false when it is true), and**
- **Type II error (not declaring model false when it is false).**

A Note re *Decisions* (cont.)

Thus, Your *decision*, requires two subjective inputs: Your prior probabilities, and the relative costs to You of outcomes.

Statisticians often focus on decision-making.

In HEP, the tradition thus far is to communicate experimental results (well) short of formal decision calculations.

It should become clear later in lectures:

Frequentist (classical) "hypothesis testing" (especially with conventions like 95% C.L. or 5σ) is not a complete theory of decision-making!

It is important to keep this in mind, since the "accept/reject" language of classical hypothesis testing (later in lectures) is too simplistic for "deciding" in important situations.

Probability, Probability Density, Likelihood

These are key building blocks in both frequentist and Bayesian statistics, and it is crucial distinguish among them.

In the following we let x be an observed or measured quantity; sometimes we use n if the observation is integer-valued and we want to emphasize that.

A "(statistical) model" is an expression specifying probabilities or probability densities for observing x.

We use μ for parameters (sometimes vector-valued) in the model. (Statistical literature prefers θ.)

Then the most common examples in HEP:

Probability, Probability Density, Likelihood (cont.)

- **Binomial probability of n_{on} successes in n_{tot} trials, each with binomial parameter ρ:**

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

- **Poisson *probability* $P(n|\mu) = \mu^n \exp(-\mu)/n!$**

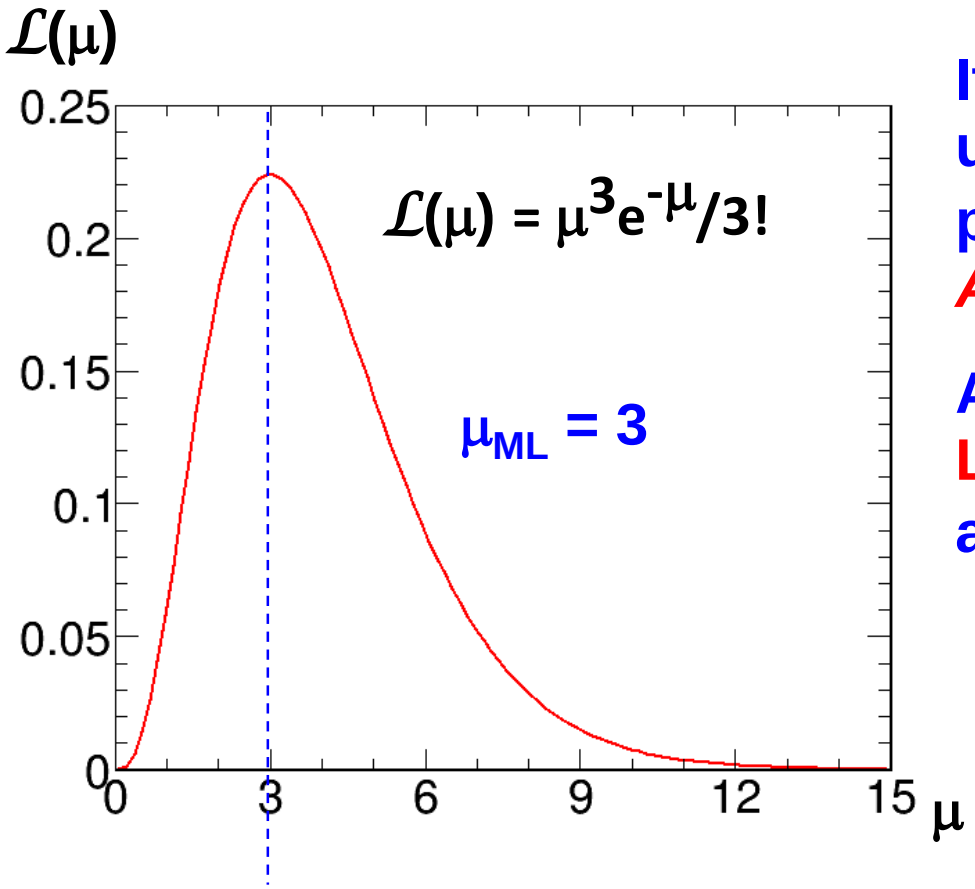
- **Gaussian *probability density function* (pdf) $p(x|\mu,\sigma)$: $p(x|\mu,\sigma)dx$ is differential of probability dP.**

In Poisson case, suppose n=3 is observed.
Substituting *observed value* n=3 into $P(n|\mu)$ yields the *likelihood function*

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

Probability, Probability Density, Likelihood (cont.)

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



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.

As we shall see,
Likelihood *Ratios* $\mathcal{L}(\mu_1)\,/\mathcal{L}(\mu_2)$
are useful and frequently used.

Notation reminder

x denotes observable(s)

More generally, x is any convenient or useful function of the observable(s), and is called a "statistic" or "test statistic"

μ denotes parameter(s)

p(x|μ) is probability/pdf characterizing everything that determines the probabilities (densities) of the observations, from laws of physics to experiment setup and protocol

p(x|μ) is called "the statistical model", or simply "the model", by statisticians.

Change of observable variable ("metric") x in pdf p(x|μ)

For pdf p(x|μ) and 1-to-1 change of variable (metric) from x to y(x), volume element modified by Jacobian. In 1D, p(y) |dy| = p(x) |dx|.

$$p(y(x)|\mu) = p(x|\mu) \,/\, |dy/dx|.$$

Jacobian modifies probability *density*, guaranties that

$$P(y(x_1) < y < y(x_2)) = P(x_1 < x < x_2),$$

(or equivalent with decreasing y(x)), i.e., guarantees that

***Probabilities are invariant under change of variable* x.**

E.g., for $x \leftrightarrow \tau$ and $y(x) \leftrightarrow \Gamma = 1/\tau$, must have

$$P(\tau \subset [\tau_1, \tau_2]) = P (\Gamma \subset [1/\tau_2, 1/\tau_1])$$

Mode of probability *density* is *not* invariant (so, e.g., criterion of maximum probability density is ill-defined).

Likelihood *ratio* $\mathcal{L}(\mu_1) / \mathcal{L}(\mu_2)$ is invariant under change of variable x to y(x). (Jacobian in denominator cancels that in numerator).

Probability Integral Transform

"…seems likely to be one of the most fruitful conceptions introduced into statistical theory in the last few years"

– Egon Pearson (1938) commenting on his father's work.

Given continuous $x \in (a,b)$, and its pdf $p(x)$, let

$$y(x) = \int_a^x p(x') \, dx' \; .$$

Then $y \in (0,1)$ and easy to show that $p(y) = 1$ (uniform) for all y. (!)

So there always exists a metric y in which the pdf is uniform.

***Many* issues become more clear (or trivial) after this transformation*. (If x is discrete, some complications.)**

Probability Integral Transform

"…seems likely to be one of the most fruitful conceptions introduced into statistical theory in the last few years"

– Egon Pearson (1938) commenting on his father's work.

Given continuous $x \in (a,b)$, and its pdf $p(x)$, let

$$y(x) = \int_a^x p(x')\,dx' \ .$$

Then $y \in (0,1)$ and easy to show that $p(y) = 1$ (uniform) for all y. (!)

So there always exists a metric y in which the pdf is uniform.

***Many* issues become more clear (or trivial) after this transformation*. (If x is discrete, some complications.)**

A look ahead, just mentioned here: The specification of a Bayesian prior pdf $p(\mu)$ for parameter μ is thus equivalent to the choice of the metric $g(\mu)$ in which the pdf is uniform. This is a *deep* issue, not always recognized by users of uniform prior pdf's in HEP!

***And the inverse transformation provides for efficient M.C. generation of p(x) starting from RAN().**

Change of parameter μ in pdf $p(x|\mu)$

The pdf for x given parameter μ=3 is the *same* as

the pdf for x given $1/\mu$=1/3, or given μ^2=9, or given any specified function of μ.

They all imply the same μ, and hence the same pdf for x.

In slightly confusing notation, that is what we mean by changing parameter from μ to $f(\mu)$, and saying that

$$p(x|f(\mu)) = p(x|\mu).$$

Thus the likelihood $\mathcal{L}(\mu)$ is *invariant (!)* under reparametrization from parameter μ to $f(\mu)$:

$$\mathcal{L}(\, f(\mu)\,) = \mathcal{L}(\mu).$$

This reinforces the fact that $\mathcal{L}(\mu)$ is *not* a pdf in μ.

Bayes' Theorem Generalized to Probability Densities

Recall $P(B|A) \propto P(A|B)\, P(B)$. For Bayesian P, continuous parameters such as μ are random variables with pdf's.

Let pdf $p(\mu|x)$ be the conditional pdf for parameter μ, given data x. As usual $p(x|\mu)$ is the conditional pdf for data x, given parameter μ. Then Bayes' Thm becomes

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

Substituting in a particular set of observed data, x_0 :

$$p(\mu|x_0) \propto p(x_0|\mu)\, p(\mu).$$

Bayes' Theorem Generalized to Probability Densities

$p(\mu|x_0) \propto p(x_0|\mu)\, p(\mu)$.

Recognizing the likelihood (variously written as $\mathcal{L}(x_0|\mu)$, $\mathcal{L}(\mu)$, or unfortunately even $\mathcal{L}(\mu|x_0)$), then

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

- $p(\mu|x_0)$ = *posterior* pdf for μ, given the results of this expt
- $\mathcal{L}(x_0|\mu)$ = Likelihood function of μ from the experiment
- $p(\mu)$ = *prior* pdf for μ, before applying the results of this expt

Note! There is one (and only one) probability density in μ on each side of the eqn, consistent with $\mathcal{L}(x_0|\mu)$ *not* being a density in μ.

Quick intro to “Bayesian” analysis

All equations up until now are true for *any* definition of probability P that obeys the axioms, including frequentist P, as long as the probabilities exist (for example if μ is sampled from an ensemble with known “prior” pdf).

The word “Bayesian” refers *not* to these equations, but to the choice of definition of P as *personal subjective degree of belief*.

Bayesian P applies to hypotheses and constants of nature (frequentist P does not), so many Bayesian-only applications.

Since Bayesian estimation *requires* a prior pdf, big issues in Bayesian estimation include:

- **What prior pdf to use, and how sensitive is the result?**
- **How to interpret posterior probability if the prior pdf is not Your personal subjective belief?**

Frequentist tools can be highly relevant to both questions!

Use of Bayesian Posterior pdf $p(\mu|x_0)$

Point estimation: **Some Bayesians use the posterior mode (aka maximum posterior density) as the point estimate of μ (though metric-dependent), others say point estimation is misguided.**

Since the Jacobian moves the mode around under change of parameter (say from lifetime τ to decay rate $\Gamma=1/\tau$), care must be used to interpret it. (Posterior median can be used in 1D.)

Use of Bayesian Posterior pdf $p(\mu|x_0)$

Point estimation: **Some Bayesians use the posterior mode (aka maximum posterior density) as the point estimate of μ (though metric-dependent), others say point estimation is misguided.**

Since the Jacobian moves the mode around under change of parameter (say from lifetime τ to decay rate $\Gamma=1/\tau$), care must be used to interpret it. (Posterior median can be used in 1D.)

Interval estimation: ***Credibility*** **of μ being in any interval $[\mu_1,\mu_2]$ can be calculated by integrating $p(\mu|x_0)$ over the interval.**

Use of Bayesian Posterior pdf $p(\mu|x_0)$

Point estimation: **Some Bayesians use the posterior mode (aka maximum posterior density) as the point estimate of μ (though metric-dependent), others say point estimation is misguided.**

Since the Jacobian moves the mode around under change of parameter (say from lifetime τ to decay rate $\Gamma=1/\tau$), care must be used to interpret it. (Posterior median can be used in 1D.)

Interval estimation: ***Credibility*** **of μ being in any interval $[\mu_1,\mu_2]$ can be calculated by integrating $p(\mu|x_0)$ over the interval.**

Hypothesis testing: **Unlike frequentist statistics, testing credibility of whether or not μ equals a particular value μ_0 is *not* performed using by examining intervals.***

One starts over with Bayesian model selection (later topic).

***assuming regular pdf p. Dirac δ–ftns in p correspond to model selection, with its issues.**

Use of Bayesian Posterior pdf $p(\mu|x_0)$

Point estimation: **Some Bayesians use the posterior mode (aka maximum posterior density) as the point estimate of μ (though metric-dependent), others say point estimation is misguided.**

Since the Jacobian moves the mode around under change of parameter (say from lifetime τ to decay rate $\Gamma=1/\tau$), care must be used to interpret it. (Posterior median can be used in 1D.)

Interval estimation: ***Credibility*** **of μ being in any interval $[\mu_1,\mu_2]$ can be calculated by integrating $p(\mu|x_0)$ over the interval.**

Hypothesis testing: **Unlike frequentist statistics, testing credibility of whether or not μ equals a particular value μ_0 is *not* performed using by examining intervals.***

One starts over with Bayesian model selection (later topic).

Decision making: **All *Decisions* about μ require not only $p(\mu|x_0)$ but also further input: the utility function.**

***assuming regular pdf p. Dirac δ–ftns in p correspond to model selection, with its issues.**

Can “subjective” be taken out of “degree of belief”?

There are compelling arguments (Savage, De Finetti et al.) that Bayesian reasoning with *personal subjective* P is the uniquely “coherent” way (with technical definition of coherent) of updating *personal* beliefs upon obtaining new data.

The huge question is: can the Bayesian formalism be used by scientists to report the results of their experiments in an “objective” way (however one defines “objective”), and does any of the glow of coherence remain when subjective P is replaced by something else?

An idea vigorously pursued by physicist Harold Jeffreys in mid-20th century:

Can one define a prior* p(μ) that *contains as little information as possible?

"Uniform Prior" Requires a Choice of Metric

The really *really* thoughtless idea*, recognized by Jeffreys as such, but dismayingly common in HEP:

Just choose prior p(μ) uniform in whatever metric you happen to be using! (UGH!)

Recall that the probability integral transform *always* allows one to find a metric in which p is uniform (for continuous μ).

Thus the question "What is the prior pdf p(μ)?" is equivalent to the question, "For what function y(μ) is p(y) uniform?"

The choice y(u) = u needs to be justified.

(It does not represent ignorance!)

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

Jeffreys's Choice of Metric for Uniform Prior

Harold Jeffreys answered the question using a prior uniform in a metric related to the Fisher information, calculated from curvature of the log-likelihood function averaged over sample space. Jeffreys priors:

- **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, $0 \leq \rho \leq 1$: $p(\rho) = \rho^{-1/2}(1 - \rho)^{-1/2} = \text{Beta}(1/2,1/2)$**

Jeffreys's Choice of Metric for Uniform Prior (cont.)

If parameter μ is changed to f(μ), the recipe for obtaining Jeffreys prior for f(μ) yields a different-looking prior that corresponds to the *same choice of uniform metric*.

So resulting resulting pdf for μ is changed to f(μ) is correctly related by Jacobian, and *probabilities* (integrals of pdfs over equivalent endpoints) using Jeffreys prior are invariant under choices of different parameterizations.

What to call such Non-Subjective Priors?

- **“*Non*informative priors”? “Uninformative priors”? Traditional among statisticians, even though *they know it is misnomer*. (You should too!)**
- **“Vague priors”? “Ignorance priors”? “Default priors”?**
- **“Reference priors”? (Unfortunately also refers to a specific recipe of Bernardo)**
- **“Objective priors”? Despite the highly questionable use of the word, Jeffreys prior and its generalization by Bernardo and Berger are now widely referred to as “objective priors”.**
- **Kass and Wasserman J. Amer. Stat. Assn. 91 1343 (1996) give the best (neutral) name in my opinion: *Priors selected by “formal rules”.***
  - ***Required reading for anyone using Bayesian methods!***

Whatever the name, prior in one metric determines it in all other metrics: be careful in choice of metric in which it is uniform!

Jeffreys's Prior (cont.)

For one-parameter models, "Jeffreys's Prior" is the most common choice among statisticians for a "default" prior -- so common that statisticians are referring to Jeffreys's prior when they say "flat prior".

The obvious generalization to multi-parameter models is problematic, so alternatives have been developed, notably by Bernardo (with Berger). In 1D, they provide a different rationale for Jeffreys's prior, namely the prior that leads to a posterior pdf that is most dominated by the likelihood.

There are many subtleties. Beware!

Jeffreys's Prior (cont.)

A key point: priors such as Jeffreys's Prior are based on the likelihood function and thus inherently derived from the *measurement apparatus and procedure*, not from thinking about the parameter!

This may seem strange, but does give advantages, particularly for frequentist (!) coverage, as discussed briefly later.

Whatever you call non-subjective priors, they do *not* represent ignorance!

Dennis V. Lindley Stat. Sci 5 85 (1990), "the mistake is to think of them [Jeffreys priors or Bernardo/Berger's reference priors] as representing ignorance"

This Lindley quote is emphasized by Christian Robert, *The Bayesian Choice*, (2007) p. 29.

Jose Bernardo: "[With non-subjective priors,] The contribution of the data in constructing the posterior of interest should be "dominant". Note that this does not mean that a non-subjective prior is a mathematical description of "ignorance". Any prior reflects some form of knowledge."

Nonetheless, Berger (1985, p. 90) argues that Bayesian analysis with noninformative priors (older name for objective priors) such as Jeffreys and Barnardo/Berger "is the single most powerful method of statistical analysis, in the sense of being the *ad hoc* method most likely to yield a sensible answer for a given investment of effort".

Priors in high dimensions

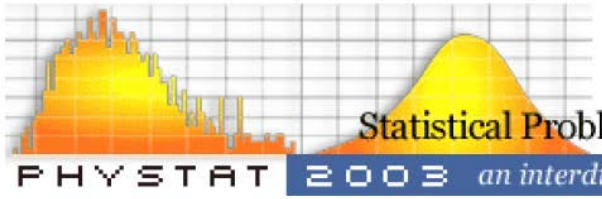
Is there a sort of informational phase space which can lead us to a sort of probability Dalitz plot? I.e., the desire is that structure in the posterior pdf represents information in the data, *not* the effect of Jacobians. *Notoriously hard problem!*

Be careful: Uniform priors push the probability away from the origin to the boundary! (Volume element in 3D goes as $r^2 dr$.)

State of the art for "objective" priors may be "reference priors" of Bernardo and Berger, but multi-D tools have been lacking.

Subjective priors also very difficult to construct in high dimensions: human intuition is poor.

- **Subjective Bayesian Michael Goldstein: "meaningful prior specification of beliefs in probabilistic form over very large possibility spaces is very difficult and may lead to a lot of arbitrariness in the specification".**



Statistical Problems in Particle Physics, Astrophysics, and Cosmology

an interdisciplinary gathering of statisticians and physicists in these fields

Bayesians, Frequentists, and Physicists

Bradley Efron

Department of Statistics and Department of Health Research and Policy, Stanford University, Stanford, CA 94305, USA

"Perhaps the most important general lesson is that the facile use of what appear to be uninformative priors is a dangerous practice in high dimensions."

Sir David Cox at PhyStat-LHC 2007

Five faces of Bayesian statistics

- empirical Bayes; number of similar parameters with a frequency distribution
- neutral (reference) priors: Laplace, Jeffreys, Jaynes, Berger and Bernardo
- information-inserting priors (evidence-based)
- personalistic priors
- technical device for generating frequentist inference

Currently in HEP, the main application is #5.
In particular, for *upper limits* we use uniform prior for Poisson mean for frequentist reasons (See my AJP paper, http://aapt.scitation.org/doi/10.1119/1.17901 .)
Unfortunately some in HEP have also added a 6th:

Cox's list, as I have seen it augmented in HEP

Six faces of Bayesian statistics

- empirical Bayes; number of similar parameters with a frequency distribution
- neutral (reference) priors: Laplace, Jeffreys, Jaynes, Berger and Bernardo
- information-inserting priors (evidence-based)
- personalistic priors
- technical device for generating frequentist inference
- Priors uniform in arbitrary variables, or in "the parameter of interest" (UGH!). This has no justification in modern subjective or objective Bayesian theory.

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.

I think too little emphasis to this is given by Bayesian advocates in HEP.

See backup slides for quotes from objective and subjective Bayesians.

Bayesian Must-Read for HEP/Astro/Cosmo (incl discussion!)

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).

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)

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

Memorable Quotes Therein from Jim Berger

"The Case for Objective Bayesian Analysis," Bayesian Analysis 1. See pp. 397, 459.

I call such analyses *pseudo-Bayes* because, while they utilize Bayesian machinery, they do not carry with them any of the guarantees of good performance that come with either true subjective analysis (with a very extensive elicitation effort) or (well-studied) objective Bayesian analysis. I will briefly discuss the problem with each of these pseudo-Bayes procedures.

***

I shall resist the temptation of saying more, because model selection is a can of worms for both objectivists and subjectivists.

Memorable Quotes Therein from Jim Berger

"The Case for Objective Bayesian Analysis," Bayesian Analysis 1. See pp. 397, 459.

I call such analyses *pseudo-Bayes* because, while they utilize Bayesian machinery, they do not carry with them any of the guarantees of good performance that come with either true subjective analysis (with a very extensive elicitation effort) or (well-studied) objective Bayesian analysis. I will briefly discuss the problem with each of these pseudo-Bayes procedures.

***

I shall resist the temptation of saying more, because model selection is a can of worms for both objectivists and subjectivists.

Pseudo-Bayes analyses pop up from time to time in HEP. Flat priors, etc. (The worst are in Model Selection.)
See my Comment at https://arxiv.org/abs/0807.1330 and my "pseudo-Bayes detection" slides 62-68 at Tokyo PhyStat-nu, http://indico.ipmu.jp/indico/event/82/session/9/contribution/16/material/slides/0.pdf .

An excellent discussion by Harrison Prosper is in Ch. 12 of Data Analysis in High Energy Physics, Ed. By O. Behnke et al.

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)**

Rather:

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

Statements are made about probabilities in *ensembles* of intervals (fraction containing 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.

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, described below, or by recipes of any origin that yield good approximations thereof.

Basic notions of confidence intervals

Basic idea in two sentences:

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".

Basic notions of confidence intervals

Basic idea in two sentences:

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".

To be well-defined, the first point needs to be supplemented:

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

2) Need also to specify what *fraction* of values of x are not considered extreme.

Basic notions of confidence intervals (cont.)

Some common ordering choices in 1D (when p(x|μ) 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 *central* quantiles of p(x|μ), with the quantiles shorter in x (least integrated probability of x) containing higher-ranked x, with lower-ranked x added as the central quantile gets longer. Gives *central* confidence intervals for μ.**

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

(4^{th} ordering, likelihood ratio used by F-C, still to come.)

Basic notions of confidence intervals (cont.)

Given model p(x|μ) 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, "extreme" fraction.

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.

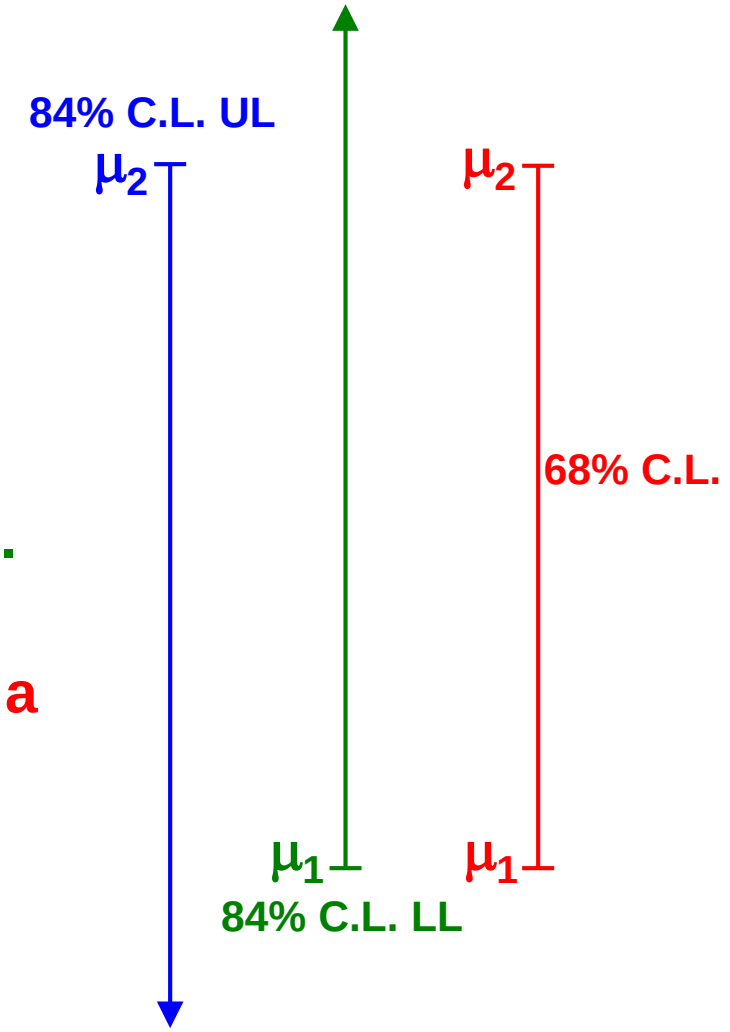
E.g.:

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 μ for which x_0 is in the central 68% quantile of x values. It is a 68% C.L. *central* confidence interval (!)

