Advanced Statistics for High Energy Physics

Bob Cousins, UCLA and CMS

Fourth CERN-Fermilab

Hadron Collider Physics Summer School

June 8, 2009

Preface

- Many of us teach advanced “data analysis” courses that last an entire academic term. What to say in two hours?
- I decided to concentrate on the “theoretical” underpinnings, i.e., on *what you must know in order to choose the right methods*, rather than on the practical implementation (details on using a particular tool). I will have a few slides on that as well.
- My rationale is that the harder part is *choosing the right methods*. C++ code with tutorials for implementing most methods exists and is becoming well documented.
- 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.

Preface (cont.)

- **But, there is no doubt that this is a dense talk – you will not pick it all up in real time (or you will be unique if you do!). It should however be extremely useful to you to study this talk, referring to the references, until you have a comfort level with the content.**
- ***Thanks* to all who helped me learn about statistical inference, beginning with Fred James over 25 years ago! (…and to Luc Demortier, Louis Lyons, and Harrison Prosper for comments on earlier versions of many of these slides).**

Outline

- Why foundations matter
- Probability, Bayes' Theorem, Probability density, Likelihood
- An note re decisions
- Simple Bayesian methods and issues
  - Subjective vs non-subjective priors
  - Sensitivity analysis
  - Priors in high dimensions
- Frequentist confidence intervals and issues
  - Neyman's construction, coverage, binomial example
- Classical hypothesis testing, duality with confidence intervals
- Goodness of fit
- Likelihood (ratio) intervals and issues
- Nuisance parameters
- Likelihood Principle
- Conditioning
- Hybrid methods: Introduction to pragmatism
- Brief note on multivariate methods
- RooStats for LHC
- Unsound statements you can now avoid

…or, more specifically, the goal of these lectures is that you understand this table on Slide 59…

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

…and that you will see what is wrong with statements like these* on slide 70:

- **“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.”**
- **“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.”**

***References available on request**

Why Foundations Matter

- In the “final analysis”, we may make approximations, take a pragmatic approach, or follow a convention. To inform such actions, it is important to have some understanding of the 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.
  In statistical inference, however, *foundational differences result in different answers*, and therefore 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 again after the resurgence of Bayesian methods in recent decades.
  I will attempt to summarize some of the things we should understand from that debate, using simple cases as examples. *Most importantly*: understand both sides!

Definitions

Much of this talk is devoted to definitions.

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 the statement,

> "The confidence level tells you how much confidence one has that the true value is in the confidence interval,"

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."

Definition of "Probability"

- Abstract mathematical probability P can be defined in terms of sets and axioms that P obeys. If the axioms are true for P, then P obeys Bayes' Theorem (see next slide)

  P(B|A) = P(A|B) P(B) / P(A).

- Two established* incarnations of P are:

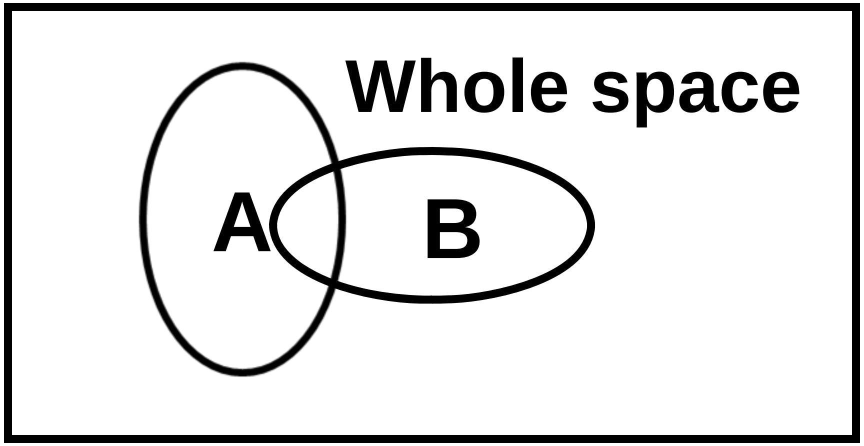
  1) *Frequentist P*: limiting frequency in ensemble of imagined repeated samples (as usually taught in Q.M.).
  P(constant of nature) and P(SUSY is true) do not exist (in a useful way) for this definition of P (at least in one universe).

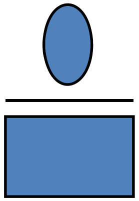
  2) *(Subjective) Bayesian P*: subjective (personalistic) *degree of belief.* (de Finetti, Savage)
  P(constant of nature) and P(SUSY is true) exist for You.
  Shown to be basis for coherent personal decision-making.

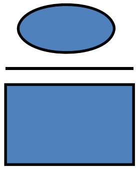
- ***It is important to be able to work with either definition of P, and to know which one you are using!***

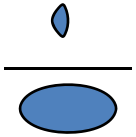
*Of course they are still argued about, but to less practical effect, I think.

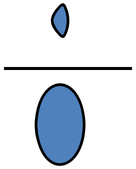
P, Conditional P, and Derivation of Bayes' Theorem in Pictures

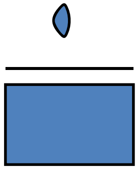


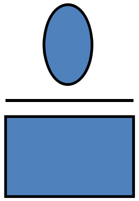
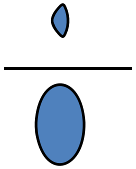
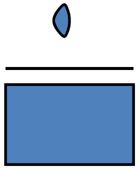
$P(A) =$ 

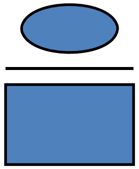
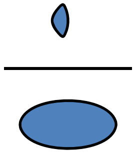
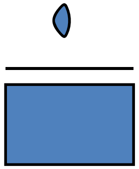
$P(B) =$ 

$P(A|B) =$ 

$P(B|A) =$ 

$P(A \cap B) =$ 

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

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

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

What is the “Whole Space”?

- Note that for probabilities to be well-defined, the “whole space” needs to be defined, which in practice introduces assumptions and restrictions.
- Thus the “whole space” itself is 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 subspace can sometimes improve the quality of statistical inference – see the discussion of “Conditioning” in later slides.
- 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, please 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(b-jet | btag) $\propto$ P(btag |b-jet) P(b-jet)

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) = $3^0e^{-3}/0!$ = 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

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

$P(0 \text{ events} \mid \text{model false}) = 1.0$

$P(>0 \text{ events} \mid \text{model true}) = 0.95$

$P(>0 \text{ events} \mid \text{model false}) = 0.0$

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:

$P(\text{model true} \mid 0 \text{ events}) \propto P(0 \text{ events} \mid \text{model true})\, P(\text{model true})$

If "model" is S.M., then still very high degree of belief after experiment! (Compare with news releases that would say "there is 5% chance the S.M. is true.")

If "model" is large extra dimensions, then low prior belief becomes even lower.

N.B. Of course this example is over-simplified; it gets more interesting when there is more than one model which predicts the signal-type events. Also when an event is seen; and when normalization factor is included.

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* whether or not to take an action (making a press release, or planning your 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* whether or not to take an action (making a press release, or planning your 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 *cost* function), which gives the relative costs (to You) of a Type I error (declaring model false when it is true) and a Type II error (not declaring model false when it is false).

Thus, Your *decision*, such as where to invest your time or money, 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. *One thing should become clear: classical "hypothesis testing" is not a complete theory of decision-making!*

Probability, Probability Density, and Likelihood

- **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 n=3 into $P(n|\mu)$ yields the *likelihood function* $\mathcal{L}(\mu) = \mu^3 \exp(-\mu)/3!$**
  - **Key point is that $\mathcal{L}(\mu)$ is *not* a probability density in μ. (It is not a density!) Area under $\mathcal{L}$ is meaningless. That's why a new word, "likelihood", was invented for this function of the parameter(s), to distinguish from a pdf in the observable(s)!**
  - **Likelihood *Ratios* $\mathcal{L}(\mu_1) /\mathcal{L}(\mu_2)$ are useful and frequently used.**

Change of variable x, change of parameter θ

- For pdf $p(x|\theta)$ and (1-to-1) change of variable from x to y(x):

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

  Jacobian modifies probability *density*, guaranties that

  $$P(\, y(x_1) < y < y(x_2)\,) \;=\; P(x_1 < x < x_2\,),$$ i.e., that

  Probabilities are invariant under change of variable x.

  - Mode of probability *density* is *not* invariant (so, e.g., criterion of maximum probability density is ill-defined).
  - Likelihood *ratio* is invariant under change of variable x. (Jacobian in denominator cancels that in numerator).

- For likelihood $\mathcal{L}(\theta)$ and reparametrization from θ to $u(\theta)$:

  $$\mathcal{L}(\theta) \;=\; \mathcal{L}(u(\theta)) \quad (!).$$

  - Likelihood $\mathcal{L}(\theta)$ is invariant under reparametrization of parameter θ (reinforcing fact that $\mathcal{L}$ is *not* a pdf in θ).

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)

Given continuous x ∈ (a,b), and its pdf p(x), let

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

Then y ∈ (0,1) and p(y) = 1 (uniform) for all y. (!)

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

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

The specification of a Bayesian prior pdf p(μ) for parameter μ is equivalent to the choice of the metric f(μ) in which the pdf is uniform. This is a *deep* issue, not always recognized as such by users of flat prior pdf's in HEP!

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

Bayes' Theorem Generalized to Probability Densities

Recall $P(B|A) \propto P(A|B)\, P(B)$.

For Bayesian P, parameters are random variables which can appear in conditional probabilities.

Let probability density function $p(x|\mu)$ be 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 :

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

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

$p(\mu)$ = prior pdf for μ, before incorporating the results of this experiment

Note that there is one (and only one) probability density in μ on each side of the equation, again consistent with the likelihood *not* being a density.

