The Large Hadron Collider and Goodness of Fit

Richard Lockhart

Simon Fraser University

DASF III, April 30 – May 1, 2010

- "The statistical model has not always been a black box."

Fraser (*Trans Roy Soc Can*, 1967) Page # -24

- "The statistical model has not always been a black box."
- "And it need not be a black box now."

Outline

For when I don't finish

- Physics problems are hard.

For when I don't finish

-22

- Physics problems are hard.
- As is not usual – you really don't want the raw data.

For when I don't finish

- Physics problems are hard.
- As is not usual – you really don't want the raw data.
- The scientific null hypothesis is a point null.

For when I don't finish

- Physics problems are hard.
- As is not usual – you really don't want the raw data.
- The scientific null hypothesis is a point null.
- But the data analytic null is more nuanced.

For when I don't finish

- Physics problems are hard.
- As is not usual – you really don't want the raw data.
- The scientific null hypothesis is a point null.
- But the data analytic null is more nuanced.
- Particle detection can be mixture of GOF with more traditional tests.

For when I don't finish

- Physics problems are hard.
- As is not usual – you really don't want the raw data.
- The scientific null hypothesis is a point null.
- But the data analytic null is more nuanced.
- Particle detection can be mixture of GOF with more traditional tests.
- Non-parametric priors lead to Neyman Pearson tests.

For when I don't finish

- Physics problems are hard.
- As is not usual – you really don't want the raw data.
- The scientific null hypothesis is a point null.
- But the data analytic null is more nuanced.
- Particle detection can be mixture of GOF with more traditional tests.
- Non-parametric priors lead to Neyman Pearson tests.
- Gaussian priors lead to quadratic tests.

For when I don't finish

- Physics problems are hard.
- As is not usual – you really don't want the raw data.
- The scientific null hypothesis is a point null.
- But the data analytic null is more nuanced.
- Particle detection can be mixture of GOF with more traditional tests.
- Non-parametric priors lead to Neyman Pearson tests.
- Gaussian priors lead to quadratic tests.
- Sensible prior: decision won't be obvious *a priori* – sample size dependence.

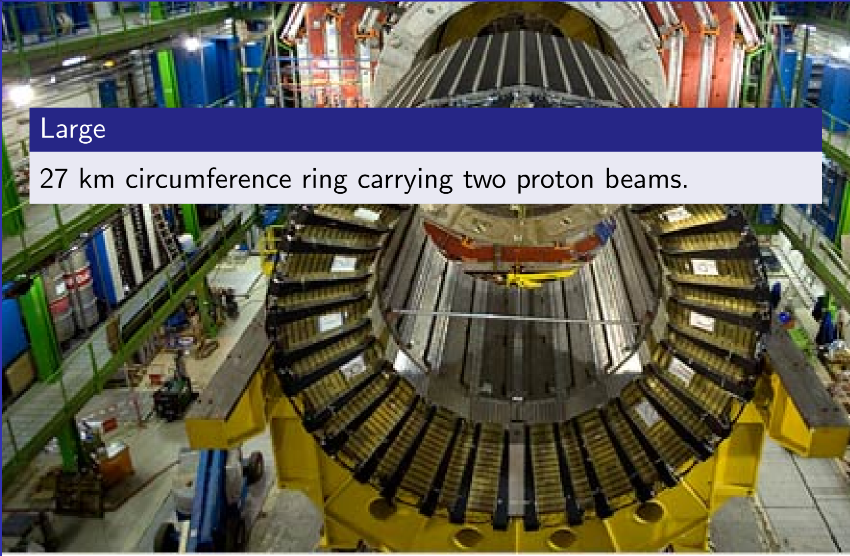
For when I don't finish

- Physics problems are hard.
- As is not usual – you really don't want the raw data.
- The scientific null hypothesis is a point null.
- But the data analytic null is more nuanced.
- Particle detection can be mixture of GOF with more traditional tests.
- Non-parametric priors lead to Neyman Pearson tests.
- Gaussian priors lead to quadratic tests.
- Sensible prior: decision won't be obvious *a priori* – sample size dependence.
- The real problems are strikingly hard – and the physicists want and expect smart solutions.

The Higgs particle

- Elementary particle; one of 5 elementary bosons.
- Existence predicted by Standard Model of particle physics.
- Not yet observed.
- Major target of experiments at Large Hadron Collider at CERN.