Examples follow, first with continuous x, then with discrete x.

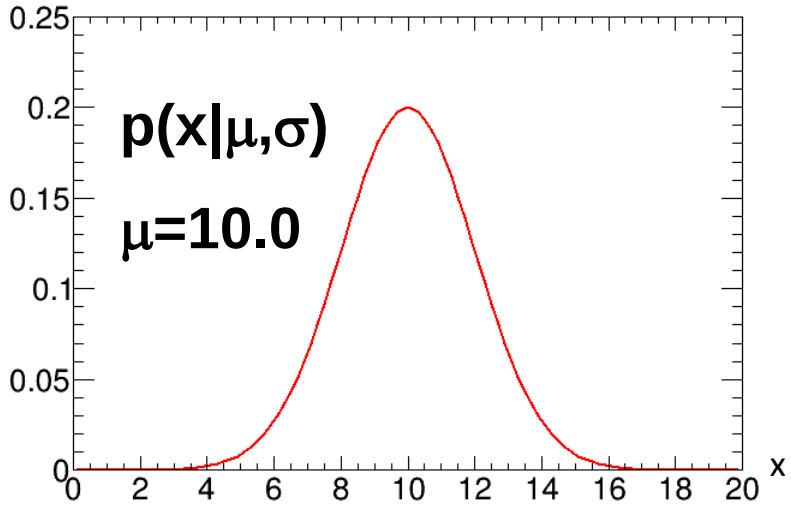


Gaussian pdf p(x|μ,σ) with σ a function of μ: σ = 0.2 μ

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

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

Plot of p(x|μ,σ) with μ=10.0, σ = 2.0 :



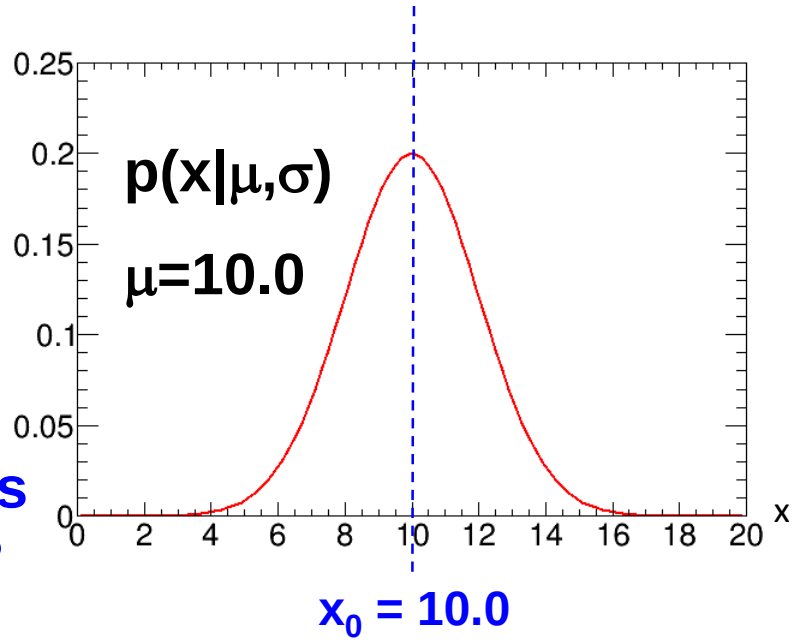
μ

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

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

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

Plot of p(x|μ,σ) with μ=10.0, σ = 2.0 :



$x_0 = 10.0$

Suppose μ is unknown, and x_0 = 10.0 is observed. What can one say about μ ?

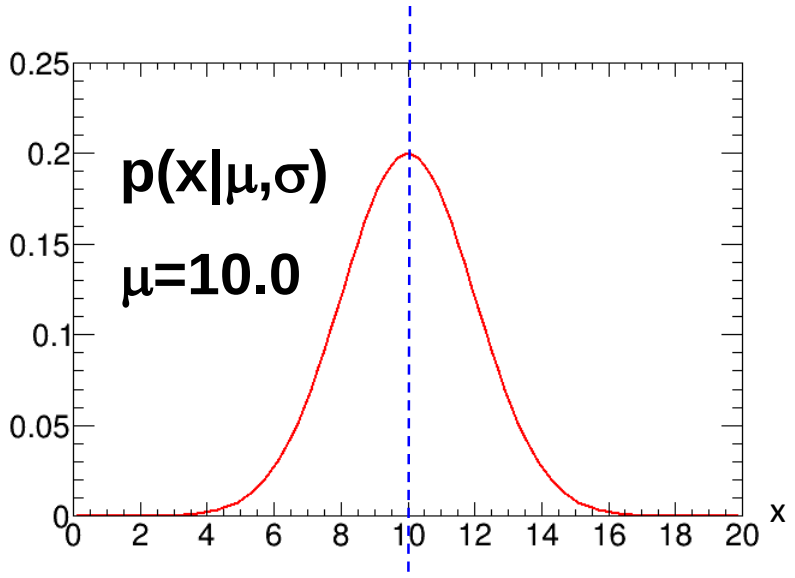
μ

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

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

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

Plot of p(x|μ,σ) with μ=10.0, σ = 2.0 :

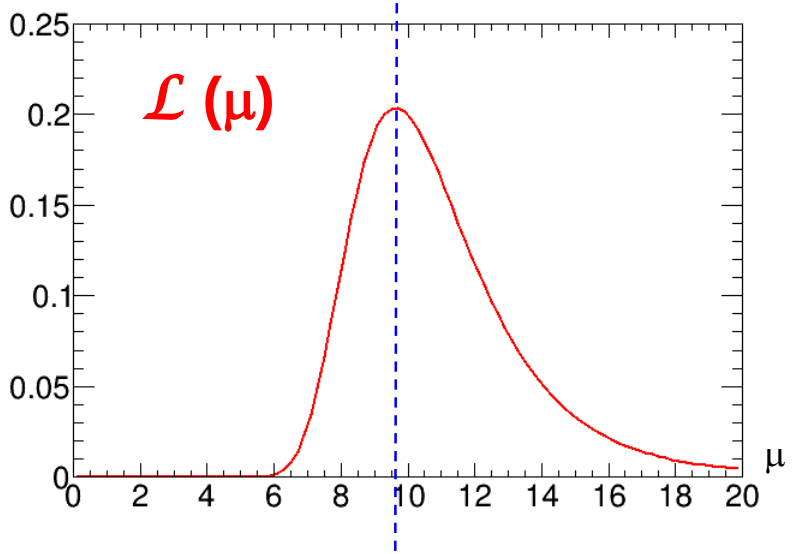


Suppose x_0 = 10.0 is observed.

$$\mathcal{L}(\mu) = \frac{1}{\sqrt{2\pi(0.2\mu)^2}}\, e^{-(x_0-\mu)^2/2(0.2\mu)^2}$$

x_0 = 10.0

Plot of $\mathcal{L}(\mu)$ for observed x_0 = 10. :



μ_{ML} = 9.63

What is *confidence interval* for μ?

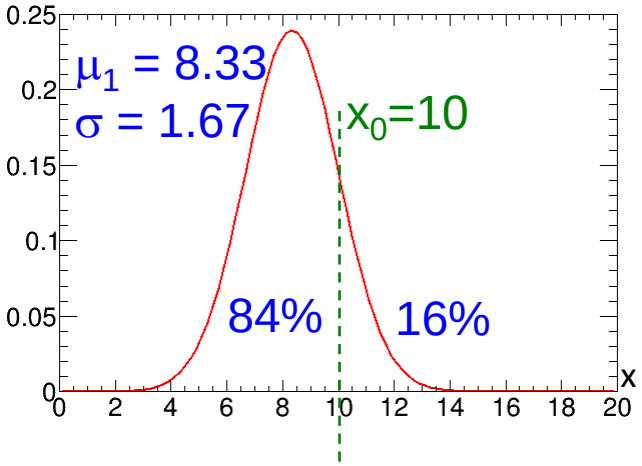
Gaussian pdf $p(x|\mu,\sigma)$ with σ a function of μ: $\sigma = 0.2\,\mu$ Observed $x_0 = 10.0$.

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.

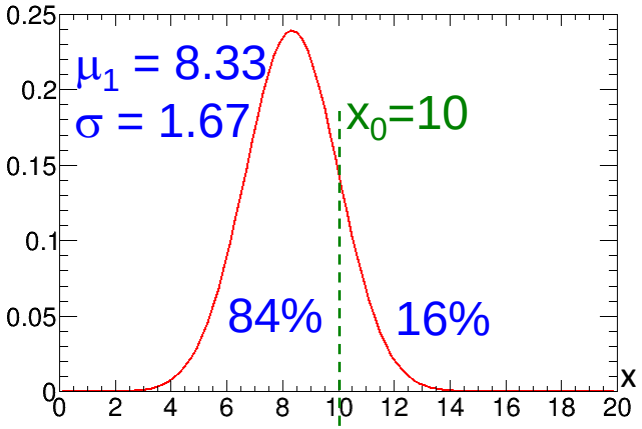


Gaussian pdf $p(x|\mu,\sigma)$ with σ a function of μ: $\sigma = 0.2\,\mu$ Observed $x_0 = 10.0$.

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

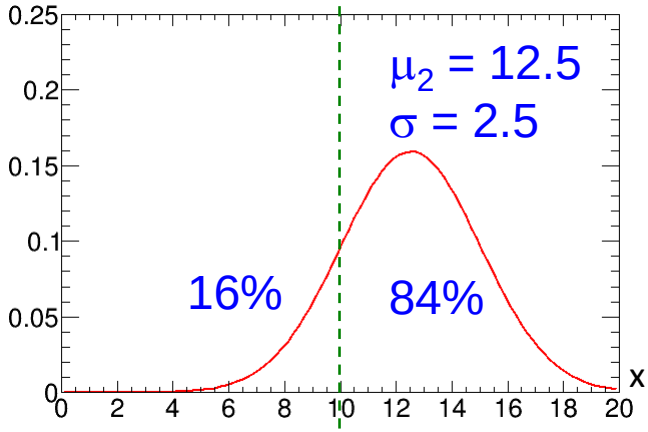
μ_1 is 84% C.L. *lower* limit for μ.



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 ("Wald-like") 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 the Wald-like 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

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

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

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

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

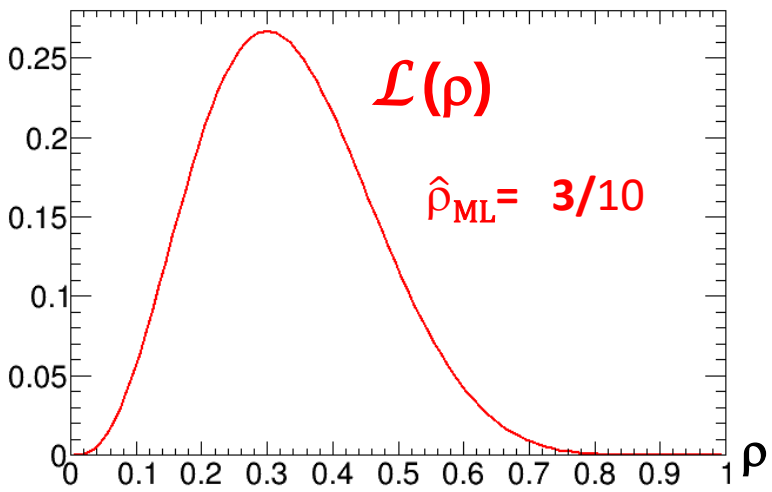
With observed successes n_{on}, the M.L. point estimate $\hat{\rho}$ of ρ is (next slide)

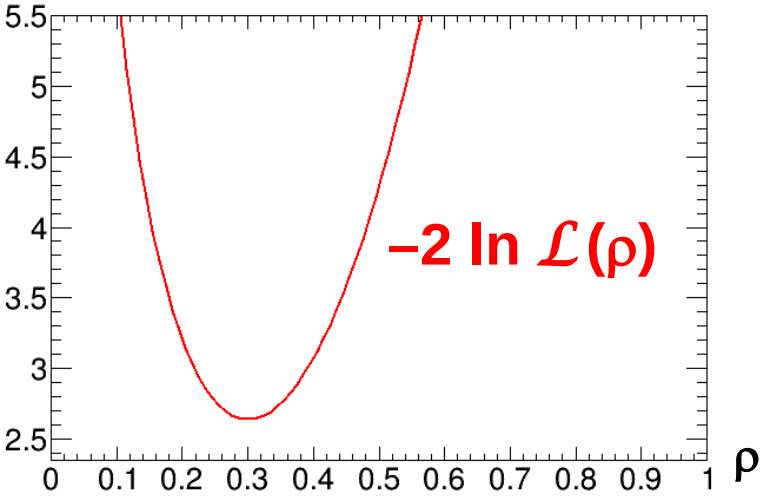
$$\hat{\rho} = n_{on} / n_{tot}\ .$$

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 $[\rho_1,\rho_2]$.

Recall shortcut above for central intervals:

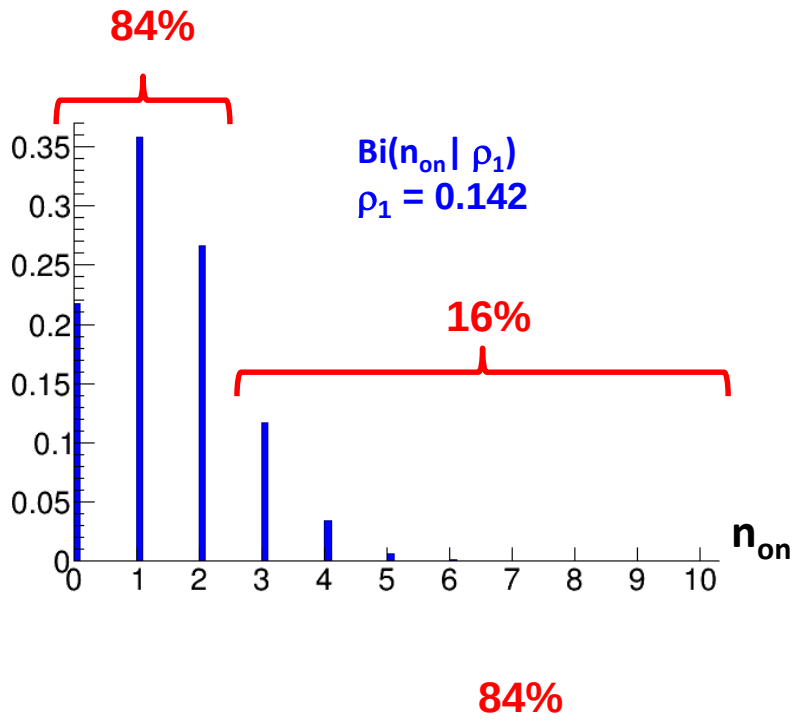
Find *lower limit* ρ_1 with C.L. = 1 – (1 – 68%)/2. = 84%

Find *upper limit* ρ_2 with C.L. = 84%

Then $[\rho_1, \rho_2]$ is 68% C.L. central confidence interval

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

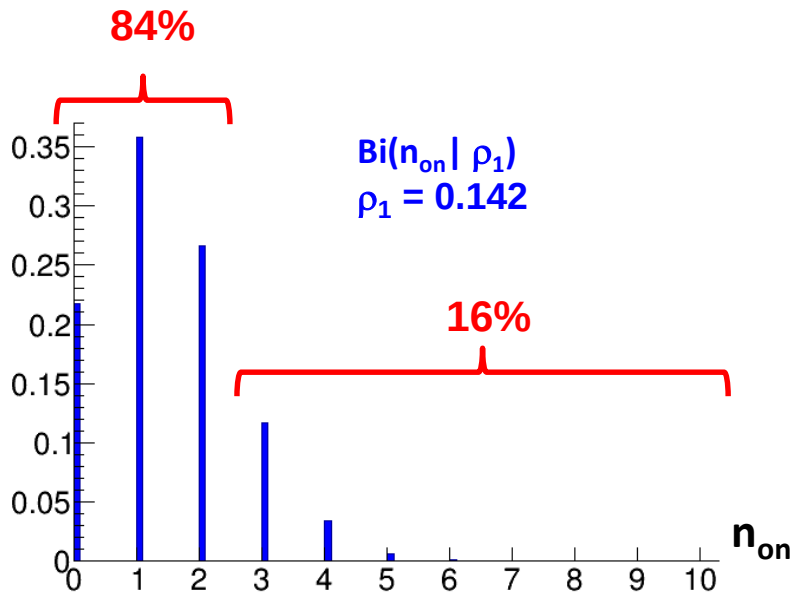
$n_{on} = 3$, $n_{tot}=10$.
Find ρ_1 such that
$Bi(n_{on} < 3 \mid \rho_1) = 84\%$
$Bi(n_{on} \geq 3 \mid \rho_1) = 16\%$
(lower limit at 84% C.L.)
Solve: $\rho_1 = 0.142$

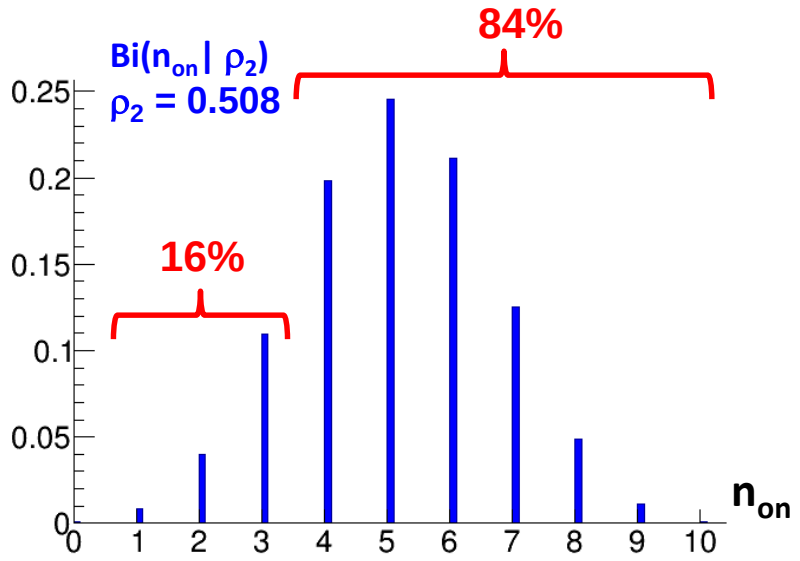


$n_{on} = 3$, $n_{tot}=10$.
Find ρ_1 such that
$Bi(n_{on} < 3 \mid \rho_1) = 84\%$
$Bi(n_{on} \geq 3 \mid \rho_1) = 16\%$
(lower limit at 84% C.L.)
Solve: $\rho_1 = 0.142$

And find ρ_2 such that
$Bi(n_{on} > 3 \mid \rho_2) = 84\%$
$Bi(n_{on} \leq 3 \mid \rho_2) = 16\%$
(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)*





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

In HEP, such Clopper-Pearson intervals are the standard for a binomial parameter

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 discreteness of x leads to an issue: C-P is criticized by some as "wastefully conservative" – see CHT 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)

Gaussian approximation for binomial conf. int.

As above, n_{on} has mean $n_{tot}\,\rho$ and rms deviation $\sqrt{n_{tot}\,\rho\,(1-\rho)}$.
One can 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 to 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*. (See CHT paper.)

Avoid the Wald interval – no reason to use it

This "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 when $n_{on} = 0$ (or $n_{on} = n_{tot}$), this gives $\hat{\sigma} = 0$.

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...)

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.* Cousins, Hyme, and Tucker, http://arxiv.org/abs/0905.3831 .
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 background exists.

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

For decades, issues with upper limits and central confidence intervals have been discussed in prototype problems in HEP:

1. **Gaussian measurement resolution near a physical boundary (e.g. neutrino mass-squared is non-negative)**
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, I mostly stick to frequentist confidence intervals in this situation.

Beyond upper/lower limits and *central* confidence intervals

More general choices for ordering x in p(x|μ):

- **For each μ, order x_0 using *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. Recall that likelihood *ratios* as in F-C are independent of metric in x since Jacobian cancels.

Beyond upper/lower limits and *central* confidence intervals

More general choices for ordering x in p(x|μ):

- **For each μ, order x_0 using *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. Recall that likelihood *ratios* as in F-C are independent of metric in x since Jacobian cancels.

Note:

Ordering x by the probability *density* p(x|μ) is *not* recommended! Recall that change of metric from x to y(x) leads to Jacobian |dy/dx| in p(y|μ) = p(x|μ) / |dy/dx|. So ordering by p(y|μ) is different than ordering by p(x|μ) and all that follows depends on arbitrary choice of metric.

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 specific interval you report" – 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 $p(x \in [x_1,x_2] \mid \mu) = \text{C.L.} = 1 - \alpha$. ("Ordering principle" for x is used to well-define.)

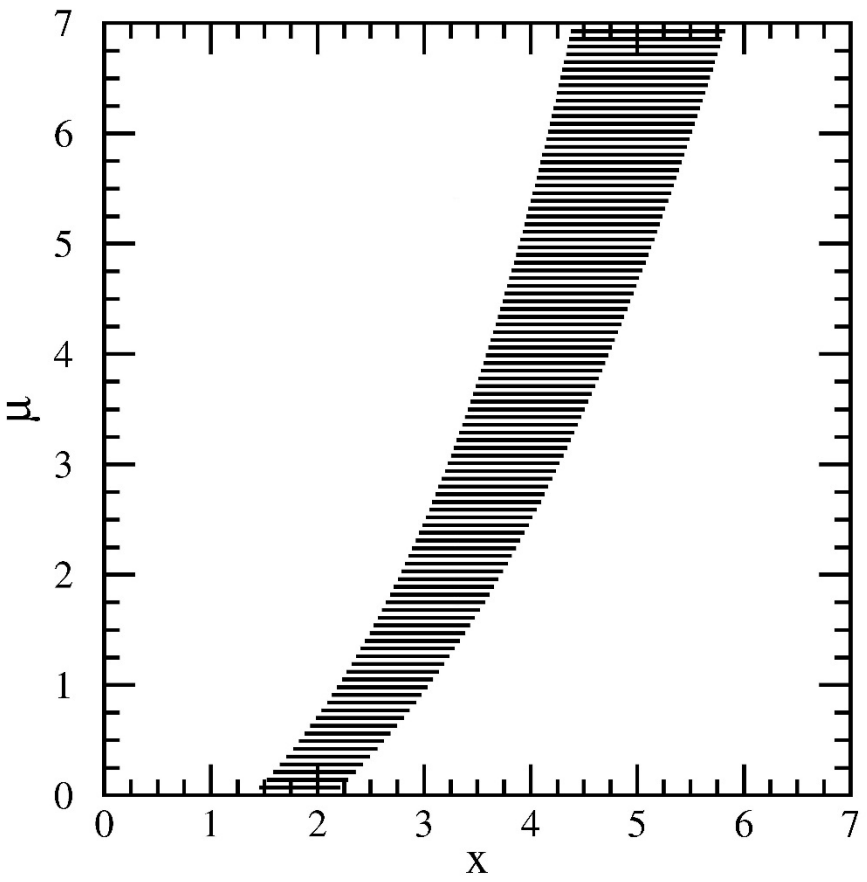


Figure from G. Feldman, R Cousins, Phys Rev D57 3873 (1998)

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 $p(x \in [x_1,x_2] \mid \mu) = \text{C.L.} = 1 - \alpha$. ("Ordering principle" for x is used to well-define.)

