PhyStat-ν 2016

A Statistical Summary

David A. van Dyk

Statistics Section, Imperial College London

Tokyo, June 2016

Disclaimer

Statisticians have "discussions" (of talks) rather than "summaries" (of conferences).

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

Other interesting topics.... but my time is limited.
E.g., Hierarchical models (WILKENSON, COLLIN)

I'm sure I'm missing important contributions!

Correct me if I mischaracterize your work!!

....and forgive me if I stand on my soap box a bit....

Outline

Outline

Deconvolving your statical thinking

First: Identify scientific objective and derive likelihood (MICHAEL)

Deriving Statistical Methods:

- minimize χ^2, likelihood-based, Bayes theorem, others?

Evaluating Statistical Methods: *(What are the operating characteristics?)*

- frequency based: coverage, error rates, mean square error
- Bayesian: complete summary of information?

(or Bayesian decision theory, MICHAEL)

Computation: (MICHAEL)

- What to compute vs. How to compute it.

... and asymptotic... again MICHAEL

Example 2: Profile or Marginalize?

Asher: Discussed the relative advantages of profiling and marginalizing.

Others confront same issue by

- Profiling in RENO analyses (SUNNY)
- or Conditioning on best fit in IceCube and SK. (JOAO PEDRO)

Example 2: Profile or Marginalize?

Asher: Discussed the relative advantages of profiling and marginalizing.

Bob: Asked about the coverage rates?

Example 2: Profile or Marginalize?

Asher: Discussed the relative advantages of profiling and marginalizing.

Bob: Asked about the coverage rates.

Biller: Might ask about likely values of parameter given *this* data.

... or about proper accounting of uncertainty.

Frequentist or Bayesian?

Do you have to choose??

- Bayes prescribes method — Frequency evaluates method.
- Frequency evaluation of Bayesian methods.
  I like intervals to include true values... at least once in a while.
- Model fitting: often little difference in fits and errors.
  ... Biller's examples not withstanding.
- Why not control detection error
  and assess probability of new physics?
- Why throw away half of your tool box?

I'm impressed with the openness of neutrino researchers to both Bayesian and Frequency based methods.

- Lots of Bayesian procedures (KABOTH, KLEESIEK, BERGSTROM, HAGEL, VILLAESCUSA-NAVARRO, SHIMIZU, CALLAND, BILLER, COLLIN, ETC.)
- My experience with cosmologists and particle physicists.

Outline

Statistical Framework for Discovery

Hypothesis Testing

H_0: The null hypothesis (e.g., no CP-violoation, $\delta = 0$)

H_1: The alternative hypothesis (e.g., CP-violation)

- Without further evidence, H_0 is presumed true.
- "Deciding" on H_1 means scientific discovery: new physics.

Other Model Hypothesis Tests (EMILIO)

Selection: May be no presumed model. (e.g., normal/inverted hierarchy)

Checking: Is data consistent with model? H_1 not required.

... a.k.a., "goodness of fit"

In all cases, there are different statistical approaches, e.g., Frequentist and Bayesian.

Statistical Criterion for Discovery

The most common criterion is the p-value,

$$\text{p-value} = \Pr\left(T(y) \geq T(y_{\text{obs}}) \mid H_0\right) \text{ with } T = \text{test statistic}$$

But....

Statistical Criterion for Discover

The most common criterion is the p-value,

$$\text{p-value} = \Pr\left(T(y) \geq T(y_{\text{obs}}) \mid H_0\right) \text{ with } T = \text{test statistic}$$

But....

Statistical Criterion for Discover

The most common criterion is the p-value,

$$\text{p-value} = \Pr\left(T(y) \geq T(y_{\text{obs}}) \mid H_0\right) \text{ with } T = \text{test statistic}$$

But....

(ASA Statement on Statistical Significance and P-values)
February 5, 2016

The Problem with P-values

The misuse of P-values:

- **Do not measure relative likelihood of hypotheses.**
- Cannot measure practical significance / effect size.
- Large p-values do not validate H_0.
- May depend on bits of H_0 that are of no interest.
- Single filter for publication / judging quality of research.
- Cherry-picking results based on p-value / publication bias.
- "Replication Crisis"

Reviewers, Editors, and Readers want a simple black-and-white rule: $p < 0.05$, or $> 5\sigma$.

