Physics 403

Model Selection and Parameter Estimation

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

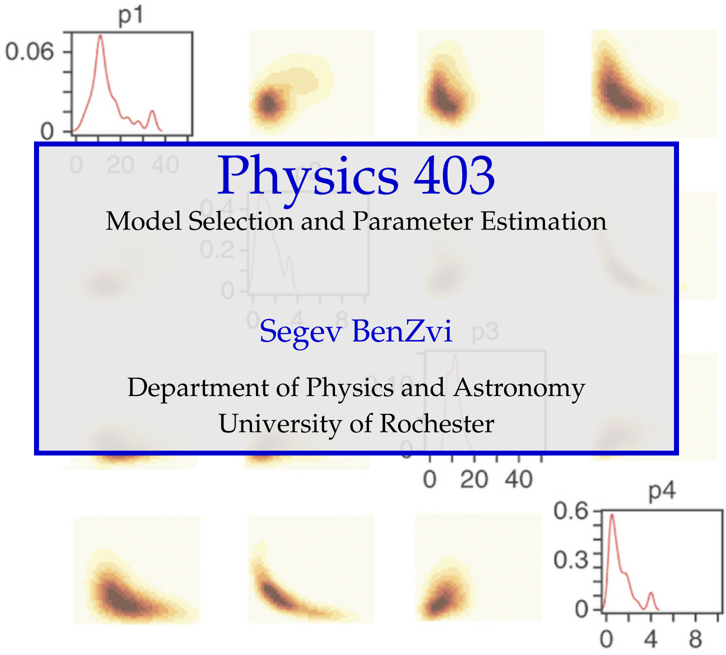


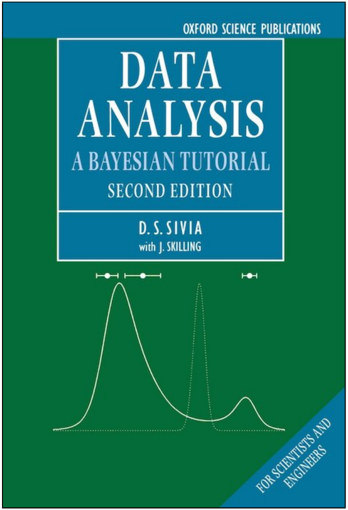
Table of Contents

1. Overview
   - Reading
   - Review of Last Class

2. Parameter Estimation: Bayesian Framework
   - Reminder of the Basics
   - Effect of the Prior
   - Marginalization
   - Comparing Models (Odds Ratio)
   - Statistical Trials
   - The Occam Factor

Reading

Sivia: Chapters 2 and 3



Last Time

Generation of pseudo-random numbers for simulation

- Simulation, data challenges, parameter estimation
- Linear Congruential Generators
- Mersenne Twister and Xorshift Generators
- Word of caution about seeding your RNG: system clock, `/dev/random`, etc.

Generating random numbers from arbitrary PDFs

- Inversion method, if PDF integrable and CDF invertible
- Acceptance/rejection method, works for most cases
- Gaussian and Poisson random numbers

Table of Contents

1. Overview
   - Reading
   - Review of Last Class

2. Parameter Estimation: Bayesian Framework
   - Reminder of the Basics
   - Effect of the Prior
   - Marginalization
   - Comparing Models (Odds Ratio)
   - Statistical Trials
   - The Occam Factor

Reminder of the Basics

Recall the basic rules of probability introduced at the start of the course:

- Sum Rule:

$$p(H|I) + p(\overline{H}|I) = 1, \qquad \sum_i p(H_i|I) = 1$$

- Product Rule:

$$p(H_i, D|I) = p(D|H_i, I)p(H_i|I) = p(H_i|D, I)p(D|I)$$

- Bayes' Theorem:

$$p(H_i|D, I) = \frac{p(D|H_i, I)p(H_i|I)}{p(D|I)}$$

- Law of total probability:

$$\sum_i p(H_i|D, I) = \frac{\sum_i p(D|H_i, I)p(H_i|I)}{p(D|I)} = 1$$

$$\therefore p(D|I) = \sum_i p(D|H_i, I)p(H_i|I)$$

Reminder of the Basics

The law of total probability has a continuous counterpart. For example, given a model M with parameters $\boldsymbol{\theta}$,

$$p(D|M) = \int_V d\boldsymbol{\theta}\, p(D|\boldsymbol{\theta}, M)\, p(\boldsymbol{\theta}|M)$$

Interpretation: the likelihood of model M is the weighted average likelihood for its parameters $\boldsymbol{\theta}$.