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

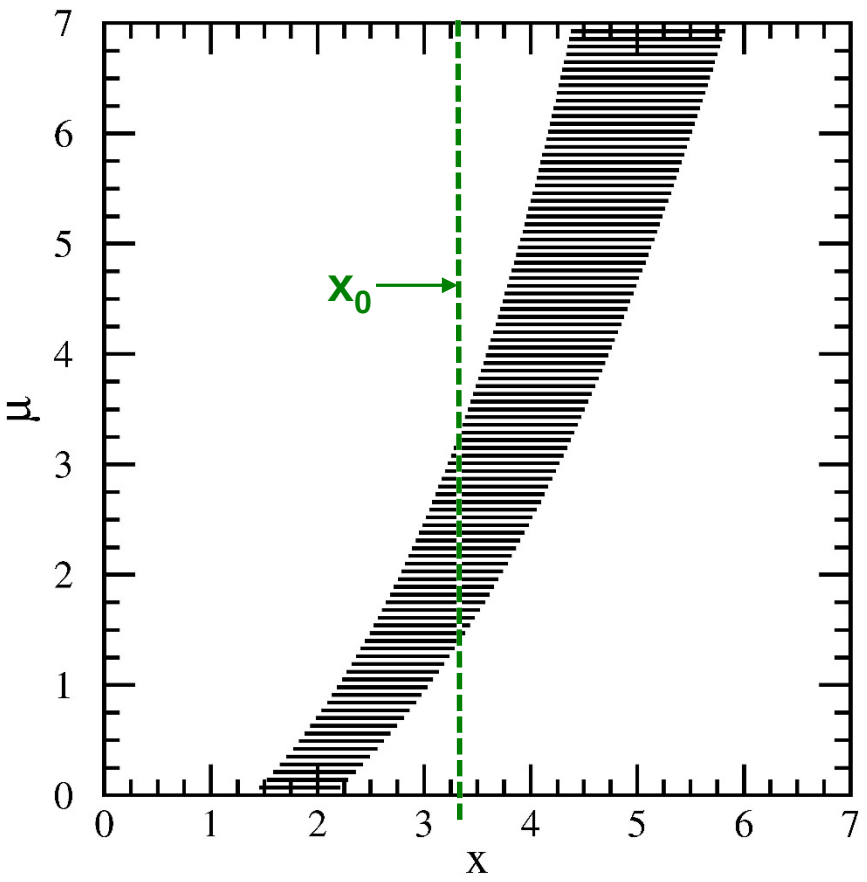
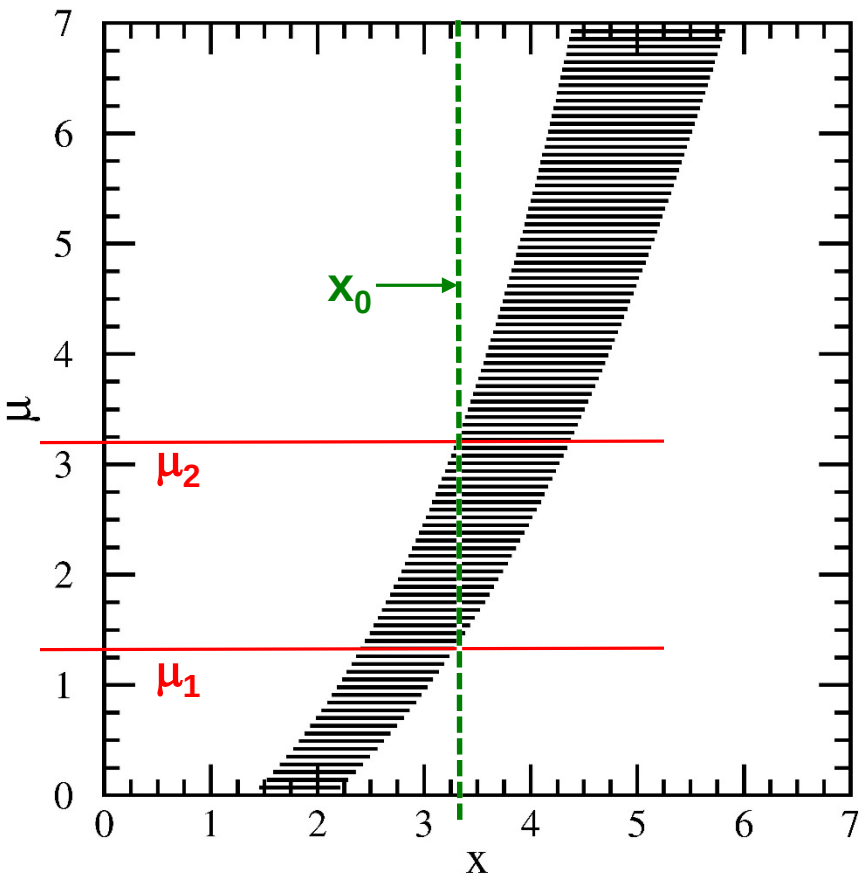


Figure from G. Feldman, R Cousins, Phys Rev D57 3873 (1998)

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. $= 1 - \alpha$. ("Ordering principle" for x is used to well-define.)

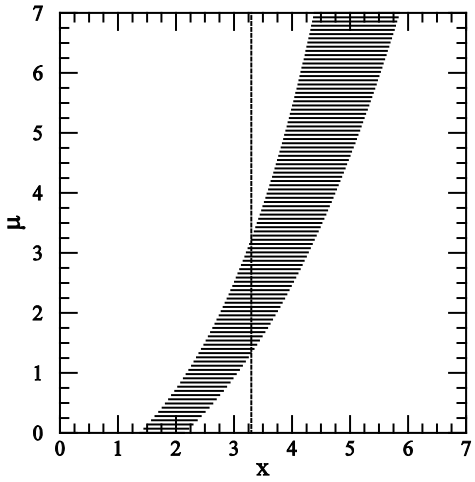
Upon observing x, obtaining the value x_0, one draws the vertical line through x_0.



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

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

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



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 $\mathcal{L}(\mu)$).

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*.

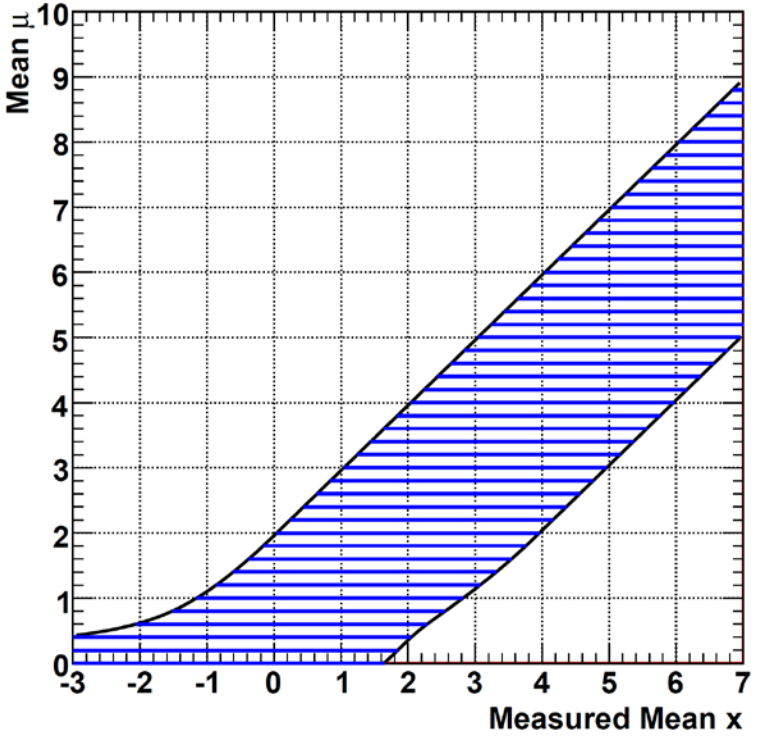
I.e., for mass μ <0, p(x|μ) does not exist: You would not know how to simulate the physics of detector response for *mass* < 0.

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



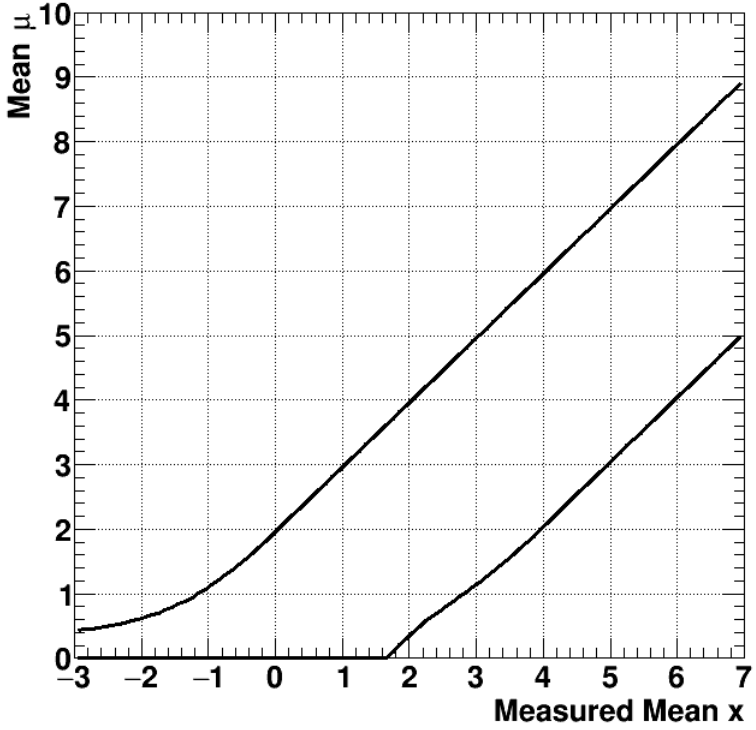
“Confidence Belt”

From the earliest days (as in 1934 Clopper-Pearson example in backup), the horizontal line segments are suppressed and only the envelope (black curves in figure), whose interior is called a *confidence belt*, is typically plotted.

Area inside black curves is“confidence belt”

I added the line segments for demonstrating the construction the F-C paper after reading Neyman’s 1937 paper, which had 2D acceptance regions shown. This practice is fortunately spreading.

Earlier works have statements that I found cryptic, such as “Notice that the confidence belt is *constructed horizontally* but *read vertically*.”



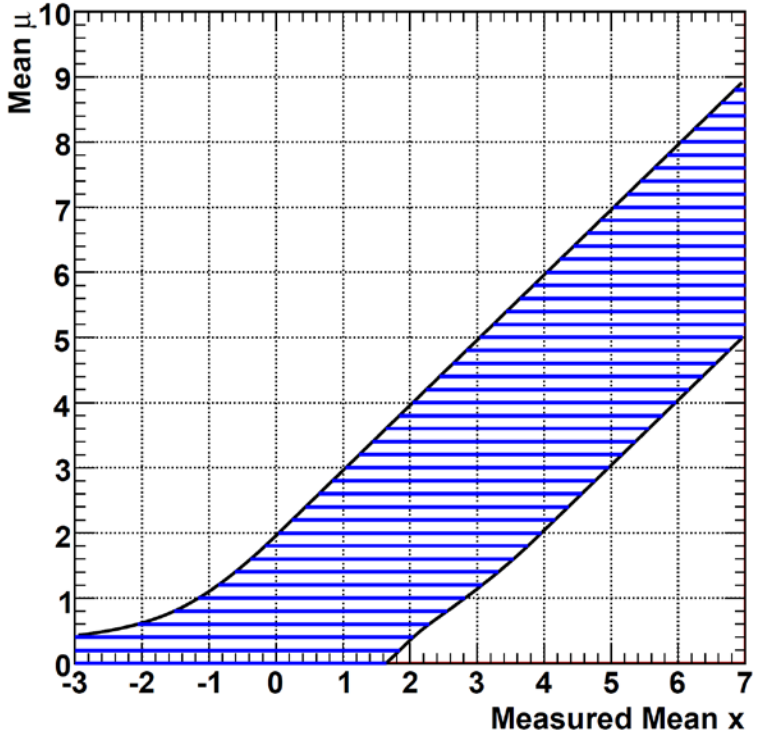
“Confidence Belt”

From the earliest days (as in 1934 Clopper-Pearson example in backup), the horizontal line segments are suppressed and only the envelope (black curves in figure), whose interior is called a *confidence belt*, is typically plotted.

Area inside black curves is“confidence belt”

I added the line segments for demonstrating the construction the F-C paper after reading Neyman’s 1937 paper, which had 2D acceptance regions shown. This practice is fortunately spreading.

Earlier works have statements that I found cryptic, such as “Notice that the confidence belt is *constructed horizontally* but *read vertically*.”



Confidence Intervals and Coverage

Recall: how is a *vector* defined in abstract math class?

Confidence Intervals and Coverage

Recall: how is a *vector* defined in abstract math class?

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.)

Confidence Intervals and Coverage

Recall: how is a *vector* defined in abstract math class?

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 (cont.)

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. (Definition of coverage)

In this (frequentist) equation, μ_t is *fixed* and unknown. 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.

* For μ_t , the probability that x_0 is in its acceptance region is C.L., by construction. For those x_0's, the vertical line will intercept μ_t's acceptance region, and so μ_t be will be put into the confidence interval.

Confidence Intervals and Coverage (cont.)

$P(\mu_t \in [\mu_1, \mu_2]) = \text{C.L.} = 1 - \alpha$. **(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!

Confidence Intervals and Coverage (cont.)

$P(\mu_t \in [\mu_1, \mu_2]) = \text{C.L.} = 1 - \alpha$. **(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!

A lot of confusion might have been avoided if Neyman had chosen the names "*coverage intervals*" and "*coverage level*"!

Maybe we can have a summit meeting treaty where frequentists stop saying "confidence" and Bayesians stop saying "noninformative"!

Confidence Intervals and Coverage (cont.)

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 C.L. least extreme values to be observed.

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

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

This appeared in early days (before Neyman's 1937 paper giving comprehensive discussion of the construction) when the concept of what we now call coverage was discussed from differing points of view of both Fisher and Neyman, with different names and arguments for justifications.

In retrospect, this paper (citing Fisher, Neyman, and Neyman's students) corresponds exactly to a Neyman construction of central confidence intervals. See backup for details.

As I have discussed, for central intervals the answers are the same as those obtained from upper and lower limits.

The same method was applied to confidence intervals for *Poisson mean* by Garwood in his 1934 thesis, and published in 1936.
Also standard in HEP when no background!
Controversy when background – see PDG RRP.

Classical Hypothesis Testing

At this point, we set aside confidence intervals for the moment and consider from the beginning the nominally different topic of hypothesis testing.

In fact we will soon find that in frequentist statistics, certain hypothesis tests will take us immediately back to confidence intervals. But first we consider the more general framework.

Frequentist hypothesis testing, often called "classical" hypothesis testing, was developed by R.A. Fisher in unfriendly competition with J. Neyman and E. Pearson. Modern testing has a mix of ideas from both.

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).

Then $p(x|\mu)$ is different for H_0 and H_1 , either because parameter μ is different, or p() itself is different.

Classical Hypothesis Testing (cont.)

For null hypothesis H_0, 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 by construction,

α = probability (with x generated according to H_0) of rejecting H_0 when it is true, i.e., false discovery claim (Type I error)

Classical Hypothesis Testing (cont.)

For null hypothesis H_0, 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). By construction:

α = probability (with x generated according to H_0) of rejecting H_0 when it is true, i.e., false discovery claim (Type I error)

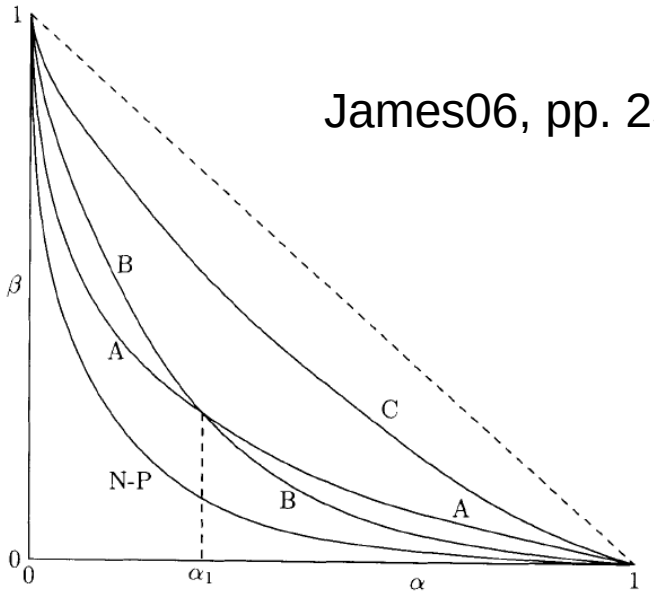
To quantity the performance of this test if H_1 is true, we define:

β = probability (with x generated according to H_1) of accepting H_0 when it is false, i.e., not claiming a discovery when there is one (Type II error)

Note: if alternative H_1 is not specified, then β is not known and optimality cannot be well-defined. The test is then called a *goodness-of-fit* test of H_0. See backup slides.

Classical Hypothesis Testing (cont.)

There is tradeoff between Type I and Type 2 errors.
Competing analysis methods can be compared by looking at graphs of β vs α at various μ, and at graphs of 1–β vs μ at various α (power function).



James06, pp. 258, 262

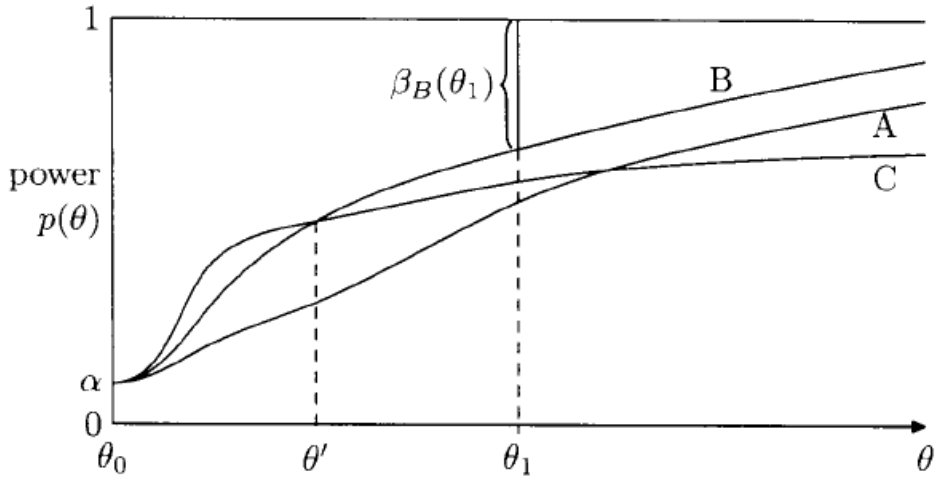
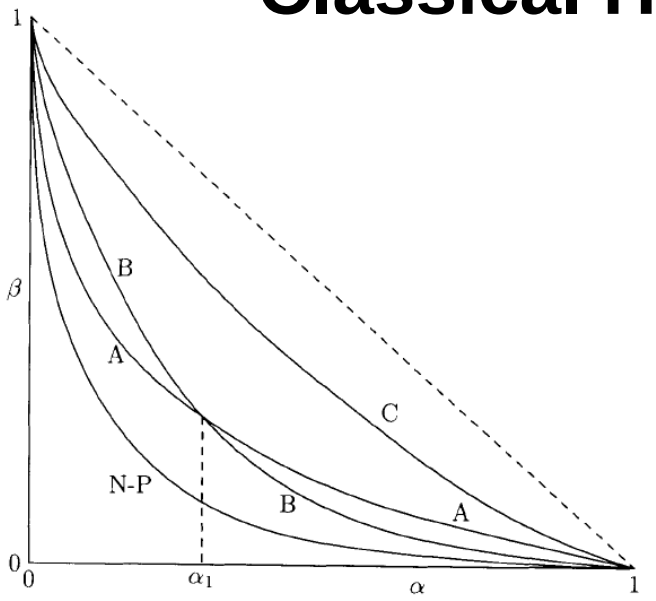


Fig. 10.3. Power functions of tests A, B, and C at significance level α. Of these three tests, B is the best for $\theta > \theta'$. For smaller values of θ, C is better.

Similar to comparing b-tagging efficiency for signal and background. Equivalent to ROC curve.

See my writeup on arxiv for detailed example.

Classical Hypothesis Testing (cont.)



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.)

The N-P language of "accept" or "reject" H_0 should not be mistaken for a complete theory of decision-making:

***Decision* on whether to declare discovery requires 2 more inputs:**

1) **Prior belief in H_0 vs H_1. (Can affect choice of α)**
2) **Cost of Type I error (false discovery claim) vs cost of Type II error (missed discovery). (Can also affect choice of α)**

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

Classical Hypothesis Testing: Simple Hypotheses

In idealized cases, a hypothesis may have no floating (unfixed) parameters.
N-P called such hypotheses *simple*, in contrast to *composite* hypotheses that have unfixed parameters.

Examples in HEP where both H_0 and H_1 are *simple* are rare, but we do have a few examples where the quantity of interest is simple in both hypotheses, and the role of unfixed parameters does not spoil the "simplicity", e.g., H_0 vs H_1 being:

"jet originated from a quark" vs "jet originated from a gluon"

spin-1 vs spin-2 for a new resonance in $\mu^+\mu^-$

$J^P=0^+$ vs $J^P=0^-$ for the Higgs-like boson

Simple Hypotheses 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 x according to the *likelihood ratio* $\lambda = \mathcal{L}(x|H_0) / \mathcal{L}(x|H_1)$.

One finds cutoff $\lambda_{cut,\alpha}$ for that α and rejects H_0 if $\lambda \leq \lambda_{cut,\alpha}$.

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

Nested Hypothesis Testing

In contrast to two disjoint simple hypotheses, it is common in HEP for H_0 to be *nested* in H_1.
This happens when there is an undetermined parameter μ in H_1 , and H_0 corresponds to a particular parameter value μ_0 (e.g., zero, 1, or ∞). So consider:

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: null $\mu_0 = 0$, alternative $\mu > 0$

$H^0 \rightarrow \gamma\gamma$ *before* discovery of this decay, μ = signal strength:
null $\mu_0 = 0$, alternative $\mu > 0$

$H^0 \rightarrow \gamma\gamma$ *after* discovery of this decay:
null μ_0 = SM prediction, alternative any other $\mu \neq \mu_0$

Nested Hypothesis Testing (cont.)

H_0: $\mu = \mu_0$ (the "point null", or "sharp hypothesis") vs
H_1: $\mu \neq \mu_0$ (the "continuous alternative").

In classical/frequentist formalism (but *not* Bayesian formalism), the theory of such nested tests maps to that of confidence intervals!

Nested Hypothesis Testing (cont.)

H_0: $\mu = \mu_0$ (the "point null", or "sharp hypothesis") vs
H_1: $\mu \neq \mu_0$ (the "continuous alternative").

In classical/frequentist formalism (but *not* Bayesian formalism), the theory of such nested tests maps to that of confidence intervals!

Intuitive argument:
Having observed data x_0, suppose the 90% C.L. confidence interval for μ is $[\mu_1,\mu_2]$.
This contains all values of μ for which observed x_0 is ranked in the *least* extreme 90% of possible outcomes x according to $p(x|\mu)$ and the ordering principle in use.

Nested Hypothesis Testing: Duality with Intervals

H_0: $\mu = \mu_0$ (the "point null", or "sharp hypothesis") vs
H_1: $\mu \neq \mu_0$ (the "continuous alternative").

In classical/frequentist formalism (but *not* Bayesian formalism), the theory of these tests maps to that of confidence intervals!

Intuitive argument:
Having observed data x_0, suppose the 90% C.L. confidence interval for μ is $[\mu_1,\mu_2]$.
This contains all values of μ for which observed x_0 is ranked in the *least* extreme 90% of possible outcomes x according to $p(x|\mu)$ and the ordering principle in use.

Now suppose we wish to test H_0 vs H_1 at Type I error prob α = 10%. We reject H_0 if x_0 is ranked in the *most* extreme 10% of x according to $p(x|\mu)$ and the ordering principle in use.

Comparing the two procedures, we see:
Reject H_0 at α=10% iff μ_0 is in 90% C.L. confidence interval $[\mu_1,\mu_2]$.

Nested Hypothesis Testing Duality

***Given* an ordering:**

Test of $\mu=\mu_0$ vs $\mu\neq\mu_0$ at significance level α
$\leftrightarrow$ 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.,

$\alpha \leftrightarrow$ 1 – C.L.
Equal-tailed test $\leftrightarrow$ central confidence intervals
One-tailed tests $\leftrightarrow$ Upper/lower limits

Use of the duality is referred to as "inverting a test" to obtain confidence intervals, and vice versa.

Nested 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 a "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 parameters! →

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

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

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,

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

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*.**

This is numerically (if not philosophically) the same as definition used e.g. by Fisher and often taught: "*p-value* is probability under H_0 of obtaining x as extreme *or more extreme* than observed x_0."

Interpreting p-values and Z-values

It is crucial to realize that that value of α (0.023 in the example) was typically *not* specified in advance, so p-values do *not* correspond to Type I error probs of experiments reporting them.

In HEP, p-value 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.

Interpreting p-values and Z-values

It is crucial to realize that that value of α (0.023 in the example) was typically *not* specified in advance, so p-values do *not* correspond to Type I error probs of experiments reporting them.

In HEP, p-value 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.