But... "Statistics is about quantifying uncertainty, not expressing certainty."
—Belancourt

5σ Discovery Threshold

5σ is required for "discovery" (LOUIS)

- High profile false discoveries led to conservative threshold
- Treat Higgs mass as known (multiple-testing) (SARA)
- "What would you have done had you had different data"
- **Calibration, systematic errors, and model misspecification**
- Of course **cranking down α does not address these issues**
- **Address systematics** via generative models? (MICHAEL)

"In particle physics, this criterion has become a convention ... but should not be interpreted literally [1] *."*

Bob: *"Two* 3.5σ *results are better than one* 5σ *result."* (SUNNY)

[1] Glossary in the *Science* review of the 2012 CMS and ATLAS discoveries.

5σ Discovery Threshold

5σ is required for "discovery" (LOUIS)

- High profile false discoveries led to conservative threshold
- Look elsewhere effect (multiple-testing)
- "What would you have done had you had different data"
- **Calibration, systematic errors, and model misspecification**
- Of course **cranking down α does not address these issues**
- **Address systematics** via generative models? (MICHAEL)

"In particle physics, this criterion has become a convention ... but should not be interpreted literally[2]*."*

Bob: *"Two* 3.5σ *results are better than one* 5σ *result."* (SUNNY)
van Dyk: *"Calibrated* 3.5σ *result is better than uncalibrated* 5σ*."*

[2] Glossary in the *Science* review of the 2012 CMS and ATLAS discoveries.

The Problem with Priors

Bayesian methods have challenges of their own.

$$\text{Bayes Factor} = \frac{p_0(y)}{p_1(y)} \text{ with } p_i(y) = \int p_i(y|\theta)p_i(\theta)d\theta.$$

$$\Pr(H_0 \mid y) = \frac{p_0(y)\pi_0}{p_0(y)\pi_0 + p_1(y)\pi_1} = \frac{\pi_0}{\pi_0 + \pi_1(\text{Bayes Factor})^{-1}}$$

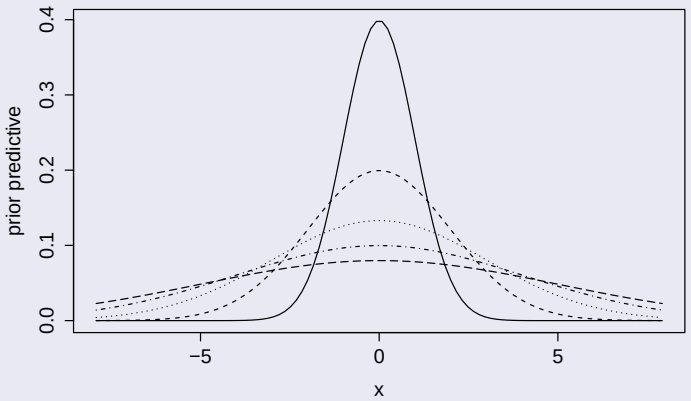
Example:

Likelihood: $y \sim \mathrm{N}(\mu, 1)$

Test: $\mu = 0$ vs $\mu \neq 0$

Prior Dist'n: $\mu \sim \mathrm{N}(0, \tau^2)$

Prior Pred.: $y \sim \mathrm{N}(0, 1 + \tau^2)$



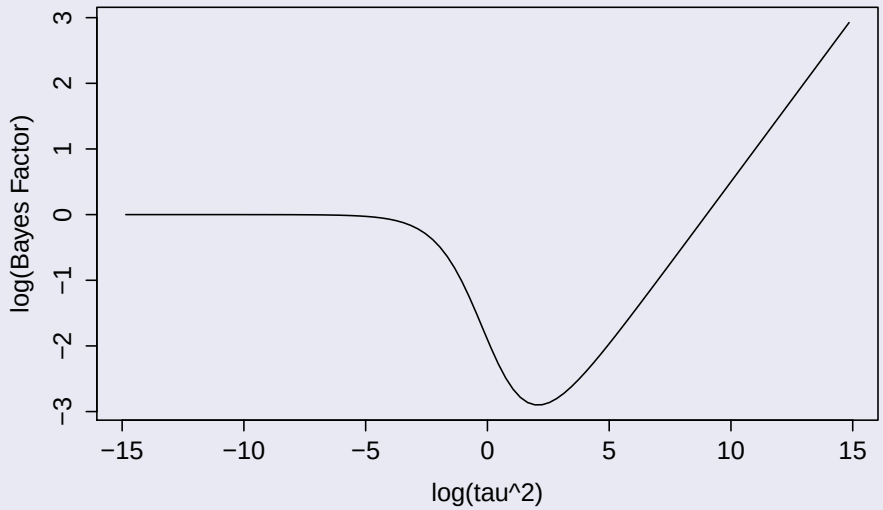
Value of $p_1(y)$ depends on τ^2!

Choice of Prior Matters!