Parameter Estimation: the determination of the values of model parameters $\boldsymbol{\theta}$ using data.

- Bayesian: evaluate the full posterior PDF $p(\boldsymbol{\theta}|D, M)$ or “best fit” summary values like the mean or mode. Uses prior $p(\boldsymbol{\theta}|M)$
- Frequentist: evaluate the best fit values from the likelihood alone
- Both approaches: give some allowed range of parameter with some probability measure (confidence interval, or credible range)

Effect of the Prior

- The presence of a prior may be a point of contention, because you can get different answers depending on the prior you choose.
- Bayesian answer: yes. So?
- The prior is how we incorporate external information about the quantities being tested
- If the posterior PDF is dominated by the prior, that just means the data are not constraining our model parameters
- **Note**: frequentists don't use priors, which in practice means that assumptions are hidden
- Best practice: report likelihoods and priors separately, and show the effect of different priors on the posterior

Coin Flipping

Example

From Sivia, Ch. 2 [1]: we walk into a casino and start betting on the outcomes of flipping a coin. (It's not a very impressive casino.)

- We don't know the probability h of getting heads, so we have to choose some $p(h|I)$.
- We do know that given h, the probability of observing heads r times in N coin flips is given by the binomial PDF

$$p(r|N,h,I) \propto h^r (1-h)^{N-r}$$

What is the effect of $p(h|I)$ on the posterior probability $p(h|N,r,I)$, the distribution of h given r heads in N tosses? From Bayes' Theorem,

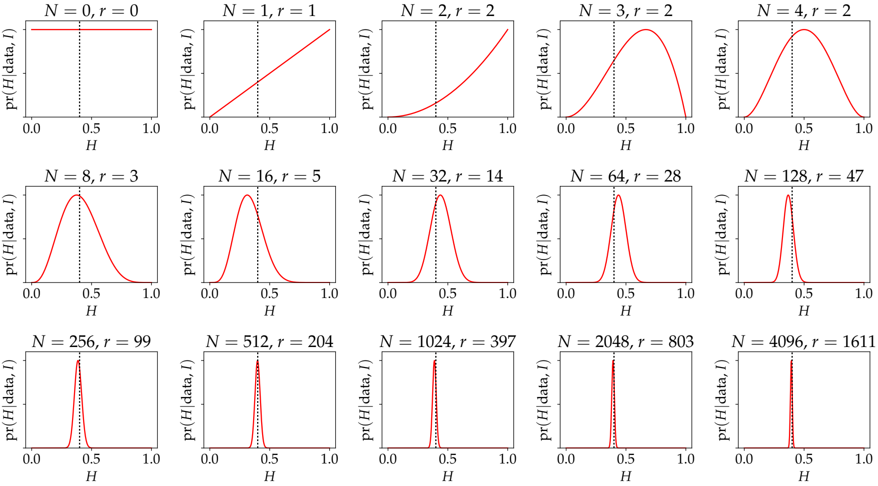
$$p(h|N,r,I) \propto p(r|h,N,I)\, p(h|I),$$

so let's try out different priors and see what happens.

Coin Flipping

Uniform "Ignorance" Prior

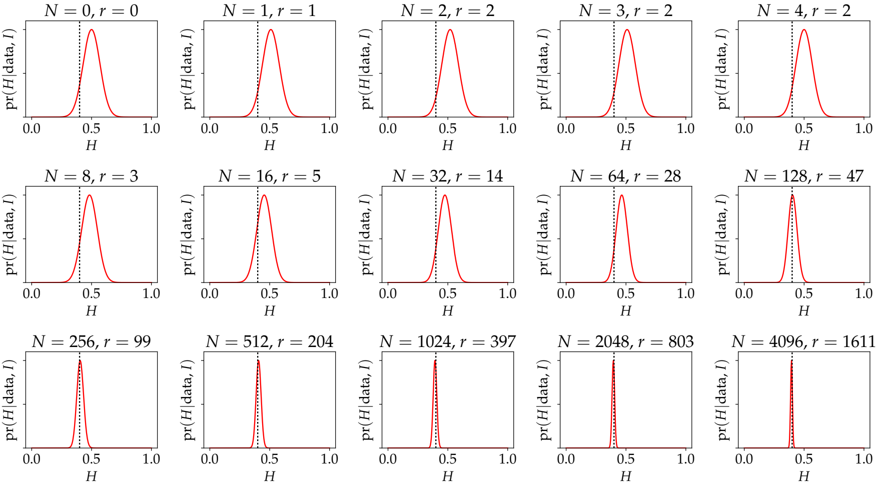
We start with no preferred value for h:



Coin Flipping

Fair Coin Prior

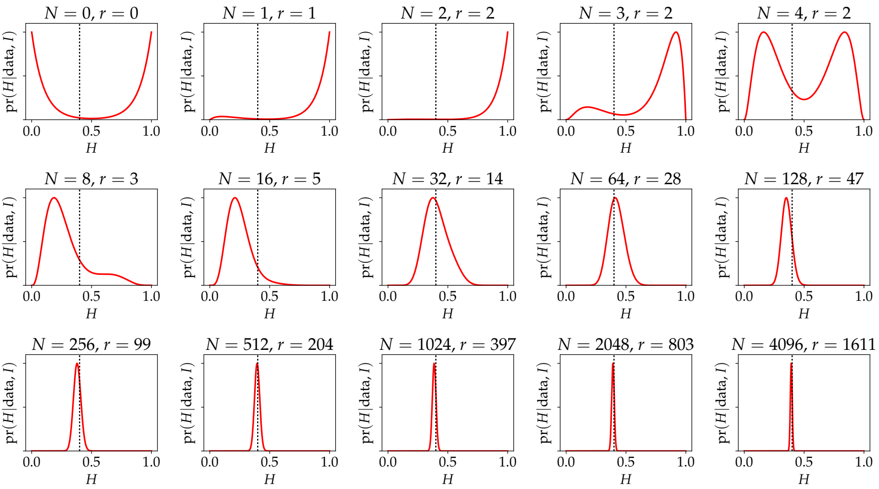
We assume the coin is fair $(h = 0.5)$ with some uncertainty:



Coin Flipping

Unfair Coin Prior

We assume the coin is very unfair, but don’t know the bias.



Marginalization

- Recall the definition of marginalization and marginal distributions: if we don't care about the effect of some parameter on a probability, we can integrate it out
- Example: for model M with parameters θ, φ, if we are only interested in θ then we can calculate the marginal PDF

$$p(\theta|D, M) = \int d\varphi \, p(\theta, \varphi|D, M)$$

- Marginalization is a general technique in Bayesian analysis that doesn't have an analog in frequentist statistics
- Terms that we don't care about are called nuisance parameters in frequentist statistics. There is no general procedure for handling them

Model Comparison

- One topic we haven't discussed yet is model comparison
- The idea: compare two competing models by calculating the probability of each model given the data D
- If we want to compare two or more alternative models M_i, then use Bayes' Theorem to calculate the posterior probability of each model:

$$p(M_i|D,I) = \frac{p(D|M_i,I)p(M_i|I)}{p(D|I)}$$

- This is analogous to parameter estimation, except instead of estimating $p(\theta|D,I)$ for a parameter, we estimate $p(M_i|D,I)$ for a model
- The math is the same, but the interpretation differs

The Odds Ratio

To select between two models, it is useful to calculate the ratio of the posterior probabilities of the models. This is called the odds ratio:

$$
\begin{aligned}
O_{ij} &= \frac{p(D|M_i,I)}{p(D|M_j,I)}\frac{p(M_i|I)}{p(M_j|I)} \\
&= B_{ij}\frac{p(M_i|I)}{p(M_j|I)}
\end{aligned}
$$

The first term is called the Bayes Factor [2, 3] and the second is called the prior odds ratio. Interpration:

- **Prior odds**: the amount by which you favor M_i over M_j *before taking data*. There is no analog in frequentist statistics.
- **Bayes Factor**: the amount that the data D causes you favor M_i over M_j. Frequentist analog: *likelihood ratio* (but frequentists can't marginalize nuisance parameters)