***Although these lectures are not "statistics in practice",**
I mention ROOT commands for *one*-tailed conversions:

```
zvalue = -TMath::NormQuantile(pvalue)
pvalue = 0.5*TMath::Erfc(zvalue/sqrt(2.0))
```

(Thanks, Igor Volobouev) Note that p-value > 0.5 means Z-value < 0.

Interpreting p-values and Z-values (cont.)

Interpretation of p-values (and hence Z-values) is a long, contentious story – beware!

Widely bashed. I give some reasons why later.

I defend their use in HEP. See https://arxiv.org/abs/1310.3791.)

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 of "probability that H_0 is true" requires prior(s)!**

Please help educate press officers and journalists! (and physicists) !

Early CMS Higgs spin-parity test of 0^+ vs. 0^-

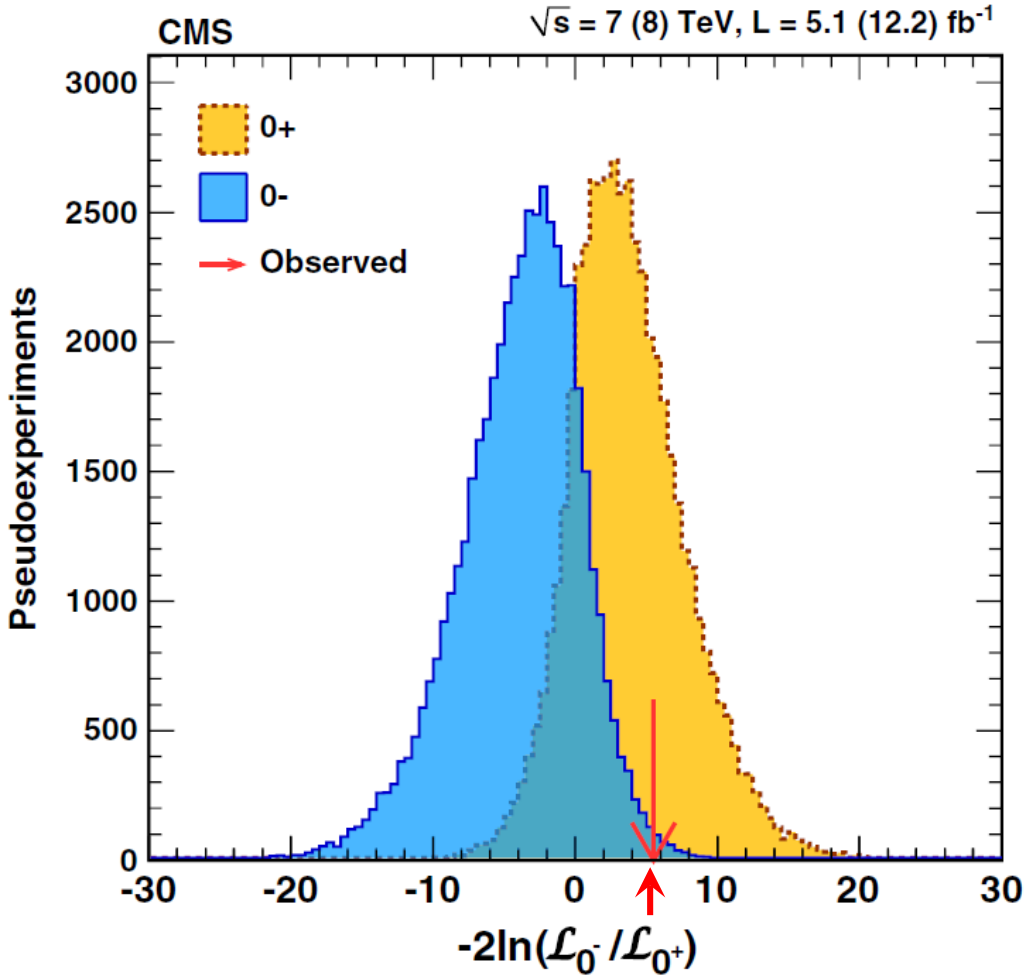


FIG. 3 (color online). Expected distribution of $-2\ln\mathcal{L}_{0^-}/\mathcal{L}_{0^+}$ under the pure pseudoscalar and pure scalar hypotheses (histograms). The arrow indicates the value determined from the observed data.

CMS, Phys. Rev. Lett. 110 (2013) 081803

Paper reported (fixing typo):

1) $-2\ln(\mathcal{L}_{0^-}/\mathcal{L}_{0^+})$ = 5.5 favoring 0^+

2) for H_0: 0^-, p-value = 0.0072

3) for H_0: 0^+, p-value = 0.7

4) CL_s = (0.0072)/(1–0.7) = 0.024, "a more conservative value for judging whether the observed data are compatible with 0^- "

N.B. See backup for figure and pointer to paper by Demortier and Lyons discussing two p-values in simple-vs-simple case.

Likelihood (Ratio) Intervals for 1 parameter

Recall: Likelihood $\mathcal{L}(\mu)$ is invariant under reparametrization from μ to $f(\mu)$: $\mathcal{L}(\mu) = \mathcal{L}(f(\mu))$.

So *likelihood ratios* $\mathcal{L}(\mu_1)/\mathcal{L}(\mu_2)$ and *log-likelihood differences* $\ln\mathcal{L}(\mu_1) - \ln\mathcal{L}(\mu_2)$ are also invariant.

After using maximum-likelihood method to obtain estimate $\hat{\mu}$ that maximizes either $\mathcal{L}(\mu)$ or $\mathcal{L}(f(\mu))$, one can obtain a likelihood interval $[\mu_1, \mu_2]$ as the union of all μ for which

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

Likelihood (Ratio) Intervals for 1 parameter

Recall: Likelihood $\mathcal{L}(\mu)$ is invariant under reparametrization from μ to f(μ): $\mathcal{L}(\mu) = \mathcal{L}(f(\mu))$.

So *likelihood ratios* $\mathcal{L}(\mu_1)/\mathcal{L}(\mu_2)$ and *log-likelihood differences* $\ln\mathcal{L}(\mu_1) - \ln\mathcal{L}(\mu_2)$ are also invariant.

After using maximum-likelihood method to obtain estimate $\hat{\mu}$ that maximizes either $\mathcal{L}(\mu)$ or $\mathcal{L}(f(\mu))$, one can obtain a likelihood interval $[\mu_1, \mu_2]$ as the union of all μ for which

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

As sample size increases (under important regularity conditions) this interval approaches a central confidence interval with C.L. corresponding to $\pm$ Z Gaussian standard deviations

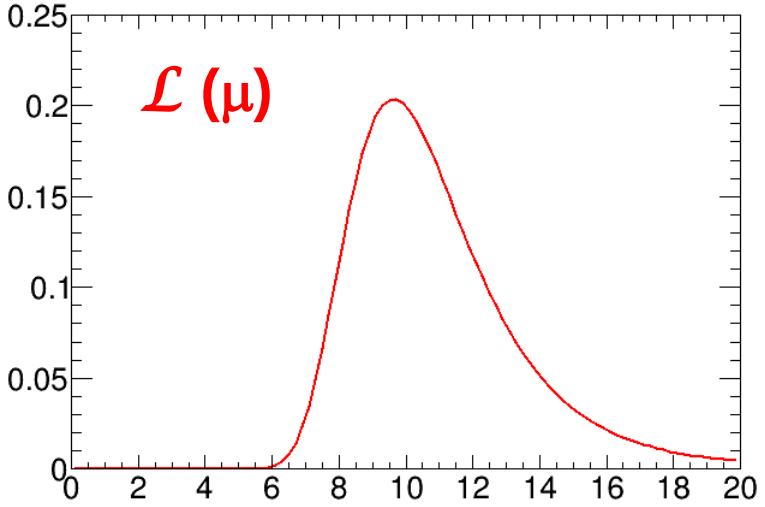
But! Regularity conditions, in particular requirement that $\hat{\mu}$ not be on the boundary, need to be carefully checked.
E.g., if $\mu \geq 0$ on physical grounds, then $\hat{\mu} = 0$ requires care.

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 $x_0 = 10.0$.

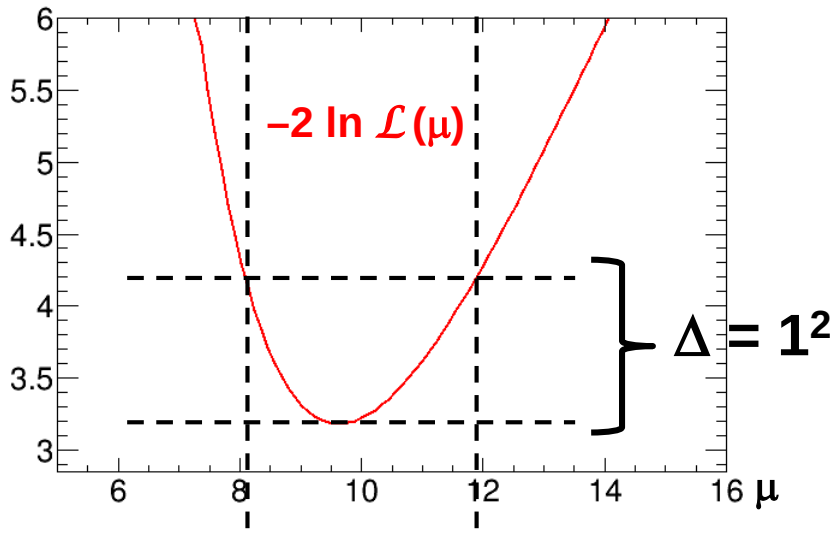
$\mu_{ML} = 9.63$



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].

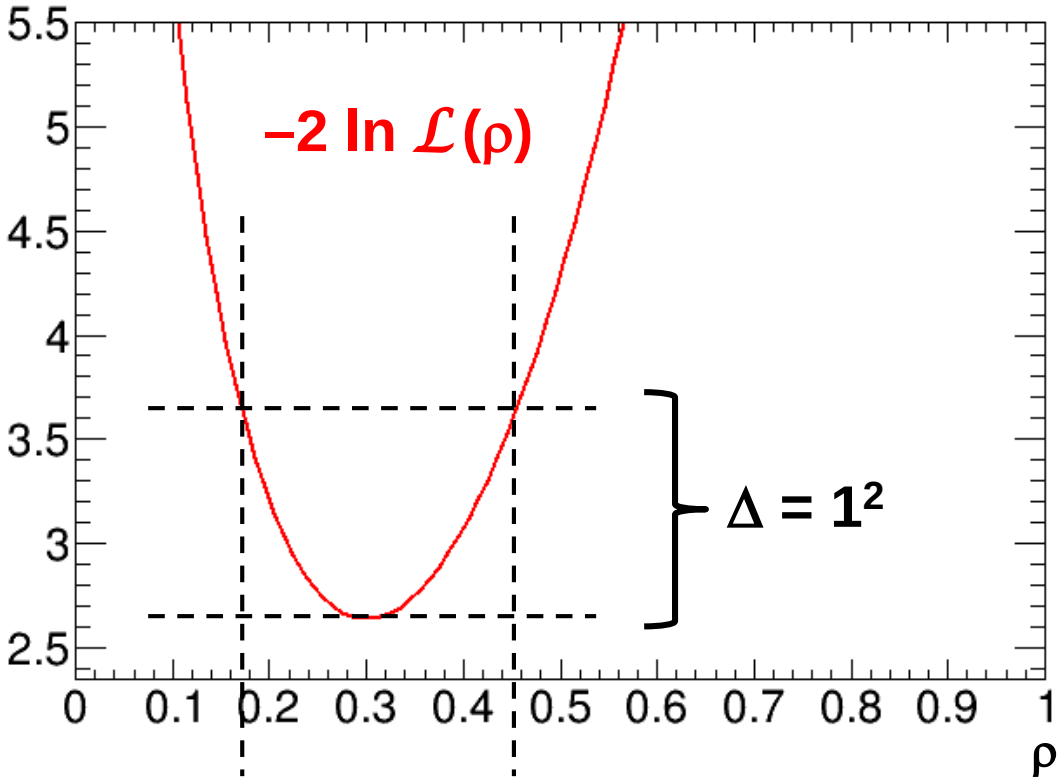


Binomial Likelihood-Ratio Interval example

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

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

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



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

Poisson Likelihood-Ratio Interval example

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

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

Recall maximum at $\mu = 3$.

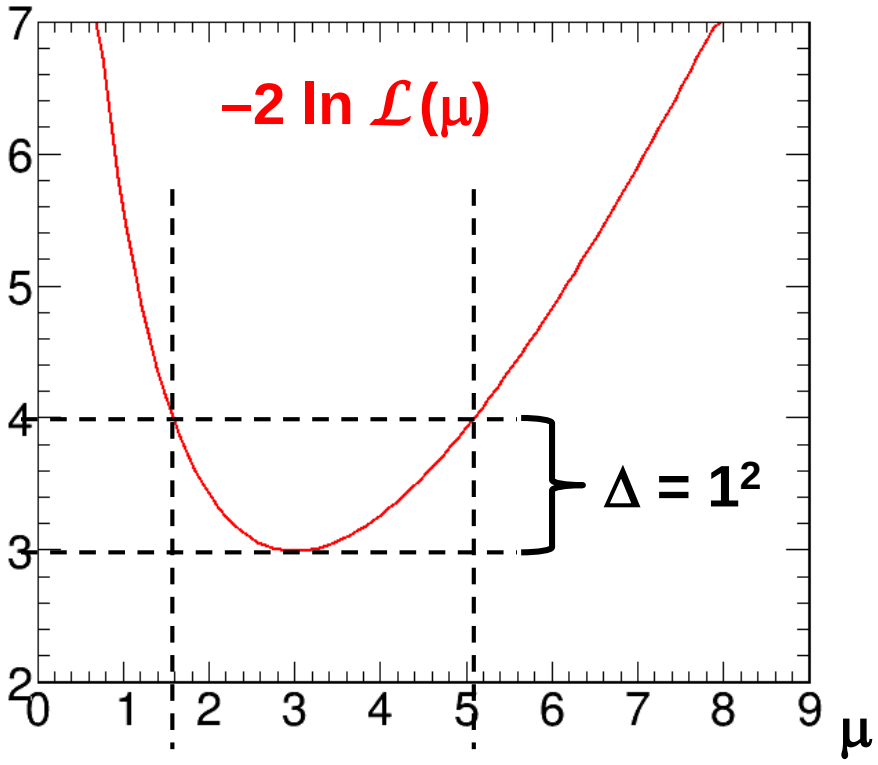
$-2 \ln \mathcal{L}(3) = 2.99$

$\Delta 2\ln\mathcal{L} = 1^2$ **yields LR interval**

$[\mu_1, \mu_2] = [1.58, 5.08]$

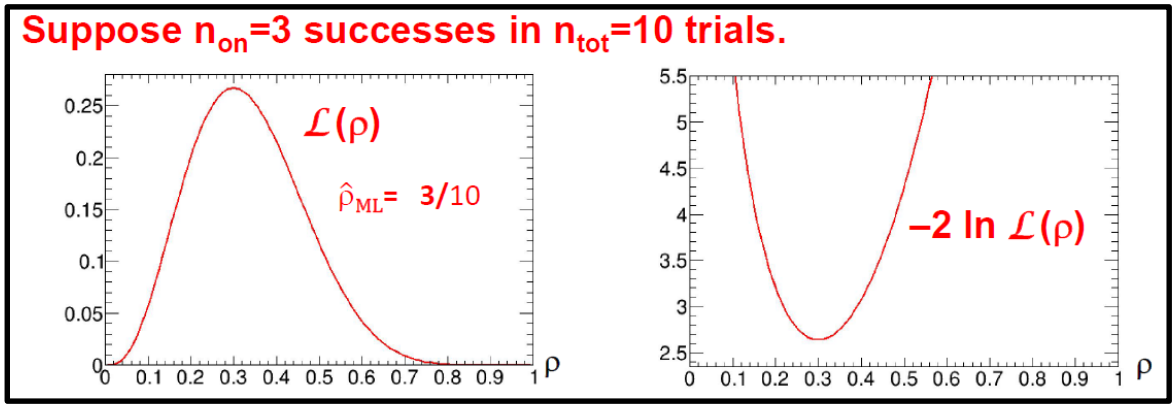
Neyman construction central:

$[\mu_1, \mu_2] = [1.37, 5.92]$



Recall 3 methods of interval construction for binomial param ρ

Bayesian and likelihood intervals: Bi(n_{on} | n_{tot} , ρ) is evaluated only at observed n_{on}=3.



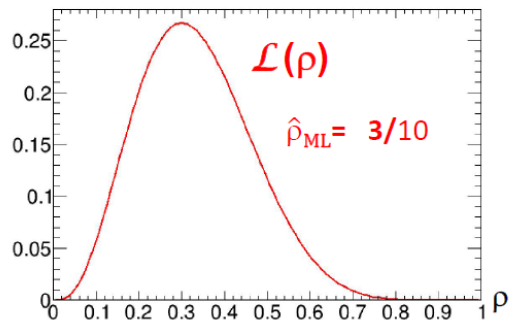
Recall 3 methods of interval construction for binomial param ρ

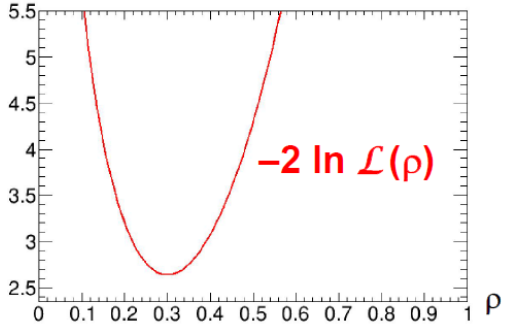
Bayesian and likelihood intervals: $Bi(n_{on} \mid n_{tot}, \rho)$ is evaluated only at observed $n_{on}=3$.

Confidence intervals use, in addition, *probabilities for values of n_{on} not observed.*

This distinction turns out to be a *huge* deal!

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

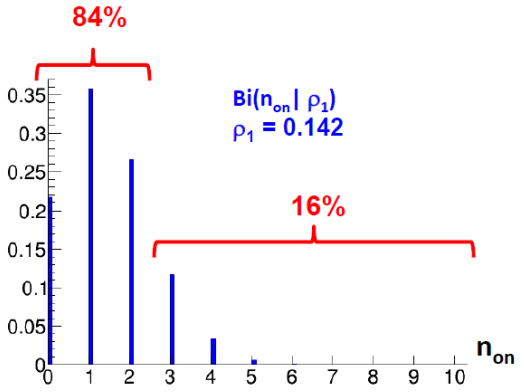


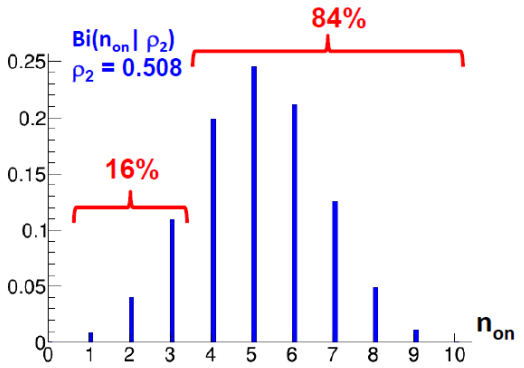


$n_{on} = 3$, $n_{tot}=10$.
Find ρ_1 such that
$Bi(n_{on} < 3 \mid \rho_1) = 84\%$
$Bi(n_{on} \geq 3 \mid \rho_1) = 16\%$
(lower limit at 84% C.L.)
Solve: $\rho_1 = 0.142$

And find ρ_2 such that
$Bi(n_{on} > 3 \mid \rho_2) = 84\%$
$Bi(n_{on} \leq 3 \mid \rho_2) = 16\%$
(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)*





Likelihood Principle

In both Bayesian methods and likelihood-ratio based methods, the probability (density) for obtaining the *data at hand* is used (via the likelihood function), *but probabilities for obtaining other data are not used!*

In contrast, in typical frequentist calculations (confidence intervals, p-values), one also uses probabilities of data that could have been observed but that was *not observed*.

The assertion that only the former is valid is captured by the *Likelihood Principle:**

If two experiments yield likelihood functions that are proportional, then Your inferences from the two experiments should be identical.

***There are various versions of the L.P., strong and weak forms etc. See Stuart99 and book by Berger and Wolpert.**

Likelihood Principle (cont.)

L.P. is built into Bayesian inference (except e.g., when Jeffreys prior leads to violation).

L.P. is violated by p-values and confidence intervals.

Jeffreys (Theory of Probability, 1961, p. 385) still seems to be unsurpassed in his ironic criticism of tail probabilities, which include probabilities of data *more extreme* than that observed:

"What the use of [the p-value] implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred."

Although practical experience indicates that the L.P. may be too restrictive, it is useful to keep in mind. When frequentist results "make no sense" or "are unphysical", in my experience the underlying reason can be traced to a bad violation of the L.P.

Likelihood Principle Example #1

The "Karmen Problem"

- You expect background events sampled from a Poisson distribution with 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 observe 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 #1

The "Karmen Problem"

- **You expect background events sampled from a Poisson distribution with 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 observe 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 Example #2

Binomial problem, famous among statisticians, translated to HEP

You want to measure the efficiency ε of some trigger selection.
You count until reaching n_{tot} =100 zero-bias events, and note that of these, m=10 passed selection.
The probability for m is binomial with binomial parameter ε :

$$Bi(m \mid n_{tot}, \varepsilon) = \frac{n_{tot}!}{m!\,(ntot-m)!}\,\varepsilon^{m}\,(1-\varepsilon)^{(ntot-m)}$$

Estimate ε = 10/100, compute binomial confidence interval for ε.

Also, plugging in the observed data, the likelihood function is

$$\mathcal{L}(\varepsilon) = \frac{100!}{10!\;90!}\,\varepsilon^{10}\,(1-\varepsilon)^{90}$$

Likelihood Principle Example #2 (cont.)

Your colleague *in a different experiment* counts zero-bias events until m=10 have *passed* her trigger. She notes that this requires n_{tot} =100 events (a coincidence).

Intuitively, ε =10/100 *over-estimates* ε because she stopped *just* upon reaching 10 passed events, and indeed an unbiased estimate of ε and confidence interval will be slightly different from the binomial case.

Relevant distribution here is (a version of) the *negative binomial*:

$$NBi(n_{tot} \mid m, \varepsilon) = \frac{(n_{tot}-1)!}{(m-1)!}\,\varepsilon^{m}\,(1-\varepsilon)^{(ntot-m)}$$

Also, plugging in the observed data, the likelihood function is

$$\mathcal{L}(\varepsilon) = \frac{99!}{9!\;90!}\,\varepsilon^{10}\,(1-\varepsilon)^{90}$$

Likelihood Principle Example #2 (cont.)

So both you and your friend observed 10 successes out of 100 trials, but with different *stopping rules*.

Your likelihood function is based on *binomial* distribution:

$$\mathcal{L}(\varepsilon) = \frac{100!}{10!\ 90!}\ \varepsilon^{10}\ (1-\varepsilon)^{90}$$

Your friend's is based on *negative binomial* distribution:

$$\mathcal{L}(\varepsilon) = \frac{99!}{9!\ 90!}\ \varepsilon^{10}\ (1-\varepsilon)^{90}$$

The two likelihoods differ by (only) a constant factor, so the (strong) LP says that inferences should be *identical*.

In contrast, frequentist inferences use probabilities of data not obtained, and result in different confidence intervals and p-values for the different stopping rules.

Likelihood Principle Example #2 (cont.)

The two efficiency measurements had a different *stopping rules*: one stopped after n_{tot} events, and the other stopped after m events passing the trigger.

Frequentist confidence intervals depend on the stopping rule; the likelihood function did not (except for an overall constant).

So Bayesians will get the same answer in both cases, unless the *prior* depends on the stopping rule.

Amusing sidebar: the Jeffreys prior is indeed different for the two distributions, so use of Jeffreys prior violates (strong) L.P.

Stopping Rule Principle

The strong L.P. typically implies, as in this example, that the inference is independent of the stopping rule! This *irrelevance* has been elevated to the "Stopping Rule Principle".

(It is sometimes amusing to ask a recent Bayesian convert if they know that they just bought the Stopping Rule Principle.)

Concepts that average/sum over the sample space, such as bias and tail probabilities, do not exist in pure Bayesian framework.

Stopping Rule Principle (cont.)

Famous quote by L.J. (Jimmie) Savage, early subjective Bayesian advocate:

"...I learned the stopping-rule principle from Professor Barnard, in conversation in the summer of 1952. Frankly, I then thought it a scandal that anyone in the profession could advance an idea so patently wrong, even as today I can scarcely believe that some people resist an idea so patently right."