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

- **Upon obtaining $p(\mu|x_0)$, the *credibility* of μ being in any interval can be calculated by integration.**
- **It is common to use the posterior mode as the estimate of μ, even though it depends on the often-arbitrary choice of metric. (Median is metric-independent but only exists in 1D.) Since the Jacobian moves the mode around under change of variable (say from decay rate Γ to lifetime $1/\Gamma$), care must be used to interpret it.**
- **To make a *decision* as to whether or not μ is in an interval or not (e.g., whether or not $\mu>0$) , one requires a further subjective input: the cost function (or utility function) for making wrong decisions.**
- **More generally, point estimation can be cast as a decision, with a cost function (J. Bernardo), but beyond scope here.**

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

- **There are compelling arguments (Savage, De Finetti) that Bayesian reasoning with 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?**
- **A bright idea, vigorously pursued by physicist Harold Jeffreys in in mid-20th century: *Can one define a prior p(μ) which contains as little information as possible, so that the posterior pdf is dominated by the likelihood?***
- **The really *really* thoughtless idea*, recognized by Jeffreys as such, but dismayingly common in HEP: just choose p(μ) uniform in whatever metric you happen to be using!**

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

"Uniform Prior" Requires a Choice of Metric

- 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 $p(\mu)$?" is equivalent to the question, "For what function $y(\mu)$ is $p(y)$ uniform?"
- "Jeffreys Prior" answers the question using a prior uniform in a metric related to the Fisher information (technical term).
  - Poisson signal mean μ, no background: $p(\mu) = 1/\mathrm{sqrt}(\mu)$
  - Poisson signal mean μ, mean background b: $p(\mu) = 1/\mathrm{sqrt}(\mu+b)$
  - Unbounded mean μ of gaussian: $p(\mu) = 1$
  - RMS deviation of a Gaussian when mean fixed: $p(\sigma) = 1/\sigma$
  - Binomial parameter ρ, $0 \leq \rho \leq 1$: $p(\rho) = \rho^{-1/2}(1 - \rho)^{-1/2} = \mathrm{Beta}(1/2,1/2)$
- Jeffreys prior yields pdfs which are consistent under transformation into different metrics.
- By a different invariance argument, one can infer $p(\mu) = 1/\mu$ for Poisson mean! (There is some arbitrariness in choice of group under which to be invariant.)

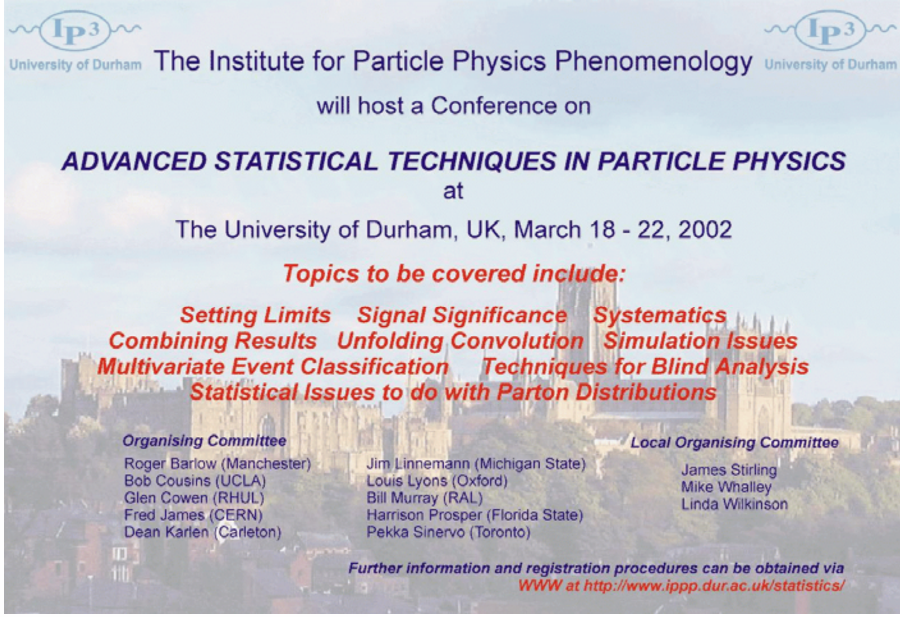
What to call Non-Subjective Priors?

- **Objective priors?**
- ***Non*informative priors? *Un*informative priors?**
- **Vague priors? Ignorance priors?**
- **Reference priors? (Unfortunately also refers to a specific recipe of Bernardo)**
- **Kass and Wasserman (Kass96), who have compiled a list of them, give the best (neutral) name in my opinion: *Priors selected by “formal rules”.***
  - ***Required reading for anyone using Bayesian methods!***
- **Whatever the name, keep in mind that choice of prior in one metric determines it in all other metrics: be careful in the choice of metric in which it is uniform!**
- **N.B. When professional statisticians refer to “flat prior”, they usually mean the Jeffreys prior.**

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.
- 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:



University of Durham The Institute for Particle Physics Phenomenology University of Durham

will host a Conference on

ADVANCED STATISTICAL TECHNIQUES IN PARTICLE PHYSICS
at
The University of Durham, UK, March 18 - 22, 2002

Topics to be covered include:

Setting Limits Signal Significance Systematics
Combining Results Unfolding Convolution Simulation Issues
Multivariate Event Classification Techniques for Blind Analysis
Statistical Issues to do with Parton Distributions

Organising Committee

Roger Barlow (Manchester)
Bob Cousins (UCLA)
Glen Cowen (RHUL)
Fred James (CERN)
Dean Karlen (Carleton)

Jim Linnemann (Michigan State)
Louis Lyons (Oxford)
Bill Murray (RAL)
Harrison Prosper (Florida State)
Pekka Sinervo (Toronto)

Local Organising Committee

James Stirling
Mike Whalley
Linda Wilkinson

Further information and registration procedures can be obtained via
WWW at http://www.ippp.dur.ac.uk/statistics/

WHY BE A BAYESIAN?

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

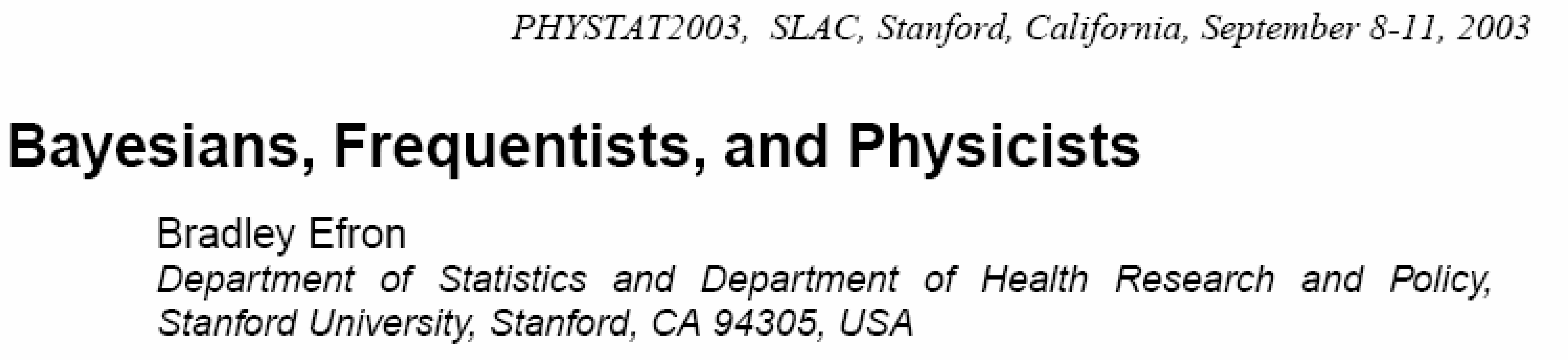
From the Proceedings: "…Again, different individuals may react differently, and the sensitivity analysis for the effect of the prior on the posterior is the analysis of the scientific community..."

From his transparencies:

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

Priors in high dimensions

- Is there a sort of informational phase space which leads 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 effect).
- For >1D, Jeffreys Prior has problems already known to Jeffreys, even with two parameters.
- State of the art for "objective" priors may be so-called "reference priors" of Bernardo, but tools seem to be lacking.
- Subjective priors also very difficult to construct in high dimensions: human intuition is poor.
  - 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".



PHYSTAT2003, SLAC, Stanford, California, September 8-11, 2003

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."

What Can Be Computed without Using a Prior?

Not P(constant of nature | data).

1) *Confidence Intervals* for parameter values, as defined in the 1930’s by Jerzy Neyman.
2) Likelihood *ratios*, the basis for a large set of techniques for point estimation, interval estimation, and hypothesis testing.

These can both be constructed using the frequentist definition of P.

I’ll introduce them, and then compare and contrast them with Bayesian methods.

Confidence Intervals

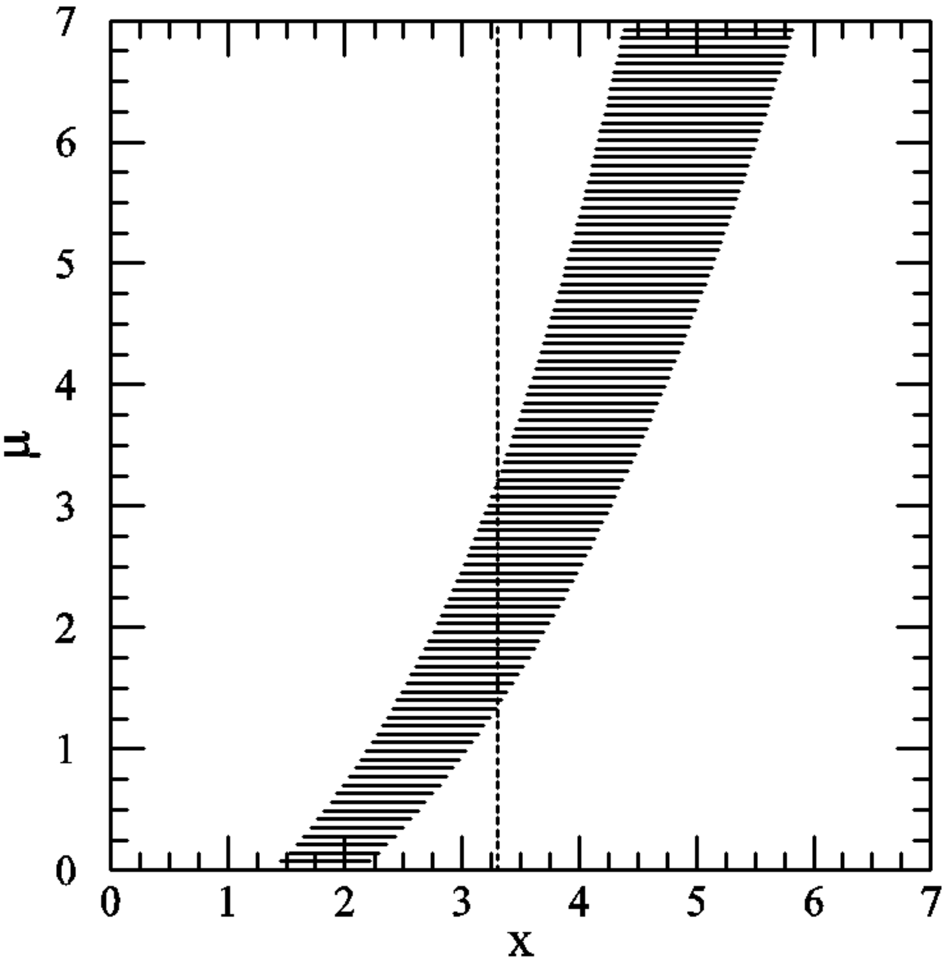
- **“Confidence intervals”, and this phrase to describe them, were invented by Jerzy Neyman in 1934-37. While statisticians mean Neyman’s intervals (or an approximation) when they say “confidence interval”, in HEP the language tends to be a little loose.**
- **I highly recommend using “confidence interval” only to describe intervals corresponding to Neyman’s construction (or good approximations thereof), described below.**
- **The next five slides contain the crucial information, but you will want to cycle through them a few times to “take home” how the construction works, since it is really ingenious – perhaps a bit *too* ingenious given how often confidence intervals are misinterpreted.**
- **In particular, you will understand that the confidence level does *not* tell you “how confident you are that the unknown true value is in the interval” – only a *subjective* Bayesian credible interval has that property!**

Neyman’s Confidence Interval construction

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

Upon performing an experiment to measure x and 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.