Bayes Factor (Asher, Johannes, Emilio, Michael)

$$H_0: \quad y \sim \mathrm{N}(0, 1).$$
$$H_A: \quad y \sim \mathrm{N}(0, 1 + \tau^2).$$

- Observe $y = 3$
- log(Bayes Factor)



Must think hard about choice of prior and report!

Bayes Factors and Likelihood Ratios

Likelihood Ratio optimizes parameters, whereas Bayes Factor marginalizes.

$$\text{Likelihood Ratio} = \frac{\max_{\theta_0} p_0(y \mid \theta_0)}{\max_{\theta_1} p_1(y \mid \theta_1)} \neq \text{Bayes Factor} = \frac{\int p_0(y \mid \theta_0)\, p(\theta_0)\, \mathrm{d}\theta_0}{\int p_1(y \mid \theta_1)\, p(\theta_1)\, \mathrm{d}\theta_1}$$

(ASHER) *....unless there are no parameters under either model.*

A Bayesian Occam's Razor (BOB,MICHAEL)

- Suppose $p(\theta_i)$ are both essentially flat over range where corresponding likelihoods are non-negligible.

$$\text{Bayes Factor} = \frac{\int p_0(y \mid \theta_0)\, p(\theta_0)\, \mathrm{d}\theta_0}{\int p_1(y \mid \theta_1)\, p(\theta_1)\, \mathrm{d}\theta_1} \approx \frac{p(\hat{\theta}_0) \int p_0(y \mid \theta_0)\, \mathrm{d}\theta_0}{p(\hat{\theta}_1) \int p_1(y \mid \theta_1)\, \mathrm{d}\theta_1}$$

- The term $p(\hat{\theta}_0)/p(\hat{\theta}_1)$ is sensitive to dimension and scale. (JOHANNES)
  - At mode, multivariate normal prior $\propto 1/|\Sigma|^{d/2}$.
- Bayes Factor penalizes larger models. *...and depends strongly on choice of prior.*
- Don't hide your priors!

Model Fitting vs Model Selection/Checking

Model Fitting

- Specify one model, fit parameters, estimate uncertainty.
- Frequency and Bayesian methods tend to agree.
- Choice of prior distribution is often not critical.
- Some *"model selection"* tasks can be accomplished via model fitting and computing
  - confidence or credible intervals.
  - probability of region in parameter space.
  - Nice example: use mixture model to avoid tests (RICHARD)
    (for event reconstruction using RJMCMC)
- Computation is often much easier too!
  (e.g., Bayes Factors / Monte Carlo p-values, aka toys)

Model fitting is always much easier.
Avoid model selection if you can!

A Taxonomy of Tests

Different types of tests involve different methods

- Simple vs Simple
- Nested I: One-sided tests
- Nested II: Precise null (with H_0 on boundary)
- Nested III: Parameters undefined under H_0
- Non-Nested

In each case we can consider relative advantages of p-values an Bayesian methods.

Simple vs Simple

Simple vs Simple

- H_0 and H_1 are fully specified: no unknown parameters.
- E.g., normal hierarchy vs inverted hierarchy.
- Bayes Factor $= \frac{p_0(y)}{p_1(y)} =$ Likelihood Ratio

P-values: log(Likelihood Ratio) $\sim$ NORMAL (SARA)

(Must use Monte Carlo to specify two null distributions.)

Bayesian: No problem with priors!

Methods give consistent results.

Everything is simple, but are the models ever fully specified in practice?

Nested I: One-sided Tests

Nested I: One-sided Tests

- $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta \geq \theta_0$.
- E.g., $H_0 : \Delta m^2_{32} \leq 0$ versus $H_1 : \Delta m^2_{32} > 0$.

P-values: p-value $= \sup_{\theta \leq \theta_0} \Pr\left(T(y) \geq T(y_{\text{obs}}) \mid \theta\right)$ *(Use Wilks Thm.)*

Bayesian: Avoid $p_0(y)$ and $p_2(y)$: $\Pr(H_0 \mid y) = \Pr(\theta \leq \theta_0 \mid y)$.

- Requires only one model and one prior specification.
- *Can incorporate external knowledge into Bayesian analysis via prior, e.g.,* $|\Delta m^2_{32}| = 2.43 \pm 0.13$.