L.J. Savage et al.,
The Foundations of Statistical Inference: A Discussion
Methuen & Co., London, 1962
Scans of "The Complete Savage Forum" on D. Mayo's web site,
http://www.phil.vt.edu/dmayo/PhilStatistics/supplementary_articles.htm

Summary of Three Ways to Make Intervals

	Bayesian Credible	Frequentist Confidence	Likelihood Ratio
Requires prior pdf?	Yes	No	No
Obeys Likelihood Principle?	Yes (exception re Jeffreys prior)	No	Yes
Random variable in "$P(\mu_t \in [\mu_1, \mu_2])$":	μ_t	μ_1, μ_2	μ_1, μ_2
Coverage guaranteed?	No	Yes (but over-coverage…)	No
Provides P(parameter|data)?	Yes	No	No

Frequentist intervals map to frequentist hypothesis tests, as previously discussed.
Bayesian approach to hypothesis testing is also called *model selection,* and is a whole other "can of worms" (J.O. Berger).

68% intervals for Poisson mean with n=3 observed

Method	Prior	Interval	Length	Coverage?
Wald, $n \pm \sqrt{(n)}$	—	(1.27, 4.73)	3.36	no
Garwood, Frequentist central	—	(1.37, 5.92)	4.55	yes
Bayesian central	1	(2.09, 5.92)	3.83	no
Bayesian central	$1/\mu$	(1.37, 4.64)	3.27	no
Bayesian central Jeffreys	$1/\sqrt{\mu}$	(1.72, 5.27)	3.55	no
Likelihood ratio	—	(1.58, 5.08)	3.50	no

Frequentist intervals over-cover due to discreteness of n.

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

68% intervals for Poisson mean with n=3 observed

Method	Prior	Interval	Length	Coverage?
Wald, $n \pm \sqrt{(n)}$	—	(1.27, 4.73)	3.36	no
Garwood, Frequentist central	—	(1.37, 5.92)	4.55	yes
Bayesian central	1	(2.09, 5.92)	3.83	no
Bayesian central	$1/\mu$	(1.37, 4.64)	3.27	no
Bayesian central Jeffreys	$1/\sqrt{\mu}$	(1.72, 5.27)	3.55	no
Likelihood ratio	—	(1.58, 5.08)	3.50	no

Bayesian *lower* limits with 1/μ prior are identical to frequentist lower limits.

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

68% intervals for Poisson mean with n=3 observed

Method	Prior	Interval	Length	Coverage?
Wald, $n \pm \sqrt{(n)}$	—	(1.27, 4.73)	3.36	no
Garwood, Frequentist central	—	(1.37, 5.92)	4.55	yes
Bayesian central	1	(2.09, 5.92)	3.83	no
Bayesian central	$1/\mu$	(1.37, 4.64)	3.27	no
Bayesian central Jeffreys	$1/\sqrt{\mu}$	(1.72, 5.27)	3.55	no
Likelihood ratio	—	(1.58, 5.08)	3.50	no

Bayesian *upper* limits with flat prior are identical to frequentist upper limits.

Since *upper* limits dominate our field, this is why flat prior for Poisson mean became so well established: it is probability matching prior for upper limits, and when background is added, becomes conservative.

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

Bayesian Hypothesis Testing (Model Selection)

Typically follows Ch. 5 of Harold Jeffreys's book:
Bayes's Theorem is applied to the models themselves after integrating out *all* parameters, including parameter of interest!

Presented too often as "logical" and therefore simple to use, with great benefits such as automatic "Occam's razor", etc.

Bayesian Hypothesis Testing (Model Selection)

Typically follows Ch. 5 of Harold Jeffreys's book:
Bayes's Theorem is applied to the models themselves after integrating out *all* parameters, including parameter of interest!

Presented too often as "logical" and therefore simple to use, with great benefits such as automatic "Occam's razor", etc.

In fact, it is full of subtleties. E.g., Jeffreys and followers use *different priors* for integrating out parameter in model selection than for *same* parameter in parameter estimation.

Here I will mainly just say: Beware! There are posted/published applications HEP that are silly (by Bayesian standards).
An example in PRL provoked me to write a Comment:
https://arxiv.org/abs/0807.1330 .

Bayesian Hypothesis Testing (Cont.)

In asymptotic limit of lots of data, your answer (e.g. probability H_0 is true, or an odds ratio called the Bayes Factor) *remains proportional to the prior pdf of parameter of interest*.

This is *totally different* behavior compared to interval estimation, where the effect of prior becomes negligible.

For testing H_0: $\mu = \mu_0$ vs H_1: $\mu \neq \mu_0$, improper priors for μ that work fine for estimation become a disaster; adding cut-off to make them proper just gives (typically arbitrary) cutoff dependence.

Bayesian Hypothesis Testing (Cont.)

For a review and comparison to p-values in discovery of Higgs boson, see my paper:

"The Jeffreys-Lindley Paradox and Discovery Criteria in High Energy Physics"

(Published in Synthese – long story)
https://arxiv.org/abs/1310.3791 .

1D parameter space, 2D observation space

Until now we have considered 1 parameter and 1 observation. Adding a second observation adds surprising subtleties.

As before, μ is parameter (often called θ by statisticians)

An experiment has two observations x_1, x_2 . These could be:

- **two samples from same $p(x|\mu)$, or**
- **samples of two different quantities from joint density $p(x_1,x_2|\mu)$.**

Neyman construction:

For each μ, use an ordering principle on the sample space (x_1,x_2) to select an acceptance region $\mathcal{A}(\mu)$ in the sample space (x_1, x_2) such that $P\big((x_1,x_2) \in \mathcal{A}(\mu)\big)$ = C.L.

In fact this was the illustration in Neyman's original paper.

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 on vertical axis and horizontal planes for 2D vectors of observables $E = (x_1, x_2)$.

Prior to experiment, acceptance regions $\mathcal{A}(\theta_1)$ in E-space planes are determined for each θ_1 (needs ordering principle) with $P(E \in \mathcal{A}(\theta_1)) = C.L.$

E′ is data actually observed in expt. Upon obtaining E′, confidence interval for θ_1 consists of all values of θ_1 for which E′ is in $\mathcal{A}(\theta_1)$.

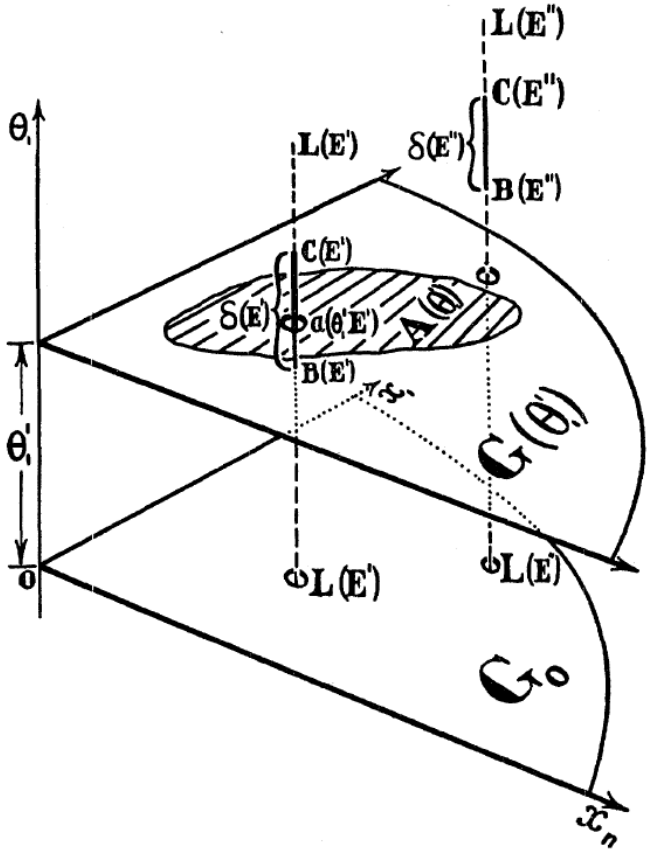


Fig. 1—The general space G.

Recall from Initial Discussion of Probability

For probabilities to be well-defined, the "whole space" needs to be defined. Can be hard for both frequentists and Bayesians!

[...]

Furthermore, it is widely accepted that restricting the "whole space" to a relevant ("conditional") subspace can sometimes improve the quality of statistical inference. The important topic of such "conditioning" in frequentist inference will be discussed in detail later.

"Later" is now!

Restricting the Sample Space Used by Frequentists

In Neyman's construction in the 2D sample space (x_1,x_2), the probabilities $P\big((x_1,x_2) \in \mathcal{A}(\mu)\big)$ associated with each acceptance region $\mathcal{A}(\mu)$ are *unconditional* probabilities with respect to the "whole" sample space of all values of (x_1,x_2).

In contrast, Bayesian inference is based on a single point in this sample space, the observed (x_1,x_2), per the Likelihood Principle.

Restricting the Sample Space Used by Frequentists

In Neyman's construction in the 2D sample space (x_1,x_2), the probabilities $P((x_1,x_2) \in \mathcal{A}(\mu))$ associated with each acceptance region $\mathcal{A}(\mu)$ are *unconditional* probabilities with respect to the "whole" sample space of all values of (x_1,x_2).

In contrast, Bayesian inference is based on a single point in this sample space, the observed (x_1,x_2), per the Likelihood Principle.

There can be a middle ground in frequentist inference, in which the probabilities $P((x_1,x_2) \in \mathcal{A}(\mu))$ are *conditional* probabilities conditioned on a function of (x_1,x_2), in effect restricting the sample space to a "recognizable subset" depending on the observed data.

The function of (x_1,x_2) used for conditional probabilities typically carries information on the *uncertainty* in the point estimate $\hat{\mu}$, but no information on $\hat{\mu}$ itself: called an *ancillary statistic*.

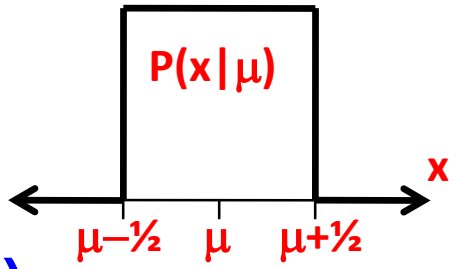
Restricting the Sample Space Used by Frequentists (cont.)

Restricting the sample space in this way is known as *conditioning (on an ancillary)*. Two famous examples:

1) **A somewhat artificial example of Welch where the conditioning arises from mathematical structure**
2) **A more physical example of Cox where the argument for conditioning seems "obvious"**

Famous example of B.L. Welch (1939)

Let $p(x|\mu) = \begin{cases} 1 \text{ if } \mu - ½ \le x \le \mu + ½ \\ 0 \text{ otherwise.} \end{cases}$



Two values x_1, x_2 are observed.

$\hat{\mu} = \bar{x} = (x_1+x_2)/2$ (Only for n=2! See Backup)

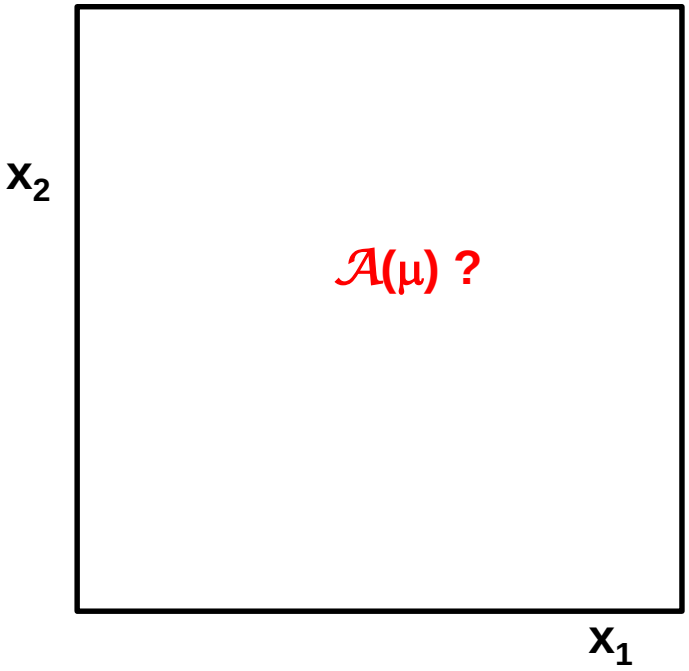
What is 68% C.L. central confidence interval for μ?

Neyman construction: Define acceptance region $\mathcal{A}(\mu)$ containing 68% of unit square of (x_1,x_2) centered on μ.
What to use?

Centrality implies symmetry.

Need something else to rank points in the plane.

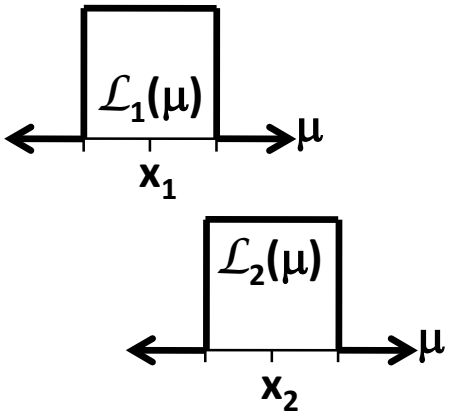
N-P Lemma gives most powerful ranking, but first let's think about some examples.



"Lucky" sample with $|x_1 - x_2|$ close to 1.

$\mathcal{L}(\mu) = \mathcal{L}_1(\mu) \times \mathcal{L}_2(\mu)$ very narrow.

Reasonable to expect small uncertainty in $\hat{\mu}$?

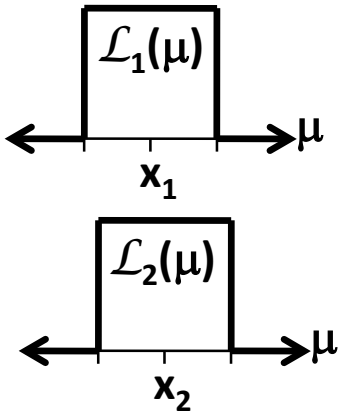


"Unlucky" sample, $|x_1 - x_2|$ close to 0.

$\mathcal{L}(\mu)$ full width close to 1;
Second observation added no useful info.

Expect 68% C.L. conf. interval 0.68 long?

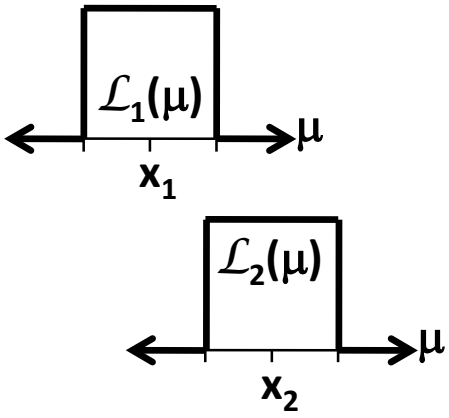
Guess reasonable answer: conf. interval centered on $\hat{\mu}$ with length $0.68(1-|x_1-x_2|)$

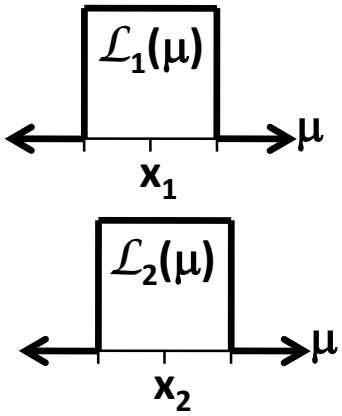


Seems reasonable for *post-data uncertainty to* depend on $|x_1 - x_2|$.

Classic example of an <u>ancillary statistic A</u>: has info on *uncertainty* on μ estimate, but no info on μ itself.

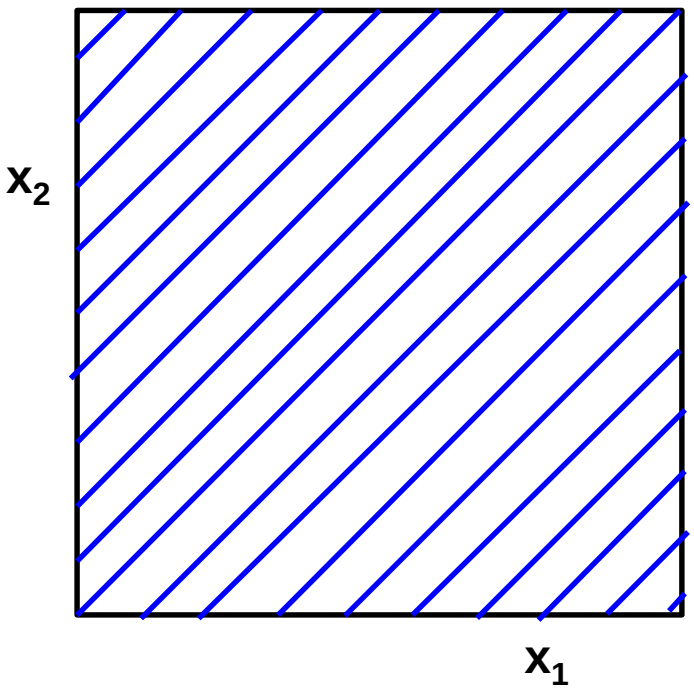
Idea dating to Fisher and before: divide the full *"unconditional" sample space* into "recognizable subsets" and report probs using the "relevant" subset rather than the whole space.





Acceptance regions partitioned by $|x_1-x_2|$

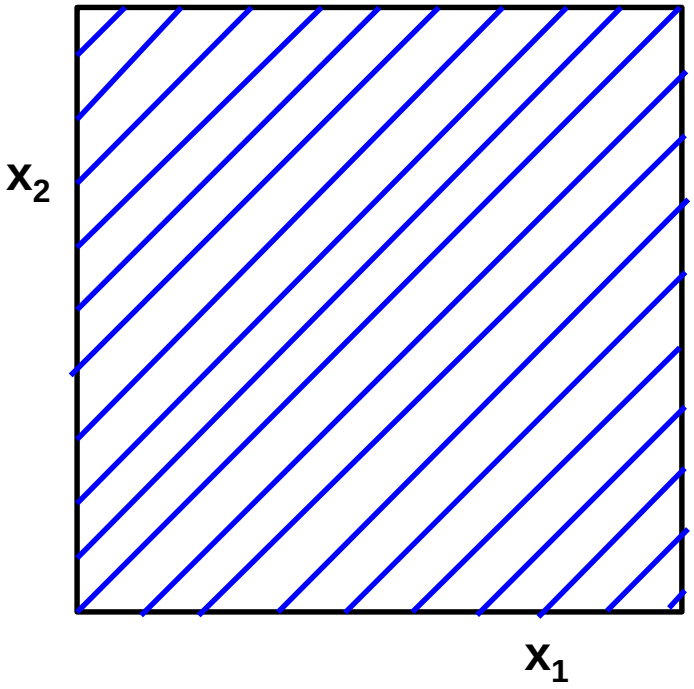
Partition full sample space via ancillary statistic A= $|x_1-x_2|$ (blue lines)



Acceptance regions partitioned by $|x_1-x_2|$

Partition full sample space via ancillary statistic A= $|x_1-x_2|$ (blue lines)

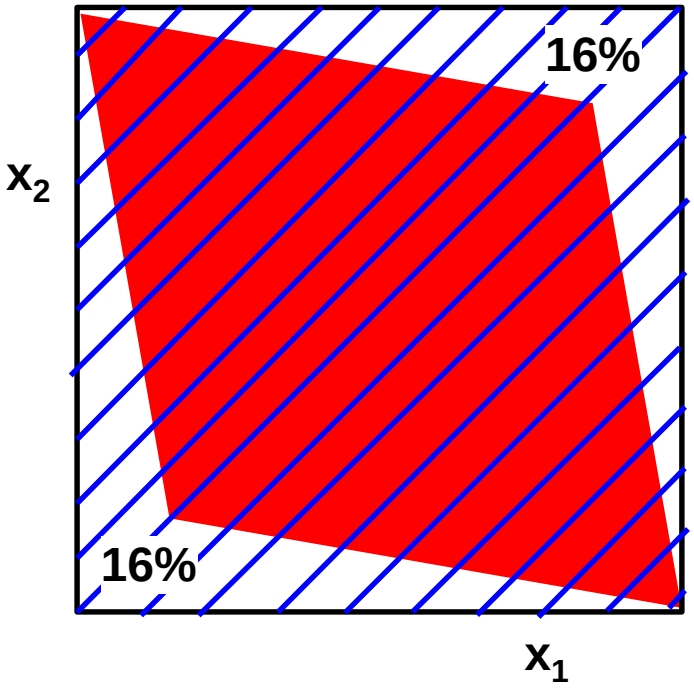
Within each partition, choose central 68% prob acceptance region



Acceptance regions partitioned by $|x_1-x_2|$

Partition full sample space via ancillary statistic A= $|x_1-x_2|$ (blue lines)

Within each partition, choose central 68% prob acceptance region (red fill)



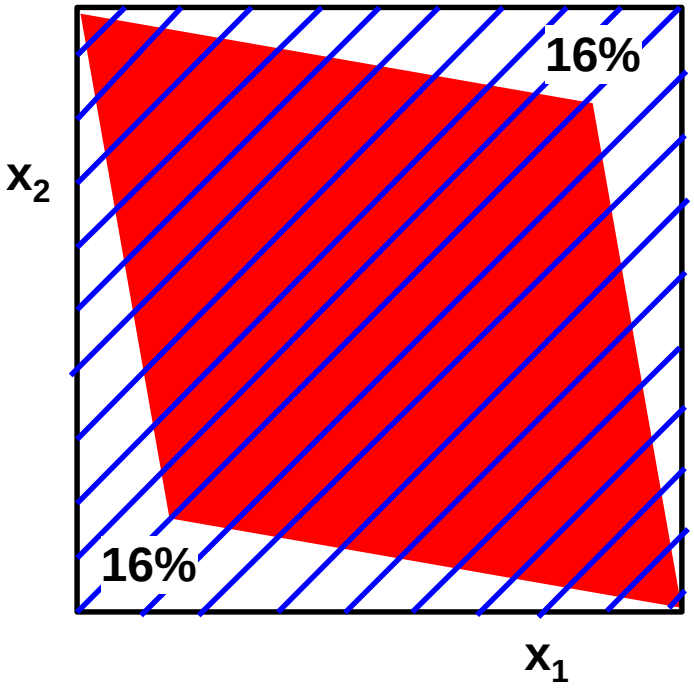
Acceptance regions partitioned by $|x_1-x_2|$

Partition full sample space via ancillary statistic A= $|x_1-x_2|$ (blue lines)

Within each partition, choose central 68% prob acceptance region (red fill)

We are thus using *conditional probabilities* (still frequentist) p(x|A,μ) in Neyman construction, with desired prob 68% *within each partition*.

Resulting $\mathcal{A}(\mu)$ fills 68% of square, so correct unconditional probability as well.



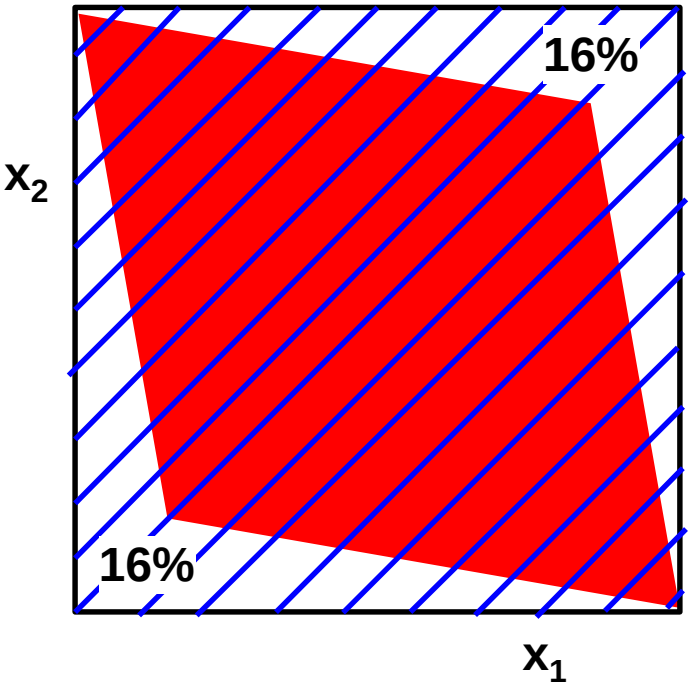
Acceptance regions partitioned by $|x_1-x_2|$

Partition full sample space via ancillary statistic A= $|x_1-x_2|$ (blue lines)

Within each partition, choose central 68% prob acceptance region (red fill)

We are thus using *conditional probabilities* (still frequentist) p(x|A,μ) in Neyman construction, with desired prob 68% *within each partition*.*

Resulting $\mathcal{A}(\mu)$ fills 68% of square, so correct unconditional probability as well.