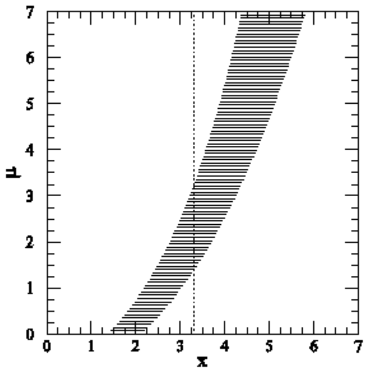


Note: x and μ need not have the same range, units, or (in generalization to higher dimensions) dimensionaliity!

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

Aside on the note regarding *x* and μ

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



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

After studying examples such as those, one learns that in the Gaussian "measurement" of a mass μ which obtains the value *x*, it is crucial to distinguish between the data *x*, which can be negative, and the mass μ, for which negative values do not exist in the model. (I.e., for which P(*x*|μ) does not exist)

Confidence Intervals and Coverage

- **Recall: in math, one defines a *vector space* as a set with certain properties, and then the definition of a *vector* is "an element of a vector space". (A vector is not defined in isolation.)**
- **Similarly, whether constructed in practice by Neyman's construction or some other technique, a *confidence interval* is defined to be "a element of a confidence set", where the *confidence set* is a set of intervals defined to have the property of frequentist *coverage* under repeated sampling:**

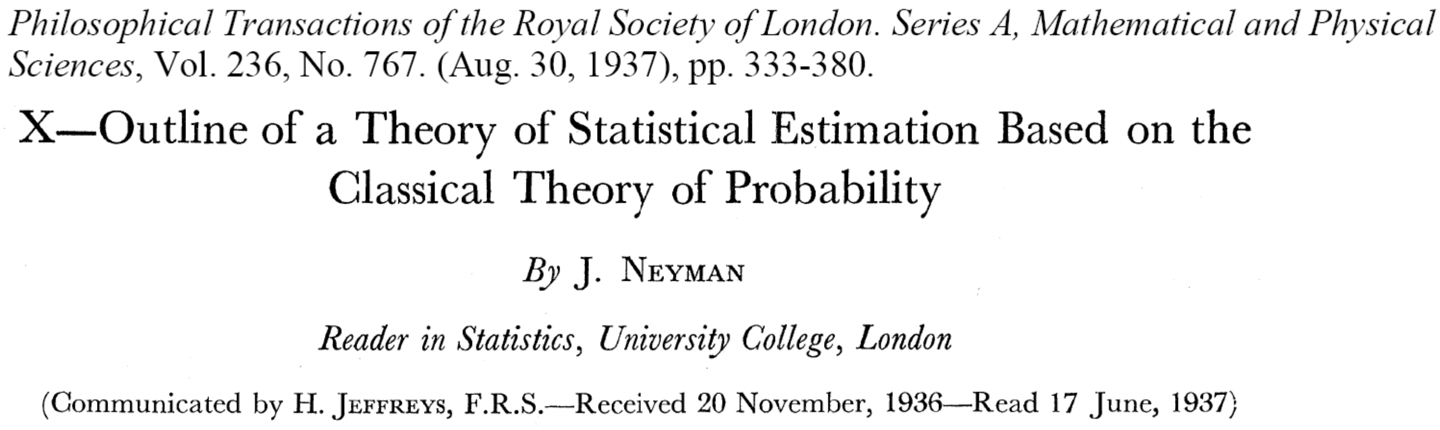
Confidence Intervals and Coverage (cont.)

- **Let the unknown true value of μ be μ_t . In repeated experiments, the confidence intervals obtained will have different endpoints [μ_1, μ_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 contruction will contain ("cover") the fixed but unknown μ_t .

  I.e., $P(\mu_t \in [\mu_1, \mu_2]) = \text{C.L.} = 1 - \alpha$.
- **One of the complaints about confidence intervals is that the consumer often forgets (if he or she ever knew) that the random variables in this equation are μ_1 and μ_2, and not μ_t, and that coverage is a property of the set, not of an individual interval!**

  Please don't forget!
- **It *is* true (in precisely the sense defined by the ordering principle used in the Neyman construction) that the confidence interval consists of those values of μ for which the observed *x* is among the most probable to be observed.**



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

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

By J. Neyman

Reader in Statistics, University College, London

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

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

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

E′ is data actually observed in expt.

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

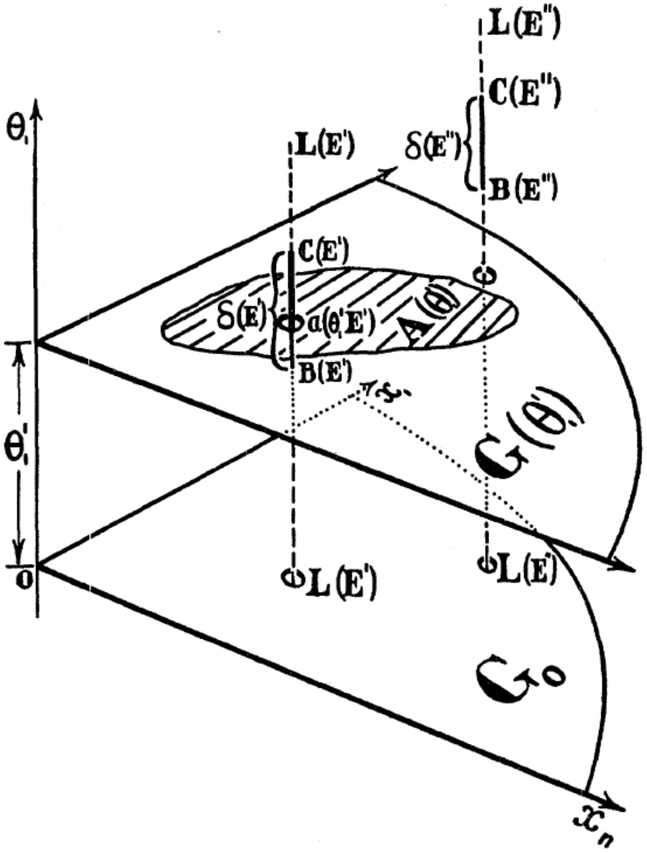
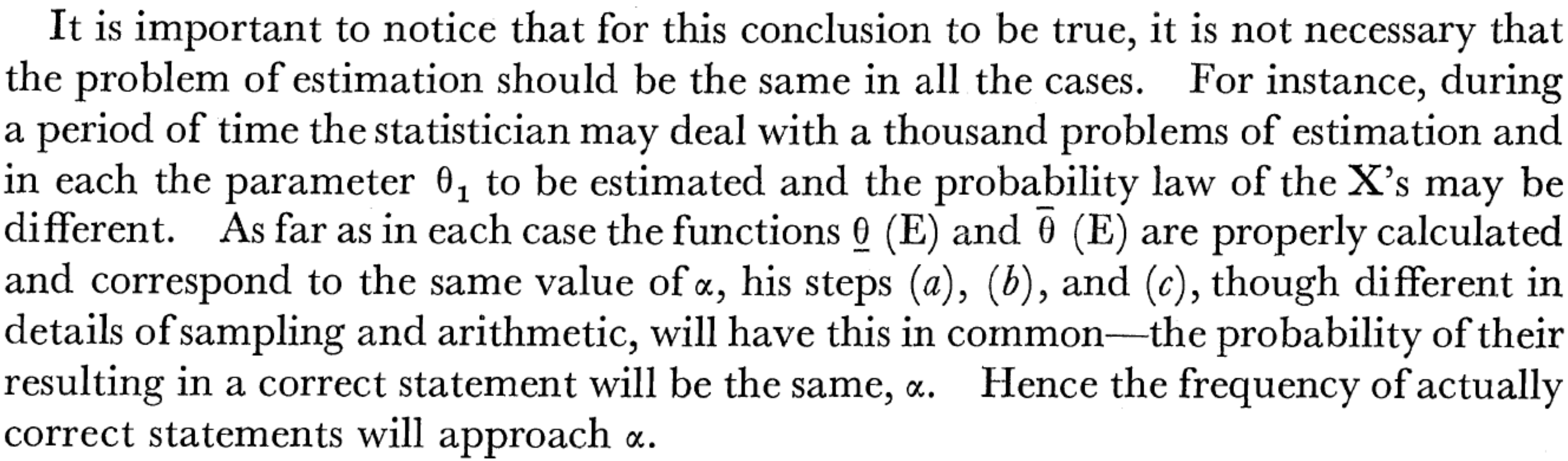


Fig. 1—The general space G.

Coverage: The experiments in the ensemble do not have to be the same.

Neyman pointed this out in his 1937 paper (in which his α is the modern 1 - α):



It is important to notice that for this conclusion to be true, it is not necessary that the problem of estimation should be the same in all the cases. For instance, during a period of time the statistician may deal with a thousand problems of estimation and in each the parameter θ_1 to be estimated and the probability law of the X's may be different. As far as in each case the functions $\underline{\theta}$ (E) and $\bar{\theta}$ (E) are properly calculated and correspond to the same value of α, his steps (*a*), (*b*), and (*c*), though different in details of sampling and arithmetic, will have this in common—the probability of their resulting in a correct statement will be the same, α. Hence the frequency of actually correct statements will approach α.

