Statistical Summary of 2010 "Significance and Discovery Claims Workshop"

David A. van Dyk

Department of Statistics, University of California, Irvine

BIRS, July 2010

Outline

Caveats

This is an incomplete summary of some of the topics that I found interesting.

I'm sure I'm missing important contributions!

Please correct me if I mischaracterize your contribution!!

Forgive me if I stand on my soap box a bit....

Outline

Why *are* we using 5σ

Are we really worried about making one Type-1 error in 1.7 million results??

No. We are worried about:

- The look elsewhere effect. (Louis Lyons)
- Calibration and systematic errors. (Richard Lockhart)

For WIMP 3σ is okay, but there is little LEE. (Henrique Araujo)

Problems with 5σ

- We don't know the actual effect of Systematics and LEE.
- "No distribution is valid to the 5σ tail!" (Cox via Lyons)
- Sampling distributions are only asymptotic approximations.
- Must calculate extreme-tail probabilities. (Michael Woodroofe)

We have **no** *idea what the actual level is.*

5σ simply sweeps the problem under the rug.

Question

We work for the null distribution of the LRT and to accurately compute extreme tail probabilities.

Why not work to crack systematics and LEE?

We could sweep the null distribution and tail probabilities under the rug and use 6 or 7σ.

It is better to face the real issues head on
(as Eliam and Ofer are with LEE).

What should we do?

- Handle the systematics and LEE directly.
- Use Neyman-Person with a realistic α or Bayesian model selection (Bob Cousins)
- Lehmann suggests comparison of α and β and using prior belief for H_0. (Bob Cousins)
- When a p-value of 10^{-8} is called back we need to figure out what went wrong! (Richard Lockhart)
- Identify systematic problems and improve procedures.

Goal: Honest frequency error rates or a calibrated Bayesian procedure.

Outline

The Problems with P-Values

With a Precise Null.

1. Replace data with "data as extreme or more extreme". Not particularly conservative. (Berger, Cousins)
2. Can vastly overstate the evidence for H_A. (Jim Berger)
3. Cannot be calibrated vis-a-vis $Pr(H_0)$. (Berger via Cousins)
4. Calibration depends on sample size, fitted model, and how sharp H_0 is.
5. Most importantly, they don't answer the question of scientific interest: "Have we discovered a Higgs Boson"

The Road to Damascus

The Road to Damascus

Bayes Factors

Challenges

- Priors really matter. They must be proper and informative. (Berger, Cousins)

Advantages

- They lay their assumptions out for all the world to see.
- Nothing need be hidden or swept under the rug.
- They are easy to interpret and answer the most relevant scientific question.

Bayes Factors don't have to be perfect, they just have to be better than p-values!

Prior Distributions with Bayes Factors

- The scale of the prior will influence the Bayes Factor. (Cousins, Berger)
- We can obtain a range of Bayes Factors using a range of priors/scales (Jim Berger)
- “There are lots of priors out there, but you can’t use them and not worry.... you can use them" (Jim Berger)
- “Subjective” and “Scientific” priors (Jim Berger)
- Reference priors (Harrison Prosper)

General Issues with Detection

- Neither Hypothesis may be true or rejectable. (Richard Lockhart)
- Model checking (and improvement!) is always in order (e.g., p_0 and p_1).
- There is no easy way out.
  1. p-values are not frequentist and are cannot be calibrated (Berger, Cousins)
  2. Neyman-Pearson gives a frequentist detection decision but says nothing about the strength of the detection.
  3. Bayesian methods require influential prior distributions.

Suggested Strategies

- Use the LRT statistic integrated over the parameters under the alternative and pick a prior under H_0 to achieve the desired level. (Richard Lockhart)
- Don't use $\mathrm{CL_S}$ with non-Poisson models. (Bill Murray)
- Use Binomial test or LRT in on/off setting. (Jim Linnemann)
- General routines for "Bayesian-Frequentist hyprid" (Kyle Cranmer).
- Report likelihood ratio along with p-value. (Bob Cousins)
- Report Interval Estimate and "Upper Limit" along with (non) detection. (David van Dyk)
- Bayes Factors / Conditional p-values (Jim Berger)

Outline

Sampling Distribution

A number of simulation analytical studies explored the null distribution of the LRT statistic

- Elliott Bloom found that the standard asymptotics don't always materialize.
- Glen Cowen described analytical results that start with the null in the interior.
- Eilam Gross showed that LRT evaluated at local modes may exhibit predictable behavior.

