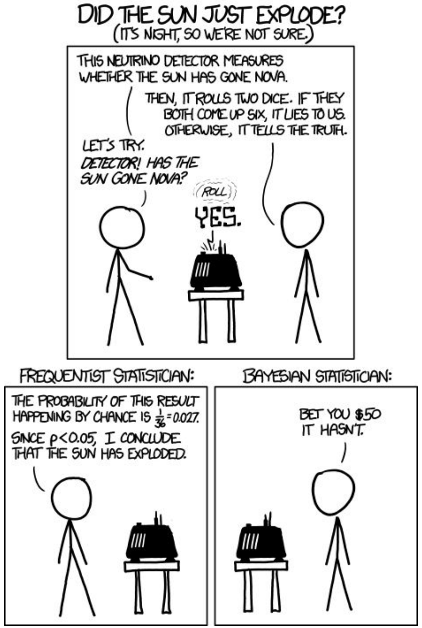
Bayesian Statistics and its Applications



Reed Watson
Physics 290E Fall 2017

Introduction

- Motivation
- Review of statistics
- Basics of Frequentist vs. Bayesian interpretations
- Pros and cons
- Nuisance parameters
- MCMC
- Examples

Motivation

- Bayesian statistics is an increasingly popular, though contentious, statistical interpretation.
- There exists confusion between Frequentist and Bayesian intervals.
- Full Bayesian treatment has been used in branching ratio studies at CDF [11], Higgs cross section limits [12], supersymmetry constraints[13].

Reverend Bayes. Source: wikipedia.org

Machinery of Statistics -Hypothesis Testing

- 2 or more hypotheses: H_0, H_1, etc. Falsely reject H_0 with frequency **α** (*significance level)*, False reject H_1 with frequency **β** (1- β is the *power)[1]*.
- Perform an experiment, obtain data ***x*****.** *Test statistic* t(**x**): characterizes deviation of **x** from expectation.
  - Neyman-Pearson Lemma: Likelihood ratio **λ** is the most powerful test statistic, but often impractical with highly correlated data.
- Significance test: f(t | H_0) is determined by data.
  - α, β, are determined beforehand, p is an outcome of the experiment.
  - $p < \alpha$ is criterion for rejection of H_0
- Parameter determination-> **ϴ** to be determined.
- Want an estimator for **ϴ** as well as a measure of uncertainty around it (in the case of a positive result), or simply an upper bound for a null result.

$$\lambda = \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)}$$

$$p = \int_{t_{obs}}^{\infty} f(t|H_0)dt$$

Probability and Intervals

Frequentist:

- P(A) means that identically repeating an experiment an infinite number of times, event A is observed with frequency P(A).
- Frequentist *confidence interval*: True observable **ϴ**. Based on data, we set a “confidence interval,” which contains ϴ with a frequency (1-α).
- **α**: *significance level.* Typically either .1 or .05 for parameter estimation.

Bayesian:

- P(A) quantifies the reasonable expectation that event A will occur given all available information.
- *Credible interval*: ϴ belongs to a probability (belief) distribution. Based on the observed data a fraction (1-α) of the distribution is within the interval.
- *prior (**π**)* and *posterior (**P**)* distributions reflect our belief about a variable before and after the experiment
- Based on Bayes Theorem:

$$P(\theta|\mathbf{x}) = \frac{P(\mathbf{x}|\theta)\pi(\theta)}{\pi(\mathbf{x})}$$

More on Bayes Theorem

Why isn’t the likelihood function it’s own inverse?

Medical Example: Some test for a disease has a 10% false positive rate. The disease is incident in 1% of population. You test positive for the disease. How likely is it that you actually have the disease?

A: 0.1 * 0.01 = 0.001

Our brains work like this, constantly updating our prior assumptions. Part of the reason why frequentist intervals are misinterpreted.