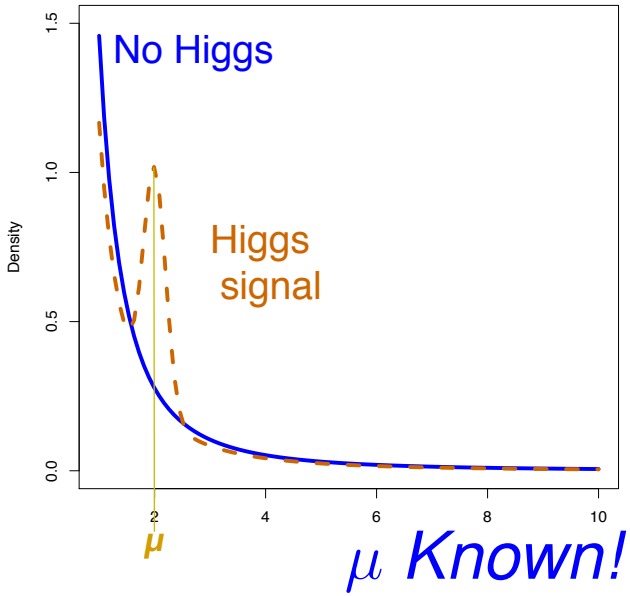
Again methods give consistent results.

Nested II: Precise Null (with H_0 on boundary)

Nested II: Precise Null

$$f(y_i|\theta) = (1-\lambda)f_0(y_i|\alpha) + \lambda f_1(y_i|\mu)$$
$$= \text{background} + \text{Higgs}$$

- f_1 is fully specified: i.e., μ is known.
- $H_0 : \lambda = 0$ (no discovery)
  $H_1 : \lambda > 0$ (discovery)
- E.g., IceCube signal search, fixed source parameters (Lu)



μ Known!

P-values: LRT: Wilks does not apply, use Chernoff.

Bayesian: Choice of prior on λ matters!
P-values $\ll \Pr(H_0 \mid y)$.

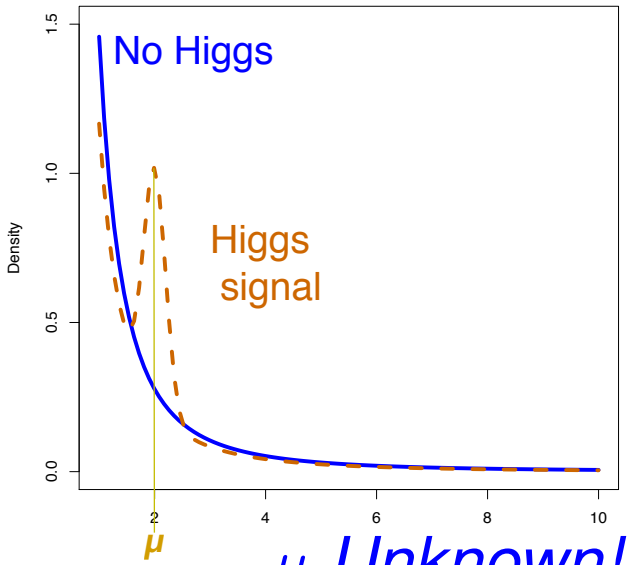
Unfortunately, p-values and $\Pr(H_0 \mid y)$ *differ significantly.*

Nested III: Precise Null (with Parameters Undefined Under H_0)

Nested III: μ undefined under H_0

$$f(y_i|\theta) \quad = \quad (1-\lambda)f_0(y_i|\alpha) + \lambda f_1(y_i|\mu)$$

- μ unknown, no value under H_0.
- $H_0 : \lambda = 0$ versus $H_1 : \lambda > 0$
- E.g., IceCube signal search, source parameters unknown (Lu)



μ *Unknown!*

P-values: Bound global p-value

- Look elsewhere effect, method of GV.
- Bonferroni versus Markov bounds. (Sara)

Bayesian: Choice of priors on λ and μ matter!

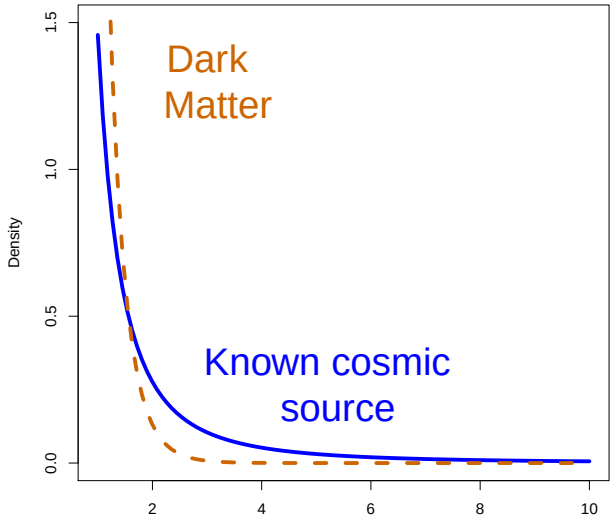
- Use prior on μ to correct for LEE.