What is going on?
Review the standard asymptotics...

Wilks (Annals of Math. Statist., 1938)

THE LARGE-SAMPLE DISTRIBUTION OF THE LIKELIHOOD RATIO FOR TESTING COMPOSITE HYPOTHESES[1]

BY S. S. WILKS

By applying the principle of maximum likelihood, J. Neyman and E. S. Pearson[2] have suggested a method for obtaining functions of observations for testing what are called *composite statistical hypotheses*, or simply *composite hypotheses*. The procedure is essentially as follows: A population K is assumed in which a variate x (x may be a vector with each component representing a variate) has a distribution function $f(x, \theta_1, \theta_2, \cdots \theta_h)$, which depends on the parameters $\theta_1, \theta_2 \cdots \theta_h$. A *simple hypothesis* is one in which the θ's have specified values. A set Ω of admissible hypotheses is considered which consists of a set of simple hypotheses. Geometrically, Ω may be represented as a region in the h-dimensional space of the θ's. A set ω of simple hypotheses is specified by taking all simple hypotheses of the set Ω for which $\theta_i = \theta_{0i}$, $i = m + 1, m + 2, \cdots h$.

Wilks (Annals of Math. Statist., 1938)

We can summarize in the

Theorem: *If a population with a variate x is distributed according to the probability function $f(x, \theta_1, \theta_2 \cdots \theta_h)$, such that optimum estimates $\tilde{\theta}_i$ of the θ_i exist which are distributed in large samples according to* (3), *then when the hypothesis H is true that $\theta_i = \theta_{0i}$, $i = m+1, m+2, \cdots h$, the distribution of $-2\log\lambda$, where λ is given by* (2) *is, except for terms of order $1/\sqrt{n}$, distributed like χ^2 with $h-m$ degrees of freedom.*

PRINCETON UNIVERSITY,
PRINCETON, N. J.

Wilks (Annals of Math. Statist., 1938)

which *optimum* estimates of the θ's exist. That is, we shall assume the existence of functions $\tilde{\theta}_i(x_1, \cdots x_n)$(maximum likelihood estimates of the θ_i) such that[3] their distribution is

$$\frac{|c_{ij}|^{\frac{1}{2}}}{(2\pi)^{h/2}} e^{-\frac{1}{2}\sum_{i,j=1}^{h} c_{ij} z_i z_j} (1 + \phi)\, dz_1 \cdots dz_h \tag{3}$$

where $z_i = (\tilde{\theta}_i - \theta_i)\sqrt{n}$, $c_{ij} = -E\left(\frac{\partial^2 \log f}{\partial\theta_i\, \partial\theta_j}\right)$, E denoting mathematical expectation, and ϕ is of order $1/\sqrt{n}$ and $||c_{ij}||$ is positive definite. Denoting (3) by

Chernoff (Annals of Math. Statist., 1954)

ON THE DISTRIBUTION OF THE LIKELIHOOD RATIO[1]

By Herman Chernoff

Stanford University

1. Summary and Introduction. A classical result due to Wilks [1] on the distribution of the likelihood ratio λ is the following. Under suitable regularity conditions, if the hypothesis that a parameter θ lies on an r-dimensional hyperplane of k-dimensional space is true, the distribution of $-2 \log \lambda$ is asymptotically that of χ^2 with $k - r$ degrees of freedom.

In many important problems it is desired to test hypotheses which are not quite of the above type. For example, one may wish to test whether θ is on one side of a hyperplane, or to test whether θ is in the positive quadrant of a two-dimensional space. The asymptotic distribution of $-2 \log \lambda$ is examined when the value of the parameter is a boundary point of both the set of θ corresponding to the hypothesis and the set of θ corresponding to the alternative.