More on Intervals

Frequentist Interval (from PDG[1])

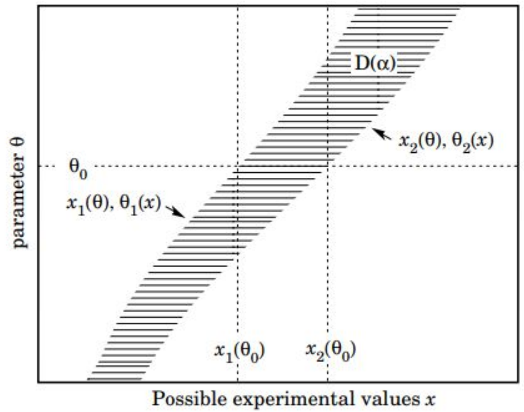
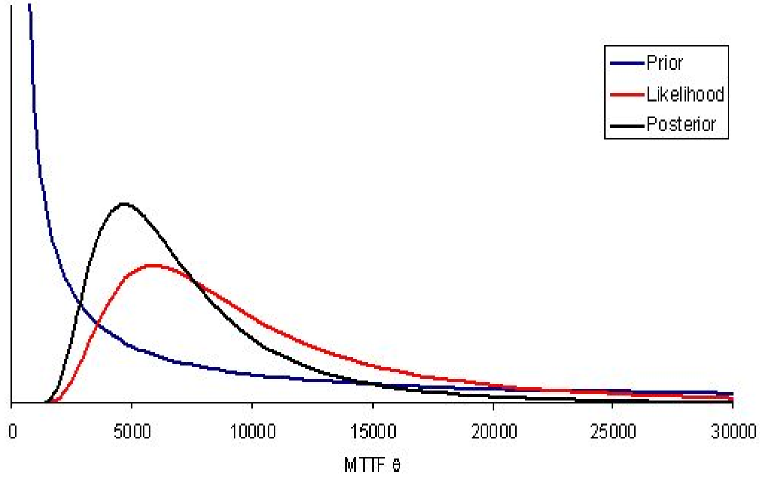


Figure 38.3: Construction of the confidence belt (see text).

Vertical lines = experimental results. Horizontal lines = confidence interval for ϴ with area α. The location is controlled by either endpoint convention or significance tests for each point (Likelihood ratio test = Feldman Cousins). For complicated models, monte carlo is used to generate the bands.

Bayesian Intervals (source: epixanalytics.com)



Likelihood is the result of experiment P(x|ϴ) . π(x) is technically in the denominator but normalization takes care of this. The interval is chosen such that the area under the curve is 1-α. Doesn't satisfy coverage in general, but under certain priors it does.

Even more comparison

Frequentist:

$$\int_{-2\ln\lambda}^{\infty} \mathcal{L}(\mathbf{x}|s) \leq \alpha$$

Bayesian:

$$1 - \alpha = \frac{\int_{-\infty}^{s_{\mathrm{up}}} \mathcal{L}(\mathbf{x}|s)\pi(s)ds}{\int_{-\infty}^{\infty} \mathcal{L}(\mathbf{x}|s)\pi(s)ds}$$

Objective vs. Subjective Bayesian Priors

- Subjective: Use all available information. Wears the cognitive aspects on its sleeves
  - . The uncertainties of the experiment are incorporated into the priors, which are then subject to defense.
  - Normalizable by construction.
  - Reflect biases that already exist.
- Objective: priors must be **noninformative**-minimal effect on posterior.
  - Mathematically: flat over areas of high likelihood, small in areas of low likelihood.
  - Ideally, given a certain type of data, everybody agrees on a type of prior, eliminating bias [4].

Source: Michael Kloran, kloran.com

Principles of objective Bayesian Priors [10]:
- Insufficient Reason
- Invariance
- (approximate) Coverage matching
- Maximal missing information
- Coherent
- Robust

Jeffrey's Prior and Fischer Information

- Obj. Bayesians desire a noninformative prior.
- Dependence on reparameterization is considered "informative."
- Prior is given by:

$$\pi(\theta) = \sqrt{\det \mathcal{I}(\theta)}$$

- Where I is the **Fischer Information Matrix.**
  - Represents the amount of information carried in the data about **ϴ.**
  - Independent of parameterization, π(**ϴ**)= π(**ϴ**^{2}), and the data **x,** depends only on likelihood function.
  -

$$\mathcal{I}_{ij}(\theta) = E[(\frac{\partial}{\partial \theta_i} \ln f(\mathbf{x}|\theta))(\frac{\partial}{\partial \theta_j} \ln f(\mathbf{x}|\theta))|\theta]$$

<u>Examples:</u>

- Gaussian with mean μ, spread σ:
  - uniform prior π(μ) =1.
  - π(σ) =1/σ.
- Poisson with rate parameter λ:

$$\pi(\lambda) = 1/\sqrt{\lambda}$$

- Bernoulli Trial with success probability γ:

$$\pi(\gamma) = 1/\sqrt{\gamma(1-\gamma)}$$

Notice that Gaussian and poisson examples are **improper**. This is okay as long as there's a cutoff (introduces bias), or the posterior is proper.