Famous Early Example of Clopper and Pearson: Confidence Intervals for a Binomial Parameter p

- **Still highly relevant. e.g., when estimating efficiencies by number of successes / number of trials and desiring an uncertainty.**
- **They can be calculated using ROOT's incomplete beta function. ***
- **Nonetheless every few years someone writes a paper claiming that frequentist confidence intervals don't exist when successes = trials (since the approximate formula based on sqrt(p(1-p) is not useful). Now you won't write something like that!**

***For the line of ROOT, and applications to the signal bin/sideband problem, see Cousins, Linnemann, and Tucker, http://arxiv.org/abs/physics/0702156, NIM A 595 (2008) 480.**

For a comprehensive study of confidence intervals for a binomial parameter and for the ratio of Poisson means, see Cousins and Tucker, http://arxiv.org/abs/0905.3831

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

THE USE OF CONFIDENCE OR FIDUCIAL LIMITS ILLUSTRATED IN THE CASE OF THE BINOMIAL.

BY C. J. CLOPPER, B.Sc., AND E. S. PEARSON, D.Sc.

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.

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

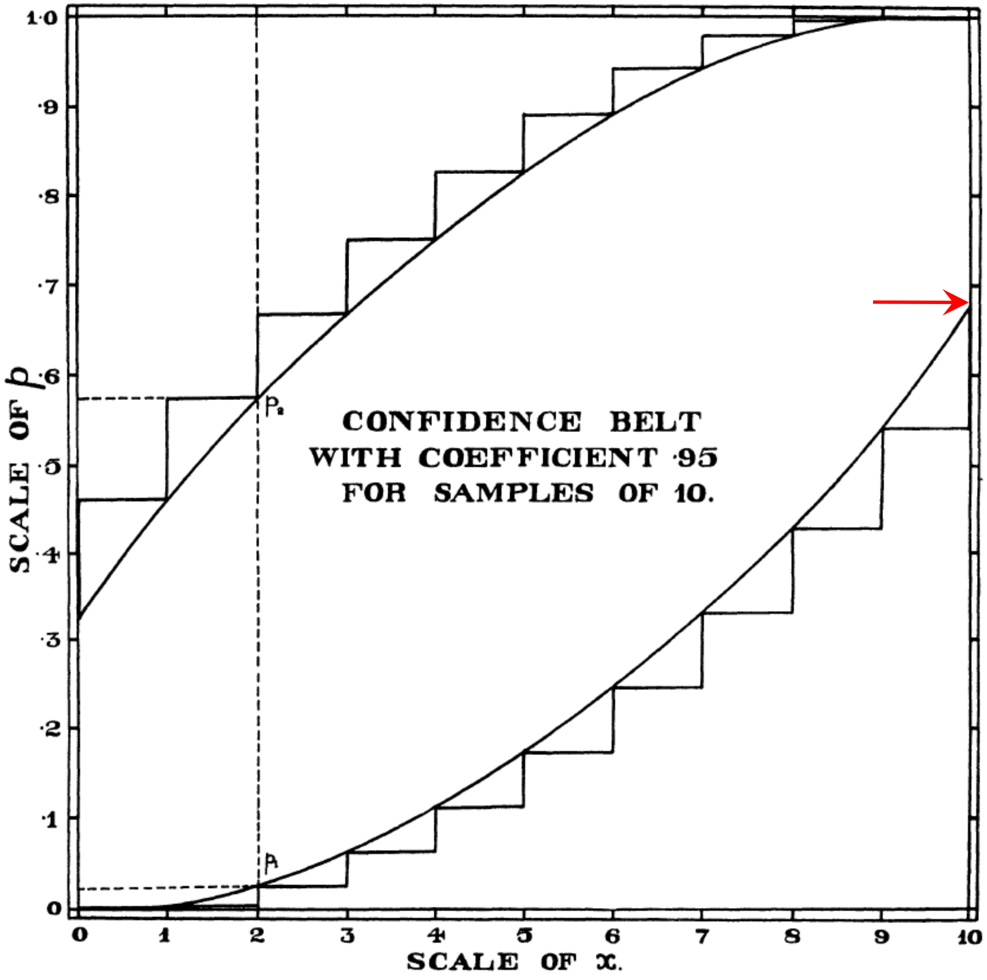


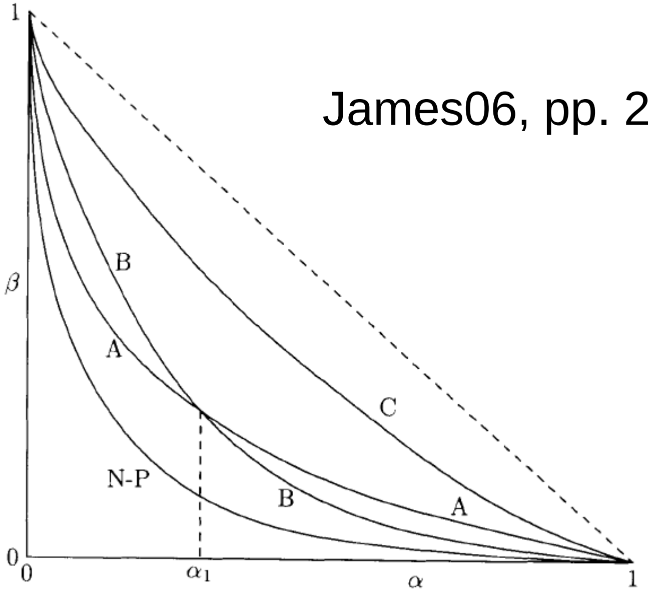
FIG. 1

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

Classical Hypothesis Testing

- In Neyman-Pearson hypothesis testing (James06), frame discussion in terms of null hypothesis H_0 = S.M., and an alternative H_1 = mSUGRA, etc.
  - α: probability (under H_0) of rejecting H_0 when it is true, i.e., false discovery claim (Type I error)
  - β: probability (under H_1) of accepting H_0 when it is false, i.e., not claiming a discovery when there is one (Type II error)
  - θ: parameters in the hypotheses
- Common for H_0 to be *nested* in H_1 to, i.e. H_0 corresponds to particular parameter values θ_0 (e.g. zero or ∞) in H_1.
- Competing analysis methods can be compared by looking at graphs of β vs α at various θ, and at graphs of β vs θ at various α (power function).
  - Similar to comparing b-tagging efficiency for signal and background, at different p_T.

Classical Hypothesis Testing (cont.)



James06, pp. 258, 262

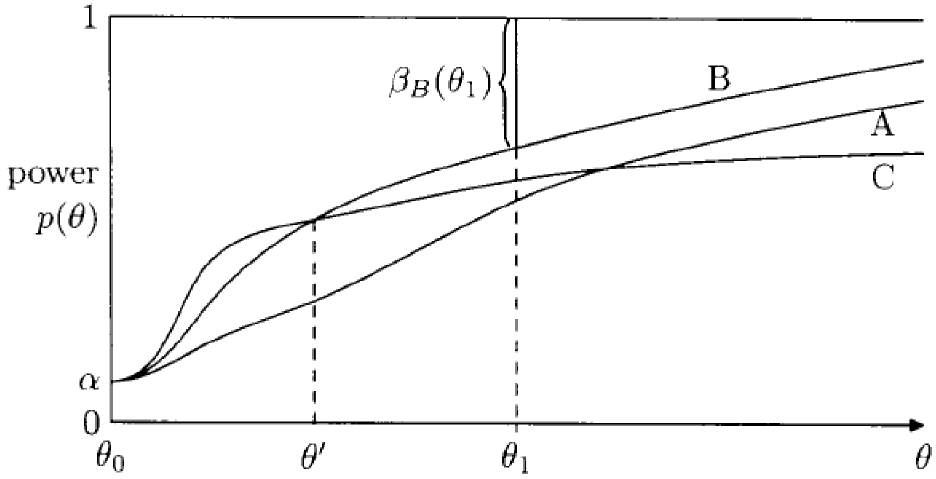


Fig. 10.3. Power functions of tests A, B, and C at significance level α. Of these three tests, B is the best for $\theta > \theta'$. For smaller values of θ, C is better.

Where to live on the β vs α curve is a *long* discussion. (Even longer when considered as number of events increases, so curve moves toward origin.) *Decision* on whether or not to declare discovery requires two more inputs:

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

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

Classical Hypothesis Testing (cont.)