Chernoff (Annals of Math. Statist., 1954)

A not-technical summary:

placed by the inverse of the information matrix. In particular, if one tests whether θ is on one side or the other of a smooth $(k-1)$-dimensional surface in k-dimensional space and θ lies on the surface, the asymptotic distribution of λ is that of a chance variable which is zero half the time and which behaves like χ^2 with one degree of freedom the other half of the time.

- Requires the MLE to converge to the truth under H_0.
- Thus, "nuisance" parameters must be identifiable under H_0.

Fan, Hung, and Wong (J. Amer. Statist. Assoc., 2000)

Geometric Understanding of Likelihood Ratio Statistics

Jianqing FAN, Hui-Nien HUNG, and Wing-Hung WONG

It is well known that twice a log-likelihood ratio statistic follows asymptotically a chi-square distribution. The result is usually understood and proved via Taylor's expansions of likelihood functions and by assuming asymptotic normality of maximum likelihood estimators (MLEs). We obtain more general results by using a different approach: the Wilks type of results hold as long as likelihood contour sets are fan-shaped. The classical Wilks theorem corresponds to the situations in which the likelihood contour sets are ellipsoidal. This provides a geometric understanding and a useful extension of the likelihood ratio theory. As a result, even if the MLEs are not asymptotically normal, the likelihood ratio statistics can still be asymptotically chi-square distributed. Our technical arguments are simple and easily understood.

1. INTRODUCTION

One of the most celebrated folk theorems in statistics is that twice the logarithm of a maximum likelihood ratio

see Examples 1 and 2 in Section 3. An additional benefit is that our technical arguments are simple and can be understood without much probability background.

We begin with the simplest case, in which the null hy-

An Example

Spectral Analysis in High-Energy Astrophysics

- We fit a power-law continuum and test for an added emission line of (a) known or (b) unknown location.

MODEL 0. There is no emission line.

MODEL 1. There in an emission line with fixed location in the spectrum, but unknown intensity.

MODEL 2. There is an emission line with unknown location and intensity.