Similarities between Bayesian and Frequentist Values

- Poisson distribution: Uniform prior with a cutoff at ϴ = 0 and b = 0 gives the frequentist upper limit (b > 0 yields *conservative* /overcovered limits)
- Symmetry of gaussian function means that flat priors give f(x | ϴ) = f(ϴ| x) and also correspond to frequentist intervals.
- Interpretation of these differ, and these only coincide for single dimensional parameters.
- *Bernstein-von-Mises Theorem:[5]* posterior pdf centered around mean is asymptotically identical to the MLE around the true value. Covariance matrices are likewise asymptotically identical.

$$1 - \alpha = \frac{\int_0^{s_{up}} (s+b)^n e^{-(s+b)} ds}{\int_0^{\infty} (s+b)^n e^{-(s+b)} ds}$$

$$p = 1 - \alpha F_{\chi^2}(2b, 2(n+1))$$

$$s_{up} = \frac{1}{2} F_{\chi^2}^{-1}(p, 2(n+1)) - b$$

Discrepancies between the interpretations

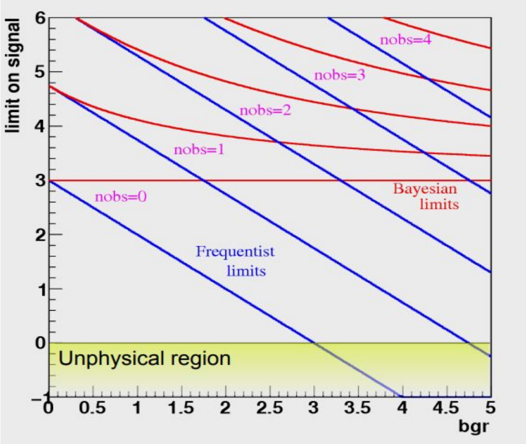
$$B_{ij} = \frac{\int f(\mathbf{x}|\theta_i, H_i)\pi(\theta_i|H_i)d\theta_\mathbf{i}}{\int f(\mathbf{x}|\theta_j, H_j)\pi(\theta_j|H_j)d\theta_\mathbf{j}}$$

- Bayes Factor: To test different hypothesis, eliminate bias in choice of prior, divide the posteriors of H_0, H_1.
- **Jeffrey-Lindley Paradox**[3]: Under certain circumstances and choices of prior, the null hypothesis can be rejected by frequentist p-values and accepted under Bayes' factor.
- Not actually a paradox:
  - Can be resolved using objective priors.
  - Frequentist asks: is H_0 consistent with data? Bayesian asks: is H_0 better than H_1 ?

Intermission: Criticism and response of Bayesians

Bayesian Criticisms:

- Subjectivity impossible to avoid.
- **Coverage** depends on priors [15].
- There really is one objective reality, not a probability distribution (veers into philosophy).



From [9]

Bayesian Response:

- Assigning a probability to anything is a good thing.
- Reference priors achieve much (but not all) in the way of objectivity.
- Coverage is an imaginary construction, you can't actually perform an infinite number of identical trials.
- More intuitive interpretation of intervals.
- Bayesian priors exclude unphysical results.

Nuisance Parameters

Nuisance parameters define the systematic uncertainties of the experiment [6]. Any uncertain value that isn't the parameter of interest is by definition a nuisance parameter. A 100% frequentist construction needs to achieve coverage for all values of v.