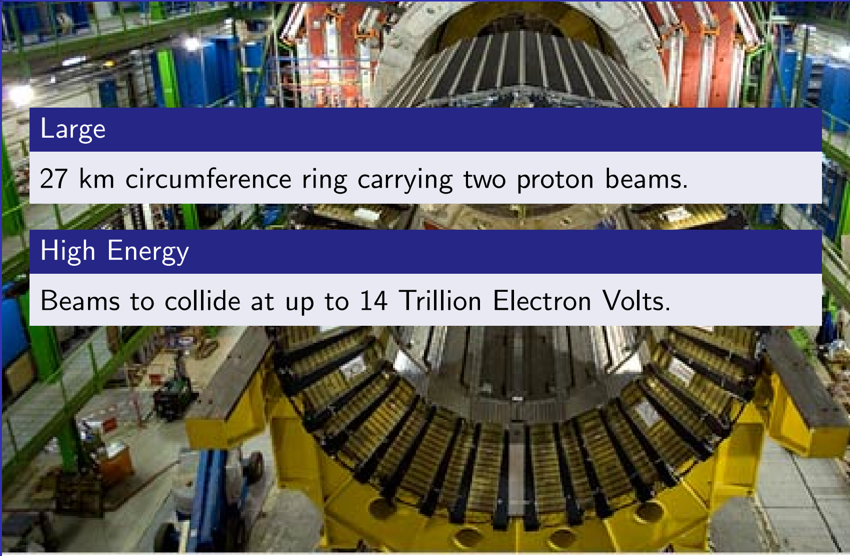
LHC



Large

27 km circumference ring carrying two proton beams.

LHC



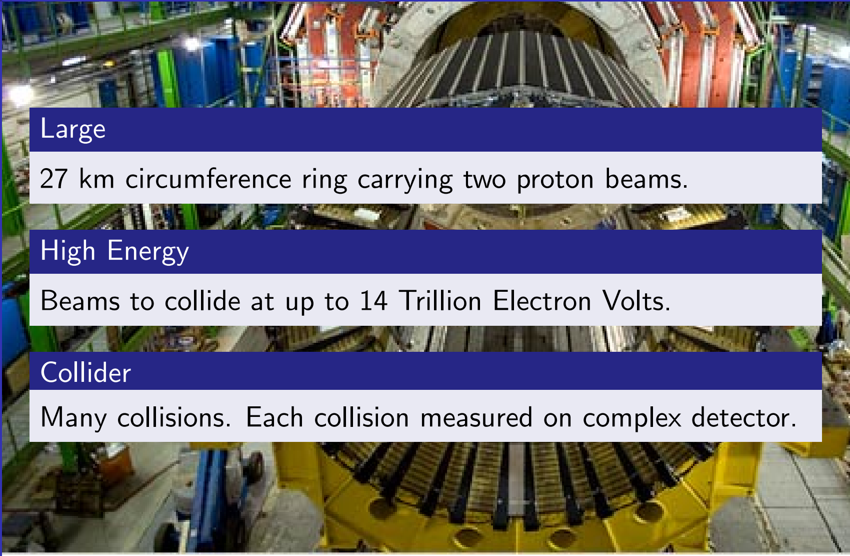
Large

27 km circumference ring carrying two proton beams.

High Energy

Beams to collide at up to 14 Trillion Electron Volts.

LHC



Large

27 km circumference ring carrying two proton beams.

High Energy

Beams to collide at up to 14 Trillion Electron Volts.

Collider

Many collisions. Each collision measured on complex detector.

A Marked Poisson Process Model

- Model data as Poisson Process of `events` in time.

A Marked Poisson Process Model

- Model data as Poisson Process of `events` in time.
- At each event measure a response X – the marks.

A Marked Poisson Process Model

- Model data as Poisson Process of `events` in time.
- At each event measure a response X – the marks.
- The event space is huge, huge, huge.

A Marked Poisson Process Model

- Model data as Poisson Process of `events` in time.
- At each event measure a response X – the marks.
- The event space is huge, huge, huge.
- Given times of events, marks are `nearly` iid.

A Marked Poisson Process Model

- Model data as Poisson Process of `events` in time.
- At each event measure a response X – the marks.
- The event space is huge, huge, huge.
- Given times of events, marks are `nearly` iid.
- Collapse data over time to get sample of N values of X_i.