$\Rightarrow$ Confidence intervals $\hat{\mu} \pm 0.34(1-|x_1-x_2|)$, as thought reasonable!

Known as "conditioning" on ancillary statistic A: Post-data, *proceed as if A had been fixed, rather than randomly sampled!*

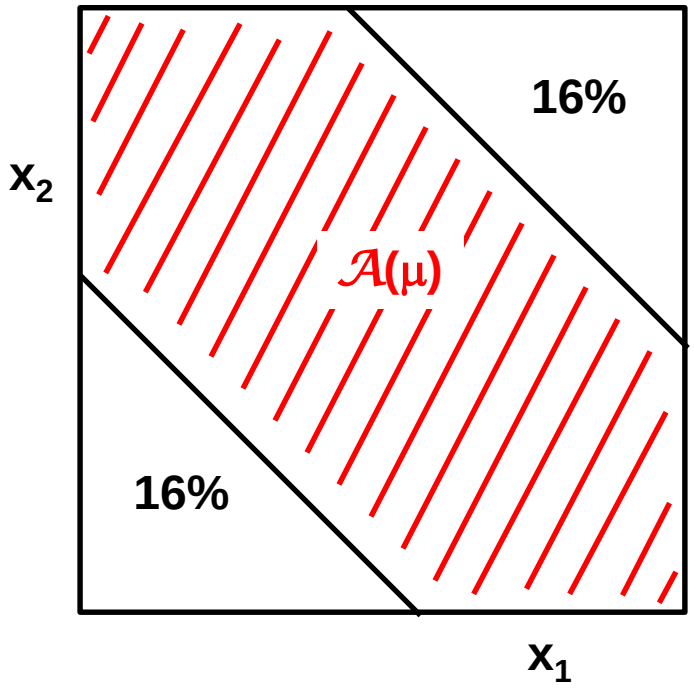
N.B. A set of measure zero has zero prob even if non-zero pdf, so in general care needed in conditioning on exact value of continuous A in p(x|A,μ).

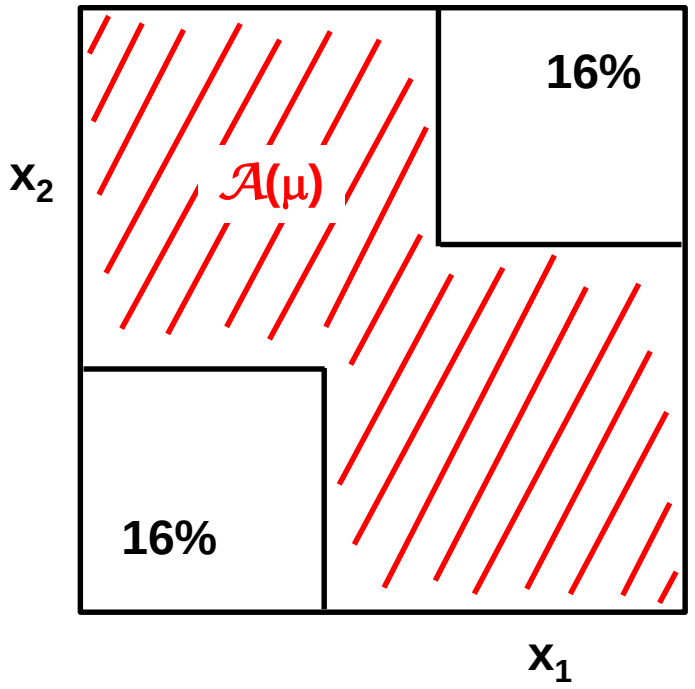
Now the catch: one can find acceptance regions $\mathcal{A}(\mu)$ that correspond to hypothesis tests with *more power* (lower Type 2 error prob β) in *the unconditional sample space*!

These have 100% coverage in the subspace where $|x_1 - x_2| \approx 1$ (narrow likelihood), while badly undercovering when $x_1 \approx x_2$.

Interval length indep of $|x_1 - x_2|$
$\Rightarrow \hat{\mu} \pm 0.22$ at 68% C.L.

Most powerful?





In fact Welch's 1939 paper argued *against* conditioning because it is less powerful in the unconditional sample space!

Neyman's position is not completely clear but he seems to have been against conditioning on ancillaries (which was Fisher's idea) when it meant an overall loss of power.*

Most modern writers use this example as an "obvious" argument *in favor* of conditioning, unless one is in "industrial" setting where unconditional ensemble is sampled repeatedly and the result for an individual sample is not of much interest.

***See J.O. Berger, https://projecteuclid.org/euclid.ss/1056397485 and**

E.L. Lehmann, http://www.jstor.org/stable/2291263 (also discussion of Cox example).

Brad Efron's talk at PhyStat-2003 gave similar example using Cauchy distribution, where ancillary statistic is curvature of $\mathcal{L}(\mu)$. He called conditional answer "correct".
(http://www.slac.stanford.edu/econf/C030908/papers/MOAT003.pdf)

As noted, conditioning on an ancillary statistic A means:

Even though A was randomly sampled in the experimental procedure, after data is obtained proceed as if A had been fixed to the value observed. Ignore sample space with all those other values of A that you could have obtained, but did not.

Famous “weighing machine” example of D.R. Cox (1958)

For measuring the mean μ with Gaussian resolution, one of two devices is selected randomly with equal probability:

- **Device #1 with σ_1**
- **Device #2 with $\sigma_2 \ll \sigma_1$.**

So in my notation, x_1 is the index (1 or 2) chosen randomly and specifying device, and x_2 is the single sample from the selected Gaussian measurement.

In Behnke13 (p. 122) Luc Demortier gives an example in HEP: μ is mass of decaying particle with probability p_h to decay hadronically (mass resolution σ_1) and probability 1-p_h to decay leptonically with mass resolution σ_2). Thus the “measuring machine” chosen randomly is the detector used to measure the decay mode chosen by QM.

Example of D.R. Cox (cont.)

- Device #1 with σ_1
- Device #2 with $\sigma_2 \ll \sigma_1$.

x_1 is the index (1 or 2) chosen randomly and specifying device;
x_2 is the single sample from the selected measurement.

So $\hat{\mu} = x_2$. What is the confidence interval?

The index x_1 is an *ancillary statistic* (gives info on uncertainty but not on point estimate), and it is reasonable (obvious?) to condition on it. I.e., we report confidence interval giving correct coverage in *subspace of measurements that used the same device we used*.

So 68% C.L. confidence interval:

- $\hat{\mu} \pm \sigma_1$ if Device #1 randomly selected
- $\hat{\mu} \pm \sigma_2$ if Device #2 randomly selected

As in Welch example, it turns out that more powerful tests (confidence intervals shorter on average!) can be found.

Example of D.R. Cox (cont.)

Demortier gives details on how average length of intervals optimized in the unconditional sample space is shorter in the HEP example. See backup for details of Cox's example.

In both cases, the idea is to undercover badly (~38% coverage for 68% C.L. intervals) when Device #1 is used, and overcover (~100% coverage) when Device #2 is used. The smaller σ of Device #2 means that average length of interval goes down, *when averaging over the entire unconditional sample space.*

One gives up power with Device #1 and uses it in Device #2.

Cox: "If, however, our object is to say `what can we learn from the data that we have', the unconditional test is surely no good."

The Welch and Cox (and Efron) examples reveal a real conflict between N-P optimization for power and conditioning to optimize relevance.

Conditionality Principle

The assertion that inference should be conditioned on an ancillary in the Welch example (where it comes out of the math) is often called the "Conditionality Principle".

Conditioning in the Cox example (a "mixture experiment" where the ancillary has physical meaning about which experiment was performed) is then the "Weak Conditionality Principle".

But note: in sufficiently complicated cases (for example if there is more than one ancillary statistic), the procedure is less clear.

In many situations, ancillary statistics do not exist, and it is not at all clear how to restrict the "whole space" to the relevant part for frequentist coverage.

The Bayesian answer: restrict the whole space to the point observed! But the price is giving up coverage.

When there are "recognizable subsets" with varying coverage, Buehler has discussed how a "conditional frequentist" can win bets against an "unconditional frequentist" – see Backup.

Conditioning in HEP

A classic example is a *measurement of the branching fraction of a particular decay mode* when the *total* number of decays N can fluctuate because the experiment design is to run for a fixed length of time. Then N is an ancillary statistic.

You perform an experiment and obtain N total decays, and then do a toy M.C. of repetitions of the experiment. Do you let N fluctuate, or do you fix it to the value observed?

It may seem that the toy M.C. should include your *complete* procedure, including fluctuations in N.

But the above arguments would point toward *conditioning on the value of the ancillary statistic actually obtained*. So your branching fraction measurement is binomial with trials N.

(Originally discussed in HEP by F. James and M. Roos, Nucl. Phys. B 172 (1980) 475. For more complete discussion, see Cousins, Hyme, Tucker, https://arxiv.org/abs/0905.3831)

Review: Conditioning and the Likelihood Principle

We have seen that *unconditional frequentists* compute probabilities with respect to the *whole sample space*.

Post-data, *conditional frequentists* try to refer to a *relevant subset* of the whole sample space (typically not easy).

We also saw that pure Bayesians refer only to the probability of the data observed (L.P.). This is literally the ultimate extreme in conditioning, *conditioning (in the continuous case) on a point of measure zero!* (You can't get any more "relevant".)

This is why coverage is not built into Bayesian answers.

Conditioning and the Likelihood Principle (cont.)

It is not surprising that Bayesians argue for the importance of *relevance* of the inference, and criticize frequentists for danger of irrelevance (and difficulty of diagnostic of irrelevance).

And it is not surprising that frequentists argue for the importance of a useful measure of "error rates", in the sense of Type 1 and Type 2 errors, coverage, etc., which may at best be estimates if L.P. is observed.

(Finally!) More than One Parameter

Generalize to two parameters μ_1 and μ_2, true values unknown.

Let data x be a multi-D vector, so the model is $p(x|\mu_1,\mu_2)$.

Observed vector value is x_0 .

First consider the desire to obtain a 2D confidence/credible *region* in the parameter space (μ_1,μ_2). All three methods discussed for intervals handle this in a straightforward (in principle) generalization:

- **Bayesian: put observed data vector x_0 into $p(x|\mu_1,\mu_2)$ to obtain the likelihood function $\mathcal{L}(\mu_1,\mu_2)$. Multiply by prior pdf $p(\mu_1,\mu_2)$ to obtain 2D posterior pdf $p(\mu_1,\mu_2|x_0)$.**
  Use posterior pdf to obtain credible regions, etc., in (μ_1,μ_2).

(Finally!) More than One Parameter

Generalize to two parameters μ_1 and μ_2, true values unknown.

Let data x be a multi-D vector, so the model is $p(x|\mu_1,\mu_2)$.

Observed vector value is x_0 .

First consider the desire to obtain a 2D confidence/credible *region* in the parameter space (μ_1,μ_2). All three methods discussed for intervals handle this in a straightforward (in principle) generalization:

- **Bayesian: put observed data vector x_0 into $p(x|\mu_1,\mu_2)$ to obtain the likelihood function $\mathcal{L}(\mu_1,\mu_2)$. Multiply by prior pdf $p(\mu_1,\mu_2)$ to obtain 2D posterior pdf $p(\mu_1,\mu_2|x_0)$. Use posterior pdf to obtain credible regions, etc., in (μ_1,μ_2).**
- **Confidence intervals: perform Neyman construction: Find acceptance regions for x as a function of (μ_1,μ_2). The 2D confidence region is union of all (μ_1,μ_2) for which x_0 is in acceptance region.**

(Finally!) More than One Parameter

Generalize to two parameters μ_1 and μ_2, true values unknown.

Let data x be a multi-D vector, so the model is $p(x|\mu_1,\mu_2)$.

Observed vector value is x_0 .

First consider the desire to obtain a 2D confidence/credible *region* in the parameter space (μ_1,μ_2). All three methods discussed for intervals handle this in a straightforward (in principle) generalization:

- **Bayesian: put observed data vector x_0 into $p(x|\mu_1,\mu_2)$ to obtain the likelihood function $\mathcal{L}(\mu_1,\mu_2)$. Multiply by prior pdf $p(\mu_1,\mu_2)$ to obtain 2D posterior pdf $p(\mu_1,\mu_2|x_0)$. Use posterior pdf to obtain credible regions, etc., in (μ_1,μ_2).**
- **Confidence intervals: perform Neyman construction: Find acceptance regions for x as a function of (μ_1,μ_2). The 2D confidence region is union of all (μ_1,μ_2) for which x_0 is in acceptance region.**
- **Likelihood regions: recall 1D method $2\ln\mathcal{L}(\hat{\mu}) - 2\ln\mathcal{L}(\mu) \leq Z^2$…**

Likelihood regions in ≥ 2D

Find global maximum of $\mathcal{L}(\mu_1,\mu_2)$, yielding point estimates $(\hat{\mu}_1,\hat{\mu}_2)$.

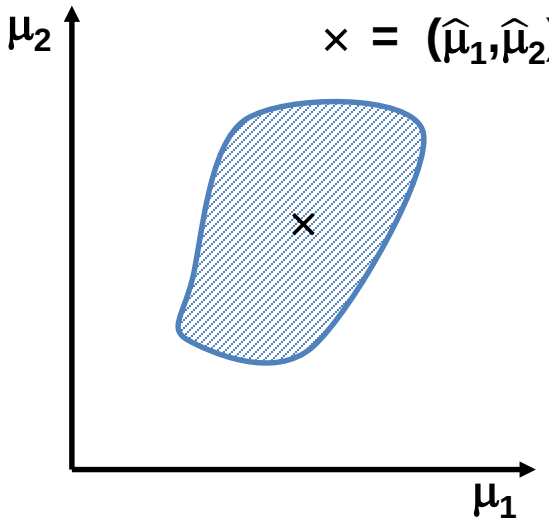
Find contour bounded by $2\Delta\ln\mathcal{L} = 2\ln\mathcal{L}(\hat{\mu}_1,\hat{\mu}_2) - 2\ln\mathcal{L}(\mu_1,\mu_2) \leq C$, where C comes from Wilks's Theorem, tabulated in PDG RPP:

Table 39.2: Values of $\Delta\chi^2$ or $2\Delta\ln L$ corresponding to a coverage probability $1-\alpha$ in the large data sample limit, for joint estimation of m parameters.

$(1-\alpha)$ (%)	$m=1$	$m=2$	$m=3$
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

http://pdg.lbl.gov/2018/reviews/rpp2018-rev-statistics.pdf

2D joint ≈ 68% C.L.
$2\Delta\ln\mathcal{L} = 2.3$



As in 1D, Wilks's Theorem is asymptotic (large N) result, with various "regularity conditions" to be satisfied.

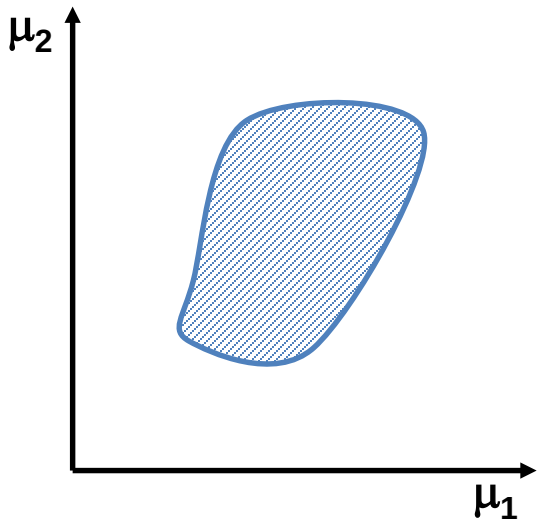
Nuisance Parameters

Frequently one is interested in considering one parameter at a time, irrespective of the value of other parameter(s).

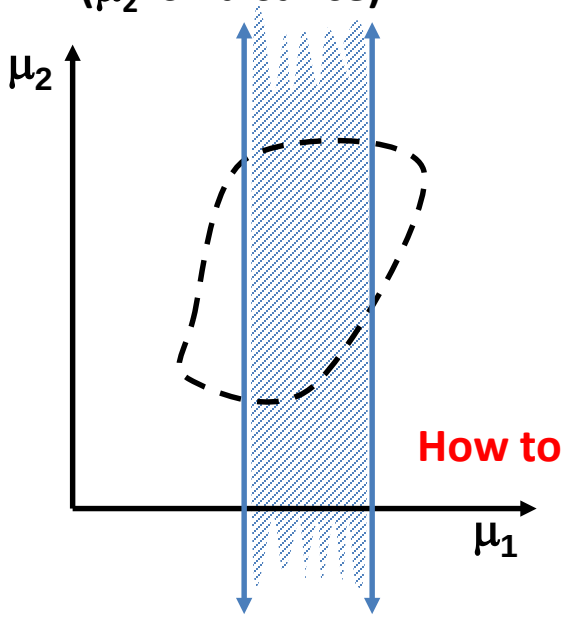
The parameter under consideration at the moment is called the "parameter of interest" and the other parameters (at that moment) are called "nuisance parameters".

E.g., if μ_1 is of interest and μ_2 and is a nuisance, one seeks a 2D confidence region that is a vertical "stripe" in the (μ_1,μ_2) plane.

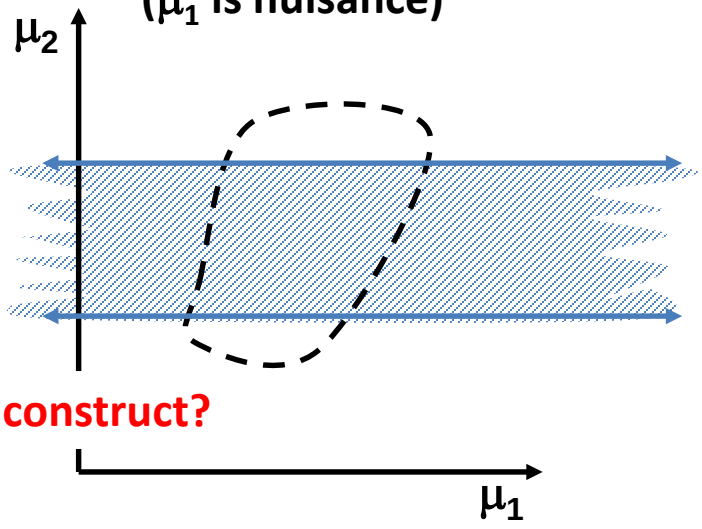
2D joint 68% C.L.



2D region to get 1D 68% C.L. interval for μ_1 (μ_2 is nuisance)



2D region to get 1D 68% C.L. interval for μ_2 (μ_1 is nuisance)



How to construct?

Systematic Uncertainties as Nuisance Parameters

A typical measurement in HEP has many subsidiary measurements of quantities not of direct physics interest, but which enter into the calculation of the physics quantity of particular interest.

E.g., if an absolute cross section is measured, one will have uncertainty in the integrated luminosity L, in the background level b, the efficiency e of detecting the signal, etc. In HEP, we call these systematic uncertainties, but statisticians (for the obvious reason) refer to L, b, and e as *nuisance parameters*.

Each of the three main classes of constructing intervals (Bayesian, Neyman confidence, likelihood ratio) has a "native" way to incorporate the uncertainty on the nuisance parameters.

But this remains a subject of frontier statistics research.

Nuisance Parameters I: Bayesian Credible Intervals

Construct a multi-D prior pdf p(parameters) for the space spanned by all parameters.

Multiply by $\mathcal{L}$(data|parameters) for the data obtained to obtained multi-D posterior pdf.

Integrate over the full subspace of all nuisance parameters (marginalization).

Thus obtain posterior pdf for the parameter of interest.

Math is reduced to the case of no nuisance parameters.

Nuisance Parameters I: Bayesian Credible Intervals

Construct a multi-D prior pdf p(parameters) for the space spanned by all parameters.

Multiply by $\mathcal{L}$(data|parameters) for the data obtained to obtained multi-D posterior pdf.

Integrate over the full subspace of all nuisance parameters (marginalization).

Thus obtain posterior pdf for the parameter of interest.

Math is reduced to the case of no nuisance parameters.

***Problems*: The multi-D prior pdf is a problem for both subjective and non-subjective priors. In HEP there has been little 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.**

As with all Bayesian analyses, how to interpret probability?

Nuisance Parameters II: Neyman Construction

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: **Typically intractable and causes overcoverage, and therefore rarely attempted.**

Nuisance Parameters II: Neyman Construction

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: **Typically intractable and causes overcoverage, and therefore rarely attempted.**

Tractability can sometimes be recovered by doing the construction in the lower dimensional space of the profile likelihood function, obtaining approximate coverage.
(This is one way to interpret the Kendall and Stuart page on likelihood ratio test with nuisances.)

Problem: Not well-studied.

Typically "elimination" is done in a way technically feasible, and the coverage studied with toy MC.

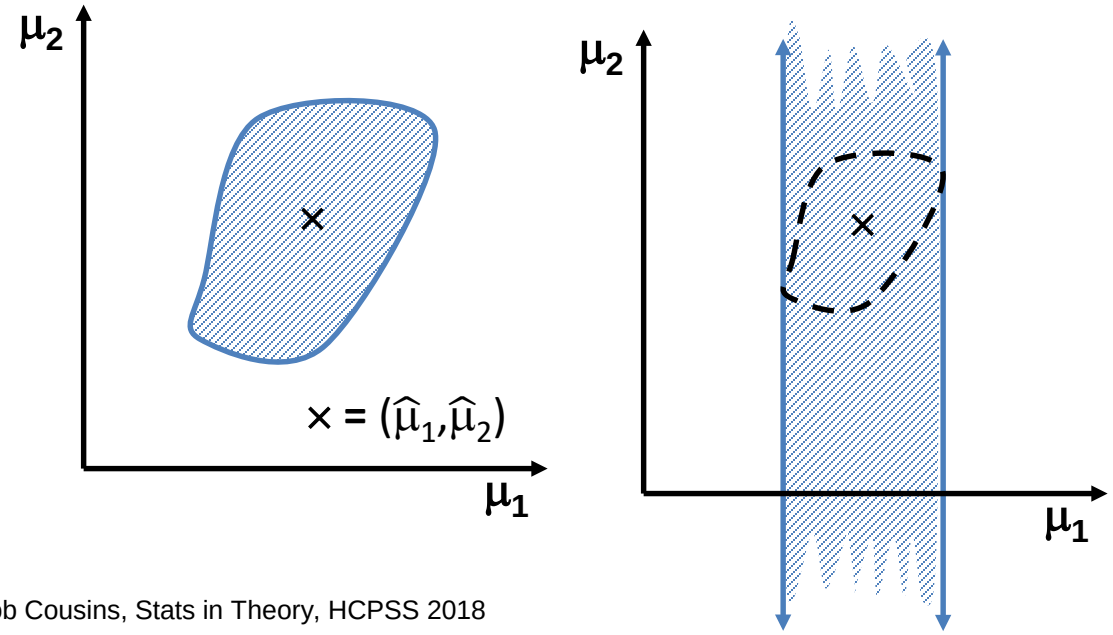
Nuisance Parameters III: Likelihood Ratio intervals

Many of us raised on MINUIT MINOS read F. James, "Interpretation of the Shape of the Likelihood Function around Its Minimum," Computer Physics Communications 20 (1980) 29.

Whereas 2D region has m=2 and hence $2\Delta\ln\mathcal{L} = 2.3$ from PDG RPP table, for 1D interval on μ_1, we first make *2D contour* with m=1 value, $2\Delta\ln\mathcal{L} = 1$ (black dashed), and then find extrema in μ_1:

2D joint ≈ 68% C.L.
$2\Delta\ln\mathcal{L} = 2.3$

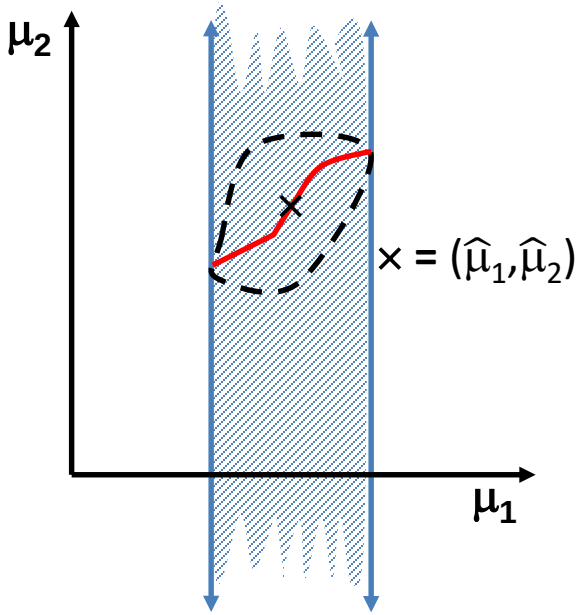
2D region to get 1D 68% C.L. interval for μ_1
$2\Delta\ln\mathcal{L} = 1$ (dashed)



...and then at the Fermilab Confidence Limits Workshop in 2000, statistician Wolfgang Rolke expressed the construction a different way:

Profile Likelihood Function

2D region to get 1D 68% C.L. interval for μ_1
$2\Delta\ln\mathcal{L} = 1$ (dashed)



Red curve is path $(\mu_1,\hat{\hat{\mu}}_2)$ along which profile $\mathcal{L}$ is evaluated

For each μ_1, find the value $\hat{\hat{\mu}}_2$ that minimizes $-2\ln\mathcal{L}(\mu_1,\hat{\hat{\mu}}_2)$.