- Frequentist method: "profile" the nuisances with the profile likelihood ratio method:

$$\lambda_P(\theta) = \frac{\mathcal{L}(\theta, \hat{\hat{\nu}})}{\mathcal{L}(\hat{\theta}, \hat{\nu})}$$

- $-2\ln\lambda_P$ Has χ^2 distribution in the limit of large statistics (Wilk's Theorem).
- Used as a replacement test statistic.
- Numerator (profile likelihood) used as a replacement likelihood in Neyman construction.

Bayesian method: "marginalise" the nuisances.

$$\mathcal{L}_m(\mathbf{x}|\theta) = \int \mathcal{L}(\mathbf{x}|\theta, \nu)\pi(\nu)d\nu$$

- Replaces the likelihood in the posterior distribution integral [2].
- **Π** is the "updated prior" after a calibration run (posterior to that experiment).
- Can also use a test statistic Q in place of **x t**o find the distribution (used in hybrid methods- coming up).

Hybrid Bayesian/Frequentist Statistics

- Extended Cousins / Highland method: nuisance parameters are integrated over, then fed into the Neyman Construction (frequentist intervals).
  - Parameters of interest are treated in a frequentist fashion, uninteresting parameters are treated in a Bayesian fashion.
  - Nuisance pdf’s are expanded in moments (mean, variance, skew, ...)
  - Performs similarly to Bayesian limits, but is overly conservative.
- CLs Method: Use the replacement test statistic
  - Conservative metric, “modified frequentist.”
  - Uses marginalisation over nuisance parameters
  - Asymptotically equivalent to Bayesian limits in single parameter case[7].
  - Prevents problem of setting limits better than experimental sensitivity (s << b).
  - Used in LHC physics and neutrino searches.

$$Q(\theta) = \frac{p_{s+b}}{1 - p_b}$$

Cousins-Highland Method

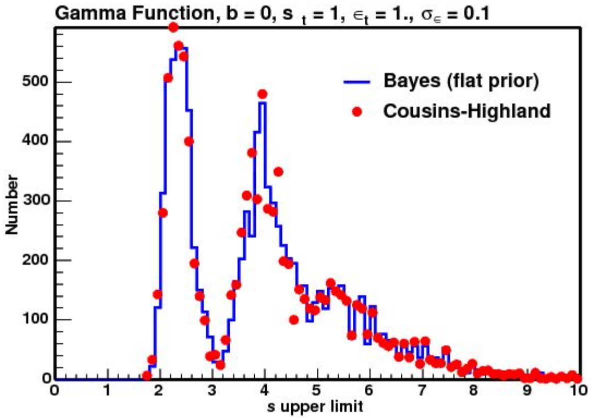


Figure 1: Distribution of limits for the case where the true signal is 1.0, the background is 0, the true efficiency is 1.0, the measurement uncertainty on the efficiency is 10%. The red histogram is the Cousins-Highland method and the blue histogram is a full Bayesian treatment with a flat prior. The peak between 2 and 3 is due to cases with zero observed events. The cases with other number of observed events give broader peaks that merge to form the rest of the distribution.