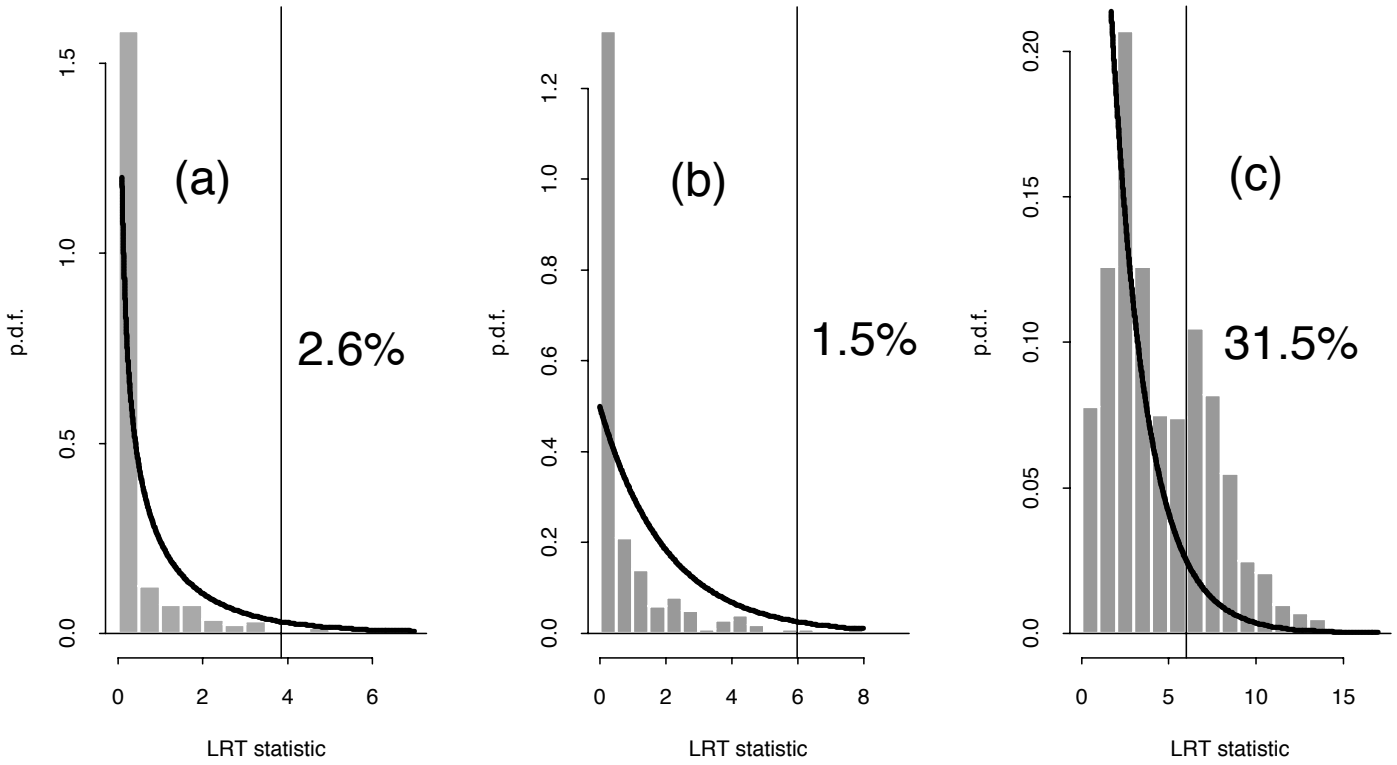
An Example

Spectral Analysis in High-Energy Astrophysics

- With known location all parameters are identified under H_0 but H_0 is on the boundary.
- With unknown location all parameters are *not* identified under H_0 and H_0 is on the boundary.
- We also tested for a (c) two-parameter absorption feature.

(Protossov, et al., 2002, ApJ)

Results



Using Wilk's Th can lead to conservative or anti- conservative results.
Chernoff's Theorem applies in (a).

Missing the global mode

- The LRT requires computing the MLE under H_0 and H_A.
- In a Monte Carlo simulation under H_0, this can be computationally challenging.
- Even if the optimizer is imperfect, however, the result is still a valid (but typically less powerful) test statistic.
- It is only required that the same computation be preformed on the real data as on the Monte Carlo sample.
- In this case the standard asymptotics do not apply.

Outline

LEE

For a frequentist analysis

- There are several sorts of "elsewhere" and we need to define them (Louis Lyons).
- But how? And where exactly? What is the real effect?

Frequency properties depend on what you would do with other data and what you might do with other data in the future.

"The absurdity of these comparisons is what makes me a Bayesian" (Steffen Lauritzen).

An easier case

Looking for a bump... in several locations.

- p-values and significance levels should be altered to account for the fact that an added model component may be anywhere in the range of the data or.... elsewhere. (Louis Lyons).
- Bob Cousins conjectures that the LEE goes as $\sqrt{n}$.
- Eilam Gross described and Ofer Vitells applied a method to adjust p-values to account for the LEE.
- This can also be done by adjusting H_A and re-calibrating.

Example

Spectral Analysis in High Energy Astrophysics:
Quasar PG1637+706.

> MODEL 0. There is no emission line.
>
> MODEL 1. There in an emission line with fixed location in the spectrum, but unknown intensity.
>
> MODEL 2. There is an emission line with unknown location and intensity.

To fit Model 2 under H_0 we use multiple starting values...
and use the *same* starts with the real data.