Make 1D plot vs μ_1 of this "profile likelihood function".

Use the m=1 threshold on $2\Delta\ln\mathcal{L}_{\text{profile}}$.

One obtains the exact same interval as "MINOS" on the left? Can you see why?

Since 2000, this statistical terminology has permeated HEP. It is also used for example in the "Kendall and Stuart" page that I showed re F-C.

Warning: Combining profile likelihoods from two experiments is unreliable. Apply profiling after combining full likelihoods.

Likelihood Ratio intervals (cont.)

Problems:

Coverage is not guaranteed, particularly at low N. By using best-fit value of the nuisance parameters corresponding to each value of the parameter of interest, this has an (underserved?) reputation for underestimating the true uncertainties.

In Poisson problems, this is partially compensated by effect due to discreteness of n, and profile likelihood (MINUIT MINOS) gives good performance in many problems. See Rolke et al., NIM A551 (2005) 493.

In some cases (for example when the are spikes in $\mathcal{L}$), marginalization may give better frequentist performance, according to statisticians.

Hybrid Techniques: Introduction to Pragmatism

Given the difficulties with all three classes of interval estimation, especially when incorporating nuisance parameters, it is common in HEP to relax foundational rigor and:

- **Treat nuisance parameters in a Bayesian way (marginalization) while treating the parameter of interest in a frequentist way.**
  Virgil Highland and I were early advocates of this for lumi uncertainty in upper limit calculation (NIM A320 (1992) 331).
  Kyle Cranmer exposed problems when used for background mean in 5σ discovery context.
  For review of background case and connection to Box's semi-Bayesian "prior predictive p-value", see NIM A595 (2008) 480, https://arxiv.org/abs/physics/0702156
- **Treat nuisance parameters by profile likelihood while treating parameter of interest another way, or**

Introduction to Pragmatism (cont.)

- **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. In fact, the statistics literature going back to 1963 has attempts to find prior pdfs that lead to posterior pdfs with good frequentist coverage: *probability matching priors. (At lowest order in 1D, it is the Jeffreys prior!)***

Evaluation of coverage with toy MC

For a single parameter of interest μ, after elimination of nuisance parameters by some approximate method and construction of intervals perhaps involving more approximations, one reports as usual the confidence interval $[\mu_1, \mu_2]$ at some C.L.

It is important to check that the approximations in the whole procedure have not materially altered the claimed coverage:

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

Typically the performance is evaluated with toy MC.

First I describe the most thorough check (very CPU intensive), and then some approximations.

In frequentist statistics, the true values of all parameters are typically *fixed but unknown*. A complete, rigorous check of coverage considers a fine multi-D grid of *all* parameters, and *for each multi-D point in the grid*, generates an ensemble of toy MC pseudo-experiments, runs the full analysis procedure, and finds the fraction of intervals covering μ_t used for that ensemble, i.e., $P(\mu_t \in [\mu_1, \mu_2]$, and compares to C.L.

Evaluation of coverage with toy MC (cont.)

Thus a thorough check of frequentist coverage includes:

1) **Fix all parameters (of interest and nuisance) to a single set of true values. For this set,**
   a) **Loop over "pseudo-experiments"**
   b) **For each pseudo-experiment, loop over events, generating each event with toy data generated from the statistical model with parameters set equal to the fixed set.**
   c) **Perform the same analysis on the toy events in the pseudo-experiment as was done for the real data.**
   d) **Find that fraction of the pseudo-experiments for which parameter(s) of interest are included in stated confidence intervals or regions.**
2) ***Repeat for a various other fixed sets of all parameters. But...*****the ideal of a fine grid is usually impractical.**

So the issue is what selection of "various other fixed sets" is adequate

Evaluation of coverage with toy MC (cont.)

The ideal of a fine grid is usually impractical.

So the issue is what selection of "various other fixed sets" is adequate.

Obviously one should check check coverage if the set of true values is equal to the global best-fit values.

Just as obviously, this may not be adequate. Some exploration is needed, particularly in directions where uncertainty on a parameter depends strongly on the parameter. One can start by varying a few critical parameters by one or two s.d., trying parameters near boundary/ies, and seeing how stable coverage is.

A Bayesian-inspired approach is to calculate a weighted average of coverage over a neighborhood parameter sets for the nuisance parameters. This requires a choice of multi-D prior (Recall problems.) Instead of fixing the true values of nuisance parameters during the toy-tossing, one samples the true parameters from the posterior distribution.

Evaluation of coverage with toy MC

Numerous studies have been done for eliminiaton of nuisances parameters in the test statistic, many concluding that results are relatively insensitive to profiling vs marginalization, so that choice can be made based on CPU time.

See for example John Conway's talk and writeup at PhyStat-2011, https://indico.cern.ch/event/107747/timetable/#all.detailed

Anecdotally, the choice for toys is more important than the choice for test statistic.

A couple looks at the statistics literature on nuisances

About 10-12 years ago, Luc Demortier and I both looked in the statistics literature regarding nuisance parameters – I thought my note was fairly thorough until I read his! Our writeups:

R.D. Cousins, Treatment of Nuisance Parameters in High Energy Physics, and Possible Justifications and Improvements in the Statistics Literature, with response by statistician Nancy Reid,

http://www.physics.ox.ac.uk/phystat05/proceedings/default.htm

Luc Demortier, P Values: What They Are and How to Use Them,

https://www-cdf.fnal.gov/~luc/statistics/cdf8662.pdf (2007)

See also his chapter in Behnke13.

Priors for nuisance parameters

It used to be (unfortunately) common practice to express, say, a 50% systematic uncertainty on a positive quantity as a Gaussian with 50% rms. Then "truncate" Gaussian by not using non-positive values.

In Bayesian calculation, interaction of uniform prior for Poisson mean and "truncated Gaussian" for systematic uncertainty in efficiency leads to integral that diverges if truncation is at origin…evaluating integral numerically will not even notice! Luc Demortier exposed this:
http://www.ippp.dur.ac.uk/old/Workshops/02/statistics/proceedings//demortier.pdf

***Recommendation:* Use log-normal or (certain) Gamma distributions instead of truncated Gaussian. Recipes at**
http://www.physics.ucla.edu/~cousins/stats/cousins_lognormal_prior.pdf

“State of the Art” in HEP

All three main classes of methods are commonly used on the parameter of interest.

- **Both marginalization and profiling are commonly used to treat nuisance parameters.**
- **Many people have the good practice of checking coverage.**
- **Too little attention is given to priors. But flat prior for Poisson mean is “safe” for *upper* limits (only!).**

A serious analysis using any of the main methods requires coding up the likelihood function.

- **Doing this (once!) with RooFit modeling language gives access to RooStats techniques for all three classes of calculations, mix-match nuisance parameter treatments.**

ATLAS and CMS Conventions

For many years, ATLAS and CMS physicists have collaborated on statistics tools (the RooStats software), and attempted to have some coherence in methods so that results could be compared, and (when worth the effort) combined.

An important development was the paper by Cowan, Cranmer, Gross and Vitells, https://arxiv.org/abs/1007.1727 , that extended asymptotic formulas to a number of cases where Wilks's theorem was not valid.

As the CCGV asymptotic formulas applied to the "fully frequentist" treatment of nuisances parameters, for consistency we tended to use that in many cases at small N as well. Toy MC is thus done in a frequentist manner (so-called parametric bootstrap).

For upper limits, there was a lot of discussion without convergence, and the two physics coordinators in 2010 decreed that CL_s be used in most cases. (See below.)

ATLAS and CMS Conventions (cont.)

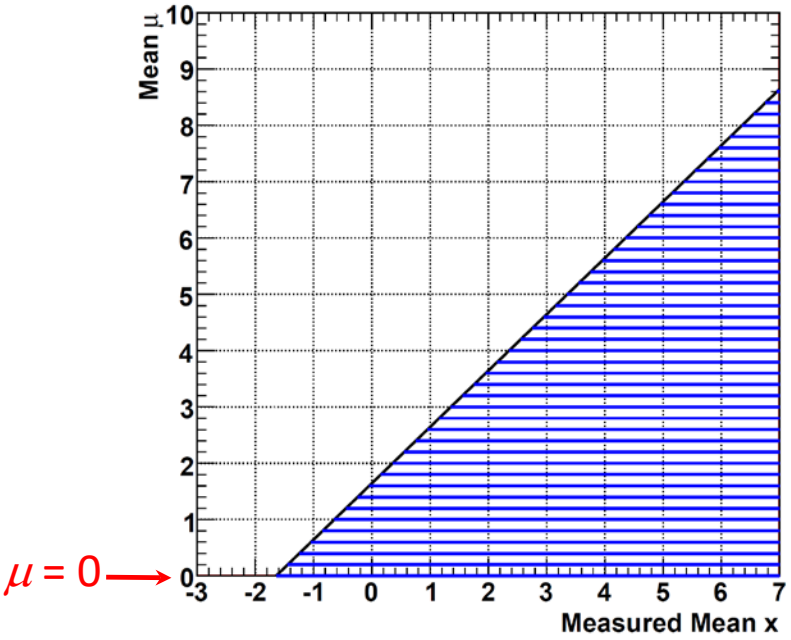
The ATLAS/CMS Higgs results followed these trends. A jointly written description is, "Procedure for the LHC Higgs boson search combination in Summer 2011," http://cds.cern.ch/record/1379837.

Many issues were further discussed and described in the ATLAS-CMS combination papers for mass and couplings, https://arxiv.org/abs/1503.07589 and https://arxiv.org/abs/1606.02266 . In particular, a lot of attention was paid to correlations.

In the last couple years, Feldman-Cousins starts to be used more, without my pushing. (Initially some at the LHC were very opposed, evidently because it could return two-sided interval not including zero when they really wanted a strict upper limit.)

Downward fluctuations in searches for excesses

Classic example: Upper limit on mean μ of Gaussian pdf for x.
Frequentist UL construction if $\mu \geq 0$ in model (σ=1) :



Frequentist 1-sided 95% C.L. Upper Limits, for α = 1–C.L. = 5%.

As observation x becomes increasingly negative, standard frequentist upper limit becomes small and then null.

For $x < -1.64\,\sigma$, the confidence interval is the *null* set!

Downward fluctuations in searches for excesses

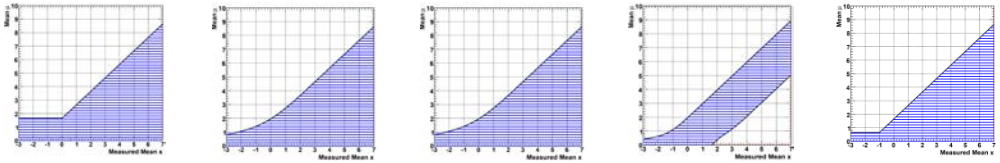
Issue acute 15-25 years ago in expts to measure ν_e mass in (tritium β decay): several measured $m_\nu^2 < 0$.

This is a *very* long story; see my "virtual talk"
http://www.physics.ucla.edu/~cousins/stats/cousins_bounded_gaussian_virtual_talk_12sep2011.pdf

and related post https://arxiv.org/abs/1109.2023 .

Contains intro to "Buehler's betting game", related to conditioning.

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



Bob Cousins

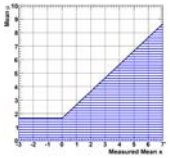
Univ. of California, Los Angeles

Virtual Talk

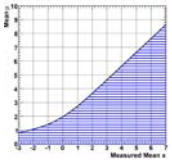
12 September 2011

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

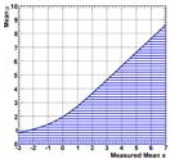
Five methods used for bounded Gaussian mean problem



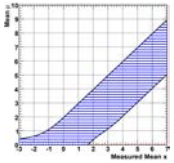
1) **1960's and beyond:**
   UL = max(x, 0) + 1.64σ



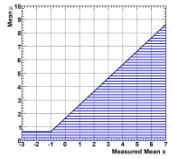
2) **1979 "PDG" (real 1986 PDG) and beyond:**
   Bayesian with uniform prior



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)

CL_S

The unfortunately named CL_S is the traditional frequentist one-tailed p-value for upper limits divided by another tail probability less than 1. The limits are thus (intentionally) conservative.

Brief definition with references in PDG RPP: http://pdg.lbl.gov/2011/reviews/rpp2011-rev-statistics.pdf .

Another technique that can be formulated in the language of statistical tests has been used to set limits on the Higgs mass from measurements at LEP [28,29]. For each value of the Higgs mass, a statistic called CL_s is determined from the ratio

$$CL_s = \frac{p\text{-value of signal plus background hypothesis}}{1 - p\text{-value of hypothesis of background only}} . \qquad (33.50)$$

The p-values in Eq. (33.50) are themselves based on a test statistic which depends in general on the signal being tested, *i.e.*, on the hypothesized Higgs mass.

In the usual procedure for constructing confidence intervals, one would exclude the signal hypothesis if the probability to obtain a value of CL_s less than the one actually observed is less than α. The LEP Higgs group has in fact followed a more conservative approach, and excludes the signal at a confidence level $1 - \alpha$ if CL_s itself (not the probability to obtain a lower CL_s value) is less than α. This prevents exclusion of a parameter value that could result from a statistical fluctuation in situations where one has no sensitivity, *e.g.*, at very high Higgs masses. The procedure results in a coverage probability that is in general greater than $1 - \alpha$. The interpretation of such intervals is discussed in Refs.[28,29].

A few more notes/history are in my arxiv post on these lectures.

CL_S inherits all issues of p-values

- What is new (non-standard statistics) in CL_S is combining the two p-values into 1 quantity.
  - This step is called "the CL_S criterion" in CMS papers.
- The p-values themselves have long existed in the statistics literature and should be designated that way (and not with the names that the inventor of named CL_S gave them). All the issues of p-values (choice of test statistic, how to eliminate nuisance parameters) of course still exist.
- LEP, Tevatron, and LHC Higgs combination groups differ in choices (!)
  - What specific likelihood ratio used in test statistic
  - Treatment of nuisance parameters
  - Ensembles used for "Toy M.C." used to get distribution of test statistic under H_0 (no Higgs) and H_1 (SM with Higgs)

My advocacy for >10 years:

Have in place tools to allow computation of results using a variety of recipes, for problems up to intermediate complexity:

- **Bayesian with analysis of sensitivity to prior**
- **Profile likelihood ratio (Minuit MINOS)**
- **Frequentist construction with approximate treatment of nuisance parameters**
- **Other "favorites" such as LEP's CL_S (an HEP invention)**

The community can (and should) then demand that a result shown with one's preferred method also be shown with the other methods, *and sampling properties studied*.

When the methods all agree, we are in asymptopic nirvana.
When the methods disagree, we are reminded that the results are answers to different questions, and we learn something! E.g.:

- **Bayesian methods can have poor frequentist properties**
- **Frequentist methods can badly violate likelihood principle**

Unsound statements you can now avoid*

- **"It makes no sense to talk about the probability density of a constant of nature."**
- **"Frequentist confidence intervals for efficiency measurements don't work when all trials give successes."**
- **"We used a uniform prior because this introduces the least bias." Or "a noninformative prioer since it contained no information."**
- **"The total number of events could fluctuate in our experiment, so *obviously* our toy Monte Carlo should let the number of events fluctuate."**
- **We used Delta-likelihood contours so there was no Gaussian approximation."**
- **"A five-sigma effect constitutes a discovery."**
- **"The confidence level tells you how much confidence one has that the true value is in the confidence interval."**
- **"We used the tail area under the likelihood function to measure the significance."**
- **"Statistics is obvious, so I prefer not to read the literature and just figure it out for myself."**

***Taken from real life**

References Cited in Talk Slides

Behnke13: O. Behnke et al., Data Analysis in High energy Physics, Wiley-VCH, 2013.

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.

Thanks again!

Thanks to many in HEP (Frederick James, Gary Feldman, Louis Lyons, Luc Demortier, + others numerous others) from whom I learned...

...many statisticians that Louis invited to PhyStat meetings -- for Bayesian statistics that was especially Jim Berger (multiple times) and Michael Goldstein...

...and to CMS Statistics Committee (Olaf Behnke et al.) for many discussions and comments on earlier versions of the slides...

...and to the authors of numerous papers from which I learned, including early (1980s) Bayesian papers by Harrison Prosper...

...and to Diego Tonelli of the LHCb experiment for comments on a earlier version of these sides.

Recommended reading

Books: I usually recommend the following progression, reading the first three cover-to-cover, and consulting the rest 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.

6) Recent book by HEP "experts": O. Behnke et al., Data Analysis in HEP, 2013

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; 7 for model selection)

BACKUP

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.

New Era for non-subjective Bayesian priors in HEP?

PHYSICAL REVIEW D **82,** 034002 (2010)

Reference priors for high energy physics

Luc Demortier
Laboratory of Experimental High Energy Physics, The Rockefeller University, New York, New York 10065, USA

Supriya Jain*
Homer L. Dodge Department of Physics and Astronomy, University of Oklahoma, Norman, Oklahoma 73019, USA

Harrison B. Prosper
Department of Physics, Florida State University, Tallahassee, Florida 32306, USA
(Received 4 February 2010; published 3 August 2010)

Bayesian inferences in high energy physics often use uniform prior distributions for parameters about which little or no information is available before data are collected. The resulting posterior distributions are therefore sensitive to the choice of parametrization for the problem and may even be improper if this choice is not carefully considered. Here we describe an extensively tested methodology, known as reference analysis, which allows one to construct parametrization-invariant priors that embody the notion of minimal informativeness in a mathematically well-defined sense. We apply this methodology to general cross section measurements and show that it yields sensible results. A recent measurement of the single-top quark cross section illustrates the relevant techniques in a realistic situation.

DOI: 10.1103/PhysRevD.82.034002 PACS numbers: 06.20.Dk, 14.65.Ha

So far use in HEP has been limited.

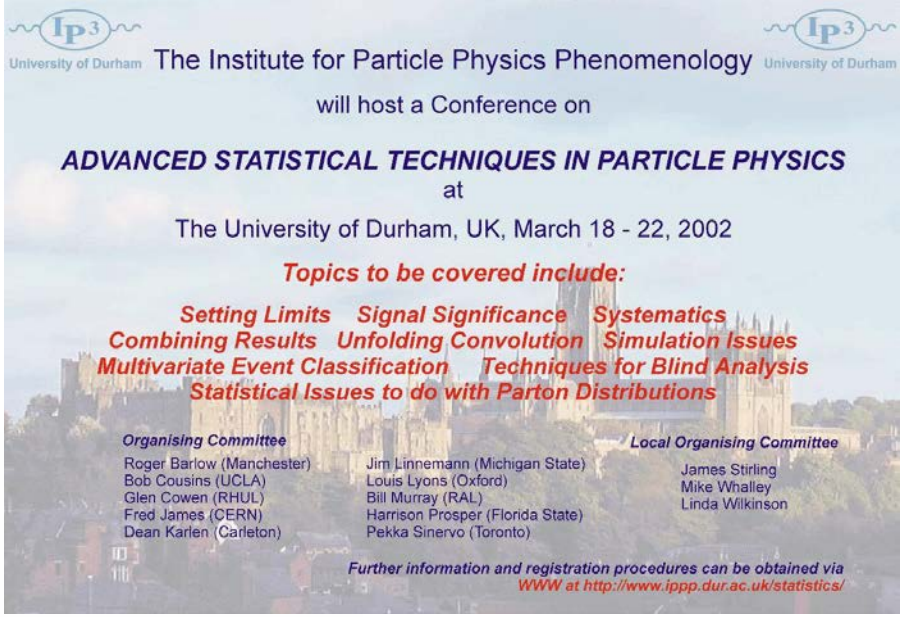
Sensitivity Analysis

An "objective Bayesian's" point of view:

"Non-subjective Bayesian analysis is just a part -- an important part, I believe – of a healthy *sensitivity* analysis to the prior choice…"

– J.M. Bernardo, J. Roy. Stat. Soc., Ser. B 41 113 (1979)

Sensitivity analysis: A subjective Bayesian's point of view:



WHY BE A BAYESIAN?

Michael Goldstein
Dept. of Mathematical Sciences, University of Durham,England

From the Proceedings: "…Again, different individuals may react differently, and the sensitivity analysis for the effect of the prior on the posterior is the analysis of the scientific community..."

From his transparencies:

"Sensitivity Analysis is at the heart of scientific Bayesianism."

Robert Kass's Questions for Classifying Kinds of Bayesians*

1. Is it important for Bayesian inferences to have good frequentist operating characteristics?
2. Does the Bayesian paradigm do anything more than produce candidate procedures, to be judged according to frequentist criteria?
3. Is there a useful role for default (a.k.a. "objective") Bayesian inferences as representing approximately subjective inferences?
4. Is it possible to interpret default Bayesian inference as anything other than approximately subjective?
5. Assuming that the data analyst has done a thorough and careful job, is it appropriate to interpret default Bayesian inferences as representing, approximately, what any reasonable person ought to think given the data and appropriate background information?
6. Is there any useful meaning to the word "objective," beyond signifying such overwhelming evidence that reasonable people will be forced to agree.
7. Is the word "objective" so easy to misunderstand that its utility in the context of Bayesian inference is, on average, negative?
8. Is it important to distinguish scientific inference from decision-making?
9. Are there scientific settings in which formal elicitation procedures are useful?

*** "Kinds of Bayesians (Comment on articles by Berger and by Goldstein)", Bayesian Analysis 1 437 (2006)**

Clopper and Pearson's construction

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! (See next slide for details.)