A Marked Poisson Process Model

- Model data as Poisson Process of `events` in time.
- At each event measure a response X – the marks.
- The event space is huge, huge, huge.
- Given times of events, marks are `nearly` iid.
- Collapse data over time to get sample of N values of X_i.
- Poisson process on the mark space.

Hypotheses

- Null hypothesis is

Hypotheses

- Null hypothesis is

  There is no such thing as a Higgs particle

Hypotheses

- Null hypothesis is

  There is no such thing as a Higgs particle

- More general null hypothesis is standard physics "The Standard Model".

Hypotheses

- Null hypothesis is

  There is no such thing as a Higgs particle

- More general null hypothesis is standard physics "The Standard Model".
- Alternative hypothesis is some other model of physics.

Hypotheses

- Null hypothesis is

  There is no such thing as a Higgs particle

- More general null hypothesis is standard physics "The Standard Model".
- Alternative hypothesis is some other model of physics.
- or perhaps

Hypotheses

- Null hypothesis is

  There is no such thing as a Higgs particle

- More general null hypothesis is standard physics "The Standard Model".
- Alternative hypothesis is some other model of physics.
- or perhaps

  The standard model is wrong

Hypotheses

- Null hypothesis is

  There is no such thing as a Higgs particle

- More general null hypothesis is standard physics "The Standard Model".
- Alternative hypothesis is some other model of physics.
- or perhaps

  The standard model is wrong

- or perhaps

Hypotheses

- Null hypothesis is

  There is no such thing as a Higgs particle

- More general null hypothesis is standard physics "The Standard Model".
- Alternative hypothesis is some other model of physics.
- or perhaps

  The standard model is wrong

- or perhaps

  The Higg's particle is not produced at LHC energies.

Statistical Translation of No Higgs

- Null hypothesis is N has Poisson(Λ) distribution and given N the X_i are iid with some density f.

Statistical Translation of No Higgs

- Null hypothesis is N has Poisson(Λ) distribution and given N the X_i are iid with some density f.
- Alternative is N has Poisson($\Lambda + M$) distribution and given N the X_i are iid with some density g given by

$$g = \frac{\Lambda}{\Lambda + M} f + \frac{M}{\Lambda + M} f^*$$

with $f^* \neq f$.

- Null hypothesis is N has Poisson(Λ) distribution and given N the X_i are iid with some density f.
- Alternative is N has Poisson($\Lambda + M$) distribution and given N the X_i are iid with some density g given by

$$g = \frac{\Lambda}{\Lambda + M} f + \frac{M}{\Lambda + M} f^*$$

  with $f^* \neq f$.
- The density f^* is the density of the marks in events which produce Higgs particles.

- With f^*, g and M known use Neyman Pearson:

$$\ell \equiv N \log(1 + M/\Lambda) + \sum_i \log\left(\frac{g(X_i)}{f(X_i)}\right).$$

Neyman Pearson

- With f^*, g and M known use Neyman Pearson:

$$\ell \equiv N \log(1 + M/\Lambda) + \sum_i \log\left(\frac{g(X_i)}{f(X_i)}\right).$$

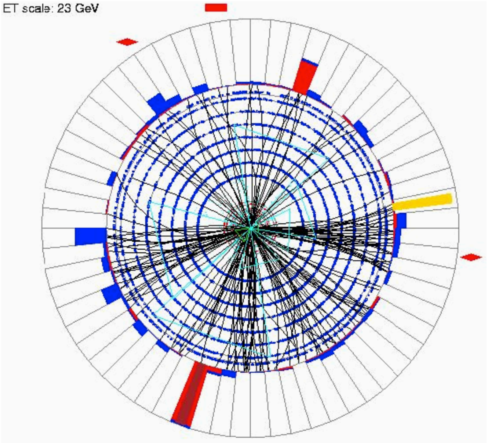
- Not much to discuss but none of the things you "know" is known exactly.

- With f^*, g and M known use Neyman Pearson:

$$\ell \equiv N \log(1 + M/\Lambda) + \sum_i \log\left(\frac{g(X_i)}{f(X_i)}\right).$$

- Not much to discuss but none of the things you "know" is known exactly.
- And there is too much data.

One Data Point



LHC & GOF
Richard Lockhart
Outline
Conclusions
LHC & Higgs
Model
Data
Bayes Power
Simulations
Real Data
Summary

The LHC experiments represent about 150 million sensors delivering data 40 million times per second. After filtering there will be about 100 collisions of interest per second.