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

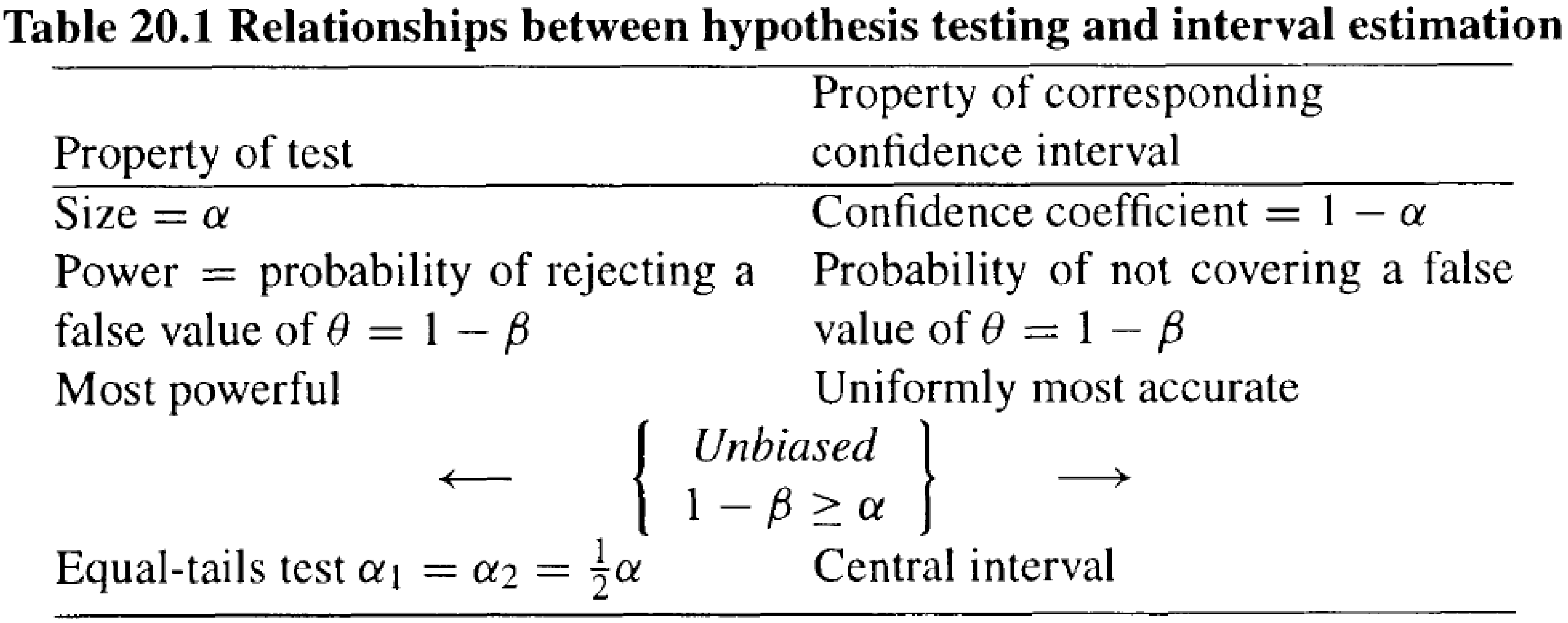


Table 20.1 Relationships between hypothesis testing and interval estimation

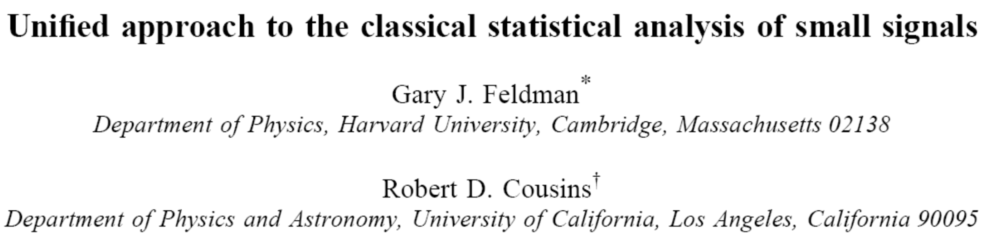
Property of test	Property of corresponding confidence interval
Size $= \alpha$	Confidence coefficient $= 1 - \alpha$
Power = probability of rejecting a false value of $\theta = 1 - \beta$	Probability of not covering a false value of $\theta = 1 - \beta$
Most powerful	Uniformly most accurate
$\longleftarrow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

"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.

Classical Hypothesis Testing (cont.)

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

Using the likelihood ratio hypothesis test, this correspondence is the basis of intervals advocated in Phys. Rev. D57 3873 (1998):



Unified approach to the classical statistical analysis of small signals

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

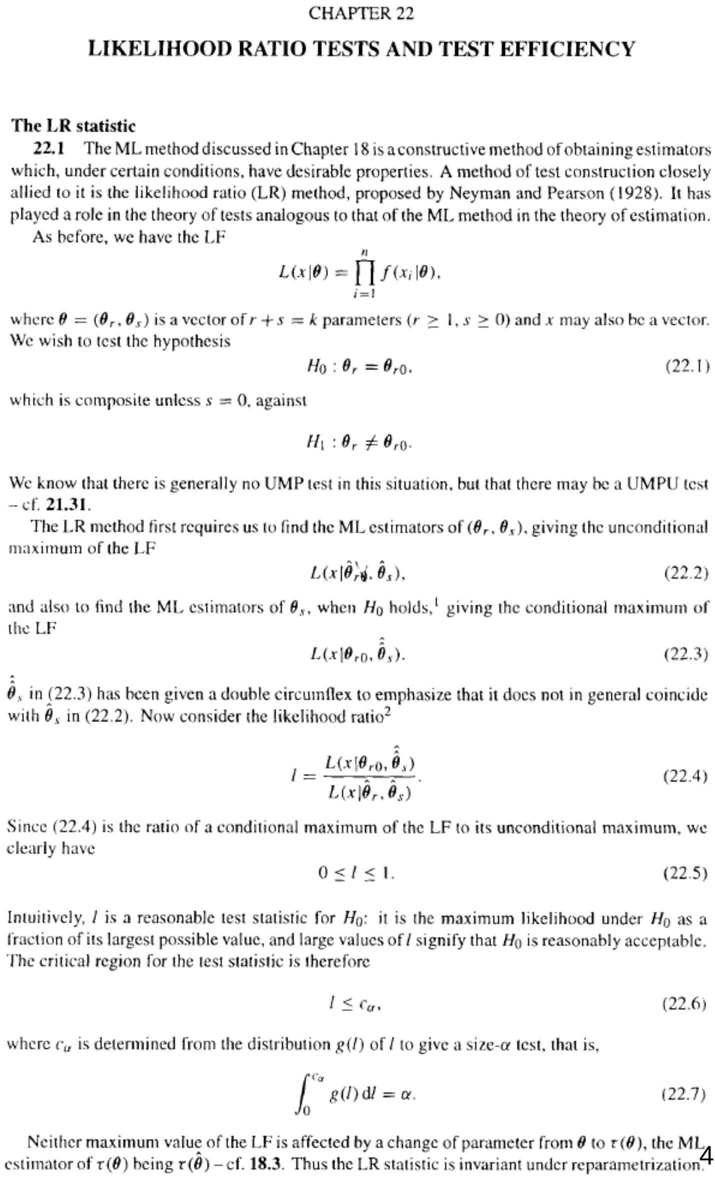
Robert D. Cousins†
Department of Physics and Astronomy, University of California, Los Angeles, California 90095

While paper was “in proof”, Gary realized that the method (including nuisance parameters) was all on 1¼ pages of “Kendall and Stuart” ! →

This was of course *good* !

It led to rapid inclusion in PDG RPP.

Today over 975 citations in SPIRES.



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

$$H_0 : \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r0}, \qquad (22.1)$$

which is composite unless $s = 0$, against

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

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

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

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

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

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

$\hat{\hat{\boldsymbol{\theta}}}_s$ 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]

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

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

$$0 \leq l \leq 1. \qquad (22.5)$$

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

$$l \leq c_\alpha, \qquad (22.6)$$

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

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

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.

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

- **If H_0 is specified but the alternative H_1 is not, then only the Type I error rate α can be calculated, since the Type II error rate β depends on a H_1. A test with this feature is called a test for *goodness-of-fit* (to H_0).**
- **The question "Which g.o.f. test is best?" is thus ill-posed. In spite of the popularity of tests with universal maps from test statistics to α (in particular χ^2 and Kolomogorov tests), they may be ill-suited for many problems (i.e., they may have poor power (1- β) against relevant alternative H_1's).**
- **In 1D, unbinned g.o.f test question is equivalent to:**
  "Given 3 numbers (e.g. neutrino mixing angles) between 0 and 1, are they consistent with three calls to RAN() ?"
  Have fun with that!
- **With the proliferation of unbinned M.L. fits, there has been a lot of discussion about an appropriate unbinned g.o.f. test in >1 dimension. See Aslan02 and references, especially book by D'Agostino and Stephens.**

Likelihood (Ratio) Intervals

- Recall from above: Likelihood $\mathcal{L}(\theta)$ is invariant under reparametrization from θ to $u(\theta)$: $\mathcal{L}(\theta) = \mathcal{L}(u(\theta))$.
  - So *likelihood ratios* $\mathcal{L}(\theta_1) / \mathcal{L}(\theta_2)$ and *log-likelihood differences* $\ln\mathcal{L}(\theta_1) - \ln\mathcal{L}(\theta_2)$ are also invariant.
- Thus, after using maximum-likelihood method to obtain estimate û which maximizes $\mathcal{L}(u)$, one can obtain a likelihood interval $[u_1,u_2]$ as the union of all u for which

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

- Asymptotically (under some regularity conditions) this interval approaches a central confidence interval with C.L. corresponding to ± Z Gaussian standard deviations
- Convergence to Gaussian is faster than you might expect. See James06 for interesting explanation why.
- But! Regularity conditions, in particular requirement that û not be on the boundary, need to be carefully checked. (E.g., if $u \geq 0$ on physical grounds, then $\hat{u}=0$ requires care.)

Likelihood-Ratio Interval example

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

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

$\Delta 2\ln\mathcal{L} = 1^2$ for approximate ±1 Gaussian standard deviation yields interval [1.58, 5.08]

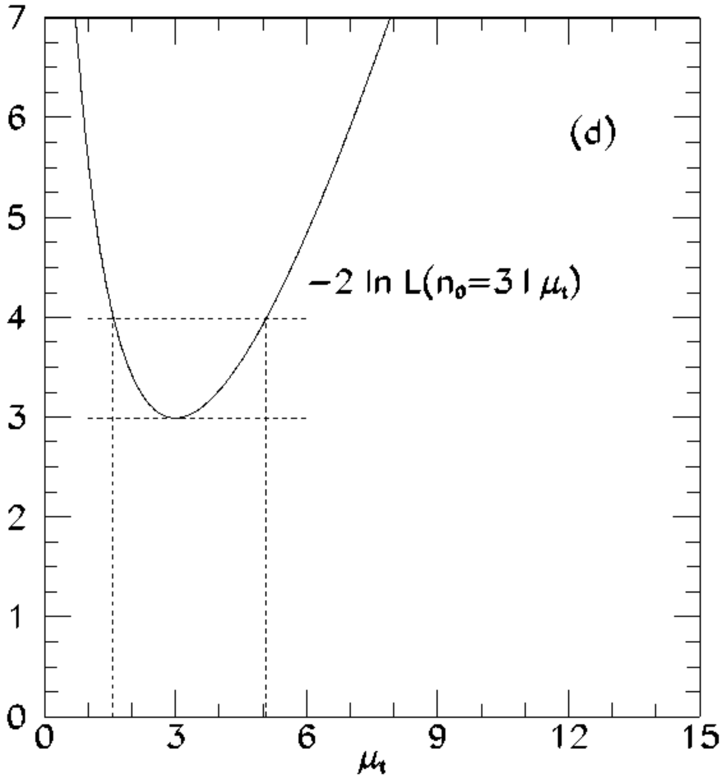


Figure from R. Cousins,
Am. J. Phys. 63 398 (1995)

More General Non-Bayesian Likelihood-Based Inference

- Beyond likelihood ratios, there are extensions which give approximations to significance level which converge faster (as n increases).
- See papers by D. Fraser and collaborators cited in Cousins05.
- I am not aware of any usage in HEP of these higher-order formulas.