The Odds Ratio

Interpreting the Bayes Factor, according to Jeffreys [2]:

B_{ij}	**Strength of Evidence**
$< 1:1$	negative (supports M_j)
$1:1$ to $3:1$	barely worth mentioning
$3:1$ to $10:1$	substantial support for M_i
$10:1$ to $30:1$	strong support for M_i
$30:1$ to $100:1$	very strong support for M_i
$> 100:1$	decisive evidence for M_i

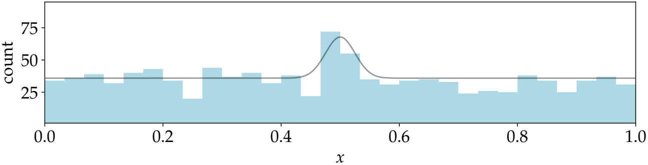
But wait, remember the "5σ rule?" That corresponds to a Gaussian *tail probability* (or **p-value**) of 6×10^{-7}. Isn't that MUCH stronger evidence than $100:1$ odds. What's going on?

Partial answer: odds ratios and p-values are not the same thing. Not to mention the "look elsewhere effect" and other sources of statistical trials

Aside: Statistical Trials

The Look Elsewhere Effect

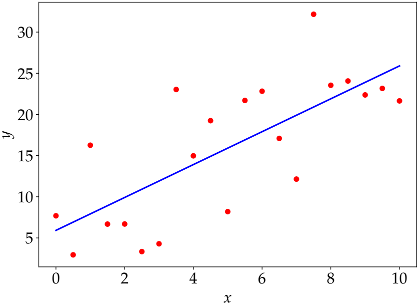
- Suppose you are looking for a spike in some data with background, e.g., a mass resonance or a spectral line, but you don't know the location of the feature, just a range of interest
- You scan over the data and find a spike which is $> 3\sigma$ above the background (p-value: $\sim 0.1\%$). Is this significant?

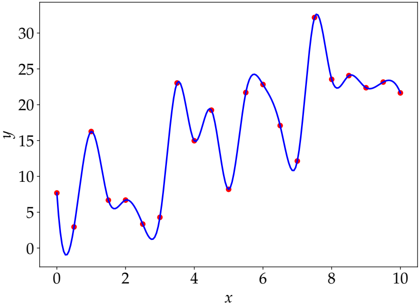


- Hang on: because location was a free parameter, you need to account for the fact that any one of the bins you looked at could have been an upward fluctuation of the background
- Conservatively, $p \to N_{\text{bins}} \times p \approx 2\%$, or $\sim 2\sigma$

Overfitting

In model selection, we want to find the model that best fits the data. By “best fit” we have some notion of minimizing the distance (by some measure) between the data and model.





The model on the right clearly has a smaller “distance” from the data than the one on the left. But is it a better model?

Occam’s Razor

Parsimony as a Problem Solving Principle

William of Ockham:

> *All things being equal, the simplest solution tends to be the best one.*

A. Einstein (paraphrased):

> *Everything should be kept as simple as possible, but no simpler.*

Question: Is there a quantifiable way of defining simplicity when choosing between two models?

Can you think of various criteria one might choose?

Statistical Trials in the Bayesian Framework

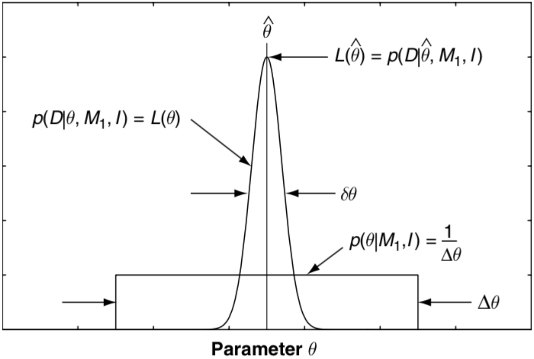
Occam's Razor

- Occam's Razor: when selecting from among competing models, generally prefer the simpler model
- For model comparison, the Bayes Factor and odds ratio have a built-in Occam's razor
- Searching for a spike: in the Bayesian framework, we would treat the location of the spike as a nuisance parameter and marginalize it (model M_1)
- Compare this to a model with no spike (M_0)
- If we did everything correctly, $p(D|M_1)$ should have extra terms compared to $p(D|M_0)$ which "penalize" it for our ignorance of the location of the spike
- So a piece of the odds ratio should account for statistical trials and favor the simpler model!