LHC

The LHC experiments represent about 150 million sensors delivering data 40 million times per second. After filtering there will be about 100 collisions of interest per second.

- For a year's worth of data

$$\Lambda \approx 10^{15}$$

and

$$M \approx 10^{3}$$

The LHC experiments represent about 150 million sensors delivering data 40 million times per second. After filtering there will be about 100 collisions of interest per second.

- For a year's worth of data

$$\Lambda \approx 10^{15}$$

  and

$$M \approx 10^3$$

- So N provides no information?

LHC

The LHC experiments represent about 150 million sensors delivering data 40 million times per second. After filtering there will be about 100 collisions of interest per second.

- For a year's worth of data

$$\Lambda \approx 10^{15}$$

and

$$M \approx 10^3$$

- So N provides no information?
- No information at all unless f^* is nothing like f.

The LHC experiments represent about 150 million sensors delivering data 40 million times per second. After filtering there will be about 100 collisions of interest per second.

- For a year's worth of data

$$\Lambda \approx 10^{15}$$

and

$$M \approx 10^{3}$$

- So N provides no information?
- No information at all unless f^* is nothing like f.
- So what do we really do?

Triggering

- In fact f^* is known or expected to be concentrated on 'small' region of mark space.

Triggering

- In fact f^* is known or expected to be concentrated on 'small' region of mark space.
- As events are registered 'triggering" algorithms determine whether or not the event is conceivably 'interesting'.

Triggering

- In fact f^* is known or expected to be concentrated on 'small' region of mark space.
- As events are registered 'triggering" algorithms determine whether or not the event is conceivably 'interesting'.
- Triggering lowers event rate from 40M per second to roughly 100 per second.

Triggering

- In fact f^* is known or expected to be concentrated on ‘small’ region of mark space.
- As events are registered ‘triggering” algorithms determine whether or not the event is conceivably ‘interesting’.
- Triggering lowers event rate from 40M per second to roughly 100 per second.
- Now apply same model to reduced part of the mark space.

Triggering

- In fact f^* is known or expected to be concentrated on 'small' region of mark space.
- As events are registered 'triggering" algorithms determine whether or not the event is conceivably 'interesting'.
- Triggering lowers event rate from 40M per second to roughly 100 per second.
- Now apply same model to reduced part of the mark space.
- Still have

$$\Lambda \approx 10^9$$

and

$$M \approx 10^3$$

or a bit less.

Cuts

- When serious analysis begins make further `cuts`.

Cuts

- When serious analysis begins make further `cuts`.
- This amounts to trimming the mark space further to reduce Λ a lot and M hopefully only a bit.

Cuts

- When serious analysis begins make further `cuts`.
- This amounts to trimming the mark space further to reduce Λ a lot and M hopefully only a bit.
- So now I imagine my problem as posed above:

$$H_o : M = 0$$

versus

$$H_1 : M = M_0, \text{density is } g.$$

Composite alternative

- Treat f^* (so too alternative g) as imprecisely known.

Composite alternative

- Treat f^* (so too alternative g) as imprecisely known.
- Test combination of two hypotheses:

$$H_{0,\text{count}} : M = 0 \quad \text{and} \quad H_{0,\text{shape}} : g = f$$

Composite alternative

- Treat f^* (so too alternative g) as imprecisely known.
- Test combination of two hypotheses:

$$H_{0,\text{count}} : M = 0 \quad \text{and} \quad H_{0,\text{shape}} : g = f$$

- Second part is goodness-of-fit.

Composite alternative

- Treat f^* (so too alternative g) as imprecisely known.
- Test combination of two hypotheses:

$$H_{0,\text{count}} : M = 0 \quad \text{and} \quad H_{0,\text{shape}} : g = f$$

- Second part is goodness-of-fit.
- GOF is somewhat disreputable.

Composite alternative

- Treat f^* (so too alternative g) as imprecisely known.
- Test combination of two hypotheses:

$$H_{0,\text{count}} : M = 0 \quad \text{and} \quad H_{0,\text{shape}} : g = f$$

- Second part is goodness-of-fit.
- GOF is somewhat disreputable.
- No really, it is.

Composite alternative

- Treat f^* (so too alternative g) as imprecisely known.
- Test combination of two hypotheses:

$$H_{0,\text{count}} : M = 0 \quad \text{and} \quad H_{0,\text{shape}} : g = f$$

- Second part is goodness-of-fit.
- GOF is somewhat disreputable.
- No really, it is.
- Ad hoc, ad hoc, ad hoc.