Results

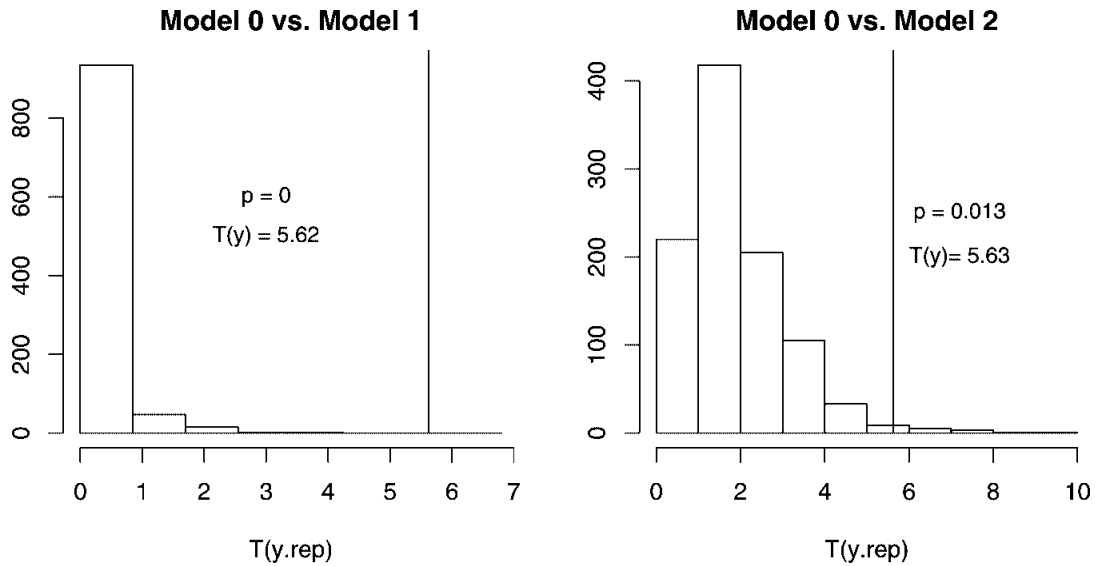


FIG. 4. *The posterior predictive check. The two histograms compare the observed likelihood ratio test statistics (vertical lines) with* 1000 *simulations from the posterior predictive distribution. The left plot is the comparison between Model* 0 *and Model* 1, *and the right plot is the comparison between Model* 0 *and Model* 2. *Both model checks indicate strong evidence for including the emission line.*

Outline

- All goodness-of-fit tests have an implicit alternative. (Richard Lockhart)
- Consider what alternatives matter and design your test accordingly. (Richard Lockhart)
- But how?
- Chad Schafer presented an impressive all purpose method for designing such tests.
- This allows us to clearly identify the alternative we are testing against.

Outline

Procedures Appear to Underestimate Uncertainty

Statistical Challenge (Jon Pumplin and Robert Thorne)

- Inconsistency between individual & combined experiments.
- Individual experiments are more variable than expected.

Suggestions

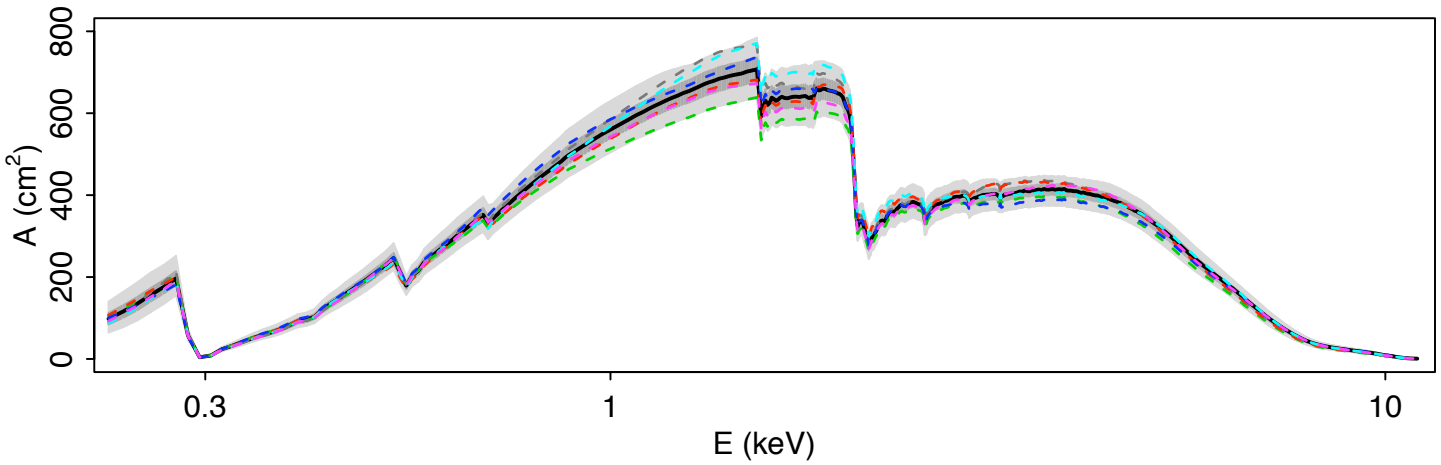
- Verify that individual experiments are really too variable.
- Compare the variance of the fitted values with Monte Carlo.
- There may be unaccounted for systematic differences.
- Random effects model could separate the experimental variability from the variance due to systematic differences.
- The data are not really exchangable....