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

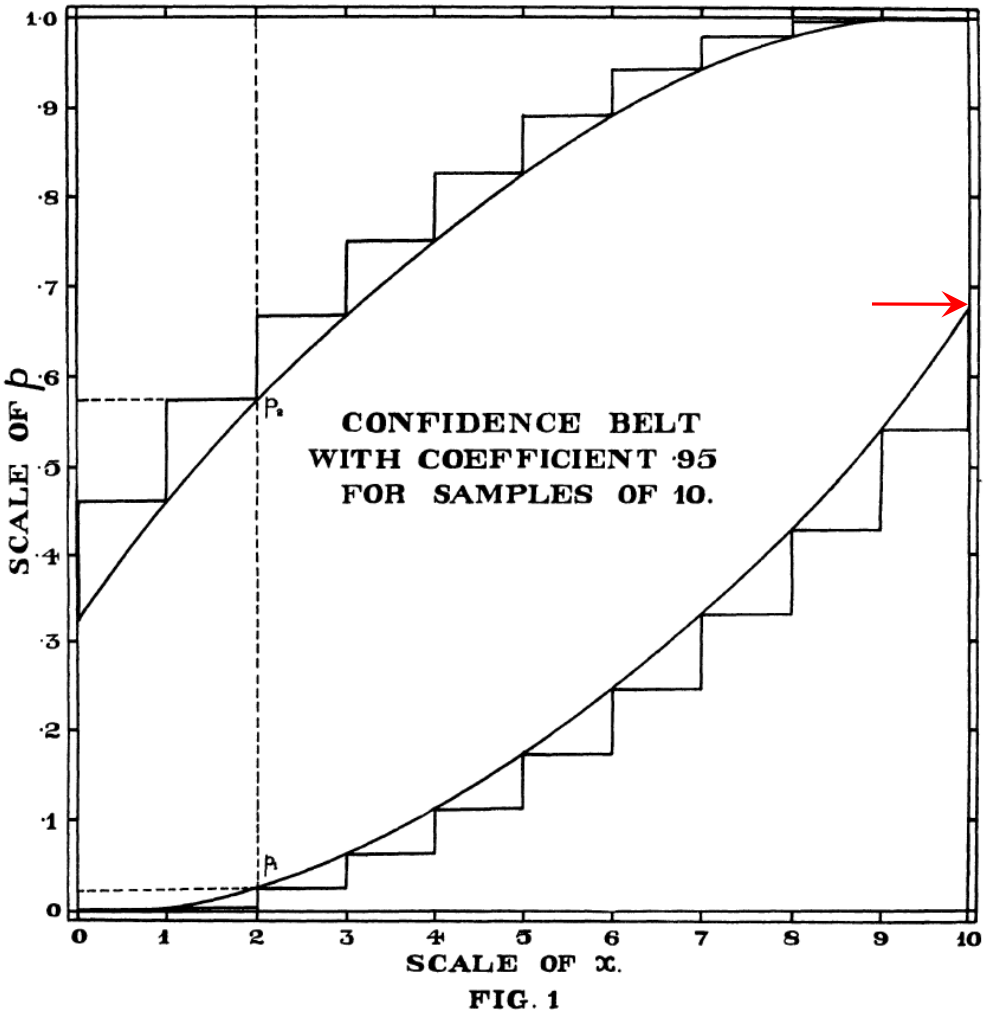


FIG. 1

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

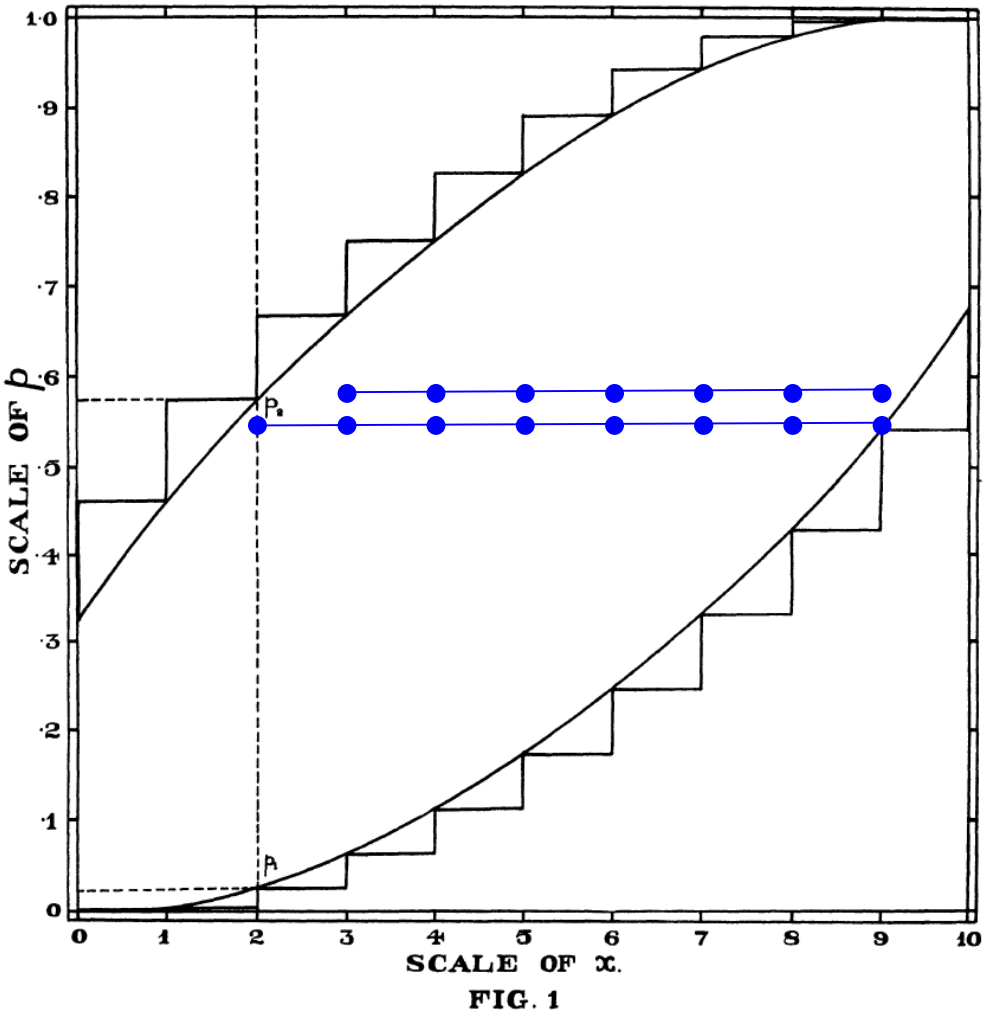
Clopper and Pearson's construction (cont.)

Partial details of construction:

Blue lines are two of the acceptance intervals having central 95% or more probability, 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.



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 probability α can be calculated, since the Type II error probability 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!

For discussion of g.o.f., see
http://cousins.web.cern.ch/cousins/ongoodness6march2016.pdf

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 Aslan and Zech, https://arxiv.org/abs/math/0207300 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).

Useful 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 on vertical axis and horizontal planes for 2D vectors of observables $E = (x_1, x_2)$.

Prior to experiment, acceptance regions $\mathcal{A}(\theta_1)$ in E-space planes are determined for each θ_1 (needs ordering principle) with $P(E \in \mathcal{A}(\theta_1)) = C.L.$

E′ is data actually observed in expt. Upon obtaining E′, confidence interval for θ_1 consists of all values of θ_1 for which E′ is in $\mathcal{A}(\theta_1)$.

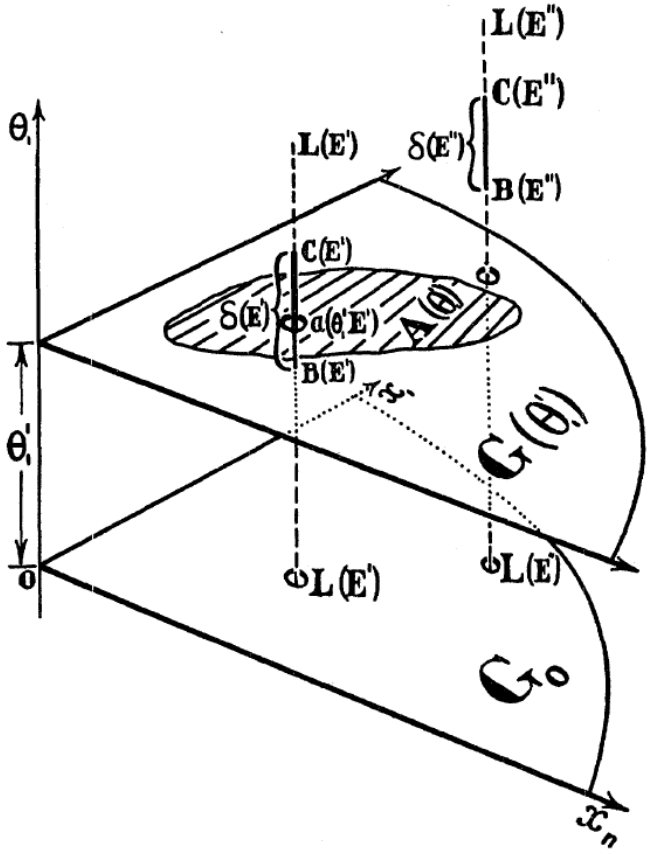


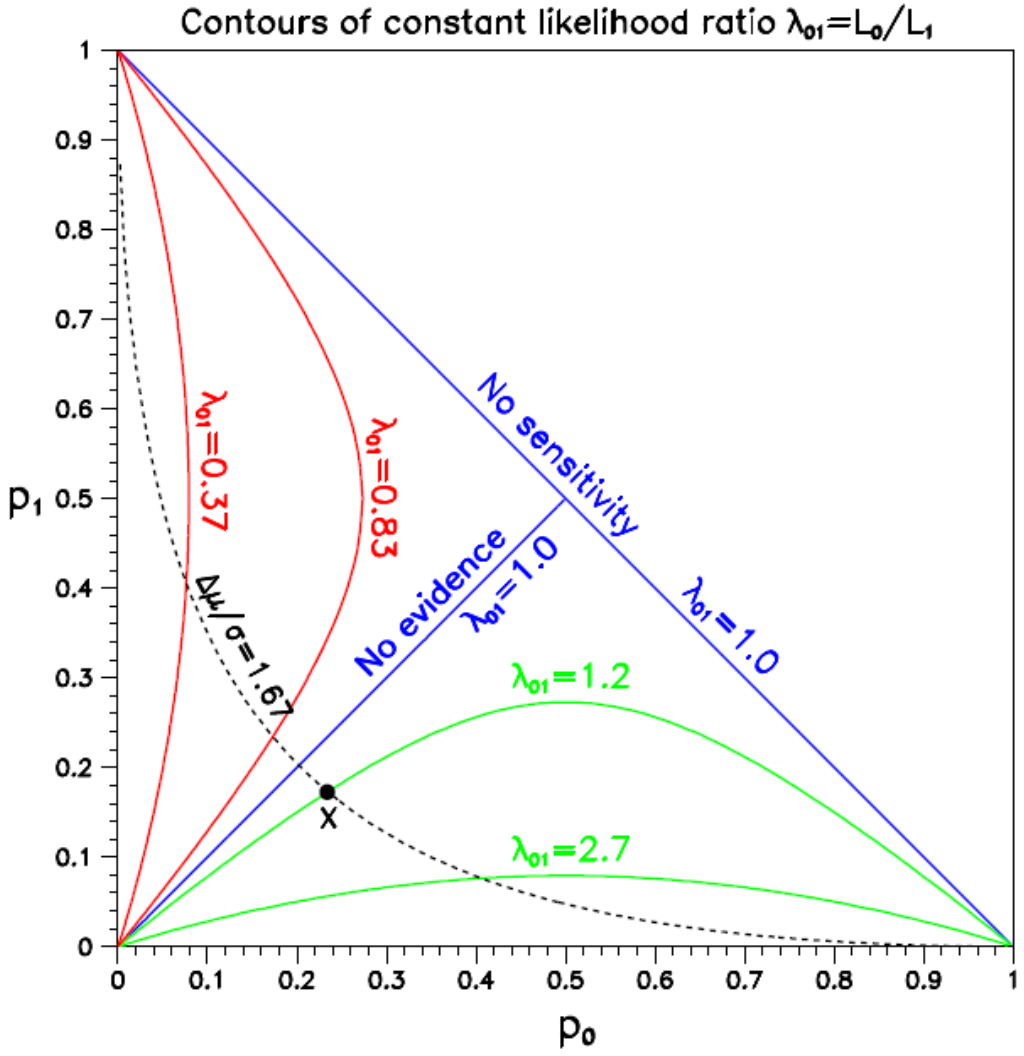
Fig. 1—The general space G.

Luc Demortier and Louis Lyons, "Testing Hypotheses in Particle Physics: Plots of p_0 versus p_1"

Test of point null vs point alternative, two Gaussians with same σ, peak separation Δμ.

At a glance can see that contours of constant λ_{01} are completely different topology from contours of e.g. p_0.

For rest of plot, you will have to read their paper or stare at it for a long time.
http://arxiv.org/abs/1408.6123



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.

Approximate Confidence Regions Using $\Delta(-\ln\mathcal{L})$
(included in appendix to MINUIT users guide)

Computer Physics Communications 20 (1980) 29–35

INTERPRETATION OF THE SHAPE OF THE LIKELIHOOD FUNCTION AROUND ITS MINIMUM

F. JAMES
Data Handling Division, CERN, Geneva, Switzerland

It often happens that the solution of a minimum problem is itself straightforward, but the calculation or interpretation of the resulting parameter uncertainties, as determined by the shape of the function at the minimum, is considerably more complicated. The purpose of this note is to clarify the most commonly encountered difficulties in parameter error determination. These difficulties may arise in connection with any fitting program, but will be discussed here with the terminology of the program MINUIT for the convenience of MINUIT users.

The most common causes of misinterpretation may be grouped into three categories:

1. Proper normalization of the user-supplied chi-square or likelihood function, and appropriate ERROR DEF.
2. Non-linearities in the problem formulation, leading to different errors being calculated by different techniques, such as MIGRAD, HESSE and MINOS.
3. Multiparameter error definition and interpretation.

All these topics are discussed in some detail by Eadie et al., which may be consulted for further details.

Table 1
Table of UP for multiparameter confidence regions

Number of parameters	Confidence level (probability contents desired inside hypercontour of "$chi^2 = chi^2$ min + UP")				
	50%	70%	90%	95%	99%
1	0.46	1.07	2.70	3.84	6.63
2	1.39	2.41	4.61	5.99	9.21
3	2.37	3.67	6.25	7.82	11.36
4	3.36	4.88	7.78	9.49	13.28
5	4.35	6.06	9.24	11.07	15.09
6	5.35	7.23	10.65	12.59	16.81
7	6.35	8.38	12.02	14.07	18.49
8	7.34	9.52	13.36	15.51	20 09
9	8.34	10.66	14.68	16.92	21.67
10	9.34	11.78	15.99	18.31	23.21
11	10.34	12.88	17.29	19.68	24.71

If FCN is –log(likelihood) instead of chi-square, all values of UP should be divided by 2.

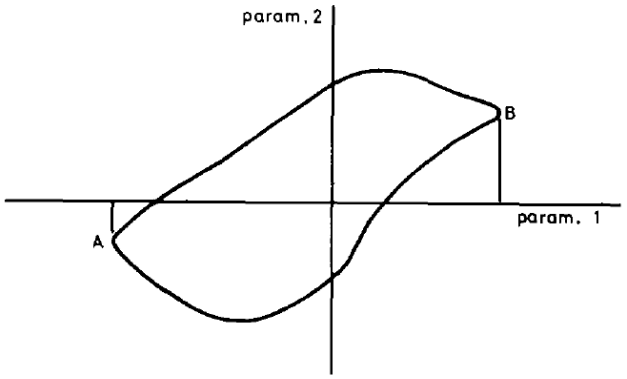


Fig. 1. MINOS errors for parameter 1.

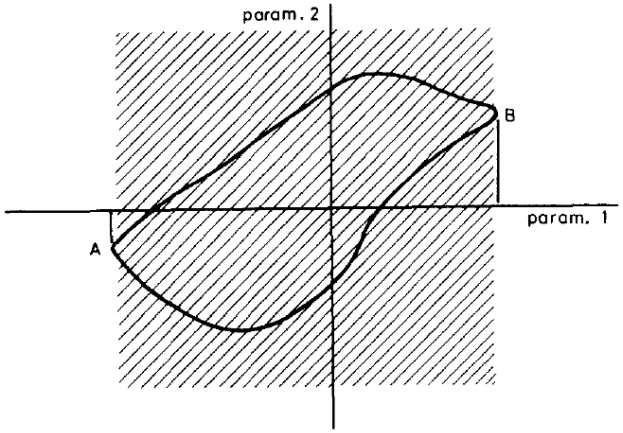


Fig. 2. MINOS error confidence region for parameter 1.

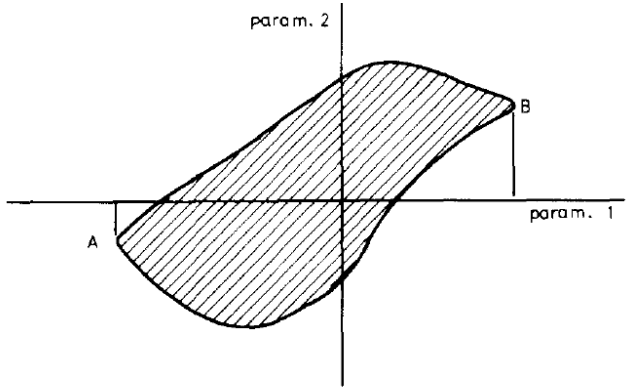


Fig. 4. Optimal confidence region for parameters 1 and 2.

Example of D.R. Cox (cont.)

Demortier gives details on how average interval length is shorter in the HEP example. Here I give Cox's discussion.

E.g., if one is testing $\mu=0$ vs $\mu=\mu_1$, with μ_1 roughly the size of σ_1 (the larger σ), consider the following 68% CL intervals:

- **$\hat{\mu} \pm (0.48)\sigma_1$ if Device #1 used (covers true μ in 37% of uses)**
- **$\hat{\mu} \pm 5\sigma_2$ if Device #2 used (covers true μ nearly 100% of uses)**

So true μ is covered in (37/2 + 100/2)% = 68% of all intervals! Unconditional (full sample space) coverage is correct, but conditional coverage is not.

Due to the smallness of σ_2 , average length of all intervals is smaller conditional intervals with independent coverage.

One gives up power with Device #1 and uses it in Device #2.

Cox: "If, however, our object is to say `what can we learn from the data that we have', the unconditional test is surely no good."

These examples reveal a real conflict between N-P optimization for power and conditioning to optimize relevance.

Look-Elsewhere Effect

In these lectures, I did not have time for the LEE.

A starting point for self-study is the discussion in:

Louis Lyons, "Comments on 'Look Elsewhere Effect' ".

https://users.physics.ox.ac.uk/~lyons/LEE_feb7_2010.pdf .

See also Section 9.2 of my paper on the Jeffreys-Lindley Paradox, https://arxiv.org/abs/1310.3791 .

An important paper is

Eilam Gross, Ofer Vitells, "Trial factors for the look elsewhere effect in high energy physics," https://arxiv.org/abs/1005.1891

Sufficiency, Conditionality, Likelihood Principles

There is a lot more to the Likelihood Principle than I had time to discuss. I omitted the important (frequentist) concept of a "sufficient statistic", due to Fisher. This is a way to describe data reduction without loss of relevant information. E.g., for testing a binomial parameter, one needs only the total numbers of successes and trials, and not the information on exactly which trials had successes. See Stuart99 for math definitions.

The sufficiency principle says (paraphrasing – there are strong and weak forms) that if the observed values of the sufficient statistic in two experiments are the same, then they constitute equivalence evidence for use in inference.

Birnbaum famously argued (1962) that the Conditionality Principle and the Sufficiency Principle imply the Likelihood Principle. Controversy continues. For recent discussion, see D. Mayo (2014), https://projecteuclid.org/euclid.ss/1408368565#toc , with comments by six statisticians and rejoinder.

Point Estimation

Point Estimation

Most of these slides are about intervals – I have not yet much about what to quote as the "measured value". Statisticians call this the "point estimate".

- **There is a huge literature on point estimation – see e.g. Ch. 7 and 8 in James06.**
- **If you are an expert on interval estimation, one approach is to use that machinery to get a point estimate.**
  - **E.g., one might take the mid-point of (say) your 68% C.L. central interval. But a better approach is probably to let the C.L. go to 0, so that your interval gets shorter and shorter, and use the limiting point. E.g. for likelihood ratio intervals, this results in the Maximum Likelihood estimate.**
- **But to give you an idea of how rich the subject is, I show a few interesting things from James06.**

Point Estimation: Traditional Desiderata

- **Consistency: Estimate converges toward true value as number of observations N increases**
- **Unbiasedness: Expectation value of estimate is equal to the true value.**
- **Efficiency: Estimate has minimum variance**
- **Minimum loss of information: (technical definition)**
- **Robustness: Insensitivity to departures from the assumed distribution**

One can add:

- **Simplicity: transparent and understandable**
- **Minimum computer time: still relevant in onlline applications, less relevant otherwise**
- **Minimum loss of physicist's time (how much weight to put on this?)**

Bias and consistency are independent properties

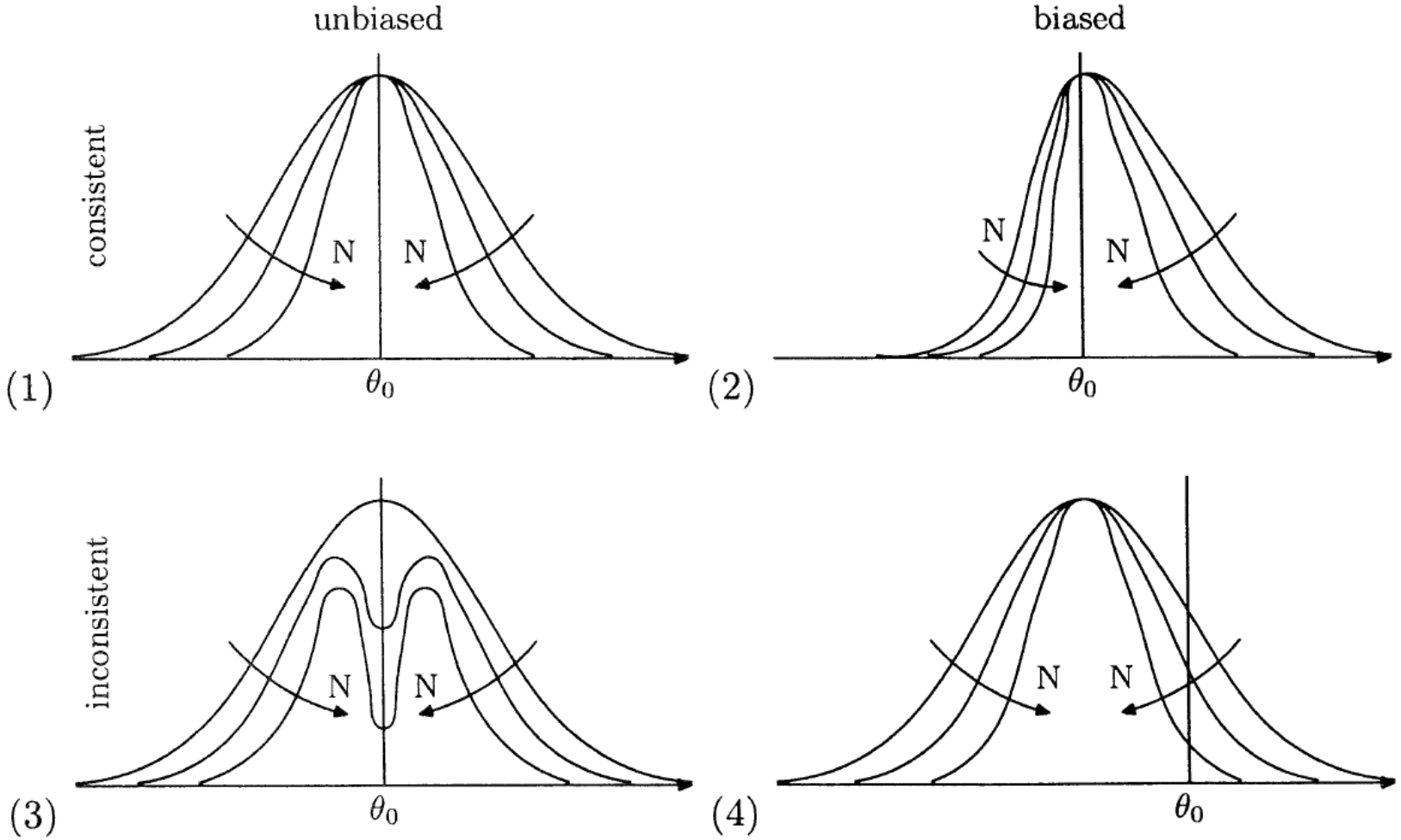


Fig. 7.2. Examples of probability density functions $f(\hat{\theta}, \theta_0)$ (not normalized) with different combinations of consistency and bias. The arrows show the direction of increasing N.

James06, p. 130

BUT (!) Other desired properties can be impossible to achieve simultaneously

- **How to choose? A thorough analysis requires further input: what are the costs of not incorporating various desiderata? Then formal *decision theory* can be used to choose estimator.**
- **In practice in HEP, *Maximum Likelihood* estimates are often used (even though they are typically not unbiased).**
  - **Consistent**
  - **Other excellent asymptotic properties (estimate is asymptotically normal)**
  - **For finite N, works well in so-called exponential family (includes Poisson, Gaussian, binomial)**
  - **Invariant under reparameterization**

Simple example illustrating diversity of point estimators (James06, p. 209)

- If $p(x|\mu) = f(x-\mu)$, where f is some pdf, then μ is called a *location parameter*. Common examples are:
  - Normal: $p \sim \exp(-(x-\mu)^2/2\sigma^2)$
  - Uniform: p = constant for $|x-\mu|<a$; p=0 otherwise
  - Cauchy: $p \sim 1/(a^2 + (x-\mu)^2)$
  - Double exponential: $p \sim \exp(-a |x-\mu|)$
- These examples are all symmetric about μ: $p(\mu+y) = p(\mu-y)$
- Suppose you are given N=11 values of x randomly sampled from $p(x|\mu)$. What estimator (function of the 11 values) gives you the "best" estimate of μ ?
- If by "best" you mean minimum variance, it is the M.L. estimate, resulting in a different formula for each!

Minimum Variance Location Estimator

Normal	**Sample mean (L_2)**
Uniform	**Midrange: mean of extreme values (L_∞)**
Cauchy	**M.L. estimate (no simple formula)**
Double-exponential	**Median: middle value (L_1)**

Three of the four are special cases of L_p , the estimator that minimizes the sum over the observations of $|x_i - \mu|^p$.
Different values of p put different emphasis on observations in the tails.
If true distribution departs from that assumed, estimate of location is no longer optimal. Sensitivity is in tails!
See nice discussion of asymptotic variance and robustness in James06, pp. 211 ff.