U.L. in Poisson Process, n=3 observed: 3 ways

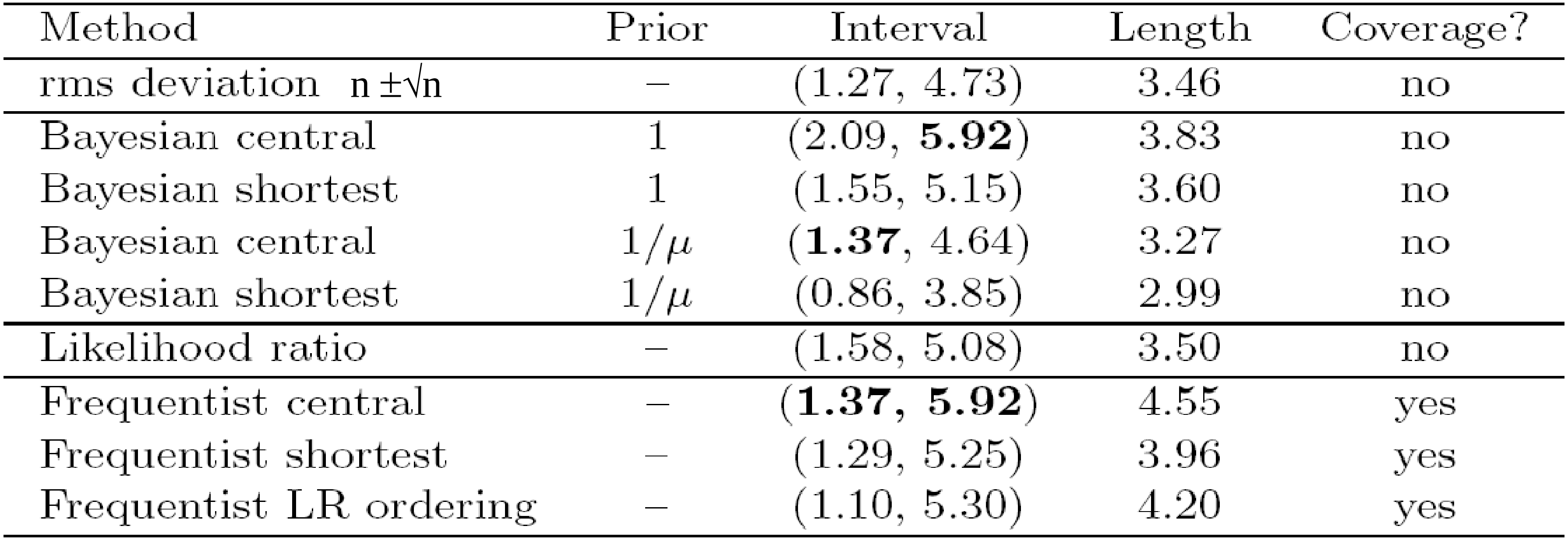
1. **Bayesian upper limit at 90% credibility:** find μ_u such that posterior probability $p(\mu > \mu_u) = 0.1$.
2. **Likelihood ratio method for approximate 90% C.L. U.L.:** find μ_u such that $\mathcal{L}(\mu_u) / \mathcal{L}(3)$ has prescribed value.
3. **Frequentist one-sided 90% C.L. upper limit:** find μ_u such that $P(n \le 3 \mid \mu_u) = 0.1$.

Deep foundational issues

- Only #3 has guaranteed ensemble properties (though issues arise with systematics.) Good ?!?
- Only #3 uses $P(n|\mu)$ for $n \neq$ observed value. Bad?!? (See below re likelihood principle)

These issues will not be resolved: aim to have software for reporting all 3 answers, and sensitivity to prior.

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



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

For the Jeffreys prior ($1/\sqrt{\mu}$), Bayesian central interval is (1.72, 5.27).

Frequentist intervals over-cover due to discreteness of n.

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

Nuisance Parameters (e.g. systematic errors)

- **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 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, likelihood ratio, Neyman confidence intervals) has a way to incorporate the uncertainty on the nuisance parameters. *But this remains a subject of frontier statistics research.***
- **After introducing the three methods, I will explain why each has problems.**

Treatment of Nuisance Parameters within Each Paradigm

- ***Bayesian credible intervals*: Construct a multi-D prior pdf P(parameters) for the space spanned by all parameters; multiply by P(data|parameters) for the data obtained; integrate over the full subspace of all nuisance parameters; you are left with the posterior pdf for the parameter of interest. The math is now reduced to the case of no nuisance parameters.**
- ***(Full) 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.**
- ***Likelihood intervals*: for each value of the parameter of interest, search the full subspace of nuisance parameters for the point at which the likelihood is maximized. Associate that value of the likelihood with that value of the parameter of interest. The set of such likelihoods is called the profile likelihood, and is a function only of the parameter of interest. The math is now reduced to the case of no nuisance parameters. (Familiar to many as MINUIT MINOS.)**

Treatment of Nuisance Parameters (Problems)

- ***Bayesian credible intervals*: the multi-D prior pdf is a problem for both subjective and non-subjective priors. In HEP there is almost no use of the favored non-subjective priors (reference priors of Bernardo and Berger), so we do not know how well they work for our problems. (The high-D integral can be a technical problem, more and more overcome by Markov Chain Monte Carlo.)**
- ***(Full) Neyman construction*: Typically intractable and causes overcoverage, and therefore rarely attempted. Tractability recovered by doing the construction in the lower dimensional space of the profile likelihood function. Not well-studied.**
- ***Likelihood intervals*: by using best-fit value of the nuisance parameters corresponding to each value of the parameter of interest, this has a reputation of 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.**
- ***It is therefore common to mix and match the various treatments!*** **For review of statistical and HEP literature, see Cousins05.**

Likelihood Principle

- As noted above, 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 (e.g., a p-value which is the probability of obtaining a value as extreme or *more extreme* than that observed), one uses probabilities of data *not seen*.
- This difference is captured by the *Likelihood Principle**: If two experiments yield likelihood functions which are proportional, then Your inferences from the two experiments should be identical.
- L.P. is built in to Bayesian inference (except e.g., when Jeffreys prior leads to violation).
- L.P. is violated by p-values and confidence intervals.
- 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.

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

Likelihood Principle Example #1

- The "Karmen Problem"
  - You expect background events sampled from a Poisson mean b=2.8, assumed known precisely.
  - For signal mean μ, the total number of events n is then sampled from Poisson mean μ+b.
  - So $P(n) = (\mu+b)^n \exp(-\mu-b)/n!$
  - Then you see no events at all! I.e., n=0.
  - $\mathcal{L}(\mu) = (\mu+b)^0 \exp(-\mu-b)/0! = \exp(-\mu) \exp(-b)$
- Note that changing b from 0 to 2.8 changes $\mathcal{L}(\mu)$ only by the constant factor exp(-b). This gets renormalized away in any Bayesian calculation, and is irrelevant for likelihood *ratios*. So for zero events observed, likelihood-based inference about signal mean μ is *independent of expected b*.
- For essentially all frequentist confidence interval constructions, the fact that n=0 is less likely for b=2.8 than for b=0 results in *narrower* confidence intervals for μ as b increases. Clear violation of the L.P.

Likelihood Principle Example #2

Binomial problem famous among statisticians (translated to HEP)

- **You want to know the trigger efficiency e. You count until reaching n=4000 zero-bias events, and note that of these, m=10 passed trigger. Estimate e = 10/4000, compute binomial confidence interval for e.**
- **Your colleague (in a different sample!) counts zero-bias events until m=10 have passed the trigger. She notes that this requires n=4000 events. Intuitively, e=10/4000 *over-estimates* e because she stopped *just* upon reaching 10 passed events. The relevant distribution is the negative binomial.**
- **Each experiment had a different *stopping rule*. Frequentist confidence intervals depend on the stopping rule.**
- **It turns out that the likelihood functions for the binomial problem and the negative binomial problem differ only by a constant! So with same n and m, (the strong version of) the L.P. demands *same* inference about e from the two stopping rules!**
- **Amusing sidebar: the Jeffreys prior is different for the two distributions, so use of Jeffreys prior violates (strong) L.P.!**

Likelihood Principle Discussion

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

- **If you are interested, read the book by Berger & Wolpert, but be prepared for the stopping rule arguments to set your head spinning.**
- ***Irrelevance* of the Stopping Rule is known as the “Stopping Rule Principle” and has been hotly debated for decades, with some famous statisticians changing their minds, e.g:**
  - **L.J. “Jimmie” Savage is widely quoted as saying in 1962, “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 resent an idea so patently right.”**

Conditioning*

- An "ancillary statistic" (see literature for precise math definition) is a function of your data which carries information about the precision of your measurement of the parameter of interest, but no info about parameter's value.
- The classic example is a branching ratio measurement in which the total number of events N can fluctuate if the expt design is to run for a fixed length of time. Then N is an ancillary statistic.
- You perform an experiment and obtain N total events, 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 there are strong arguments, going back to Fisher, that inference should be based on probabilities *conditional on the value of the ancillary statistic actually obtained*!

***See Read95 for a review.**

Conditioning (cont.)

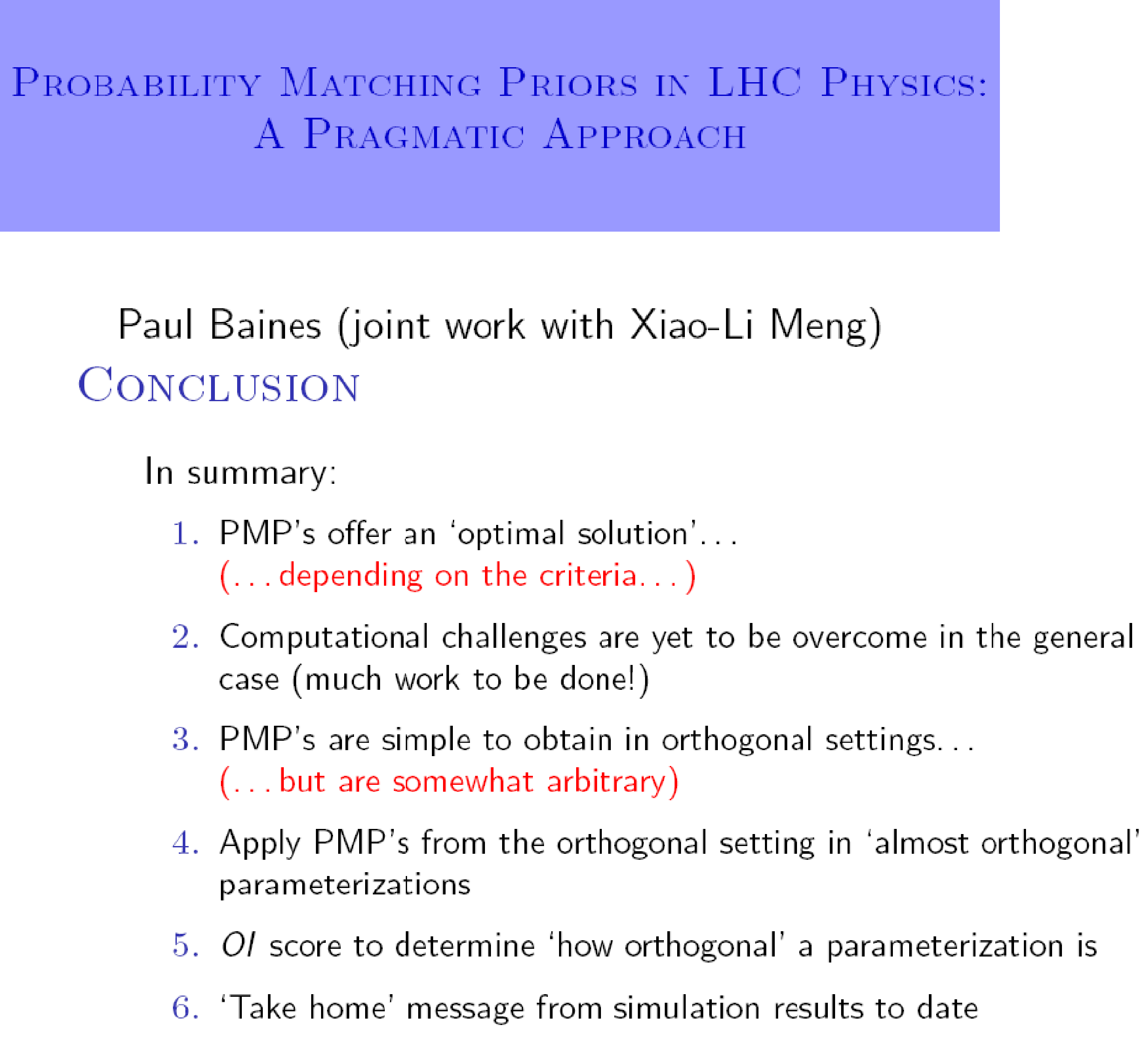
- The 1958 thought expt of David R. Cox focused the issue:
  - Your procedure for weighing an object consists of flipping a coin to decide whether to use a weighing machine with a 10% error or one with a 1% error; and then measuring the weight. (Coin flip result is ancillary stat.)
  - Then "surely" the error you quote for your measurement should reflect which weighing machine you actually used, and not the average error of the "whole space" of all measurements!
  - But classical most powerful Neyman-Pearson hypothesis test uses the whole space!
- In more complicated 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.
- In methods obeying the likelihood principle, in effect one conditions on the exact data obtained, giving up the frequentist coverage criterion for the guarantee of relevance.