A Bayesian trapped in frequentist world

- Must carry out fixed level α test.

A Bayesian trapped in frequentist world -10

- Must carry out fixed level α test.
- Must publish a protocol.

A Bayesian trapped in frequentist world

- Must carry out fixed level α test.
- Must publish a protocol.
- Wants to reject H_o.

A Bayesian trapped in frequentist world -10

- Must carry out fixed level α test.
- Must publish a protocol.
- Wants to reject H_o.
- Uses prior on alternative to design Neyman-Pearson test.

A Bayesian trapped in frequentist world

- Must carry out fixed level α test.
- Must publish a protocol.
- Wants to reject H_o.
- Uses prior on alternative to design Neyman-Pearson test.
- Maximizes expected power.

A Bayesian trapped in frequentist world

- Must carry out fixed level α test.
- Must publish a protocol.
- Wants to reject H_o.
- Uses prior on alternative to design Neyman-Pearson test.
- Maximizes expected power.

A frequentist can use the idea to design tests.

Priors on Densities

Stochastic Process Prior

Think of unknown density g as a random function ($=$ stochastic process) which happens to be positive and integrate to 1.

Stochastic Process Prior

Think of unknown density g as a random function (= stochastic process) which happens to be positive and integrate to 1.

If g is random then the joint density of $\mathbf{X} \equiv (X_1, \ldots, X_n)$ at the point $\mathbf{x} \equiv (x_1, \ldots, x_m)$ is

$$\Psi(\mathbf{x}) \equiv \mathrm{E}\left\{g(x_1) \cdots g(x_n)\right\}$$

where it is g not the x_is that are being averaged over!

Stochastic Process Prior

Think of unknown density g as a random function (= stochastic process) which happens to be positive and integrate to 1.

If g is random then the joint density of $\mathbf{X} \equiv (X_1, \ldots, X_n)$ at the point $\mathbf{x} \equiv (x_1, \ldots, x_m)$ is

$$\Psi(\mathbf{x}) \equiv \mathrm{E}\left\{g(x_1) \cdots g(x_n)\right\}$$

where it is g not the x_is that are being averaged over!
So NP Likelihood ratio test statistic is

$$(1 + M/\Lambda)^N e^{-\Lambda\Psi(\mathbf{X})}.$$

Priors on Densities

Stochastic Process Prior

Think of unknown density g as a random function ($=$ stochastic process) which happens to be positive and integrate to 1.

If g is random then the joint density of $\mathbf{X} \equiv (X_1, \ldots, X_n)$ at the point $\mathbf{x} \equiv (x_1, \ldots, x_m)$ is

$$\Psi(\mathbf{x}) \equiv \mathrm{E}\left\{g(x_1) \cdots g(x_n)\right\}$$

where it is g not the x_is that are being averaged over!
So NP Likelihood ratio test statistic is

$$(1 + M/\Lambda)^N e^{-\Lambda\Psi(\mathbf{X})}.$$

Now I need computable examples.

Gaussian Priors on log Densities

- Idea: $\epsilon Z(x)$ Gaussian process approximating (up to location and scale) log likelihood ratio.

Gaussian Priors on log Densities

- Idea: $\epsilon Z(x)$ Gaussian process approximating (up to location and scale) log likelihood ratio.
- Problem:

$$\int \exp\{\epsilon Z(x)\} f(x) dx \neq 1$$

so define our random density by

$$g(x) = \exp\{\epsilon Z(x)\} f(x) / \int \exp\{\epsilon Z(x)\} f(x) dx.$$

Gaussian Priors on log Densities

- Idea: $\epsilon Z(x)$ Gaussian process approximating (up to location and scale) log likelihood ratio.
- Problem:

$$\int \exp\{\epsilon Z(x)\} f(x) dx \neq 1$$

so define our random density by

$$g(x) = \exp\{\epsilon Z(x)\} f(x) \Big/ \int \exp\{\epsilon Z(x)\} f(x) dx.$$

- Additive invariance: wlog take $\int Z(x) f(x) dx = 0$.