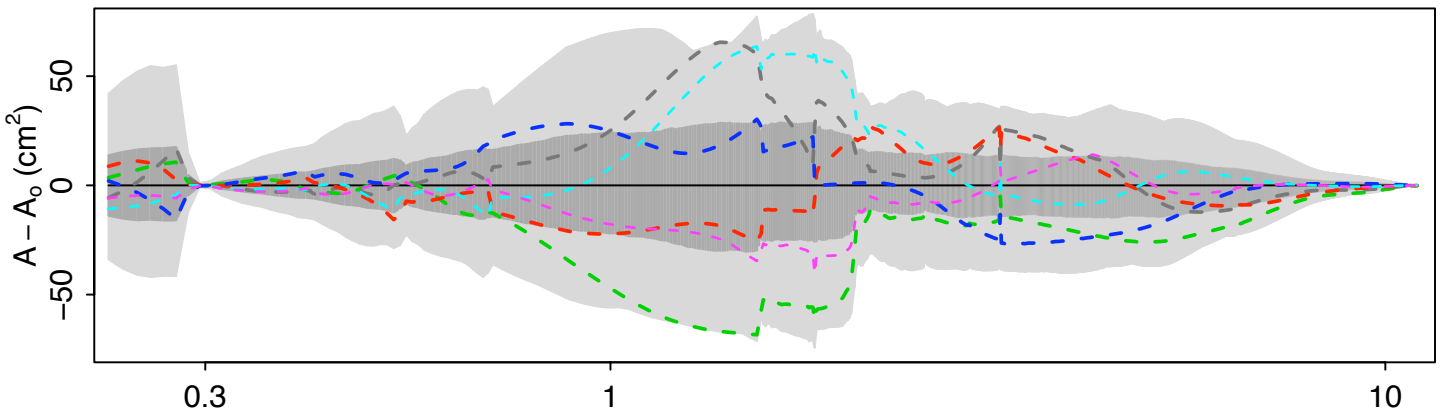
Outline

Dealing with Systematics

- "If you can't account for systematics, no number of σ will help" (Richard Lockhart)
- But how can we account for systematics?
- Example from High-Energy Astrophysics (Lee et al., 2010, under revision for ApJ)

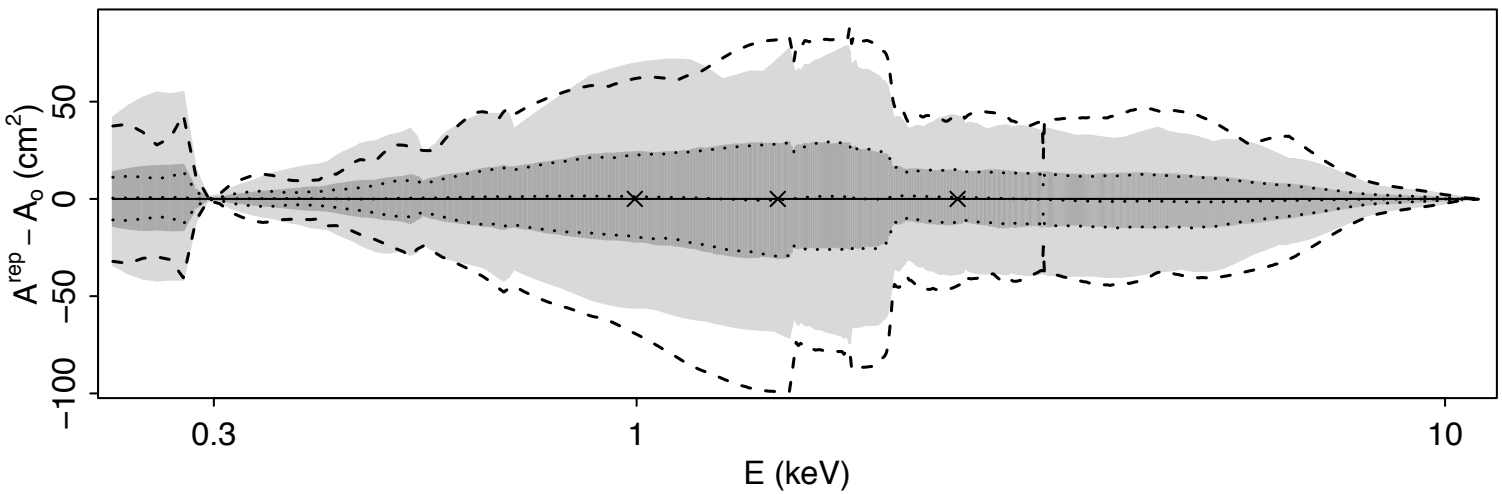
Effective Area Curves and Calibration Sample