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

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 while treating the parameter of interest in a frequentist way, or
  - Treat nuisance parameters by profile likelihood while treating parameter of interest another way, or
  - Use the Bayesian framework (even without the priors recommended by statisticians), but evaluate the frequentist performance. In effect (as in profile likelihood) one gets approximate coverage while respecting the L.P.
- In fact, the statistics literature going back to 1963 has attempts to find prior pdfs which lead to posterior pdfs with good frequentist coverage: *probability matching priors.* *(At lowest order in 1D, it is the Jeffreys prior!)*



Probability Matching Priors in LHC Physics: A Pragmatic Approach

Paul Baines (joint work with Xiao-Li Meng)

Conclusion

In summary:

1. PMP's offer an 'optimal solution'. . .
   (. . . depending on the criteria. . .)
2. Computational challenges are yet to be overcome in the general case (much work to be done!)
3. PMP's are simple to obtain in orthogonal settings. . .
   (. . . but are somewhat arbitrary)
4. Apply PMP's from the orthogonal setting in 'almost orthogonal' parameterizations
5. *OI* score to determine 'how orthogonal' a parameterization is
6. 'Take home' message from simulation results to date

PHYSTAT-LHC Workshop

on

Statistical Issues for LHC Physics

CERN Geneva June 27-29, 2007

This Workshop will address statistical topics relevant for LHC Physics analyses.Issues related to discovery, and the associated problems arising from systematic uncertainties, will feature prominently.

Contacts

Louis Lyons	l.lyons@physics.ox.ac.uk
Albert De Roeck	Albert.de.Roeck@cern.ch

Conference secretaries

Dorothée Denise	Dorothee.Denise@cern.ch
Kate Ross	Kate.Ross@cern.ch

Further information and registration at http://cern.ch/phystat-lhc

- **'The Holy Grail of PMP'**: A general framework to implement first and second order PMP's (unlikely anytime soon. . .)

Workshop Homepage

Workshop on Confidence Limits

27-28 March, 2000

Fermilab 1-West Conference Room

Jim Berger:

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

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

What should be the view today;

Objective Bayesian analysis is the best frequentist tool around.

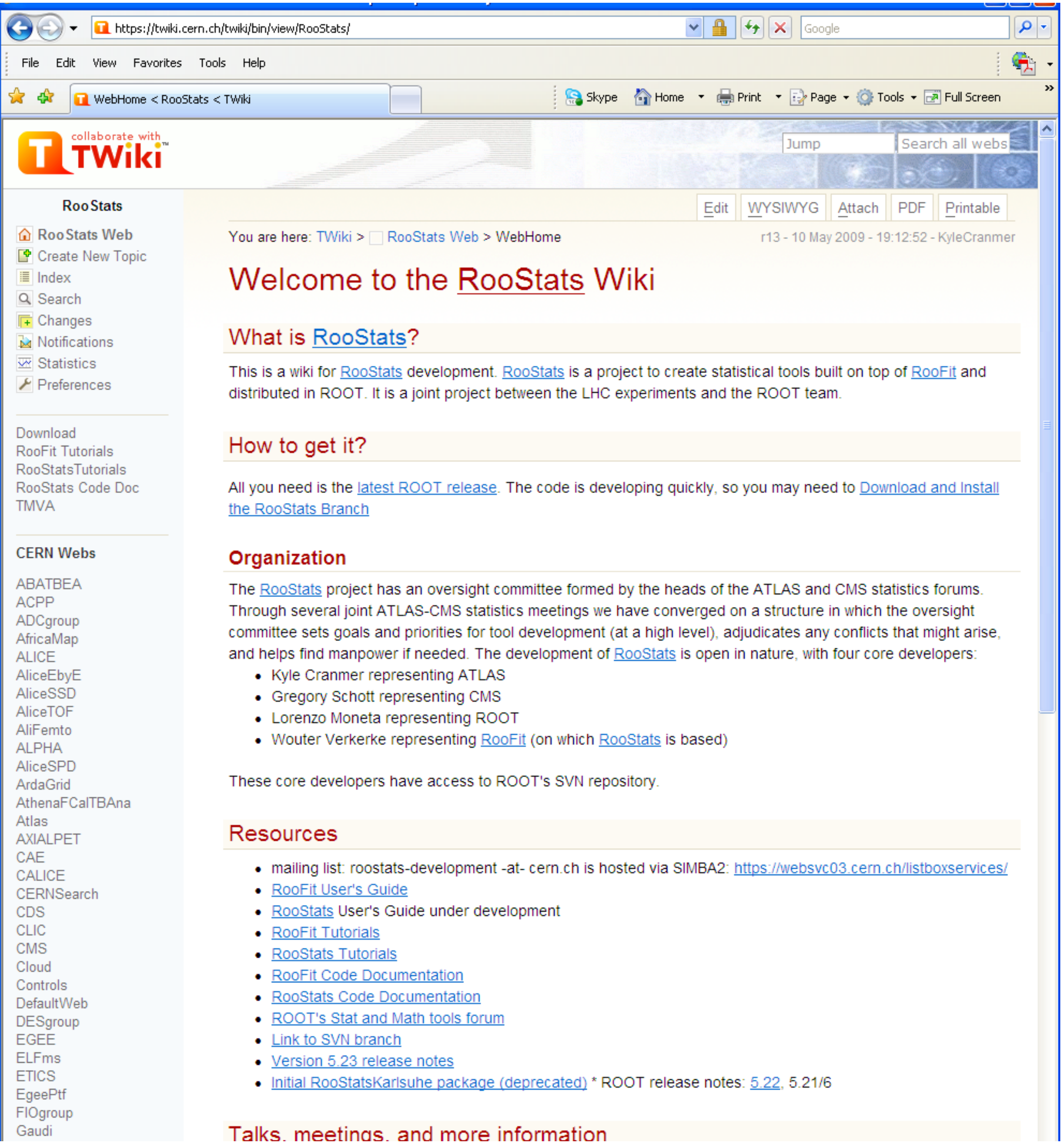
Multivariate Methods

- For deep reasons, *Likelihood Ratios* are fundamental to distinguishing between hypotheses in both frequentist and Bayesian statistics.
- Unfortunately, in multiple dimensions, it can be *extremely* difficult for a human to write down the correct likelihood functions, with all the correlations and non-linearities.
- In recent decades, there has been an explosion of methods of *machine learning*, algorithms whereby the computer can construct what is in effect a numerical substitute for the likelihood ratio.
- We have an expert at this HCPSS (Andreas Hoecker), so I just mention one of many resources, the book, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* http://www-stat.stanford.edu/~tibs/ElemStatLearn/

Goal for the LHC a Few Years Ago

- **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**
  - **Frequentist construction with approximate treatment of nuisance parameters**
  - **Profile likelihood ratio (Minuit MINOS)**
  - **Other "favorites" such as LEP's $\mathrm{CL_S}$ (which is an HEP invention)**
- **The community can 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 learn something!**
  - **The results are answers to different questions.**
  - **Bayesian methods can have poor frequentist properties**
  - **Frequentist methods can badly violate likelihood principle**

ATLAS/CMS/ROOT Project: RooStats built on RooFit



You are here: TWiki > RooStats Web > WebHome

r13 - 10 May 2009 - 19:12:52 - KyleCranmer

Welcome to the RooStats Wiki

What is RooStats?

This is a wiki for RooStats development. RooStats is a project to create statistical tools built on top of RooFit and distributed in ROOT. It is a joint project between the LHC experiments and the ROOT team.

How to get it?

All you need is the latest ROOT release. The code is developing quickly, so you may need to Download and Install the RooStats Branch

Organization

The RooStats project has an oversight committee formed by the heads of the ATLAS and CMS statistics forums. Through several joint ATLAS-CMS statistics meetings we have converged on a structure in which the oversight committee sets goals and priorities for tool development (at a high level), adjudicates any conflicts that might arise, and helps find manpower if needed. The development of RooStats is open in nature, with four core developers:

- Kyle Cranmer representing ATLAS
- Gregory Schott representing CMS
- Lorenzo Moneta representing ROOT
- Wouter Verkerke representing RooFit (on which RooStats is based)

These core developers have access to ROOT's SVN repository.

Resources

- mailing list: roostats-development -at- cern.ch is hosted via SIMBA2: https://websvc03.cern.ch/listboxservices/
- RooFit User's Guide
- RooStats User's Guide under development
- RooFit Tutorials
- RooStats Tutorials
- RooFit Code Documentation
- RooStats Code Documentation
- ROOT's Stat and Math tools forum
- Link to SVN branch
- Version 5.23 release notes
- Initial RooStatsKarlsuhe package (deprecated) * ROOT release notes: 5.22, 5.21/6

Talks, meetings, and more information

Core developers:
K. Cranmer (ATLAS)
Gregory Schott (CMS)
Wouter Verkerke (RooFit)
Lorenzo Moneta (ROOT)
Open project, all welcome to contribute.