Gaussian Priors on log Densities

- Idea: $\epsilon Z(x)$ Gaussian process approximating (up to location and scale) log likelihood ratio.
- Problem:
$$\int \exp\{\epsilon Z(x)\} f(x) dx \neq 1$$
so define our random density by
$$g(x) = \exp\{\epsilon Z(x)\} f(x) / \int \exp\{\epsilon Z(x)\} f(x) dx.$$
- Additive invariance: wlog take $\int Z(x) f(x) dx = 0$.
- Problem: denominator. Solution: choose ϵ depending on sample size.

Principles used here

1. Statistical tests are useful when $0.05 < \text{power} < 1$: alternatives of interest are neither indetectably nor grossly different from the null hypothesis.

Principles used here

1. Statistical tests are useful when $0.05 < \text{power} < 1$: alternatives of interest are neither indetectably nor grossly different from the null hypothesis.
2. Good tests are designed to be sensitive to alternatives likely to arise in practice.

Principles used here

1. Statistical tests are useful when $0.05 < \text{power} < 1$: alternatives of interest are neither indetectably nor grossly different from the null hypothesis.
2. Good tests are designed to be sensitive to alternatives likely to arise in practice.
3. The purpose of computing large sample limits is approximation.

Principles used here

1. Statistical tests are useful when $0.05 < \text{power} < 1$: alternatives of interest are neither indetectably nor grossly different from the null hypothesis.
2. Good tests are designed to be sensitive to alternatives likely to arise in practice.
3. The purpose of computing large sample limits is approximation.
4. Data structure and model at hand can be embedded in **any** convenient sequence to get approximation.

Principles used here

1. Statistical tests are useful when $0.05 < \text{power} < 1$: alternatives of interest are neither indetectably nor grossly different from the null hypothesis.
2. Good tests are designed to be sensitive to alternatives likely to arise in practice.
3. The purpose of computing large sample limits is approximation.
4. Data structure and model at hand can be embedded in **any** convenient sequence to get approximation.
5. Pick sequence to get quick convergence to computable limit!

I believe the problem will be interesting -6

- Principle: design tests for good properties when alternative detectably but not obviously different from null.
- Precise large sample version: take $\epsilon = a/\sqrt{M}$.
- Log-likelihood ratio given Z is approximately

$$\sum \log\{g(x_i)/f(x_i)\} \approx W \equiv \frac{a\sum Z(x_i)}{\sqrt{M}} - a^2 \int Z^2(u)f(u)du/2$$

Now: compute marginal joint density of

$$X_1, \ldots, X_n$$

by averaging density over Z.

Unexplained result of computation

Test rejects for large values of

$$e^{-M}(1 + M/\Lambda)^N \exp\{\text{GOF Statistic}(N, M)\}$$

where “GOF statistic” is generalization of EDF type tests.

Unexplained result of computation

Test rejects for large values of

$$e^{-M}(1 + M/\Lambda)^N \exp\{\text{GOF Statistic}(N, M)\}$$

where "GOF statistic" is generalization of EDF type tests. Actual formula is a secret.

Unexplained result of computation

Test rejects for large values of

$$e^{-M}(1 + M/\Lambda)^N \exp\{\text{GOF Statistic}(N, M)\}$$

where "GOF statistic" is generalization of EDF type tests.
Actual formula is a secret.
Because it is hideous.

Unexplained result of computation

Test rejects for large values of

$$e^{-M}(1 + M/\Lambda)^N \exp\{\text{GOF Statistic}(N, M)\}$$

where "GOF statistic" is generalization of EDF type tests.
Actual formula is a secret.
Because it is hideous.
But I do have a formula in terms of eigenvalues, eigenfunctions, non-centrality parameters and data.

Unexplained result of computation

Test rejects for large values of

$$e^{-M}(1 + M/\Lambda)^N \exp\{\text{GOF Statistic}(N, M)\}$$

where "GOF statistic" is generalization of EDF type tests.
Actual formula is a secret.
Because it is hideous.
But I do have a formula in terms of eigenvalues, eigenfunctions, non-centrality parameters and data.
Large sample approximate P-values are computable.

Composite Null: One Example

- Specify alternative in two pieces.
- Alternative $g(x) = f(x, \theta) \exp\{\epsilon Z(x, \theta)\}$.
- Apply prior π_1 to θ.
- Prior on alternative decomposed into conditional alternative given θ averaged over θ.
- Null: apply prior π_0.