Occam's Razor

The Bayesian Framework

Let's be more explicit. Imagine M_1 has a single parameter θ (e.g., the location of a spike) which is unknown. M_0 has θ fixed at θ_0.


$\hat{\theta}$
$L(\hat{\theta}) = p(D|\hat{\theta}, M_1, I)$
$p(D|\theta, M_1, I) = L(\theta)$
$\delta\theta$
$p(\theta|M_1, I) = \frac{1}{\Delta\theta}$
$\Delta\theta$
Parameter θ

Suppose our prior on θ is uniform in model M_1. I.e., we don't know what it is, just that it lies in some range $\Delta\theta$. And suppose the data tell us a lot about θ, so $p(D|\theta, M_1, I)$ is very peaked about $\hat{\theta}$ with width $\delta\theta$.

Occam's Razor

The Bayesian Framework

The "global likelihood" of the data given M_1 (independent of θ) is

$$\begin{aligned}
p(D|M_1,I) &= \int d\theta\, p(D|\theta,M_1,I)\, p(\theta|M_1,I) \\
&= \int d\theta\, p(D|\theta,M_1,I)\, \frac{1}{\Delta\theta} \\
&\approx p(D|\hat{\theta},M_1,I)\, \delta\theta \frac{1}{\Delta\theta}
\end{aligned}$$

Since M_0 has no free parameters, its global likelihood is

$$\begin{aligned}
p(D|M_0,I) &= \int d\theta\, p(D|\theta,M_1,I)\, \delta(\theta-\theta_0) \\
&= p(D|\theta_0,M_1,I)
\end{aligned}$$

I.e., it's just the likelihood of model M_1 with θ fixed.

Occam's Razor

The Bayesian Framework

Putting it all together, the Bayes factor in favor of the more complex model M_1 is

$$\begin{aligned} B_{10} &\approx \frac{p(D|\hat{\theta}, M_1, I)}{p(D|\theta_0, M_1, I)} \frac{\delta\theta}{\Delta\theta} \\ &= \frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\theta_0)} \frac{\delta\theta}{\Delta\theta} \end{aligned}$$

The first term is a likelihood ratio, which favors M_1 because of the strong peak at $\hat{\theta}$.

But the second term penalizes M_1 since $\delta\theta < \Delta\theta$. In other words, M_1 is penalized because of the wasted parameter space that gets ruled out by the data.

The Occam Factor

- Generalizing from this specific problem, we can express any likelihood of data D given a model M as the maximum value of its likelihood times an Occam factor:

$$p(D|M,I) = \mathcal{L}_{\text{max}}\Omega_{\theta}$$

- The Occam factor corrects the likelihood for the statistical trials incurred by scanning the parameter space for $\hat{\theta}$.
- The odds ratio automatically accounts for these factors. It is in this way that the Bayesian framework prevents overfitting of data with arbitrarily complicated models.
- **Note**: in frequentist statistics, statistical penalties are more of a kluge. There are many ways to calculate them (e.g., the N_{bins} factor used earlier) but no simple framework.

Summary

- The formalism for parameter estimation and model selection in Bayesian statistics is mathematically the same
- We estimate parameters by looking at the PDF and its maximum likelihood (same as frequentist approach)
- We perform model selection by computing an odds ratio and making a decision about the odds. In frequentist approach: a likelihood ratio test, or Neyman-Pearson test
- The odds ratio has a built-in Occam factor that accounts for "scanning" for the best value in a parameter space
- Marginalization gives us a uniform way of handling unknown *nuisance parameters*, including systematic uncertainties

References I

[1] D.S. Sivia and John Skilling. *Data Analysis: A Bayesian Tutorial*. New York: Oxford University Press, 1998.

[2] Harold Jeffreys. *The Theory of Probability*. 3rd ed. Oxford, 1961.

[3] Robert E. Kass and Adrian E. Raftery. “Bayes Factors”. In: *J. Am. Stat. Assoc.* 90.430 (1995), pp. 773–795. URL: `http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572`.