Again, p-values and $\Pr(H_0 \mid y)$ *differ significantly.*

Non-nested models

Non-nested models

- two *parameterized* non-nested models
- H_0: γ-ray energy of known cosmic sources
  H_1: γ-ray energy of dark matter.
- H_0: normal hierarchy
  H_1: inverted hierarchy .
- Is there a *null* model?



P-values: Embed in mixture model and bound global p-value or Monte Carlo (toys). (SARA)

Bayesian: No problems in principle. (DISCUSSED BELOW!)

... but choice of prior may cause difficulties.

Again, p-values and $\Pr(H_0 \mid y)$ *may differ significantly.*

Outline

Normal Hierarchy versus Inverted Hierarchy

Non-nested parameterized models (YUFENG, EMILIO, RONALD, ETC.)

We can compare:

H_0 : normal hierarchy	$H_0 : \Delta m^2_{32} \leq 0$	$H_0 : \lambda = 0$
H_1 : inverted hierarchy	$H_1 : \Delta m^2_{32} > 0$	$H_1 : \lambda > 0$

where $f(y) = \lambda f(y \mid \mathrm{NH}) + (1-\lambda) f(y \mid \mathrm{IH})$.

Methods:

- P-value using combined model (SARA, JOHANNES, EMILIO)
- Bayes Factors (JOHANNES, EMILIO)
  - Not highly prior dependent (JOHANNES)
  - "Priors must be agreed on, especially for m^2_{23}" (MARCO)
- Posterior predictive p-values. (power??) (DAVIDE)

Normal versus Inverted Hierarchy: p-value

Non-nested parameterized models

H_0 : normal hierarchy $\quad H_0 : \Delta m^2_{32} \leq 0 \quad H_0 : \lambda = 0$

H_1 : inverted hierarchy $\quad H_1 : \Delta m^2_{32} > 0 \quad H_1 : \lambda > 0$

Computing a p-value using LRT

- If there are no nuisance parameters, it's easy. But....
- With nuisance parameters (δ or θ_{23}), no general solution.
  - *What is null distribution of $\hat{\delta}$ when fitting H_1?* (RONALD)
  - Single nuisance parameter (δ), LRT is Gaussian (EMILIO)
  - Still low power owing to degeneracy. (DAVIDE, EMILIO)
  - More nuisance parameters, LRT dist'n unknown. (EMILIO)

Are we back to Monte Carlo? 5σ??

... as Ronald nicely illustrated.

Normal versus Inverted Hierarchy: Easier Way?

Non-nested parameterized models

H_0 : normal hierarchy	$H_0 : \Delta m^2_{32} \leq 0$	$H_0 : \lambda = 0$
H_1 : inverted hierarchy	$H_1 : \Delta m^2_{32} > 0$	$H_1 : \lambda > 0$

Is there an easier solution??

Why not just compute $\Pr(H_0 \mid y) = \Pr(\Delta m^2_{32} \leq 0 \mid y)$?

In this case Bayes Factor is particularly easy because

$$\text{Bayes Factor} = \frac{\Pr(\Delta m^2_{32} \leq 0 \mid y)\pi_1}{\Pr(\Delta m^2_{32} > 0 \mid y)\pi_0}$$

Only one model and one prior, easy to compute, not sensitive to prior... what's not to like?

Bayesian solution is easier in this case.

CP-violation

Test: $H_0 : \delta \in \{0, \pi\}$ versus $H_1 : \delta \notin \{0, \pi\}$

p-value

- Chernoff's theorem applies...
  but insufficient data for asymptotics.
- Monte Carlo required, challenges for global fits. (JOHANNES)
- More data required! (For 5σ??)

Bayes Factor

- Computed via Savage-Dickey. (JOHANNES)
- Sensitive to prior on δ, but finite support.

Again, Bayesian solution is easier (with limited data).

Still Easier: *Report a confidence interval for δ.*

But Feldman-Cousins not feasible for global fits...

Strategies

What is a physicists to do?

- Controlling false discovery is critical in physical sciences.
- Comparing p-values with a predetermined significant level can control false discovery.... *if used with care, e.g., no cherry picking!*
- When confronted with small p-values researchers *...even statisticians!!...* may believe H_0 is unlikely.
- Bayesian solutions can better quantify likelihood of H_0 / H_1.
- **Solution:** Compute both *global* p-value *and* Bayes Factor.

Careful

1. *global corrections for p-values*
2. *choice and validation of prior distributions*

remain challenging!