Structure of Likelihood Ratio -3

- Neyman-Pearson likelihood ratio is product of two terms.
- Conditional joint density of $X_1, \ldots, X_n$ given θ is

$$\Psi(\mathbf{x}, \theta)$$

computed as before for $Z(x, \theta)$.
- First term in likelihood ratio is posterior expectation

$$\Psi(\mathbf{x}) \equiv \int \Psi\left(\mathbf{x}, \theta\right) \pi_1(d\theta|\mathbf{x}).$$

where $\pi_1(d\theta|\mathbf{x})$ is posterior computed under null for prior π_1.
- Second is ratio of marginals under null:

$$\frac{\int \prod f(x_i, \theta)\pi_1(\theta)d\theta}{\int \prod f(x_i, \theta)\pi_0(\theta)d\theta}.$$

- Adjust π_0 to get level α.

Simulation Results

Simulation Results -2

"Not today, sir."

Data Analysis

Data Analysis

Are you kidding?

The Real(er) Problem

- The density f^* is computed by measurement, approximation, and simulation.

The Real(er) Problem

- The density f^* is computed by measurement, approximation, and simulation.
- The density g is computed by measurement, approximation, and simulation – from extreme tails.

The Real(er) Problem

- The density f^* is computed by measurement, approximation, and simulation.
- The density g is computed by measurement, approximation, and simulation – from extreme tails.
- The intensity M is an unknown parameter of interest.

The Real(er) Problem

- The density f^* is computed by measurement, approximation, and simulation.
- The density g is computed by measurement, approximation, and simulation – from extreme tails.
- The intensity M is an unknown parameter of interest.
- In fact M is product of parameter of interest `cross-section` and other things — prior information available for the other things.

The Real(er) Problem

- The density f^* is computed by measurement, approximation, and simulation.
- The density g is computed by measurement, approximation, and simulation – from extreme tails.
- The intensity M is an unknown parameter of interest.
- In fact M is product of parameter of interest `cross-section` and other things — prior information available for the other things.
- But Bayes on cross-section not tolerable to physicists.

The Real(er) Problem

- The density f^* is computed by measurement, approximation, and simulation.
- The density g is computed by measurement, approximation, and simulation – from extreme tails.
- The intensity M is an unknown parameter of interest.
- In fact M is product of parameter of interest `cross-section` and other things — prior information available for the other things.
- But Bayes on cross-section not tolerable to physicists.
- Null intensity Λ is computed by approximation and simulation.

The Real(er) Problem

- The density f^* is computed by measurement, approximation, and simulation.
- The density g is computed by measurement, approximation, and simulation – from extreme tails.
- The intensity M is an unknown parameter of interest.
- In fact M is product of parameter of interest `cross-section` and other things — prior information available for the other things.
- But Bayes on cross-section not tolerable to physicists.
- Null intensity Λ is computed by approximation and simulation.
- The analysis will be carried out over several energy ranges; multiple comparisons.

- An experimenter wanting to provide evidence convincing to others should use prior on alternative and publish a data analysis protocol.

Summary, questions, extensions 1

- An experimenter wanting to provide evidence convincing to others should use prior on alternative and publish a data analysis protocol.
- A fairly natural prior leads to pooling quadratic GOF tests with a test based on N.

Summary, questions, extensions

- An experimenter wanting to provide evidence convincing to others should use prior on alternative and publish a data analysis protocol.
- A fairly natural prior leads to pooling quadratic GOF tests with a test based on N.
- I need to find more priors.

Summary, questions, extensions

- An experimenter wanting to provide evidence convincing to others should use prior on alternative and publish a data analysis protocol.
- A fairly natural prior leads to pooling quadratic GOF tests with a test based on N.
- I need to find more priors.
- The data analysis null hypothesis is uncertain; is Bayes avoidable?

Summary, questions, extensions

- An experimenter wanting to provide evidence convincing to others should use prior on alternative and publish a data analysis protocol.
- A fairly natural prior leads to pooling quadratic GOF tests with a test based on N.
- I need to find more priors.
- The data analysis null hypothesis is uncertain; is Bayes avoidable?
- What is the foundational status of "systematics"?

Summary, questions, extensions

- An experimenter wanting to provide evidence convincing to others should use prior on alternative and publish a data analysis protocol.
- A fairly natural prior leads to pooling quadratic GOF tests with a test based on N.
- I need to find more priors.
- The data analysis null hypothesis is uncertain; is Bayes avoidable?
- What is the foundational status of "systematics"?
- Blind analysis?