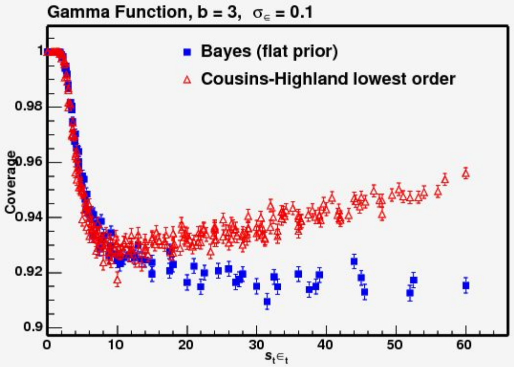


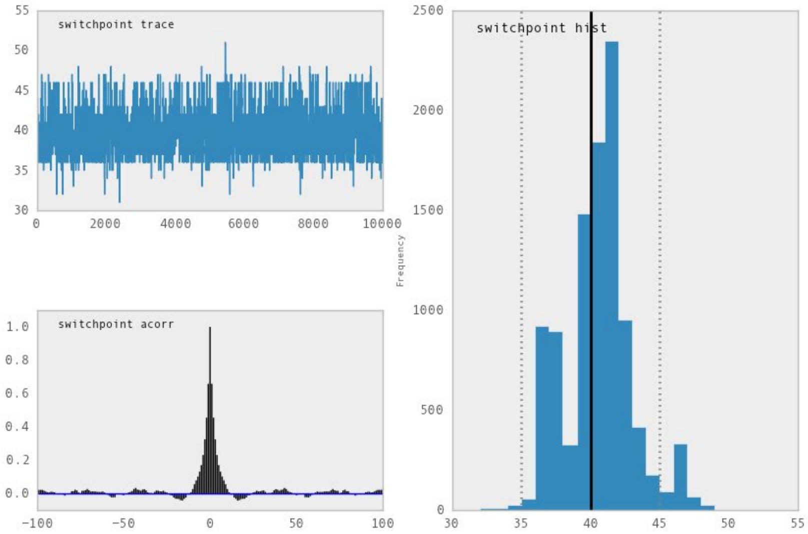
Figure 4: Coverage as a function of the product of the true efficiency and the true signal. The red points are the Cousins-Highland first order approximation, and the blue points are a full Bayesian treatment with a flat prior. The background is 3.0 and the measurement uncertainty on the efficiency is 10% for all points.

CDF [8]

Sampling posterior: Markov Chain Monte Carlo

- Bayes factor integrals with improper priors often lack analytic solutions and have large dimensionality.
- Solution: numerical integration with **MCMC**.
- Uses the **Metropolis-Hastings Algorithm** to sample the posterior distribution.
  - Start with a sample x.
  - Propose a new sample x’ bases on *jumping distribution Q(x’|x)*.
  - If posterior f(x’ | y) > f(x | y), accept it. Otherwise, accept it with frequency f(x’|y)/ f(x|y).
  - Continue until desired trace length filled.
- Issues:
  - Usually start far from minimum. Solution: Throw away the first N samples, called “burn-in.”
  - X_n correlated with x_{n+1}, so we “thin” the trace by stepping through in steps of length d.
  - Sometimes the Q(x’|x) is altered on the fly to improve acceptance ratio. “Simulated annealing.”
- Implemented in the PyMC package with huge configurability.

MCMC, continued.



From pymc-devs.github.io

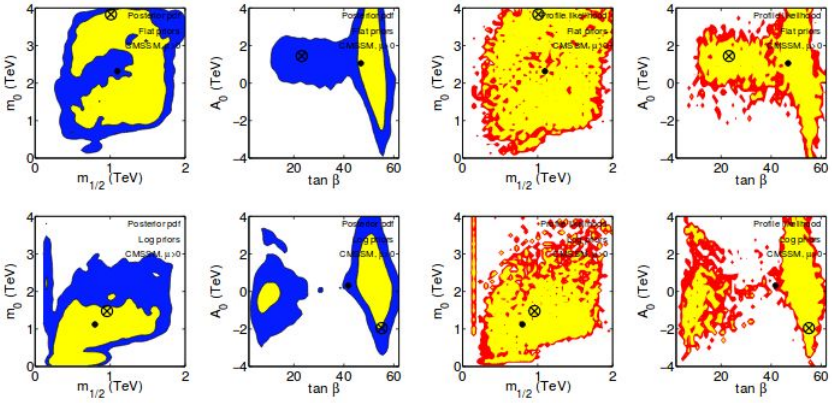


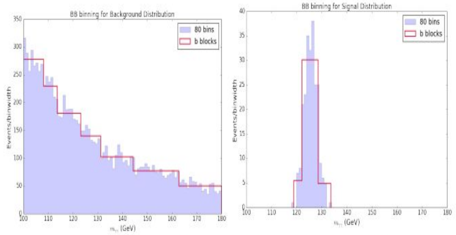
Figure 6: Posterior pdf (left two columns) and profile likelihood (right two columns) for flat priors (top row) and log priors (bottom row) for a scan including SM nuisance parameters constraints, collider limits on Higgs and superpartner masses and the WMAP5 CDM abundance determination (PHYS+NUIS+COLL+CDM). The inner and outer contours enclose respective 68% and 95% joint regions for both statistics. The posterior pdf has been smoothed with a Gaussian kernel of 1 bin width for display purposes. The cross gives the best-fit point, the filled circle is the posterior mean.