Included in ROOT production releases since v5.22, more soon to come

Example macros in $ROOTSYS/tutorials/roostats

RooFit *extensively* documented, RooStats manual catching up, code doc in ROOT.

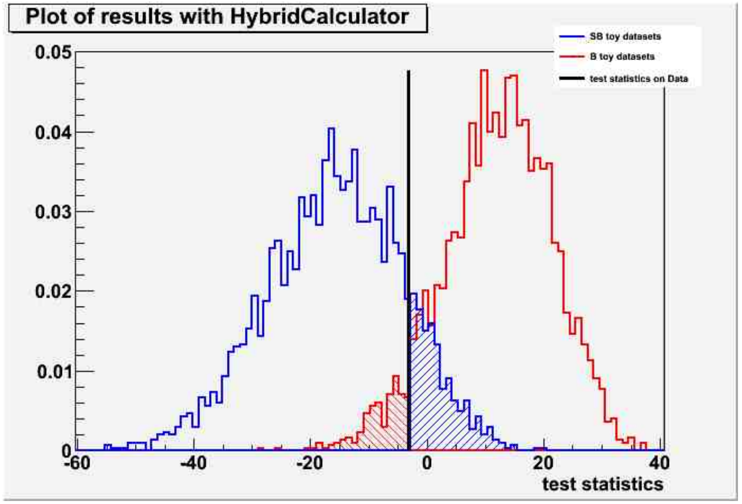
https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome

Running RooStats in Latest (dev) ROOT 5.23.04

CERN lxplus machines: examples run with a few lines, e.g.:

```
setenv ROOTSYS /afs/cern.ch/sw/lcg/app/releases/ROOT/5.23.04/slc4_ia32_gcc34/root
setenv LD_LIBRARY_PATH ${ROOTSYS}/lib
setenv PATH ${PATH}:${ROOTSYS}/bin
root -x $ROOTSYS/tutorials/roostats/rs201_hybridcalculator.C
```

Problem considered: Is histogram all background or is there a signal superimposed?
H_0 is uniform background “B” with Poisson mean estimated to be 40±10 events
H_1 is has Gaussian signal in addition to background (“SB”), Poisson mean 20 events.
Calculates test statistic in frequentist way for signal mean, but treats 25% uncertainty on background in Bayesian way. Produces distributions of test statistics under H_0 and H_1 on, from which tail probabilities are calculated and printed.



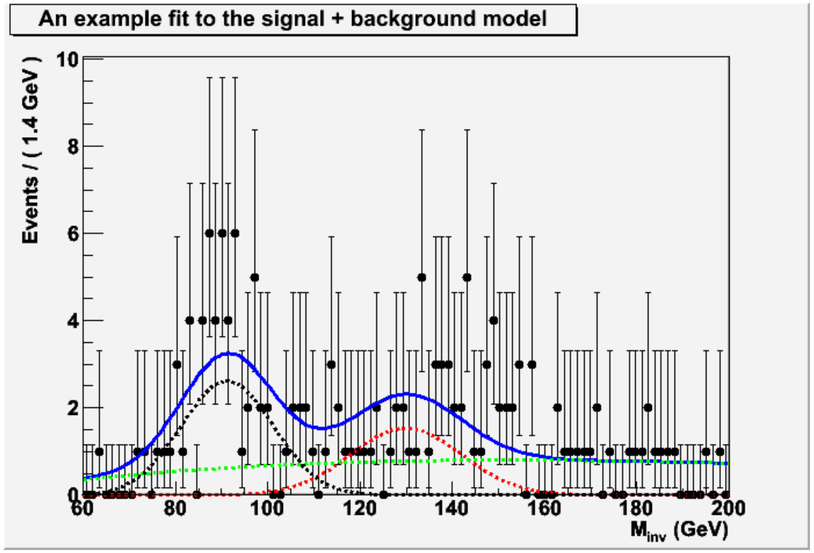
Code in this example is well-commented but to fully understand it requires consulting the code doc as well, some knowledge of Roofit and RooStats.

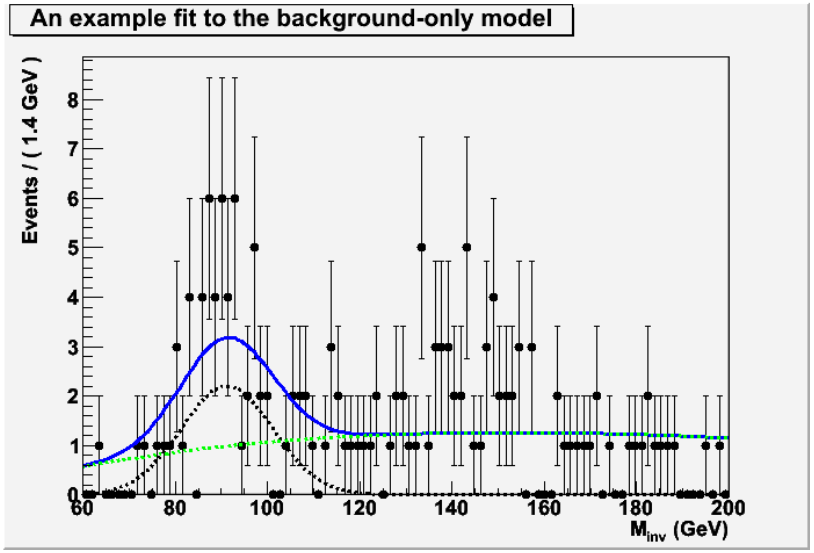
Another example in /tutorial/roostats/: rs102_hypotestwithshapes.C:

Search for a new particle by studying an invariant mass distribution.

The macro creates a simple signal model (Higgs) and two background models (smooth and Z^0), which are added to a RooWorkspace.

The macro creates a toy dataset, and then uses a RooStats ProfileLikleihoodCalculator to do a hypothesis test of the background-only and signal+background hypotheses. In this example, shape uncertainties are not taken into account, but normalization uncertainties are.





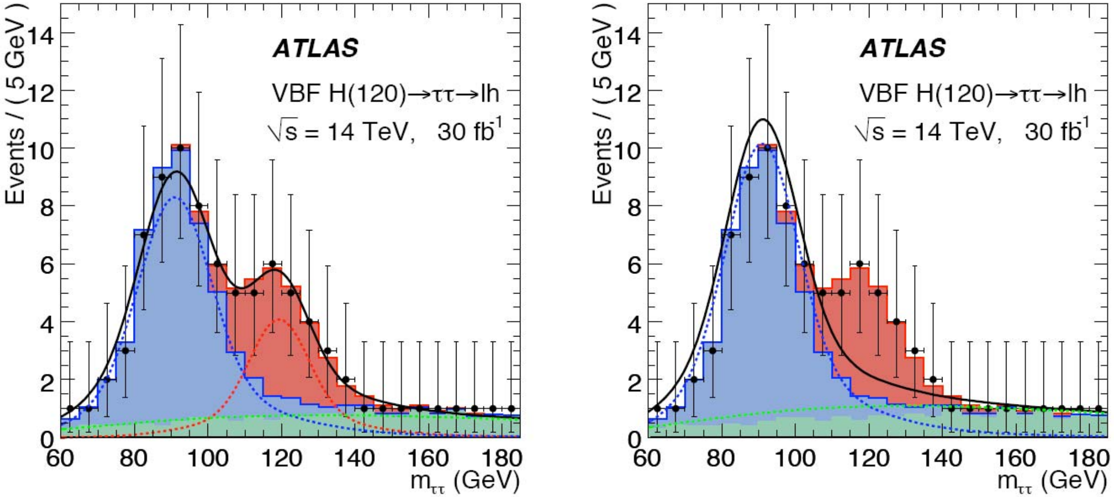
Returns:

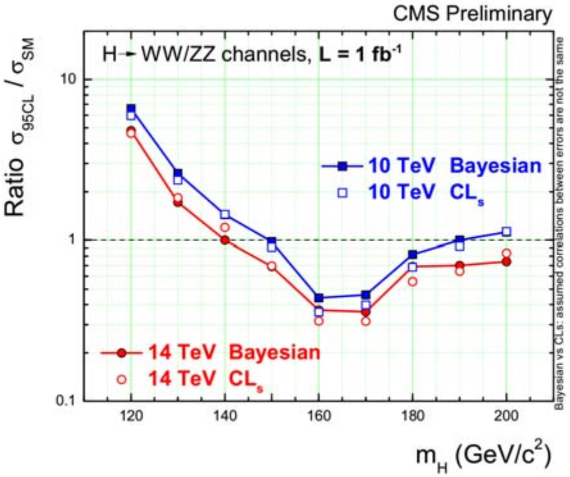
- p-value for the null hypothesis: 9.8e-05
- corresponding to equivalent Z = 3.72 Gaussian sigmas

Beyond tutorials: **The code in the previous example came from an ATLAS Higgs study,**

Search for the Standard Model Higgs Boson via Vector Boson Fusion Production Process in the Di-Tau Channels**,**

in the recent ATLAS physics performance book, http://arxiv.org/abs/0901.0512 .





Similarly, some "approved" results in CMS (curves marked $\mathrm{CL_S}$) were obtained with CMS's RooStats development code, since ported to production. Others methods now being ported as well.

RooStats Outlook

- Goal is within reach to have implementation of the three main approaches for a number of typical applications.
- Since December 2008, part of ROOT production releases: Many technical, organizational, and interface hurdles have been resolved, developers can now concentrate on implementation and validation for specific problems.
- Whole framework is designed for combining results:
  - From several channels within an experiment
  - From CMS and ATLAS (and more)
- "Workspace" concept allows electronic publication of the likelihood function, etc.
- Some rather complex combinations have already been implemented, e.g., for Higgs searches.
- In active development, more developers welcome, further validation crucial.

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.”**
- **“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.”**

***References available on request**

Recommended reading

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

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

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

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

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

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

By now there are many many web pages with lists of statistics references – Google on your favorite topic. For Bayesian work, a must-read Bayesian reading list is the set of citations in my Comment, Phys. Rev. Lett. 101 029101 (2008).

Huge literature on multi-variate analysis; e.g http://www-stat.stanford.edu/~jhf/, including the book Elements of Statistical Learning.

References Cited in Talk Slides

Aslan02: B. Aslan, G. Zech, “Comparison of different goodness-of-fit tests”, Conference on Advanced Statistical Techniques in Particle Physics, Durham, England, 18-22 Mar 2002.

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

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

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

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

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

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