Representation of Calibration Sample

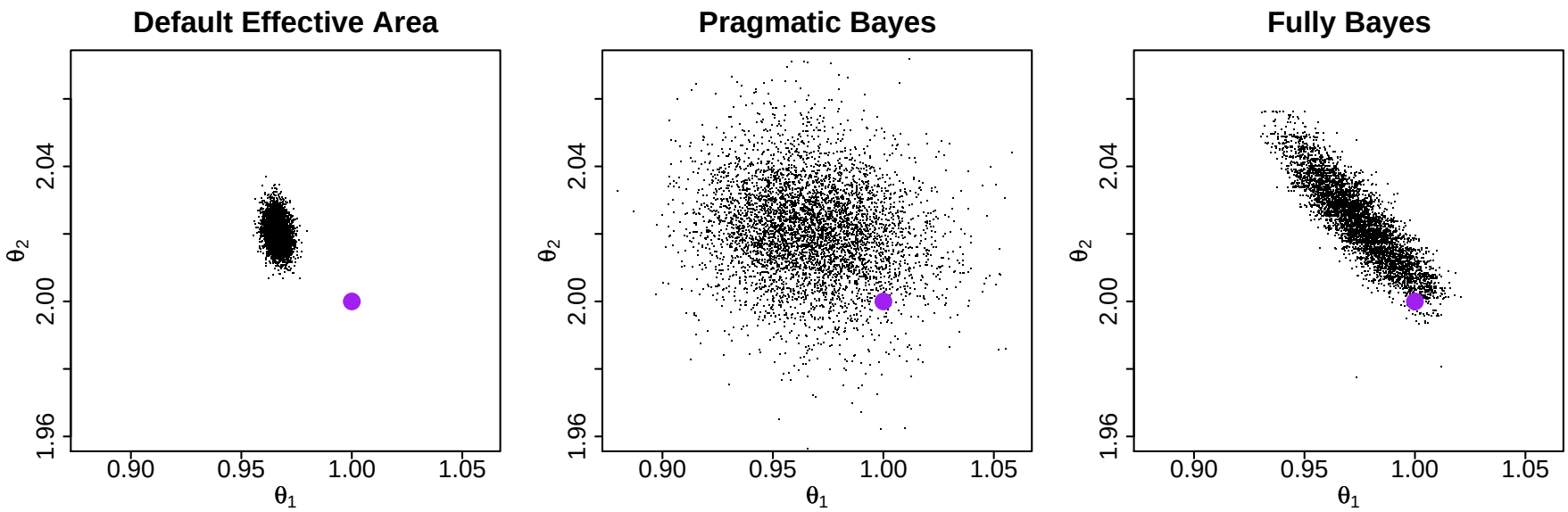
- We summarize the Calibration sample using Principle Component Analysis.
- Effective Area Curves can then be sampled as needed.
- Compare 67% and full ranges of original and PCA sample:



Methods

- Calibration sample is a prior for the effective area curve.
- Repeat standard analysis with random sample of curves.
- Combine results using Multiple Imputation combining rules.
- Alternatively, embed into Bayesian model &fit with MCMC.
- Assume data are independent of curve to simplify computation.
- Sample effective area curves from prior, not given data.

Results


Default Effective Area
Pragmatic Bayes
Fully Bayes
θ2
θ1
1.96
2.00
2.04
0.90
0.95
1.00
1.05