From [16]

Application - Bayesian Blocks

- Selects nonuniform bin widths for histograms.
    - Useful on log plots where be become Poisson-dominated.
- Goes through blocks, maximizes the fitness of the bin edges.
    - Fitness determined through Cash statistic $N \ln \lambda - \lambda T$, which is similar to χ^2 but works better for low counts/bin.
- A prior on the number of blocks penalizes overfitting.
- “Bayesian” because it iteratively updates the likelihood and has a prior.
- Implementation found in python package Astro-ML

FIG. 4: Fits to the background and signal distributions.



(a) Bayesian Blocks and uniform binning of the background-only toy data

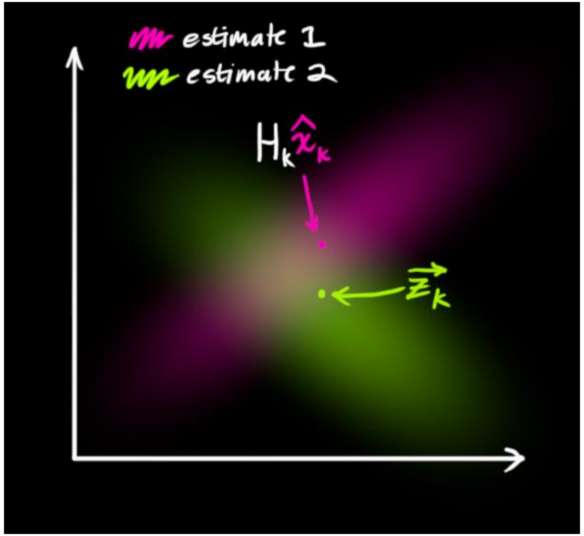
(b) Bayesian Blocks and uniform binning of the signal-only toy data. The signal shape is reduced to three bins (*i.e.*, four change points)

FIG. 5: Performance of BB algorithm for background-only and signal-only toy datasets.

Source: [15]

Application - Kalman Filter

- Imprecise phase space measurements (some uncertainty)
- Use stream of incoming data to predict next step (prior distribution).
  - Underlying Markovian process.
- Compare to reality and update the model (posterior distribution).
- Gets a better idea of state than single measurement precision alone.



Overlap of prediction and sensor estimates in a Kalman filter. Source: bzarg.org.

Summary

- Bayesian statistics uses a differing definition of probability to approach the same problems as classical statistics.
- Has intuitive interpretations of both limits and straightforward handling of nuisance parameters.
- It is subjective, at times mathematically inelegant, and fails to have coverage properties.
- Bayesians say that these aren't really problems, and frequentists have incorporated Bayesian strategies into hybrid methods.

Bibliography

1. C. Patrignani *et al.* (Particle Data Group), Chin. Phys. C, **40**, 100001 (2016) and 2017 update.
2. E. Gross. *LHC Statistics for Pedestrians*. Weizmann Institute, Rehovot, Israel.
3. Arxiv:1503.04098v1 [math.ST]
4. Arxiv:1002.1111v2 [stat.AP]
5. Doob, Joseph L. (1949). "Application of the theory of martingales". *Colloq. Intern. du C.N.R.S (Paris)*. **13**: 23–27.
6. arxiv:1301.1273v1 [physics.data-an]
7. arXiv:1404.1340 [stat.ME]
8. C. Blocker, “Interval Estimation in the Presence of Nuisance Parameters 2. Cousins and Highland Method”, CDF Statistics Committee, 2006.
9. S. Schmidt, “Limits in High Energy Physics,” Talk given at Terascale Lecture School, 2013
10. L. Demortier “Objective Bayesian Statistics for Poisson Processes,” CDF memo 2005.
11. arXiv:0901.3803 [hep-ex]
12. arXiv:hep-ex/0503039
13. arXiv:1111.6098 [hep-ph]
14. R. Cousins, “Why isn’t every physicist a Bayesian?” *American Journal of Physics Teachers* 1995.
15. arXiv: 1709.00810 [physics.data-an]
16. R. Trotta, F. Feroz, M. P. Hobson, L. Roszkowski and R. Ruiz de Austri, JHEP 0812 (2008) 024 [arXiv:0809.3792 [hep-ph]].